A **CRC Press** FREEBOOK

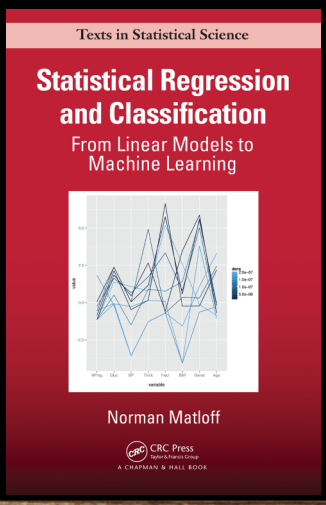
Linear Models and Extensions

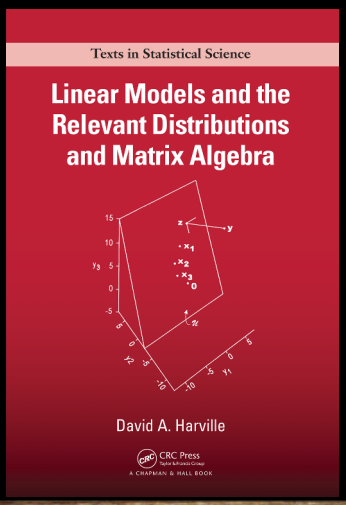
A CRC Press FreeBook

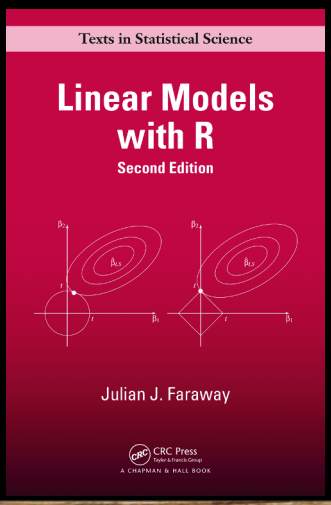
CRC Press
Taylor & Francis Group

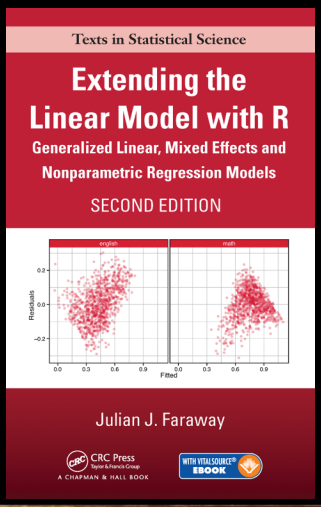
TABLE OF CONTENTS

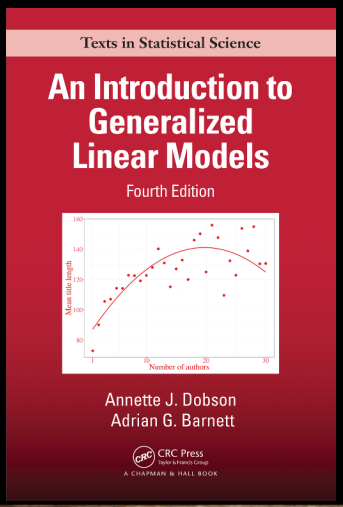
READ THE LATEST ON LINEAR MODELS AND EXTENSIONS WITH THESE KEY TITLES

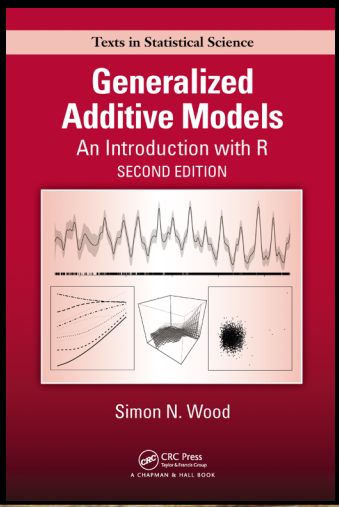












VISIT WWW.CRCPRESS.COM

TO BROWSE OUR FULL RANGE OF TITLES

Introduction

A linear model describes quantitative response in terms of a linear combination of predictors. You can use a linear model to make predictions or explain the relationship between the response and the predictors. Linear models and their extensions, i.e. generalized linear models (GLMs), generalized additive models, etc. are very flexible and widely used in applications in physical science, engineering, social science and business. Linear models are part of the core of Statistics and understanding them well is crucial to a broader competence in the practice of statistics.1

Linear Models are now taught in every graduate program in statistics, sometimes at undergraduate level, and increasingly as a standard course in other fields. CRC Press is the leading publisher of textbooks for Linear Models and their Extensions, with books aimed specifically at statistics graduate students, students in the physical sciences, and self-studying practitioners. This CRC Press Freebook presents six chapters from some of our leading textbooks in the field.

Setting the Stage (Chapter 1) from Statistical Regression and Classification: From Linear Models to Machine Learning

Statistical Regression and Classification: From Linear Models to Machine Learning takes an innovative look at the traditional statistical regression course, presenting a contemporary treatment in line with today's applications and users. Though some statistical learning methods are introduced, the primary methodology used is linear and generalized linear parametric models, covering both the Description and Prediction goals of regression methods. Chapter 1 sets the stage for the book, previewing many of the major concepts to be presented in later chapters.

Estimation and Prediction: Classical Approach (Chapter 5) from Linear Models and the Relevant Distributions and Matrix Algebra

This book provides in-depth and detailed coverage of the use of linear statistical models as a basis for parametric and predictive inference. Under a linear model, the data are regarded as the observed values of random variables, and the expected values of these random variables are linear combinations of a parametric vector β. A linear combination of the elements of β is either estimable or nonestimable. According to the Gauss-Markov theorem, the least squares estimator of an estimable linear combination is (under certain conditions) the best linear unbiased estimator; it is also the best linear translation-equivariant estimator. Least squares estimates can be computed from a solution to the normal equations or conjugate normal equations; alternatively, they can be computed from a QR decomposition of the model matrix.

Introduction

Insurance Redlining – A Complete Example (Chapter 12) from Linear Models with R, 2e

Like its widely praised, best-selling predecessor, this edition combines statistics and R to seamlessly give a coherent exposition of the practice of linear modeling. The text offers up-to-date insight on essential data analysis topics, from estimation, inference, and prediction to missing data, factorial models, and block designs. Numerous examples illustrate how to apply the different methods using R.

This chapter presents a relatively complete data analysis. The example is interesting because it illustrates several of the ambiguities and difficulties encountered in statistical practice.

Random Effects (Chapter 10) from Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, 2e

Since the publication of the bestselling, highly recommended first edition, R has considerably expanded both in popularity and in the number of packages available. This revision takes advantage of the greater functionality now available in R and substantially revises and adds several topics.

This chapter looks at random effects, which are random variables. In contrast, a fixed effect is an unknown constant that one tries to estimate from the data. It does not make sense to estimate a random effect; instead, one tries to estimate the parameters that describe the distribution of this random effect.

Poisson Regression and Log-Linear Models (Chapter 9) from An Introduction to Generalized Linear Models, 4e

This new edition of a bestseller has been updated with new sections on non-linear associations, strategies for model selection, and good statistical practice. Like its predecessor, this edition presents the theoretical background of generalized linear models (GLMs) before focusing on methods for analyzing particular kinds of data. Using popular statistical software programs, this concise and accessible text illustrates practical approaches to estimation, model fitting, and model comparisons.

This chapter gives an overview of Poisson regression and log-linear models. It includes numerical examples to illustrate the concepts and methods, including model checking and inference.

Introducing GAMs (Chapter 4) from Generalized Additive Models: An Introduction with R, 2e

Established as one of the leading references on generalized additive models (GAMs), this is the only book on the topic to be introductory in nature with a wealth of practical examples and software implementation. It is self-contained, providing the necessary background in linear models, linear mixed models, and generalized linear models (GLMs), before presenting a balanced treatment of the theory and applications of GAMs and related models. Use of R software helps explain the theory and illustrates the practical application of the methodology.

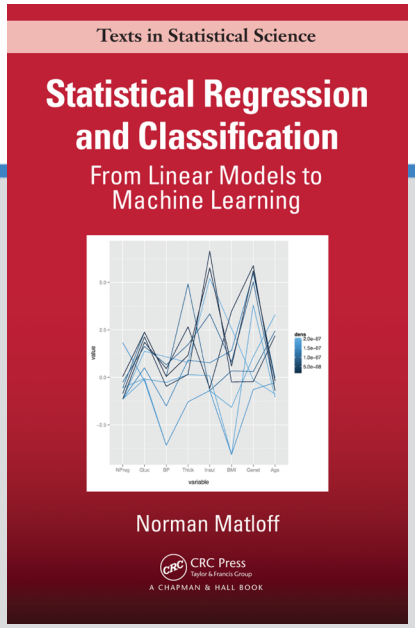
This chapter introduces GAMs by having the reader 'build their own' GAM using R. This provides a useful way of quickly gaining a rather solid familiarity with the fundamentals of the GAM framework presented in the book. Once the basic framework is mastered, the theory and examples in later chapters can be used to fill in the details and apply the methods.

1. Faraway - Linear Models with R. Chapman & Hall/CRC (2015)

CHAPTER

1

SETTING THE STAGE



Chapter 1

Setting the Stage

This chapter will set the stage for the book, previewing many of the major concepts to be presented in later chapters. The material here will be referenced repeatedly throughout the book.

1.1 Example: Predicting Bike-Sharing Activity

Let's start with a well-known dataset, **Bike Sharing**, from the Machine Learning Repository at the University of California, Irvine.[1] Here we have daily/hourly data on the number of riders, weather conditions, day-of-week, month and so on. Regression analysis, which relates the mean of one variable to the values of one or more other variables, may turn out to be useful to us in at least two ways:

- **Prediction:**

  The managers of the bike-sharing system may wish to predict ridership, say for the following question:

  > Tomorrow, Sunday, is expected to be sunny and cool, say 62 degrees Fahrenheit. We may wish to predict the number of riders, so that we can get some idea as to how many bikes will need repair. We may try to predict ridership, given the

[1] Available at *https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset*.

weather conditions, day of the week, time of year and so on.

Some bike-sharing services actually have trucks to move large numbers of bikes to locations that are expected to have high demand. Prediction would be even more useful here.

- **Description:**

  We may be interested in determining what factors affect ridership. How much effect, for instance, does wind speed have in influencing whether people wish to borrow a bike?

These twin goals, Prediction and Description, will arise frequently in this book. Choice of methodology will often depend on the goal in the given application.

1.2 Example of the Prediction Goal: Body Fat

Prediction is difficult, especially about the future — baseball great, Yogi Berra

The great baseball player Yogi Berra was often given to malapropisms, one of which supposedly was the quote above. But there is more than a grain of truth to this, because indeed we may wish to "predict" the present or even the past.

For example, consiser the **bodyfat** data set, available in the R package, **mfp**, available on CRAN [5]. (See Section 1.20.1 for information on CRAN packages, a number of which will be used in this book.) Direct measurment of body fat is expensive and unwieldy, as it involves underwater weighing. Thus it would be highly desirable to "predict" that quantity from easily measurable variables such as height, age, weight, abdomen circumference and so on.

In scientific studies of ancient times, there may be similar situations in which we "predict" past unknown quantities from present known ones.

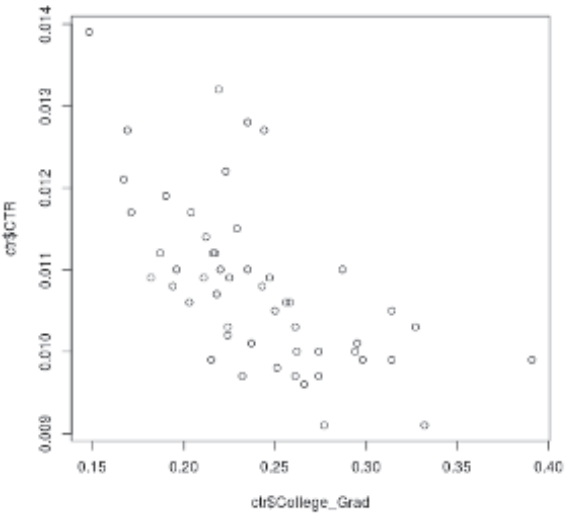


Figure 1.1: Click rate vs. college rate

1.3 Example: Who Clicks Web Ads?

One of the most common applications of machine learning methods is in marketing. Sellers wish to learn which types of people might be interested in a given product. The reader is probably familiar with Amazon's *recommender system*, in which the viewer who indicates interest in a given book, say, is shown a list of similar books.[2]

We will discuss recommender systems at several points in this book, beginning with Section 3.2.4. A more general issue is the *click-through rate* (CTR), meaning the proportion of viewers of a Web page who click on a particular ad on the page. A simple but very engaging example was discussed online [53]. The data consist of one observation per state of the U.S.[3] There was one predictor, the proportion of college graduates in the state, and a response variable, the CTR.

[2] As a consumer, I used to ignore these, but now with the sharp decline in the number of bricks-and-mortar bookstores which I could browse, I now often find Amazon's suggestions useful.

[3] We use the classical statistical term **observation** here, meaning a single data point, in this case data for a single state. In the machine learning community, it is common to use the term **case**.

A plot of the data, click rate vs. college rate, is in Figure 1.1. There definitely seems to be something happening here, with a visible downward trend to the points. But how do we quantify that? One approach to learning what relation, if any, educational level has to CTR would be to use regression analysis. We will see how to do so in Section 1.8.

1.4 Approach to Prediction

Even without any knowledge of statistics, many people would find it reasonable to predict via subpopulation means. In the above bike-sharing example, say, this would work as follows.

Think of the "population" of all days, past, present and future, and their associated values of number of riders, weather variables and so on.[4] Our data set is considered a sample from this population. Now consider the subpopulation consisting of all days with the given conditions: Sundays, sunny skies and 62-degree temperatures.

It is intuitive that:

> A reasonable prediction for tomorrow's ridership would be the mean ridership among all days in the subpopulation of Sundays with sunny skies and 62-degree temperatures.

In fact, such a strategy is optimal, in the sense that it minimizes our expected squared prediction error, as discussed in Section 1.19.3 of the Mathematical Complements section at the end of this chapter. But what is important for now is to note that in the above prediction rule, we are dealing with a *conditional* mean: Mean ridership, *given* day of the week is Sunday, skies are sunny, and temperature is 62.

Note too that we can only calculate an *estimated* conditional mean. We wish we had the true population value, but since our data is only a sample, we must always keep in mind that we are just working with estimates.

[4] This is a somewhat slippery notion, because there may be systemic differences from the present and the distant past and distant future, but let's suppose we've resolved that by limiting our time range.

1.5 A Note about E(), Samples and Populations

To make this more mathematically precise, note carefully that in this book, as with many other books, the *expected value* functional $E()$ refers to the population mean. Say we are studying personal income, I, for some population, and we choose a person at random from that population. Then $E(I)$ is not only the mean of that random variable, but much more importantly, it is the mean income of all people in that population.

Similarly, we can define conditional means, i.e., means of subpopulations. Say G is gender. Then the conditional expected value, $E(I \mid G = \text{ male})$ is the mean income of all men in the population.

To illustrate this in the bike-sharing context, let's define some variables:

- R, the number of riders
- W, the day of the week
- S, the sky conditions, e.g., sunny
- T, the temperature

We would like our prediction Q to be the conditional mean,

$$Q = E(R \mid W = \text{ Sunday}, S = \text{ sunny}, T = 62) \tag{1.1}$$

There is one major problem, though: We don't know the value of the right-hand side of (1.1). All we know is what is in our sample data, whereas the right-side of (1.1) is a population value, and thus unknown.

The difference between sample and population is of course at the very core of statistics. In an election opinion survey, for instance, we wish to know p, the proportion of people in the population who plan to vote for Candidate Jones. But typically only 1200 people are sampled, and we calculate the proportion of Jones supporters among them, $\widehat{p}$, using that as our estimate of p. (Note that the "hat" notation $\widehat{\ }$ is the traditional one for "estimate of.") This is why the news reports on these polls always include the *margin of error*.[5]

[5] This is actually the radius of a 95% confidence interval for p.

Similarly, though we would like to know the value of E(R | W = Sunday, S = sunny, T = 62), **it is an unknown population value, and thus must be estimated from our sample data**, which we'll do later in this chapter.

Readers will greatly profit from constantly keeping in mind this distinction between populations and samples.

Another point is that in statistics, the populations are often rather conceptual in nature. On the one hand, in the election poll example above, there is a concrete population involved, the population of all voters. On the other hand, consider the bike rider data in Section 1.1. Here we can think of our data as being a sample from the population of all bikeshare users, past, present and future.

Before going on, a bit of terminology, again to be used throughout the book: We will refer to the quantity to be predicted, e.g., R above, as the *response variable*, and the quantities used in prediction, e.g., W, S and T above, as the *predictor variables*. Other popular terms are *dependent variable* for the response and *independent variables* or *regressors* for the predictors. The machine learning community uses the term *features* rather than *predictors*.

1.6 Example of the Description Goal: Do Baseball Players Gain Weight As They Age?

Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less — Marie Curie

Though the bike-sharing data set is the main example in this chapter, it is rather sophisticated for introductory material. Thus we will set it aside temporarily, and bring in a simpler data set for now. We'll return to the bike-sharing example in Section 1.15.

This new dataset involves 1015 major league baseball players, courtesy of the UCLA Statistics Department. You can obtain the data as the data set **mlb** in **freqparcoord**, a CRAN package authored by Yingkang Xie and myself [104].[6] The variables of interest to us here are player weight W, height H and age A, especially the first two.

[6] We use the latter version of the dataset here, in which we have removed the Designated Hitters.

Here are the first few records:

```
> library(freqparcoord)
> data(mlb)
> head(mlb)
                Name Team          Position Height
1   Adam_Donachie  BAL           Catcher     74
2        Paul_Bako  BAL           Catcher     74
3 Ramon_Hernandez  BAL           Catcher     72
4    Kevin_Millar  BAL  First_Baseman        72
5      Chris_Gomez  BAL  First_Baseman        73
6   Brian_Roberts  BAL Second_Baseman        69
  Weight   Age PosCategory
1    180 22.99    Catcher
2    215 34.69    Catcher
3    210 30.78    Catcher
4    210 35.43  Infielder
5    188 35.71  Infielder
6    176 29.39  Infielder
```

1.6.1 Prediction vs. Description

Recall the Prediction and Description goals of regression analysis, discussed in Section 1.1. With the baseball player data, we may be more interested in the Description goal, such as:

> Ahtletes strive to keep physically fit. Yet even they may gain weight over time, as do people in the general population. To what degree does this occur with the baseball players? This question can be answered by performing a regression analysis of weight against height and age, which we'll do in Section 1.9.1.2.[7]

On the other hand, there doesn't seem to be much of a Prediction goal here. It is hard to imagine much need to predict a player's weight. One example of this, though, is working with missing data, in which we wish to predict any value that might be unavailable.

However, for the purposes of explaining the concepts, we will often phrase things in a Prediction context. In the baseball player example, it will turn

[7] The phrasing here, "regression analysis of ... against ...," is commonly used in this field. The quantity before "against" is the response variable, and the ones following are the predictors.

out that by trying to predict weight, we can deduce effects of height and age. In particular, we can answer the question posed above concerning weight gain over time.

So, suppose we will have a continuing stream of players for whom we only know height (we'll bring in the age variable later), and need to predict their weights. Again, we will use the conditional mean to do so. For a player of height 72 inches, for example, our prediction might be

$$\widehat{W} = E(W \mid H = 72) \tag{1.2}$$

Again, though, this is a population value, and all we have is sample data. How will we estimate $E(W \mid H = 72)$ from that data?

First, some important notation: Recalling that μ is the traditional Greek letter to use for a population mean, let's now use it to denote a function that gives us subpopulation means:

> For any height t, define
>
> $$\mu(t) = E(W \mid H = t) \tag{1.3}$$
>
> which is the mean weight of all people in the population who are of height t.
>
> Since we can vary t, this is indeed a function, and it is known as *the regression function of W on H*.

So, $\mu(72.12)$ is the mean population weight of all players of height 72.12, $\mu(73.88)$ is the mean population weight of all players of height 73.88, and so on. These means are population values and thus unknown, but they do exist.

So, to predict the weight of a 71.6-inch-tall player, we would use $\mu(71.6)$ — if we knew that value, which we don't, since once again this is a population value while we only have sample data. So, we need to estimate that value from the (height, weight) pairs in our sample data, which we will denote by $(H_1, W_1), ...(H_{1015}, W_{1015})$. How might we do that? In the next two sections, we will explore ways to form our estimate, $\widehat{\mu}(t)$. (Keep in mind that for now, we are simply exploring, especially in the first of the following two sections.)

1.6.2 A First Estimator

Our height data is only measured to the nearest inch, so instead of estimating values like $\mu(71.6)$, we'll settle for $\mu(72)$ and so on. A very natural estimate for $\mu(72)$, again using the "hat" symbol to indicate "estimate of," is the mean weight among all players in our sample for whom height is 72, i.e.

$$\widehat{\mu}(72) = \text{ mean of all } W_i \text{ such that } H_i = 72 \tag{1.4}$$

R's **tapply()** can give us all the $\widehat{\mu}(t)$ at once:

```
> library(freqparcoord)
> data(mlb)
> muhats <- tapply(mlb$Weight,mlb$Height,mean)
> muhats
      67       68       69       70       71       72
172.5000 173.8571 179.9474 183.0980 190.3596 192.5600
      73       74       75       76       77       78
196.7716 202.4566 208.7161 214.1386 216.7273 220.4444
      79       80       81       82       83
218.0714 237.4000 245.0000 240.5000 260.0000
```

In case you are not familiar with **tapply()**, here is what just happened. We asked R to partition the Weight variable into groups according to values of the Height variable, and then compute the mean weight in each group. So, the mean weight of people of height 72 in our sample was 192.5600. In other words, we would set $\widehat{\mu}(72) = 192.5600$, $\widehat{\mu}(74) = 202.4566$, and so on. (More detail on **tapply()** is given in the Computational Complements section at the end of this chapter.)

Since we are simply performing the elementary statistics operation of estimating population means from samples, we can form confidence intervals (CIs). For this, we'll need the "n" and sample standard deviation for each height group:

```
> tapply(mlb$Weight,mlb$Height,length)
 67  68  69  70  71  72  73  74  75  76  77  78
  2   7  19  51  89 150 162 173 155 101  55  27
 79  80  81  82  83
 14   5   2   2   1
> tapply(mlb$Weight,mlb$Height,sd)
      67       68       69       70       71       72
10.60660 22.08641 15.32055 13.54143 16.43461 17.56349
```

```
      73        74        75        76        77        78
16.41249  18.10418  18.27451  19.98151  18.48669  14.44974
      79        80        81        82        83
28.17108  10.89954  21.21320  13.43503        NA
```

Here is how that first call to **tapply()** worked. Recall that this function partitions the data by the Height variables, resulting in a weight vector for each height value. We need to specify a function to apply to each of the resulting vectors, which in this case we choose to be R's **length()** function. The latter then gives us the count of weights for each height value, the "n" that we need to form a CI. By the way, the NA value is due to there being only one player with height 83, which is makes life impossible for **sd()**, as it divides from "n-1."

An approximate 95% CI for $\mu(72)$, for example, is then

$$190.3596 \pm 1.96 \frac{17.56349}{\sqrt{150}} \tag{1.5}$$

or about (187.6,193.2).

The above analysis takes what is called a *nonparametric* approach. To see why, let's proceed to a parametric one, in the next section.

1.6.3 A Possibly Better Estimator, Using a Linear Model

All models are wrong, but some are useful — famed statistician George Box

[In spite of] innumerable twists and turns, the Yellow River flows east — Confucious

So far, we have assumed nothing about the shape that $\mu(t)$ would have, if it were plotted on a graph. Again, it is unknown, but the function does exist, and thus it does correspond to *some* curve. But we might consider making an assumption on the shape of this unknown curve. That might seem odd, but you'll see below that this is a very powerful, intuitively reasonable idea.

Toward this end, let's plot those values of $\widehat{\mu}(t)$ we found above. We run

```
> plot(67:83,muhats)
```

producing Figure 1.2.

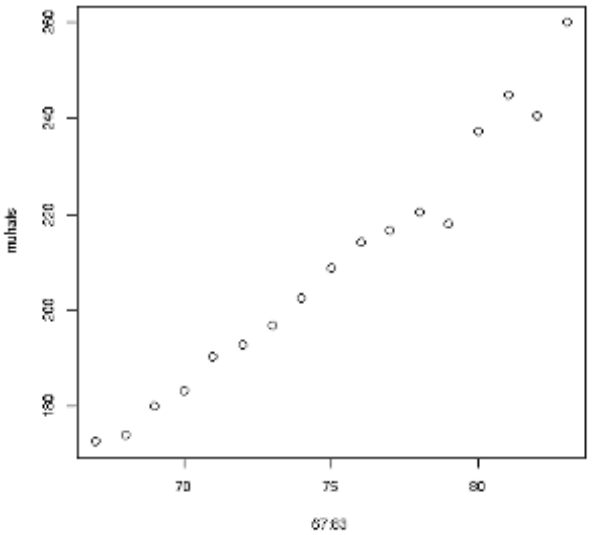


Figure 1.2: Plotted $\widehat{\mu}(t)$

Interestingly, the points in this plot seem to be near a straight line. Just like the quote of Confucious above concerning the Yellow River, visually we see something like a linear trend, in spite of the "twists and turns" of the data in the plot. This suggests that our unknown function $\widehat{\mu}(t)$ has a linear form, i.e., that

$$\mu(t) = c + dt \tag{1.6}$$

for some constants c and d, over the range of t appropriate to human heights. Or, in English,

$$\text{mean weight} = c + d \times \text{ height} \tag{1.7}$$

Don't forget the word *mean* here! We are assuming that the *mean* weights in the various height subpopulations have the form (1.6), NOT that weight itself is this function of height, which can't be true.

This is called a *parametric* model for $\mu(t)$, with parameters c and d. We will use this below to estimate $\mu(t)$. Our earlier estimation approach, in

Section 1.6.2, is called *nonparametric*. It is also called *assumption-free* or *model-free*, since it made no assumption at all about the shape of the $\mu(t)$ curve.

Note the following carefully:

- Figure 1.2 suggests that our straight-line model for $\mu(t)$ may be less accurate at very small and very large values of t. This is hard to say, though, since we have rather few data points in those two regions, as seen in our earlier R calculations; there is only one person of height 83, for instance.

  But again, in this chapter we are simply exploring, so let's assume for now that the straight-line model for $\widehat{\mu}(t)$ is reasonably accurate. We will discuss in Chapter 6 how to assess the validity of this model.

- Since $\mu(t)$ is a population function, the constants c and d are population values, thus unknown. However, we can estimate them from our sample data. We do so using R's **lm()** ("linear model") function:[8]

```
> lmout <- lm(mlb$Weight ~ mlb$Height)
> lmout
Call:
lm(formula = mlb$Weight ~ mlb$Height)

Coefficients:
(Intercept)    mlb$Height
    -151.133         4.783
```

This gives $\widehat{c} = -151.133$ and $\widehat{d} = 4.783$. We can superimpose the fitted line to Figure 1.2, using R's **abline()** function, which adds a line with specified slope and intercept to the currently-displayed plot:

```
> abline(coef=coef(lmout))
```

The result is shown in Figure 1.3.

Note carefully that we do not expect the line to fit the points exactly. On the contrary, the line is only an estimate of $\mu(t)$, the condiitonal *mean* of weight given height, not weight itself.

[8] Details on how the estimation is done will be given in Chapter 2.

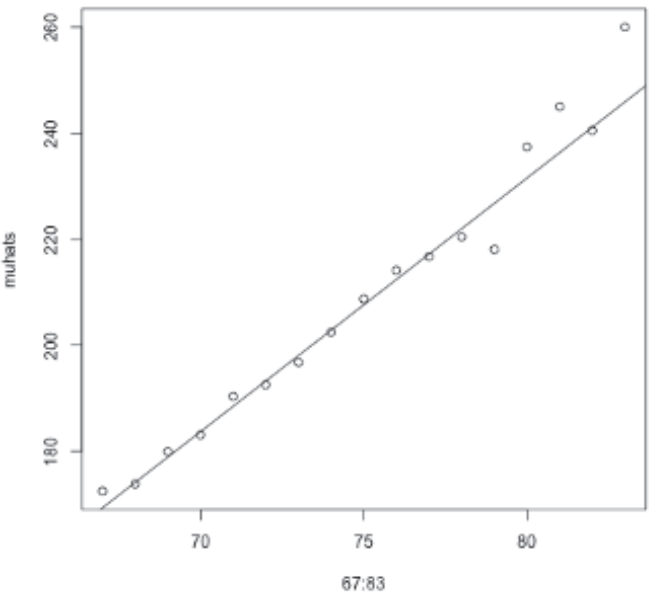


Figure 1.3: Turkish student evaluations

We would then set, for instance (using the "check" instead of the hat, so as to distinguish from our previous estimator)

$$\check{\mu}(72) = -151.133 + 4.783 \times 72 = 193.2666 \tag{1.8}$$

So, using this model, we would predict a slightly heavier weight than our earlier prediction.

By the way, we need not type the above expression into R by hand. Here is why: Writing the expression in matrix-multiply form, it is

$$(-151.133, 4.783) \begin{pmatrix} 1 \\ 72 \end{pmatrix} \tag{1.9}$$

Be sure to see the need for that 1 in the second factor; it is used to pick up the -151.133. Now let's use that matrix form to show how we can conveniently compute that value in R:[9]

The key is that we can exploit the fact that R's **coef()** function fetches the coefficients c and d for us:

```
> coef(lmout)
(Intercept)   mlb$Height
-151.133291     4.783332
```

Recalling that the matrix-times-matrix operation in R is specified via the %*% operator, we can now obtain our estimated value of $\mu(72)$ as

```
> coef(lmout) %*% c(1,72)
          [,1]
[1,] 193.2666
```

We can form a confidence interval from this too, which for the 95% level will be

$$\check{\mu}(72) \pm 1.96 \text{ s.e.}[(\check{\mu}(72)] \tag{1.10}$$

where *s.e.*[] signifies *standard error*, the estimated standard deviation of an estimator. Here $\check{\mu}(72)$, being based on our random sample data, is itself random, i.e., it will vary from sample to sample. It thus has a standard deviation, which we call the standard error. We will see later that s.e.$[(\check{\mu}(72)]$ is obtainable using the R **vcov()** function:

```
> tmp <- c(1,72)
> sqrt(tmp %*% vcov(lmout) %*% tmp)
          [,1]
[1,] 0.6859655
> 193.2666 + 1.96 * 0.6859655
[1] 194.6111
> 193.2666 - 1.96 * 0.6859655
[1] 191.9221
```

(More detail on **vcov()** and **coef()** as R functions is presented in Section 1.20.4 in the Computational Complements section at the end of this chapter.)

[9] In order to gain a solid understanding of the concepts, we will refrain from using R's **predict()** function for now. It will be introduced later, though, in Section 1.10.3.

So, an approximate 95% CI for $\mu(72)$ under this model would be about (191.9,194.6).

1.7 Parametric vs. Nonparametric Models

Now here is a major point: The CI we obtained from our linear model, (191.9,194.6), was narrower than what the nonparametric approach gave us, (187.6,193.2); the former has width of about 2.7, while the latter's is 5.6. In other words:

> A parametric model is — if it is (approximately) valid — more powerful than a nonparametric one, yielding estimates of a regression function that tend to be more accurate than what the nonparametric approach gives us. This should translate to more accurate prediction as well.

Why should the linear model be more effective? Here is some intuition, say for estimating $\mu(72)$: As will be seen in Chapter 2, the **lm()** function uses *all* of the data to estimate the regression coefficients. In our case here, all 1015 data points played a role in the computation of $\check{\mu}(72)$, whereas only 150 of our observations were used in calculating our nonparametric estimate $\widehat{\mu}(72)$. The former, being based on much more data, should tend to be more accurate.[10]

On the other hand, in some settings it may be difficult to find a valid parametric model, in which case a nonparametric approach may be much more effective. *This interplay between parametric and nonparametric models will be a recurring theme in this book.*

1.8 Example: Click-Through Rate

Let's try a linear regression model on the CTR data in Section 1.3. The file can be downloaded from the link in [53].

```
> ctr <- read.table('State_CTR_Date.txt',
    header=TRUE,sep='\t')
```

[10] Note the phrase *tend to* here. As you know, in statistics one usually cannot say that one estimator is always better than another, because anomalous samples do have some nonzero probability of occurring.

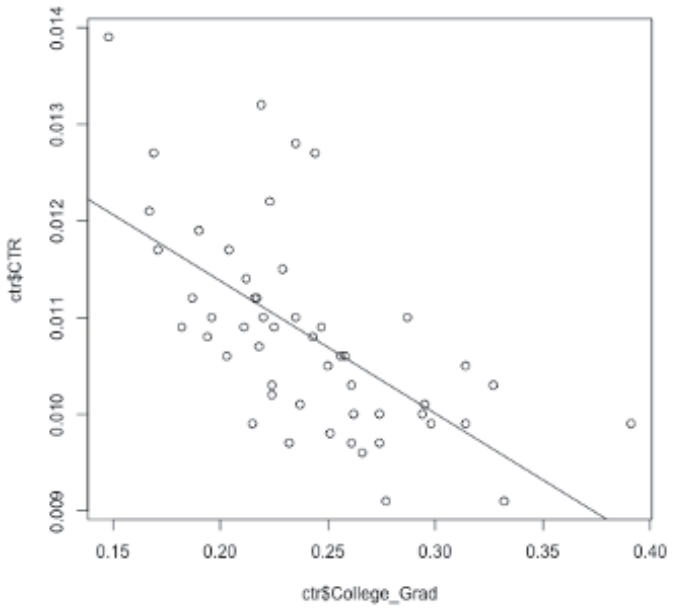


Figure 1.4: CTR data and fitted line

```
> lmout <- lm(ctr$CTR ~ ctr$College_Grad)
> lmout
...
Coefficients:
      (Intercept)   ctr$College_Grad
          0.01412           -0.01373
...
```

A scatter plot of the data, with the fitted line superimposed, is shown in Figure 1.4. It was generated by the code

```
> plot(ctr$College_Grad, ctr$CTR)
> abline(coef=coef(lmout))
```

The relation between education and CTR is interesting, but let's put this in perspective, by considering the standard deviation of **College_Grad**:

```
> sd(ctr$College_Grad)
[1] 0.04749804
```

So, a "typical" difference between one state and another is something like 0.05. Multiplying by the -0.01373 figure above, this translates to a difference in click-through rate from state to state of about 0.0005. This is certainly not enough to have any practical meaning.

So, putting aside such issues as whether our data constitute a sample from some "population" of potential states, the data suggest that there is really no substantial relation between educational level and CTR. The original blog post on this data, noting the negative value of $\widehat{d}$, cautioned that though this seems to indicate that the more-educated people click less, "correlation is not causation." Good advice, but it's equally important to note here that even if the effect is causal, it is tiny.

1.9 Several Predictor Variables

Now let's predict weight from height and age. We first need some notation.

Say we are predicting a response variable Y from variables $X^{(1)}, ..., X^{(k)}$. The regression function is now defined to be

$$\mu(t_1, ..., t_k) = E(Y \mid X^{(1)} = t_1, ..., X^{(k)} = t_k) \tag{1.11}$$

In other words, $\mu(t_1, ..., t_k)$ is the mean Y among all units (people, cars, whatever) in the population for which $X^{(1)} = t_1, ..., X^{(k)} = t_k$.

In our baseball data, Y, $X^{(1)}$ and $X^{(2)}$ might be weight, height and age, respectively. Then $\mu(72, 25)$ would be the population mean weight among all players of height 72 and age 25.

We will often use a vector notation

$$\mu(t) = E(Y \mid X = t) \tag{1.12}$$

with $t = (t_1, ..., t_k)'$ and $X = (X^{(1)}, ..., X^{(k)})'$, where $'$ denotes matrix transpose.[11]

[11] Our vectors in this book are column vectors. However, since they occupy a lot of

1.9.1 Multipredictor Linear Models

Let's consider a parametric model for the baseball data,

$$\text{mean weight} = c + d \times \text{ height} + e \times \text{ age} \tag{1.14}$$

1.9.1.1 Estimation of Coefficients

We can again use **lm()** to obtain sample estimates of c, d and e:

```
> lm(mlb$Weight ~ mlb$Height + mlb$Age)
...
Coefficients:
(Intercept)    mlb$Height        mlb$Age
   -187.6382        4.9236         0.9115
```

Note that the notation mlb**$**Weight ~mlb**$**Height + mlb**$**Age simply means "predict weight from height and age." The variable to be predicted is specified to the left of the tilde, and the predictor variables are written to the right of it. The + does not mean addition.

A shorter formulation is

```
> lm(Weight ~ Height + Age, data=mlb)
```

You can see that if we have many predictors, this notation is more compact and convenient.

And, shorter still, we could write

```
> lm(Weight ~ ., data=mlb[,4:6])
```

Here the period means "all the other variables." Since we are restricting the data to be columns 4 and 6 of **mlb**, Height and Age, the period means those two variables.

---

space on a page, we will often show them as transposes of rows. For instance, we will often write $(5, 12, 13)'$ instead of

$$\begin{pmatrix} 5 \\ 12 \\ 13 \end{pmatrix} \tag{1.13}$$

So, the output shows us the estimated coefficientsis, e.g., $\widehat{d} = 4.9236$. Our estimated regression function is

$$\widehat{\mu}(t_1, t_2) = -187.6382 + 4.9236\ t_1 + 0.9115\ t_2 \tag{1.15}$$

where t_1 and t_2 are height and age, respectively.

Setting $t_1 = 72$ and $t_2 = 25$, we find that

$$\widehat{\mu}(72, 25) = 189.6485 \tag{1.16}$$

and we would predict the weight of a 72-inch tall, age 25 player to be about 190 pounds.

1.9.1.2 The Description Goal

It was mentioned in Section 1.1 that regression analysis generally has one or both of two goals, Prediction and Description. In light of the latter, some brief comments on the magnitudes of the estimated coefficientsis would be useful at this point:

- We estimate that, on average (a key qualifier), each extra inch in height corresponds to almost 5 pounds of additional weight.
- We estimate that, on average, each extra year of age corresponds to almost a pound in extra weight.

That second item is an example of the Description goal in regression analysis. We may be interested in whether baseball players gain weight as they age, like "normal" people do. Athletes generally make great efforts to stay fit, but we may ask how well they succeed in this. The data here seem to indicate that baseball players indeed are prone to some degree of "weight creep" over time.

1.9.2 Nonparametric Regression Estimation: k-NN

Now let's drop the linear model assumption (1.14), and estimate our regression function "from scratch." So this will be a model-free approach, thus termed *nonparametric* as explained earlier.

Our analysis in Section 1.6.2 was model-free. But here we will need to broaden our approach, as follows.

1.9.2.1 Looking at Nearby Points

Again say we wish to estimate, using our data, the value of $\mu(72, 25)$. A potential problem is that there likely will not be any data points in our sample that exactly match those numbers, quite unlike the situation in (1.4), where $\widehat{\mu}(72)$ was based on 150 data points. Let's check:

```
> z <- mlb[mlb$Height == 72 & mlb$Age == 25,]
> z
[1] Name        Team        Position
[4] Height      Weight      Age
[7] PosCategory
<0 rows> (or 0-length row.names)
```

(Recall that in R, we use a single ampersand when "and-ing" vector quantities, but use a double one for ordinary logical expressions.)

So, indeed there were no data points matching the 72 and 25 numbers. Since the ages are recorded to the nearest 0.01 year, this result is not surprising. But at any rate we thus cannot set $\widehat{\mu}(72, 25)$ to be the mean weight among our sample data points satisfying those conditions, as we did in Section 1.6.2. And even if we had had a few data points of that nature, that would not have been enough to obtain an accurate estimate $\widehat{\mu}(72, 25)$.

Instead, we use data points that are *close* to the desired prediction point. Again taking the weight/height/age case as a first example, this means that we would estimate $\mu(72, 25)$ by the average weight in our sample data among those data points for which height is *near* 72 and age is *near* 25.

1.9.2.2 Measures of Nearness

Nearness is generally defined as *Euclidean distance*:

$$\text{distance}[(s_1, s_2, ..., s_k), (t_1, t_2, ..., t_k)] = \sqrt{((s_1 - t_1)^2 + ... + (s_k - t_k)^2} \tag{1.17}$$

For instance, the distance from a player in our sample of height 72.5 and

age 24.2 to the point (72,25) would be

$$\sqrt{(72.5-72)^2+(24.2-25)^2} = 0.9434 \tag{1.18}$$

Note that the Euclidean distance between $s = (s_1, ..., s_k)$ and $t = (t_1, ..., t_k)$ is simply the Euclidean norm of the difference $s - t$ (Section A.1).

1.9.2.3 The k-NN Method, and Tuning Parameters

The *k-Nearest Neighbor* (k-NN) method for estimating regression functions is simple: Find the k data points in our sample that are closest to the desired prediction point, and average their values of the response variable Y.

A question arises as to how to choose the value of k. Too large a value means we are including "nonrepresentative" data points, but too small a value gives us too few points to average for a good estimate. We will return to this question later, but will note that due to this nature of k, we will call k a *tuning* parameter. Various tuning parameters will come up in this book.

1.9.2.4 Nearest-Neighbor Analysis in the regtools Package

We will use the k-NN functions in my **regtools** package, available on CRAN [97]. The main computation is performed by **knnest()**, with preparatory nearest-neighbor computation done by **preprocessx()**. The call forms are

```
preprocessx(x,kmax,xval=FALSE)
knnest(y,xdata,k,nearf=meany)
```

In the first, **x** is our predictor variable data, one column per predictor. The argument **kmax** specifies the maximum value of k we wish to use (we might try several), and **xval** refers to cross-validation, a concept to be introduced later in this chapter. The essence of **preprocessx()** is to find the **kmax** nearest neighbors of each observation in our dataset, i.e., row of **x**.

The arguments of **knnest()** are as follows. The vector **y** is our response variable data; **xdata** is the output of **preprocessx()**; **k** is the number of nearest neighbors we wish to use. The argument **nearf** specifies the function we wish to be applied to the Y values of the neighbors; the default is the mean, but instead we could for instance specify the median. (This flexibility will be useful in other ways as well.)

The return value from **knnest()** is an object of class **'knn'**.

1.9.2.5 Example: Baseball Player Data

There is also a **predict** function associated with **knnest()**, with call form

```
predict(kout,predpts,needtoscale)
```

Here **kout** is the return value of a call to **knnest()**, and each row of **regestpts** is a point at which we wish to estimate the regression function. Also, if the points to be predicted are not in our original data, we need to set **needtoscale** to TRUE.

For example, let's estimate $\mu(72, 25)$, based on the 20 nearest neighbors at each point.

```
> data(mlb)
> library(regtools)
> xd <- preprocessx(mlb[,c(4,6)],20)
> kout <- knnest(mlb[,5],xd,20)
> predict(kout,c(72,25),TRUE)
187.4
```

So we would predict the weight of a 72-inches tall, age 25 player to be about 187 pounds, not much different — in this instance — from what we obtained earlier with the linear model.

1.10 After Fitting a Model, How Do We Use It for Prediction?

As noted, our goal in regression analysis could be either Prediction or Description (or both). How specifically does the former case work?

1.10.1 Parametric Settings

The parametric case is the simpler one. We fit our data, write down the result, and then use that result in the future whenever we are called upon to do a prediction.

Recall Section 1.9.1.1. It was mentioned there that in that setting, we probably are not interested in the Prediction goal, but just as an illustration,

suppose we do wish to predict. We fit our model to our data — called our *training data* — resulting in our estimated regression function, (1.15). From now on, whenever we need to predict a player's weight, given his height and age, we simply plug those values into (1.15).

1.10.2 Nonparametric Settings

The nonparametric case is a little more involved, because we have no explicit equation like (1.15). Nevertheless, we use our training data in the same way. For instance, say we need to predict the weight of a player whose height and age are 73.2 and 26.5, respectively. Our predicted value will then be $\widehat{\mu}(73.2, 26.5)$. To obtain that, we go back to our training data, find the k nearest points to (73.2,26.5), and average the weights of those k players. We would go through this process each time we are called upon to perform a prediction.

A variation:

A slightly different approach, which is used in **regtools**, is as follows. Denote our training set data as $(X_1, Y_1), ..., (X_n, Y_n)$, where again the X_i are typically vectors, e.g., (height,age). We estimate our regression function at each of the points X_i, forming $\widehat{\mu}(X_i), i = 1, ..., n$. Then, when faced with a new case (X, Y) for which Y is unknown, we find the *single* closest X_i to X, and guess Y to be 1 or 0, depending on whether $\widehat{\mu}(X_i) > 0.5$. Since $\widehat{\mu}(X_i)$ already incorporates the neighborhood-averaging operation, doing so for our new point would be largely redundant. Using only the single closest point saves both computation time and storage space.

1.10.3 The Generic predict() Function

Consider this code:

```
> lmout <- lm(Weight ~ Height + Age,data=mlb)
> predict(lmout,data.frame(Height = 72, Age = 25))
       1
189.6493
```

We fit the model as in Section 1.9.1.1, and then predicted the weight of a player who is 72 inches tall and age 25. We use $\widehat{\mu}(72, 25)$ for this, which of course we could obtain as

```
> coef(lmout) %*% c(1,72,25)
          [,1]
```

```
[1,] 189.6493
```

But the **predict()** function is simpler and more explicitly reflects what we want to accomplish.

By the way, **predict** is a *generic* function. This means that R will *dispatch* a call to **predict()** to a function specific to the given class. In this case, **lmout** above is of class **'lm'**, so the function ultimately executed above is **predict.lm‘**. Similarly, in Section 1.9.2.5, the call to **predict()** goes to **predict.knn()**. More details are in Section 1.20.4.

IMPORTANT NOTE: To use **predict()** with **lm()**, the latter must be called in the **data =** form shown above, and the new data to be predicted must be a data frame with the same column names.

1.11 Overfitting, and the Variance-Bias Tradeoff

One major concern in model development is *overfitting*, meaning to fit such an elaborate model that it "captures the noise rather than the signal." This description is often heard these days, but it is vague and potentially misleading. We will discuss it in detail in Chapter 9, but it is of such importance that we introduce it here in this prologue chapter.

The point is that, after fitting our model, we are concerned that it may fit our training data well but not predict well on new data in the future.[12] Let's look into this further:

1.11.1 Intuition

To see how overfitting may occur, consider the famous *bias-variance tradeoff*, illustrated in the following example. Again, keep in mind that the treatment will at this point just be intuitive, not mathematical.

Long ago, when I was just finishing my doctoral study, I had my first experience with statistical consulting. A chain of hospitals was interested in comparing the levels of quality of care given to heart attack patients at its various locations. A problem was noticed by the chain regarding straight comparison of raw survival rates: One of the locations served a

[12] Note that this assumes that nothing changes in the system under study between the time we collect our training data and the time we do future predictions.

largely elderly population, and since this demographic presumably has more difficulty surviving a heart attack, this particular hospital may misleadingly appear to be giving inferior care.

An analyst who may not realize the age issue here would thus be biasing the results. The term "bias" here doesn't mean deliberate distortion of the analysis, just that the model has a systemic bias, i.e., it is "skewed," in the common vernacular. And it is permanent bias, in the sense that it won't disappear, no matter how large a sample we take.

Such a situation, in which an important variable is not included in the analysis, is said to be *underfitted*. By adding more predictor variables in a regression model, in this case age, we are reducing bias.

Or, suppose we use a regression model that is linear in our predictors, but the true regression function is nonlinear. This is bias too, and again it won't go away even if we make the sample size huge. This is often called *model bias* by statisticians; the economists call the model *misspecified*.

On the other hand, we must keep in mind that our data is a sample from a population. In the hospital example, for instance, the patients on which we have data can be considered a sample from the (somewhat conceptual) population of all patients at this hospital, past, present and future. A different sample would produce different regression coefficient estimates. In other words, there is variability in those coefficients from one sample to another, i.e., variance. We hope that that variance is small, which gives us confidence that the sample we have is representative.

But the more predictor variables we have, the more collective variability there is in the inputs to our regression calculations, and thus the larger the variances of the estimated coefficients.[13] If those variances are large enough, the bias-reducing benefit of using a lot of predictors may be overwhelmed by the increased variability of the results. This is called *overfitting*.

In other words:

> In deciding how many (and which) predictors to use, we have a tradeoff. The richer our model, the less bias, but the higher the variance.

In Section 1.19.2 it is shown that for any statistical estimator $\widehat{\theta}$ (that has finite variance),

$$\text{mean squred error} = \text{squared bias} + \text{variance}$$

[13] I wish to thank Ariel Shin for this interpretation.

Our estimator here is $\widehat{\mu}(t)$. This shows the tradeoff: Adding variables, such as age in the hospital example, reduces squared bias but increases variance. Or, equivalently, removing variables reduces variance but exacerbates bias. It may, for example, be beneficial to accept a little bias in exchange for a sizable reduction in variance, which we may achieve by removing some predictors from our model.

The trick is to somehow find a "happy medium," easier said than done. Chapter 9 will cover this in depth, but for now, we introduce a common method for approaching the problem:

1.11.2 Example: Student Evaluations of Instructors

In Section 9.9.5 we will analyze a dataset consisting of student evaluations of instructors. Let's defer the technical details until then, but here is a sneak preview.

The main variables here consist of 28 questions on the instructor, such as "The quizzes, assignments, projects and exams contributed to helping the learning." The student gives a rating of 1 to 5 on each question.

Figure 1.5 describes this data, plotting frequency of occurrence against the questions (the 28, plus 4 others at the beginning). Again, don't worry about the details now, but it basically shows there are 3 kinds of instructors: one kind gets very righ ratings on the 28 questions, across the board; one kind gets consistently medium-high ratings; and the third kind gets low ratings across all the questions.

This indicates that we might reduce those 28 questions to just one, in fact any one of the 28.

1.12 Cross-Validation

The proof of the pudding is in the eating — old English saying

Toward that end, i.e., proof via "eating," it is common to artificially create a set of "new" data and try things out there. Instead of using all of our collected data as our training set, we set aside part of it to serve as simulated "new" data. This is called the *validation set* or *test set*. The remainder will be our actual training data. In other words, we randomly partition our original data, taking one part as our training set and the other part to

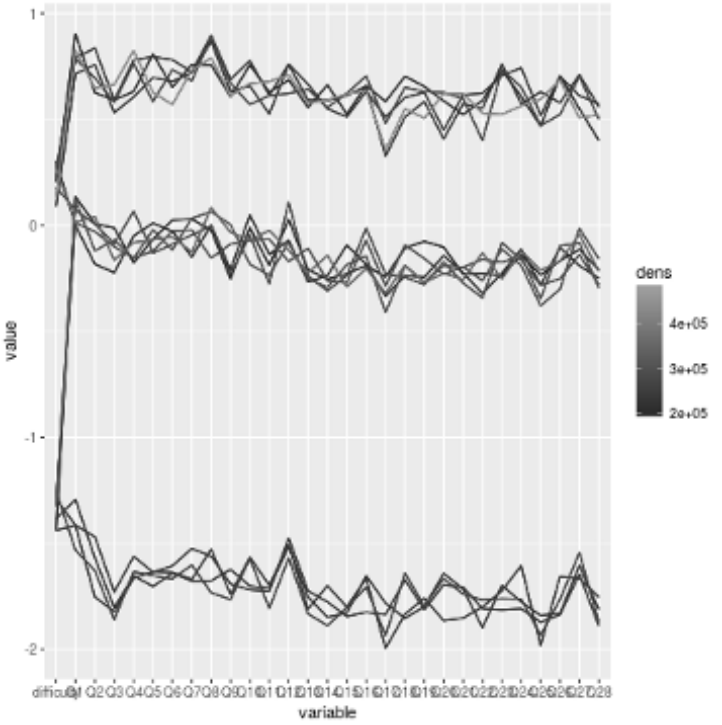


Figure 1.5: Turkish student evaluations (see color insert)

play the role of new data. We fit our model, or models, to the training set, then do prediction on the test set, pretending its response variable values are unknown. We then compare to the real values. This will give us an idea of how well our models will predict in the future. The method is called *cross-validation*.

The above description is a little vague, and since there is nothing like code to clarify the meaning of an algorithm, let's develop some. Here first is code to do the random partitioning of **data**, with a proportion **p** to go to the training set:

```
xvalpart <- function(data,p) {
   n <- nrow(data)
   ntrain <- round(p*n)
```

```
    trainidxs <- sample(1:n,ntrain,replace=FALSE)
    list(train=data[trainidxs,],
        valid=data[-trainidxs,])
}
```

R uses **-** in array indices for exclusion, e.g.,

```
> x <- c(5,12,13)
> x[-2]
[1]  5 13
```

Thus, using the expression **-trainidxs** above gives us the validation cases.

Now to perform cross-validation, we'll consider the parametric and non-parametric cases separately, in the next two sections.

1.12.1 Linear Model Case

To do cross-validation for linear models, we could use this code.[14]

1.12.1.1 The Code

```
# arguments:
#
#     data:  full data
#     ycol:  column number of resp. var.
#     predvars:  column numbers of predictors
#     p:  prop. for training set
#     meanabs:  see 'value' below

# value:  if meanabs is TRUE, the mean absolute
#         prediction error; otherwise, an R list
#         containing pred., real Y

xvallm <- function(data,ycol,predvars,p,meanabs=TRUE){
   tmp <- xvalpart(data,p)
   train <- tmp$train
   valid <- tmp$valid
   # fit model to training data
```

[14] There are sophisticated packages on CRAN for this, such as **cvTools** [4]. But to keep things simple, and to better understand the concepts, we will write our own code. Similarly, as mentioned, we will not use R's **predict()** function for the time being.

```
    trainy <- train[,ycol]
    trainpreds <- train[,predvars]
    # using matrix form in lm() call
    trainpreds <- as.matrix(trainpreds)
    lmout <- lm(trainy ~ trainpreds)
    # apply fitted model to validation data; note
    # that %*% works only on matrices, not data frames
    validpreds <- as.matrix(valid[,predvars])
    predy <- cbind(1,validpreds)%*% coef(lmout)
    realy <- valid[,ycol]
    if (meanabs) return(mean(abs(predy - realy)))
    list(predy = predy, realy = realy)
}
```

1.12.1.2 Applying the Code

Let's try cross-validtion on the weight/height/age data, using mean absolute prediction error as our criterion for prediction accuracy:

```
library(freqparcoord)
data(mlb)
xvallm(mlb,5,c(4,6),2/3)
```

```
> xvallm(mlb,5,c(4,6),2/3)
[1] 13.38045
```

So, on average we would be off by about 13 pounds. We might improve upon this by using the data's Position variable, but we'll leave that for later.

Keep in mind the randomness, though. We randomly split the data, and would get a different result if we were to run the code again. This point is explored in Exercise 1 at the end of this chapter. Also, we will later discuss an extension, i*r-fold cross-validation*, in Section 2.9.6.

1.12.2 k-NN Case

Here is the code for performing cross-validation for k-NN:

```
# arguments:
#
#    data:  full data
#    ycol:  column number of resp. var.
```

```
#    k:  number of nearest neighbors
#    p:  prop. for training set
#    meanabs:  see 'value' below

# value:  if meanabs is TRUE, the mean absolute
#         prediction error; otherwise, an R list
#         containing pred., real Y

xvalknn <-
      function(data,ycol,predvars,k,p,meanabs=TRUE){
   # cull out just Y and the Xs
   data <- data[,c(predvars,ycol)]
   ycol <- length(predvars) + 1
   tmp <- xvalpart(data,p)
   train <- tmp$train
   valid <- tmp$valid
   valid <- as.matrix(valid)
   xd <- preprocessx(train[,-ycol],k)
   kout <- knnest(train[,ycol],xd,k)
   predy <- predict(kout,valid[,-ycol],TRUE)
   realy <- valid[,ycol]
   if (meanabs) return(mean(abs(predy - realy)))
   list(predy = predy, realy = realy)
}
```

So, how well does k-NN predict?

```
> library(regtools)
> set.seed(9999)
> xvalknn(mlb,5,c(4,6),25,2/3)
[1] 14.32817
```

The two methods gave similar results. However, not only must we keep in mind the randomness of the partitioning of the data, but we also must recognize that this output above depended on choosing a value of 25 for **k**, the number of nearest neighbors. We could have tried other values of **k**, and in fact could have used cross-validation to choose the "best" value.

1.12.3 Choosing the Partition Sizes

One other problem, of course, is that we did have a random partition of our data. A different one might have given substantially different results.

In addition, there is the matter of choosing the sizes of the training and validation sets (e.g., via the argument **p** in **xvalpart()**). We have a classical tradeoff at work here: Let k be the size of our training set. If we make k too large, the validation set will be too small for an accurate measure of prediction accuracy. We won't have that problem if we set k to a smaller size, but then we are measuring the predictive ability of only k observations, whereas in the end we will be using all n observations for predicting new data.

The *Leaving One-Out Method* and its generalizations solves this problem, albeit at the expense of much more computation. It will be presented in Section 2.9.5.

1.13 Important Note on Tuning Parameters

Recall how k-NN works: To predict a new case for which $X = t$ but Y is unknown, we look at the our existing data. We find the k closest neighbors to t, then average their Y values. That average becomes our predicted value for the new case.

We refer to k as a tuning parameter, to be chosen by the user. Many methods have multiple tuning parameters, making the choice a challenge. One can of course choose their values using cross validation, and in fact the **caret** package includes methods to automate the process, simultaneously optimizing over many tuning parameters.

But cross-validation can have its own overfitting problems (Section 9.3.2). One should not be lulled into a false sense of security.

The late Leo Breiman was suspicious of tuning parameters, and famously praised one regression method (*boosting*), as "the best off-the-shelf method" available — meaning that the method works well without tweaking tuning parameters. His statement may have been overinterpreted regarding the boosting method, but the key point here is that Breiman was not a fan of tuning parameters.

A nice description of Breiman's view was given in an obituary by Michael Jordan, who noted [78],

> Another preferred piece of Breimanesque terminology was "off-the-shelf," again a rather physical metaphor. Leo tended to be suspicious of "free parameters;" procedures should work with little or no "tuning."

Breiman's concerns about tuning parameters extended to choosing those parameters via cross-validation. The latter is an important tool, but should be used with a healthy dose of skepticism.

This issue will come up often, since many commonly-used methods do have various tuning parameters; some have multiple, complex tuning parameters.

Again, this point must be kept in mind:

> Optimizing for a tuning parameter is **inherently prone to overfitting**, as we are optimizing for particular data. If we have multiple tuning parameters, the potential for overfitting is compounded.

1.14 Rough Rule of Thumb

The issue of how many predictors to use to simultaneously avoid overfitting and still produce a good model is nuanced, and in fact this is still not fully resolved. Chapter 9 will be devoted to this complex matter.

Until then, though it is worth using the following:[15]

> **Rough Rule of Thumb (Tukey):** For a data set consisting of n observations, use fewer than $\sqrt{(n)}$ predictors.

1.15 Example: Bike-Sharing Data

We now return to the bike-sharing data (Section 1.1). Our little excursion to the simpler data set, involving baseball player weights and heights, helped introduce the concepts in a less complex setting. The bike-sharing data set is more complicated in several ways:

- **Complication (a):** It has more potential predictor variables.
- **Complication (b):** It includes some *nominal* (or *categorical*) variables, such as Day of Week. The latter is technically numeric, 0 through 6, but those codes are just names. Hence the term *nominal*. In R, by the way, the formal term for such variables is *factors*.

[15] Unfortunately, reference unknown.

The problem is that there is no reason, for instance, that Sunday, Thursday and Friday should have an ordinal relation in terms of ridership just because, say, $0 < 4 < 5$.

- **Complication (c):** It has some potentially nonlinear relations. For instance, people don't like to ride bikes in freezing weather, but they are not keen on riding on really hot days either. Thus we might suspect that the relation of ridership to temperature rises at first, eventually reaching a peak, but declines somewhat as the temperature increases further.

Now that we know some of the basic issues from analyzing the baseball data, we can treat this more complicated data set.

Let's read in the bike-sharing data. We'll look at one of the files in that dataset, **day.csv**. We'll restrict attention to the first year,[16] and since we will focus on the registered riders, let's shorten the name for convenience:

```
> shar <- read.csv("day.csv",header=TRUE)
> shar <- shar[1:365,]
> names(shar)[15] <- "reg"
```

1.15.1 Linear Modeling of $\mu(t)$

In view of Complication (c) above, the inclusion of the word *linear* in the title of our current section might seem contradictory. But one must look carefully at *what* is linear or not, and we will see shortly that, yes, we can use linear models to analyze nonlinear relations.

Let's first check whether the ridership/temperature relation seems nonlinear, as we have speculated:

```
plot(shar$temp,shar$reg)
```

The result is shown in Figure 1.6.

There seem to be some interesting groupings among the data, likely due to the other variables, but putting those aside for now, the plot does seem to suggest that ridership is slightly associated with temperature in the "first rising, then later falling" form as we had guessed.

[16] There appears to have been some systemic change in the second year, and while this could be modeled, we'll keep things simple by considering only the first year.

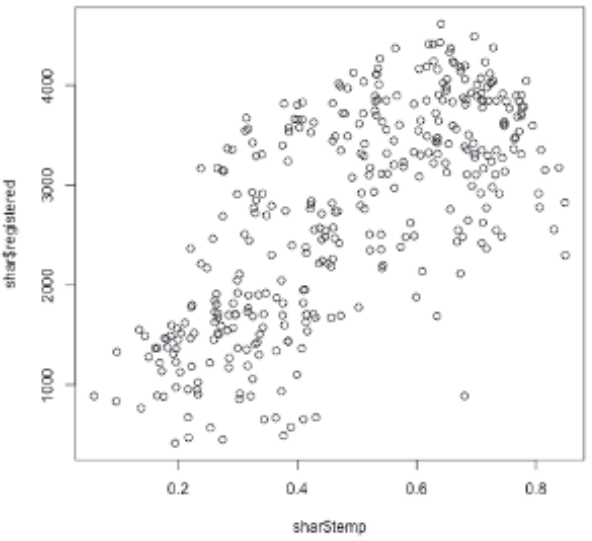


Figure 1.6: Ridership vs. temperature

Thus a linear model of the form

$$\text{mean ridership} = c + d \times \text{ temperature} \tag{1.19}$$

would seem inappropriate. But don't give up so quickly! A model like

$$\text{mean ridership} = c + d \times \text{ temperature} + e \times \text{ temperature}^2 \tag{1.20}$$

i.e., with a temperature-squared term added, might work fine. A negative value for e would give us the "first up, then later down" behavior we want our model to have.

And there is good news — the model (1.20) is actually linear! We say that the expression is *linear in the parameters*, even though it is nonlinear with respect to the temperature variable. This means that if we multiply each of c, d and e by, say, 8, then the values of the left and right sides of the equation both increase eightfold.

Anotber way to see this is that in calling **lm()**, we can simply regard squared temperature as a new variable:

```
> shar$temp2 <- shar$temp^2
> lm(shar$reg ~ shar$temp + shar$temp2)

Call:
lm(formula = shar$reg ~ shar$temp + shar$temp2)

Coefficients:
(Intercept)     shar$temp    shar$temp2
     -378.9        9841.8       -6169.8
```

And note that, sure enough, the coefficient of the squared term, $\widehat{e} = -6169.8$, did indeed turn out to be negative.

Of course, we want to predict from many variables, not just temperature, so let's now turn to Complication (b) cited earlier, the presence of nominal data. This is not much of a problem either.

Such situations are generally handled by setting up what are called *indicator variables* or *dummy variables*. The former term alludes to the fact that our variable will *indicate* whether a certain condition holds or not, with 1 coding the yes case and 0 indicating no.

We could, for instance, set up such a variable for Tuesday data:

```
> shar$tues <- as.integer(shar$weekday == 2)
```

Indeed, we could define six variables like this, one for each of the days Monday through Saturday. Note that Sunday would then be indicated indirectly, via the other variables all having the value 0. A direct Sunday variable would be redundant, and in fact would present mathematical problems, as we'll see in Chapter 8. (Actually, R's **lm()** function can deal with factor variables directly, as shown in Section 9.7.5.1. But we take the more basic route here, in order to make sure the underlying principles are clear.)

However, let's opt for a simpler analysis, in which we distinguish only between weekend days and weekdays, i.e. define a dummy variable that is 1 for Monday through Friday, and 0 for the other days. Actually, those who assembled the data set already defined such a variable, which they named **workingday**.[17]

[17] More specifically, a value of 1 for this variable indicates that the day is in the Monday-Friday range *and* it is not a holiday.

We incorporate this into our linear model:

$$\text{mean reg} = c + d \times \text{ temp} + e \times \text{ temp}^2 + f \text{ workingday} \tag{1.21}$$

There are several other dummy variables that we could add to our model, but for this introductory example let's define just one more:

```
> shar$clearday <- as.integer(shar$weathersit == 1)
```

So, our regression model will be

$$\begin{aligned} \text{mean reg} &= \beta_0 + \beta_1 \text{ temp} + \beta_2 \text{ temp}^2 \\ &+ \beta_3 \text{ workingday } + \beta_4 \text{ clearday} \end{aligned} \tag{1.22}$$

As is traditional, here we have used subscripted versions of the Greek letter β to denote our equation coefficients, rather than c, d and so on.

So, let's run this through **lm()**:

```
> lmout <- lm(reg ~ temp+temp2+workingday+clearday ,
      data = shar)
```

The return value of **lm()**, assigned here to **lmout**, is a very complicated R object, of class **'lm'**. We shouldn't inspect it in detail now, but let's at least print the object, which in R's interactive mode can be done simply by typing the name, which automatically calls **print()** on the object:[18]

```
> lmout
...
...
Coefficients:
(Intercept)         temp        temp2   workingday
    -1362.6      11059.2      -7636.4        686.0
   clearday
      518.9
```

Remember, the population function $\mu(t)$ is unnown, so the β_i are unknown. The above coefficients are merely sample-based estimates. For example, using our usual "hat" notation to mean "estimate of," we have that

$$\widehat{\beta}_3 = 686.0 \tag{1.23}$$

[18] See more detail on this in Section 1.20.4.

The estimated regression function is then

$$\widehat{\mu}(t_1, t_2, t_3, t_4) = -1362.6 + 11059.2t_1 - 7636.4t_2 + 686.0t_3 + 518.9t_4 \quad (1.24)$$

where $t_2 = t_1^2$.

So, what should we predict for the number of riders on the type of day described at the outset of this chapter — Sunday, sunny, 62 degrees Fahrenheit? First, note that the designers of the data set have scaled the **temp** variable to [0,1], as

$$\frac{\text{Celsius temperature} - \text{minimum}}{\text{maximum} - \text{minimum}} \quad (1.25)$$

where the minimum and maximum here were -8 and 39, respectively. This form may be easier to understand, as it is expressed in terms of where the given temperature fits on the normal range of temperatures. A Fahrenheit temperature of 62 degrees corresponds to a scaled value of 0.525. So, our predicted number of riders is

```
> coef(lmout) %*% c(1,0.525,0.525^2,0,1)
          [,1]
[1,] 2857.677
```

So, our predicted number of riders for sunny, 62-degree Sundays will be about 2858. How does that compare to the average day?

```
> mean(shar$reg)
[1] 2728.359
```

So, we would predict a somewhat above-average level of ridership.

As noted earlier, one can also form confidence intervals and perform significance tests on the β_i. We'll go into this in Chapter 2, but some brief comments on the magnitudes and signs of the $\widehat{\beta}_i$ are useful at this point:

- As noted, the estimated coefficient of **temp2** is negative, consistent with our intuition. Note, though, that it is actually less negative than when we predicted **reg** from only temperature and its square. This change is typical, and will be discussed in detail in Chapter 7.
- The estimated coefficient for **workingday** is positive. This too matches our intuition, as presumably many of the registered riders use the

bikes to commute to work. The value of the estimate here, 686.0, indicates that, for fixed temperature and weather conditions, weekdays tend to have close to 700 more registered riders than weekends.

- Similarly, the coefficient of **clearday** suggests that for fixed temperature and day of the week, there are about 519 more riders on clear days than on other days.

1.15.2 Nonparametric Analysis

Let's see what k-NN gives us as our predicted value for sunny, 62-degree Sundays, say with $k = 20$:

```
> shar1 <-
      shar[,c('workingday','temp','reg','clearday')]
> xd <- preprocessx(shar1[,-3],20)
> kout <- knnest(shar1$reg,xd,20)
> predict(kout,c(0,0.525,1),TRUE)
2881.8
```

This is again similar to what the linear model gave us. This probably means that the linear model was pretty good, but we will discuss this in detail in Chapter 6.

1.16 Interaction Terms, Including Quadratics

Let's take another look at (1.22), specifically the term involving the variable **workingday**, a dummy indicating a nonholiday Monday through Friday. Our estimate for β_3 turned out to be 686.0, meaning that, holding temperature and the other variables fixed, there is a mean increase of about 686.0 riders on working days.

But look at our model, (1.22). The (estimated) values of the right-hand side will differ by 686.0 for working vs. nonworking days, no matter what the temperature is. In other words, the working day effect is the same on low-temprerature days as on warmer days. For a broader model that does not make this assumption, we could add an *interaction term*, consisting of a product of **workingday** and **temp**:

$$\begin{aligned} \text{mean reg} &= \beta_0 + \beta_1 \text{ temp} + \beta_2 \text{ temp}^2 \\ &+ \beta_3 \text{ workingday} + \beta_4 \text{ clearday} \qquad (1.26) \\ &+ \beta_5 \text{ temp} \times \text{workingday} \qquad (1.27) \end{aligned}$$

Note that the temp^2 term is also an interaction term, the interaction of the **temp** variable with itself.

How does this model work? Let's illustrate it with a new data set.

1.16.1 Example: Salaries of Female Programmers and Engineers

This data is from the 2000 U.S. Census, consisting of 20,090 programmers and engineers in the Silicon Valley area. The data set is included in the **freqparcoord** package on CRAN [104]. Suppose we are working toward a Description goal, specifically the effects of gender on wage income.

As with our bike-sharing data, we'll add a quadratic term, in this case on the age variable, reflecting the fact that many older programmers and engineers encounter trouble finding work [108]. Let's restrict our analysis to workers having at least a Bachelor's degree, and look at the variables **age**, **age2**, **sex** (coded 1 for male, 2 for female), **wkswrked** (number of weeks worked), **ms**, **phd** and **wageinc** (wage income). Other than an age^2 term, we'll start out with no interaction terms.

```
> library(freqparcoord)
> data(prgeng)
> prgeng$age2 <- prgeng$age^2
> edu <- prgeng$educ
> prgeng$ms <- as.integer(edu == 14)
> prgeng$phd <- as.integer(edu == 16)
> prgeng$fem <- prgeng$sex - 1
> tmp <- prgeng[edu >= 13,]
> pe <- tmp[,c(1,12,9,13,14,15,8)]
> pe <- as.matrix(pe)
```

Our model is

$$\begin{aligned} \text{mean wageinc} &= \beta_0 + \beta_1 \text{ age} + \beta_2 \text{ age}^2 + \beta_3 \text{ wkswrkd} \\ &+ \beta_4 \text{ ms} + \beta_5 \text{ phd} \\ &+ \beta_6 \text{ fem} \end{aligned} \tag{1.28}$$

We find the following:

```
> lm(wageinc ~
      age+age2+wkswrkd+ms+phd+fem, data=prgeng)
...
Coefficients:
(Intercept)          age         age2      wkswrkd
  -81136.70      3900.35       -40.33      1196.39
         ms          phd          fem
   15431.07     23183.97    -11484.49
```

The model probably could use some refining, for example variables we have omitted, such as occupation. But as a preliminary statement, the results are striking in terms of gender: With age, education and so on held constant, women are estimated to have incomes about $11,484 lower than comparable men.

But this analysis implicitly assumes that the female wage deficit is, for instance, uniform across educational levels. To see this, consider (1.28). Being female makes a β_6 difference, no matter what the values of **ms** and **phd** are. (For that matter, this is true of **age** too, though we won't model that here for simplicity.) To generalize our model in this regard, let's define two interaction variables, the product of **ms** and **fem**, and the product of **phd** and **fem**.

Our model is now

$$\begin{aligned} \text{mean wageinc} &= \beta_0 + \beta_1 \text{ age} + \beta_2 \text{ age}^2 + \beta_3 \text{ wkswrkd} \\ &+ \beta_4 \text{ ms} + \beta_5 \text{ phd} \\ &+ \beta_6 \text{ fem} + \beta_7 \text{ msfem} + \beta_8 \text{ phdfem} \end{aligned} \tag{1.29}$$

So, now instead of there being a single number for the "female effect," β_6, we how have three:

- Female effect for holders of a Bachelor's degree: β_6

- Female effect for Master's degree holders: $\beta_6 + \beta_7$
- Female effect for PhD degree holders $\beta_6 + \beta_8$

So, let's rerun the regression analysis:

```
> prgeng$msfem <- prgeng$ms * prgeng$fem
> prgeng$phdfem <- prgeng$phd * prgeng$fem
> lm(wageinc ~
      age+age2+wkswrkd+ms+phd+fem+msfem+phdfem,
      data=prgeng)
...
Coefficients:
(Intercept)          age          age2       wkswrkd
  -81216.78      3894.32       -40.29       1195.31
         ms          phd          fem         msfem
   16433.67     25325.31    -10276.80     -4157.25
     phdfem
  -14061.64
```

Let's compute the estimated values of the female effects, first for a worker with less than a graduate degree. This is -10276.80. For the Master's case, the mean female effect is estimated to be -10276.80 - 4157.25 = -14434.05. For a PhD, the figure is -10276.80 - 14861.64 = -25138.44. In other words, Once one factors in educational level, the gender gap is seen to be even worse than before.

Thus we still have many questions to answer, especially since we haven't considered other types of interactions yet. This story is not over yet, and will be pursued in detail in Chapter 7.

Rather than creating the interaction terms "manually" as is done here, one can use R colon operator, e.g., **ms:fem**, which automates the process. This was not done above, so as to ensure that the reader fully understands the meaning of interaction terms. But this is how it would go:

```
> lm(wageinc ~ age+age2+wkswrkd+ms+phd+fem+
      ms:fem+phd:fem, data=prgeng)
...
Coefficients:
(Intercept)          age          age2
  -81216.78      3894.32       -40.29
    wkswrkd           ms          phd
    1195.31     16433.67     25325.31
```

```
          fem        ms:fem        phd:fem
   -10276.80      -4157.25      -14061.64
```

For information on the colon and related operators, type **?formula** at the R prompt.

1.16.2 Fitting Separate Models

Suppose we have a model that includes a dummy predictor D, and we form interaction terms between D and other predictors. In essence, this is the same as fitting two regression models without interaction terms, one for the subpopulation $D = 1$ and the other for $D = 0$. To see this, consider again the census data above.

To keep things simple, let's just one other predictor, the age variable, and take D to be the dummy variable for female:

```
> data(prgeng)
> prgeng$fem <- prgeng$sex - 1
> fm <- which(prgeng$fem == 1)
> male <- prgeng[-fm,]  # data from male subpop
> female <- prgeng[fm,]  # data from female subpop
> lm(wageinc ~ age,data=male)
Coefficients:
(Intercept)          age
    44313.2        486.2
> lm(wageinc ~ age,data=female)
Coefficients:
(Intercept)          age
      30551          503
> lm(wageinc ~ age+fem+age*fem,data=prgeng)
Coefficients:
(Intercept)          age          fem      age:fem
    44313.2        486.2     -13761.7         16.8
```

Look at that last result. For a female worker, **fem** and **age:fem** would be equal to 1 and **age**, respectively. That means the coefficent for **age** would be $486.2 + 16.8 = 503$, which matches the 503 value obtained from running **lm()** with **data = female**. For a male worker, **fem** and **age:fem** would both be 0, and the **age** coefficent is then 486.2, matching the **lm()** results for the **male** data. The intercept terms match similarly.

The reader may be surprised that the estimated age coefficient is higher

for the women than the men. The problem is that the intercept term is much lower for women, and the line for men is above that for women for all reasonable values of age. At age 50, for instance, the estimated mean for men is $44313.2 + 486.2 \times 50 = 68623.2$, while for women it is $30551 + 503 \times 50 = 55701$.

If our goal is Description, running separate regression models like this may be much easier to interpret. This is highly encouraged. However, things become unwieldy if we have multiple dummies; if there are d of them, we must fit 2^d separate models.

1.16.3 Saving Your Work

Readers who are running the book's examples on their computers may find it convenient to use R's **save()** and **load()** functions. Our **pe** data above will be used again at various points in the book, so it is worthwhile to save it:

```
> save(pe,file='pe.save')
```

Later — days, weeks, whatever — you can reload it by simply typing

```
load('pe.save')
```

Your old **pe** object will now be back in memory. This is a lot easier than reloading the original **prgeng** data, adding the **fem**, **ms** and **phd** variables, etc.

1.16.4 Higher-Order Polynomial Models

Theoretically, we need not stop with quadratic terms. We could add cubic terms, quartic terms and so on. Indeed, the famous Stone-Weierstrass Theorem [123] says that any continuous function can be approximated to any desired accuracy by some high-order polynomial.

But this is not practical. In addition to the problem of overfitting there are numerical issues. In other words, roundoff errors in the computation would render it meaningless at some point, and indeed **lm()** will refuse to compute if it senses a situation like this. See Exercise 1 in Chapter 8.

1.17 Classification Techniques

Recall the hospital example in Section 1.11.1. There the response variable is nominal, represented by a dummy variable taking the values 1 and 0, depending on whether the patient survives or not. This is referred to as a *classification problem*, because we are trying to predict which class the population unit belongs to — in this case, whether the patient will belong to the survival or nonsurvival class. We could set up dummy variables for each of the hospital branches, and use these to assess whether some were doing a better job than others, while correcting for variations in age distribution from one branch to another. (Thus our goal here is Description rather than directly Prediction itself.)

The point is that we are predicting a 1-0 variable. In a marketing context, we might be predicting which customers are more likely to purchase a certain product. In a computer vision context, we may want to predict whether an image contains a certain object. In the future, if we are fortunate enough to develop relevant data, we might even try our hand at predicting earthquakes.

Classification applications are extremely common. And in many cases there are more than two classes, such as in identifying many different printed characters in computer vision.

In a number of applications, it is desirable to actually convert a problem with a numeric response variable into a classification problem. For instance, there may be some legal or contractual aspect that comes into play when our variable V is above a certain level $\mathbf{c}$, and we are only interested in whether the requirement is satisfied. We could replace V with a new variable

$$Y = \begin{cases} 1, & \text{if } V > c \\ 0, & \text{if } V \leq c \end{cases} \tag{1.30}$$

Classification methods will play a major role in this book.

1.17.1 It's a Regression Problem!

Recall that the regression function is the conditional mean:

$$\mu(t) = E(Y \mid X = t) \tag{1.31}$$

(As usual, X and t may be vector-valued.) In the classification case, Y is an indicator variable, which implies that we know its mean is the probability that $Y = 1$ (Section 1.19.1). In other words,

$$\mu(t) = P(Y = 1 \mid X = t) \tag{1.32}$$

The great implication of this is that *the extensive knowledge about regression analysis developed over the years can be applied to the classification problem.*

One intuitive strategy would be to guess that $Y = 1$ if the conditional probability of 1 is greater than 0.5, and guess 0 otherwise. In other words,

$$\text{guess for } Y = \begin{cases} 1, & \text{if } \mu(X) > 0.5 \\ 0, & \text{if } \mu(X) \leq 0.5 \end{cases} \tag{1.33}$$

It turns out that this strategy is optimal, in that it minimizes the overall misclassification error rate (see Section 1.19.4 in the Mathematical Complements portion of this chapter). However, it should be noted that this is not the only possible criterion that might be used. We'll return to this issue in Chapter 5.

As before, note that (1.32) is a population quantity. We'll need to estimate it from our sample data.

1.17.2 Example: Bike-Sharing Data

Let's take as our example the situation in which ridership is above 3500 bikes, which we will call HighUsage:

```
> shar$highuse <- as.integer(shar$reg > 3500)
```

We'll try to predict that variable. Let's again use our earlier example, of a Sunday, clear weather, 62 degrees. Should we guess that this will be a High Usage day?

We can use our k-NN approach just as before. Indeed, we don't need to re-run **preprocessx()**.

```
> kout <- knnest(as.integer(shar1$reg > 3500),xd,20)
> predict(kout,c(0,0.525,1),TRUE)
0.1
```

We estimate that there is a 10% chance of that day having HighUsage.

The parametric case is a little more involved. A model like

$$\begin{aligned} \text{probability of HighUsage} &= \beta_0 + \beta_1 \text{ temp} + \beta_2 \text{ temp}^2 \\ &+ \beta_3 \text{ workingday } + \beta_4 \text{ clearday} \end{aligned} \tag{1.34}$$

could be used, but would not be very satisfying. The left-hand side of (1.34), as a probability, should be in [0,1], but the right-hand side could in principle fall far outside that range.

Instead, the most common model for conditional probability is *logistic regression*:

$$\begin{aligned} \text{probability of HighUsage} &= \ell(\beta_0 + \beta_1 \text{ temp} + \beta_2 \text{ temp}^2 \\ &+ \beta_3 \text{ workingday } + \beta_4 \text{ clearday}) \end{aligned} \tag{1.35}$$

where $\ell(s)$ is the *logistic function*,

$$\ell(s) = \frac{1}{1+e^{-s}} \tag{1.36}$$

Our model, then, is

$$\mu(t_1,t_2,t_3,t_4) = \frac{1}{1+e^{-(\beta_0+\beta_1 t_1+\beta_2 t_2+\beta_3 t_3+\beta_4 t_4)}} \tag{1.37}$$

where t_1 is temperature, t_2 is the square of temperature, and so on. We wish to estimate $\mu(62, 62^2, 0, 1)$.

Note the form of the curve, shown in Figure 1.7 The appeal of this model is clear at a glance: First, the logistic function produces a value in [0,1], as appropriate for modeling a probability. Second, it is a monotone increasing function in each of the variables in (1.35), just as was the case in (1.22) for predicting our numeric variable, **reg**. Other motivations for using the logistic model will be discussed in Chapter 4.

R provides the **glm()** ("generalized linear model") function for several nonlinear model families, including the logistic,[19] which is designated via **family = binomial**:

[19] Often called "logit," by the way.

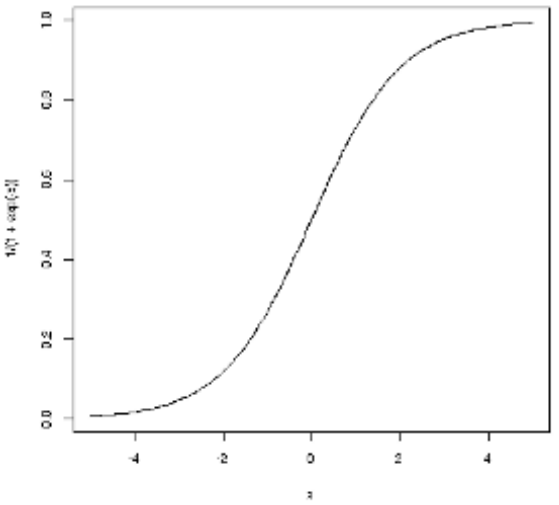


Figure 1.7: Logistic function

```
> shar$highuse <- as.integer(shar$reg > 3500)
> glmout <- glm(highuse ~
      temp+temp2+workingday+clearday,
      data=shar,family=binomial)
> tmp <- coef(glmout) %*% c(1,0.525,0.525^2,0,1)
> 1/(1+exp(-tmp))
           [,1]
[1,] 0.1010449
```

So, our parametric model gives an almost identical result here to the one arising from k-NN, about a 10% probability of HighUsage.

1.18 Crucial Advice: Don't Automate, Participate!

Data science should not be a "spectator sport"; the methodology is effective only if the users *participate*. Avoid ceding the decision making to the computer output. For example:

- Statistical significance does not imply practical importance, and conversely.
- A model is just that — just an approximation to reality, hopefully useful but never exact.
- Don't rely solely on variable selection algorithms to choose your model (Chapter 9).
- "Read directions before use" — make sure you understand what a method really does before employing it.

1.19 Mathematical Complements

1.19.1 Indicator Random Variables

A random variable W is an indicator variable, if it is equal to 1 or 0, depending on whether a certain event Q occurs or not. Two simple properties are very useful:

- $EW = P(Q)$

  This follows from

$$EW = 1 \cdot P(Q) + 0 \cdot P(\text{not } Q) = P(Q) \tag{1.38}$$

- $Var(W) = P(Q) \cdot [1 - P(Q)]$

  True because

$$Var(W) = E(W^2) - (EW)^2 = E(W) - E(W^2) = EW(1 - EW) \tag{1.39}$$

  where the second equality stems from $W^2 = W$ (remember, W is either 1 or 0). Then use the first bullet above!

1.19.2 Mean Squared Error of an Estimator

Say we are estimating some unknown population value θ, using an estimator $\widehat{\theta}$ based on our sample data. Then a natural measure of the accuracy of

our estimator is the *Mean Squared Error* (MSE),

$$E[(\widehat{\theta} - \theta)^2] \tag{1.40}$$

This is the squared distance from our estimator to the true value, averaged over all possible samples.

Let's rewrite the quantity on which we are taking the expected value:

$$\left(\widehat{\theta} - \theta\right)^2 = \left(\widehat{\theta} - E\widehat{\theta} + E\widehat{\theta} - \theta\right)^2 = (\widehat{\theta} - E\widehat{\theta})^2 + (E\widehat{\theta} - \theta)^2 + 2(\widehat{\theta} - E\widehat{\theta})(E\widehat{\theta} - \theta) \tag{1.41}$$

Look at the three terms on the far right of (1.41). The expected value of the first is $Var(\widehat{\theta})$, by definition of variance.

As to the second term, $E\widehat{\theta} - \theta$ is the *bias* of $\widehat{\theta}$, the tendency of $\widehat{\theta}$ to over- or underestimate θ over all possible samples.

What about the third term? Note first that $E\widehat{\theta} - \theta$ is a constant, thus factoring out of the expectation. But for what remains,

$$E(\widehat{\theta} - E\widehat{\theta}) = 0 \tag{1.42}$$

Taking the expected value of both sides of (1.41), and taking the above remarks into account, we have

$$\begin{aligned} \text{MSE}(\widehat{\theta}) &= Var(\widehat{\theta}) + (E\widehat{\theta} - \theta)^2 && (1.43) \\ &= \text{variance} + \text{bias}^2 && (1.44) \end{aligned}$$

In other words:

> The MSE of $\widehat{\theta}$ is equal to the variance of $\widehat{\theta}$ plus squared bias of $\widehat{\theta}$.

1.19.3 μ(t) Minimizes Mean Squared Prediction Error

Claim: *Consider all the functions f() with which we might predict Y from X, i.e.,* $\widehat{Y} = f(X)$. *The one that minimizes mean squared prediction error,* $E[(Y - f(X))^2]$, *is the regression function,* $\mu(t) = E(Y \mid X = t)$.

(Note that the above involves population quantities, not samples. Consider the quantity $E[(Y-f(X))^2]$, for instance. It is the mean squared prediction error (MSPE) over all (X,Y) pairs in the population.)

To derive this, first ask, for any (finite-variance) random variable W, what number c minimizes the quantity $E[(W-c)^2]$? The answer is $c = EW$. To see this, write

$$E[(W-c)^2] = E(W^2 - 2cW + c^2] = E(W^2) - 2cEW + c^2 \tag{1.45}$$

Setting to 0 the derivative of the right-hand side with respect to c, we find that indeed, $c = EW$.

Now use the Law of Total Expectation (Section 1.19.5):

$$\text{MSPE} = E[(Y - f(X))^2] = E\left[E((Y - f(X))^2|X)\right] \tag{1.46}$$

In the inner expectation, X is a constant, and from the statement following (1.45) we know that the minimizing value of $f(X)$ is "EW," in this case $E(Y|X)$, i.e. $\mu(X)$. Since that minimizes the inner expectation for any $\mathbf{X}$, the overall expectation is minimized too.

1.19.4 $\mu(t)$ Minimizes the Misclassification Rate

We are concerned here with the classification context. It shows that if we know the population distribution — we don't, but are going through this exercise to guide our intuition — the conditional mean provides the optimal action in the classification context.

Remember, in this context, $\mu(t) = P(Y \mid X = t)$, i.e. the conditional mean reduces to the conditional probability. Now plug in X for t, and we have the following.

Claim: *Consider all rules based on X that produce a guess $\widehat{Y}$, taking on values 0 and 1. The one that minimizes the overall misclassification rate $P(\widehat{Y} \neq Y)$ is*

$$\widehat{Y} = \begin{cases} 1, & \text{if } \mu(X) > 0.5 \\ 0, & \text{if } \mu(X) \leq 0.5 \end{cases} \tag{1.47}$$

The claim is completely intuitive, almost trivial: After observing X, how

should we guess Y? If conditionally Y has a greater than 50% chance of being 1, then guess it to be 1!

(Note: In some settings, a "false positive" may be worse than a "false negative," or *vice versa*. The reader should ponder how to modify the material here for such a situation. We'll return to this issue in Chapter 5.)

Think of this simple situation: There is a biased coin, with *known* probability of heads p. The coin will be tossed once, and we are supposed to guess the outcome.

Let's name your guess g (a nonrandom constant), and let C denote the as-yet-unknown outcome of the toss (1 for heads, 0 for tails). Then the reader should check that, no matter whether we choose 0 or 1 for g, the probability that we guess correctly is

$$\begin{aligned} P(C = g) &= P(C = 1)g + P(C = 0)(1 - g) && (1.48) \\ &= pg + (1 - p)(1 - g) && (1.49) \\ &= [2p - 1]g + 1 - p && (1.50) \end{aligned}$$

Now remember, p is known. How should we choose g, 0 or 1, in order to maximize (1.50), the probability that our guess is correct? Inspecting (1.50) shows that maximizing that expression will depend on whether $2p-1$ is positive or negative, i.e., whether $p > 0.5$ or not. In the former case we should choose $g = 1$, while in the latter case g should be chosen to be 0.

The above reasoning gives us the very intuitive — actually trivial, when expressed in English — result:

> If the coin is biased toward heads, we should guess heads. If the coin is biased toward tails, we should guess tails.

Now to show the original claim, we use The Law of Total Expectation. This will be discussed in detail in Section 1.19.5, but for now, it says this:

$$E(V) = E[E(V|U)] \tag{1.51}$$

i.e. the expected value of a conditional random variable is the unconditional expectation. In the case where V is an indicator random variable, the above reduces to

$$P(A) = E[P(A \mid U)] \tag{1.52}$$

Returning to our original claim, write

$$P(\widehat{Y} = Y) = E\left[P(\widehat{Y} = Y \mid X)\right] \tag{1.53}$$

In that inner probability, "p" is

$$P(Y = 1 \mid X) = \mu(X) \tag{1.54}$$

which completes the proof.

1.19.5 Some Properties of Conditional Expectation

Since the regression function is defined as a conditional expected value, as in (1.3), for mathematical analysis we'll need some properties. First, a definition.

1.19.5.1 Conditional Expectation As a Random Variable

For any random variables U and V with defined expectation, either of which could be vector-valued, define a new random variable W, as follows. First note that the conditional expectation of V given $U = t$ is a function of t,

$$\mu(t) = E(V \mid U = t) \tag{1.55}$$

This is an ordinary function, just like, say, $\sqrt{t}$. But we can turn that ordinary function into a random variable by plugging in a random variable, say Q, for t: $R = \sqrt{Q}$ is a random variable. Thinking along these lines, we define the *random variable* version of conditional expectation accordingly. In the *function* $\mu(t)$ in (1.55), we plug in U for t:

$$W = E(V|U) = \mu(U) \tag{1.56}$$

This W is a random variable. As a simple example, say we choose a number U at random from the numbers 1 through 5. We then randomly choose a second number V, from the numbers 1 through U. Then

$$\mu(t) = E(V \mid U = t) = \frac{1+t}{2} \tag{1.57}$$

We now form a new random variable $W = (1 + U)/2$.

And, since W is a random variable, we can talk of *its* expected value, which turns out to be an elegant result:

1.19.5.2 The Law of Total Expectation

A property of conditional expected value, proven in many undergraduate probability texts, is

$$E(V) = EW = E[E(V \mid U)] \tag{1.58}$$

The foreboding appearance of this equation belies the fact that it is actually quite intuitive, as follows. Say you want to compute the mean height of all people in the U.S., and you already have available the mean heights in each of the 50 states. You cannot simply take the straight average of those state mean heights, because you need to give more weight to the more populous states. In other words, the national mean height is a *weighted* average of the state means, with the weight for each state being its proportion of the national population.

In (1.58), this corresponds to having V as height and U as state. State coding is an integer-valued random variable, ranging from 1 to 50, so we have

$$\begin{aligned} EV &= E[E(V \mid U)] && (1.59)\\ &= EW && (1.60)\\ &= \sum_{i=1}^{50} P(U = i)\ E(V \mid U = i) && (1.61) \end{aligned}$$

The left-hand side, EV, is the overall mean height in the nation; $E(V \mid U = i)$ is the mean height in state i; and the weights in the weighted average are the proportions of the national population in each state, $P(U = i)$.

Not only can we look at the mean of W, but also its variance. By using the various familiar properties of mean and variance, one can derive a similar relation for variance:

1.19.5.3 Law of Total Variance

For scalar V,

$$Var(V) = E[Var(V|U)] + Var[E(V|U)] \tag{1.62}$$

One might initially guess that we only need the first term. To obtain the national variance in height, we would take the weighted average of the state variances. But this would not take into account that the mean heights vary from state to state, thus also contributing to the national variance in height, hence the second term.

This is proven in Section 2.12.8.3.

1.19.5.4 Tower Property

Now consider conditioning on two variables, say U_1 and U_2. One can show that

$$E\left[E(V|U_1,U_2) \mid U_1\right] = E(V \mid U_1) \tag{1.63}$$

Here is an intuitive interpretation of that in the height example above. Take V, U_1 and U_2 to be height, state and gender, respectively, so that $E(V|U_1,U_2)$ is the mean height of all people in a certain state and of a certain gender. If we then take the mean of all these values for a certain state — i.e. take the average of the two gender-specific means in the state — we get the mean height in the state without regard to gender.

Again, note that we take the straight average of the two gender-specific means, because the two genders have equal proportions. If, say, U_2 were race instead of gender, we would need to compute a *weighted* average of the race-specific means, with the weights being the proportions of the various races in the given state.

This is proven in Section 7.8.1.

1.19.5.5 Geometric View

There is an elegant way to view all of this in terms of abstract vector spaces — (1.58) becomes the Pythagorean Theorem! — which we will address later in Mathematical Complements Sections 2.12.8 and 7.8.1.

1.20 Computational Complements

1.20.1 CRAN Packages

There are thousands of useful contributed R packages available on CRAN, the Comprehensive R Archive Network, *https://cran.r-project.org*. The easiest way to install them is from R's interactive mode, e.g.

```
> install.packages('freqparcoord','~/R')
```

Here I have instructed R to download the **freqparcoord** package, installing it in **~/R**, the directory where I like to store my packages.

(If you are using RStudio or some other indirect interface to R, all this can be done from a menu, rather than using **installing.packages**.)

Official R parlance is *package*, not *library*, even though ironically one loads a package using the **library()** function! For instance,

```
> library(freqparcoord)
```

One can learn about the package in various ways. After loading it, for instance, you can list its objects, such as

```
> ls('package:freqparcoord')
[1] "freqparcoord" "knndens"      "knnreg"
"posjitter"      "regdiag"
[6] "regdiagbas"    "rmixmvnorm"    "smoothz"
"smoothzpred"
```

where we see objects (functions here) **knndens()** and so on. There is the **help()** function, e.g.

```
> help(package=freqparcoord)

Information on package freqparcoord

Description:

Package:        freqparcoord
Version:        1.1.0
Author:         Norm Matloff <normmatloff@gmail.com>
                and Yingkang Xie
                <yingkang.xie@gmail.com>
Maintainer:     Norm Matloff <normmatloff@gmail.com>
```

...

Some packages have *vignettes*, extended tutorials. Type

```
> vignette()
```

to see what's available.

1.20.2 The Function tapply() and Its Cousins

In Section 1.6.2 we had occasion to use R's `tapply()`, a highly useful feature of the language. To explain it, let's start with useful function, **split()**.

Consider this tiny data frame:

```
> x
  gender height
1      m     66
2      f     67
3      m     72
4      f     63
```

Now let's split by gender:

```
> xs <- split(x,x$gender)
> xs
$f
  gender height
2      f     67
4      f     63
5      f     63

$m
  gender height
1      m     66
3      m     72
```

Note the types of the objects:

- **xs** is an R list
- **xs$f** and **xs$m** are data frames, the male and female subsets of **x**

We *could* then find the mean heights for each gender this way:

```
> mean(xs$f$height)
[1] 64.33333
> mean(xs$m$height)
[1] 69
```

But with **tapply()**, we can combine the two operations:

```
> tapply(x$height,x$gender,mean)
       f        m
64.33333 69.00000
```

The first argument of `tapply()` must be a vector, but the function that is applied can be vector-valued. Say we want to find not only the mean but also the standard deviation. We can do this:

```
> tapply(x$height,x$gender,function(w) c(mean(w),sd(w)))
$f
[1] 64.333333  2.309401

$m
[1] 69.000000  4.242641
```

Here our function, which we defined "on the spot," within our call to **tapply()**, produces a vector of two components. We asked **tapply()** to call that function on our vector of heights, doing so separately for each gender.

As noted in the title of this section, **tapply()** has "cousins." Here is a brief overview of some of them:

```
# form a matrix by binding the rows (1,2) and (3,4)
> m <- rbind(1:2,3:4)
> m
     [,1] [,2]
[1,]    1    2
[2,]    3    4
# apply the sum() function to each row
> apply(m,1,sum)
[1] 3 7
# apply the sum() function to each column
> apply(m,2,sum)
[1] 4 6

> l <- list(a = c(3,8), b = 12)
> l
$a
```

```
[1] 3 8
$b
[1] 12
# apply sum() to each element of the list,
# forming a new list
> lapply(l,sum)
$a
[1] 11
$b
[1] 12
# do the same, but try to reduce the result
# to a vector
> sapply(l,sum)
 a  b
11 12
```

1.20.3 The Innards of the k-NN Code

Here are simplified versions of the code:

```
preprocessx <- function(x,kmax,xval=FALSE) {
   result$x <- x
   tmp <- FNN::get.knnx(data=x, query=x, k=kmax+xval)
   nni <- tmp$nn.index
   result$idxs <- nni[,(1+xval):ncol(nni)]
   result$xval <- xval
   result$kmax <- kmax
   class(result) <- 'preknn'
   result
}
```

The code is essentially just a wrapper for calls to the **FNN** package on CRAN, which does nearest-neighbor computation.

```
knnest <- function(y,xdata,k,nearf=meany)
{
   idxs <- xdata$idxs
   idx <- idxs [,1:k]
   # set idxrows[[i]] to row i of idx, the indices of
   # the neighbors of the i-th observation
   idxrows <- matrixtolist(1,idx)
   # now do the kNN smoothing
```

```
    # first, form the neighborhoods
    x <- xdata$x
    xy <- cbind(x,y)
    nycol <- ncol(y)  # how many cols in xy are y?
    # ftn to form one neighborhood (x and y vals)
    form1nbhd <-  function(idxrow) xy[idxrow,]
    # now form all the neighborhoods
    nearxy <-
       lapply(idxrows,function(idxrow) xy[idxrow,])
    # now nearxy[[i]] is the rows of x corresponding to
    # neighbors of x[i,], together with the associated
    # Y values

    # now find the estimated regression function values
    # at each point in the training set
    regest <- sapply(1:nrow(x),
       function(i) nearf(x[i,],nearxy[[i]]))
    regest <-
       if (nycol > 1) t(regest) else as.matrix(regest)
    xdata$regest <- regest
    xdata$nycol <- nycol
    xdata$y <- y
    xdata$k <- k
    class(xdata) <- 'knn'
    xdata
}
```

1.20.4 Function Dispatch

The return value from a call to **lm()** is an object of R's S3 class structure; the class, not surprisingly, is named **'lm'**. It turns out that the functions **coef()** and **vcov()** mentioned in this chapter are actually related to this class, as follows.

Recall our usage, on the baseball player data:

```
> lmout <- lm(mlb$Weight ~ mlb$Height)
> coef(lmout) %*% c(1,72)
          [,1]
[1,] 193.2666
```

The call to **coef** extracted the vector of estimated regression coefficents

(which we also could have obtained as **lmout$coefficents**). But here is what happened behind the scenes:

The R function **coef()** is a *generic function*, which means it's just a placeholder, not a "real" function. When we call it, the R interpreter says,

> This is a generic function, so I need to relay this call to the one associated with this class, **'lm'**. That means I need to check whether we have a function **coef.lm()**. Oh, yes we do, so let's call that.

That relaying action is referred to in R terminology as the original call being *dispatched* to **coef.lm()**.

This is a nice convenience. Consider another generic R function, **plot()**. No matter what object we are working with, the odds are that some kind of plotting function has been written for it. We can just call **plot()** on the given object, and leave it to R to find the proper call. (This includes the **'lm'** class; try it on our **lmout** above!)

Similarly, there are a number of R classes on which **coef()** is defined, and the same is true for **vcov()**.

One generic function we will use quite often, and indeed have already used in this chapter, is **summary()**. As its name implies, it summarizes (what the function's author believes) are the most important characteristics of the object. So, when this generic function is called on an **'lm'** object, the call is dispatched to **summary.lm()**, yielding estimated coefficients, standard errors and so on.

Another generic function to be used often here is **predict()**, from Section 1.10.3. In the example there, **lmout** was of class **'lm'**, so the call to **predict()** was dispatched to **predict.lm()**.

1.21 Centering and Scaling

It is common in many statistical methods to *center and scale* the data. Here we subtract from each variable the sample mean of that variable. This process is called *centering*. Typically one also *scales* each predictor, i.e. divides each predictor by its sample standard deviation. Now all variables will have mean 0 and standard deviation 1.

It is clear that this is very useful for k-NN regression. Consider the ex-

ample later in this chapter involving Census data. Without at least scaling, variables that are very large, such as income, would dominate the nearest-neighbor computations, and small but important variables such as age would essentially be ignored. The **knnest()** function that we will be using does do centering and scaling as preprocessing for the predictor variables.

In a parametric setting such as linear models, centering and scaling has the goal of reducing numerical roundoff error.

In R, the centering/scaling operation is done with the **scale()** function. In order to be able to reverse the process later, the means and standard deviations are recorded as R *attributes*:

```
> m <- rbind(1:2,3:4)
> m
     [,1] [,2]
[1,]    1    2
[2,]    3    4
> m1 <- scale(m)
> m1
           [,1]       [,2]
[1,] -0.7071068 -0.7071068
[2,]  0.7071068  0.7071068
attr(,"scaled:center")
[1] 2 3
attr(,"scaled:scale")
[1] 1.414214 1.414214
> attr(m1,'scaled:center')
[1] 2 3
```

1.22 Exercises: Data, Code and Math Problems

Data problems:

1. In Section 1.12.1.2, the reader was reminded that the results of a cross-validation are random, due to the random partitioning into training and test sets. Try doing several runs of the linear and k-NN code in that section, comparing results.

2. Extend (1.28) to include interaction terms for age and gender, and age^2

and gender. Run the new model, and find the estimated effect of being female, for a 32-year-old person with a Master's degree.

3. Consider the **bodyfat** data mentioned in Section 1.2. Use **lm()** to form a prediction equation for **density** from the other variables (skipping the first three), and comment on whether use of indirect methods in this way seems feasible.

4. In Section 1.19.5.2, we gave this intuitive explanation:

> In other words, the national mean height is a *weighted* average of the state means, with the weight for each state being its proportion of the national population. Replace state by gender in the following.

(a) Write English prose that relates the overall mean height of people and the gender-specific mean heights.

(b) Write English prose that relates the overall proportion of people taller than 70 inches to the gender-specific proportions.

Mini-CRAN and other computational problems:

5. In Section 1.12, we used R's negative-index capability to form the training/test set partitioning. Show how we could use the R function **setdiff()** to do this as an alternate approach.

6. We saw in this chapter, e.g., in Figure 1.3, how R's **abline()** function can be used to add a straight line to a plot. What about adding a quadratic function?

(a) Write an R function with call form

```
abccurve(coef,xint)
```

where **coef** is a vector of the coeficients a, b and c in the polynomial

$$a + bt + ct^2 \tag{1.64}$$

and **xint** is a 2-element vector that gives the range of the horizontal axis for t. The function superimposes the quadratic curve onto the existing graph. Hint: Use R's **curve()** function.

(b) Fit a quadratic model to the click-through data, and use your **abccurve()** function on the scatter plot for that data.

Math problems:

7. Suppose the joint density of (X, Y) is $3s^2e^{-st}$, $1 < s < 2, 0 < t < \infty$. Find the regression function $\mu(s) = E(Y|X = s)$.

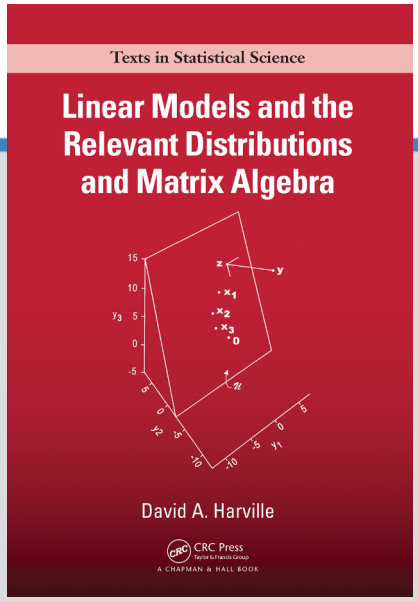
8. For (X, Y) in the notation of Section 1.19.3, show that the predicted value $\mu(X)$ and the predicton error $Y - \mu(X)$ are uncorrelated.

9. Suppose X is a scalar random variable with density g. We are interested in the nearest neighbors to a point t, based on a random sample $X_1, ..., X_n$ from g. Find L_k, the cumulative distribution function of the distance of the k^{th}-nearest neighbor to t.

CHAPTER

2

CLASSICAL APPROACH



This chapter is excerpted from

Linear Models and the Relevant Distributions and Matrix Algebra

by David A. Harville.

Learn more

5

Estimation and Prediction: Classical Approach

Models of the form of the general linear model, and in particular those of the form of the Gauss–Markov or Aitken model, are often used to obtain point estimates of the unobservable quantities represented by various parametric functions. In many cases, the parametric functions are ones that are expressible in the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$, where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_P)'$ is a P-dimensional column vector of constants, or equivalently ones that are expressible in the form $\sum_{j=1}^{P} \lambda_j \beta_j$. Models of the form of the G–M, Aitken, or general linear model may also be used to obtain predictions for future quantities; these would be future quantities that are represented by unobservable random variables with expected values of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$. The emphasis in this chapter is on the G–M model (in which the only parameter other than $\beta_1, \beta_2, \ldots, \beta_P$ is the standard deviation σ) and on what might be regarded as a classical approach to estimation and prediction.

5.1 Linearity and Unbiasedness

Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows the G–M, Aitken, or general linear model, and consider the estimation of a parametric function of the form $\boldsymbol{\lambda}'\boldsymbol{\beta} = \sum_{j=1}^{P} \lambda_j \beta_j$, where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_P)'$. Is it "possible" to estimate $\boldsymbol{\lambda}'\boldsymbol{\beta}$ from the available information, and if so, which estimator is best and in what sense is it best? One way to judge the "goodness" of an estimator is on the basis of its mean squared error (MSE) or its root mean squared error (root MSE)—the root MSE is the square root of the MSE. When a function $t(\mathbf{y})$ of $\mathbf{y}$ is regarded as an estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$, its MSE is (by definition) $\mathrm{E}\{[t(\mathbf{y}) - \boldsymbol{\lambda}'\boldsymbol{\beta}]^2\}$.

The information about the distribution of $\mathbf{y}$ provided by the G–M, Aitken, or general linear model is limited; it is confined to information about $\mathrm{E}(\mathbf{y})$ and $\mathrm{var}(\mathbf{y})$. The evaluation and comparison of potential estimators are greatly facilitated by restricting attention to estimators that are of a relatively simple form or that satisfy certain criteria and/or by making assumptions about the distribution of $\mathbf{y}$ that go beyond those inherent in the G–M, Aitken, or general linear model. If the evaluations and comparisons are to be meaningful, the restrictions need to be ones that have appeal in their own right, and the assumptions need to be realistic.

An estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$, say an estimator $t(\mathbf{y})$, is said to be *linear* if it is expressible in the form

$$t(\mathbf{y}) = c + \sum_{i=1}^{N} a_i y_i,$$

where c and $a_1, a_2, \ldots, a_N$ are constants, or equivalently if it is expressible in the form

$$t(\mathbf{y}) = c + \mathbf{a}'\mathbf{y},$$

where c is a constant and $\mathbf{a} = (a_1, a_2, \ldots, a_N)'$ is an N-dimensional column vector of constants. Linear estimators of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ are of a relatively simple form, which makes them readily amenable to evaluation, comparison, and interpretation. Accordingly, it is convenient and of some interest to obtain results on the estimation of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ in the special case where consideration is restricted to linear estimators.

Attention is sometimes restricted to estimators that are unbiased. By definition, an estimator $t(\mathbf{y})$ of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is unbiased if $\mathrm{E}[t(\mathbf{y})] = \boldsymbol{\lambda}'\boldsymbol{\beta}$. If $t(\mathbf{y})$ is an unbiased estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$, then

$$\mathrm{E}\{[t(\mathbf{y}) - \boldsymbol{\lambda}'\boldsymbol{\beta}]^2\} = \mathrm{var}[t(\mathbf{y})], \tag{1.1}$$

that is, its MSE equals its variance.

In the case of a linear estimator $c + \mathbf{a}'\mathbf{y}$, the expected value of the estimator is

$$\mathrm{E}(c + \mathbf{a}'\mathbf{y}) = c + \mathbf{a}'\mathrm{E}(\mathbf{y}) = c + \mathbf{a}'\mathbf{X}\boldsymbol{\beta}. \tag{1.2}$$

Accordingly, $c + \mathbf{a}'\mathbf{y}$ is an unbiased estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ if and only if, for every P-dimensional column vector $\underline{\boldsymbol{\beta}}$,

$$c + \mathbf{a}'\mathbf{X}\underline{\boldsymbol{\beta}} = \boldsymbol{\lambda}'\underline{\boldsymbol{\beta}}. \tag{1.3}$$

Clearly, a sufficient condition for the unbiasedness of the linear estimator $c + \mathbf{a}'\mathbf{y}$ is

$$c = 0 \quad \text{and} \quad \mathbf{a}'\mathbf{X} = \boldsymbol{\lambda}' \tag{1.4}$$

or, equivalently,

$$c = 0 \quad \text{and} \quad \mathbf{X}'\mathbf{a} = \boldsymbol{\lambda}. \tag{1.5}$$

This condition is also a necessary condition for the unbiasedness of $c + \mathbf{a}'\mathbf{y}$ as is evident upon observing that if equality (1.3) holds for every column vector $\underline{\boldsymbol{\beta}}$ in $\mathcal{R}^P$, then it holds in particular when $\underline{\boldsymbol{\beta}}$ is taken to be the $P \times 1$ null vector $\mathbf{0}$ (so that $c = 0$) and when (for each integer j between 1 and P, inclusive) $\underline{\boldsymbol{\beta}}$ is taken to be the jth column of $\mathbf{I}_P$ (so that the jth element of $\mathbf{a}'\mathbf{X}$ equals the jth element of $\boldsymbol{\lambda}'$).

In the special case of a linear unbiased estimator $\mathbf{a}'\mathbf{y}$, expression (1.1) for the MSE of an unbiased estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ simplifies to

$$\mathrm{E}[(\mathbf{a}'\mathbf{y} - \boldsymbol{\lambda}'\boldsymbol{\beta})^2] = \mathbf{a}'\mathrm{var}(\mathbf{y})\mathbf{a}. \tag{1.6}$$

5.2 Translation Equivariance

Suppose (as in Section 5.1) that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows the Gauss–Markov, Aitken, or general linear model and that we wish to estimate a parametric function of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$. Attention is sometimes restricted to estimators that are unbiased. However, unbiasedness is not the only criterion that could be used to restrict the quantity of estimators under consideration. Another possible criterion is *translation equivariance* (also known as location equivariance).

Let $\mathbf{k}$ represent a P-dimensional column vector of known constants, and define $\mathbf{z} = \mathbf{y} + \mathbf{X}\mathbf{k}$. The vector $\mathbf{z}$, like the vector $\mathbf{y}$, is an N-dimensional observable random vector. Moreover,

$$\mathbf{z} = \mathbf{X}\boldsymbol{\tau} + \mathbf{e}, \tag{2.1}$$

where $\boldsymbol{\tau} = \boldsymbol{\beta} + \mathbf{k}$. Accordingly, $\mathbf{z}$ follows a G–M, Aitken, or general linear model that is identical in all respects to the model followed by $\mathbf{y}$, except that the role of the parameter vector $\boldsymbol{\beta}$ is played by a vector [represented by $\boldsymbol{\tau}$ in equality (2.1)] that has a different interpretation.

It can be argued that an estimator, say $t(\mathbf{y})$, of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ should be such that the results obtained in using $t(\mathbf{y})$ to estimate $\boldsymbol{\lambda}'\boldsymbol{\beta}$ are consistent with those obtained in using $t(\mathbf{z})$ to estimate the corresponding parametric function ($\boldsymbol{\lambda}'\boldsymbol{\tau} = \boldsymbol{\lambda}'\boldsymbol{\beta} + \boldsymbol{\lambda}'\mathbf{k}$). Here, the consistency is in the sense that

$$t(\mathbf{y}) + \boldsymbol{\lambda}'\mathbf{k} = t(\mathbf{z})$$

or, equivalently, that

$$t(\mathbf{y}) + \boldsymbol{\lambda}'\mathbf{k} = t(\mathbf{y} + \mathbf{X}\mathbf{k}). \tag{2.2}$$

When applied to a linear estimator $c + \mathbf{a}'\mathbf{y}$, condition (2.2) becomes

$$c + \mathbf{a}'\mathbf{y} + \boldsymbol{\lambda}'\mathbf{k} = c + \mathbf{a}'(\mathbf{y} + \mathbf{X}\mathbf{k}),$$

which (after some simplification) can be restated as

$$\mathbf{a}'\mathbf{X}\mathbf{k} = \boldsymbol{\lambda}'\mathbf{k}. \tag{2.3}$$

The estimator $t(\mathbf{y})$ is said to be *translation equivariant* if it is such that condition (2.2) is satisfied for every $\mathbf{k} \in \mathcal{R}^P$ (and for every value of $\mathbf{y}$). Accordingly, the linear estimator $c + \mathbf{a}'\mathbf{y}$ is translation equivariant if and only if condition (2.3) is satisfied for every $\mathbf{k} \in \mathcal{R}^P$ or, equivalently, if and only if

$$\mathbf{a}'\mathbf{X} = \boldsymbol{\lambda}'. \tag{2.4}$$

Observe (in light of the results of Section 5.1) that condition (2.4) is identical to one of the conditions needed for unbiasedness—for unbiasedness, we also need the condition $c = 0$. Thus, the motivation for requiring that the coefficient vector $\mathbf{a}'$ in the linear estimator $c + \mathbf{a}'\mathbf{y}$ satisfy the condition $\mathbf{a}'\mathbf{X} = \boldsymbol{\lambda}'$ can come from a desire to achieve unbiasedness or translation equivariance or both.

5.3 Estimability

Suppose (as in Sections 5.1 and 5.2) that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows the G–M, Aitken, or general linear model, and consider the estimation of a parametric function that is expressible in the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ or $\sum_{j=1}^{P} \lambda_j \beta_j$, where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_P)'$ is a $P \times 1$ vector of coefficients. If there exists a linear unbiased estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ [i.e., if there exists a constant c and an $N \times 1$ vector of constants $\mathbf{a}$ such that $\mathrm{E}(c + \mathbf{a}'\mathbf{y}) = \boldsymbol{\lambda}'\boldsymbol{\beta}$], then $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is said to be *estimable*. Otherwise (if no such estimator exists), $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is said to be *nonestimable*.

If $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, then the data provide at least some information about $\boldsymbol{\lambda}'\boldsymbol{\beta}$. Estimability can be of critical importance in the design of an experiment. If the data from the experiment are to be regarded as having originated from a G–M, Aitken, or general linear model and if the quantities of interest are to be formulated as parametric functions of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ (as is common practice), then it is imperative that every one of the relevant functions be estimable.

It follows immediately from the results of Section 5.1 that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable if and only if there exists an $N \times 1$ vector $\mathbf{a}$ such that

$$\boldsymbol{\lambda}' = \mathbf{a}'\mathbf{X} \tag{3.1}$$

or, equivalently, such that

$$\boldsymbol{\lambda} = \mathbf{X}'\mathbf{a}. \tag{3.2}$$

Thus, for $\boldsymbol{\lambda}'\boldsymbol{\beta}$ to be estimable (under the G–M, Aitken, or general linear model), it is necessary and sufficient that

$$\boldsymbol{\lambda}' \in \mathcal{R}(\mathbf{X}) \tag{3.3}$$

or, equivalently, that

$$\boldsymbol{\lambda} \in \mathcal{C}(\mathbf{X}'). \tag{3.4}$$

Note that it follows from the very definition of estimability [as well as from condition (3.1)] that if $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, then there exists an $N \times 1$ vector $\mathbf{a}$ such that

$$\boldsymbol{\lambda}'\boldsymbol{\beta} = \mathbf{a}'\,\mathrm{E}(\mathbf{y}). \tag{3.5}$$

Thus, if $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, it is interpretable in terms of the expected values of $y_1, y_2, \ldots, y_N$. In fact, if $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, it may be expressible in the form (3.5) for each of a number of different

choices of $\mathbf{a}$ and, consequently, it may have multiple interpretations in terms of the expected values of $y_1, y_2, \ldots, y_N$.

Two basic and readily verifiable observations about linear combinations of parametric functions of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ are as follows:

(1) linear combinations of estimable functions are estimable; and

(2) linear combinations of nonestimable functions are not necessarily nonestimable.

How many "essentially different" estimable functions are there? Let $\boldsymbol{\lambda}_1'\boldsymbol{\beta}, \boldsymbol{\lambda}_2'\boldsymbol{\beta}, \ldots, \boldsymbol{\lambda}_K'\boldsymbol{\beta}$ represent K (where K is an arbitrary positive integer) linear combinations of the elements of $\boldsymbol{\beta}$. These linear combinations are said to be linearly independent if their coefficient vectors $\boldsymbol{\lambda}_1', \boldsymbol{\lambda}_2', \ldots, \boldsymbol{\lambda}_K'$ are linearly independent vectors. A question as to whether $\boldsymbol{\lambda}_1'\boldsymbol{\beta}, \boldsymbol{\lambda}_2'\boldsymbol{\beta}, \ldots, \boldsymbol{\lambda}_K'\boldsymbol{\beta}$ are essentially different can be made precise by taking *essentially different* to mean *linearly independent*.

Letting $R = \text{rank}(\mathbf{X})$, some basic and readily verifiable observations about linearly independent parametric functions of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ and about their estimability or nonestimability are as follows:

(1) there exists a set of R linearly independent estimable functions;

(2) no set of estimable functions contains more than R linearly independent estimable functions; and

(3) if the model is not of full rank (i.e., if $R < P$), then at least one and, in fact, at least $P - R$ of the individual parameters $\beta_1, \beta_2, \ldots, \beta_P$ are nonestimable.

When the model matrix $\mathbf{X}$ has full column rank P, the model is said to be of *full rank*. In the special case of a full-rank model, $\mathcal{R}(\mathbf{X}) = \mathcal{R}^P$, and every parametric function of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable.

Note that the existence of an $N \times 1$ vector $\mathbf{a}$ that satisfies equality (3.2) is equivalent to the consistency of a linear system (in an $N \times 1$ vector $\mathbf{a}$ of unknowns), namely, the linear system with coefficient matrix $\mathbf{X}'$ (which is of dimensions $P \times N$) and with right side $\boldsymbol{\lambda}$. The significance of this equivalence is that any result on the consistency of a linear system can be readily translated into a result on the estimability of the parametric function $\boldsymbol{\lambda}'\boldsymbol{\beta}$. Consider, in particular, Theorem 2.11.1. Upon applying this theorem [and observing that $(\mathbf{X}^-)'$ is a generalized inverse of $\mathbf{X}'$], we find that for $\boldsymbol{\lambda}'\boldsymbol{\beta}$ to be estimable, it is necessary and sufficient that

$$\boldsymbol{\lambda}'\mathbf{X}^-\mathbf{X} = \boldsymbol{\lambda}' \tag{3.6}$$

or, equivalently, that

$$\boldsymbol{\lambda}'(\mathbf{I} - \mathbf{X}^-\mathbf{X}) = \mathbf{0}. \tag{3.7}$$

If $\text{rank}(\mathbf{X}) = P$, then (in light of Lemma 2.10.3) $\mathbf{X}^-\mathbf{X} = \mathbf{I}$. Thus, in the special case of a full-rank model, conditions (3.6) and (3.7) are vacuous.

a. A result on the consistency of a linear system

The following result on the consistency of a linear system can be used to obtain additional results on the estimability of a parametric function of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$.

Theorem 5.3.1. A linear system $\mathbf{AX} = \mathbf{B}$ (in $\mathbf{X}$) is consistent if and only if $\mathbf{k}'\mathbf{B} = \mathbf{0}$ for every column vector $\mathbf{k}$ (of compatible dimension) such that $\mathbf{k}'\mathbf{A} = \mathbf{0}$.

Proof. Denote by M the number of rows in $\mathbf{A}$ (and in $\mathbf{B}$), let $\mathbf{k}$ represent an M-dimensional column vector, and observe (in light of Corollary 2.11.4 and Lemma 2.10.10) that

$$\begin{aligned}
\mathbf{k}'\mathbf{A} = \mathbf{0} &\Leftrightarrow \mathbf{A}'\mathbf{k} = \mathbf{0} \\
&\Leftrightarrow \mathbf{k} \in \mathcal{N}(\mathbf{A}') \\
&\Leftrightarrow \mathbf{k} \in \mathcal{C}[\mathbf{I} - (\mathbf{A}^-)'\mathbf{A}'] \\
&\Leftrightarrow \mathbf{k} = [\mathbf{I} - (\mathbf{A}^-)'\mathbf{A}']\mathbf{r} \text{ for some } M \times 1 \text{ vector } \mathbf{r} \\
&\Leftrightarrow \mathbf{k}' = \mathbf{r}'(\mathbf{I} - \mathbf{A}\mathbf{A}^-) \text{ for some } M \times 1 \text{ vector } \mathbf{r}.
\end{aligned}$$

Thus, recalling Lemma 2.2.2, we find that

$$\begin{aligned}\mathbf{k}'\mathbf{B} = \mathbf{0} \text{ for every } \mathbf{k} \text{ such that } \mathbf{k}'\mathbf{A} = \mathbf{0} \\ \Leftrightarrow \quad \mathbf{r}'(\mathbf{I} - \mathbf{A}\mathbf{A}^-)\mathbf{B} = \mathbf{0} \text{ for every } M \times 1 \text{ vector } \mathbf{r} \\ \Leftrightarrow \quad (\mathbf{I} - \mathbf{A}\mathbf{A}^-)\mathbf{B} = \mathbf{0}.\end{aligned}$$

And based on Theorem 2.11.1, we conclude that the linear system $\mathbf{A}\mathbf{X} = \mathbf{B}$ is consistent if and only if $\mathbf{k}'\mathbf{B} = \mathbf{0}$ for every $\mathbf{k}$ such that $\mathbf{k}'\mathbf{A} = \mathbf{0}$. Q.E.D.

Theorem 5.3.1 establishes that the consistency of the linear system $\mathbf{A}\mathbf{X} = \mathbf{B}$ is equivalent to a condition that is sometimes referred to as compatibility; the linear system $\mathbf{A}\mathbf{X} = \mathbf{B}$ is said to be compatible if every linear relationship that exists among the rows of the coefficient matrix $\mathbf{A}$ also exists among the rows of the right side $\mathbf{B}$ (in the sense that $\mathbf{k}'\mathbf{A} = \mathbf{0} \Rightarrow \mathbf{k}'\mathbf{B} = \mathbf{0}$). The proof presented herein differs from that presented in *Matrix Algebra from a Statistician's Perspective* (Harville 1997, sec. 7.3); it makes use of results on generalized inverses.

b. Some alternative necessary and sufficient conditions for estimability

Let us now consider further the estimability of a parametric function of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ (under the G–M, Aitken, or general linear model). As noted earlier, $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable if and only if the linear system $\mathbf{X}'\mathbf{a} = \boldsymbol{\lambda}$ (in $\mathbf{a}$) is consistent. Accordingly, it follows from Theorem 5.3.1 that for $\boldsymbol{\lambda}'\boldsymbol{\beta}$ to be estimable, it is necessary and sufficient that

$$\mathbf{k}'\boldsymbol{\lambda} = 0 \text{ for every } P \times 1 \text{ vector } \mathbf{k} \text{ such that } \mathbf{k}'\mathbf{X}' = \mathbf{0} \tag{3.8}$$

or, equivalently, that

$$\mathbf{k}'\boldsymbol{\lambda} = 0 \text{ for every } P \times 1 \text{ vector } \mathbf{k} \text{ in } \mathcal{N}(\mathbf{X}). \tag{3.9}$$

Let $S = \dim[\mathcal{N}(\mathbf{X})]$. And observe (in light of Lemma 2.11.5) that

$$S = P - \operatorname{rank}(\mathbf{X}).$$

Unless the model is of full rank [in which case $S = 0$, $\mathcal{N}(\mathbf{X}) = \{\mathbf{0}\}$, and conditions (3.8) and (3.9) are vacuous], condition (3.9) comprises an infinite number of equalities—there is one equality for each vector $\mathbf{k}$ in the S-dimensional linear space $\mathcal{N}(\mathbf{X})$. Fortunately, all but S of the equalities that form condition (3.9) can be eliminated without affecting the necessity or sufficiency of the condition.

Let $\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_S$ represent any S linearly independent vectors in $\mathcal{N}(\mathbf{X})$, that is, any S linearly independent (P-dimensional) column vectors such that $\mathbf{X}\mathbf{k}_1 = \mathbf{X}\mathbf{k}_2 = \cdots = \mathbf{X}\mathbf{k}_S = \mathbf{0}$. Then, for $\boldsymbol{\lambda}'\boldsymbol{\beta}$ to be estimable, it is necessary and sufficient that

$$\mathbf{k}_1'\boldsymbol{\lambda} = \mathbf{k}_2'\boldsymbol{\lambda} = \cdots = \mathbf{k}_S'\boldsymbol{\lambda} = 0. \tag{3.10}$$

To verify the necessity and sufficiency of condition (3.10), it suffices to establish that condition (3.10) is equivalent to condition (3.9). In fact, it is enough to establish that condition (3.10) implies condition (3.9)—that condition (3.9) implies condition (3.10) is obvious. Accordingly, let $\mathbf{k}$ represent an arbitrary member of $\mathcal{N}(\mathbf{X})$. And observe (in light of Theorem 2.4.11) that the set $\{\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_S\}$ is a basis for $\mathcal{N}(\mathbf{X})$, implying the existence of scalars $a_1, a_2, \ldots, a_S$ such that

$$\mathbf{k} = a_1\mathbf{k}_1 + a_2\mathbf{k}_2 + \cdots + a_S\mathbf{k}_S$$

and hence such that

$$\mathbf{k}'\boldsymbol{\lambda} = a_1\mathbf{k}_1'\boldsymbol{\lambda} + a_2\mathbf{k}_2'\boldsymbol{\lambda} + \cdots + a_S\mathbf{k}_S'\boldsymbol{\lambda}.$$

Thus, if $\mathbf{k}_1'\boldsymbol{\lambda} = \mathbf{k}_2'\boldsymbol{\lambda} = \cdots = \mathbf{k}_S'\boldsymbol{\lambda} = 0$, then $\mathbf{k}'\boldsymbol{\lambda} = 0$, leading to the conclusion that condition (3.10) implies condition (3.9).

Condition (3.10) comprises only S of the infinite number of equalities that form condition (3.9), making it much easier to administer than condition (3.9).

c. A related concept: identifiability

Let us continue to suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M, Aitken, or general linear model. Recall that $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$. Does a parametric function of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ have a fixed value for each value of $\mathrm{E}(\mathbf{y})$? This question can be restated more formally as follows: is $\boldsymbol{\lambda}'\boldsymbol{\beta}_1 = \boldsymbol{\lambda}'\boldsymbol{\beta}_2$ for every pair of P-dimensional column vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ such that $\mathbf{X}\boldsymbol{\beta}_1 = \mathbf{X}\boldsymbol{\beta}_2$? Or, equivalently, is $\mathbf{X}\boldsymbol{\beta}_1 \neq \mathbf{X}\boldsymbol{\beta}_2$ for every pair of P-dimensional column vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ such that $\boldsymbol{\lambda}'\boldsymbol{\beta}_1 \neq \boldsymbol{\lambda}'\boldsymbol{\beta}_2$? Unless the model is of full rank, the answer depends on the coefficient vector $\boldsymbol{\lambda}'$. When the answer is yes, the parametric function $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is said to be *identifiable*—this terminology is consistent with that of Hinkelmann and Kempthorne (2008, sec. 4.4).

The parametric function $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is identifiable if and only if it is estimable. To see this, suppose that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable. Then, $\boldsymbol{\lambda}' = \mathbf{a}'\mathbf{X}$ for some column vector $\mathbf{a}$. And for any P-dimensional column vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ such that $\mathbf{X}\boldsymbol{\beta}_1 = \mathbf{X}\boldsymbol{\beta}_2$,

$$\boldsymbol{\lambda}'\boldsymbol{\beta}_1 = \mathbf{a}'\mathbf{X}\boldsymbol{\beta}_1 = \mathbf{a}'\mathbf{X}\boldsymbol{\beta}_2 = \boldsymbol{\lambda}'\boldsymbol{\beta}_2.$$

Accordingly, $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is identifiable.

Conversely, suppose that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is identifiable. Then, $\boldsymbol{\lambda}'\boldsymbol{\beta}_1 = \boldsymbol{\lambda}'\mathbf{0}$ for every P-dimensional column vector $\boldsymbol{\beta}_1$ such that $\mathbf{X}\boldsymbol{\beta}_1 = \mathbf{X}\mathbf{0}$, or equivalently $\boldsymbol{\lambda}'\mathbf{k} = 0$ for every vector $\mathbf{k}$ in $\mathfrak{N}(\mathbf{X})$. And based on the results on estimability established in Subsection b (and on the observation that $\boldsymbol{\lambda}'\mathbf{k} = \mathbf{k}'\boldsymbol{\lambda}$), we conclude that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable.

d. Polynomials (in 1 variable)

Suppose that $\mathbf{y}$ is an N-dimensional observable random vector that follows a G–M, Aitken, or general linear model. Suppose further that there is a single explanatory variable, so that $C = 1$ and $\mathbf{u} = (u_1)$. And (for the sake of simplicity) let us write u for u_1 or (depending on the context) for $\mathbf{u}$.

Let us consider the case (considered initially in Section 4.2a) where $\delta(u)$ is a polynomial. Specifically, let us consider the case where

$$\delta(u) = \beta_1 + \beta_2 u + \beta_3 u^2 + \cdots + \beta_P u^{P-1}. \tag{3.11}$$

Under what circumstances are all P of the coefficients $\beta_1, \beta_2, \ldots, \beta_P$ estimable? Or, equivalently, under what circumstances is the model of full rank? The answer to this question can be established with the help of a result on a kind of matrix known as a Vandermonde matrix.

Vandermonde matrices. A Vandermonde matrix is a square matrix $\mathbf{A}$ of the general form

$$\mathbf{A} = \begin{pmatrix} 1 & t_1 & t_1^2 & \ldots & t_1^{K-1} \\ 1 & t_2 & t_2^2 & \ldots & t_2^{K-1} \\ 1 & t_3 & t_3^2 & \ldots & t_3^{K-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_K & t_K^2 & \ldots & t_K^{K-1} \end{pmatrix},$$

where $t_1, t_2, t_3, \ldots, t_K$ are arbitrary scalars. The determinant of a Vandermonde matrix is obtainable from the formula

$$|\mathbf{A}| = \prod_{\substack{i,j \\ (j<i)}} (t_i - t_j). \tag{3.12}$$

For a derivation of this formula, refer, for example, to Harville (1997, sec. 13.6).

Denote by R the number of distinct values represented among $t_1, t_2, t_3, \ldots, t_K$. Then, it follows from result (3.12) that

$$R = K \quad \Leftrightarrow \quad |\mathbf{A}| \neq 0$$

and hence (in light of Theorem 2.14.20) that

$$R = K \quad \Leftrightarrow \quad \mathbf{A} \text{ is nonsingular.} \tag{3.13}$$

Rank of the model matrix. We are now in a position to determine the rank of the model matrix $\mathbf{X}$ (of a G–M, Aitken, or general linear model) in the special case where $C = 1$ and where $\delta(u)$ is the polynomial (3.11) (which is of degree $P - 1$ in the lone explanatory variable u). Denote by D the number of distinct values of u represented among the N values of u corresponding to the N observable random variables $y_1, y_2, \ldots, y_N$. And take $i_1, i_2, \ldots, i_D$ $(i_1 < i_2 < \cdots < i_D)$ to be integers between 1 and N, inclusive, such that the values of u corresponding to $y_{i_1}, y_{i_2}, \ldots, y_{i_D}$ are distinct. Each of the N rows of $\mathbf{X}$ is either among its $i_1, i_2, \ldots, i_D$th rows or is a duplicate of one of those rows. Thus, $\mathcal{R}(\mathbf{X})$ is spanned by the $i_1, i_2, \ldots, i_D$th rows of $\mathbf{X}$, and it follows that $\text{rank}(\mathbf{X}) \leq D$ and hence [since $\text{rank}(\mathbf{X}) \leq P$] that $\text{rank}(\mathbf{X}) \leq M$, where $M = \min(D, P)$. Moreover, it follows from result (3.13) that the $M \times M$ submatrix of $\mathbf{X}$ formed from its $i_1, i_2, \ldots, i_M$th rows and its first M columns is nonsingular, implying (in light of Theorem 2.4.19) that $\text{rank}(\mathbf{X}) \geq M$. And we conclude that

$$\text{rank}(\mathbf{X}) = \min(D, P). \tag{3.14}$$

In light of result (3.14), it is evident that [in the special case where $\delta(u)$ is the $(P-1)$-degree polynomial (3.11)] the model is of full rank if and only if $D \geq P$, that is, if and only if at least P of the N values of u (the N values of u corresponding to $y_1, y_2, \ldots, y_N$) are distinct. When the model is of full rank, all P of the coefficients $\beta_1, \beta_2, \ldots, \beta_P$ [in the $(P-1)$-degree polynomial] are estimable.

In the application to the ouabain data (of Section 4.2b), there are 4 distinct values of the explanatory variable, representing the 4 different rates of injection (or perhaps the logarithms of the 4 different rates). Accordingly, if $\delta(u)$ were taken to be a polynomial of the form (3.11), the model would be of full rank if and only if the degree of the polynomial were taken to be 3 or less.

e. An illustration: mixture data

Consider an application of the G–M, Aitken, or general linear model in which the C explanatory variables $u_1, u_2, \ldots, u_C$ represent the proportionate amounts of the ingredients in a mixture. By their very nature, the explanatory variables are such that

$$\sum_{j=1}^{C} u_j = 1. \tag{3.15}$$

Now, suppose that

$$\delta(\mathbf{u}) = \beta_1 + \beta_2 u_1 + \beta_3 u_2 + \cdots + \beta_{C+1} u_C \tag{3.16}$$

(in which case, $P = C + 1$). And let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{C+1}$ represent the first through last columns of the model matrix $\mathbf{X}$. Then,

$$\mathbf{x}_1 = \sum_{j=2}^{C+1} \mathbf{x}_j \tag{3.17}$$

or, equivalently,

$$\mathbf{X}\begin{pmatrix} -1 \\ \mathbf{1}_C \end{pmatrix} = \mathbf{0}. \tag{3.18}$$

Suppose, for example, that the mixtures are fruit juices, each of which is a blend of watermelon juice, orange juice, pineapple juice, and grapefruit juice (in which case, the data points might be flavor scores). Suppose further that $N = 6$ and that the data are those obtained for the following 6 blends:

u_1	u_2	u_3	u_4
0.8	0.2	0	0
0.2	0	0.8	0
0.4	0	0	0.6
0.5	0.1	0.4	0
0.6	0.1	0	0.3
0.3	0	0.4	0.3

What parametric functions of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ are estimable? In light of equality (3.18), $\text{rank}(\mathbf{X}) \leq C$, and it follows from the results of Subsection b that for $\boldsymbol{\lambda}'\boldsymbol{\beta}$ to be estimable, it is necessary that

$$\begin{pmatrix} -1 \\ \mathbf{1}_C \end{pmatrix}' \boldsymbol{\lambda} = 0 \tag{3.19}$$

or, equivalently, that

$$\sum_{j=2}^{C+1} \lambda_j = \lambda_1. \tag{3.20}$$

Is this condition sufficient as well as necessary? The answer to this question depends on $\text{rank}(\mathbf{X})$. It follows from the results of Subsection b that if $\text{rank}(\mathbf{X}) = C$, then condition (3.20) is sufficient (as well as necessary) for the estimability of $\boldsymbol{\lambda}'\boldsymbol{\beta}$; however, if $\text{rank}(\mathbf{X}) < C$, then condition (3.20) is not (in and of itself) sufficient.

In the case of the 6 blends of the 4 juices,

$$\text{rank}(\mathbf{X}) = 3 = C - 1 < C.$$

To see this, observe that the last 3 columns of $\mathbf{X}$ (which contain the values of u_2, u_3, and u_4, respectively) are linearly independent, so that $\text{rank}(\mathbf{X}) \geq 3$. Observe also that each of the 6 blends of the 4 juices is such that

$$u_4 = 1.5u_1 - 6u_2 - 0.375u_3,$$

so that (in the case of the 6 blends of the 4 juices)

$$\mathbf{x}_5 = 1.5\mathbf{x}_2 - 6\mathbf{x}_3 - 0.375\mathbf{x}_4, \tag{3.21}$$

which [together with result (3.17)] implies that the first and last columns of $\mathbf{X}$ are expressible as linear combinations of the other 3 columns and hence that $\text{rank}(\mathbf{X}) \leq 3$. Thus, $\text{rank}(\mathbf{X}) = 3$.

To obtain conditions that are both necessary and sufficient for the estimability of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ (from the information provided by the data on the 6 blends of the 4 juices), observe that equality (3.21) can be reexpressed in the form

$$\mathbf{X}\begin{pmatrix} 0 \\ 1.5 \\ -6 \\ -0.375 \\ -1 \end{pmatrix} = \mathbf{0}.$$

And it follows from the results of Subsection b that for $\boldsymbol{\lambda}'\boldsymbol{\beta}$ to be estimable, it is necessary that

$$\begin{pmatrix} 0 \\ 1.5 \\ -6 \\ -0.375 \\ -1 \end{pmatrix}' \boldsymbol{\lambda} = 0 \tag{3.22}$$

or, equivalently, that

$$\lambda_5 = 1.5\lambda_2 - 6\lambda_3 - 0.375\lambda_4. \tag{3.23}$$

Moreover, together, the two conditions (3.19) and (3.22) or, equivalently, the two conditions (3.20) and (3.23) are sufficient (as well as necessary) for the estimability of $\boldsymbol{\lambda}'\boldsymbol{\beta}$.

The example provided by the 6 blends of the 4 juices is one in which $\text{rank}(\mathbf{X}) = C - 1\ (= P - 2)$. It is easy to construct examples in which $\text{rank}(\mathbf{X}) = C\ (= P - 1)$ and to do so for any $N\ (\geq C)$—in light of result (3.17) or (3.18), $\text{rank}(\mathbf{X})$ cannot be larger than C. In what is perhaps the simplest way to construct such an example, we can take any C of the blends corresponding to the N data points to be the C pure blends, each of which consists entirely of one ingredient. This approach results in a model matrix $\mathbf{X}$ whose N rows include the vectors $(1, 1, 0, 0, \ldots, 0, 0)$, $(1, 0, 1, 0, \ldots, 0, 0)$, $\ldots$, $(1, 0, 0, 0, \ldots, 0, 1)$. Clearly, these C vectors are linearly independent, implying that $\text{rank}(\mathbf{X}) \geq C$ and hence [since $\text{rank}(\mathbf{X}) \leq C$] that $\text{rank}(\mathbf{X}) = C$.

When $\text{rank}(\mathbf{X}) = C$, the condition $\begin{pmatrix} -1 \\ \mathbf{1}_C \end{pmatrix}' \boldsymbol{\lambda} = 0$, or equivalently the condition $\sum_{j=2}^{C+1} \lambda_j = \lambda_1$, is sufficient (as well as necessary) for the estimability of $\boldsymbol{\lambda}'\boldsymbol{\beta}$. It is worth noting that this condition is not satisfied by any of the individual parameters $\beta_1, \beta_2, \ldots, \beta_{C+1}$ and, consequently, none of these parameters is estimable. Thus, we have established (by means of an example) that not only is it possible for all P of the individual parameters $\beta_1, \beta_2, \ldots, \beta_P$ of a G–M, Aitken, or general linear model to be nonestimable, but it is possible even if $\text{rank}(\mathbf{X}) = P - 1$.

f. Inherent versus noninherent restrictions on estimability

Let us continue to consider the estimability of parametric functions of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ (under the G–M, Aitken, or general linear model). Unless the model is of full rank, estimability is restricted to a proper subset of these parametric functions. The restriction is in the form of restrictions on the coefficient vector $\boldsymbol{\lambda}'$. As discussed in Subsection b, a necessary and sufficient condition for $\boldsymbol{\lambda}'\boldsymbol{\beta}$ to be estimable is that

$$\mathbf{k}'\boldsymbol{\lambda} = 0 \tag{3.24}$$

for every $P \times 1$ vector $\mathbf{k}$ in $\mathfrak{N}(\mathbf{X})$.

It can be helpful to think of each of the restrictions of the form (3.24) as being either an inherent restriction or a noninherent restriction. Let U represent a set consisting of C-dimensional column vectors that are considered to be "feasible" values of the vector $\mathbf{u}$ of explanatory variables—a value might be deemed infeasible either because of the nature of the explanatory variables or because of the "limitations" of the model. The set U is assumed to include the values $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$ of $\mathbf{u}$ that correspond to the N data points.

A restriction of the form (3.24) is an *inherent restriction* if it would be applicable regardless of the number of data points and regardless of the values of $\mathbf{u}$ (in U) corresponding to the data points. Otherwise, the restriction is a *noninherent restriction*. A parametric function of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ that fails to satisfy an inherent restriction tends not to be conceptually meaningful. Accordingly, the nonestimability of such a function tends not to be of concern.

A parametric function of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ that is nonestimable but that would be estimable if it were not for the presence of noninherent restrictions tends to be conceptually meaningful. Its nonestimability may or may not be of concern, depending on whether or not it represents a quantity of interest.

It is informative to consider these concepts in the context of the illustrative example introduced and discussed in Subsection e. Accordingly, suppose that the C explanatory variables $u_1, u_2, \ldots, u_C$ represent the proportionate amounts of the ingredients in a mixture, in which case they are subject to the constraint $\sum_{j=1}^{C} u_j = 1$. And suppose that $\delta(\mathbf{u})$ is of the form (3.16). Then, as discussed in Subsection e, not all parametric functions of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ are estimable; for $\boldsymbol{\lambda}'\boldsymbol{\beta}$ to be estimable, it is necessary that

$$\lambda_1 - \sum_{j=2}^{C+1} \lambda_j = 0 \tag{3.25}$$

or, equivalently, that

$$\sum_{j=2}^{C+1} \lambda_j = \lambda_1. \tag{3.26}$$

In this setting, the set U of feasible values of $\mathbf{u}$ would be the set

$$\{\mathbf{u} \ : \ u_j \geq 0 \text{ (for } j = 1, 2, \ldots, C) \text{ and } \textstyle\sum_{j=1}^{C} u_j = 1\} \tag{3.27}$$

(or, perhaps, a nondegenerate subset of that set). Geometrically, the form of the set (3.27) is that of a $(C-1)$-dimensional simplex. Regardless of the number of data points and regardless of which values of $\mathbf{u}$ correspond to the data points, the model matrix $\mathbf{X}$ would be such that $\mathbf{X}\begin{pmatrix} -1 \\ \mathbf{1}_C \end{pmatrix} = \mathbf{0}$ and, consequently, condition (3.25), or equivalently condition (3.26), would be a necessary condition for the estimability of $\boldsymbol{\lambda}'\boldsymbol{\beta}$. Accordingly, condition (3.26) constitutes an inherent restriction on the estimability of $\boldsymbol{\lambda}'\boldsymbol{\beta}$.

As discussed in Subsection e, the parametric functions for which condition (3.26) is not satisfied include all $C + 1$ of the individual parameters $\beta_1, \beta_2, \ldots, \beta_{C+1}$. If the explanatory variables $u_1, u_2, \ldots, u_C$ were unrestricted, the individual parameters would have meaningful interpretations, emanating from the observation that $\beta_1 = \delta(\mathbf{0})$ and that (for $j = 1, 2, \ldots, C$) β_{j+1} equals the change in $\delta(\mathbf{u})$ effected by a unit change in the jth explanatory variable u_j when the other $C - 1$ explanatory variables are held constant. However, those interpretations are rendered meaningless by the restriction of the vector $\mathbf{u}$ of explanatory variables to the set (3.27). The interpretation of β_1 emanating from the observation that $\beta_1 = \delta(\mathbf{0})$ is meaningless because there is no mixture for which $\mathbf{u} = \mathbf{0}$. The interpretations of $\beta_2, \beta_3, \ldots, \beta_{C+1}$ in terms of the change in $\delta(\mathbf{u})$ effected by changing one of the explanatory variables while holding the others constant are also meaningless. By their very nature, the explanatory variables $u_1, u_2, \ldots, u_C$ are such that $\sum_{j=1}^{C} u_j = 1$, making it impossible to change one of the explanatory variables without changing any of the others.

Subsection e includes an example in which $N = 6$ and $C = 4$ and in which the mixtures consist of blends of juices. In that example, $\mathbf{X}\begin{pmatrix} 0 \\ 1.5 \\ -6 \\ -0.375 \\ -1 \end{pmatrix} = \mathbf{0}$, so that for a parametric function of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ to be estimable, it is necessary that

$$1.5\lambda_2 - 6\lambda_3 - 0.375\lambda_4 - \lambda_5 = 0 \tag{3.28}$$

or, equivalently, that

$$\lambda_5 = 1.5\lambda_2 - 6\lambda_3 - 0.375\lambda_4. \tag{3.29}$$

Condition (3.29) constitutes a noninherent restriction on the estimability of $\boldsymbol{\lambda}'\boldsymbol{\beta}$. It would not be applicable if, for instance, the value of $\mathbf{u}$ corresponding to the fourth of the example's 6 data points were $(0.5, 0.25, 0.25, 0)'$ rather than $(0.5, 0.1, 0.4, 0)'$ [in which case, the first 4 rows of the model matrix $\mathbf{X}$ would be linearly independent and rank$(\mathbf{X})$ would be 4 rather than 3]. Parametric functions of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ that satisfy restriction (3.26), but not restriction (3.29), are conceptually meaningful. Because of restriction (3.29), the only mixtures for which the "response function" $\delta(\mathbf{u})$ would be estimable from the 6 data points in the example are those for which

$$u_4 = 1.5u_1 - 6u_2 - 0.375u_3.$$

If it had been the case that the only restriction on estimability were that determined by the inherent restriction (3.26), then the value of $\delta(\mathbf{u})$ would have been estimable for every mixture [i.e., for every $\mathbf{u}$ in the set (3.27)].

Noninherent restrictions on the estimability of parametric functions of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ may be encountered in cases where the data are "observational" in nature. They may also be encountered in cases where the data come from a designed experiment or a sample survey. The extent to which their presence is of concern would seem to depend on which parametric functions are rendered nonestimable and on the extent to which those functions are of interest.

In the case of data from a designed experiment or a sample survey, the presence of noninherent restrictions may be either inadvertent or by intent. Their presence may be attributable to problems in execution or design. Or when the affected parametric functions are ones that are considered to be of little importance and/or of negligible size, the presence of noninherent restrictions may be viewed as an acceptable consequence of an attempt to make the best possible use of limited resources.

5.4 The Method of Least Squares

Let $\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_N$ represent N data points, and let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$ represent the corresponding values of a C-dimensional column vector $\mathbf{u}$ of explanatory variables. Consider the problem of choosing a function $\underline{\delta}(\mathbf{u})$ of $\mathbf{u}$ on the basis of how well the values $\underline{\delta}(\mathbf{u}_1), \underline{\delta}(\mathbf{u}_2), \ldots, \underline{\delta}(\mathbf{u}_N)$ approximate $\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_N$. Let Δ represent the set of candidates from which the function $\underline{\delta}(\cdot)$ is to be chosen.

Which member of Δ provides the best approximation? One way of addressing this question is through the minimization [for $\underline{\delta}(\cdot) \in \Delta$] of the norm of the N-dimensional column vector with elements $\underline{y}_1 - \underline{\delta}(\mathbf{u}_1), \underline{y}_2 - \underline{\delta}(\mathbf{u}_2), \ldots, \underline{y}_N - \underline{\delta}(\mathbf{u}_N)$. When the norm is taken to be the usual norm, this method is equivalent to minimizing $\sum_{i=1}^{N} [\underline{y}_i - \underline{\delta}(\mathbf{u}_i)]^2$, and is referred to as the *method of least squares* or simply as *least squares*.

The origins of the method of least squares are a matter of some dispute, but date back at least to an 1805 publication by Adrien Marie Legendre. For discussions of the history of the method, refer, for example, to Plackett (1972) and Stigler (1986, chap. 1).

The focus herein is on the method of least squares as applied to settings in which the data points are regarded as the values of observable random variables that follow a G–M, Aitken, or general linear model. In such a setting, the method of least squares can be used to obtain estimates of estimable functions (of the model's parameters) of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$. In the special case of the G–M model, least squares estimators have certain optimal properties.

In what follows, consideration of the method of least squares is restricted to the special case where Δ consists of those functions (of $\mathbf{u}$) that are expressible as linear combinations of P specified functions $\delta_1(\cdot), \delta_2(\cdot), \ldots, \delta_P(\cdot)$. Thus, $\underline{\delta}(\cdot)$ is a member of Δ if $\underline{\delta}(\mathbf{u})$ is expressible in the form

$$\underline{\delta}(\mathbf{u}) = b_1\delta_1(\mathbf{u}) + b_2\delta_2(\mathbf{u}) + \cdots + b_P\delta_P(\mathbf{u}), \tag{4.1}$$

where $b_1, b_2, \ldots, b_P$ are arbitrary scalars. When $\underline{\delta}(\cdot)$ is such that $\underline{\delta}(\mathbf{u})$ is expressible in the form (4.1),

$$\sum_{i=1}^{N} [\underline{y}_i - \underline{\delta}(\mathbf{u}_i)]^2 = \sum_{i=1}^{N} \Big(\underline{y}_i - \sum_{j=1}^{P} x_{ij}b_j\Big)^2,$$

where (for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, P$) $x_{ij} = \delta_j(\mathbf{u}_i)$. Accordingly, in the special case under consideration, the minimization of $\sum_{i=1}^{N} [\underline{y}_i - \underline{\delta}(\mathbf{u}_i)]^2$ with respect to $\underline{\delta}(\cdot)$ [for $\underline{\delta}(\cdot) \in \Delta$] is equivalent to the minimization of $\sum_{i=1}^{N} \big(\underline{y}_i - \sum_{j=1}^{P} x_{ij}b_j\big)^2$ with respect to $b_1, b_2, \ldots, b_P$. Moreover, upon letting $\underline{\mathbf{y}} = (\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_N)'$ and $\mathbf{b} = (b_1, b_2, \ldots, b_P)'$ and taking $\mathbf{X}$ to be the $N \times P$ matrix with ijth element x_{ij},

$$\sum_{i=1}^{N} \Big(\underline{y}_i - \sum_{j=1}^{P} x_{ij}b_j\Big)^2 = (\underline{\mathbf{y}} - \mathbf{X}\mathbf{b})'(\underline{\mathbf{y}} - \mathbf{X}\mathbf{b}), \tag{4.2}$$

so that the minimization of $\sum_{i=1}^{N}\left(\underset{\sim}{y}_i - \sum_{j=1}^{P} x_{ij}b_j\right)^2$ with respect to $b_1, b_2, \ldots, b_P$ is equivalent to the minimization of $(\underset{\sim}{\mathbf{y}} - \mathbf{X}\mathbf{b})'(\underset{\sim}{\mathbf{y}} - \mathbf{X}\mathbf{b})$ with respect to the P-dimensional vector $\mathbf{b}$ (where $\mathbf{b}$ is an arbitrary member of $\mathcal{R}^P$).

We wish to obtain a solution to the problem of minimizing the quantity $\sum_{i=1}^{N}\left(\underset{\sim}{y}_i - \sum_{j=1}^{P} x_{ij}b_j\right)^2$ with respect to $b_1, b_2, \ldots, b_P$. Conditions that are necessary for this quantity to attain a minimum value can be obtained by differentiating with respect to $b_1, b_2, \ldots, b_P$ and by equating the resultant partial derivatives to 0. Or in what can be regarded as an appealing variation on this approach, we can reformulate the minimization problem in matrix notation [as the problem of minimizing $(\underset{\sim}{\mathbf{y}}-\mathbf{X}\mathbf{b})'(\underset{\sim}{\mathbf{y}}-\mathbf{X}\mathbf{b})$ with respect to $\mathbf{b}$] and take advantage of some basic results on vector differentiation.

a. Some basic results on vector differentiation

Let $\mathbf{x} = (x_1, x_2, \ldots, x_M)'$ represent an M-dimensional column vector of variables. And let $f(\mathbf{x})$ represent a function of $\mathbf{x}$. Write $\dfrac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$, or simply $\dfrac{\partial f}{\partial \mathbf{x}}$, for the M-dimensional column vector whose jth element is the (first-order) partial derivative $\dfrac{\partial f(\mathbf{x})}{\partial x_j}$ of f with respect to x_j; this vector may be referred to as the *derivative of* $f(\mathbf{x})$ *with respect to* $\mathbf{x}$ and is sometimes called the *gradient vector* of f. And write $\dfrac{\partial f(\mathbf{x})}{\partial \mathbf{x}'}$, or $\dfrac{\partial f}{\partial \mathbf{x}'}$, for $\left[\dfrac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right]'$; this vector may be referred to as the *derivative of* $f(\mathbf{x})$ *with respect to* $\mathbf{x}'$. Further, write $\dfrac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}'}$, or $\dfrac{\partial^2 f}{\partial \mathbf{x}\partial \mathbf{x}'}$, for the $M \times M$ matrix whose ijth element is the second-order partial derivative $\dfrac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}$—or, when $j = i$, $\dfrac{\partial^2 f(\mathbf{x})}{\partial x_i^2}$—of f with respect to x_i and x_j, and refer to this matrix as the *Hessian matrix* of f.

The formulas for obtaining the partial derivatives of a linear combination of functions, a product of functions, and a ratio of two functions with respect to a single variable extend in an altogether straightforward way to vector differentiation. In particular, in the case of the product of two functions $f(\mathbf{x})$ and $g(\mathbf{x})$ of $\mathbf{x}$,

$$\frac{\partial f(\mathbf{x})g(\mathbf{x})}{\partial \mathbf{x}} = f(\mathbf{x})\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} + g(\mathbf{x})\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}. \tag{4.3}$$

Refer, for example, to Harville (1997, sec. 15.2) for further particulars.

Denote by $\mathbf{a} = (a_1, a_2, \ldots, a_M)'$ an M-dimensional column vector of constants and by $\mathbf{A} = \{a_{ik}\}$ an $M \times M$ matrix of constants. And consider the differentiation of the linear form $\mathbf{a}'\mathbf{x} = \sum_i a_i x_i$ and of the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i,k} a_{ik}x_i x_k$. The first-order partial derivatives of $\mathbf{a}'\mathbf{x}$ and the first- and second-order partial derivatives of $\mathbf{x}'\mathbf{A}\mathbf{x}$ are

$$\frac{\partial\, \mathbf{a}'\mathbf{x}}{\partial x_j} = a_j \qquad (j = 1, 2, \ldots, M), \tag{4.4}$$

$$\frac{\partial\, \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial x_j} = \sum_i a_{ij}x_i + \sum_k a_{jk}x_k \qquad (j = 1, 2, \ldots, M), \tag{4.5}$$

and

$$\frac{\partial^2\, \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial x_s \partial x_j} = a_{sj} + a_{js} \qquad (s, j = 1, 2, \ldots, M), \tag{4.6}$$

as can be verified via a relatively simple exercise—refer, e.g., to Harville (1997, sec. 15.3) for the details.

In light of result (4.4), the gradient vector of $\mathbf{a}'\mathbf{x}$ is

$$\frac{\partial\, \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}. \tag{4.7}$$

And in light of result (4.5), the gradient vector of $\mathbf{x}'\mathbf{A}\mathbf{x}$ is

$$\frac{\partial\, \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}')\mathbf{x} \tag{4.8}$$

(as is evident upon observing that $\sum_k a_{jk}x_k$ is the jth element of the column vector $\mathbf{A}\mathbf{x}$ and $\sum_i a_{ij}x_i$ is the jth element of $\mathbf{A}'\mathbf{x}$). Further, in light of result (4.6), the Hessian matrix of $\mathbf{x}'\mathbf{A}\mathbf{x}$ is

$$\frac{\partial^2\, \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}\partial \mathbf{x}'} = \mathbf{A} + \mathbf{A}'. \tag{4.9}$$

In the special case where $\mathbf{A}$ is symmetric, results (4.8) and (4.9) simplify to

$$\frac{\partial\, \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x} \tag{4.10}$$

and

$$\frac{\partial^2\, \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}\partial \mathbf{x}'} = 2\mathbf{A}. \tag{4.11}$$

Let $\mathbf{f}(\mathbf{x})$ represent a $(P \times 1)$-dimensional vector-valued function of the M-dimensional (column) vector $\mathbf{x} = (x_1, x_2, \ldots, x_M)'$, and denote by $f_1(\mathbf{x})$, $f_2(\mathbf{x})$, $\ldots$, $f_P(\mathbf{x})$ the P functions of $\mathbf{x}$ that constitute the first, second, ..., Pth elements of $\mathbf{f}(\mathbf{x})$. Let us write $\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}'}$, or simply $\frac{\partial \mathbf{f}}{\partial \mathbf{x}'}$, for the $P \times M$ matrix whose sjth element is $\frac{\partial f_s}{\partial x_j}$; this matrix may be referred to as the *derivative of* $\mathbf{f}(\mathbf{x})$ *with respect to* $\mathbf{x}'$ and is sometimes called the *Jacobian matrix* of $\mathbf{f}$. And let us write $\frac{\partial \mathbf{f}'(\mathbf{x})}{\partial \mathbf{x}}$, or simply $\frac{\partial \mathbf{f}'}{\partial \mathbf{x}}$, for the $M \times P$ matrix whose jsth element is $\frac{\partial f_s}{\partial x_j}$ or, equivalently, whose sth column is $\frac{\partial f_s}{\partial \mathbf{x}}$; this matrix may be referred to as the *derivative of* $\mathbf{f}'(\mathbf{x})$ *with respect to* $\mathbf{x}$ and is sometimes called the *gradient matrix* of $\mathbf{f}$. Note that for any $P \times M$ matrix of constants $\mathbf{B}$,

$$\frac{\partial (\mathbf{B}\mathbf{x})}{\partial \mathbf{x}'} = \mathbf{B} \quad \text{and} \quad \frac{\partial (\mathbf{B}\mathbf{x})'}{\partial \mathbf{x}} = \mathbf{B}'. \tag{4.12}$$

b. Solution to the least squares minimization problem

Let us now consider the implications of the results of Subsection a as applied to the minimization (with respect to the $P \times 1$ vector $\mathbf{b} = \{b_j\}$) of $(\underline{\mathbf{y}} - \mathbf{X}\mathbf{b})'(\underline{\mathbf{y}} - \mathbf{X}\mathbf{b})$, where $\underline{\mathbf{y}} = \{\underline{y}_i\}$ is an $N \times 1$ data vector and $\mathbf{X} = \{x_{ij}\}$ an $N \times P$ matrix. Take $q(\cdot)$ to be the function (of $\mathbf{b}$) defined by $q(\mathbf{b}) = (\underline{\mathbf{y}} - \mathbf{X}\mathbf{b})'(\underline{\mathbf{y}} - \mathbf{X}\mathbf{b})$. And observe that

$$q(\mathbf{b}) = \underline{\mathbf{y}}'\underline{\mathbf{y}} - 2(\mathbf{X}'\underline{\mathbf{y}})'\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}.$$

Then, as an immediate application of formulas (4.7), (4.10), and (4.11), we have that

$$\frac{\partial\, q(\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{X}'\underline{\mathbf{y}} + 2\mathbf{X}'\mathbf{X}\mathbf{b} \tag{4.13}$$

and

$$\frac{\partial^2 q(\mathbf{b})}{\partial \mathbf{b}\partial \mathbf{b}'} = 2\mathbf{X}'\mathbf{X}. \tag{4.14}$$

It follows from basic results on unconstrained minimization (e.g., Luenberger and Ye 2016, sec. 7.1) that a necessary condition for $q(\mathbf{b})$ to attain a minimum value at a point $\tilde{\mathbf{b}}$ is that $\frac{\partial\, q(\mathbf{b})}{\partial \mathbf{b}} = \mathbf{0}$ at $\mathbf{b} = \tilde{\mathbf{b}}$. Clearly,

$$\frac{\partial q(\mathbf{b})}{\partial \mathbf{b}} = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\underset{\sim}{\mathbf{y}}.$$

Thus, a necessary condition for $q(\mathbf{b})$ to attain a minimum value at a point $\tilde{\mathbf{b}}$ is that $\tilde{\mathbf{b}}$ constitute a solution to the linear system

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\underset{\sim}{\mathbf{y}} \tag{4.15}$$

(in the $P \times 1$ vector $\mathbf{b}$). Linear system (4.15) consists of P equations; collectively, these equations are known as the *normal equations*.

The normal equations are consistent, as can be verified by, for example, observing [in light of equality (2.12.2)] that $\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\underset{\sim}{\mathbf{y}} = \mathbf{X}'\underset{\sim}{\mathbf{y}}$ [which implies that $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\underset{\sim}{\mathbf{y}}$ is a solution to the normal equations]. Moreover, for $q(\mathbf{b})$ to attain a minimum value at a point $\tilde{\mathbf{b}}$, it is sufficient, as well as necessary, that $\tilde{\mathbf{b}}$ constitute a solution to the normal equations. That is, the set of points at which $q(\mathbf{b})$ attains a minimum value equals the solution set of linear system (4.15).

Let us verify that every solution to the normal equations is a point at which $q(\mathbf{b})$ attains a minimum value. There are at least two different ways of accomplishing the verification. We could start with the observation that $\dfrac{\partial^2 q(\mathbf{b})}{\partial \mathbf{b}\partial \mathbf{b}'}$ is a nonnegative definite matrix and that, as a consequence, $q(\mathbf{b})$ is a convex function (e.g., Luenberger and Ye 2016, sec. 7.4). We could then take advantage of a general result on the minimization of convex functions (e.g., Luenberger and Ye 2016, sec. 7.5) to conclude that for any $P \times 1$ vector $\tilde{\mathbf{b}}$ such that $\dfrac{\partial q(\mathbf{b})}{\partial \mathbf{b}} = \mathbf{0}$ at $\mathbf{b} = \tilde{\mathbf{b}}$ or, equivalently, for any solution $\tilde{\mathbf{b}}$ to the normal equations, $q(\mathbf{b})$ attains a minimum value at $\mathbf{b} = \tilde{\mathbf{b}}$.

An alternative way of accomplishing the verification is to do so directly (without making any demands on the reader's knowledge of the literature on optimization). For any solution $\tilde{\mathbf{b}}$ to the normal equations,

$$q(\mathbf{b}) = q(\tilde{\mathbf{b}}) + [\mathbf{X}(\mathbf{b} - \tilde{\mathbf{b}})]'\mathbf{X}(\mathbf{b} - \tilde{\mathbf{b}}) \geq q(\tilde{\mathbf{b}}), \tag{4.16}$$

as is evident upon observing that

$$q(\mathbf{b}) = [\underset{\sim}{\mathbf{y}} - \mathbf{X}\tilde{\mathbf{b}} - \mathbf{X}(\mathbf{b} - \tilde{\mathbf{b}})]'[\underset{\sim}{\mathbf{y}} - \mathbf{X}\tilde{\mathbf{b}} - \mathbf{X}(\mathbf{b} - \tilde{\mathbf{b}})],$$

that

$$[\mathbf{X}(\mathbf{b} - \tilde{\mathbf{b}})]'(\underset{\sim}{\mathbf{y}} - \mathbf{X}\tilde{\mathbf{b}}) = (\mathbf{b} - \tilde{\mathbf{b}})'(\mathbf{X}'\underset{\sim}{\mathbf{y}} - \mathbf{X}'\mathbf{X}\tilde{\mathbf{b}}) = 0,$$

that

$$(\underset{\sim}{\mathbf{y}} - \mathbf{X}\tilde{\mathbf{b}})'\mathbf{X}(\mathbf{b} - \tilde{\mathbf{b}}) = [(\underset{\sim}{\mathbf{y}} - \mathbf{X}\tilde{\mathbf{b}})'\mathbf{X}(\mathbf{b} - \tilde{\mathbf{b}})]' = [\mathbf{X}(\mathbf{b} - \tilde{\mathbf{b}})]'(\underset{\sim}{\mathbf{y}} - \mathbf{X}\tilde{\mathbf{b}}),$$

and that $[\mathbf{X}(\mathbf{b} - \tilde{\mathbf{b}})]'\mathbf{X}(\mathbf{b} - \tilde{\mathbf{b}})$ is a sum of squares [of the elements of $\mathbf{X}(\mathbf{b} - \tilde{\mathbf{b}})$]. And it follows immediately from result (4.16) that $q(\mathbf{b})$ attains a minimum value at $\mathbf{b} = \tilde{\mathbf{b}}$.

Result (4.16) also serves to confirm that any point at which $q(\mathbf{b})$ attains a minimum value is a solution to the normal equations (or, equivalently, a point at which $\dfrac{\partial q(\mathbf{b})}{\partial \mathbf{b}} = \mathbf{0}$). To see this, observe that

$$\begin{aligned} q(\mathbf{b}) = q(\tilde{\mathbf{b}}) \quad &\Rightarrow \quad [\mathbf{X}(\mathbf{b} - \tilde{\mathbf{b}})]'\mathbf{X}(\mathbf{b} - \tilde{\mathbf{b}}) = 0 \\ &\Rightarrow \quad \mathbf{X}(\mathbf{b} - \tilde{\mathbf{b}}) = \mathbf{0} \\ &\Rightarrow \quad \mathbf{X}\mathbf{b} = \mathbf{X}\tilde{\mathbf{b}} \\ &\Rightarrow \quad \mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{X}\tilde{\mathbf{b}} = \mathbf{X}'\underset{\sim}{\mathbf{y}}. \end{aligned} \tag{4.17}$$

The value of the vector $\mathbf{X}\tilde{\mathbf{b}}$ is the same for any solution $\tilde{\mathbf{b}}$ to the normal equations or, equivalently, for any vector $\tilde{\mathbf{b}}$ at which $q(\mathbf{b})$ attains its minimum value. That is, for any two solutions $\tilde{\mathbf{b}}_1$ and $\tilde{\mathbf{b}}_2$ to the normal equations [or any two vectors $\tilde{\mathbf{b}}_1$ and $\tilde{\mathbf{b}}_2$ at which $q(\mathbf{b})$ attains its minimum value],

$$\mathbf{X}\tilde{\mathbf{b}}_1 = \mathbf{X}\tilde{\mathbf{b}}_2. \tag{4.18}$$

Equality (4.18) can be established by applying result (4.17) or, alternatively, by observing that $\mathbf{X}'\mathbf{X}\tilde{\mathbf{b}}_1 = \mathbf{X}'\underset{\sim}{\mathbf{y}} = \mathbf{X}'\mathbf{X}\tilde{\mathbf{b}}_2$ and then observing (in light of Corollary 2.3.4) that $\mathbf{X}'\mathbf{X}\tilde{\mathbf{b}}_1 = \mathbf{X}'\mathbf{X}\tilde{\mathbf{b}}_2 \Rightarrow$

$\mathbf{X}\tilde{\mathbf{b}}_1 = \mathbf{X}\tilde{\mathbf{b}}_2$. The vector $(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\underline{\mathbf{y}}$ is a solution to the normal equations. Thus, as a variation on result (4.18), we have that for any solution $\tilde{\mathbf{b}}$ to the normal equations or, equivalently, for any vector $\tilde{\mathbf{b}}$ at which $q(\mathbf{b})$ attains its minimum value,

$$\mathbf{X}\tilde{\mathbf{b}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\underline{\mathbf{y}} = \mathbf{P}_{\mathbf{X}}\underline{\mathbf{y}} \tag{4.19}$$

[where $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$].

Let $\tilde{\mathbf{b}}$ represent any solution to the normal equations. Then, the minimum value of $q(\mathbf{b})$ is

$$q(\tilde{\mathbf{b}}) = (\underline{\mathbf{y}} - \mathbf{X}\tilde{\mathbf{b}})'(\underline{\mathbf{y}} - \mathbf{X}\tilde{\mathbf{b}}).$$

This value is reexpressible in the form

$$q(\tilde{\mathbf{b}}) = \underline{\mathbf{y}}'(\underline{\mathbf{y}} - \mathbf{X}\tilde{\mathbf{b}}) \tag{4.20}$$

and in the form

$$q(\tilde{\mathbf{b}}) = \underline{\mathbf{y}}'\underline{\mathbf{y}} - \tilde{\mathbf{b}}'\mathbf{X}'\underline{\mathbf{y}}, \tag{4.21}$$

as is evident upon observing that

$$(\mathbf{X}\tilde{\mathbf{b}})'(\underline{\mathbf{y}} - \mathbf{X}\tilde{\mathbf{b}}) = \tilde{\mathbf{b}}'(\mathbf{X}'\underline{\mathbf{y}} - \mathbf{X}'\mathbf{X}\tilde{\mathbf{b}}) = \tilde{\mathbf{b}}'\mathbf{0} = 0$$

[and that $\underline{\mathbf{y}}'\mathbf{X}\tilde{\mathbf{b}} = (\underline{\mathbf{y}}'\mathbf{X}\tilde{\mathbf{b}})' = \tilde{\mathbf{b}}'\mathbf{X}'\underline{\mathbf{y}}$]. Moreover, in light of results (4.19) and (4.20), the minimum value of $q(\mathbf{b})$ is also expressible as

$$q(\tilde{\mathbf{b}}) = \underline{\mathbf{y}}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\underline{\mathbf{y}}. \tag{4.22}$$

Note that expression (4.22) is a quadratic form (in $\underline{\mathbf{y}}$), the matrix of which is $\mathbf{I} - \mathbf{P}_{\mathbf{X}}$.

In summary, we have established the following:

(1) The function $q(\mathbf{b}) = (\underline{\mathbf{y}} - \mathbf{X}\mathbf{b})'(\underline{\mathbf{y}} - \mathbf{X}\mathbf{b})$ attains a minimum value at a point $\tilde{\mathbf{b}}$ if and only if $\tilde{\mathbf{b}}$ is a solution to the linear system $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\underline{\mathbf{y}}$ (comprising the normal equations).

(2) The linear system $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\underline{\mathbf{y}}$ is consistent.

(3) $\mathbf{X}\tilde{\mathbf{b}}_1 = \mathbf{X}\tilde{\mathbf{b}}_2$ for any two solutions $\tilde{\mathbf{b}}_1$ and $\tilde{\mathbf{b}}_2$ to the linear system $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\underline{\mathbf{y}}$ or, equivalently, for any two vectors $\tilde{\mathbf{b}}_1$ amd $\tilde{\mathbf{b}}_2$ at which $q(\mathbf{b})$ attains its minimum value.

(3′) $\mathbf{X}\tilde{\mathbf{b}} = \mathbf{P}_{\mathbf{X}}\mathbf{y}$ for any solution $\tilde{\mathbf{b}}$ to the linear system $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\underline{\mathbf{y}}$ or, equivalently, for any vector $\tilde{\mathbf{b}}$ at which $q(\mathbf{b})$ attains its minimum value.

(4) For any solution $\tilde{\mathbf{b}}$ to the linear system $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\underline{\mathbf{y}}$,

$$\min_{\mathbf{b}} q(\mathbf{b}) = q(\tilde{\mathbf{b}}) = \underline{\mathbf{y}}'(\underline{\mathbf{y}} - \mathbf{X}\tilde{\mathbf{b}}) = \underline{\mathbf{y}}'\underline{\mathbf{y}} - \tilde{\mathbf{b}}'\mathbf{X}'\underline{\mathbf{y}} = \underline{\mathbf{y}}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\underline{\mathbf{y}}.$$

c. Least squares estimators of estimable functions

Suppose the setting is one in which N data points are regarded as the respective values of observable random variables $y_1, y_2, \ldots, y_N$ that follow a G–M, Aitken, or general linear model. And consider the method of least squares as applied to the estimation of an estimable parametric function of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$. Take the functions $\delta_1(\cdot), \delta_2(\cdot), \ldots, \delta_P(\cdot)$ (and the number of such functions P) to be those for which (under the G–M, Aitken, or general linear model)

$$\mathrm{E}(y_i) = \beta_1\delta_1(\mathbf{u}_i) + \beta_2\delta_2(\mathbf{u}_i) + \cdots + \beta_P\delta_P(\mathbf{u}_i) \qquad (i = 1, 2, \ldots, N).$$

Further, let $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$, and continue to take $\mathbf{X}$ to be the $N \times P$ matrix with ijth element $x_{ij} = \delta_j(\mathbf{u}_i)$.

By definition, the *least squares estimator* of an estimable function $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is the function, say $\ell(\mathbf{y})$, of $\mathbf{y}$ whose value at $\mathbf{y} = \underline{\mathbf{y}}$ (an arbitrary $N \times 1$ vector) is taken to be $\boldsymbol{\lambda}'\tilde{\mathbf{b}}$, where $\tilde{\mathbf{b}}$ is any solution to the linear system

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\underline{\mathbf{y}} \tag{4.23}$$

(in the $P \times 1$ vector $\mathbf{b}$), comprising the normal equations. Unless $\text{rank}(\mathbf{X}) = P$, there are an infinite number of solutions to the normal equations and hence an infinite number of choices for $\tilde{\mathbf{b}}$. Nevertheless, $\boldsymbol{\lambda}'\tilde{\mathbf{b}}$ is uniquely defined; that is, $\boldsymbol{\lambda}'\tilde{\mathbf{b}}$ is invariant to the choice of $\tilde{\mathbf{b}}$. To see this, let $\tilde{\mathbf{b}}_1$ and $\tilde{\mathbf{b}}_2$ represent any two solutions to linear system (4.23), and observe (in light of the results of Subsection b and Section 5.3) that $\mathbf{X}\tilde{\mathbf{b}}_1 = \mathbf{X}\tilde{\mathbf{b}}_2$ and that (because of the estimability of $\boldsymbol{\lambda}'\boldsymbol{\beta}$) $\boldsymbol{\lambda}' = \mathbf{a}'\mathbf{X}$ for some $N \times 1$ vector $\mathbf{a}$, so that

$$\boldsymbol{\lambda}'\tilde{\mathbf{b}}_1 = \mathbf{a}'\mathbf{X}\tilde{\mathbf{b}}_1 = \mathbf{a}'\mathbf{X}\tilde{\mathbf{b}}_2 = \boldsymbol{\lambda}'\tilde{\mathbf{b}}_2.$$

The solutions to linear system (4.23) include the vector $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\underline{\mathbf{y}}$. Thus, among the representations for the least squares estimator $\ell(\mathbf{y})$ of an estimable function $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is the representation

$$\ell(\mathbf{y}) = \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}, \tag{4.24}$$

so the least squares estimator is a linear estimator. In the special case where $\mathbf{X}$ is of full column rank P, linear system (4.23) has the unique solution $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. And in that special case, expression (4.24) becomes

$$\ell(\mathbf{y}) = \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Some further results on estimability. Let us continue to suppose that $\mathbf{y}$ is a column vector of observable random variables $y_1, y_2, \ldots, y_N$ that follow a G–M, Aitken, or general linear model. And let us consider further the subject of Section 5.3, namely, the estimability of a parametric function of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ $(= \sum_{j=1}^{P} \lambda_j \beta_j)$.

In Section 5.3, the estimability of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ was related to various characteristics of the model matrix $\mathbf{X}$. A number of conditions were set forth, each of which is necessary and sufficient for estimability. Those conditions can be restated in terms of various characteristics of the $P \times P$ matrix $\mathbf{X}'\mathbf{X}$, which is the coefficient matrix of the normal equations. Their restatement is based on the following results:

$$\mathcal{R}(\mathbf{X}'\mathbf{X}) = \mathcal{R}(\mathbf{X}); \tag{4.25}$$

$$\text{rank}(\mathbf{X}'\mathbf{X}) = \text{rank}(\mathbf{X}); \tag{4.26}$$

$$\mathbf{X}[(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}']\mathbf{X} = \mathbf{X} \tag{4.27}$$

[i.e., $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ is a generalized inverse of $\mathbf{X}$]; and

$$\mathbf{k}'\mathbf{X}'\mathbf{X} = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{k}'\mathbf{X}' = \mathbf{0} \tag{4.28}$$

(where $\mathbf{k}$ is an arbitrary $P \times 1$ vector) or, equivalently,

$$\mathbf{X}'\mathbf{X}\mathbf{k} = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{X}\mathbf{k} = \mathbf{0}$$

or, also equivalently,

$$\mathcal{N}(\mathbf{X}'\mathbf{X}) = \mathcal{N}(\mathbf{X})$$

—results (4.25), (4.26), and (4.27) were established in Section 2.12, and result (4.28) is a consequence of Corollary 2.3.4.

In light of results (4.25), (4.26), (4.27), and (4.28), it follows from the results of Section 5.3 that each of the following conditions is necessary and sufficient for the estimability of $\boldsymbol{\lambda}'\boldsymbol{\beta}$:

(1) $\boldsymbol{\lambda}' \in \mathcal{R}(\mathbf{X}'\mathbf{X})$;

(2) $\boldsymbol{\lambda}' = \mathbf{r}'\mathbf{X}'\mathbf{X}$ for some $P \times 1$ vector $\mathbf{r}$;

(3) $\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \boldsymbol{\lambda}'$ or, equivalently,

(3′) $\boldsymbol{\lambda}'[\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}] = \mathbf{0}$;

(4) $\mathbf{k}'\boldsymbol{\lambda} = 0$ for every $P \times 1$ vector $\mathbf{k}$ such that $\mathbf{k}'\mathbf{X}'\mathbf{X} = \mathbf{0}$ or, equivalently,

(4′) $\mathbf{k}'\boldsymbol{\lambda} = 0$ for every $P \times 1$ vector $\mathbf{k}$ in $\mathfrak{N}(\mathbf{X}'\mathbf{X})$;

(5) $\mathbf{k}_1'\boldsymbol{\lambda} = \mathbf{k}_2'\boldsymbol{\lambda} = \cdots = \mathbf{k}_S'\boldsymbol{\lambda} = 0$,

where $S = N - \text{rank}(\mathbf{X}'\mathbf{X})$ and where $\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_S$ are any S linearly independent vectors in $\mathfrak{N}(\mathbf{X}'\mathbf{X})$ [i.e., any S linearly independent (P-dimensional) column vectors such that $\mathbf{X}'\mathbf{X}\mathbf{k}_1 = \mathbf{X}'\mathbf{X}\mathbf{k}_2 = \cdots = \mathbf{X}'\mathbf{X}\mathbf{k}_S = \mathbf{0}$].

Conditions (1), (2), (3), and (3′) are stated in terms of the row vector $\boldsymbol{\lambda}'$. Alternative versions of what are essentially the same conditions can be obtained by restating the conditions in terms of the column vector $\boldsymbol{\lambda}$ (and by observing that $[(\mathbf{X}'\mathbf{X})^-]'$, like $(\mathbf{X}'\mathbf{X})^-$ itself, is a generalized inverse of $\mathbf{X}'\mathbf{X}$). The alternative versions are as follows:

(1) $\boldsymbol{\lambda} \in \mathcal{C}(\mathbf{X}'\mathbf{X})$;

(2) $\boldsymbol{\lambda} = \mathbf{X}'\mathbf{X}\mathbf{r}$ for some $P \times 1$ vector $\mathbf{r}$;

(3) $\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\boldsymbol{\lambda} = \boldsymbol{\lambda}$;

(3′) $[\mathbf{I} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-]\boldsymbol{\lambda} = \mathbf{0}$.

As in the case of the original versions, each of these conditions is necessary and sufficient for $\boldsymbol{\lambda}'\boldsymbol{\beta}$ to be estimable.

Note that the P-dimensional random column vector $\mathbf{X}'\mathbf{y}$, defined by the right side of the normal equations, is such that

$$\text{E}(\mathbf{X}'\mathbf{y}) = \mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \tag{4.29}$$

And observe that if $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, then (in light of our results on estimability) there exists a $P \times 1$ vector $\mathbf{r}$ such that

$$\boldsymbol{\lambda}'\boldsymbol{\beta} = \mathbf{r}'\text{E}(\mathbf{X}'\mathbf{y}). \tag{4.30}$$

Thus, an estimable function is interpretable in terms of the expected values of the P elements of $\mathbf{X}'\mathbf{y}$. In fact, unless $\text{rank}(\mathbf{X}) = P$, an estimable function $\boldsymbol{\lambda}'\boldsymbol{\beta}$ will have multiple representations of the form (4.30) and hence multiple interpretations in terms of the expected values of the elements of $\mathbf{X}'\mathbf{y}$.

As a further implication of result (4.29), we have that, for any $P \times 1$ vector $\mathbf{r}$ of constants,

$$\text{E}(\mathbf{r}'\mathbf{X}'\mathbf{y}) = \mathbf{r}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.$$

And upon observing that the least squares estimator of $\mathbf{r}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$ is $\mathbf{r}'\mathbf{X}'\mathbf{y}$, it follows that any linear combination of the elements of the vector $\mathbf{X}'\mathbf{y}$ (defined by the right side of the normal equations) is the least squares estimator of its expected value.

Conjugate normal equations. Let us resume our discussion of least squares estimation, taking the setting to that in which N data points are regarded as the respective values of the elements of an $N \times 1$ observable random vector $\mathbf{y}$ that follows a G–M, Aitken, or general linear model. Corresponding to any linear combination $\boldsymbol{\lambda}'\boldsymbol{\beta}$ (of the elements of the vector $\boldsymbol{\beta}$) is the linear system

$$\mathbf{X}'\mathbf{X}\mathbf{r} = \boldsymbol{\lambda}, \tag{4.31}$$

comprising P equations in a $P \times 1$ vector $\mathbf{r}$ of unknowns. The coefficient matrix $\mathbf{X}'\mathbf{X}$ of this linear system is the same as that of the linear system

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\underline{\mathbf{y}}, \tag{4.32}$$

comprising the normal equations; however, the right side of linear system (4.31) is the coefficient vector $\boldsymbol{\lambda}$, while that of linear system (4.32) is $\mathbf{X}'\underline{\mathbf{y}}$. The P equations that form linear system (4.31) are sometimes referred to (collectively) as the *conjugate normal equations*.

It follows from the results of Part 1 of the present subsection that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable if and only if the conjugate normal equations are consistent. Now, suppose that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, and consider the least squares estimator $\ell(\mathbf{y})$ of $\boldsymbol{\lambda}'\boldsymbol{\beta}$. The value $\ell(\underline{\mathbf{y}})$ of $\ell(\mathbf{y})$ at $\mathbf{y} = \underline{\mathbf{y}}$ is expressible in terms of any

solution to the normal equations; for any solution $\tilde{\mathbf{b}}$ to linear system (4.32),

$$\ell(\underline{\mathbf{y}}) = \boldsymbol{\lambda}'\tilde{\mathbf{b}}. \tag{4.33}$$

The value of $\ell(\mathbf{y})$ is also expressible in terms of any solution to the conjugate normal equations. For any solution $\tilde{\mathbf{r}}$ to linear system (4.31) [and any solution $\tilde{\mathbf{b}}$ to linear system (4.32)], we find that

$$\ell(\underline{\mathbf{y}}) = \boldsymbol{\lambda}'\tilde{\mathbf{b}} = (\mathbf{X}'\mathbf{X}\tilde{\mathbf{r}})'\tilde{\mathbf{b}} = \tilde{\mathbf{r}}'\mathbf{X}'\mathbf{X}\tilde{\mathbf{b}} = \tilde{\mathbf{r}}'\mathbf{X}'\underline{\mathbf{y}}. \tag{4.34}$$

The upshot of result (4.34) is that the roles of the normal equations and the conjugate normal equations are (in a certain sense) interchangeable. The least squares estimate $\ell(\underline{\mathbf{y}})$ can be obtained by forming the (usual) inner product $\boldsymbol{\lambda}'\tilde{\mathbf{b}}$ of a solution $\tilde{\mathbf{b}}$ to the normal equations and of the right side $\boldsymbol{\lambda}$ of the conjugate normal equations. Or, alternatively, it can be obtained by forming the inner product $\tilde{\mathbf{r}}'\mathbf{X}'\underline{\mathbf{y}}$ of a solution $\tilde{\mathbf{r}}$ to the conjugate normal equations and of the right side $\mathbf{X}'\underline{\mathbf{y}}$ of the normal equations.

The general form and expected values, variances, and covariances of least squares estimators. Let us continue to take $\mathbf{y}$ to be an N-dimensional observable random (column) vector that follows a G–M, Aitken, or general linear model. And let us consider the general form and expected values, variances, and covariances of least squares estimators of estimable linear combinations of the elements of $\boldsymbol{\beta}$.

Suppose that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is an estimable linear combination. Then, in light of the results of Part 2 of the present subsection, the least squares estimator $\ell(\mathbf{y})$ of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is expressible in the form

$$\ell(\mathbf{y}) = \tilde{\mathbf{r}}'\mathbf{X}'\mathbf{y}, \tag{4.35}$$

where $\tilde{\mathbf{r}}$ is any solution to the conjugate normal equations $\mathbf{X}'\mathbf{X}\mathbf{r} = \boldsymbol{\lambda}$. It follows immediately that the least squares estimator is a linear estimator, in confirmation of what was established earlier (in the introductory part of the present subsection) via a different approach. Moreover,

$$\mathrm{E}(\tilde{\mathbf{r}}'\mathbf{X}'\mathbf{y}) = \tilde{\mathbf{r}}'\mathbf{X}'\mathrm{E}(\mathbf{y}) = \tilde{\mathbf{r}}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X}\tilde{\mathbf{r}})'\boldsymbol{\beta} = \boldsymbol{\lambda}'\boldsymbol{\beta}. \tag{4.36}$$

Thus, the least squares estimator is a linear unbiased estimator.

The vector $\mathbf{X}\tilde{\mathbf{r}}$ is invariant to the choice of the solution $\tilde{\mathbf{r}}$ to the conjugate normal equations (as is evident, e.g., from Corollary 2.3.4). The solutions to the conjugate normal equations include the vector $(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda}$, and {because $[(\mathbf{X}'\mathbf{X})^{-}]'$, like $(\mathbf{X}'\mathbf{X})^{-}$ itself, is a generalized inverse of $\mathbf{X}'\mathbf{X}$}they also include the vector $[(\mathbf{X}'\mathbf{X})^{-}]'\boldsymbol{\lambda}$. Accordingly,

$$\mathbf{X}\tilde{\mathbf{r}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda} = \mathbf{X}[(\mathbf{X}'\mathbf{X})^{-}]'\boldsymbol{\lambda}. \tag{4.37}$$

Under the general linear model, the variance of the least squares estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is [in light of result (4.37) and the equality $\tilde{\mathbf{r}}'\mathbf{X}' = (\mathbf{X}\tilde{\mathbf{r}})'$] expressible as

$$\mathrm{var}(\tilde{\mathbf{r}}'\mathbf{X}'\mathbf{y}) = \tilde{\mathbf{r}}'\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})\mathbf{X}\tilde{\mathbf{r}} = \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda}. \tag{4.38}$$

Result (4.38) can be extended. Suppose that $\boldsymbol{\lambda}_1'\boldsymbol{\beta}$ and $\boldsymbol{\lambda}_2'\boldsymbol{\beta}$ are two estimable linear combinations of the elements of $\boldsymbol{\beta}$. Then, the least squares estimator of $\boldsymbol{\lambda}_1'\boldsymbol{\beta}$ equals $\tilde{\mathbf{r}}_1'\mathbf{X}'\mathbf{y}$ and that of $\boldsymbol{\lambda}_2'\boldsymbol{\beta}$ equals $\tilde{\mathbf{r}}_2'\mathbf{X}'\mathbf{y}$. Here, $\tilde{\mathbf{r}}_1$ is any solution to the linear system $\mathbf{X}'\mathbf{X}\mathbf{r}_1 = \boldsymbol{\lambda}_1$ (in $\mathbf{r}_1$) and $\tilde{\mathbf{r}}_2$ any solution to the linear system $\mathbf{X}'\mathbf{X}\mathbf{r}_2 = \boldsymbol{\lambda}_2$ (in $\mathbf{r}_2$). And under the general linear model,

$$\mathrm{cov}(\tilde{\mathbf{r}}_1'\mathbf{X}'\mathbf{y},\ \tilde{\mathbf{r}}_2'\mathbf{X}'\mathbf{y}) = \tilde{\mathbf{r}}_1'\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})\mathbf{X}\tilde{\mathbf{r}}_2 = \boldsymbol{\lambda}_1'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda}_2. \tag{4.39}$$

In the special case of the Aitken model, result (4.38) "simplifies" to

$$\mathrm{var}(\tilde{\mathbf{r}}'\mathbf{X}'\mathbf{y}) = \sigma^2\tilde{\mathbf{r}}'\mathbf{X}'\mathbf{H}\mathbf{X}\tilde{\mathbf{r}} = \sigma^2\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{H}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda}, \tag{4.40}$$

and result (4.39) to

$$\text{cov}(\tilde{\mathbf{r}}_1'\mathbf{X}'\mathbf{y},\ \tilde{\mathbf{r}}_2'\mathbf{X}'\mathbf{y}) = \sigma^2\tilde{\mathbf{r}}_1'\mathbf{X}'\mathbf{H}\mathbf{X}\tilde{\mathbf{r}}_2 = \sigma^2\boldsymbol{\lambda}_1'(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{H}\mathbf{X}(\mathbf{X}'\mathbf{X})^-\boldsymbol{\lambda}_2. \tag{4.41}$$

Under the G–M model, considerable further simplification is possible, and various additional representations are obtainable. Specifically, we find that (under the G–M model)

$$\text{var}(\tilde{\mathbf{r}}'\mathbf{X}'\mathbf{y}) = \sigma^2\tilde{\mathbf{r}}'\mathbf{X}'\mathbf{X}\tilde{\mathbf{r}} = \sigma^2\tilde{\mathbf{r}}'\boldsymbol{\lambda} = \sigma^2\boldsymbol{\lambda}'\tilde{\mathbf{r}} \tag{4.42}$$

$$= \sigma^2\tilde{\mathbf{r}}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X}\tilde{\mathbf{r}} = \sigma^2\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^-\boldsymbol{\lambda}, \tag{4.43}$$

and, similarly,

$$\text{cov}(\tilde{\mathbf{r}}_1'\mathbf{X}'\mathbf{y},\ \tilde{\mathbf{r}}_2'\mathbf{X}'\mathbf{y}) = \sigma^2\tilde{\mathbf{r}}_1'\mathbf{X}'\mathbf{X}\tilde{\mathbf{r}}_2 = \sigma^2\tilde{\mathbf{r}}_1'\boldsymbol{\lambda}_2 = \sigma^2\boldsymbol{\lambda}_1'\tilde{\mathbf{r}}_2 \tag{4.44}$$

$$= \sigma^2\tilde{\mathbf{r}}_1'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X}\tilde{\mathbf{r}}_2 = \sigma^2\boldsymbol{\lambda}_1'(\mathbf{X}'\mathbf{X})^-\boldsymbol{\lambda}_2. \tag{4.45}$$

d. The geometry of least squares

It can be informative to consider the method of least squares from a geometrical perspective. As a preliminary to doing so, let us extend some of the basic definitions of plane and solid geometry from $\mathcal{R}^2$ and $\mathcal{R}^3$, where they can be interpreted visually, to $\mathcal{R}^M$ (where M may be greater than 3).

Geometrically-related definitions. The *inner* (or dot) *product* of two M-dimensional column vectors, say $\mathbf{x} = \{x_i\}$ and $\mathbf{y} = \{y_i\}$, is denoted by the symbol $\mathbf{x}\bullet\mathbf{y}$. For a general definition of the inner product of two M-dimensional column vectors (or two $M \times N$ matrices), refer to Section 2.4f. The usual inner product is that for which

$$\mathbf{x}\bullet\mathbf{y} = \mathbf{y}'\mathbf{x} = \mathbf{x}'\mathbf{y} = \sum_{i=1}^{M} x_i y_i. \tag{4.46}$$

The usual inner product is (in the special cases $M = 2$ and $M = 3$) the inner product customarily employed in plane and solid geometry.

The definition of an inner product underlies the definitions of various other quantities. Consider, in particular, the *norm* (also known as the length or magnitude) of an M-dimensional column vector $\mathbf{x} = \{x_i\}$. The norm of $\mathbf{x}$ is denoted by the symbol $\|\mathbf{x}\|$. By definition,

$$\|\mathbf{x}\| = (\mathbf{x}\bullet\mathbf{x})^{1/2}.$$

When the inner product is taken to be the usual inner product,

$$\|\mathbf{x}\| = (\mathbf{x}'\mathbf{x})^{1/2} = \Big(\sum_{i=1}^{M} x_i^2\Big)^{1/2}, \tag{4.47}$$

and the norm is referred to as the usual norm.

The *distance* between two M-dimensional column vectors $\mathbf{x} = \{x_i\}$ and $\mathbf{y} = \{y_i\}$ is defined to be the norm $\|\mathbf{x}-\mathbf{y}\|$ of the difference $\mathbf{x}-\mathbf{y}$ between $\mathbf{x}$ and $\mathbf{y}$. In the case of the usual inner product,

$$\|\mathbf{x}-\mathbf{y}\| = [(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})]^{1/2} = \Big[\sum_{i=1}^{M}(x_i-y_i)^2\Big]^{1/2}. \tag{4.48}$$

The *angle* between two nonnull M-dimensional column vectors $\mathbf{x} = \{x_i\}$ and $\mathbf{y} = \{y_i\}$ is defined indirectly in terms of its cosine. Specifically, the angle between $\mathbf{x}$ and $\mathbf{y}$ is the angle θ $(0 \le \theta \le \pi)$ defined by

$$\cos\theta = \frac{\mathbf{x}\bullet\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|} \tag{4.49}$$

—it follows from Theorem 2.4.21 (the Cauchy–Schwarz inequality) that $-1 \le \dfrac{\mathbf{x}\bullet\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} \le 1$. In the

case of the usual inner product (and usual norm), equality (4.49) can be reexpressed in the form

$$\cos\theta = \frac{\mathbf{x}'\mathbf{y}}{(\mathbf{x}'\mathbf{x})^{1/2}(\mathbf{y}'\mathbf{y})^{1/2}} = \frac{\sum_{i=1}^{M} x_i y_i}{\left(\sum_{i=1}^{M} x_i^2\right)^{1/2}\left(\sum_{i=1}^{M} y_i^2\right)^{1/2}}. \tag{4.50}$$

By definition, two M-dimensional column vectors $\mathbf{x} = \{x_i\}$ and $\mathbf{y} = \{y_i\}$ are *orthogonal* (or perpendicular) to each other if $\mathbf{x} \bullet \mathbf{y} = 0$. Thus, when the inner product is taken to be the usual inner product, $\mathbf{x}$ and $\mathbf{y}$ are orthogonal to each other if $\mathbf{x}'\mathbf{y} = 0$ or, equivalently, if $\sum_{i=1}^{M} x_i y_i = 0$. The statement that $\mathbf{x}$ and $\mathbf{y}$ are orthogonal to each other is sometimes abbreviated to the statement that $\mathbf{x}$ and $\mathbf{y}$ are orthogonal. Clearly, two nonnull vectors are orthogonal if and only if the angle between them is $\pi/2$ (90°) or, equivalently, the cosine of that angle is 0.

If an M-dimensional column vector $\mathbf{x}$ is orthogonal to every vector in a subspace $\mathcal{U}$ of M-dimensional column vectors, $\mathbf{x}$ is said to be orthogonal to $\mathcal{U}$. The set consisting of all M-dimensional column vectors that are orthogonal to the subspace $\mathcal{U}$ is called the *orthogonal complement* of $\mathcal{U}$ and is denoted by the symbol $\mathcal{U}^{\perp}$. The set $\mathcal{U}^{\perp}$ is a linear space (as can be readily verified). When $\mathcal{U} = \mathcal{C}(\mathbf{X})$ (where $\mathbf{X}$ is a matrix), we may write $\mathcal{C}^{\perp}(\mathbf{X})$ for $\mathcal{U}^{\perp}$.

Least squares revisited: the projection and decomposition of the data vector. Denote by $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$ an N-dimensional column vector of data points—this notation differs somewhat from that employed earlier in the section (which included an underline). Further, suppose that $y_1, y_2, \ldots, y_N$ are accompanied by the corresponding values $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$ of a C-dimensional column vector $\mathbf{u}$ of explanatory variables. Let us consider the approximation of $y_1, y_2, \ldots, y_N$ by $\underline{\delta}(\mathbf{u}_1), \underline{\delta}(\mathbf{u}_2), \ldots, \underline{\delta}(\mathbf{u}_N)$, where $\underline{\delta}(\mathbf{u})$ is a function of $\mathbf{u}$. Which of the possible choices for the function $\underline{\delta}(\cdot)$ results in the "best" approximation (and in what sense)? In particular, which results in the best approximation when the choice for $\underline{\delta}(\cdot)$ is restricted to functions (of $\mathbf{u}$) that are expressible as linear combinations of P specified functions $\delta_1(\cdot), \delta_2(\cdot), \ldots, \delta_P(\cdot)$; that is, when the choice is restricted to those functions that are expressible in the form

$$\underline{\delta}(\mathbf{u}) = b_1\delta_1(\mathbf{u}) + b_2\delta_2(\mathbf{u}) + \cdots + b_P\delta_P(\mathbf{u}), \tag{4.51}$$

where $b_1, b_2, \ldots, b_P$ are arbitrary scalars.

In the method of least squares, the function $\underline{\delta}(\cdot)$ is chosen so as to minimize the quantity $\{\sum_{i=1}^{N}[y_i - \underline{\delta}(\mathbf{u}_i)]^2\}^{1/2}$. This quantity is the (usual) norm of the N-dimensional vector whose elements are the individual errors of approximation $y_1 - \underline{\delta}(\mathbf{u}_1), y_2 - \underline{\delta}(\mathbf{u}_2), \ldots, y_N - \underline{\delta}(\mathbf{u}_N)$. It is interpretable as the (ordinary) distance between the N-dimensional data vector $\mathbf{y}$ and the N-dimensional vector whose elements are the approximations $\underline{\delta}(\mathbf{u}_1), \underline{\delta}(\mathbf{u}_2), \ldots, \underline{\delta}(\mathbf{u}_N)$.

Suppose now that $\underline{\delta}(\cdot)$ is taken to be of the form (4.51). And (for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, P$) define $x_{ij} = \delta_j(\mathbf{u}_i)$. Further, let $\mathbf{X}$ represent the $N \times P$ matrix with ijth element x_{ij}, and (for $j = 1, 2, \ldots, P$) take $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{Nj})'$ (so that $\mathbf{x}_j$ is the jth column of $\mathbf{X}$). Then, as discussed earlier (in the introductory part of the present section),

$$\sum_{i=1}^{N}[y_i - \underline{\delta}(\mathbf{u}_i)]^2 = \sum_{i=1}^{N}\Big(y_i - \sum_{j=1}^{P} x_{ij}b_j\Big)^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}), \tag{4.52}$$

where $\mathbf{b} = (b_1, b_2, \ldots, b_P)'$. Note [in connection with result (4.52)] that

$$(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \Big(\mathbf{y} - \sum_{j=1}^{P} b_j\mathbf{x}_j\Big)'\Big(\mathbf{y} - \sum_{j=1}^{P} b_j\mathbf{x}_j\Big).$$

In light of result (4.52), the minimization problem that gives rise to the method of least squares can be regarded as that of minimizing $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$ [or, equivalently, that of minimizing the (usual) norm of $\mathbf{y} - \mathbf{X}\mathbf{b}$] with respect to the $P \times 1$ vector $\mathbf{b}$. As previously indicated (in Subsection

b), $(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})$ attains a minimum value at a $P \times 1$ vector $\tilde{\mathbf{b}}$ if and only if $\tilde{\mathbf{b}}$ is a solution to the normal equations $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$.

Some further insights into the method of least squares can be obtained by transforming the underlying minimization problem into a more geometrically meaningful form. Let $\mathcal{U} = \mathcal{C}(\mathbf{X})$, and observe that an N-dimensional column vector $\mathbf{w}$ is a member of $\mathcal{U}$ if and only if $\mathbf{w} = \mathbf{X}\mathbf{b}$ for some $\mathbf{b}$, in which case the elements of $\mathbf{b}$ are the "coordinates" of $\mathbf{w}$ with respect to the spanning set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_P\}$. Accordingly, the problem of minimizing $(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})$ with respect to $\mathbf{b}$ can be reformulated as the "coordinate-free" problem of minimizing $(\mathbf{y}-\mathbf{w})'(\mathbf{y}-\mathbf{w})$ with respect to $\mathbf{w}$, where $\mathbf{w}$ is an arbitrary member of the linear space $\mathcal{U}$. The latter problem depends on the matrix $\mathbf{X}$ only through its column space. From a geometrical perspective, the problem is that of finding the vector in the subspace $\mathcal{U}$ (of $\mathcal{R}^N$) that is "closest" to the data vector $\mathbf{y}$.

It follows from the results of Subsection b that $(\mathbf{y}-\mathbf{w})'(\mathbf{y}-\mathbf{w})$ attains a minimum value over the subspace $\mathcal{U}$, doing so at a unique point $\mathbf{z}$ that is expressible as

$$\mathbf{z} = \mathbf{X}\tilde{\mathbf{b}}, \tag{4.53}$$

where $\tilde{\mathbf{b}}$ is any solution to the normal equations $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$, and also as

$$\mathbf{z} = \mathbf{P}_{\mathbf{X}}\mathbf{y}. \tag{4.54}$$

Taking (here and in the remainder of the present subsection) the inner product to be the usual inner product, the vector $\mathbf{z}$ is such that $\mathbf{y}-\mathbf{z} \in \mathcal{U}^{\perp}$; that is, the difference between $\mathbf{y}$ and $\mathbf{z}$ is orthogonal to every vector in $\mathcal{U}$. To see this, let $\mathbf{a}$ represent an arbitrary member of $\mathcal{U}$ $[= \mathcal{C}(\mathbf{X})]$, and observe that $\mathbf{a} = \mathbf{X}\mathbf{r}$ for some $P \times 1$ vector $\mathbf{r}$ and hence that

$$\mathbf{a}'(\mathbf{y}-\mathbf{z}) = \mathbf{r}'\mathbf{X}'(\mathbf{y}-\mathbf{X}\tilde{\mathbf{b}}) = \mathbf{r}'(\mathbf{X}'\mathbf{y}-\mathbf{X}'\mathbf{X}\tilde{\mathbf{b}}) = \mathbf{r}'(\mathbf{X}'\mathbf{y}-\mathbf{X}'\mathbf{y}) = \mathbf{r}'\mathbf{0} = 0.$$

Moreover, there is no member $\mathbf{w}$ of $\mathcal{U}$ other than $\mathbf{z}$ for which $\mathbf{y}-\mathbf{w} \in \mathcal{U}^{\perp}$, as is evident upon observing that if $\mathbf{w} \in \mathcal{U}$ and $\mathbf{y}-\mathbf{w} \in \mathcal{U}^{\perp}$, then $\mathbf{w}-\mathbf{z} \in \mathcal{U}$ and

$$\mathbf{w}-\mathbf{z} = (\mathbf{y}-\mathbf{z}) - (\mathbf{y}-\mathbf{w}) \in \mathcal{U}^{\perp}$$

(so that the vector $\mathbf{w}-\mathbf{z}$ is orthogonal to itself), implying that

$$(\mathbf{w}-\mathbf{z})'(\mathbf{w}-\mathbf{z}) = 0$$

and hence that $\mathbf{w}-\mathbf{z} = \mathbf{0}$ or, equivalently, that

$$\mathbf{w} = \mathbf{z}.$$

In summary, there is a unique vector $\mathbf{w}$ in $\mathcal{U}$ such that $\mathbf{y}-\mathbf{w} \in \mathcal{U}^{\perp}$, namely, the vector $\mathbf{z}$. This vector is referred to as the *orthogonal projection* of $\mathbf{y}$ on $\mathcal{U}$ or simply as the *projection* of $\mathbf{y}$ on $\mathcal{U}$. As previously indicated, the matrix $\mathbf{P}_{\mathbf{X}}$ is referred to as a projection matrix; the reason why is apparent from expression (4.54).

Conceptually, the point $\mathbf{z}$ in $\mathcal{R}^N$ at which $(\mathbf{y}-\mathbf{w})'(\mathbf{y}-\mathbf{w})$ attains its minimum value for $\mathbf{w} \in \mathcal{U}$ is obtainable by "projecting" the point $\mathbf{y}$ onto the surface $\mathcal{U}$. The point in $\mathcal{R}^N$ located by this operation is such that the "line" formed by joining that point with the point $\mathbf{y}$ is orthogonal (perpendicular) to the surface $\mathcal{U}$.

Corresponding to the projection $\mathbf{z}$ of $\mathbf{y}$ on $\mathcal{U}$ is the decomposition

$$\mathbf{y} = \mathbf{z} + \mathbf{d}, \tag{4.55}$$

where $\mathbf{d} = \mathbf{y}-\mathbf{z}$. The first component of this decomposition is a member of the linear space $\mathcal{U}$ $[= \mathcal{C}(\mathbf{X})]$, and the second component is a member of the orthogonal complement $\mathcal{U}^{\perp}$ $[= \mathcal{C}^{\perp}(\mathbf{X})]$ of

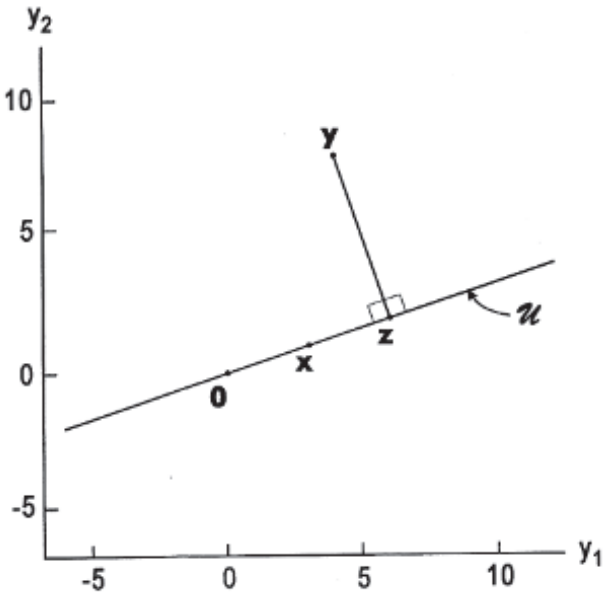


FIGURE 5.1. The projection $\mathbf{z}$ of the 2-dimensional data vector $\mathbf{y} = (4, 8)'$ on the 1-dimensional linear space $\mathcal{U} = \mathcal{C}(\mathbf{X})$, where $\mathbf{X} = \mathbf{x} = (3, 1)'$.

$\mathcal{U}$. In this context, the linear space $\mathcal{U}$ is sometimes referred to as the *estimation space*—logically, it could also be referred to as the approximation space—and the linear space $\mathcal{U}^{\perp}$ is sometimes referred to as the *error space*. Decomposition (4.55) is unique; if $\mathbf{y}$ is expressed as the sum of two components, the first of which is in the estimation space $\mathcal{U}$ and the second of which is in the error space $\mathcal{U}^{\perp}$, then necessarily the first component equals $\mathbf{z}$ and the second equals $\mathbf{d}$ $(= \mathbf{y} - \mathbf{z})$.

Example: N = 2. Suppose that $N = 2$, that $\mathbf{y} = (4, 8)'$, and that $\mathbf{X} = \mathbf{x}$, where $\mathbf{x}$ is the 2-dimensional column vector $\mathbf{x} = (3, 1)'$ (in which case $P = 1$). Then, the linear system $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$ becomes $(10)\mathbf{b} = (20)$, which has the unique solution $\mathbf{b} = (2)$. Thus, the projection of $\mathbf{y}$ on the linear space $\mathcal{U}$ $[= \mathcal{C}(\mathbf{X})]$ is the vector

$$\mathbf{z} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}(2) = \begin{pmatrix} 6 \\ 2 \end{pmatrix},$$

as depicted in Figure 5.1.

Example: N = 3. Suppose that $N = 3$, that $\mathbf{y} = (3, -38/5, 74/5)'$, and that $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$, where

$$\mathbf{x}_1 = \begin{pmatrix} 0 \\ 3 \\ 6 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} -2 \\ 2 \\ 4 \end{pmatrix}, \quad \text{and} \quad \mathbf{x}_3 = \begin{pmatrix} -2 \\ 1 \\ 2 \end{pmatrix}.$$

Clearly, $\mathbf{x}_1$ and $\mathbf{x}_2$ are linearly independent, and $\mathbf{x}_3 = \mathbf{x}_2 - (1/3)\mathbf{x}_1$. Thus, the linear space $\mathcal{U}$ $[= \mathcal{C}(\mathbf{X})]$ is of dimension 2.

The normal equations $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$ become

$$\begin{pmatrix} 45 & 30 & 15 \\ 30 & 24 & 14 \\ 15 & 14 & 9 \end{pmatrix}\mathbf{b} = \begin{pmatrix} 66 \\ 38 \\ 16 \end{pmatrix}.$$

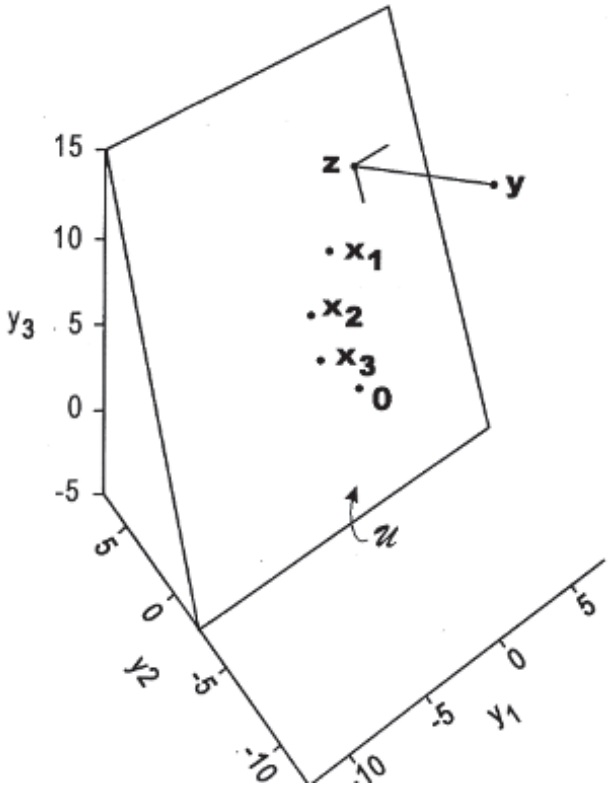


FIGURE 5.2. The projection $\mathbf{z}$ of the 3-dimensional data vector $\mathbf{y} = (3, -38/5, 74/5)'$ on the 2-dimensional linear space $\mathcal{U} = \mathcal{C}(\mathbf{X})$, where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$, with $\mathbf{x}_1 = (0, 3, 6)'$, $\mathbf{x}_2 = (-2, 2, 4)'$, and $\mathbf{x}_3 = (-2, 1, 2)'$.

One solution to these equations is the vector $(32/15, -1/2, -1)'$. Thus, the projection of $\mathbf{y}$ on the linear space $\mathcal{U}$ $[= \mathcal{C}(\mathbf{X})]$ is

$$\mathbf{z} = \begin{pmatrix} 0 & -2 & -2 \\ 3 & 2 & 1 \\ 6 & 4 & 2 \end{pmatrix} \begin{pmatrix} 32/15 \\ -1/2 \\ -1 \end{pmatrix} = \begin{pmatrix} 3 \\ 22/5 \\ 44/5 \end{pmatrix},$$

as depicted in Figure 5.2.

e. Least squares computations

Let us continue to take $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$ to be an N-dimensional vector of data points and to suppose that $y_1, y_2, \ldots, y_N$ are accompanied by the values $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$ of a C-dimensional (column) vector $\mathbf{u}$ of explanatory variables. And let us continue to consider the approximation of $y_1, y_2, \ldots, y_N$ by $\underline{\delta}(\mathbf{u}_1), \underline{\delta}(\mathbf{u}_2), \ldots, \underline{\delta}(\mathbf{u}_N)$, where $\underline{\delta}(\cdot)$ is a function (of $\mathbf{u}$) that is expressible in the form of a linear combination

$$\underline{\delta}(\mathbf{u}) = b_1\delta_1(\mathbf{u}) + b_2\delta_2(\mathbf{u}) + \cdots + b_P\delta_P(\mathbf{u})$$

of P specified functions $\delta_1(\cdot), \delta_2(\cdot), \ldots, \delta_P(\cdot)$. Further, take $\mathbf{b} = (b_1, b_2, \ldots, b_P)'$ to be the $P \times 1$ vector of coefficients, and take $\mathbf{X}$ to be the $N \times P$ matrix whose ijth element x_{ij} is defined by $x_{ij} = \delta_j(\mathbf{u}_i)$. In the method of least squares, the value of $\mathbf{b}$ is taken to be a value at which the quantity $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$ attains a minimum value.

As discussed in Subsection b, $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$ attains a minimum value at a $P \times 1$ vector $\tilde{\mathbf{b}}$ if and only if $\tilde{\mathbf{b}}$ is a solution to the normal equations $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$. Accordingly, the least squares computations can be carried out by forming and solving the normal equations. Alternatively, by making use of various results on the decomposition of a matrix (as applied to the matrix $\mathbf{X}$), they can be carried out in a way that does not require the formation of the normal equations. The alternative approach can be advantageous from the standpoint of numerical accuracy, though any such advantage typically comes at the expense of greater demands on computing resources. In many implementations of the alternative approach, the underlying decomposition of the matrix $\mathbf{X}$ is taken to be a decomposition that is known as the QR decomposition.

QR decomposition of a matrix. Any $M \times N$ matrix $\mathbf{A}$ of full column rank N is expressible in the form

$$\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1, \tag{4.56}$$

where $\mathbf{Q}_1$ is an $M \times N$ matrix with orthonormal columns and $\mathbf{R}_1$ is an upper triangular matrix with (strictly) positive diagonal elements. Moreover, the matrices $\mathbf{Q}_1$ and $\mathbf{R}_1$ are unique. The existence of a decomposition of the form (4.56) can be established by, for example, applying Gram–Schmidt orthogonalization. For a proof of the existence and uniqueness of a decomposition of the form (4.56), refer, for example, to Harville (1997, sec. 6.4).

As a variation on expression (4.56), we have the expression

$$\mathbf{A} = \mathbf{Q}\mathbf{R}, \tag{4.57}$$

where $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2)$ is an $M \times M$ orthogonal matrix and where $\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{pmatrix}$. The columns of the $M \times (M-N)$ submatrix $\mathbf{Q}_2$ are any $M-N$ M-dimensional column vectors that together with the N columns of $\mathbf{Q}_1$ form an orthonormal basis for $\mathcal{R}^M$.

Either of the two decompositions (4.56) and (4.57) might be referred to as a *QR decomposition*.

QR decomposition as a basis for least squares computations. Let us now resume our discussion of the computational aspects of the minimization of $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$. Assume that the $N \times P$ matrix $\mathbf{X}$ is of full column rank P—discussion of the general case where $\text{rank}(\mathbf{X})$ may be less than P is deferred until the final part of the present subsection. Consider the QR decomposition of $\mathbf{X}$. That is, consider a decomposition of $\mathbf{X}$ of the form

$$\mathbf{X} = \mathbf{Q}_1\mathbf{R}_1, \tag{4.58}$$

where $\mathbf{Q}_1$ is an $N \times P$ matrix with orthonormal columns and $\mathbf{R}_1$ is an upper triangular matrix with (strictly) positive diagonal elements, or of the related form

$$\mathbf{X} = \mathbf{Q}\mathbf{R}, \tag{4.59}$$

where $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2)$ is an $N \times N$ orthogonal matrix and where $\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{pmatrix}$.

Let $\mathbf{z} = \mathbf{Q}'\mathbf{y}$, and partition $\mathbf{z}$ as $\mathbf{z} = \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix}$, where $\mathbf{z}_1 = \mathbf{Q}_1'\mathbf{y}$ and $\mathbf{z}_2 = \mathbf{Q}_2'\mathbf{y}$. Then,

$$\begin{aligned} \mathbf{y} - \mathbf{X}\mathbf{b} &= \mathbf{Q}(\mathbf{z} - \mathbf{R}\mathbf{b}) \\ &= \mathbf{Q}_1(\mathbf{z}_1 - \mathbf{R}_1\mathbf{b}) + \mathbf{Q}_2\mathbf{z}_2. \end{aligned} \tag{4.60}$$

And

$$\begin{aligned} (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) &= (\mathbf{z} - \mathbf{R}\mathbf{b})'\mathbf{Q}'\mathbf{Q}(\mathbf{z} - \mathbf{R}\mathbf{b}) \\ &= (\mathbf{z} - \mathbf{R}\mathbf{b})'(\mathbf{z} - \mathbf{R}\mathbf{b}) \\ &= (\mathbf{z}_1 - \mathbf{R}_1\mathbf{b})'(\mathbf{z}_1 - \mathbf{R}_1\mathbf{b}) + \mathbf{z}_2'\mathbf{z}_2. \end{aligned} \tag{4.61}$$

Now, consider the linear system

$$\mathbf{R}_1\mathbf{b} = \mathbf{z}_1 \tag{4.62}$$

(in the vector $\mathbf{b}$), whose coefficient matrix is $\mathbf{R}_1$ and right side is $\mathbf{z}_1$. It follows from elementary results on triangular matrices (e.g., Harville 1997, corollary 8.5.6) that $\mathbf{R}_1$ is nonsingular. Thus, linear system (4.62) has a unique solution, say $\tilde{\mathbf{b}}$. Clearly, the first term of expression (4.61) equals 0 if $\mathbf{b} = \tilde{\mathbf{b}}$, and is greater than 0 otherwise (i.e., if $\mathbf{b} \neq \tilde{\mathbf{b}}$). And because the second term of expression (4.61) does not depend on $\mathbf{b}$, we conclude that $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$ attains a minimum value of $\mathbf{z}_2'\mathbf{z}_2$ and does so uniquely at the point $\tilde{\mathbf{b}}$. Moreover, in light of result (4.60),

$$\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}} = \mathbf{Q}_2\mathbf{z}_2. \tag{4.63}$$

These results serve as the basis for an alternative approach to the least squares computations (not requiring the formation of the normal equations). In the alternative approach, the formation of the matrix $\mathbf{R}_1$ and the vector $\mathbf{z}_1$ are at the heart of the computations. Their formation can be accomplished through the use of Householder transformations (reflections) or Givens transformations (rotations) or through the use of a modified Gram–Schmidt procedure—refer, for example, to Golub and Van Loan (2013, chap. 5) for a detailed discussion. The value $\tilde{\mathbf{b}}$ at which $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$ attains its minimum value is determined from $\mathbf{R}_1$ and $\mathbf{z}_1$ by solving linear system (4.62), doing so in a way that exploits the triangularity of $\mathbf{R}_1$—refer, for example, to Harville (1997, sec. 11.8) for a discussion of the solution of a linear system with a triangular coefficient matrix.

Our results on the alternative approach to the least squares computations can be extended to the general case where the matrix $\mathbf{X}$ is not necessarily of full column rank. The extension requires some familiarity with a type of matrix called a permutation matrix.

Permutation matrices. A *permutation matrix* is a square matrix whose columns can be obtained by permuting (rearranging) the columns of an identity matrix. Thus, letting $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$ represent the first, second, ..., Nth columns, respectively, of $\mathbf{I}_N$, an $N \times N$ permutation matrix is a matrix of the general form $(\mathbf{u}_{k_1}, \mathbf{u}_{k_2}, \ldots, \mathbf{u}_{k_N})$, where $k_1, k_2, \ldots, k_N$ is an arbitrary permutation of the first N positive integers $1, 2, \ldots, N$. For example, one permutation matrix of order $N = 3$ is the 3×3 matrix

$$(\mathbf{u}_3, \mathbf{u}_1, \mathbf{u}_2) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix},$$

whose columns are the third, first, and second columns, respectively, of $\mathbf{I}_3$. Clearly, the columns of any permutation matrix form an orthonormal (with respect to the usual inner product) set, and hence any permutation matrix is an orthogonal matrix.

The jth element of the k_jth row of the $N \times N$ permutation matrix $(\mathbf{u}_{k_1}, \mathbf{u}_{k_2}, \ldots, \mathbf{u}_{k_N})$ is 1, and its other $N-1$ elements are 0. That is, the jth row $\mathbf{u}_j'$ of $\mathbf{I}_N$ is the k_jth row of $(\mathbf{u}_{k_1}, \mathbf{u}_{k_2}, \ldots, \mathbf{u}_{k_N})$ or, equivalently, the jth column $\mathbf{u}_j$ of $\mathbf{I}_N$ is the k_jth column of $(\mathbf{u}_{k_1}, \mathbf{u}_{k_2}, \ldots, \mathbf{u}_{k_N})'$. Thus, the transpose of any permutation matrix is itself a permutation matrix. Further, the rows of any permutation matrix are a permutation of the rows of an identity matrix and, conversely, any matrix whose rows can be obtained by permuting the rows of an identity matrix is a permutation matrix.

The effect of postmultiplying an $M \times N$ matrix $\mathbf{A}$ by an $N \times N$ permutation matrix $\mathbf{P}$ is to permute the columns of $\mathbf{A}$ in the same way that the columns of $\mathbf{I}_N$ were permuted in forming $\mathbf{P}$. Thus, if $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N$ are the first, second, ..., Nth columns of $\mathbf{A}$, the first, second, ..., Nth columns of the product $\mathbf{A}(\mathbf{u}_{k_1}, \mathbf{u}_{k_2}, \ldots, \mathbf{u}_{k_N})$ of $\mathbf{A}$ and the $N \times N$ permutation matrix $(\mathbf{u}_{k_1}, \mathbf{u}_{k_2}, \ldots, \mathbf{u}_{k_N})$ are $\mathbf{a}_{k_1}, \mathbf{a}_{k_2}, \ldots, \mathbf{a}_{k_N}$, respectively. Further, the first, second ..., Nth columns $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N$ of $\mathbf{A}$ are the $k_1, k_2, \ldots, k_N$th columns, respectively, of the product $\mathbf{A}(\mathbf{u}_{k_1}, \mathbf{u}_{k_2}, \ldots, \mathbf{u}_{k_N})'$ of $\mathbf{A}$ and the permutation matrix $(\mathbf{u}_{k_1}, \mathbf{u}_{k_2}, \ldots, \mathbf{u}_{k_N})'$. When $N = 3$, we have, for example, that

$$\mathbf{A}(\mathbf{u}_3, \mathbf{u}_1, \mathbf{u}_2) = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = (\mathbf{a}_3, \mathbf{a}_1, \mathbf{a}_2)$$

and

$$\mathbf{A}(\mathbf{u}_3, \mathbf{u}_1, \mathbf{u}_2)' = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} = (\mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_1).$$

Similarly, the effect of premultiplying an $N \times M$ matrix $\mathbf{A}$ by an $N \times N$ permutation matrix is to permute the rows of $\mathbf{A}$. If the first, second, ..., Nth rows of $\mathbf{A}$ are $\mathbf{a}_1', \mathbf{a}_2', \ldots, \mathbf{a}_N'$, respectively, then the first, second, ..., Nth rows of the product $(\mathbf{u}_{k_1}, \mathbf{u}_{k_2}, \ldots, \mathbf{u}_{k_N})'\mathbf{A}$ of the permutation matrix $(\mathbf{u}_{k_1}, \mathbf{u}_{k_2}, \ldots, \mathbf{u}_{k_N})'$ and $\mathbf{A}$ are $\mathbf{a}_{k_1}', \mathbf{a}_{k_2}', \ldots, \mathbf{a}_{k_N}'$, respectively, and $\mathbf{a}_1', \mathbf{a}_2', \ldots, \mathbf{a}_N'$ are the $k_1, k_2, \ldots, k_N$th rows, respectively, of $(\mathbf{u}_{k_1}, \mathbf{u}_{k_2}, \ldots, \mathbf{u}_{k_N})\mathbf{A}$. When $N = 3$, we have, for example, that

$$(\mathbf{u}_3, \mathbf{u}_1, \mathbf{u}_2)'\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \mathbf{a}_3' \end{pmatrix} = \begin{pmatrix} \mathbf{a}_3' \\ \mathbf{a}_1' \\ \mathbf{a}_2' \end{pmatrix}$$

and

$$(\mathbf{u}_3, \mathbf{u}_1, \mathbf{u}_2)\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}\begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \mathbf{a}_3' \end{pmatrix} = \begin{pmatrix} \mathbf{a}_2' \\ \mathbf{a}_3' \\ \mathbf{a}_1' \end{pmatrix}.$$

Alternative approach to least squares computations: general case. Let us now extend our initial results on the alternative approach to the least squares computations. Accordingly, suppose that we wish to minimize the quantity $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$ and that $\text{rank}(\mathbf{X}) = K$, where K is possibly less than P—our initial results (the results of Part 2) were obtained under the simplifying assumption that the $N \times P$ matrix $\mathbf{X}$ is of full column rank P.

Let $\mathbf{L}$ represent any $P \times P$ permutation matrix such that the first K columns of the $N \times P$ matrix $\mathbf{X}\mathbf{L}$ are linearly independent, and partition $\mathbf{L}$ as $\mathbf{L} = (\mathbf{L}_1, \mathbf{L}_2)$, where $\mathbf{L}_1$ is of dimensions $P \times K$. Then,

$$\mathbf{X}\mathbf{L} = (\mathbf{X}\mathbf{L}_1, \mathbf{X}\mathbf{L}_2),$$

and $\mathbf{X}\mathbf{L}_1$ is of full column rank K. Decompose $\mathbf{X}\mathbf{L}_1$ as

$$\mathbf{X}\mathbf{L}_1 = \mathbf{Q}_1\mathbf{R}_1, \tag{4.64}$$

where $\mathbf{Q}_1$ is an $N \times K$ matrix with orthonormal columns and $\mathbf{R}_1$ is an upper triangular matrix with (strictly) positive diagonal elements. And observe that the columns of $\mathbf{Q}_1$ form a basis for $\mathcal{C}(\mathbf{X}\mathbf{L})$ $[= \mathcal{C}(\mathbf{X}\mathbf{L}_1)]$, so that

$$\mathbf{X}\mathbf{L}_2 = \mathbf{Q}_1\mathbf{R}_2, \tag{4.65}$$

for some matrix $\mathbf{R}_2$. Together, results (4.64) and (4.65) imply that

$$\mathbf{X}\mathbf{L} = \mathbf{Q}_1(\mathbf{R}_1, \mathbf{R}_2)$$

and also that

$$\mathbf{X}\mathbf{L} = \mathbf{Q}\mathbf{R},$$

where $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2)$ is an $N \times N$ orthogonal matrix and where $\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & \mathbf{R}_2 \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$. Or, equivalently,

$$\mathbf{X} = \mathbf{Q}_1(\mathbf{R}_1, \mathbf{R}_2)\mathbf{L}' = \mathbf{Q}_1\mathbf{R}_1\mathbf{L}_1' + \mathbf{Q}_1\mathbf{R}_2\mathbf{L}_2'$$

and

$$\mathbf{X} = \mathbf{Q}\mathbf{R}\mathbf{L}'. \tag{4.66}$$

Let $\mathbf{h} = \mathbf{L}'\mathbf{b}$, and partition $\mathbf{h}$ as $\mathbf{h} = \begin{pmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{pmatrix}$, where $\mathbf{h}_1 = \mathbf{L}_1'\mathbf{b}$ and $\mathbf{h}_2 = \mathbf{L}_2'\mathbf{b}$. Further, let

$\mathbf{z} = \mathbf{Q}'\mathbf{y}$, and partition $\mathbf{z}$ as $\mathbf{z} = \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix}$, where $\mathbf{z}_1 = \mathbf{Q}_1'\mathbf{y}$ and $\mathbf{z}_2 = \mathbf{Q}_2'\mathbf{y}$. Then,

$$\begin{aligned} \mathbf{y} - \mathbf{X}\mathbf{b} &= \mathbf{Q}(\mathbf{z} - \mathbf{R}\mathbf{h}) \\ &= \mathbf{Q}_1(\mathbf{z}_1 - \mathbf{R}_1\mathbf{h}_1 - \mathbf{R}_2\mathbf{h}_2) + \mathbf{Q}_2\mathbf{z}_2, \end{aligned} \tag{4.67}$$

which is a generalization of result (4.60). And

$$\begin{aligned} (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) &= (\mathbf{z} - \mathbf{R}\mathbf{h})'\mathbf{Q}'\mathbf{Q}(\mathbf{z} - \mathbf{R}\mathbf{h}) \\ &= (\mathbf{z} - \mathbf{R}\mathbf{h})'(\mathbf{z} - \mathbf{R}\mathbf{h}) \\ &= (\mathbf{z}_1 - \mathbf{R}_1\mathbf{h}_1 - \mathbf{R}_2\mathbf{h}_2)'(\mathbf{z}_1 - \mathbf{R}_1\mathbf{h}_1 - \mathbf{R}_2\mathbf{h}_2) + \mathbf{z}_2'\mathbf{z}_2, \end{aligned} \tag{4.68}$$

which is a generalization of result (4.61).

Now, consider the minimization of $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$ with respect to the transformed vector $\mathbf{h}$ $(= \mathbf{L}'\mathbf{b})$. It follows from result (4.68) that $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$ attains a minimum value of $\mathbf{z}_2'\mathbf{z}_2$ and that it does so at those values of $\mathbf{h}$ for which the first term of expression (4.68) equals 0 or, equivalently, at those values for which $\mathbf{z}_1 - \mathbf{R}_1\mathbf{h}_1 - \mathbf{R}_2\mathbf{h}_2 = \mathbf{0}$. Accordingly, $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$ attains a minimum value of $\mathbf{z}_2'\mathbf{z}_2$ at values $\tilde{\mathbf{h}}_1$ and $\tilde{\mathbf{h}}_2$ of $\mathbf{h}_1$ and $\mathbf{h}_2$, respectively, if and only if $\mathbf{R}_1\tilde{\mathbf{h}}_1 = \mathbf{z}_1 - \mathbf{R}_2\tilde{\mathbf{h}}_2$ or, equivalently, if and only if $\tilde{\mathbf{h}}_1$ is the solution to the linear system

$$\mathbf{R}_1\mathbf{h}_1 = \mathbf{z}_1 - \mathbf{R}_2\tilde{\mathbf{h}}_2 \tag{4.69}$$

(in the vector $\mathbf{h}_1$). Thus, an arbitrary one of the values of $\mathbf{h}$ at which $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$ attains a minimum value is obtained by assigning $\mathbf{h}_2$ an arbitrary value $\tilde{\mathbf{h}}_2$ and by then taking the value $\tilde{\mathbf{h}}_1$ of $\mathbf{h}_1$ to be the solution to linear system (4.69)—the matrix $\mathbf{R}_1$ is nonsingular, so that $\tilde{\mathbf{h}}_1$ is uniquely determined by $\tilde{\mathbf{h}}_2$. In particular, we could take the value of $\mathbf{h}_2$ to be $\mathbf{0}$, and take the value of $\mathbf{h}_1$ to be the (unique) solution to the linear system $\mathbf{R}_1\mathbf{h}_1 = \mathbf{z}_1$.

We conclude that $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$ attains a minimum value of $\mathbf{z}_2'\mathbf{z}_2$ and that it does so at a value $\tilde{\mathbf{b}}$ of $\mathbf{b}$ if and only if for some $(P-K) \times 1$ vector $\tilde{\mathbf{h}}_2$, $\tilde{\mathbf{b}} = \mathbf{L}_1\tilde{\mathbf{h}}_1 + \mathbf{L}_2\tilde{\mathbf{h}}_2$, where $\tilde{\mathbf{h}}_1$ is the solution to linear system (4.69). Note [in light of result (4.67)] that for any such (minimizing) value $\tilde{\mathbf{b}}$ of $\mathbf{b}$,

$$\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}} = \mathbf{Q}_2\mathbf{z}_2. \tag{4.70}$$

These results generalize the results obtained earlier (in Part 2) for the special case where the rank K of the $N \times P$ matrix $\mathbf{X}$ equals P. They provide a basis for extending the alternative approach to the least squares computations to the general case (where K may be less than P). As in the special case, the formation of the matrix $\mathbf{R}_1$ and the vector $\mathbf{z}_1$ are at the heart of the computations. (And, as in the special case, the formation of $\mathbf{R}_1$ and $\mathbf{z}_1$ can be accomplished via any of several procedures devised for that purpose.) In the general case, there is also a need to determine the permutation matrix $\mathbf{L}$ (i.e., to identify K linearly independent columns of $\mathbf{X}$) and possibly the matrix $\mathbf{R}_2$—if the value of $\mathbf{h}_2$ is taken to be $\mathbf{0}$, then $\mathbf{R}_2$ is not needed. A value $\tilde{\mathbf{b}}$ at which $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$ attains its minimum value is determined from $\mathbf{R}_1$, $\mathbf{z}_1$, $\mathbf{L}$, and possibly $\mathbf{R}_2$ by taking $\tilde{\mathbf{h}}_2$ to be any $(P-K) \times 1$ vector, by computing the solution $\tilde{\mathbf{h}}_1$ to linear system (4.69), and by setting $\tilde{\mathbf{b}} = \mathbf{L}_1\tilde{\mathbf{h}}_1 + \mathbf{L}_2\tilde{\mathbf{h}}_2$.

5.5 Best Linear Unbiased or Translation-Equivariant Estimation of Estimable Functions (under the G–M Model)

Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M, Aitken, or general linear model. And consider the least squares estimator $\ell(\mathbf{y})$ of an estimable linear combination $\boldsymbol{\lambda}'\boldsymbol{\beta}$ of

the elements of the parametric vector $\boldsymbol{\beta}$. The least squares estimator is a linear estimator, as was demonstrated in Section 5.4c—refer to representation (4.24) or (4.35). Moreover, the least squares estimator is an unbiased estimator. Its unbiasedness can be established directly by verifying that $\mathrm{E}[\ell(\mathbf{y})] = \boldsymbol{\lambda}'\boldsymbol{\beta}$, as was done in Section 5.4c—refer to result (4.36). Alternatively, its unbiasedness can be established by applying the following result (from Section 5.1) on the unbiasedness of linear estimators: for an estimator of the form $c + \mathbf{a}'\mathbf{y}$ to be an unbiased estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$, it is necessary and sufficient that

$$c = 0 \quad \text{and} \quad \mathbf{X}'\mathbf{a} = \boldsymbol{\lambda}. \tag{5.1}$$

Upon observing [in light of result (4.35)] that $\ell(\mathbf{y}) = (\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y}$, where $\tilde{\mathbf{r}}$ is any solution to the conjugate normal equations $\mathbf{X}'\mathbf{X}\mathbf{r} = \boldsymbol{\lambda}$, it follows immediately from the sufficiency of condition (5.1) that $\ell(\mathbf{y})$ is an unbiased estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$.

The least squares estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is translation equivariant as well as unbiased. To see this, recall (from Section 5.2) that for an estimator of the form $c + \mathbf{a}'\mathbf{y}$ to be a translation-equivariant estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$, it is necessary and sufficient that $\mathbf{a}'\mathbf{X} = \boldsymbol{\lambda}'$ or, equivalently, that $\mathbf{X}'\mathbf{a} = \boldsymbol{\lambda}$. The translation equivariance of the least squares estimator $\ell(\mathbf{y})$ [which is expressible in the form $\ell(\mathbf{y}) = (\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y}$] follows from the sufficiency of the condition $\mathbf{X}'\mathbf{a} = \boldsymbol{\lambda}$ in much the same way that its unbiasedness follows from the sufficiency of condition (5.1).

When $\mathbf{y}$ follows a G–M model, the least squares estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is superior to other linear unbiased or translation-equivariant estimators in a sense that is to be discussed in Subsections a and b. More generally (when $\mathbf{y}$ follows an Aitken or general linear model), this superiority is confined to special cases. These special cases include, of course, G–M models, but also a limited number of other models.

a. Gauss–Markov theorem

In the special case where $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M model, the least squares estimator of an estimable linear combination $\boldsymbol{\lambda}'\boldsymbol{\beta}$ of the elements of the parametric vector $\boldsymbol{\beta}$ is the best linear unbiased estimator in the sense described in the following theorem.

Theorem 5.5.1 (Gauss–Markov theorem)**.** Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M, Aitken, or general linear model, and suppose that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is an estimable linear combination of the elements of the parametric vector $\boldsymbol{\beta}$. Then, the least squares estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is a linear unbiased estimator. Moreover, in the special case where $\mathbf{y}$ follows a G–M model, the variance (and hence the mean squared error) of the least squares estimator is uniformly smaller than that of any other linear unbiased estimator.

Proof. That the least squares estimator is a linear unbiased estimator was established earlier (in the introductory part of the present section). Now, take $c + \mathbf{a}'\mathbf{y}$ to be an arbitrary linear unbiased estimator of the estimable linear combination $\boldsymbol{\lambda}'\boldsymbol{\beta}$, in which case

$$c = 0 \quad \text{and} \quad \mathbf{X}'\mathbf{a} = \boldsymbol{\lambda}$$

(as noted earlier). And recall that the least squares estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is expressible in the form $(\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y}$, where $\tilde{\mathbf{r}}$ is any solution to the conjugate normal equations $\mathbf{X}'\mathbf{X}\mathbf{r} = \boldsymbol{\lambda}$.

In the special case where $\mathbf{y}$ follows a G–M model, we find that

$$\begin{aligned}\operatorname{cov}[(\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y},\ c + \mathbf{a}'\mathbf{y} - (\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y}] &= \tilde{\mathbf{r}}'\mathbf{X}'(\sigma^2\mathbf{I})(\mathbf{a} - \mathbf{X}\tilde{\mathbf{r}}) \\ &= \sigma^2\tilde{\mathbf{r}}'(\mathbf{X}'\mathbf{a} - \mathbf{X}'\mathbf{X}\tilde{\mathbf{r}}) \\ &= \sigma^2\tilde{\mathbf{r}}'(\boldsymbol{\lambda} - \boldsymbol{\lambda}) = 0.\end{aligned}$$

Thus, in that special case,

$$\begin{aligned}\operatorname{var}(c+\mathbf{a}'\mathbf{y}) &= \operatorname{var}[(\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y} + c + \mathbf{a}'\mathbf{y} - (\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y}] \\ &= \operatorname{var}[(\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y}] + \operatorname{var}[c + \mathbf{a}'\mathbf{y} - (\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y}] \\ &\qquad + 2\operatorname{cov}[(\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y},\ c + \mathbf{a}'\mathbf{y} - (\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y}] \\ &= \operatorname{var}[(\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y}] + \operatorname{var}[c + \mathbf{a}'\mathbf{y} - (\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y}] \\ &\geq \operatorname{var}[(\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y}], \end{aligned} \tag{5.2}$$

with equality holding if and only if $\operatorname{var}[c + \mathbf{a}'\mathbf{y} - (\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y}] = 0$. Moreover, in the special case of the G–M model,

$$\begin{aligned}\operatorname{var}[c + \mathbf{a}'\mathbf{y} - (\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y}] &= (\mathbf{a} - \mathbf{X}\tilde{\mathbf{r}})'(\sigma^2\mathbf{I})(\mathbf{a} - \mathbf{X}\tilde{\mathbf{r}}) \\ &= \sigma^2(\mathbf{a} - \mathbf{X}\tilde{\mathbf{r}})'(\mathbf{a} - \mathbf{X}\tilde{\mathbf{r}}),\end{aligned}$$

so that equality holds in inequality (5.2) if and only if $\mathbf{a} - \mathbf{X}\tilde{\mathbf{r}} = \mathbf{0}$ or, equivalently, if and only if $\mathbf{a} = \mathbf{X}\tilde{\mathbf{r}}$. We conclude that in the special case of the G–M model, the variance of the least squares estimator is uniformly smaller than that of any other linear unbiased estimator. Q.E.D.

Theorem 5.5.1 (in one form or another) has come to be known as the Gauss–Markov theorem (in honor of the contributions of Carl Friedrich Gauss and Andrei Andreevich Markov). It is one of the most famous theoretical results in all of statistics. Seal (1967, sec. 3) considered this result from a historical perspective. That Gauss's name has come to be attached to the result of Theorem 5.5.1 seems altogether appropriate. The case for the attachment of Markov's name appears to be much weaker.

It is customary (both in the present setting and in general) to refer to a linear unbiased estimator that has minimum variance among all linear unbiased estimators as a BLUE (an acronym for best linear unbiased estimator or estimation). If $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M model, then (according to the Gauss–Markov theorem) the least squares estimator of an estimable linear combination $\boldsymbol{\lambda}'\boldsymbol{\beta}$ of the elements of the parametric vector $\boldsymbol{\beta}$ is the unique BLUE of $\boldsymbol{\lambda}'\boldsymbol{\beta}$. Albert (1972, sec. 6.1), in a comment he characterized as jocular, suggested that the least squares estimator of an estimable linear combination could be referred to as a TRUE (an acronym for tiniest residual unbiased estimator). Accordingly, when the least squares estimator is a BLUE, it could be referred to as a TRUE-BLUE—someone who is unswervingly loyal or faithful is said to be true-blue.

b. A corollary

Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M, Aitken, or general linear model, and take $\boldsymbol{\lambda}'\boldsymbol{\beta}$ to be an estimable linear combination of the elements of the parametric vector $\boldsymbol{\beta}$. Further, let $\ell(\mathbf{y}) = (\mathbf{X}\tilde{\mathbf{r}})'\mathbf{y}$, where $\tilde{\mathbf{r}}$ is any solution to the conjugate normal equations $\mathbf{X}'\mathbf{X}\mathbf{r} = \boldsymbol{\lambda}$ [so that $\ell(\mathbf{y})$ is the least squares estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$]. And let $c + \mathbf{a}'\mathbf{y}$ represent an arbitrary linear translation-equivariant estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ or, equivalently, any estimator of the form $c + \mathbf{a}'\mathbf{y}$ that satisfies the condition $\mathbf{a}'\mathbf{X} = \boldsymbol{\lambda}'$; and recall (from the introductory part of the present section) that the least squares estimator is a linear translation-equivariant estimator.

Clearly, $\mathrm{E}(\mathbf{a}'\mathbf{y}) = \boldsymbol{\lambda}'\boldsymbol{\beta}$, that is, $\mathbf{a}'\mathbf{y}$ is an unbiased estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$. And, as a consequence, the MSE (mean squared error) of $c + \mathbf{a}'\mathbf{y}$ is

$$\begin{aligned}\mathrm{E}[(c + \mathbf{a}'\mathbf{y} - \boldsymbol{\lambda}'\boldsymbol{\beta})^2] &= c^2 + \mathrm{E}[(\mathbf{a}'\mathbf{y} - \boldsymbol{\lambda}'\boldsymbol{\beta})^2] + 2c\,\mathrm{E}(\mathbf{a}'\mathbf{y} - \boldsymbol{\lambda}'\boldsymbol{\beta}) \\ &= c^2 + \operatorname{var}(\mathbf{a}'\mathbf{y}) \\ &\geq \operatorname{var}(\mathbf{a}'\mathbf{y}),\end{aligned}$$

with equality holding if and only if $c = 0$ and hence if and only if $c + \mathbf{a}'\mathbf{y} = \mathbf{a}'\mathbf{y}$. Moreover, in the special case where $\mathbf{y}$ follows a G–M model, it follows from the Gauss–Markov theorem that

$$\text{var}(\mathbf{a}'\mathbf{y}) \geq \text{var}[\ell(\mathbf{y})],$$

with equality holding if and only if $\mathbf{a}'\mathbf{y} = \ell(\mathbf{y})$, that is, if and only if $\mathbf{a}'\mathbf{y}$ is the least squares estimator. Accordingly, in that special case, the MSE of the least squares estimator is uniformly smaller than the MSE of any other linear translation-equivariant estimator.

In summary, we have the following result, the main part of which can be regarded as a corollary of the Gauss–Markov theorem.

Corollary 5.5.2. Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M, Aitken, or general linear model, and suppose that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is an estimable linear combination of the elements of the parametric vector $\boldsymbol{\beta}$. Then, the least squares estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is a linear translation-equivariant estimator. Moreover, in the special case where $\mathbf{y}$ follows a G–M model, the mean squared error of the least squares estimator is uniformly smaller than that of any other linear translation-equivariant estimator.

5.6 Simultaneous Estimation

Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M, Aitken, or general linear model. And consider the estimation of estimable linear combinations of the elements of the parametric vector $\boldsymbol{\beta}$. Specifically, suppose that we wish to estimate a finite number M of such linear combinations, say $\boldsymbol{\lambda}_1'\boldsymbol{\beta}, \boldsymbol{\lambda}_2'\boldsymbol{\beta}, \ldots, \boldsymbol{\lambda}_M'\boldsymbol{\beta}$ (and perhaps some or all linear combinations of $\boldsymbol{\lambda}_1'\boldsymbol{\beta}, \boldsymbol{\lambda}_2'\boldsymbol{\beta}, \ldots, \boldsymbol{\lambda}_M'\boldsymbol{\beta}$). The Gauss–Markov theorem is relevant to the estimation of these linear combinations when the linear combinations are considered individually. However, that each of these linear combinations is to be estimated simultaneously with $M-1$ or more other linear combinations is not reflected in the criterion employed in the Gauss–Markov theorem. In this section, we obtain some results that account explicitly for the simultaneous estimation of the various linear combinations.

Let $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_M)$, so that $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is the M-dimensional column vector whose elements are the M linear combinations $\boldsymbol{\lambda}_1'\boldsymbol{\beta}, \boldsymbol{\lambda}_2'\boldsymbol{\beta}, \ldots, \boldsymbol{\lambda}_M'\boldsymbol{\beta}$—when all M linear combinations $\boldsymbol{\lambda}_1'\boldsymbol{\beta}, \boldsymbol{\lambda}_2'\boldsymbol{\beta}, \ldots, \boldsymbol{\lambda}_M'\boldsymbol{\beta}$ are estimable (as is being assumed), the vector $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is said to be estimable. By definition,the least squares estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is the $M \times 1$ vector $\boldsymbol{\ell}(\mathbf{y}) = [\ell_1(\mathbf{y}), \ell_2(\mathbf{y}), \ldots, \ell_M(\mathbf{y})]'$, where $\ell_1(\mathbf{y}), \ell_2(\mathbf{y}), \ldots, \ell_M(\mathbf{y})$ are the least squares estimators of $\boldsymbol{\lambda}_1'\boldsymbol{\beta}, \boldsymbol{\lambda}_2'\boldsymbol{\beta}, \ldots, \boldsymbol{\lambda}_M'\boldsymbol{\beta}$, respectively. In light of result (4.24), the least squares estimator is expressible as

$$\boldsymbol{\ell}(\mathbf{y}) = \boldsymbol{\Lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}. \tag{6.1}$$

And, in light of result (4.35), it is also expressible as

$$\boldsymbol{\ell}(\mathbf{y}) = \tilde{\mathbf{R}}'\mathbf{X}'\mathbf{y}, \tag{6.2}$$

where $\tilde{\mathbf{R}}$ is any solution to the linear system $\mathbf{X}'\mathbf{X}\mathbf{R} = \boldsymbol{\Lambda}$ (in the $P \times M$ matrix $\mathbf{R}$).

We have that

$$\text{E}[\boldsymbol{\ell}(\mathbf{y})] = \boldsymbol{\Lambda}'\boldsymbol{\beta}, \tag{6.3}$$

as is evident upon taking the expected value of expression (6.2) or, alternatively, upon using result (4.36) to establish that each element of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ equals the corresponding element of $\text{E}[\boldsymbol{\ell}(\mathbf{y})]$.

The least squares estimators $\ell_1(\mathbf{y}), \ell_2(\mathbf{y}), \ldots, \ell_M(\mathbf{y})$ of $\boldsymbol{\lambda}_1'\boldsymbol{\beta}, \boldsymbol{\lambda}_2'\boldsymbol{\beta}, \ldots, \boldsymbol{\lambda}_M'\boldsymbol{\beta}$ have the following basic property: the least squares estimator of $\sum_{j=1}^{M} k_j\boldsymbol{\lambda}_j'\boldsymbol{\beta}$ (where $k_1, k_2, \ldots, k_M$ are arbitrary constants) is $\sum_{j=1}^{M} k_j\ell_j(\mathbf{y})$—recall (from Section 5.3) that linear combinations of estimable functions are estimable. Upon letting $\mathbf{k} = (k_1, k_2, \ldots, k_M)'$, this property can be restated (in matrix notation)

as follows: the least squares estimator of $\mathbf{k}'\boldsymbol{\Lambda}'\boldsymbol{\beta}$ [$= (\boldsymbol{\Lambda}\mathbf{k})'\boldsymbol{\beta}$] is $\mathbf{k}'\boldsymbol{\ell}(\mathbf{y})$. In light of results (4.24) and (6.1), this property can be readily verified by observing that the least squares estimator of $(\boldsymbol{\Lambda}\mathbf{k})'\boldsymbol{\beta}$ is

$$(\boldsymbol{\Lambda}\mathbf{k})'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} = \mathbf{k}'\boldsymbol{\Lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} = \mathbf{k}'\boldsymbol{\ell}(\mathbf{y}).$$

Alternatively, in light of results (4.35) and (6.2), it can be verified by observing that (for any solution $\tilde{\mathbf{R}}$ to $\mathbf{X}'\mathbf{X}\mathbf{R} = \boldsymbol{\Lambda}$) $\tilde{\mathbf{R}}\mathbf{k}$ is a solution to the linear system $\mathbf{X}'\mathbf{X}\mathbf{r} = \boldsymbol{\Lambda}\mathbf{k}$ and hence that the least squares estimator of $(\boldsymbol{\Lambda}\mathbf{k})'\boldsymbol{\beta}$ is

$$(\tilde{\mathbf{R}}\mathbf{k})'\mathbf{X}'\mathbf{y} = \mathbf{k}'\tilde{\mathbf{R}}'\mathbf{X}'\mathbf{y} = \mathbf{k}'\boldsymbol{\ell}(\mathbf{y}).$$

Under the general linear model, the variance-covariance matrix of the least squares estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is expressible as

$$\text{var}[\boldsymbol{\ell}(\mathbf{y})] = \tilde{\mathbf{R}}'\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})\mathbf{X}\tilde{\mathbf{R}} = \boldsymbol{\Lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\Lambda} \tag{6.4}$$

(where $\tilde{\mathbf{R}}$ is any solution to $\mathbf{X}'\mathbf{X}\mathbf{R} = \boldsymbol{\Lambda}$). Result (6.4) can be deduced from result (4.39): start with the expressions for $\text{var}[\ell_i(\mathbf{y}),\ \ell_j(\mathbf{y})]$ obtained by applying result (4.39), and then observe that these expressions are essentially the same as the ijth elements of the expressions for $\text{var}[\boldsymbol{\ell}(\mathbf{y})]$ given by result (6.4) ($i, j = 1, 2, \ldots, M$). In the special case of the Aitken model, result (6.4) "simplifies" to

$$\text{var}[\boldsymbol{\ell}(\mathbf{y})] = \sigma^2\tilde{\mathbf{R}}'\mathbf{X}'\mathbf{H}\mathbf{X}\tilde{\mathbf{R}} = \sigma^2\boldsymbol{\Lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{H}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\Lambda}. \tag{6.5}$$

And in the further special case of the G–M model, we find that

$$\text{var}[\boldsymbol{\ell}(\mathbf{y})] = \sigma^2\tilde{\mathbf{R}}'\mathbf{X}'\mathbf{X}\tilde{\mathbf{R}} = \sigma^2\tilde{\mathbf{R}}'\boldsymbol{\Lambda} = \sigma^2\boldsymbol{\Lambda}'\tilde{\mathbf{R}} \tag{6.6}$$

$$= \sigma^2\tilde{\mathbf{R}}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}\tilde{\mathbf{R}} = \sigma^2\boldsymbol{\Lambda}'(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\Lambda}. \tag{6.7}$$

a. Best linear unbiased estimation

Let us consider further the estimation of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ (based on an $N \times 1$ observable random vector $\mathbf{y}$ that follows a G–M, Aitken, or general linear model). An estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is said to be a linear estimator if each of its elements is a linear estimator (of the corresponding element of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$) or, equivalently, if it is expressible in the form $\mathbf{c} + \mathbf{A}'\mathbf{y}$ (where $\mathbf{c}$ is an $M \times 1$ vector of constants and $\mathbf{A}$ an $N \times M$ matrix of constants). An estimator $\mathbf{t}(\mathbf{y})$ of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is said to be unbiased if each of its elements is an unbiased estimator of the corresponding element of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ or, equivalently, if $\text{E}[\mathbf{t}(\mathbf{y})] = \boldsymbol{\Lambda}'\boldsymbol{\beta}$. It follows from the results of Section 5.1 that for a linear estimator $\mathbf{c} + \mathbf{A}'\mathbf{y}$ of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ to be an unbiased estimator, it is necessary and sufficient that

$$\mathbf{c} = \mathbf{0} \quad \text{and} \quad \mathbf{A}'\mathbf{X} = \boldsymbol{\Lambda}'. \tag{6.8}$$

Moreover, it follows from what was established earlier (e.g., in Section 5.5) that the least squares estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is linear and unbiased.

How does the variance-covariance matrix of the least squares estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ compare with the variance-covariance matrix of other linear unbiased estimators of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$? The Gauss–Markov theorem implies that in the special case where $\mathbf{y}$ follows a G–M model, at least some of the diagonal elements of the variance-covariance matrix of the least squares estimator are (strictly) less than (and the others are equal to) the corresponding diagonal elements of the variance-covariance matrix of any other linear unbiased estimator. The following theorem makes a stronger statement.

Theorem 5.6.1. Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M, Aitken, or general linear model, take $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ to be any $M \times 1$ vector of estimable linear combinations of the elements of the parametric vector $\boldsymbol{\beta}$, denote by $\boldsymbol{\ell}(\mathbf{y})$ the least squares estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$, and let $\mathbf{c} + \mathbf{A}'\mathbf{y}$ represent an arbitrary linear unbiased estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ (or, equivalently, any estimator of the form

$\mathbf{c}+\mathbf{A}'\mathbf{y}$ such that $\mathbf{c}=\mathbf{0}$ and $\mathbf{A}'\mathbf{X}=\boldsymbol{\Lambda}'$). Then, (the least squares estimator) $\boldsymbol{\ell}(\mathbf{y})$ is a linear unbiased estimator. Moreover, in the special case where $\mathbf{y}$ follows a G–M model, $\text{var}(\mathbf{c}+\mathbf{A}'\mathbf{y})-\text{var}[\boldsymbol{\ell}(\mathbf{y})]$ is a nonnegative definite matrix, and $\text{var}(\mathbf{c}+\mathbf{A}'\mathbf{y})-\text{var}[\boldsymbol{\ell}(\mathbf{y})]=\mathbf{0}$ or, equivalently, $\text{var}(\mathbf{c}+\mathbf{A}'\mathbf{y})=\text{var}[\boldsymbol{\ell}(\mathbf{y})]$ if and only if $\mathbf{c}+\mathbf{A}'\mathbf{y}=\boldsymbol{\ell}(\mathbf{y})$.

Proof. That $\boldsymbol{\ell}(\mathbf{y})$ is a linear unbiased estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ follows from what was established in Section 5.5 (as was noted previously). Now, suppose that $\mathbf{y}$ follows a G–M model, and consider the quadratic form

$$\mathbf{k}'\{\text{var}(\mathbf{c}+\mathbf{A}'\mathbf{y})-\text{var}[\boldsymbol{\ell}(\mathbf{y})]\}\mathbf{k} \tag{6.9}$$

(in an M-dimensional column vector $\mathbf{k}$). Clearly, the quadratic form (6.9) is reexpressible as

$$\mathbf{k}'\{\text{var}(\mathbf{c}+\mathbf{A}'\mathbf{y})-\text{var}[\boldsymbol{\ell}(\mathbf{y})]\}\mathbf{k}=\text{var}[\mathbf{k}'\mathbf{c}+(\mathbf{A}\mathbf{k})'\mathbf{y}]-\text{var}[\mathbf{k}'\boldsymbol{\ell}(\mathbf{y})]. \tag{6.10}$$

Moreover, $\mathbf{k}'\mathbf{c}+(\mathbf{A}\mathbf{k})'\mathbf{y}$ is a linear unbiased estimator of $(\boldsymbol{\Lambda}\mathbf{k})'\boldsymbol{\beta}$; the unbiasedness of which can be verified simply by observing that $\text{E}[\mathbf{k}'\mathbf{c}+(\mathbf{A}\mathbf{k})'\mathbf{y}]=\mathbf{k}'\,\text{E}(\mathbf{c}+\mathbf{A}'\mathbf{y})=\mathbf{k}'\boldsymbol{\Lambda}'\boldsymbol{\beta}=(\boldsymbol{\Lambda}\mathbf{k})'\boldsymbol{\beta}$ or, alternatively, by observing [in light of the sufficiency of condition (1.4)] that $\mathbf{k}'\mathbf{c}=\mathbf{k}'\mathbf{0}=0$ and that $(\mathbf{A}\mathbf{k})'\mathbf{X}=\mathbf{k}'\mathbf{A}'\mathbf{X}=\mathbf{k}'\boldsymbol{\Lambda}'=(\boldsymbol{\Lambda}\mathbf{k})'$. And as discussed in the introductory part of the present section, $\mathbf{k}'\boldsymbol{\ell}(\mathbf{y})$ is the least squares estimator of $(\boldsymbol{\Lambda}\mathbf{k})'\boldsymbol{\beta}$. Thus, it follows from the Gauss–Markov theorem that $\text{var}[\mathbf{k}'\boldsymbol{\ell}(\mathbf{y})]\leq\text{var}[\mathbf{k}'\mathbf{c}+(\mathbf{A}\mathbf{k})'\mathbf{y}]$ or, equivalently, that

$$\text{var}[\mathbf{k}'\mathbf{c}+(\mathbf{A}\mathbf{k})'\mathbf{y}]-\text{var}[\mathbf{k}'\boldsymbol{\ell}(\mathbf{y})]\geq 0. \tag{6.11}$$

Together, results (6.10) and (6.11) imply that the quadratic form $\mathbf{k}'\{\text{var}(\mathbf{c}+\mathbf{A}'\mathbf{y})-\text{var}[\boldsymbol{\ell}(\mathbf{y})]\}\mathbf{k}$ is nonnegative definite and hence that the matrix $\text{var}(\mathbf{c}+\mathbf{A}'\mathbf{y})-\text{var}[\boldsymbol{\ell}(\mathbf{y})]$ is nonnegative definite.

As a further implication of the Gauss–Markov theorem, we have that $\text{var}[\mathbf{k}'\mathbf{c}+(\mathbf{A}\mathbf{k})'\mathbf{y}]=\text{var}[\mathbf{k}'\boldsymbol{\ell}(\mathbf{y})]$ or, equivalently, that equality holds in inequality (6.11) if and only if $\mathbf{k}'\mathbf{c}+(\mathbf{A}\mathbf{k})'\mathbf{y}=\mathbf{k}'\boldsymbol{\ell}(\mathbf{y})$, leading [in light of equality (6.10) and Corollary 2.13.4] to the conclusion that $\text{var}(\mathbf{c}+\mathbf{A}'\mathbf{y})-\text{var}[\boldsymbol{\ell}(\mathbf{y})]=\mathbf{0}$ if and only if $\mathbf{k}'\mathbf{c}+(\mathbf{A}\mathbf{k})'\mathbf{y}=\mathbf{k}'\boldsymbol{\ell}(\mathbf{y})$ for every $\mathbf{k}$ and hence if and only if $\mathbf{c}+\mathbf{A}'\mathbf{y}=\boldsymbol{\ell}(\mathbf{y})$. Q.E.D.

Suppose (in connection with Theorem 5.6.1) that $\mathbf{y}$ follows a G–M model, in which case $\text{var}(\mathbf{c}+\mathbf{A}'\mathbf{y})-\text{var}[\boldsymbol{\ell}(\mathbf{y})]$ is nonnegative definite. Then, $\text{var}(\mathbf{c}+\mathbf{A}'\mathbf{y})-\text{var}[\boldsymbol{\ell}(\mathbf{y})]=\mathbf{R}'\mathbf{R}$ for some matrix $\mathbf{R}$ (as is evident from Corollary 2.13.25). And upon recalling Lemma 2.3.2 and observing that $\mathbf{R}'\mathbf{R}$ equals $\mathbf{0}$ if and only if all M of its diagonal elements equal 0 and upon letting (for $j=1,2,\ldots,M$) c_j represent the jth element of $\mathbf{c}$, $\mathbf{a}_j$ the jth column of $\mathbf{A}$, and $\ell_j(\mathbf{y})$ the jth element of $\boldsymbol{\ell}(\mathbf{y})$, it follows that

$$\begin{aligned}\text{var}(\mathbf{c}+\mathbf{A}'\mathbf{y})-\text{var}[\boldsymbol{\ell}(\mathbf{y})]=\mathbf{0}&\\ \Leftrightarrow\ \text{var}(c_j+\mathbf{a}_j'\mathbf{y})-\text{var}[\ell_j(\mathbf{y})]=0\ \ (j=1,2,\ldots,M)&\\ \Leftrightarrow\ \text{tr}\{\text{var}(\mathbf{c}+\mathbf{A}'\mathbf{y})-\text{var}[\boldsymbol{\ell}(\mathbf{y})]\}=0&\end{aligned}$$

or, equivalently, that

$$\begin{aligned}\text{var}(\mathbf{c}+\mathbf{A}'\mathbf{y})=\text{var}[\boldsymbol{\ell}(\mathbf{y})]\ \ &\Leftrightarrow\ \ \text{var}(c_j+\mathbf{a}_j'\mathbf{y})=\text{var}[\ell_j(\mathbf{y})]\ \ (j=1,2,\ldots,M)\\ &\Leftrightarrow\ \ \text{tr}[\text{var}(\mathbf{c}+\mathbf{A}'\mathbf{y})]=\text{tr}\{\text{var}[\boldsymbol{\ell}(\mathbf{y})]\}.\end{aligned}$$

Because the diagonal elements of a nonnegative definite matrix are inherently nonnegative (as evidenced by Corollary 2.13.14), the following result can be regarded as a corollary of Theorem 5.6.1.

Corollary 5.6.2. Suppose that $\mathbf{y}$ is an $N\times 1$ observable random vector that follows a G–M model, take $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ to be any $M\times 1$ vector of estimable linear combinations of the elements of the parametric vector $\boldsymbol{\beta}$, denote by $\boldsymbol{\ell}(\mathbf{y})$ the least squares estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$, and let $\mathbf{c}+\mathbf{A}'\mathbf{y}$ represent an arbitrary linear unbiased estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$. Then,

$$\text{tr}\{\text{var}[\boldsymbol{\ell}(\mathbf{y})]\}\leq\text{tr}[\text{var}(\mathbf{c}+\mathbf{A}'\mathbf{y})],$$

with equality holding if and only if $\mathbf{c} + \mathbf{A}'\mathbf{y} = \boldsymbol{\ell}(\mathbf{y})$.

Alternatively, Corollary 5.6.2 can be established as an almost immediate consequence of the Gauss–Markov theorem. A more substantial implication of Theorem 5.6.1 is provided by the following corollary.

Corollary 5.6.3. Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M model, take $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ to be any $M \times 1$ vector of estimable linear combinations of the elements of the parametric vector $\boldsymbol{\beta}$, denote by $\boldsymbol{\ell}(\mathbf{y})$ the least squares estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$, and let $\mathbf{c} + \mathbf{A}'\mathbf{y}$ represent an arbitrary linear unbiased estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$. Then,

$$\det\{\text{var}[\boldsymbol{\ell}(\mathbf{y})]\} \leq \det[\text{var}(\mathbf{c} + \mathbf{A}'\mathbf{y})],$$

with equality holding if and only if $\text{rank}(\mathbf{A}) < M$ (in which case both sides of the inequality equal 0) or $\mathbf{c} + \mathbf{A}'\mathbf{y} = \boldsymbol{\ell}(\mathbf{y})$.

Corollary 5.6.3 can be derived from Theorem 5.6.1 by applying the following result on determinants: for any $M \times M$ symmetric nonnegative definite matrix $\mathbf{B}$ and for any $M \times M$ symmetric matrix $\mathbf{C}$ such that $\mathbf{C} - \mathbf{B}$ is nonnegative definite, $|\mathbf{C}| \geq |\mathbf{B}|$, with equality holding if and only if $\mathbf{C}$ is singular or $\mathbf{C} = \mathbf{B}$—for a proof of this result, refer, e.g., to Harville (1997, corollary 18.1.8). Specifically, the application is that obtained by setting $\mathbf{B} = \text{var}[\boldsymbol{\ell}(\mathbf{y})]$ and $\mathbf{C} = \text{var}(\mathbf{c} + \mathbf{A}'\mathbf{y})$.

The determinant of the variance-covariance matrix of a vector-valued estimator is sometimes referred to as the generalized variance of the estimator. Accordingly, the result of Corollary 5.6.3 implies that (under the G–M model) the least squares estimator of the M-dimensional vector $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is a best linear unbiased estimator in the sense that its generalized variance is less than or equal to that of any other linear unbiased estimator—if $\text{rank}(\boldsymbol{\Lambda}) = M$, the generalized variance of the least squares estimator is (strictly) less than that of any other linear unbiased estimator.

b. Best linear translation-equivariant estimation

Let us continue to consider the estimation of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ (based on an $N \times 1$ observable random vector $\mathbf{y}$ that follows a G–M, Aitken, or general linear model). An estimator, say $\mathbf{t}(\mathbf{y})$, of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is said to be translation equivariant if the elements of $\mathbf{t}(\mathbf{y})$ are translation-equivariant estimators of the corresponding elements of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$. Thus, in light of the discussion of Section 5.2, a necessary and sufficient condition for $\mathbf{t}(\mathbf{y})$ to be a translation-equivariant estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is that

$$\mathbf{t}(\mathbf{y}) + \boldsymbol{\Lambda}'\mathbf{k} = \mathbf{t}(\mathbf{y} + \mathbf{X}\mathbf{k}) \tag{6.12}$$

for every $P \times 1$ vector $\mathbf{k}$ (and for every value of $\mathbf{y}$). Further, for a linear estimator $\mathbf{c} + \mathbf{A}'\mathbf{y}$ to be a translation-equivariant estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$, it is necessary and sufficient that

$$\mathbf{A}'\mathbf{X} = \boldsymbol{\Lambda}'. \tag{6.13}$$

Moreover, it follows from what was established earlier (in Section 5.5) that the least squares estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is translation equivariant.

As what can be regarded as an additional corollary of Theorem 5.6.1, we have the following result.

Corollary 5.6.4. Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M, Aitken, or general linear model, take $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ to be any $M \times 1$ vector of estimable linear combinations of the elements of the parametric vector $\boldsymbol{\beta}$, denote by $\boldsymbol{\ell}(\mathbf{y})$ the least squares estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$, and let $\mathbf{c} + \mathbf{A}'\mathbf{y}$ represent an arbitrary linear translation-equivariant estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ (or, equivalently, any estimator of the form $\mathbf{c} + \mathbf{A}'\mathbf{y}$ such that $\mathbf{A}'\mathbf{X} = \boldsymbol{\Lambda}'$). Then, (the least squares estimator) $\boldsymbol{\ell}(\mathbf{y})$ is a linear translation-equivariant estimator. Moreover, in the special case where $\mathbf{y}$ follows a G–M model, the difference

$$\text{E}[(\mathbf{c}+\mathbf{A}'\mathbf{y}-\boldsymbol{\Lambda}'\boldsymbol{\beta})(\mathbf{c}+\mathbf{A}'\mathbf{y}-\boldsymbol{\Lambda}'\boldsymbol{\beta})'] - \text{E}\{[\boldsymbol{\ell}(\mathbf{y})-\boldsymbol{\Lambda}'\boldsymbol{\beta}][\boldsymbol{\ell}(\mathbf{y})-\boldsymbol{\Lambda}'\boldsymbol{\beta}]'\} \tag{6.14}$$

between the mean-squared-error matrices of $\mathbf{c} + \mathbf{A}'\mathbf{y}$ and $\boldsymbol{\ell}(\mathbf{y})$ is a nonnegative definite matrix, and this difference equals $\mathbf{0}$ if and only if $\mathbf{c} + \mathbf{A}'\mathbf{y} = \boldsymbol{\ell}(\mathbf{y})$.

The nonnegative definiteness of the matrix (6.14) and the condition $[\mathbf{c} + \mathbf{A}'\mathbf{y} = \boldsymbol{\ell}(\mathbf{y})]$ under which it equals $\mathbf{0}$ follow from Theorem 5.6.1 in much the same way that the main part of Corollary 5.5.2 follows from the Gauss–Markov theorem.

5.7 Estimation of Variability and Covariability

Suppose that $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$ is an $N \times 1$ observable random vector that follows a G–M, Aitken, or general linear model. Then, the N diagonal elements $\text{var}(y_1), \text{var}(y_2), \ldots, \text{var}(y_N)$ of the matrix $\text{var}(\mathbf{y})$ represent the underlying variability and the $N(N-1)$ off-diagonal elements $\text{cov}(y_i, y_j)$ $(j \neq i = 1, 2, \ldots, N)$ represent the underlying covariability. In the case of the G–M model, $\text{var}(\mathbf{y}) = \sigma^2\mathbf{I}$, so that $y_1, y_2, \ldots, y_N$ are uncorrelated with a common (strictly) positive variance σ^2 (of unknown value). In the case of the Aitken model, $\text{var}(\mathbf{y}) = \sigma^2\mathbf{H}$ (where $\mathbf{H}$ is a known symmetric nonnegative definite matrix), so that the variances and covariances of $y_1, y_2, \ldots, y_N$ are known up to the (unknown) value of a (strictly) positive scalar multiple σ^2. And, in the case of the general linear model, $\text{var}(\mathbf{y}) = \mathbf{V}(\boldsymbol{\theta})$ [where $\mathbf{V}(\boldsymbol{\theta})$ is a symmetric nonnegative definite matrix whose elements are known functions of a $T \times 1$ parameter vector $\boldsymbol{\theta}$], so that the variances and covariances of $y_1, y_2, \ldots, y_N$ are known up to the (unknown) value of $\boldsymbol{\theta}$.

The matrix $\text{var}(\mathbf{y})$ is of interest because its value determines the variances and covariances of the least squares estimators of estimable linear combinations of the elements of the parametric vector $\boldsymbol{\beta}$. Moreover, the underlying variability and covariability [represented by the diagonal and off-diagonal elements of $\text{var}(\mathbf{y})$] may be of interest in their own right.

In the present section, some initial results are obtained on the estimation of variability and covariability. The emphasis is on results that are specific to the estimation of σ^2 under the G–M model. As a preliminary, formulas are derived for the expected values and variances and covariances of quadratic forms (in a random vector) and for the covariance of a quadratic form and a linear form. And, prior to that, some matrix operations that enter in various of those formulas are introduced and briefly discussed.

a. Some matrix operations

Vec of a matrix. Let $\mathbf{A}$ represent an $M \times N$ matrix. It is sometimes convenient to rearrange the elements of $\mathbf{A}$ in the form of an MN-dimensional column vector. The conventional way of doing so is to successively place the first, second, ..., Nth columns $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N$ of $\mathbf{A}$ one under the other, giving the column vector

$$\begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_N \end{pmatrix}. \tag{7.1}$$

The vector (7.1) is referred to as the *vec* of $\mathbf{A}$, and is denoted by the symbol $\text{vec}(\mathbf{A})$ or (when the parentheses are not needed for clarity) by $\text{vec}\,\mathbf{A}$. By definition, the ijth element of $\mathbf{A}$ is the $[(j-1)M + i]$th element of $\text{vec}\,\mathbf{A}$.

Vech of a symmetric matrix. Let $\mathbf{A} = \{a_{ij}\}$ represent an $N \times N$ symmetric matrix. The values of all N^2 elements of $\mathbf{A}$ can be determined from the values of those $N(N+1)/2$ elements that are on or below the diagonal [or, alternatively, from those $N(N+1)/2$ elements that are on or above the

diagonal]. Accordingly, in rearranging the elements of $\mathbf{A}$ in the form of a vector (as in forming the vec), we may wish to exclude the $N(N-1)/2$ "duplicate" elements. Thus, as an alternative to the vec of $\mathbf{A}$, we may wish to consider the $N(N+1)/2$-dimensional column vector

$$\begin{pmatrix} \mathbf{a}_1^* \\ \mathbf{a}_2^* \\ \vdots \\ \mathbf{a}_N^* \end{pmatrix}, \tag{7.2}$$

where (for $i = 1, 2, \ldots, N$) $\mathbf{a}_i^* = (a_{ii}, a_{i+1,i}, \ldots, a_{Ni})'$ is the subvector of the ith column of $\mathbf{A}$ obtained by striking out its first $i-1$ elements. The vector (7.2) is referred to as the *vech* of $\mathbf{A}$ and is denoted by the symbol vech($\mathbf{A}$) or vech $\mathbf{A}$. For $N = 1$, $N = 2$, and $N = 3$,

$$\text{vech}\,\mathbf{A} = (a_{11}),\ \text{vech}\,\mathbf{A} = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{22} \end{pmatrix},\ \text{and vech}\,\mathbf{A} = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \\ a_{22} \\ a_{32} \\ a_{33} \end{pmatrix},\ \text{respectively.}$$

Every element of $\mathbf{A}$, and hence every element of vec $\mathbf{A}$, is either an element of vech $\mathbf{A}$ or a "duplicate" of an element of vech $\mathbf{A}$. Thus, there exists a unique $[N^2 \times N(N+1)/2]$-dimensional matrix, to be denoted by the symbol $\mathbf{G}_N$, such that (for every $N \times N$ symmetric matrix $\mathbf{A}$)

$$\text{vec}\,\mathbf{A} = \mathbf{G}_N\,\text{vech}\,\mathbf{A}.$$

This matrix is called the *duplication matrix*. Clearly,

$$\mathbf{G}_1 = (1),\ \mathbf{G}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},\ \text{and}\ \mathbf{G}_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Note that

$$\text{rank}\,\mathbf{G}_N = N(N+1)/2 \tag{7.3}$$

(so that $\mathbf{G}_N$ is of full column rank), as is evident upon observing that every row of $\mathbf{I}_{N(N+1)/2}$ is a row of $\mathbf{G}_N$ and hence that $\mathbf{G}_N$ contains $N(N+1)/2$ linearly independent rows.

Kronecker product. The *Kronecker product* of two matrices, say an $M \times N$ matrix $\mathbf{A} = \{a_{ij}\}$ and a $P \times Q$ matrix $\mathbf{B} = \{b_{ij}\}$, is denoted by the symbol $\mathbf{A} \otimes \mathbf{B}$ and is defined to be the $MP \times NQ$ matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \ldots & a_{1N}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \ldots & a_{2N}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{M1}\mathbf{B} & a_{M2}\mathbf{B} & \ldots & a_{MN}\mathbf{B} \end{pmatrix}$$

obtained by replacing (for $i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, N$) the ij element of $\mathbf{A}$ with the $P \times Q$ matrix $a_{ij}\mathbf{B}$. Thus, the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ can be regarded as a partitioned matrix, comprising M rows and N columns of $(P \times Q)$-dimensional blocks, the ijth of which is $a_{ij}\mathbf{B}$.

Among the various properties of the Kronecker product operation is the following: for any matrices $\mathbf{A}$ and $\mathbf{B}$,

$$(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}' \tag{7.4}$$

—for a verification of equality (7.4), refer, e.g., to Harville (1997, sec. 16.1).

Two formulas. There are two formulas that will be convenient to have at our disposal; one of these is for the vec of a product of three matrices and the other is for the trace of the product of four matrices. The two formulas are as follows. For any $M \times N$ matrix $\mathbf{A}$, $N \times P$ matrix $\mathbf{B}$, and $P \times Q$ matrix $\mathbf{C}$,

$$\text{vec}\,\mathbf{ABC} = (\mathbf{C}' \otimes \mathbf{A})\,\text{vec}\,\mathbf{B}. \tag{7.5}$$

And for any $M \times N$ matrix $\mathbf{A}$, $M \times P$ matrix $\mathbf{B}$, $P \times Q$ matrix $\mathbf{C}$, and $N \times Q$ matrix $\mathbf{D}$,

$$\text{tr}(\mathbf{A}'\mathbf{BCD}') = (\text{vec}\,\mathbf{A})'(\mathbf{D} \otimes \mathbf{B})\,\text{vec}\,\mathbf{C}. \tag{7.6}$$

For a derivation of formulas (7.5) and (7.6), refer, for example, to Harville (1997, sec. 16.2).

b. Expected values and variances of quadratic forms (and their covariances with each other and with linear forms)

Suppose that $\mathbf{x}$ is an N-dimensional random column vector. Then, it is customary to refer to a linear combination, say $\mathbf{a}'\mathbf{x}$, of the elements of $\mathbf{x}$ (where $\mathbf{a}$ is an $N \times 1$ vector of constants) as a linear form (in $\mathbf{x}$).

Formulas for the expected values and the variances and covariances of linear forms are available from the results of Sections 3.1 and 3.2. If the random vector $\mathbf{x}$ has a mean vector $\boldsymbol{\mu}$, then the expected value of a linear form $\mathbf{a}'\mathbf{x}$ (in $\mathbf{x}$) is expressible as

$$\text{E}(\mathbf{a}'\mathbf{x}) = \mathbf{a}'\boldsymbol{\mu}. \tag{7.7}$$

And if, in addition, $\mathbf{x}$ has a variance-covariance matrix $\boldsymbol{\Sigma}$, then the variance of $\mathbf{a}'\mathbf{x}$ is expressible as

$$\text{var}(\mathbf{a}'\mathbf{x}) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}, \tag{7.8}$$

and, more generally, the covariance of $\mathbf{a}'\mathbf{x}$ and $\mathbf{b}'\mathbf{x}$ (where $\mathbf{b}'\mathbf{x}$ is a second linear form in $\mathbf{x}$) is expressible as

$$\text{cov}(\mathbf{a}'\mathbf{x},\, \mathbf{b}'\mathbf{x}) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{b}. \tag{7.9}$$

In what follows, these results are extended by obtaining formulas for the expected values and variances of quadratic forms (in a random column vector) and formulas for the covariances of the quadratic forms with each other and with linear forms.

Main results. The main results are presented in a series of three theorems.

Theorem 5.7.1. Let $\mathbf{x}$ represent an N-dimensional random column vector having mean vector $\boldsymbol{\mu} = \{\mu_i\}$ and variance-covariance matrix $\boldsymbol{\Sigma} = \{\sigma_{ij}\}$, and take $\mathbf{A} = \{a_{ij}\}$ to be an $N \times N$ matrix of constants. Then,

$$\text{E}(\mathbf{x}'\mathbf{Ax}) = \sum_{i,j} a_{ij}(\sigma_{ij} + \mu_i\mu_j) \tag{7.10}$$

$$= \text{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}. \tag{7.11}$$

Proof. Letting x_i represent the ith element of $\mathbf{x}$ $(i = 1, 2, \ldots, N)$, we find that

$$\begin{aligned}\mathrm{E}(\mathbf{x}'\mathbf{A}\mathbf{x}) = \mathrm{E}\Big(\sum_{i,j} a_{ij}x_i x_j\Big) &= \sum_{i,j} a_{ij}\mathrm{E}(x_i x_j)\\ &= \sum_{i,j} a_{ij}(\sigma_{ij} + \mu_i\mu_j)\\ &= \sum_i \Big(\sum_j a_{ij}\sigma_{ji}\Big) + \sum_{i,j} a_{ij}\mu_i\mu_j\\ &= \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}.\end{aligned}$$

Q.E.D.

Theorem 5.7.2. Let $\mathbf{x}$ represent an N-dimensional random column vector having mean vector $\boldsymbol{\mu} = \{\mu_i\}$, variance-covariance matrix $\boldsymbol{\Sigma} = \{\sigma_{ij}\}$, and third central moments $\lambda_{ijk} = \mathrm{E}[(x_i - \mu_i)(x_j - \mu_j)(x_k - \mu_k)]$ $(i, j, k = 1, 2, \ldots, N)$, and take $\mathbf{b} = \{b_i\}$ to be an N-dimensional column vector of constants and $\mathbf{A} = \{a_{ij}\}$ to be an $N \times N$ symmetric matrix of constants. Then,

$$\mathrm{cov}(\mathbf{b}'\mathbf{x},\, \mathbf{x}'\mathbf{A}\mathbf{x}) = \sum_{i,j,k} b_i a_{jk}(\lambda_{ijk} + 2\mu_j\sigma_{ik}) \tag{7.12}$$

$$= \mathbf{b}'\boldsymbol{\Lambda}\,\mathrm{vec}\,\mathbf{A} + 2\mathbf{b}'\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}, \tag{7.13}$$

where $\boldsymbol{\Lambda}$ is an $N \times N^2$ matrix whose entry for the ith row and jkth column [i.e., column $(j-1)N+k$] is λ_{ijk}.

Proof. Letting $\mathbf{z} = \{z_i\} = \mathbf{x} - \boldsymbol{\mu}$ (in which case $\mathbf{x} = \mathbf{z} + \boldsymbol{\mu}$) and using Theorem 5.7.1 [and observing that $\mathbf{z}'\mathbf{A}\boldsymbol{\mu} = (\mathbf{z}'\mathbf{A}\boldsymbol{\mu})' = \boldsymbol{\mu}'\mathbf{A}\mathbf{z}$ and that $\mathbf{b}'\mathbf{z} = \mathbf{z}'\mathbf{b}$], we find that

$$\begin{aligned}\mathrm{cov}(\mathbf{b}'\mathbf{x},\, \mathbf{x}'\mathbf{A}\mathbf{x}) &= \mathrm{cov}(\mathbf{b}'\mathbf{z},\, \mathbf{x}'\mathbf{A}\mathbf{x}) = \mathrm{E}[(\mathbf{b}'\mathbf{z})\mathbf{x}'\mathbf{A}\mathbf{x}]\\ &= \mathrm{E}[(\mathbf{b}'\mathbf{z})(\mathbf{z}'\mathbf{A}\mathbf{z} + 2\boldsymbol{\mu}'\mathbf{A}\mathbf{z} + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu})]\\ &= \mathrm{E}\Big[\Big(\sum_i b_i z_i\Big)\Big(\sum_{j,k} a_{jk}z_j z_k\Big)\Big] + 2\,\mathrm{E}[\mathbf{z}'\mathbf{b}\boldsymbol{\mu}'\mathbf{A}\mathbf{z}] + 0\\ &= \mathrm{E}\Big(\sum_{i,j,k} b_i a_{jk} z_i z_j z_k\Big) + 2\sum_{i,k} b_i\Big(\sum_j \mu_j a_{jk}\Big)\sigma_{ik}\\ &= \sum_{i,j,k} b_i a_{jk}(\lambda_{ijk} + 2\mu_j\sigma_{ik})\\ &= \sum_{j,k}\Big(\sum_i b_i\lambda_{ijk}\Big)a_{kj} + 2\sum_k\Big(\sum_i b_i\sigma_{ik}\Big)\Big(\sum_j a_{kj}\mu_j\Big)\\ &= \mathbf{b}'\boldsymbol{\Lambda}\,\mathrm{vec}\,\mathbf{A} + 2\mathbf{b}'\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}.\end{aligned}$$

Q.E.D.

If the distribution of $\mathbf{x}$ is MVN, then $\boldsymbol{\Lambda} = \mathbf{0}$ (as is evident from the results of Section 3.5n). More generally, if the distribution of $\mathbf{x}$ is symmetric [in the sense that $-(\mathbf{x}-\boldsymbol{\mu}) \sim \mathbf{x} - \boldsymbol{\mu}$], then $\boldsymbol{\Lambda} = \mathbf{0}$ [as is evident upon observing that if the distribution of $\mathbf{x}$ is symmetric, then (for all i, j, and k) $-\lambda_{ijk} = \lambda_{ijk}$]. When $\boldsymbol{\Lambda} = \mathbf{0}$, formula (7.13) simplifies to

$$\mathrm{cov}(\mathbf{b}'\mathbf{x},\, \mathbf{x}'\mathbf{A}\mathbf{x}) = 2\mathbf{b}'\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}. \tag{7.14}$$

Thus, if the distribution of $\mathbf{x}$ is symmetric and its mean vector is null, then any linear form in $\mathbf{x}$ is uncorrelated with any quadratic form.

Theorem 5.7.3. Let $\mathbf{x}$ represent an N-dimensional random column vector having mean vector $\boldsymbol{\mu} = \{\mu_i\}$, variance-covariance matrix $\boldsymbol{\Sigma} = \{\sigma_{ij}\}$, third central moments $\lambda_{ijk} = \mathrm{E}[(x_i - \mu_i)$

$(x_j-\mu_j)(x_k-\mu_k)]$ $(i,j,k=1,2,\ldots,N)$, and fourth central moments $\gamma_{ijkm}=\mathrm{E}[(x_i-\mu_i)(x_j-\mu_j)(x_k-\mu_k)(x_m-\mu_m)]$ $(i,j,k,m=1,2,\ldots,N)$; and take $\mathbf{A}=\{a_{ij}\}$ and $\mathbf{H}=\{h_{ij}\}$ to be $N\times N$ symmetric matrices of constants. Then,

$$\begin{aligned}\operatorname{cov}(\mathbf{x}'\mathbf{A}\mathbf{x},\,\mathbf{x}'\mathbf{H}\mathbf{x})&=\sum_{i,j,k,m}a_{ij}h_{km}[(\gamma_{ijkm}-\sigma_{ij}\sigma_{km}-\sigma_{ik}\sigma_{jm}-\sigma_{im}\sigma_{jk})\\&\qquad+2\mu_k\lambda_{ijm}+2\mu_i\lambda_{jkm}+2\sigma_{ik}\sigma_{jm}+4\mu_j\mu_k\sigma_{im}] \qquad (7.15)\\&=(\operatorname{vec}\mathbf{A})'\boldsymbol{\Omega}\operatorname{vec}\mathbf{H}+2\boldsymbol{\mu}'\mathbf{H}\boldsymbol{\Lambda}\operatorname{vec}\mathbf{A}+2\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Lambda}\operatorname{vec}\mathbf{H}\\&\qquad+2\operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\Sigma})+4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\mu}, \qquad (7.16)\end{aligned}$$

where $\boldsymbol{\Omega}$ is an $N^2\times N^2$ matrix whose entry for the ijth row [row $(i-1)N+j$] and kmth column [column $(k-1)N+m$] is $\gamma_{ijkm}-\sigma_{ij}\sigma_{km}-\sigma_{ik}\sigma_{jm}-\sigma_{im}\sigma_{jk}$ and where $\boldsymbol{\Lambda}$ is an $N\times N^2$ matrix whose entry for the jth row and kmth column [column $(k-1)N+m$] is λ_{jkm}.

Proof. Letting $\mathbf{z}=\{z_i\}=\mathbf{x}-\boldsymbol{\mu}$ (in which case $\mathbf{x}=\mathbf{z}+\boldsymbol{\mu}$) and using Theorems 5.7.1 and 5.7.2 [and observing that $\mathbf{z}'\mathbf{A}\boldsymbol{\mu}=(\mathbf{z}'\mathbf{A}\boldsymbol{\mu})'=\boldsymbol{\mu}'\mathbf{A}\mathbf{z}$ and similarly that $\mathbf{z}'\mathbf{H}\boldsymbol{\mu}=\boldsymbol{\mu}'\mathbf{H}\mathbf{z}$), we find that

$$\begin{aligned}&\operatorname{cov}(\mathbf{x}'\mathbf{A}\mathbf{x},\,\mathbf{x}'\mathbf{H}\mathbf{x})\\&\quad=\mathrm{E}[(\mathbf{x}'\mathbf{A}\mathbf{x})(\mathbf{x}'\mathbf{H}\mathbf{x})]-\mathrm{E}(\mathbf{x}'\mathbf{A}\mathbf{x})\,\mathrm{E}(\mathbf{x}'\mathbf{H}\mathbf{x})\\&\quad=\mathrm{E}[(\mathbf{z}+\boldsymbol{\mu})'\mathbf{A}(\mathbf{z}+\boldsymbol{\mu})(\mathbf{z}+\boldsymbol{\mu})'\mathbf{H}(\mathbf{z}+\boldsymbol{\mu})]\\&\qquad\qquad-[\mathrm{E}(\mathbf{z}'\mathbf{A}\mathbf{z})+\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}][\mathrm{E}(\mathbf{z}'\mathbf{H}\mathbf{z})+\boldsymbol{\mu}'\mathbf{H}\boldsymbol{\mu}]\\&\quad=\mathrm{E}[(\mathbf{z}'\mathbf{A}\mathbf{z})(\mathbf{z}'\mathbf{H}\mathbf{z})]+2\,\mathrm{E}[(\mathbf{z}'\mathbf{A}\mathbf{z})(\boldsymbol{\mu}'\mathbf{H}\mathbf{z})]+2\,\mathrm{E}[(\boldsymbol{\mu}'\mathbf{A}\mathbf{z})(\mathbf{z}'\mathbf{H}\mathbf{z})]\\&\qquad+4\,\mathrm{E}[(\mathbf{z}'\mathbf{A}\boldsymbol{\mu})(\boldsymbol{\mu}'\mathbf{H}\mathbf{z})]+2\,\mathrm{E}[(\mathbf{z}'\mathbf{A}\boldsymbol{\mu})(\boldsymbol{\mu}'\mathbf{H}\boldsymbol{\mu})]+2\,\mathrm{E}[(\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu})(\boldsymbol{\mu}'\mathbf{H}\mathbf{z})]\\&\qquad-\mathrm{E}(\mathbf{z}'\mathbf{A}\mathbf{z})\,\mathrm{E}(\mathbf{z}'\mathbf{H}\mathbf{z})\\&\quad=\mathrm{E}[(\mathbf{z}'\mathbf{A}\mathbf{z})(\mathbf{z}'\mathbf{H}\mathbf{z})]+2\operatorname{cov}(\boldsymbol{\mu}'\mathbf{H}\mathbf{z},\,\mathbf{z}'\mathbf{A}\mathbf{z})+2\operatorname{cov}(\boldsymbol{\mu}'\mathbf{A}\mathbf{z},\,\mathbf{z}'\mathbf{H}\mathbf{z})\\&\qquad\qquad+4\,\mathrm{E}(\mathbf{z}'\mathbf{A}\boldsymbol{\mu}\boldsymbol{\mu}'\mathbf{H}\mathbf{z})+0+0-\mathrm{E}(\mathbf{z}'\mathbf{A}\mathbf{z})\,\mathrm{E}(\mathbf{z}'\mathbf{H}\mathbf{z})\\&\quad=\mathrm{E}\Big(\sum_{i,j,k,m}a_{ij}h_{km}z_iz_jz_kz_m\Big)\\&\qquad+2\sum_{i,j,m}a_{ij}\Big(\sum_k\mu_kh_{km}\Big)\lambda_{ijm}+2\sum_{j,k,m}h_{km}\Big(\sum_i\mu_ia_{ij}\Big)\lambda_{jkm}\\&\qquad+4\sum_{i,m}\Big(\sum_j a_{ij}\mu_j\Big)\Big(\sum_k h_{km}\mu_k\Big)\sigma_{im}-\sum_{i,j}a_{ij}\sigma_{ij}\sum_{k,m}h_{km}\sigma_{km}\\&\quad=\sum_{i,j,k,m}a_{ij}h_{km}[\gamma_{ijkm}+2\mu_k\lambda_{ijm}+2\mu_i\lambda_{jkm}+4\mu_j\mu_k\sigma_{im}-\sigma_{ij}\sigma_{km}]\\&\quad=\sum_{i,j,k,m}a_{ij}h_{km}[(\gamma_{ijkm}-\sigma_{ij}\sigma_{km}-\sigma_{ik}\sigma_{jm}-\sigma_{im}\sigma_{jk})\\&\qquad\qquad+2\mu_k\lambda_{ijm}+2\mu_i\lambda_{jkm}+2\sigma_{ik}\sigma_{jm}+4\mu_j\mu_k\sigma_{im}]\\&\quad=\sum_{i,j,k,m}a_{ji}h_{mk}(\gamma_{ijkm}-\sigma_{ij}\sigma_{km}-\sigma_{ik}\sigma_{jm}-\sigma_{im}\sigma_{jk})\\&\qquad+2\sum_{i,j}\Big[\sum_m\Big(\sum_k\mu_kh_{km}\Big)\lambda_{mij}\Big]a_{ji}+2\sum_{k,m}\Big[\sum_j\Big(\sum_i\mu_ia_{ij}\Big)\lambda_{jkm}\Big]h_{mk}\\&\qquad\qquad+2\sum_i\sum_m\Big(\sum_j a_{ij}\sigma_{jm}\Big)\Big(\sum_k h_{mk}\sigma_{ki}\Big)\\&\qquad\qquad+4\sum_{i,m}\Big(\sum_j\mu_ja_{ji}\Big)\Big(\sum_k h_{mk}\mu_k\Big)\sigma_{im}\end{aligned}$$

$$= (\operatorname{vec}\mathbf{A})'\boldsymbol{\Omega}\operatorname{vec}\mathbf{H} + 2\boldsymbol{\mu}'\mathbf{H}\boldsymbol{\Lambda}\operatorname{vec}\mathbf{A} + 2\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Lambda}\operatorname{vec}\mathbf{H} + 2\operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\Sigma}) + 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\mu}.$$

Q.E.D.

As a special case of formula (7.16) (that where $\mathbf{H} = \mathbf{A}$), we have that

$$\operatorname{var}(\mathbf{x}'\mathbf{A}\mathbf{x}) = (\operatorname{vec}\mathbf{A})'\boldsymbol{\Omega}\operatorname{vec}\mathbf{A} + 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Lambda}\operatorname{vec}\mathbf{A} + 2\operatorname{tr}[(\mathbf{A}\boldsymbol{\Sigma})^2] + 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}. \tag{7.17}$$

If the distribution of $\mathbf{x}$ is symmetric, then formula (7.16) simplifies to

$$\operatorname{cov}(\mathbf{x}'\mathbf{A}\mathbf{x},\ \mathbf{x}'\mathbf{H}\mathbf{x}) = (\operatorname{vec}\mathbf{A})'\boldsymbol{\Omega}\operatorname{vec}\mathbf{H} + 2\operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\Sigma}) + 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\mu}. \tag{7.18}$$

If the distribution of $\mathbf{x}$ is MVN, then it is symmetric and, in addition, it is such that $\boldsymbol{\Omega} = \mathbf{0}$ (as is evident from the results of Section 3.5n), in which case there is a further simplification to

$$\operatorname{cov}(\mathbf{x}'\mathbf{A}\mathbf{x},\ \mathbf{x}'\mathbf{H}\mathbf{x}) = 2\operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\Sigma}) + 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\mu}, \tag{7.19}$$

or, in the special case where $\mathbf{H} = \mathbf{A}$, to

$$\operatorname{var}(\mathbf{x}'\mathbf{A}\mathbf{x}) = 2\operatorname{tr}[(\mathbf{A}\boldsymbol{\Sigma})^2] + 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}. \tag{7.20}$$

The formulas of Theorems 5.7.2 and 5.7.3 were derived under the assumption that the matrix $\mathbf{A}$ of the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ is symmetric and (in the case of Theorem 5.7.3) the assumption that the matrix $\mathbf{H}$ of the quadratic form $\mathbf{x}'\mathbf{H}\mathbf{x}$ is symmetric. Note that whether or not $\mathbf{A}$ and/or $\mathbf{H}$ are symmetric, it would be the case that $\mathbf{x}'\mathbf{A}\mathbf{x} = \frac{1}{2}\mathbf{x}'(\mathbf{A}+\mathbf{A}')\mathbf{x}$ and that $\mathbf{x}'\mathbf{H}\mathbf{x} = \frac{1}{2}\mathbf{x}'(\mathbf{H}+\mathbf{H}')\mathbf{x}$. Thus, the formulas of Theorems 5.7.2 and 5.7.3 could be extended to the case where the matrices of the quadratic forms are possibly nonsymmetric simply by substituting $\frac{1}{2}(\mathbf{A}+\mathbf{A}')$ for $\mathbf{A}$ and (in the case of Theorem 5.7.3) $\frac{1}{2}(\mathbf{H}+\mathbf{H}')$ for $\mathbf{H}$.

Some alternative representations. By making use of the vec and vech operations, the expressions provided by the formulas of Theorems 5.7.1, 5.7.2, and 5.7.3 can be recast in ways that are informative about the nature of the dependence of the expressions on the elements of the matrices of the quadratic forms.

An alternative to the matrix expression (7.11) provided by Theorem 5.7.1 for the expected value of the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ is as follows:

$$\mathrm{E}(\mathbf{x}'\mathbf{A}\mathbf{x}) = [\operatorname{vec}(\boldsymbol{\Sigma}) + (\boldsymbol{\mu}\otimes\boldsymbol{\mu})]'\operatorname{vec}\mathbf{A} \tag{7.21}$$

[as can be readily verified from expression (7.10)]. Expression (7.21) is a linear form in $\operatorname{vec}\mathbf{A}$; that is, it is a linear combination of the elements of $\operatorname{vec}\mathbf{A}$ (which are the elements of $\mathbf{A}$). If $\mathbf{A}$ is symmetric, then expression (7.21) can be restated as follows:

$$\mathrm{E}(\mathbf{x}'\mathbf{A}\mathbf{x}) = [\operatorname{vec}(\boldsymbol{\Sigma}) + (\boldsymbol{\mu}\otimes\boldsymbol{\mu})]'\mathbf{G}_N\operatorname{vech}\mathbf{A} \tag{7.22}$$

(where $\mathbf{G}_N$ is the duplication matrix). Expression (7.22) is a linear form in $\operatorname{vech}\mathbf{A}$, the elements of which are $N(N+1)/2$ nonredundant elements of $\mathbf{A}$—if $\mathbf{A}$ is symmetric, $N(N-1)/2$ of its elements are redundant.

Now, consider the matrix expression (7.13) provided by Theorem 5.7.2 for the covariance of the linear form $\mathbf{b}'\mathbf{x}$ and the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$. Making use of result (7.6), we find that the second term of expression (7.13) can be reexpressed as follows:

$$2\mathbf{b}'\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu} = 2\operatorname{tr}(\mathbf{b}'\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}) = 2\operatorname{tr}[\mathbf{b}'\boldsymbol{\Sigma}\mathbf{A}(\boldsymbol{\mu}')'] = 2(\operatorname{vec}\mathbf{b})'(\boldsymbol{\mu}'\otimes\boldsymbol{\Sigma})\operatorname{vec}\mathbf{A} = 2\mathbf{b}'(\boldsymbol{\mu}'\otimes\boldsymbol{\Sigma})\operatorname{vec}\mathbf{A}.$$

Thus, formula (7.13) can be restated as follows:

$$\begin{aligned}\operatorname{cov}(\mathbf{b}'\mathbf{x},\ \mathbf{x}'\mathbf{A}\mathbf{x}) &= \mathbf{b}'[\boldsymbol{\Lambda} + 2(\boldsymbol{\mu}'\otimes\boldsymbol{\Sigma})]\operatorname{vec}\mathbf{A} \\ &= \mathbf{b}'[\boldsymbol{\Lambda} + 2(\boldsymbol{\mu}'\otimes\boldsymbol{\Sigma})]\mathbf{G}_N\operatorname{vech}\mathbf{A}.\end{aligned} \tag{7.23}$$

Expression (7.23) is a bilinear form in the N-dimensional column vector $\mathbf{b}$ and the $N(N+1)/2$-dimensional column vector $\operatorname{vech}\mathbf{A}$, that is, for any particular value of $\mathbf{b}$, it is a linear form in $\operatorname{vech}\mathbf{A}$, and for any particular value of $\operatorname{vech}\mathbf{A}$, it is a linear form in $\mathbf{b}$.

Further, consider the matrix expression (7.16) provided by Theorem 5.7.3 for the covariance of the two quadratic forms $\mathbf{x}'\mathbf{A}\mathbf{x}$ and $\mathbf{x}'\mathbf{H}\mathbf{x}$. Making use of results (7.5) and (7.4), we find that the second and third terms of expression (7.16) can be reexpressed as follows:

$$\begin{aligned} 2\boldsymbol{\mu}'\mathbf{H}\boldsymbol{\Lambda}\operatorname{vec}\mathbf{A} &= 2[\boldsymbol{\mu}'\mathbf{H}\boldsymbol{\Lambda}\operatorname{vec}\mathbf{A}]' \\ &= 2(\operatorname{vec}\mathbf{A})'\boldsymbol{\Lambda}'\mathbf{H}\boldsymbol{\mu} \\ &= 2(\operatorname{vec}\mathbf{A})'\operatorname{vec}(\boldsymbol{\Lambda}'\mathbf{H}\boldsymbol{\mu}) \\ &= 2(\operatorname{vec}\mathbf{A})'(\boldsymbol{\mu}'\otimes\boldsymbol{\Lambda}')\operatorname{vec}\mathbf{H} \\ &= 2(\operatorname{vec}\mathbf{A})'(\boldsymbol{\mu}\otimes\boldsymbol{\Lambda})'\operatorname{vec}\mathbf{H} \end{aligned} \tag{7.24}$$

and

$$\begin{aligned} 2\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Lambda}\operatorname{vec}\mathbf{H} &= 2(\boldsymbol{\Lambda}'\mathbf{A}\boldsymbol{\mu})'\operatorname{vec}\mathbf{H} \\ &= 2[\operatorname{vec}(\boldsymbol{\Lambda}'\mathbf{A}\boldsymbol{\mu})]'\operatorname{vec}\mathbf{H} \\ &= 2[(\boldsymbol{\mu}'\otimes\boldsymbol{\Lambda}')\operatorname{vec}\mathbf{A}]'\operatorname{vec}\mathbf{H} \\ &= 2(\operatorname{vec}\mathbf{A})'(\boldsymbol{\mu}\otimes\boldsymbol{\Lambda})\operatorname{vec}\mathbf{H}. \end{aligned} \tag{7.25}$$

Moreover, making use of result (7.6) (and Lemma 2.3.1), the fourth and fifth terms of expression (7.16) are reexpressible as

$$2\operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\Sigma}) = 2\operatorname{tr}(\mathbf{A}'\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\Sigma}') = 2(\operatorname{vec}\mathbf{A})'(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\operatorname{vec}\mathbf{H} \tag{7.26}$$

and

$$\begin{aligned} 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\mu} &= 4\operatorname{tr}(\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\mu}) \\ &= 4\operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\mu}\boldsymbol{\mu}') \\ &= 4\operatorname{tr}[\mathbf{A}'\boldsymbol{\Sigma}\mathbf{H}(\boldsymbol{\mu}\boldsymbol{\mu}')'] \\ &= 4(\operatorname{vec}\mathbf{A})'[(\boldsymbol{\mu}\boldsymbol{\mu}')\otimes\boldsymbol{\Sigma}]\operatorname{vec}\mathbf{H}. \end{aligned} \tag{7.27}$$

And based on results (7.24), (7.25), (7.26), and (7.27), formula (7.16) can be restated as follows:

$$\begin{aligned} \operatorname{cov}(\mathbf{x}'\mathbf{A}\mathbf{x},\ \mathbf{x}'\mathbf{H}\mathbf{x}) &= (\operatorname{vec}\mathbf{A})'\{\boldsymbol{\Omega} + 2(\boldsymbol{\mu}\otimes\boldsymbol{\Lambda})' + 2(\boldsymbol{\mu}\otimes\boldsymbol{\Lambda}) \\ &\qquad + 2(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}) + 4[(\boldsymbol{\mu}\boldsymbol{\mu}')\otimes\boldsymbol{\Sigma}]\}\operatorname{vec}\mathbf{H} \qquad (7.28) \\ &= (\operatorname{vech}\mathbf{A})'\mathbf{G}_N'\{\boldsymbol{\Omega} + 2(\boldsymbol{\mu}\otimes\boldsymbol{\Lambda})' + 2(\boldsymbol{\mu}\otimes\boldsymbol{\Lambda}) \\ &\qquad + 2(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}) + 4[(\boldsymbol{\mu}\boldsymbol{\mu}')\otimes\boldsymbol{\Sigma}]\}\mathbf{G}_N\operatorname{vech}\mathbf{H}. \qquad (7.29) \end{aligned}$$

In result (7.29), $\operatorname{cov}(\mathbf{x}'\mathbf{A}\mathbf{x},\ \mathbf{x}'\mathbf{H}\mathbf{x})$ is expressed as a bilinear form in the $N(N+1)/2$-dimensional column vectors $\operatorname{vech}\mathbf{A}$ and $\operatorname{vech}\mathbf{H}$. As a special case of result (7.29) (that where $\mathbf{H} = \mathbf{A}$), we have the result

$$\begin{aligned} \operatorname{var}(\mathbf{x}'\mathbf{A}\mathbf{x}) = (\operatorname{vech}\mathbf{A})'\mathbf{G}_N'\{&\boldsymbol{\Omega} + 2(\boldsymbol{\mu}\otimes\boldsymbol{\Lambda})' + 2(\boldsymbol{\mu}\otimes\boldsymbol{\Lambda}) \\ &+ 2(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}) + 4[(\boldsymbol{\mu}\boldsymbol{\mu}')\otimes\boldsymbol{\Sigma}]\}\mathbf{G}_N\operatorname{vech}\mathbf{A}, \end{aligned} \tag{7.30}$$

in which $\operatorname{var}(\mathbf{x}'\mathbf{A}\mathbf{x})$ is expressed as a quadratic form in the $N(N+1)/2$-dimensional column vector $\operatorname{vech}\mathbf{A}$.

c. Estimation of σ^2 (under the G–M model)

Let us add to our earlier discussion of the method of least squares by introducing some notation, terminology, and results that are relevant to making inferences about variability and covariability.

Suppose that $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$ is an $N \times 1$ observable random vector that follows a G–M, Aitken, or general linear model. And let

$$\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{P}_{\mathbf{X}}\mathbf{y} \quad [= (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}]$$

[where $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$], so that (for $i = 1, 2, \ldots, N$) the ith element of $\tilde{\mathbf{e}}$ is the difference between y_i and the least squares estimator of $\mathrm{E}(y_i)$. It is customary to refer to the elements of $\tilde{\mathbf{e}}$ (or to their observed values) as *least squares residuals*, or simply as *residuals*, and to refer to $\tilde{\mathbf{e}}$ itself (or to its observed value) as the *residual vector*.

Upon applying formulas (3.1.7) and (3.2.47) and observing (in light of Theorem 2.12.2) that $\mathbf{P}_{\mathbf{X}}\mathbf{X} = \mathbf{X}$ and that $\mathbf{P}_{\mathbf{X}}$ is symmetric and idempotent, we find that

$$\mathrm{E}(\tilde{\mathbf{e}}) = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{X}\boldsymbol{\beta} = (\mathbf{X} - \mathbf{X})\boldsymbol{\beta} = \mathbf{0} \tag{7.31}$$

and that, in the special case where $\mathbf{y}$ follows a G–M model,

$$\mathrm{var}(\tilde{\mathbf{e}}) = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})(\sigma^2\mathbf{I})(\mathbf{I} - \mathbf{P}_{\mathbf{X}})' = \sigma^2(\mathbf{I} - \mathbf{P}_{\mathbf{X}}). \tag{7.32}$$

Corresponding to the vector $\tilde{\mathbf{e}}$ is the quantity $\tilde{\mathbf{e}}'\tilde{\mathbf{e}}$, which is customarily referred to as the *residual sum of squares*. It follows from the results of Section 5.4b (on least squares minimization) that, for every value of $\mathbf{y}$,

$$\tilde{\mathbf{e}}'\tilde{\mathbf{e}} = \min_{\mathbf{b}} (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}). \tag{7.33}$$

Moreover,

$$\tilde{\mathbf{e}}'\tilde{\mathbf{e}} = \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}, \tag{7.34}$$

as is evident from the results of Section 5.4b or upon observing (in light of the symmetry and idempotency of $\mathbf{P}_{\mathbf{X}}$) that $(\mathbf{I} - \mathbf{P}_{\mathbf{X}})'(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) = \mathbf{I} - \mathbf{P}_{\mathbf{X}}$.

In what follows (i.e., in the remainder of Subsection c), it is supposed that $\mathbf{y}$ follows a G–M model, and the emphasis is on the estimation of the parameter σ^2.

An unbiased estimator. The expected value of the residual sum of squares can be derived by applying formula (7.11) (for the expected value of a quadratic form) to expression (7.34) (which is a quadratic form in the random vector $\mathbf{y}$). Recalling that $\mathbf{P}_{\mathbf{X}}\mathbf{X} = \mathbf{X}$, we find that

$$\begin{aligned}
\mathrm{E}(\tilde{\mathbf{e}}'\tilde{\mathbf{e}}) &= \mathrm{E}[\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}] \\
&= \mathrm{tr}[(\mathbf{I} - \mathbf{P}_{\mathbf{X}})(\sigma^2\mathbf{I})] + (\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{X}\boldsymbol{\beta} \\
&= \sigma^2\,\mathrm{tr}(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) + 0 \\
&= \sigma^2[N - \mathrm{tr}(\mathbf{P}_{\mathbf{X}})].
\end{aligned}$$

Moreover, because $\mathbf{P}_{\mathbf{X}}$ is idempotent and because $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ is a generalized inverse of $\mathbf{X}$—refer to Part (1) of Theorem 2.12.2—we have (in light of Corollary 2.8.3 and Lemma 2.10.13) that

$$\mathrm{tr}(\mathbf{P}_{\mathbf{X}}) = \mathrm{rank}(\mathbf{P}_{\mathbf{X}}) = \mathrm{rank}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'] = \mathrm{rank}\,\mathbf{X}. \tag{7.35}$$

Thus, the expected value of the residual sum of squares is

$$\mathrm{E}(\tilde{\mathbf{e}}'\tilde{\mathbf{e}}) = \sigma^2(N - \mathrm{rank}\,\mathbf{X}). \tag{7.36}$$

Assume that the rank of the model matrix $\mathbf{X}$ is (strictly) less than N. Then, upon dividing the residual sum of squares by $N - \mathrm{rank}\,\mathbf{X}$, we obtain the quantity

$$\hat{\sigma}^2 = \frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}{N - \mathrm{rank}\,\mathbf{X}}.$$

Clearly,

$$\mathrm{E}(\hat{\sigma}^2) = \sigma^2, \tag{7.37}$$

that is, the quantity $\hat{\sigma}^2$ obtained by dividing the residual sum of squares by $N - \text{rank}\,\mathbf{X}$ is an unbiased estimator of the parameter σ^2.

Let us find the variance of the estimator $\hat{\sigma}^2$. Suppose that the fourth-order moments of the distribution of the vector $\mathbf{e} = (e_1, e_2, \ldots, e_N)'$ of residual effects are such that (for $i, j, k, m = 1, 2, \ldots, N$)

$$\mathrm{E}(e_i e_j e_k e_m) = \begin{cases} 3\sigma^4 & \text{if } m = k = j = i, \\ \sigma^4 & \text{if } j = i \text{ and } m = k \neq i, \text{ if } k = i \text{ and } m = j \neq i, \\ & \quad \text{or if } m = i \text{ and } k = j \neq i, \\ 0 & \text{otherwise} \end{cases} \tag{7.38}$$

(as would be the case if the distribution of $\mathbf{e}$ were MVN). Further, let $\mathbf{\Lambda}$ represent the $N \times N^2$ matrix whose entry for the jth row and kmth column [column $(k-1)N + m$] is $\mathrm{E}(e_j e_k e_m)$. Then, upon applying formula (7.17) and once again making use of the properties of the $\mathbf{P}_{\mathbf{X}}$ matrix (set forth in Theorem 2.12.2), we find that

$$\begin{aligned} \mathrm{var}(\tilde{\mathbf{e}}'\tilde{\mathbf{e}}) &= \mathrm{var}[\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}] \\ &= 4\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{\Lambda}\,\mathrm{vec}(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) + 2\,\mathrm{tr}[(\mathbf{I} - \mathbf{P}_{\mathbf{X}})(\sigma^2\mathbf{I})(\mathbf{I} - \mathbf{P}_{\mathbf{X}})(\sigma^2\mathbf{I})] \\ &\qquad + 4\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})(\sigma^2\mathbf{I})(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{X}\boldsymbol{\beta} \\ &= 0 + 2\sigma^4\,\mathrm{tr}(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) + 0 \\ &= 2\sigma^4[N - \mathrm{tr}(\mathbf{P}_{\mathbf{X}})]. \end{aligned}$$

And upon applying result (7.35), we conclude that

$$\mathrm{var}(\tilde{\mathbf{e}}'\tilde{\mathbf{e}}) = 2\sigma^4(N - \text{rank}\,\mathbf{X}). \tag{7.39}$$

Moreover, as a particular implication of result (7.39), we have that

$$\mathrm{var}(\hat{\sigma}^2) = \frac{\mathrm{var}(\tilde{\mathbf{e}}'\tilde{\mathbf{e}})}{(N - \text{rank}\,\mathbf{X})^2} = \frac{2\sigma^4}{N - \text{rank}\,\mathbf{X}}. \tag{7.40}$$

The Hodges–Lehmann estimator. The estimator $\hat{\sigma}^2$ is of the general form

$$\frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}{k}, \tag{7.41}$$

where k is a (strictly) positive constant. It is the estimator of the form (7.41) obtained by taking $k = N - \text{rank}\,\mathbf{X}$. Taking $k = N - \text{rank}\,\mathbf{X}$ achieves unbiasedness. Nevertheless, it can be of interest to consider other choices for k.

Let us derive the MSE (mean squared error) of the estimator (7.41). And, in doing so, let us continue to suppose that the fourth-order moments of the distribution of the vector $\mathbf{e} = (e_1, e_2, \ldots, e_N)'$ of residual effects are such that (for $i, j, k, m = 1, 2, \ldots, N$) $\mathrm{E}(e_i e_j e_k e_m)$ satisfies condition (7.38) (as would be the case if the distribution of $\mathbf{e}$ were MVN).

The MSE of the estimator (7.41) can be regarded as a function, say $m(k)$, of the scalar k. Making

use of results (7.36) and (7.39), we find that

$$\begin{aligned} m(k) &= \text{var}\left(\frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}{k}\right) + \left[\text{E}\left(\frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}{k}\right) - \sigma^2\right]^2 \\ &= \frac{1}{k^2}\left\{\text{var}(\tilde{\mathbf{e}}'\tilde{\mathbf{e}}) + [\text{E}(\tilde{\mathbf{e}}'\tilde{\mathbf{e}}) - k\sigma^2]^2\right\} \\ &= \frac{\sigma^4}{k^2}\left\{2(N - \text{rank}\,\mathbf{X}) + [N - \text{rank}(\mathbf{X}) - k]^2\right\} \qquad (7.42) \\ &= \sigma^4\left\{\frac{(N - \text{rank}\,\mathbf{X})[N - \text{rank}(\mathbf{X}) + 2 - 2k]}{k^2} + 1\right\}. \qquad (7.43) \end{aligned}$$

For what choice of k does $m(k)$ attain its minimum value? Upon differentiating $m(k)$ and engaging in some algebraic simplification, we find that

$$\frac{dm(k)}{dk} = \frac{-2\sigma^4(N - \text{rank}\,\mathbf{X})[N - \text{rank}(\mathbf{X}) + 2 - k]}{k^3},$$

so that $\frac{dm(k)}{dk} < 0$ if $k < N - \text{rank}(\mathbf{X}) + 2$, $\frac{dm(k)}{dk} = 0$ if $k = N - \text{rank}(\mathbf{X}) + 2$, and $\frac{dm(k)}{dk} > 0$ if $k > N - \text{rank}(\mathbf{X}) + 2$. Thus, $m(k)$ is a decreasing function of k over the interval $0 < k \leq N - \text{rank}(\mathbf{X}) + 2$, is an increasing function over the interval $k \geq N - \text{rank}(\mathbf{X}) + 2$, and attains its minimum value at $k = N - \text{rank}(\mathbf{X}) + 2$.

We conclude that among estimators of σ^2 of the form (7.41), the estimator

$$\frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}{N - \text{rank}(\mathbf{X}) + 2} \qquad (7.44)$$

has minimum MSE. The estimator (7.44) is sometimes referred to as the *Hodges–Lehmann estimator*. In light of results (7.36) and (7.42), it has a bias of

$$\begin{aligned} \text{E}\left[\frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}{N - \text{rank}(\mathbf{X}) + 2}\right] - \sigma^2 &= \sigma^2\left[\frac{N - \text{rank}\,\mathbf{X}}{N - \text{rank}(\mathbf{X}) + 2} - 1\right] \\ &= \frac{-2\sigma^2}{N - \text{rank}(\mathbf{X}) + 2} \qquad (7.45) \end{aligned}$$

and an MSE of

$$\frac{\sigma^4[2(N - \text{rank}\,\mathbf{X}) + (-2)^2]}{[N - \text{rank}(\mathbf{X}) + 2]^2} = \frac{2\sigma^4}{N - \text{rank}(\mathbf{X}) + 2}. \qquad (7.46)$$

By way of comparison, the unbiased estimator $\hat{\sigma}^2$ (obtained by taking $k = N - \text{rank}\,\mathbf{X}$) has an MSE of $2\sigma^4/(N - \text{rank}\,\mathbf{X})$.

Statistical independence. Let us conclude the present subsection (Subsection c) with some results pertaining to least squares estimators of estimable linear combinations of the elements of the parametric vector $\boldsymbol{\beta}$. The least squares estimator of any such linear combination is expressible as $\mathbf{r}'\mathbf{X}'\mathbf{y}$ for some $P \times 1$ vector $\mathbf{r}$ of constants; more generally, the M-dimensional column vector whose elements are the least squares estimators of M such linear combinations is expressible as $\mathbf{R}'\mathbf{X}'\mathbf{y}$ for some $P \times M$ matrix $\mathbf{R}$ of constants. Making use of formula (3.2.46) and of Parts (4) and (2) of Theorem 2.12.2, we find that

$$\text{cov}(\mathbf{r}'\mathbf{X}'\mathbf{y},\, \tilde{\mathbf{e}}) = \mathbf{r}'\mathbf{X}'(\sigma^2\mathbf{I})(\mathbf{I} - \mathbf{P}_{\mathbf{X}})' = \sigma^2\mathbf{r}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) = \mathbf{0} \qquad (7.47)$$

and, similarly (and more generally), that

$$\text{cov}(\mathbf{R}'\mathbf{X}'\mathbf{y},\ \tilde{\mathbf{e}}) = \mathbf{0}. \tag{7.48}$$

Thus, the least squares estimator $\mathbf{r}'\mathbf{X}'\mathbf{y}$ and the residual vector $\tilde{\mathbf{e}}$ are uncorrelated. And, more generally, the vector $\mathbf{R}'\mathbf{X}'\mathbf{y}$ of least squares estimators and the residual vector $\tilde{\mathbf{e}}$ are uncorrelated.

Is the least squares estimator $\mathbf{r}'\mathbf{X}'\mathbf{y}$ uncorrelated with the residual sum of squares $\tilde{\mathbf{e}}'\tilde{\mathbf{e}}$? Or, equivalently, is $\mathbf{r}'\mathbf{X}'\mathbf{y}$ uncorrelated with an estimator of σ^2 of the form (7.41), including the unbiased estimator $\hat{\sigma}^2$ (and the Hodges–Lehmann estimator)? Assuming the model is such that the distribution of the vector $\mathbf{e} = (e_1, e_2, \ldots, e_N)'$ of residual effects has third-order moments $\lambda_{ijk} = \text{E}(e_i e_j e_k)$ $(i, j, k = 1, 2, \ldots, N)$ and making use of formula (7.13), we find that

$$\text{cov}(\mathbf{r}'\mathbf{X}'\mathbf{y},\ \tilde{\mathbf{e}}'\tilde{\mathbf{e}}) = \mathbf{r}'\mathbf{X}'\boldsymbol{\Lambda}\ \text{vec}(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) + 2\mathbf{r}'\mathbf{X}'(\sigma^2\mathbf{I})(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{X}\boldsymbol{\beta}, \tag{7.49}$$

where $\boldsymbol{\Lambda}$ is an $N \times N^2$ matrix whose entry for the ith row and jkth column [column $(j-1)N + k$] is λ_{ijk}. The second term of expression (7.49) equals 0, as is evident upon recalling that $\mathbf{P}_{\mathbf{X}}\mathbf{X} = \mathbf{X}$, and the first term equals 0 if $\boldsymbol{\Lambda} = \mathbf{0}$, as would be the case if the distribution of $\mathbf{e}$ were MVN or, more generally, if the distribution of $\mathbf{e}$ were symmetric. Thus, if the distribution of $\mathbf{e}$ is symmetric, then

$$\text{cov}(\mathbf{r}'\mathbf{X}'\mathbf{y},\ \tilde{\mathbf{e}}'\tilde{\mathbf{e}}) = 0 \tag{7.50}$$

and, more generally,

$$\text{cov}(\mathbf{R}'\mathbf{X}'\mathbf{y},\ \tilde{\mathbf{e}}'\tilde{\mathbf{e}}) = \mathbf{0}. \tag{7.51}$$

Accordingly, if the distribution of $\mathbf{e}$ is symmetric, $\mathbf{r}'\mathbf{X}'\mathbf{y}$ and $\mathbf{R}'\mathbf{X}'\mathbf{y}$ are uncorrelated with any estimator of σ^2 of the form (7.41), including the unbiased estimator $\hat{\sigma}^2$ (and the Hodges–Lehmann estimator).

Are the vector $\mathbf{R}'\mathbf{X}'\mathbf{y}$ of least squares estimators and the residual vector $\tilde{\mathbf{e}}$ statistically independent (as well as uncorrelated)? If the model is such that the distribution of $\mathbf{e}$ is MVN (in which case the distribution of $\mathbf{y}$ is also MVN), then it follows from Corollary 3.5.6 that the answer is yes. That is, if the model is such that the distribution of $\mathbf{e}$ is MVN, then $\tilde{\mathbf{e}}$ is distributed independently of $\mathbf{R}'\mathbf{X}'\mathbf{y}$ (and, in particular, $\tilde{\mathbf{e}}$ is distributed independently of $\mathbf{r}'\mathbf{X}'\mathbf{y}$). Moreover, $\tilde{\mathbf{e}}$ being distributed independently of $\mathbf{R}'\mathbf{X}'\mathbf{y}$ implies that "any" function of $\tilde{\mathbf{e}}$ is distributed independently of $\mathbf{R}'\mathbf{X}'\mathbf{y}$—refer, e.g., to Casella and Berger (2002, theorem 4.6.12). Accordingly, if the distribution of $\mathbf{e}$ is MVN, then the residual sum of squares $\tilde{\mathbf{e}}'\tilde{\mathbf{e}}$ is distributed independently of $\mathbf{R}'\mathbf{X}'\mathbf{y}$ and any estimator of σ^2 of the form (7.41) (including the unbiased estimator $\hat{\sigma}^2$ and the Hodges–Lehmann estimator) is distributed independently of $\mathbf{R}'\mathbf{X}'\mathbf{y}$.

d. Translation invariance

Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M, Aitken, or general linear model. And suppose that we wish to make inferences about σ^2 (in the case of a G–M or Aitken model) or $\boldsymbol{\theta}$ (in the case of a general linear model) or about various functions of σ^2 or $\boldsymbol{\theta}$. In making such inferences, it is common practice to restrict attention to procedures that depend on the value of $\mathbf{y}$ only through the value of a (possibly vector-valued) statistic having a property known as translation invariance (or location invariance).

Proceeding as in Section 5.2 (in discussing the translation-equivariant estimation of a parametric function of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$), let $\mathbf{k}$ represent a P-dimensional column vector of known constants, and define $\mathbf{z} = \mathbf{y} + \mathbf{X}\mathbf{k}$. Then, $\mathbf{z} = \mathbf{X}\boldsymbol{\tau} + \mathbf{e}$, where $\boldsymbol{\tau} = \boldsymbol{\beta} + \mathbf{k}$. And $\mathbf{z}$ can be regarded as an $N \times 1$ observable random vector that follows a G–M, Aitken, or general linear model that is identical in all respects to the model followed by $\mathbf{y}$, except that the role of the parametric vector $\boldsymbol{\beta}$ is played by a vector (represented by $\boldsymbol{\tau}$) having a different interpretation. It can be argued that inferences about σ^2 or $\boldsymbol{\theta}$, or about functions of σ^2 or $\boldsymbol{\theta}$, should be made on the basis of a statistical procedure that

depends on the value of $\mathbf{y}$ only through the value of a (possibly vector-valued) statistic $\mathbf{h}(\mathbf{y})$ that, for every $\mathbf{k} \in \mathcal{R}^P$ (and for every value of $\mathbf{y}$), satisfies the condition

$$\mathbf{h}(\mathbf{y}) = \mathbf{h}(\mathbf{z})$$

or, equivalently, the condition

$$\mathbf{h}(\mathbf{y}) = \mathbf{h}(\mathbf{y} + \mathbf{X}\mathbf{k}). \tag{7.52}$$

Any statistic $\mathbf{h}(\mathbf{y})$ that satisfies condition (7.52) and that does so for every $\mathbf{k} \in \mathcal{R}^P$ (and for every value of $\mathbf{y}$) is said to be *translation invariant*.

If the statistic $\mathbf{h}(\mathbf{y})$ is translation invariant, then

$$\mathbf{h}(\mathbf{y}) = \mathbf{h}[\mathbf{y} + \mathbf{X}(-\boldsymbol{\beta})] = \mathbf{h}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{h}(\mathbf{e}). \tag{7.53}$$

Thus, the statistical properties of a statistical procedure that depends on the value of $\mathbf{y}$ only through the value of a translation-invariant statistic $\mathbf{h}(\mathbf{y})$ are completely determined by the distribution of the vector $\mathbf{e}$ of residual effects. They do not depend on the vector $\boldsymbol{\beta}$.

Let us now consider condition (7.52) in the special case where $\mathbf{h}(\mathbf{y})$ is a scalar-valued statistic $h(\mathbf{y})$ of the form

$$h(\mathbf{y}) = \mathbf{y}'\mathbf{A}\mathbf{y},$$

where $\mathbf{A}$ is a symmetric matrix of constants. In this special case,

$$\begin{aligned} \mathbf{h}(\mathbf{y}) = \mathbf{h}(\mathbf{y} + \mathbf{X}\mathbf{k}) \quad &\Leftrightarrow \quad \mathbf{y}'\mathbf{A}\mathbf{X}\mathbf{k} + \mathbf{k}'\mathbf{X}'\mathbf{A}\mathbf{y} + \mathbf{k}'\mathbf{X}'\mathbf{A}\mathbf{X}\mathbf{k} = 0 \\ &\Leftrightarrow \quad 2\mathbf{y}'\mathbf{A}\mathbf{X}\mathbf{k} = -\mathbf{k}'\mathbf{X}'\mathbf{A}\mathbf{X}\mathbf{k}. \end{aligned} \tag{7.54}$$

For condition (7.54) to be satisfied for every $\mathbf{k} \in \mathcal{R}^P$ (and for every value of $\mathbf{y}$), it is sufficient that $\mathbf{A}\mathbf{X} = \mathbf{0}$. It is also necessary. To see this, suppose that condition (7.54) is satisfied for every $\mathbf{k} \in \mathcal{R}^P$ (and for every value of $\mathbf{y}$). Then, upon setting $\mathbf{y} = \mathbf{0}$ in condition (7.54), we find that $\mathbf{k}'\mathbf{X}'\mathbf{A}\mathbf{X}\mathbf{k} = 0$ for every $\mathbf{k} \in \mathcal{R}^P$, implying (in light of Corollary 2.13.4) that $\mathbf{X}'\mathbf{A}\mathbf{X} = \mathbf{0}$. Thus, $\mathbf{y}'\mathbf{A}\mathbf{X}\mathbf{k} = 0$ for every $\mathbf{k} \in \mathcal{R}^P$ (and every value of $\mathbf{y}$), implying that every element of $\mathbf{A}\mathbf{X}$ equals 0 and hence that $\mathbf{A}\mathbf{X} = \mathbf{0}$.

In summary, we have established that the quadratic form $\mathbf{y}'\mathbf{A}\mathbf{y}$ (where $\mathbf{A}$ is a symmetric matrix of constants) is a translation-invariant statistic if and only if the matrix $\mathbf{A}$ of the quadratic form satisfies the condition

$$\mathbf{A}\mathbf{X} = \mathbf{0}. \tag{7.55}$$

Adopting the same notation and terminology as in Subsection c, consider the concept of translation invariance as applied to the residual vector $\tilde{\mathbf{e}}$ and to the residual sum of squares $\tilde{\mathbf{e}}'\tilde{\mathbf{e}}$. Recall that $\tilde{\mathbf{e}}$ is expressible as $\tilde{\mathbf{e}} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$ and $\tilde{\mathbf{e}}'\tilde{\mathbf{e}}$ as $\tilde{\mathbf{e}}'\tilde{\mathbf{e}} = \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$. Recall also that $\mathbf{P}_{\mathbf{X}}\mathbf{X} = \mathbf{X}$ and hence that $(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{X} = \mathbf{0}$. Thus, for any $P \times 1$ vector $\mathbf{k}$ (and for any value of $\mathbf{y}$),

$$(\mathbf{I} - \mathbf{P}_{\mathbf{X}})(\mathbf{y} + \mathbf{X}\mathbf{k}) = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}.$$

And it follows that $\tilde{\mathbf{e}}$ is translation invariant. Moreover, $\tilde{\mathbf{e}}'\tilde{\mathbf{e}}$ is also translation invariant, as is evident upon observing that it depends on $\mathbf{y}$ only through the value of $\tilde{\mathbf{e}}$ or, alternatively, upon applying condition (7.55) (with $\mathbf{A} = \mathbf{I} - \mathbf{P}_{\mathbf{X}}$)—that condition (7.55) is applicable is evident upon recalling that $\mathbf{P}_{\mathbf{X}}$ is symmetric and hence that the matrix $\mathbf{I} - \mathbf{P}_{\mathbf{X}}$ of the quadratic form $\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$ is symmetric.

Let us now specialize by supposing that $\mathbf{y}$ follows a G–M model, and let us add to the results obtained in Subsection c (on the estimation of σ^2) by obtaining some results on translation-invariant estimation. Since the residual sum of squares $\tilde{\mathbf{e}}'\tilde{\mathbf{e}}$ is translation invariant, any estimator of σ^2 of the form (7.41) is translation invariant. In particular, the unbiased estimator $\hat{\sigma}^2$ is translation invariant (and the Hodges–Lehmann estimator is translation invariant).

A quadratic form $\mathbf{y}'\mathbf{A}\mathbf{y}$ in the observable random vector $\mathbf{y}$ (where $\mathbf{A}$ is a symmetric matrix of constants) is an unbiased estimator of σ^2 and is translation invariant if and only if

$$\mathrm{E}(\mathbf{y}'\mathbf{A}\mathbf{y}) = \sigma^2 \qquad \text{and} \qquad \mathbf{A}\mathbf{X} = \mathbf{0} \tag{7.56}$$

(in which case the quadratic form is referred to as a quadratic unbiased translation-invariant estimator). As an application of formula (7.11), we have that

$$\mathrm{E}(\mathbf{y}'\mathbf{A}\mathbf{y}) = \mathrm{tr}[\mathbf{A}(\sigma^2\mathbf{I})] + \boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta} = \sigma^2\,\mathrm{tr}(\mathbf{A}) + \boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta}. \tag{7.57}$$

In light of result (7.57), condition (7.56) is equivalent to the condition

$$\mathrm{tr}(\mathbf{A}) = 1 \quad \text{and} \quad \mathbf{A}\mathbf{X} = \mathbf{0}. \tag{7.58}$$

Thus, the quadratic form $\mathbf{y}'\mathbf{A}\mathbf{y}$ is a quadratic unbiased translation-invariant estimator of σ^2 if and only if the matrix $\mathbf{A}$ of the quadratic form satisfies condition (7.58).

Clearly, the estimator $\hat{\sigma}^2$ [which is expressible in the form $\hat{\sigma}^2 = \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$] is a quadratic unbiased translation-invariant estimator of σ^2. In fact, if the fourth-order moments of the distribution of the vector $\mathbf{e} = (e_1, e_2, \ldots, e_N)'$ of residual effects are such that (for $i, j, k, m = 1, 2, \ldots, N$) $\mathrm{E}(e_i e_j e_k e_m)$ satisfies condition (7.38) (as would be the case if the distribution of $\mathbf{e}$ were MVN), then the estimator $\hat{\sigma}^2$ has minimum variance (and hence minimum MSE) among all quadratic unbiased translation-invariant estimators of σ^2, as we now proceed to show.

Suppose that the fourth-order moments of the distribution of the vector $\mathbf{e} = (e_1, e_2, \ldots, e_N)'$ are such that (for $i, j, k, m = 1, 2, \ldots, N$) $\mathrm{E}(e_i e_j e_k e_m)$ satisfies condition (7.38). And denote by $\boldsymbol{\Lambda}$ the $N \times N^2$ matrix whose entry for the jth row and kmth column [column $(k-1)N + m$] is $\mathrm{E}(e_j e_k e_m)$. Then, for any quadratic unbiased translation-invariant estimator $\mathbf{y}'\mathbf{A}\mathbf{y}$ of σ^2 (where $\mathbf{A}$ is symmetric), we find [upon applying formula (7.17) and observing that $\mathbf{A}\mathbf{X} = \mathbf{0}$] that

$$\begin{aligned}\mathrm{var}(\mathbf{y}'\mathbf{A}\mathbf{y}) &= 4\boldsymbol{\beta}'(\mathbf{A}\mathbf{X})'\boldsymbol{\Lambda}\,\mathrm{vec}\,\mathbf{A} + 2\sigma^4\,\mathrm{tr}(\mathbf{A}^2) + 4\sigma^2\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{A}\mathbf{X}\boldsymbol{\beta} \\ &= 0 + 2\sigma^4\,\mathrm{tr}(\mathbf{A}^2) + 0 = 2\sigma^4\,\mathrm{tr}(\mathbf{A}^2).\end{aligned} \tag{7.59}$$

Let $\mathbf{R} = \mathbf{A} - \dfrac{1}{N - \mathrm{rank}\,\mathbf{X}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})$, so that

$$\mathbf{A} = \frac{1}{N - \mathrm{rank}\,\mathbf{X}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) + \mathbf{R}. \tag{7.60}$$

Further, observe that (since $\mathbf{P}_{\mathbf{X}}$ is symmetric) $\mathbf{R}' = \mathbf{R}$, that (since $\mathbf{A}\mathbf{X} = \mathbf{0}$ and $\mathbf{P}_{\mathbf{X}}\mathbf{X} = \mathbf{X}$)

$$\mathbf{R}\mathbf{X} = \mathbf{A}\mathbf{X} - \frac{1}{N - \mathrm{rank}\,\mathbf{X}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{X} = \mathbf{0} - \mathbf{0} = \mathbf{0},$$

and that

$$\mathbf{X}'\mathbf{R} = \mathbf{X}'\mathbf{R}' = (\mathbf{R}\mathbf{X})' = \mathbf{0}' = \mathbf{0}.$$

Accordingly, upon substituting expression (7.60) for $\mathbf{A}$ (and recalling that $\mathbf{P}_{\mathbf{X}}$ is idempotent), we find that

$$\mathbf{A}^2 = \frac{1}{(N - \mathrm{rank}\,\mathbf{X})^2}(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) + \frac{2}{N - \mathrm{rank}\,\mathbf{X}}\mathbf{R} + \mathbf{R}'\mathbf{R}. \tag{7.61}$$

Moreover, because $\mathrm{tr}(\mathbf{A}) = 1$, we have [in light of result (7.35)] that

$$\mathrm{tr}(\mathbf{R}) = \mathrm{tr}(\mathbf{A}) - \frac{1}{N - \mathrm{rank}\,\mathbf{X}}\,\mathrm{tr}(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) = 1 - 1 = 0. \tag{7.62}$$

And upon substituting expression (7.61) for $\mathbf{A}^2$ in expression (7.59) and making use of results (7.35) and (7.62), we find that

$$\begin{aligned}\mathrm{var}(\mathbf{y}'\mathbf{A}\mathbf{y}) &= 2\sigma^4\left[\frac{1}{(N - \mathrm{rank}\,\mathbf{X})^2}\,\mathrm{tr}(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) + \frac{2}{N - \mathrm{rank}\,\mathbf{X}}\,\mathrm{tr}(\mathbf{R}) + \mathrm{tr}(\mathbf{R}'\mathbf{R})\right] \\ &= 2\sigma^4\left[\frac{1}{N - \mathrm{rank}\,\mathbf{X}} + \mathrm{tr}(\mathbf{R}'\mathbf{R})\right].\end{aligned} \tag{7.63}$$

Finally, upon observing that $\mathrm{tr}(\mathbf{R}'\mathbf{R}) = \sum_{i,j} r_{ij}^2$, where (for $i, j = 1, 2, \ldots, N$) r_{ij} is the ijth element of $\mathbf{R}$, we conclude that $\mathrm{var}(\mathbf{y}'\mathbf{A}\mathbf{y})$ attains a minimum value of $2\sigma^4/(N - \mathrm{rank}\,\mathbf{X})$ and does so uniquely when $\mathbf{R} = \mathbf{0}$ or, equivalently, when $\mathbf{A} = \dfrac{1}{N - \mathrm{rank}\,\mathbf{X}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})$ (i.e., when $\mathbf{y}'\mathbf{A}\mathbf{y} = \hat{\sigma}^2$).

5.8 Best (Minimum-Variance) Unbiased Estimation

Take $\mathbf{y}$ to be an $N \times 1$ observable random vector that follows a G–M model, and consider the estimation of an estimable linear combination $\boldsymbol{\lambda}'\boldsymbol{\beta}$ of the elements of the parametric vector $\boldsymbol{\beta}$ and consider also the estimation of the parameter σ^2. In Section 5.5a, it was determined that the least squares estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ has minimum variance among all linear unbiased estimators. And in Section 5.7d, it was determined that the estimator $\hat{\sigma}^2 = \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}/(N - \text{rank}\,\mathbf{X})$ has minimum variance among all quadratic unbiased translation-invariant estimators of σ^2 [provided that the fourth-order moments of the distribution of the vector $\mathbf{e} = (e_1, e_2, \ldots, e_N)'$ of residual effects are such that (for $i, j, k, m = 1, 2, \ldots, N$) $\mathrm{E}(e_i e_j e_k e_m)$ satisfies condition (7.38)].

If the distribution of $\mathbf{e}$ is assumed to be MVN, something more can be said. It can be shown that under the assumption of multivariate normality, $\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$ form a complete sufficient statistic—refer, e.g., to Casella and Berger (2002, def. 6.2.21) or to Schervish (1995, def. 2.34) for the definition of completeness—in which case "any" function, say $t[\mathbf{X}'\mathbf{y},\ \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}]$, of $\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$ is a best (minimum-variance) unbiased estimator of $\mathrm{E}\{t[\mathbf{X}'\mathbf{y},\ \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}]\}$ (e.g., Schervish 1995, theorem 5.5; Casella and Berger 2002, theorem 7.3.23). It follows, in particular, that under the assumption of multivariate normality, the least squares estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ has minimum variance among all unbiased estimators (linear or not) and the estimator $\hat{\sigma}^2$ has minimum variance among all unbiased estimators of σ^2 (quadratic and/or translation invariant or not).

Let us assume that the distribution of $\mathbf{e}$ is MVN, and verify that (under the assumption of multivariate normality) $\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$ form a complete sufficient statistic. Let us begin by introducing a transformation of $\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$ that facilitates the verification.

Define $K = \text{rank}\,\mathbf{X}$. And observe that there exists an $N \times K$ matrix, say $\mathbf{W}$, whose columns form a basis for $\mathcal{C}(\mathbf{X})$. Observe also that $\mathbf{W} = \mathbf{X}\mathbf{R}$ for some matrix $\mathbf{R}$ and that $\mathbf{X} = \mathbf{W}\mathbf{S}$ for some $(K \times P)$ matrix $\mathbf{S}$ (of rank K). Moreover, $\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$ are expressible in terms of the $(K \times 1)$ vector $\mathbf{W}'\mathbf{y}$ and the sum of squares $\mathbf{y}'\mathbf{y}$; we have that

$$\mathbf{X}'\mathbf{y} = \mathbf{S}'\mathbf{W}'\mathbf{y} \quad \text{and} \quad \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y} = \mathbf{y}'\mathbf{y} - (\mathbf{W}'\mathbf{y})'\mathbf{S}(\mathbf{X}'\mathbf{X})^{-}\mathbf{S}'\mathbf{W}'\mathbf{y}. \tag{8.1}$$

Conversely, $\mathbf{W}'\mathbf{y}$ and $\mathbf{y}'\mathbf{y}$ are expressible in terms of $\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$; we have that

$$\mathbf{W}'\mathbf{y} = \mathbf{R}'\mathbf{X}'\mathbf{y} \quad \text{and} \quad \mathbf{y}'\mathbf{y} = \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y} + (\mathbf{X}'\mathbf{y})'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}. \tag{8.2}$$

Thus, corresponding to any function $g[\mathbf{X}'\mathbf{y},\ \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}]$ of $\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$, there is a function, say $g_*(\mathbf{W}'\mathbf{y},\ \mathbf{y}'\mathbf{y})$, of $\mathbf{W}'\mathbf{y}$ and $\mathbf{y}'\mathbf{y}$ such that $g_*(\mathbf{W}'\mathbf{y},\ \mathbf{y}'\mathbf{y}) = g[\mathbf{X}'\mathbf{y},\ \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}]$ for every value of $\mathbf{y}$; namely, the function $g_*(\mathbf{W}'\mathbf{y},\ \mathbf{y}'\mathbf{y})$ defined by

$$g_*(\mathbf{W}'\mathbf{y},\ \mathbf{y}'\mathbf{y}) = g[\mathbf{S}'\mathbf{W}'\mathbf{y},\ \mathbf{y}'\mathbf{y} - (\mathbf{W}'\mathbf{y})'\mathbf{S}(\mathbf{X}'\mathbf{X})^{-}\mathbf{S}'\mathbf{W}'\mathbf{y}].$$

Similarly, corresponding to any function $h(\mathbf{W}'\mathbf{y},\ \mathbf{y}'\mathbf{y})$, of $\mathbf{W}'\mathbf{y}$ and $\mathbf{y}'\mathbf{y}$, there is a function, say $h_*[\mathbf{X}'\mathbf{y},\ \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}]$, of $\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$ such that $h_*[\mathbf{X}'\mathbf{y},\ \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}] = h(\mathbf{W}'\mathbf{y},\ \mathbf{y}'\mathbf{y})$ for every value of $\mathbf{y}$; namely, the function $h_*[\mathbf{X}'\mathbf{y},\ \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}]$ defined by

$$h_*[\mathbf{X}'\mathbf{y},\ \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}] = h[\mathbf{R}'\mathbf{X}'\mathbf{y},\ \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y} + (\mathbf{X}'\mathbf{y})'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}].$$

Now, suppose that $\mathbf{W}'\mathbf{y}$ and $\mathbf{y}'\mathbf{y}$ form a complete sufficient statistic. Then, it follows from result (8.2) that $\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$ form a sufficient statistic. Moreover, if $\mathrm{E}\{g[\mathbf{X}'\mathbf{y},\ \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}]\} = 0$, then $\mathrm{E}[g_*(\mathbf{W}'\mathbf{y},\ \mathbf{y}'\mathbf{y})] = 0$, implying that $\Pr[g_*(\mathbf{W}'\mathbf{y},\ \mathbf{y}'\mathbf{y}) = 0] = 1$ and hence that $\Pr\{g[\mathbf{X}'\mathbf{y},\ \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}] = 0\} = 1$. Thus, $\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$ form a complete statistic.

Conversely, suppose that $\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$ form a complete sufficient statistic. Then, it follows from result (8.1) that $\mathbf{W}'\mathbf{y}$ and $\mathbf{y}'\mathbf{y}$ form a sufficient statistic. Moreover, if $\mathrm{E}[h(\mathbf{W}'\mathbf{y},\ \mathbf{y}'\mathbf{y})] = 0$, then $\mathrm{E}\{h_*[\mathbf{X}'\mathbf{y},\ \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}]\} = 0$, implying that $\Pr\{h_*[\mathbf{X}'\mathbf{y},\ \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}] = 0\} = 1$ and hence

that $\Pr[h(\mathbf{W}'\mathbf{y}, \mathbf{y}'\mathbf{y}) = 0] = 1$. Thus, $\mathbf{W}'\mathbf{y}$ and $\mathbf{y}'\mathbf{y}$ form a complete statistic.

At this point, we have established that $\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$ form a complete sufficient statistic if and only if $\mathbf{W}'\mathbf{y}$ and $\mathbf{y}'\mathbf{y}$ form a complete sufficient statistic. Thus, for purposes of verifying that $\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$ form a complete sufficient statistic, it suffices to consider the sufficiency and the completeness of the statistic formed by $\mathbf{W}'\mathbf{y}$ and $\mathbf{y}'\mathbf{y}$. In that regard, the probability density function of $\mathbf{y}$, say $f(\cdot)$, is expressible as follows:

$$\begin{aligned} f(\mathbf{y}) &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta})\right] \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right] \exp\left[-\frac{1}{2\sigma^2}\mathbf{y}'\mathbf{y} + \left(\frac{1}{\sigma^2}\mathbf{S}\boldsymbol{\beta}\right)'\mathbf{W}'\mathbf{y}\right]. \end{aligned} \tag{8.3}$$

Based on a well-known result on complete sufficient statistics for exponential families of distributions [a result that is theorem 2.74 in Schervish's (1995) book], it follows from result (8.3) that $\mathbf{W}'\mathbf{y}$ and $\mathbf{y}'\mathbf{y}$ form a complete sufficient statistic—to establish that the result on complete sufficient statistics for exponential families is applicable, it suffices to observe [in connection with expression (8.3)] that the parametric function $-1/(2\sigma^2)$ and the $(K \times 1)$ vector $(1/\sigma^2)\mathbf{S}\boldsymbol{\beta}$ of parametric functions are such that, for any (strictly) negative scalar c and any $K \times 1$ vector $\mathbf{d}$, $-1/(2\sigma^2) = c$ and $(1/\sigma^2)\mathbf{S}\boldsymbol{\beta} = \mathbf{d}$ for some value of σ^2 and some value of $\boldsymbol{\beta}$ (as is evident upon noting that $\mathbf{S}$ contains K linearly independent columns). It remains only to observe that since $\mathbf{W}'\mathbf{y}$ and $\mathbf{y}'\mathbf{y}$ form a complete sufficient statistic, $\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$ form a complete sufficient statistic.

5.9 Likelihood-Based Methods

A likelihood-based method, known as maximum likelihood (ML), can be used to estimate functions of the parameters (σ and the elements of $\boldsymbol{\beta}$) of the G–M or Aitken model. More generally, it can be used to estimate the parameters (the elements of $\boldsymbol{\beta}$ and of $\boldsymbol{\theta}$) of the general linear model. The use of this method requires an assumption that the distribution of the vector $\mathbf{e}$ of residual effects is known up to the value of σ (in the special case of a G–M or Aitken model) or up to the value of $\boldsymbol{\theta}$ (in the case of a general linear model). Typically, the distribution of $\mathbf{e}$ is taken to be MVN (multivariate normal).

a. (Ordinary) maximum likelihood estimation

It is convenient and instructive to begin by considering ML estimation in the relatively simple case of a G–M model.

G–M model. Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M model. Suppose further that the distribution of the vector $\mathbf{e}$ of residual effects is MVN. Then, $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

Let $f(\cdot\,; \boldsymbol{\beta}, \sigma)$ represent the probability density function (pdf) of the distribution of $\mathbf{y}$, and denote by $\underline{\mathbf{y}}$ the observed value of $\mathbf{y}$. Then, by definition, the likelihood function is the function, say $L(\boldsymbol{\beta}, \sigma; \underline{\mathbf{y}})$, of the parameters (which consist of σ and the elements of $\boldsymbol{\beta}$) defined (for $\boldsymbol{\beta} \in \Re^P$ and $\sigma > 0$) by $L(\boldsymbol{\beta}, \sigma; \underline{\mathbf{y}}) = f(\underline{\mathbf{y}}; \boldsymbol{\beta}, \sigma)$. Accordingly,

$$L(\boldsymbol{\beta}, \sigma; \underline{\mathbf{y}}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}(\underline{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})'(\underline{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})\right]. \tag{9.1}$$

And the log-likelihood function, say $\ell(\boldsymbol{\beta}, \sigma; \underset{\sim}{\mathbf{y}})$ [which, by definition, is the function obtained by equating $\ell(\boldsymbol{\beta}, \sigma; \underset{\sim}{\mathbf{y}})$ to the logarithm of the likelihood function, i.e., to $\log L(\boldsymbol{\beta}, \sigma; \underset{\sim}{\mathbf{y}})$], is expressible as

$$\ell(\boldsymbol{\beta}, \sigma; \underset{\sim}{\mathbf{y}}) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\underset{\sim}{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})'(\underset{\sim}{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}). \tag{9.2}$$

Now, consider the maximization of the likelihood function $L(\boldsymbol{\beta}, \sigma; \underset{\sim}{\mathbf{y}})$ or, equivalently, of the log-likelihood function $\ell(\boldsymbol{\beta}, \sigma; \underset{\sim}{\mathbf{y}})$. Irrespective of the value of σ, $\ell(\boldsymbol{\beta}, \sigma; \underset{\sim}{\mathbf{y}})$ attains its maximum value with respect to $\boldsymbol{\beta}$ at any value of $\boldsymbol{\beta}$ that minimizes $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Thus, in light of the results of Section 5.4b (on least squares minimization), $\ell(\boldsymbol{\beta}, \sigma; \underset{\sim}{\mathbf{y}})$ attains its maximum value with respect to $\boldsymbol{\beta}$ at a point $\tilde{\boldsymbol{\beta}}$ if and only if

$$\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{X}'\underset{\sim}{\mathbf{y}}, \tag{9.3}$$

that is, if and only if $\tilde{\boldsymbol{\beta}}$ is a solution to the normal equations.

Letting $\tilde{\boldsymbol{\beta}}$ represent any $P \times 1$ vector that satisfies condition (9.3), it remains to consider the maximization of $\ell(\tilde{\boldsymbol{\beta}}, \sigma; \underset{\sim}{\mathbf{y}})$ with respect to σ. In that regard, take $g(\sigma)$ to be a function of σ of the form

$$g(\sigma) = a - \frac{K}{2}\log\sigma^2 - \frac{c}{2\sigma^2}, \tag{9.4}$$

where a is a constant, c is a (strictly) positive constant, and K is a (strictly) positive integer. And observe that, unless $\underset{\sim}{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{0}$ (which is an event of probability 0), $\ell(\tilde{\boldsymbol{\beta}}, \sigma; \underset{\sim}{\mathbf{y}})$ is of the form (9.4); in the special case where $a = -(N/2)\log(2\pi)$, $c = (\underset{\sim}{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\underset{\sim}{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}})$, and $K = N$, $g(\sigma) = \ell(\tilde{\boldsymbol{\beta}}, \sigma; \underset{\sim}{\mathbf{y}})$. Clearly,

$$\frac{dg(\sigma)}{d\sigma} = -\frac{K}{\sigma} + \frac{c}{\sigma^3} = -\frac{K}{\sigma^3}\left(\sigma^2 - \frac{c}{K}\right).$$

Thus, $dg(\sigma)/d\sigma > 0$ if $\sigma^2 < c/K$, $dg(\sigma)/d\sigma = 0$ if $\sigma^2 = c/K$, and $dg(\sigma)/d\sigma < 0$ if $\sigma^2 > c/K$, so that $g(\sigma)$ is an increasing function of σ for $\sigma < \sqrt{c/K}$, is a decreasing function for $\sigma > \sqrt{c/K}$, and attains its maximum value at $\sigma = \sqrt{c/K}$.

Unless the model is of full rank (i.e., unless $\operatorname{rank}\mathbf{X} = P$), there are an infinite number of solutions to the normal equations and hence an infinite number of values of $\boldsymbol{\beta}$ that maximize $\ell(\boldsymbol{\beta}, \sigma; \underset{\sim}{\mathbf{y}})$. However, the value of an estimable linear combination $\boldsymbol{\lambda}'\boldsymbol{\beta}$ of the elements of $\boldsymbol{\beta}$ is the same for every value of $\boldsymbol{\beta}$ that maximizes $\ell(\boldsymbol{\beta}, \sigma; \underset{\sim}{\mathbf{y}})$—recall (from the results of Section 5.4 on the method of least squares) that $\boldsymbol{\lambda}'\tilde{\mathbf{b}}$ has the same value for every solution $\tilde{\mathbf{b}}$ to the normal equations.

In effect, we have established that the least squares estimator of any estimable linear combination of the elements of $\boldsymbol{\beta}$ is also the ML estimator. Moreover, since condition (9.3) can be satisfied by taking $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\underset{\sim}{\mathbf{y}}$, the ML estimator of σ^2 (the square root of which is the ML estimator of σ) is the estimator

$$\frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}{N}, \tag{9.5}$$

where $\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{P}_{\mathbf{X}}\mathbf{y}$. Like the unbiased estimator $\tilde{\mathbf{e}}'\tilde{\mathbf{e}}/(N - \operatorname{rank}\mathbf{X})$ and the Hodges–Lehmann estimator $\tilde{\mathbf{e}}'\tilde{\mathbf{e}}/[N - \operatorname{rank}(\mathbf{X}) + 2]$, the ML estimator of σ^2 is of the form (7.41).

A result on minimization and some results on matrices. As a preliminary to considering ML estimation as applied to a general linear model (or an Aitken model), it is convenient to establish the following result on minimization.

Theorem 5.9.1. Let $\mathbf{b}$ represent a $P \times 1$ vector of (unconstrained) variables, and define $f(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})'\mathbf{W}(\mathbf{y} - \mathbf{X}\mathbf{b})$, where $\mathbf{W}$ is an $N \times N$ symmetric nonnegative definite matrix, $\mathbf{X}$ is an $N \times P$ matrix, and $\mathbf{y}$ is an $N \times 1$ vector. Then, the linear system $\mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{W}\mathbf{y}$ (in $\mathbf{b}$) is consistent. Further, $f(\mathbf{b})$ attains its minimum value at a point $\tilde{\mathbf{b}}$ if and only if $\tilde{\mathbf{b}}$ is a solution to $\mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{W}\mathbf{y}$, in which case $f(\tilde{\mathbf{b}}) = \mathbf{y}'\mathbf{W}\mathbf{y} - \tilde{\mathbf{b}}'\mathbf{X}'\mathbf{W}\mathbf{y}$.

Proof. Let $\mathbf{R}$ represent a matrix such that $\mathbf{W} = \mathbf{R}'\mathbf{R}$—the existence of such a matrix is guaranteed by Corollary 2.13.25. Then, upon letting $\mathbf{t} = \mathbf{R}\mathbf{y}$ and $\mathbf{U} = \mathbf{R}\mathbf{X}$, $f(\mathbf{b})$ is expressible as $f(\mathbf{b}) = (\mathbf{t} - \mathbf{U}\mathbf{b})'(\mathbf{t} - \mathbf{U}\mathbf{b})$. Moreover, it follows from the results of Section 5.4b (on least squares minimization) that the linear system $\mathbf{U}'\mathbf{U}\mathbf{b} = \mathbf{U}'\mathbf{t}$ (in $\mathbf{b}$) is consistent and that $(\mathbf{t} - \mathbf{U}\mathbf{b})'(\mathbf{t} - \mathbf{U}\mathbf{b})$ attains its minimum value at a point $\tilde{\mathbf{b}}$ if and only if $\tilde{\mathbf{b}}$ is a solution to $\mathbf{U}'\mathbf{U}\mathbf{b} = \mathbf{U}'\mathbf{t}$, in which case

$$(\mathbf{t} - \mathbf{U}\tilde{\mathbf{b}})'(\mathbf{t} - \mathbf{U}\tilde{\mathbf{b}}) = \mathbf{t}'\mathbf{t} - \tilde{\mathbf{b}}'\mathbf{U}'\mathbf{t}.$$

It remains only to observe that $\mathbf{U}'\mathbf{U} = \mathbf{X}'\mathbf{W}\mathbf{X}$, that $\mathbf{U}'\mathbf{t} = \mathbf{X}'\mathbf{W}\mathbf{y}$, and that $\mathbf{t}'\mathbf{t} = \mathbf{y}'\mathbf{W}\mathbf{y}$. Q.E.D.

In addition to Theorem 5.9.1, it is convenient to have at our disposal the following lemma, which can be regarded as a generalization of Lemma 2.12.1.

Lemma 5.9.2. For any $N \times P$ matrix $\mathbf{X}$ and any $N \times N$ symmetric nonnegative definite matrix $\mathbf{W}$,

$$\mathcal{R}(\mathbf{X}'\mathbf{W}\mathbf{X}) = \mathcal{R}(\mathbf{W}\mathbf{X}), \quad \mathcal{C}(\mathbf{X}'\mathbf{W}\mathbf{X}) = \mathcal{C}(\mathbf{X}'\mathbf{W}), \quad \text{and} \quad \operatorname{rank}(\mathbf{X}'\mathbf{W}\mathbf{X}) = \operatorname{rank}(\mathbf{W}\mathbf{X}).$$

Proof. In light of Corollary 2.13.25, $\mathbf{W} = \mathbf{R}'\mathbf{R}$ for some matrix $\mathbf{R}$. And upon observing that $\mathbf{X}'\mathbf{W}\mathbf{X} = (\mathbf{R}\mathbf{X})'\mathbf{R}\mathbf{X}$ and making use of Corollary 2.4.4 and Lemma 2.12.1, we find that

$$\mathcal{R}(\mathbf{W}\mathbf{X}) = \mathcal{R}(\mathbf{R}'\mathbf{R}\mathbf{X}) \subset \mathcal{R}(\mathbf{R}\mathbf{X}) = \mathcal{R}[(\mathbf{R}\mathbf{X})'\mathbf{R}\mathbf{X}] = \mathcal{R}(\mathbf{X}'\mathbf{W}\mathbf{X}) \subset \mathcal{R}(\mathbf{W}\mathbf{X})$$

and hence that $\mathcal{R}(\mathbf{X}'\mathbf{W}\mathbf{X}) = \mathcal{R}(\mathbf{W}\mathbf{X})$. Moreover, that $\mathcal{R}(\mathbf{X}'\mathbf{W}\mathbf{X}) = \mathcal{R}(\mathbf{W}\mathbf{X})$ implies that $\operatorname{rank}(\mathbf{X}'\mathbf{W}\mathbf{X}) = \operatorname{rank}(\mathbf{W}\mathbf{X})$ and, in light of Lemma 2.4.6, that $\mathcal{C}(\mathbf{X}'\mathbf{W}\mathbf{X}) = \mathcal{C}(\mathbf{X}'\mathbf{W})$. Q.E.D.

In the special case of Lemma 5.9.2 where $\mathbf{W}$ is a (symmetric) positive definite matrix (and hence is nonsingular), it follows from Corollary 2.5.6 that $\mathcal{R}(\mathbf{W}\mathbf{X}) = \mathcal{R}(\mathbf{X})$, $\mathcal{C}(\mathbf{X}'\mathbf{W}) = \mathcal{C}(\mathbf{X}')$, and $\operatorname{rank}(\mathbf{W}\mathbf{X}) = \operatorname{rank}(\mathbf{X})$. Thus, we have the following corollary, which (like Lemma 5.9.2 itself) can be regarded as a generalization of Lemma 2.12.1.

Corollary 5.9.3. For any $N \times P$ matrix $\mathbf{X}$ and any $N \times N$ symmetric positive definite matrix $\mathbf{W}$,

$$\mathcal{R}(\mathbf{X}'\mathbf{W}\mathbf{X}) = \mathcal{R}(\mathbf{X}), \quad \mathcal{C}(\mathbf{X}'\mathbf{W}\mathbf{X}) = \mathcal{C}(\mathbf{X}'), \quad \text{and} \quad \operatorname{rank}(\mathbf{X}'\mathbf{W}\mathbf{X}) = \operatorname{rank}(\mathbf{X}).$$

As an additional corollary of Lemma 5.9.2, we have the following result.

Corollary 5.9.4. For any $N \times P$ matrix $\mathbf{X}$ and any $N \times N$ symmetric nonnegative definite matrix $\mathbf{W}$,

$$\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}\mathbf{X} = \mathbf{W}\mathbf{X} \quad \text{and} \quad \mathbf{X}'\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W} = \mathbf{X}'\mathbf{W}.$$

Proof. In light of Lemmas 5.9.2 and 2.4.3, $\mathbf{W}\mathbf{X} = \mathbf{L}'\mathbf{X}'\mathbf{W}\mathbf{X}$ for some $P \times N$ matrix $\mathbf{L}$. Thus,

$$\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}\mathbf{X} = \mathbf{L}'\mathbf{X}'\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}\mathbf{X} = \mathbf{L}'\mathbf{X}'\mathbf{W}\mathbf{X} = \mathbf{W}\mathbf{X}$$

and [since $\mathbf{X}'\mathbf{W} = (\mathbf{W}\mathbf{X})' = (\mathbf{L}'\mathbf{X}'\mathbf{W}\mathbf{X})' = \mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{L}$]

$$\mathbf{X}'\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W} = \mathbf{X}'\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{L} = \mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{L} = \mathbf{X}'\mathbf{W}.$$

Q.E.D.

General linear model. Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a general linear model. Suppose further that the distribution of the vector $\mathbf{e}$ of residual effects is MVN, so that $\mathbf{y} \sim N[\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\theta})]$. And suppose that $\mathbf{V}(\boldsymbol{\theta})$ is of rank N (for every $\boldsymbol{\theta} \in \Theta$).

Let us consider the ML estimation of functions of the model's parameters (which consist of the elements $\beta_1, \beta_2, \ldots, \beta_P$ of the vector $\boldsymbol{\beta}$ and the elements $\theta_1, \theta_2, \ldots, \theta_T$ of the vector $\boldsymbol{\theta}$). Let $f(\cdot; \boldsymbol{\beta}, \boldsymbol{\theta})$ represent the pdf of the distribution of $\mathbf{y}$, and denote by $\underline{\mathbf{y}}$ the observed value of $\mathbf{y}$. Then, the likelihood function is the function, say $L(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}})$, of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ defined (for $\boldsymbol{\beta} \in \mathcal{R}^P$ and $\boldsymbol{\theta} \in \Theta$) by $L(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}}) = f(\underline{\mathbf{y}}; \boldsymbol{\beta}, \boldsymbol{\theta})$. Accordingly,

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}}) = \frac{1}{(2\pi)^{N/2}|\mathbf{V}(\boldsymbol{\theta})|^{1/2}} \exp\left\{-\frac{1}{2}(\underline{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})'[\mathbf{V}(\boldsymbol{\theta})]^{-1}(\underline{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})\right\}. \tag{9.6}$$

And the log-likelihood function, say $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}})$, is expressible as

$$\ell(\boldsymbol{\beta},\boldsymbol{\theta};\underline{\mathbf{y}}) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2}(\underline{\mathbf{y}}-\mathbf{X}\boldsymbol{\beta})'[\mathbf{V}(\boldsymbol{\theta})]^{-1}(\underline{\mathbf{y}}-\mathbf{X}\boldsymbol{\beta}). \tag{9.7}$$

Maximum likelihood estimates are obtained by maximizing $L(\boldsymbol{\beta},\boldsymbol{\theta};\underline{\mathbf{y}})$ or, equivalently, $\ell(\boldsymbol{\beta},\boldsymbol{\theta};\underline{\mathbf{y}})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$: if $L(\boldsymbol{\beta},\boldsymbol{\theta};\underline{\mathbf{y}})$ or $\ell(\boldsymbol{\beta},\boldsymbol{\theta};\underline{\mathbf{y}})$ attains its maximum value at values $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\theta}}$ (of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively), then an ML estimate of a function, say $h(\boldsymbol{\beta},\boldsymbol{\theta})$, of $\boldsymbol{\beta}$ and/or $\boldsymbol{\theta}$ is provided by the quantity $h(\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\theta}})$ obtained by substituting $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\theta}}$ for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. In considering the maximization of the likelihood or log-likelihood function, it is helpful to begin by regarding the value of $\boldsymbol{\theta}$ as "fixed" and considering the maximization of the likelihood or log-likelihood function with respect to $\boldsymbol{\beta}$ alone.

Observe [in light of result (4.5.5) and Corollary 2.13.12] that (regardless of the value of $\boldsymbol{\theta}$) $[\mathbf{V}(\boldsymbol{\theta})]^{-1}$ is a symmetric positive definite matrix. Accordingly, it follows from Theorem 5.9.1 that for any particular value of $\boldsymbol{\theta}$, the linear system

$$\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\mathbf{b} = \mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\underline{\mathbf{y}} \tag{9.8}$$

(in the $P\times 1$ vector $\mathbf{b}$) is consistent. Further, $(\underline{\mathbf{y}}-\mathbf{X}\boldsymbol{\beta})'[\mathbf{V}(\boldsymbol{\theta})]^{-1}(\underline{\mathbf{y}}-\mathbf{X}\boldsymbol{\beta})$ attains its minimum value, or equivalently $-(1/2)(\underline{\mathbf{y}}-\mathbf{X}\boldsymbol{\beta})'[\mathbf{V}(\boldsymbol{\theta})]^{-1}(\underline{\mathbf{y}}-\mathbf{X}\boldsymbol{\beta})$ attains its maximum value, at a value $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ of $\boldsymbol{\beta}$ if and only if $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ is a solution to linear system (9.8), that is, if and only if

$$\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\underline{\mathbf{y}}, \tag{9.9}$$

in which case

$$\begin{aligned}\max_{\boldsymbol{\beta}\in\mathcal{R}^P} -\frac{1}{2}(\underline{\mathbf{y}}-\mathbf{X}\boldsymbol{\beta})'[\mathbf{V}(\boldsymbol{\theta})]^{-1}(\underline{\mathbf{y}}-\mathbf{X}\boldsymbol{\beta}) &\\ = -\frac{1}{2}[\underline{\mathbf{y}}-\mathbf{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})]'[\mathbf{V}(\boldsymbol{\theta})]^{-1}[\underline{\mathbf{y}}-\mathbf{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})] &\\ = -\frac{1}{2}\{\underline{\mathbf{y}}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\underline{\mathbf{y}} - [\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})]'\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\underline{\mathbf{y}}\}.&\end{aligned} \tag{9.10}$$

Now, suppose that (for $\boldsymbol{\theta}\in\Theta$) $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ satisfies condition (9.9). Then, for any matrix $\mathbf{A}$ such that $\mathcal{R}(\mathbf{A})\subset\mathcal{R}(\mathbf{X})$, the value of $\mathbf{A}\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ (at any particular value of $\boldsymbol{\theta}$) does not depend on the choice of $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$, as is evident upon observing (in light of Corollary 5.9.3) that $\mathbf{A} = \mathbf{T}(\boldsymbol{\theta})\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}$ for some matrix-valued function $\mathbf{T}(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ and hence that

$$\mathbf{A}\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \mathbf{T}(\boldsymbol{\theta})\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \mathbf{T}(\boldsymbol{\theta})\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\underline{\mathbf{y}}.$$

Thus, $\mathbf{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ does not depend on the choice of $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$, and for any estimable linear combination $\boldsymbol{\lambda}'\boldsymbol{\beta}$ of the elements of $\boldsymbol{\beta}$, $\boldsymbol{\lambda}'\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ does not depend on the choice of $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$. Among the possible choices for $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ are the vector $\{\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\}^{-}\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\underline{\mathbf{y}}$ and the vector $(\{\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\}^{-})'\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\underline{\mathbf{y}}$.

Define

$$L_*(\boldsymbol{\theta};\underline{\mathbf{y}}) = L[\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\theta};\underline{\mathbf{y}}] \quad\text{and}\quad \ell_*(\boldsymbol{\theta};\underline{\mathbf{y}}) = \ell[\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\theta};\underline{\mathbf{y}}]\ [=\log L_*(\boldsymbol{\theta};\underline{\mathbf{y}})]. \tag{9.11}$$

Then,

$$L_*(\boldsymbol{\theta};\underline{\mathbf{y}}) = \max_{\boldsymbol{\beta}\in\mathcal{R}^P} L(\boldsymbol{\beta},\boldsymbol{\theta};\underline{\mathbf{y}}) \quad\text{and}\quad \ell_*(\boldsymbol{\theta};\underline{\mathbf{y}}) = \max_{\boldsymbol{\beta}\in\mathcal{R}^P} \ell(\boldsymbol{\beta},\boldsymbol{\theta};\underline{\mathbf{y}}), \tag{9.12}$$

so that $L_*(\boldsymbol{\theta};\underline{\mathbf{y}})$ is a profile likelihood function and $\ell_*(\boldsymbol{\theta};\underline{\mathbf{y}})$ is a profile log-likelihood function—refer, e.g., to Severini (2000, sec 4.6) for the definition of a profile likelihood or profile log-likelihood function. Moreover,

$$\ell_*(\boldsymbol{\theta};\underline{\mathbf{y}}) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2}[\underline{\mathbf{y}}-\mathbf{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})]'[\mathbf{V}(\boldsymbol{\theta})]^{-1}[\underline{\mathbf{y}}-\mathbf{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})] \tag{9.13}$$

$$= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2}\{\underline{\mathbf{y}}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\underline{\mathbf{y}} - [\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})]'\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\underline{\mathbf{y}}\} \tag{9.14}$$

$$\begin{aligned}= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}(\boldsymbol{\theta})| &\\ -\frac{1}{2}\underline{\mathbf{y}}'\big([\mathbf{V}(\boldsymbol{\theta})]^{-1} - [\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\{\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\}^{-}\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\big)\underline{\mathbf{y}}.&\end{aligned} \tag{9.15}$$

Result (9.12) is significant from a computational standpoint. It "reduces" the problem of maximizing $L(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}})$ or $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ to that of maximizing $L_*(\boldsymbol{\theta}; \underline{\mathbf{y}})$ or $\ell_*(\boldsymbol{\theta}; \underline{\mathbf{y}})$ with respect to $\boldsymbol{\theta}$ alone. Values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ at which $L(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}})$ or $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}})$ attains its maximum value can be obtained by taking the value of $\boldsymbol{\theta}$ to be a value, say $\tilde{\boldsymbol{\theta}}$, at which $L_*(\boldsymbol{\theta}; \underline{\mathbf{y}})$ or $\ell_*(\boldsymbol{\theta}; \underline{\mathbf{y}})$ attains its maximum value and by then taking the value of $\boldsymbol{\beta}$ to be a solution $\tilde{\boldsymbol{\beta}}(\tilde{\boldsymbol{\theta}})$ to the linear system

$$\mathbf{X}'[\mathbf{V}(\tilde{\boldsymbol{\theta}})]^{-1}\mathbf{X}\mathbf{b} = \mathbf{X}'[\mathbf{V}(\tilde{\boldsymbol{\theta}})]^{-1}\underline{\mathbf{y}}$$

(in the $P \times 1$ vector $\mathbf{b}$).

In general, a solution to the problem of maximizing $\ell_*(\boldsymbol{\theta}; \underline{\mathbf{y}})$ is not obtainable in "closed form"; rather, the maximization must be accomplished numerically via an iterative procedure—the discussion of such procedures is deferred until later in the book. Nevertheless, there are special cases where the maximization of $\ell_*(\boldsymbol{\theta}; \underline{\mathbf{y}})$, and hence that of $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}})$, can be accomplished without resort to indirect (iterative) numerical methods. Indirect numerical methods are not needed in the special case where $\mathbf{y}$ follows a G–M model; that special case was discussed in Part 1 of the present subsection. More generally, indirect numerical methods are not needed in the special case where $\mathbf{y}$ follows an Aitken model, as is to be demonstrated in what follows.

Aitken model. Suppose that $\mathbf{y}$ follows an Aitken model (and that $\mathbf{H}$ is nonsingular and that the distribution of $\mathbf{e}$ is MVN). And regard the Aitken model as the special case of the general linear model where $T = 1$ (i.e., where $\boldsymbol{\theta}$ has only 1 element), where $\theta_1 = \sigma$, and where $\mathbf{V}(\boldsymbol{\theta}) = \sigma^2\mathbf{H}$. In that special case, linear system (9.8) is equivalent to the linear system

$$\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{H}^{-1}\underline{\mathbf{y}} \tag{9.16}$$

—the equivalence is in the sense that both linear systems have the same set of solutions. The equations comprising linear system (9.16) are known as the Aitken equations. When $\mathbf{H} = \mathbf{I}$ (i.e., when the model is a G–M model), the linear system (9.16) of Aitken equations simplifies to the linear system $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\underline{\mathbf{y}}$ of normal equations.

In this setting, we are free to choose the vector $\tilde{\beta}(\boldsymbol{\theta})$ in such a way that it has the same value for every value of $\boldsymbol{\theta}$. Accordingly, for every value of $\boldsymbol{\theta}$, take $\tilde{\beta}(\boldsymbol{\theta})$ to be $\tilde{\beta}$, where $\tilde{\beta}$ is any solution to the Aitken equations. Then, writing σ for $\boldsymbol{\theta}$, the profile log-likelihood function $\ell_*(\sigma; \underline{\mathbf{y}})$ is expressible as

$$\ell_*(\sigma; \underline{\mathbf{y}}) = \ell(\tilde{\beta}, \sigma; \underline{\mathbf{y}}) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2}\log|\mathbf{H}| - \frac{1}{2\sigma^2}(\underline{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{H}^{-1}(\underline{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}}).$$

Unless $\underline{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{0}$ (which is an event of probability 0), $\ell_*(\sigma; \underline{\mathbf{y}})$ is of the form of the function $g(\sigma)$ defined (in Part 1 of the present subsection) by equality (9.4); upon setting $a = -(N/2)\log(2\pi) - (1/2)\log|\mathbf{H}|$, $c = (\underline{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{H}^{-1}(\underline{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}})$, and $K = N$, $g(\sigma) = \ell_*(\sigma; \underline{\mathbf{y}})$. Thus, it follows from the results of Part 1 that $\ell_*(\sigma; \underline{\mathbf{y}})$ attains its maximum value when σ^2 equals

$$\frac{(\underline{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{H}^{-1}(\underline{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}})}{N}. \tag{9.17}$$

And we conclude that $\ell(\boldsymbol{\beta}, \sigma; \underline{\mathbf{y}})$ attains its maximum value when $\boldsymbol{\beta}$ equals $\tilde{\boldsymbol{\beta}}$ and when σ^2 equals the quantity (9.17). This conclusion serves to generalize the conclusion reached in Part 1, where it was determined that in the special case of the G–M model, the log-likelihood function attains its maximum value when β equals a solution, say $\tilde{\boldsymbol{\beta}}$, to the normal equations (i.e., to the linear system $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\underline{\mathbf{y}}$) and when σ^2 equals $(\underline{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\underline{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}})/N$.

b. Restricted or residual maximum likelihood estimation (REML estimation)

Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a general linear model. Suppose further that the distribution of the vector $\mathbf{e}$ of residual effects is MVN, so that $\mathbf{y} \sim N[\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\theta})]$.

And let $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}})$ represent the log-likelihood function [where $\underline{\mathbf{y}}$ is the observed value of $\mathbf{y}$ and where $\mathbf{V}(\boldsymbol{\theta})$ is assumed to be of rank N (for every $\boldsymbol{\theta} \in \Theta$)]. This function has the representation (9.7).

Suppose that $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\theta}}$ are values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ at which $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}})$ attains its maximum value. And observe that $\tilde{\boldsymbol{\theta}}$ is a value of $\boldsymbol{\theta}$ at which $\ell(\tilde{\boldsymbol{\beta}}, \boldsymbol{\theta}; \underline{\mathbf{y}})$ attains its maximum value. There is an implication that $\tilde{\boldsymbol{\theta}}$ is identical to the value of $\boldsymbol{\theta}$ that would be obtained from maximizing the likelihood function under a supposition that $\boldsymbol{\beta}$ is a known $(P \times 1)$ vector (rather than a vector of unknown parameters) and under the further supposition that $\boldsymbol{\beta}$ equals $\tilde{\boldsymbol{\beta}}$ (or, perhaps more precisely, $\mathbf{X}\boldsymbol{\beta}$ equals $\mathbf{X}\tilde{\boldsymbol{\beta}}$). Thus, in a certain sense, maximum likelihood estimators of functions of $\boldsymbol{\theta}$ fail to account for the estimation of $\boldsymbol{\beta}$. This failure can be disconcerting and can have undesirable consequences.

It is informative to consider the manifestation of this phenomenon in the relatively simple special case of a G–M model. In that special case, the use of maximum likelihood estimation results in σ^2 being estimated by the quantity (9.5), in which the residual sum of squares is divided by N rather than by $N - \operatorname{rank}\mathbf{X}$ as in the case of the unbiased estimator [or by $N - \operatorname{rank}(\mathbf{X}) + 2$ as in the case of the Hodges–Lehmann estimator].

The failure of ML estimators of functions of $\boldsymbol{\theta}$ to account for the estimation of $\boldsymbol{\beta}$ has led to the widespread use of a variant of maximum likelihood that has come to be known by the acronym REML (which is regarded by some as standing for restricted maximum likelihood and by others as standing for residual maximum likelihood). In REML, inferences about functions of $\boldsymbol{\theta}$ are based on the likelihood function associated with a vector of what are sometimes called error contrasts.

An *error contrast* is a linear unbiased estimator of 0, that is, a linear combination, say $\mathbf{r}'\mathbf{y}$, of the elements of $\mathbf{y}$ such that $\mathrm{E}(\mathbf{r}'\mathbf{y}) = 0$ or, equivalently, such that $\mathbf{X}'\mathbf{r} = \mathbf{0}$. Thus, $\mathbf{r}'\mathbf{y}$ is an error contrast if and only if $\mathbf{r} \in \mathfrak{N}(\mathbf{X}')$. Moreover, in light of Lemma 2.11.5,

$$\dim[\mathfrak{N}(\mathbf{X}'] = N - \operatorname{rank}(\mathbf{X}') = N - \operatorname{rank}\mathbf{X}.$$

And it follows that there exists a set of $N - \operatorname{rank}\mathbf{X}$ linearly independent error contrasts and that no set of error contrasts contains more than $N - \operatorname{rank}\mathbf{X}$ linearly independent error contrasts.

Accordingly, let $\mathbf{R}$ represent an $N \times (N - \operatorname{rank}\mathbf{X})$ matrix (of constants) of full column rank $N - \operatorname{rank}\mathbf{X}$ such that $\mathbf{X}'\mathbf{R} = \mathbf{0}$ [or, equivalently, an $N \times (N - \operatorname{rank}\mathbf{X})$ matrix whose columns are linearly independent members of the null space $\mathfrak{N}(\mathbf{X}')$ of $\mathbf{X}'$]. And take $\mathbf{z}$ to be the $(N - \operatorname{rank}\mathbf{X}) \times 1$ vector defined by $\mathbf{z} = \mathbf{R}'\mathbf{y}$ (so that the elements of $\mathbf{z}$ are $N - \operatorname{rank}\mathbf{X}$ linearly independent error contrasts). Then, $\mathbf{z} \sim N[\mathbf{0}, \mathbf{R}'\mathbf{V}(\boldsymbol{\theta})\mathbf{R}]$, and [in light of the assumption that $\mathbf{V}(\boldsymbol{\theta})$ is nonsingular and in light of Theorem 2.13.10] $\mathbf{R}'\mathbf{V}(\boldsymbol{\theta})\mathbf{R}$ is nonsingular. Further, let $f(\cdot; \boldsymbol{\theta})$ represent the pdf of the distribution of $\mathbf{z}$, and take $L(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ to be the function of $\boldsymbol{\theta}$ defined (for $\boldsymbol{\theta} \in \Theta$) by $L(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}}) = f(\mathbf{R}'\underline{\mathbf{y}}; \boldsymbol{\theta})$. The function $L(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ is a likelihood function; it is the likelihood function obtained by regarding the observed value of $\mathbf{z}$ as the data vector. In REML, the inferences about functions of $\boldsymbol{\theta}$ are based on the likelihood function $L(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ [or on a likelihood function that is equivalent to $L(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ in the sense that it differs from $L(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ by no more than a multiplicative constant].

It is worth noting that the use of REML results in the same inferences regardless of the choice of the matrix $\mathbf{R}$. To see that REML has this property, let $\mathbf{R}_1$ and $\mathbf{R}_2$ represent any two choices for $\mathbf{R}$, that is, take $\mathbf{R}_1$ and $\mathbf{R}_2$ to be any two $N \times (N - \operatorname{rank}\mathbf{X})$ matrices of full column rank such that $\mathbf{X}'\mathbf{R}_1 = \mathbf{X}'\mathbf{R}_2 = \mathbf{0}$. Further, define $\mathbf{z}_1 = \mathbf{R}_1'\mathbf{y}$ and $\mathbf{z}_2 = \mathbf{R}_2'\mathbf{y}$. And let $f_1(\cdot; \boldsymbol{\theta})$ represent the pdf of the distribution of $\mathbf{z}_1$ and $f_2(\cdot; \boldsymbol{\theta})$ the pdf of the distribution of $\mathbf{z}_2$; and take $L_1(\boldsymbol{\theta}; \mathbf{R}_1'\underline{\mathbf{y}})$ and $L_2(\boldsymbol{\theta}; \mathbf{R}_2'\underline{\mathbf{y}})$ to be the functions of $\boldsymbol{\theta}$ defined by $L_1(\boldsymbol{\theta}; \mathbf{R}_1'\underline{\mathbf{y}}) = f_1(\mathbf{R}_1'\underline{\mathbf{y}}; \boldsymbol{\theta})$ and $L_2(\boldsymbol{\theta}; \mathbf{R}_2'\underline{\mathbf{y}}) = f_2(\mathbf{R}_2'\underline{\mathbf{y}}; \boldsymbol{\theta})$.

There exists an $(N - \operatorname{rank}\mathbf{X}) \times (N - \operatorname{rank}\mathbf{X})$ matrix $\mathbf{A}$ such that $\mathbf{R}_2 = \mathbf{R}_1\mathbf{A}$, as is evident upon observing that the columns of each of the two matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ form a basis for the $(N - \operatorname{rank}\mathbf{X})$-dimensional linear space $\mathfrak{N}(\mathbf{X}')$; necessarily, $\mathbf{A}$ is nonsingular. Moreover, the pdf's of the distributions of $\mathbf{z}_1$ and $\mathbf{z}_2$ are such that (for every value of $\mathbf{z}_1$)

$$f_1(\mathbf{z}_1) = |\det \mathbf{A}| f_2(\mathbf{A}'\mathbf{z}_1)$$

—this relationship can be verified directly from formula (3.5.32) for the pdf of an MVN distribution or simply by observing that $\mathbf{z}_2 = \mathbf{A}'\mathbf{z}_1$ and making use of standard results (e.g., Bickel and Doksum 2001, sec. B.2) on a change of variables. Thus,

$$L_2(\boldsymbol{\theta}; \mathbf{R}_2'\underset{\sim}{\mathbf{y}}) = f_2(\mathbf{R}_2'\underset{\sim}{\mathbf{y}}; \boldsymbol{\theta}) = f_2(\mathbf{A}'\mathbf{R}_1'\underset{\sim}{\mathbf{y}}; \boldsymbol{\theta}) = |\det \mathbf{A}|^{-1} f_1(\mathbf{R}_1'\underset{\sim}{\mathbf{y}}; \boldsymbol{\theta}) = |\det \mathbf{A}|^{-1} L_1(\boldsymbol{\theta}; \mathbf{R}_1'\underset{\sim}{\mathbf{y}}).$$

We conclude that the two likelihood functions $L_1(\boldsymbol{\theta}; \mathbf{R}_1'\underset{\sim}{\mathbf{y}})$ and $L_2(\boldsymbol{\theta}; \mathbf{R}_2'\underset{\sim}{\mathbf{y}})$ differ from each other by no more than a multiplicative constant and hence that they are equivalent.

The $(N - \operatorname{rank}\mathbf{X})$-dimensional vector $\mathbf{z} = \mathbf{R}'\mathbf{y}$ of error contrasts is translation invariant, as is evident upon observing that for every $P \times 1$ vector $\mathbf{k}$ (and every value of $\mathbf{y}$),

$$\mathbf{R}'(\mathbf{y} + \mathbf{X}\mathbf{k}) = \mathbf{R}'\mathbf{y} + (\mathbf{X}'\mathbf{R})'\mathbf{k} = \mathbf{R}'\mathbf{y} + \mathbf{0}\mathbf{k} = \mathbf{R}'\mathbf{y}.$$

In fact, $\mathbf{z}$ is a maximal invariant: in the present context, a (possibly vector-valued) statistic $\mathbf{h}(\mathbf{y})$ is said to be a *maximal invariant* if it is invariant and if corresponding to each pair of values $\mathbf{y}_1$ and $\mathbf{y}_2$ of $\mathbf{y}$ such that $\mathbf{h}(\mathbf{y}_2) = \mathbf{h}(\mathbf{y}_1)$, there exists a $P \times 1$ vector $\mathbf{k}$ such that $\mathbf{y}_2 = \mathbf{y}_1 + \mathbf{X}\mathbf{k}$—refer, e.g., to Lehmann and Romano (2005b, sec 6.2) for a general definition (of a maximal invariant).

To confirm that $\mathbf{z}$ is a maximal invariant, take $\mathbf{y}_1$ and $\mathbf{y}_2$ to be any pair of values of $\mathbf{y}$ such that $\mathbf{R}'\mathbf{y}_2 = \mathbf{R}'\mathbf{y}_1$. And observe that $\mathbf{y}_2 = \mathbf{y}_1 + (\mathbf{y}_2 - \mathbf{y}_1)$ and that $\mathbf{y}_2 - \mathbf{y}_1 \in \mathfrak{N}(\mathbf{R}')$. Observe also (in light of Lemma 2.11.5) that $\dim[\mathfrak{N}(\mathbf{R}')] = \operatorname{rank}\mathbf{X}$. Moreover, $\mathbf{R}'\mathbf{X} = (\mathbf{X}'\mathbf{R})' = \mathbf{0}$, implying (in light of Lemma 2.4.2) that $\mathcal{C}(\mathbf{X}) \subset \mathfrak{N}(\mathbf{R}')$ and hence (in light of Theorem 2.4.10) that $\mathcal{C}(\mathbf{X}) = \mathfrak{N}(\mathbf{R}')$. Thus, the linear space $\mathfrak{N}(\mathbf{R}')$ is spanned by the columns of $\mathbf{X}$, leading to the conclusion that there exists a $P \times 1$ vector $\mathbf{k}$ such that $\mathbf{y}_2 - \mathbf{y}_1 = \mathbf{X}\mathbf{k}$ and hence such that $\mathbf{y}_2 = \mathbf{y}_1 + \mathbf{X}\mathbf{k}$.

That $\mathbf{z} = \mathbf{R}'\mathbf{y}$ is a maximal invariant is of interest because any maximal invariant, say $\mathbf{h}(\mathbf{y})$, has (in the present context) the following property: a (possibly vector-valued) statistic, say $\mathbf{g}(\mathbf{y})$, is translation invariant if and only if $\mathbf{g}(\mathbf{y})$ depends on the value of $\mathbf{y}$ only through $\mathbf{h}(\mathbf{y})$, that is, if and only if there exists a function $\mathbf{s}(\cdot)$ such that $\mathbf{g}(\mathbf{y}) = \mathbf{s}[\mathbf{h}(\mathbf{y})]$ (for every value of $\mathbf{y}$). To see that $\mathbf{h}(\mathbf{y})$ has this property, observe that if [for some function $\mathbf{s}(\cdot)$] $\mathbf{g}(\mathbf{y}) = \mathbf{s}[\mathbf{h}(\mathbf{y})]$ (for every value of $\mathbf{y}$), then (for every $P \times 1$ vector $\mathbf{k}$)

$$\mathbf{g}(\mathbf{y} + \mathbf{X}\mathbf{k}) = \mathbf{s}[\mathbf{h}(\mathbf{y} + \mathbf{X}\mathbf{k})] = \mathbf{s}[\mathbf{h}(\mathbf{y})] = \mathbf{g}(\mathbf{y}),$$

so that $\mathbf{g}(\mathbf{y})$ is translation invariant. Conversely, if $\mathbf{g}(\mathbf{y})$ is translation invariant and if $\mathbf{y}_1$ and $\mathbf{y}_2$ are any pair of values of $\mathbf{y}$ such that $\mathbf{h}(\mathbf{y}_2) = \mathbf{h}(\mathbf{y}_1)$, then $\mathbf{y}_2 = \mathbf{y}_1 + \mathbf{X}\mathbf{k}$ for some vector $\mathbf{k}$ and, consequently, $\mathbf{g}(\mathbf{y}_2) = \mathbf{g}(\mathbf{y}_1 + \mathbf{X}\mathbf{k}) = \mathbf{g}(\mathbf{y}_1)$.

The vector $\mathbf{z}$ consists of $N - \operatorname{rank}\mathbf{X}$ linearly independent linear combinations of the elements of the $N \times 1$ vector $\mathbf{y}$. Suppose that we introduce an additional $\operatorname{rank}\mathbf{X}$ linear combinations in the form of the $(\operatorname{rank}\mathbf{X}) \times 1$ vector $\mathbf{u}$ defined by $\mathbf{u} = \mathbf{X}_*'\mathbf{y}$, where $\mathbf{X}_*$ is any $N \times (\operatorname{rank}\mathbf{X})$ matrix (of constants) whose columns are linearly independent columns of $\mathbf{X}$ or, more generally, whose columns form a basis for $\mathcal{C}(\mathbf{X})$. Then,

$$\begin{pmatrix} \mathbf{u} \\ \mathbf{z} \end{pmatrix} = (\mathbf{X}_*, \mathbf{R})'\mathbf{y}.$$

And (since $\mathbf{X}_* = \mathbf{X}\mathbf{A}$ for some matrix $\mathbf{A}$)

$$\begin{aligned} \operatorname{rank}(\mathbf{X}_*, \mathbf{R}) &= \operatorname{rank}[(\mathbf{X}_*, \mathbf{R})'(\mathbf{X}_*, \mathbf{R})] \\ &= \operatorname{rank}\operatorname{diag}(\mathbf{X}_*'\mathbf{X}_*, \mathbf{R}'\mathbf{R}) \\ &= \operatorname{rank}(\mathbf{X}_*'\mathbf{X}_*) + \operatorname{rank}(\mathbf{R}'\mathbf{R}) \\ &= \operatorname{rank}\mathbf{X}_* + \operatorname{rank}\mathbf{R} \\ &= \operatorname{rank}(\mathbf{X}) + N - \operatorname{rank}(\mathbf{X}) = N. \end{aligned} \tag{9.18}$$

Accordingly, the likelihood function that would result from regarding the observed value $(\mathbf{X}_*, \mathbf{R})'\underset{\sim}{\mathbf{y}}$ of $\begin{pmatrix} \mathbf{u} \\ \mathbf{z} \end{pmatrix}$ as the data vector differs by no more than a multiplicative constant from that obtained by

regarding the observed value $\underline{\mathbf{y}}$ of $\mathbf{y}$ as the data vector (as can be readily verified). When viewed in this context, the likelihood function that is employed in REML can be regarded as what is known as a marginal likelihood—refer, e.g., to Pawitan (2001, sec. 10.3) for the definition of a marginal likelihood.

The vector $\tilde{\mathbf{e}} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$ [where $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$] is the vector of (least squares) residuals. Observe [in light of Theorem 2.12.2 and Lemma 2.8.4] that $\mathbf{X}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) = \mathbf{0}$ and that

$$\operatorname{rank}(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) = N - \operatorname{rank}\mathbf{P}_{\mathbf{X}} = N - \operatorname{rank}\mathbf{X}. \tag{9.19}$$

Thus, among the choices for the $N \times (N - \operatorname{rank}\mathbf{X})$ matrix $\mathbf{R}$ (of full column rank $N - \operatorname{rank}\mathbf{X}$ such that $\mathbf{X}'\mathbf{R} = \mathbf{0}$) is any $N \times (N - \operatorname{rank}\mathbf{X})$ matrix whose columns are a linearly independent subset of the columns of the (symmetric) matrix $\mathbf{I} - \mathbf{P}_{\mathbf{X}}$. For any such choice of $\mathbf{R}$, the elements of the $(N - \operatorname{rank}\mathbf{X}) \times 1$ vector $\mathbf{z} = \mathbf{R}'\mathbf{y}$ consist of linearly independent (least squares) residuals.

The letters R and E in the acronym REML can be regarded as representing either restricted or residual. REML is restricted ML in the sense that in the formation of the likelihood function, the data are restricted to those inherent in the values of the $N - \operatorname{rank}\mathbf{X}$ linearly independent error contrasts. REML is residual ML in the sense that the $N - \operatorname{rank}\mathbf{X}$ linearly independent error contrasts can be taken to be (least squares) residuals.

It might seem as though the use of REML would result in the loss of some information about functions of $\boldsymbol{\theta}$. However, in at least one regard, there is no loss of information. Consider the profile likelihood function $L_*(\cdot\,;\underline{\mathbf{y}})$ or profile log-likelihood function $\ell_*(\cdot\,;\underline{\mathbf{y}})$ of definition (9.11)—the (ordinary) ML estimate of a function of $\boldsymbol{\theta}$ is obtained from a value of $\boldsymbol{\theta}$ at which $L_*(\boldsymbol{\theta};\underline{\mathbf{y}})$ or $\ell_*(\boldsymbol{\theta};\underline{\mathbf{y}})$ attains its maximum value. The identity of the function $L_*(\cdot\,;\underline{\mathbf{y}})$ or, equivalently, that of the function $\ell_*(\cdot\,;\underline{\mathbf{y}})$ can be determined solely from knowledge of the observed value $\mathbf{R}'\underline{\mathbf{y}}$ of the vector $\mathbf{z}$ of error contrasts; complete knowledge of the observed value $\underline{\mathbf{y}}$ of $\mathbf{y}$ is not required. Thus, the (ordinary) ML estimator of a function of $\boldsymbol{\theta}$ (like the REML estimator) depends on the value of $\mathbf{y}$ only through the value of the vector of error contrasts.

Let us verify that the identity of the function $\ell_*(\cdot\,;\underline{\mathbf{y}})$ is determinable solely from knowledge of $\mathbf{R}'\underline{\mathbf{y}}$. Let $\underline{\tilde{\mathbf{e}}} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\underline{\mathbf{y}}$, and observe (in light of Theorem 2.12.2) that $\mathbf{X}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})' = \mathbf{0}$, implying [since the columns of $\mathbf{R}$ form a basis for $\mathfrak{N}(\mathbf{X}')$] that $(\mathbf{I} - \mathbf{P}_{\mathbf{X}})' = \mathbf{R}\mathbf{K}$ for some matrix $\mathbf{K}$ and hence that

$$\underline{\tilde{\mathbf{e}}} = (\mathbf{R}\mathbf{K})'\underline{\mathbf{y}} = \mathbf{K}'\mathbf{R}'\underline{\mathbf{y}} \tag{9.20}$$

—$\underline{\tilde{\mathbf{e}}}$ is the observed value of the vector $\tilde{\mathbf{e}} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$. Moreover, upon observing [in light of result (2.5.5) and Corollary 2.13.12] that $[\mathbf{V}(\boldsymbol{\theta})]^{-1}$ is a symmetric positive definite matrix, it follows from Corollary 5.9.4 that

$$\big([\mathbf{V}(\boldsymbol{\theta})]^{-1} - [\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\{\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\}^{-}\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\big)\mathbf{X} = \mathbf{0}$$

and that

$$\mathbf{X}'\big([\mathbf{V}(\boldsymbol{\theta})]^{-1} - [\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\{\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\}^{-}\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\big) = \mathbf{0}.$$

And as a consequence, formula (9.15) for $\ell_*(\boldsymbol{\theta};\underline{\mathbf{y}})$ can be reexpressed as follows:

$$\ell_*(\boldsymbol{\theta};\underline{\mathbf{y}}) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2}\underline{\tilde{\mathbf{e}}}'\big([\mathbf{V}(\boldsymbol{\theta})]^{-1} - [\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\{\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\}^{-}\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\big)\underline{\tilde{\mathbf{e}}}. \tag{9.21}$$

Together, results (9.21) and (9.20) imply that the identity of the function $\ell_*(\cdot\,;\underline{\mathbf{y}})$ is determinable solely from knowledge of $\mathbf{R}'\underline{\mathbf{y}}$.

Some results on symmetric idempotent matrices and on null spaces. As a preliminary to considering REML in the special case of a G–M model, it is helpful to establish the following three results on symmetric idempotent matrices and on null spaces.

Theorem 5.9.5. Every symmetric idempotent matrix is nonnegative definite. Moreover, if $\mathbf{A}$ is an $N \times N$ symmetric idempotent matrix of rank $R > 0$, then there exists an $N \times R$ matrix $\mathbf{Q}$ such that $\mathbf{A} = \mathbf{QQ}'$, and, for any such $N \times R$ matrix $\mathbf{Q}$, $\text{rank}\,\mathbf{Q} = R$ and $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$. And, conversely, for any $N \times R$ matrix $\mathbf{Q}$ such that $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$, $\mathbf{QQ}'$ is an $N \times N$ symmetric idempotent matrix of rank R.

Proof. Suppose that $\mathbf{A}$ is an $N \times N$ symmetric idempotent matrix of rank R (≥ 0). Then, $\mathbf{A} = \mathbf{A}^2 = \mathbf{A}'\mathbf{A}$, and it follows from Corollary 2.13.15 that $\mathbf{A}$ is nonnegative definite. Moreover, assuming that $R > 0$, it follows from Corollary 2.13.23 that there exists an $N \times R$ matrix $\mathbf{Q}$ such that $\mathbf{A} = \mathbf{QQ}'$. And for any such $N \times R$ matrix $\mathbf{Q}$, we find, in light of Lemma 2.12.1 [and result (2.4.1)], that $\text{rank}\,\mathbf{Q} = R$ and that $\mathbf{Q}'\mathbf{Q}$ is nonsingular and, in addition, we find that

$$\mathbf{Q}'\mathbf{QQ}'\mathbf{QQ}'\mathbf{Q} = \mathbf{Q}'\mathbf{A}^2\mathbf{Q} = \mathbf{Q}'\mathbf{AQ} = \mathbf{Q}'\mathbf{QQ}'\mathbf{Q} \tag{9.22}$$

and hence [upon premultiplying and postmultiplying both sides of equality (9.22) by $(\mathbf{Q}'\mathbf{Q})^{-1}$] that $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$.

Conversely, suppose that $\mathbf{Q}$ is an $N \times R$ matrix such that $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$. Then, upon observing that $\mathbf{QQ}' = \mathbf{P}_{\mathbf{Q}}$ and (in light of Lemma 2.12.1) that

$$\text{rank}(\mathbf{Q}) = \text{rank}(\mathbf{Q}'\mathbf{Q}) = \text{rank}(\mathbf{I}_R) = R,$$

it follows from Theorem 2.12.2 that $\mathbf{QQ}'$ is a symmetric idempotent matrix of rank R. Q.E.D.

Theorem 5.9.6. Let $\mathbf{X}$ represent an $N \times P$ matrix of rank R ($< N$). Then, $\mathbf{I} - \mathbf{P}_{\mathbf{X}}$ is an $N \times N$ symmetric idempotent matrix of rank $N-R$, and there exists an $N \times (N-R)$ matrix $\mathbf{Q}$ such that $\mathbf{I} - \mathbf{P}_{\mathbf{X}} = \mathbf{QQ}'$. Moreover, for any $N \times (N-R)$ matrix $\mathbf{Q}$, $\mathbf{I} - \mathbf{P}_{\mathbf{X}} = \mathbf{QQ}'$ if and only if $\mathbf{X}'\mathbf{Q} = \mathbf{0}$ and $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$ (in which case $\mathbf{Q}$ is of full column rank $N-R$).

Proof. In light of Theorem 2.12.2 and Lemmas 2.8.1 and 2.8.4, it is clear that $\mathbf{I}-\mathbf{P}_{\mathbf{X}}$ is a symmetric idempotent matrix of rank $N-R$. And in light of Theorem 5.9.5, there exists an $N \times (N-R)$ matrix $\mathbf{Q}$ such that $\mathbf{I} - \mathbf{P}_{\mathbf{X}} = \mathbf{QQ}'$.

Now, suppose that $\mathbf{Q}$ is any $N \times (N-R)$ matrix such that $\mathbf{X}'\mathbf{Q} = \mathbf{0}$ and $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$. Then, $\mathbf{QQ}'$ is a symmetric idempotent matrix of rank $N-R$ (as is evident from Theorem 5.9.5), and $\mathbf{P}_{\mathbf{X}}\mathbf{Q} = \mathbf{0}$ and $\mathbf{Q}'\mathbf{P}_{\mathbf{X}} = \mathbf{0}$. And, consequently, $\mathbf{I}-\mathbf{P}_{\mathbf{X}}-\mathbf{QQ}'$ is a symmetric idempotent matrix. Further, making use of Corollary 2.8.3 and Lemma 2.12.1, we find that

$$\begin{aligned}\text{rank}(\mathbf{I} - \mathbf{P}_{\mathbf{X}} - \mathbf{QQ}') &= \text{tr}(\mathbf{I} - \mathbf{P}_{\mathbf{X}} - \mathbf{QQ}') \\ &= \text{tr}(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) - \text{tr}(\mathbf{QQ}') \\ &= \text{rank}(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) - \text{rank}(\mathbf{QQ}') \\ &= N-R-(N-R) = 0,\end{aligned}$$

implying that $\mathbf{I} - \mathbf{P}_{\mathbf{X}} - \mathbf{QQ}' = \mathbf{0}$ and hence that $\mathbf{I} - \mathbf{P}_{\mathbf{X}} = \mathbf{QQ}'$.

Conversely, suppose that $\mathbf{Q}$ is any $N \times (N-R)$ matrix such that $\mathbf{I} - \mathbf{P}_{\mathbf{X}} = \mathbf{QQ}'$. Then, according to Theorem 5.9.5, $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$. Moreover, making use of Theorem 2.12.2, we find that

$$\mathbf{X}'\mathbf{QQ}' = \mathbf{X}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) = \mathbf{0},$$

implying (in light of Corollary 2.3.4) that $\mathbf{X}'\mathbf{Q} = \mathbf{0}$. Q.E.D.

Lemma 5.9.7. Let $\mathbf{X}$ represent an $N \times P$ matrix of rank R ($< N$). Then, for any $N \times (N-R)$ matrix $\mathbf{Q}$, $\mathbf{X}'\mathbf{Q} = \mathbf{0}$ and $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$ if and only if the columns of $\mathbf{Q}$ form an orthonormal basis for $\mathfrak{N}(\mathbf{X}')$.

Proof. If the columns of $\mathbf{Q}$ form an orthonormal basis for $\mathfrak{N}(\mathbf{X}')$, then clearly $\mathbf{X}'\mathbf{Q} = \mathbf{0}$ and $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$. Conversely, suppose that $\mathbf{X}'\mathbf{Q} = \mathbf{0}$ and $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$. Then, clearly, the $N-R$ columns of $\mathbf{Q}$ are orthonormal, and each of them is contained in $\mathfrak{N}(\mathbf{X}')$. And since orthonormal vectors are linearly independent (as is evident from Lemma 2.4.22) and since (according to Lemma 2.11.5)

$\dim[\mathfrak{N}(\mathbf{X}')] = N - R$, it follows from Theorem 2.4.11 that the columns of $\mathbf{Q}$ form a basis for $\mathfrak{N}(\mathbf{X}')$. Q.E.D.

REML in the special case of a G–M model. Let us consider REML in the special case where the $N \times 1$ observable random vector $\mathbf{y}$ follows a G–M model. And in doing so, let us continue to suppose that the distribution of the vector $\mathbf{e}$ of residual effects is MVN. Then, $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

What is the REML estimator of σ^2, and how does it compare with other estimators of σ^2, including the (ordinary) ML estimator (which was derived in Subsection a)? These questions can be readily answered by making a judicious choice for the $N \times (N - \operatorname{rank}\mathbf{X})$ matrix $\mathbf{R}$ (of full column rank $N - \operatorname{rank}\mathbf{X}$) such that $\mathbf{X}'\mathbf{R} = \mathbf{0}$.

Let $\mathbf{Q}$ represent an $N \times (N - \operatorname{rank}\mathbf{X})$ matrix whose columns form an orthonormal basis for $\mathfrak{N}(\mathbf{X}')$. Or, equivalently (in light of Lemma 5.9.7), take $\mathbf{Q}$ to be an $N \times (N - \operatorname{rank}\mathbf{X})$ matrix such that $\mathbf{X}'\mathbf{Q} = \mathbf{0}$ and $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$. And observe (in light of Theorem 5.9.6) that

$$\mathbf{I} - \mathbf{P}_{\mathbf{X}} = \mathbf{Q}\mathbf{Q}'$$

(and that $\mathbf{Q}$ is of full column rank).

Suppose that in implementing REML, we set $\mathbf{R} = \mathbf{Q}$—clearly, that is a legitimate choice for $\mathbf{R}$. Then, $\mathbf{z} = \mathbf{Q}'\mathbf{y} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. And, letting $\underline{\mathbf{y}}$ represent the observed value of $\mathbf{y}$, the log-likelihood function that results from regarding the observed value $\mathbf{Q}'\underline{\mathbf{y}}$ of $\mathbf{z}$ as the data vector is the function $\ell(\sigma, \mathbf{Q}'\underline{\mathbf{y}})$ of σ given by

$$\begin{aligned}
\ell(\sigma, \mathbf{Q}'\underline{\mathbf{y}}) &= -\frac{N - \operatorname{rank}\mathbf{X}}{2}\log(2\pi) - \frac{1}{2}\log\left|\sigma^2\mathbf{I}_{N-\operatorname{rank}\mathbf{X}}\right| - \frac{1}{2}\underline{\mathbf{y}}'\mathbf{Q}(\sigma^2\mathbf{I})^{-1}\mathbf{Q}'\underline{\mathbf{y}} \\
&= -\frac{N - \operatorname{rank}\mathbf{X}}{2}\log(2\pi) - \frac{N - \operatorname{rank}\mathbf{X}}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\underline{\mathbf{y}}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\underline{\mathbf{y}} \\
&= -\frac{N - \operatorname{rank}\mathbf{X}}{2}\log(2\pi) - \frac{N - \operatorname{rank}\mathbf{X}}{2}\log\sigma^2 - \frac{1}{2\sigma^2}[(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\underline{\mathbf{y}}]'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\underline{\mathbf{y}}. \qquad (9.23)
\end{aligned}$$

Unless $(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\underline{\mathbf{y}} = \mathbf{0}$ (which is an event of probability 0), $\ell(\sigma, \mathbf{Q}'\underline{\mathbf{y}})$ is of the form of the function $g(\sigma)$ defined by equality (9.4); upon setting $a = -[(N - \operatorname{rank}\mathbf{X})/2]\log(2\pi)$, $c = [(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\underline{\mathbf{y}}]'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\underline{\mathbf{y}}$, and $K = N - \operatorname{rank}\mathbf{X}$, $g(\sigma) = \ell(\sigma, \mathbf{Q}'\underline{\mathbf{y}})$. Accordingly, it follows from the results of Part 1 of Subsection a that $\ell(\sigma, \mathbf{Q}'\underline{\mathbf{y}})$ attains its maximum value when σ^2 equals

$$\frac{[(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\underline{\mathbf{y}}]'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\underline{\mathbf{y}}}{N - \operatorname{rank}\mathbf{X}}.$$

Thus, the REML estimator of σ^2 is the estimator

$$\frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}{N - \operatorname{rank}\mathbf{X}}, \qquad (9.24)$$

where $\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{P}_{\mathbf{X}}\mathbf{y}$.

The REML estimator (9.24) is of the form (7.41) considered in Section 5.7c; it is the estimator of the form (7.41) that is unbiased. Unlike the (ordinary) ML estimator $\tilde{\mathbf{e}}'\tilde{\mathbf{e}}/N$ [which was derived in Part 1 of Subsection a and is also of the form (7.41)], it "accounts for the estimation of $\boldsymbol{\beta}$"; in the REML estimation of σ^2, the residual sum of squares $\tilde{\mathbf{e}}'\tilde{\mathbf{e}}$ is divided by $N - \operatorname{rank}\mathbf{X}$ rather than by N.

A matrix lemma. Preliminary to the further discussion of REML, it is convenient to establish the following lemma.

Lemma 5.9.8. Let $\mathbf{A}$ represent a $Q \times S$ matrix. Then, for any $K \times Q$ matrix $\mathbf{C}$ of full column rank Q and any $S \times T$ matrix $\mathbf{B}$ of full row rank S, $\mathbf{B}(\mathbf{C}\mathbf{A}\mathbf{B})^{-}\mathbf{C}$ is a generalized inverse of $\mathbf{A}$.

Proof. In light of Lemma 2.5.1, $\mathbf{C}$ has a left inverse, say $\mathbf{L}$, and $\mathbf{B}$ has a right inverse, say $\mathbf{R}$. And it follows that

$$\mathbf{A}\mathbf{B}(\mathbf{C}\mathbf{A}\mathbf{B})^{-}\mathbf{C}\mathbf{A} = \mathbf{I}\mathbf{A}\mathbf{B}(\mathbf{C}\mathbf{A}\mathbf{B})^{-}\mathbf{C}\mathbf{A}\mathbf{I} = \mathbf{L}\mathbf{C}\mathbf{A}\mathbf{B}(\mathbf{C}\mathbf{A}\mathbf{B})^{-}\mathbf{C}\mathbf{A}\mathbf{B}\mathbf{R} = \mathbf{L}\mathbf{C}\mathbf{A}\mathbf{B}\mathbf{R} = \mathbf{I}\mathbf{A}\mathbf{I} = \mathbf{A}.$$

Thus, $\mathbf{B}(\mathbf{CAB})^{-}\mathbf{C}$ is a generalized inverse of $\mathbf{A}$. Q.E.D.

Note that in the special case where $\mathbf{A}$ is nonsingular (i.e., the special case where $\mathbf{A}$ is a $Q \times Q$ matrix of rank Q), the result of Lemma 5.9.8 can be restated as follows:

$$\mathbf{A}^{-1} = \mathbf{B}(\mathbf{CAB})^{-}\mathbf{C}. \tag{9.25}$$

An informative and computationally useful expression for the REML log-likelihood function. Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a general linear model. Suppose further that the distribution of the vector $\mathbf{e}$ of residual effects is MVN and that the variance-covariance matrix $\mathbf{V}(\boldsymbol{\theta})$ of $\mathbf{e}$ is nonsingular (for every $\boldsymbol{\theta} \in \Theta$). And let $\mathbf{z} = \mathbf{R}'\mathbf{y}$, where $\mathbf{R}$ is an $N \times (N - \text{rank}\,\mathbf{X})$ matrix of full column rank $N - \text{rank}\,\mathbf{X}$ such that $\mathbf{X}'\mathbf{R} = \mathbf{0}$, and denote by $\underline{\mathbf{y}}$ the observed value of $\mathbf{y}$.

In REML, inferences about functions of $\boldsymbol{\theta}$ are based on the likelihood function $L(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ obtained by regarding the observed value $\mathbf{R}'\underline{\mathbf{y}}$ of $\mathbf{z}$ as the data vector. Corresponding to $L(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ is the log-likelihood function $\ell(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}}) = \log L(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$. We have that $\mathbf{z} \sim N[\mathbf{0}, \mathbf{R}'\mathbf{V}(\boldsymbol{\theta})\mathbf{R}]$, and it follows that

$$\ell(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}}) = -\frac{N - \text{rank}\,\mathbf{X}}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{R}'\mathbf{V}(\boldsymbol{\theta})\mathbf{R}| - \frac{1}{2}\underline{\mathbf{y}}'\mathbf{R}[\mathbf{R}'\mathbf{V}(\boldsymbol{\theta})\mathbf{R}]^{-1}\mathbf{R}'\underline{\mathbf{y}} \tag{9.26}$$

—recall that $\mathbf{R}'\mathbf{V}(\boldsymbol{\theta})\mathbf{R}$ is nonsingular.

REML estimates of functions of $\boldsymbol{\theta}$ are obtained from a value, say $\hat{\boldsymbol{\theta}}$, of $\boldsymbol{\theta}$ at which $L(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ or, equivalently, $\ell(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ attains its maximum value. By way of comparison, (ordinary) ML estimates of such functions are obtained from a value, say $\tilde{\boldsymbol{\theta}}$, at which the profile likelihood function $L_*(\boldsymbol{\theta}; \underline{\mathbf{y}})$ or profile log-likelihood function $\ell_*(\boldsymbol{\theta}; \underline{\mathbf{y}})$ attains its maximum value; the (ordinary) ML estimate of a function $h(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ is $h(\tilde{\boldsymbol{\theta}})$, whereas the REML estimate is $h(\hat{\boldsymbol{\theta}})$. It is of potential interest to compare $\ell(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ with $\ell_*(\boldsymbol{\theta}; \underline{\mathbf{y}})$. Expressions for $\ell_*(\boldsymbol{\theta}; \underline{\mathbf{y}})$ are given by results (9.13), (9.14), and (9.15). However, expression (9.26) [for $\ell(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$] is not of a form that facilitates meaningful comparisons with any of those expressions. Moreover, depending on the nature of the variance-covariance matrix $\mathbf{V}(\boldsymbol{\theta})$ (and on the choice of the matrix $\mathbf{R}$), expression (9.26) may not be well-suited for computational purposes [such as in computing the values of $\ell(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ corresponding to various values of $\boldsymbol{\theta}$].

For purposes of obtaining a more useful expression for $\ell(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$, take $\mathbf{S}$ to be any matrix (with N rows) whose columns span $\mathcal{C}(\mathbf{X})$, that is, any matrix such that $\mathcal{C}(\mathbf{S}) = \mathcal{C}(\mathbf{X})$ (in which case, $\mathbf{S} = \mathbf{XA}$ for some matrix $\mathbf{A}$). And, temporarily (for the sake of simplicity) writing $\mathbf{V}$ for $\mathbf{V}(\boldsymbol{\theta})$, observe that

$$(\mathbf{V}^{-1}\mathbf{S},\ \mathbf{R})'\mathbf{V}(\mathbf{V}^{-1}\mathbf{S},\ \mathbf{R}) = \text{diag}(\mathbf{S}'\mathbf{V}^{-1}\mathbf{S},\ \mathbf{R}'\mathbf{V}\mathbf{R}) \tag{9.27}$$

and [in light of result (2.5.5), Corollary 2.13.12, and Corollary 5.9.3] that

$$\begin{aligned}\text{rank}[(\mathbf{V}^{-1}\mathbf{S},\ \mathbf{R})'\mathbf{V}(\mathbf{V}^{-1}\mathbf{S},\ \mathbf{R})] &= \text{rank}[\text{diag}(\mathbf{S}'\mathbf{V}^{-1}\mathbf{S},\ \mathbf{R}'\mathbf{V}\mathbf{R})] \\ &= \text{rank}(\mathbf{S}'\mathbf{V}^{-1}\mathbf{S}) + \text{rank}(\mathbf{R}'\mathbf{V}\mathbf{R}) \\ &= \text{rank}(\mathbf{S}) + \text{rank}(\mathbf{R}) \\ &= \text{rank}(\mathbf{X}) + N - \text{rank}(\mathbf{X}) = N. \end{aligned} \tag{9.28}$$

Result (9.28) implies (in light of Corollary 5.9.3) that

$$\text{rank}\,(\mathbf{V}^{-1}\mathbf{S},\ \mathbf{R}) = N \tag{9.29}$$

or, equivalently, that $(\mathbf{V}^{-1}\mathbf{S},\ \mathbf{R})$ is of full row rank. Thus, upon applying formula (9.25), it follows from result (9.27) that

$$\begin{aligned}\mathbf{V}^{-1} &= (\mathbf{V}^{-1}\mathbf{S},\ \mathbf{R})\,\text{diag}[(\mathbf{S}'\mathbf{V}^{-1}\mathbf{S})^{-},\ (\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}](\mathbf{V}^{-1}\mathbf{S},\ \mathbf{R})' \\ &= \mathbf{V}^{-1}\mathbf{S}(\mathbf{S}'\mathbf{V}^{-1}\mathbf{S})^{-}\mathbf{S}'\mathbf{V}^{-1} + \mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}' \end{aligned} \tag{9.30}$$

and hence that

$$\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}' = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{S}(\mathbf{S}'\mathbf{V}^{-1}\mathbf{S})^{-}\mathbf{S}'\mathbf{V}^{-1}. \tag{9.31}$$

Moreover, as a special case of equality (9.31) (that where $\mathbf{S} = \mathbf{X}$), we obtain the following expression for $\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}'$ [a quantity which appears in the 3rd term of expression (9.26) for $\ell(\boldsymbol{\theta}; \mathbf{R}'\underset{\sim}{\mathbf{y}})$]:

$$\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}' = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}. \tag{9.32}$$

Now, consider the quantity $|\mathbf{R}'\mathbf{V}\mathbf{R}|$ [which appears in the 2nd term of expression (9.26)]. Take $\mathbf{X}_*$ to be any $N \times (\text{rank}\,\mathbf{X})$ matrix whose columns are linearly independent columns of $\mathbf{X}$ or, more generally, whose columns form a basis for $\mathcal{C}(\mathbf{X})$ (in which case, $\mathbf{X}_* = \mathbf{X}\mathbf{A}$ for some matrix $\mathbf{A}$). Observing that

$$(\mathbf{X}_*, \mathbf{R})'(\mathbf{X}_*, \mathbf{R}) = \text{diag}(\mathbf{X}_*'\mathbf{X}_*, \mathbf{R}'\mathbf{R})$$

and making use of basic properties of determinants, we find that

$$\begin{aligned} |(\mathbf{X}_*, \mathbf{R})'\mathbf{V}(\mathbf{X}_*, \mathbf{R})| &= |(\mathbf{X}_*, \mathbf{R})'||(\mathbf{X}_*, \mathbf{R})||\mathbf{V}| \\ &= |(\mathbf{X}_*, \mathbf{R})'(\mathbf{X}_*, \mathbf{R})||\mathbf{V}| \\ &= |\text{diag}(\mathbf{X}_*'\mathbf{X}_*, \mathbf{R}'\mathbf{R})||\mathbf{V}| \\ &= |\mathbf{X}_*'\mathbf{X}_*||\mathbf{R}'\mathbf{R}||\mathbf{V}|. \end{aligned} \tag{9.33}$$

And making use of formula (2.14.29) for the determinant of a partitioned matrix, we find that

$$\begin{aligned} |(\mathbf{X}_*, \mathbf{R})'\mathbf{V}(\mathbf{X}_*, \mathbf{R})| &= \begin{vmatrix} \mathbf{X}_*'\mathbf{V}\mathbf{X}_* & \mathbf{X}_*'\mathbf{V}\mathbf{R} \\ \mathbf{R}'\mathbf{V}\mathbf{X}_* & \mathbf{R}'\mathbf{V}\mathbf{R} \end{vmatrix} \\ &= |\mathbf{R}'\mathbf{V}\mathbf{R}||\mathbf{X}_*'\mathbf{V}\mathbf{X}_* - \mathbf{X}_*'\mathbf{V}\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}'\mathbf{V}\mathbf{X}_*| \\ &= |\mathbf{R}'\mathbf{V}\mathbf{R}||\mathbf{X}_*'[\mathbf{V} - \mathbf{V}\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}'\mathbf{V}]\mathbf{X}_*|. \end{aligned} \tag{9.34}$$

Moreover, as a special case of equality (9.30) (that where $\mathbf{S} = \mathbf{X}_*$), we have (since, in light of Corollary 5.9.3, $\mathbf{X}_*'\mathbf{V}^{-1}\mathbf{X}_*$ is nonsingular) that

$$\mathbf{V}^{-1} = \mathbf{V}^{-1}\mathbf{X}_*(\mathbf{X}_*'\mathbf{V}^{-1}\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{V}^{-1} + \mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}'$$

and (upon premultiplying and postmultiplying by $\mathbf{V}$ and rearranging terms) that

$$\mathbf{V} - \mathbf{V}\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}'\mathbf{V} = \mathbf{X}_*(\mathbf{X}_*'\mathbf{V}^{-1}\mathbf{X}_*)^{-1}\mathbf{X}_*'. \tag{9.35}$$

Upon replacing $\mathbf{V} - \mathbf{V}\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}'\mathbf{V}$ with expression (9.35), result (9.34) simplifies as follows:

$$\begin{aligned} |(\mathbf{X}_*, \mathbf{R})'\mathbf{V}(\mathbf{X}_*, \mathbf{R})| &= |\mathbf{R}'\mathbf{V}\mathbf{R}||\mathbf{X}_*'\mathbf{X}_*(\mathbf{X}_*'\mathbf{V}^{-1}\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{X}_*| \\ &= |\mathbf{R}'\mathbf{V}\mathbf{R}||\mathbf{X}_*'\mathbf{X}_*|^2/|\mathbf{X}_*'\mathbf{V}^{-1}\mathbf{X}_*|. \end{aligned} \tag{9.36}$$

It remains to equate expressions (9.33) and (9.36); doing so leads to the following expression for $|\mathbf{R}'\mathbf{V}\mathbf{R}|$:

$$|\mathbf{R}'\mathbf{V}\mathbf{R}| = |\mathbf{R}'\mathbf{R}||\mathbf{V}||\mathbf{X}_*'\mathbf{V}^{-1}\mathbf{X}_*|/|\mathbf{X}_*'\mathbf{X}_*|. \tag{9.37}$$

Upon substituting expressions (9.32) and (9.37) [for $\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}'$ and $|\mathbf{R}'\mathbf{V}\mathbf{R}|$] into expression (9.26), we find that the REML log-likelihood function $\ell(\boldsymbol{\theta}; \mathbf{R}'\underset{\sim}{\mathbf{y}})$ is reexpressible as follows:

$$\begin{aligned} \ell(\boldsymbol{\theta}; \mathbf{R}'\underset{\sim}{\mathbf{y}}) = &-\frac{N - \text{rank}\,\mathbf{X}}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{R}'\mathbf{R}| + \frac{1}{2}\log|\mathbf{X}_*'\mathbf{X}_*| \\ &- \frac{1}{2}\log|\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2}\log|\mathbf{X}_*'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}_*| \\ &- \frac{1}{2}\underset{\sim}{\mathbf{y}}'\big([\mathbf{V}(\boldsymbol{\theta})]^{-1} - [\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\{\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\}^{-}\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\big)\underset{\sim}{\mathbf{y}}. \end{aligned} \tag{9.38}$$

If $\mathbf{R}$ is taken to be a matrix whose columns form an orthonormal basis for $\mathfrak{N}(\mathbf{X}')$, then the second term of expression (9.38) equals 0; similarly, if $\mathbf{X}_*$ is taken to be a matrix whose columns form an orthonormal basis for $\mathcal{C}(\mathbf{X})$, then the third term of expression (9.38) equals 0. However, what is more important is that the choice of $\mathbf{R}$ affects expression (9.38) only through its second term, which is a constant (i.e., does not involve $\boldsymbol{\theta}$). And for any two choices of $\mathbf{X}_*$, say $\mathbf{X}_1$ and $\mathbf{X}_2$, $\mathbf{X}_2 = \mathbf{X}_1\mathbf{B}$ for some matrix $\mathbf{B}$ (which is necessarily nonsingular), implying that

$$-\frac{1}{2}\log|\mathbf{X}_2'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}_2| = -\frac{1}{2}\log|\mathbf{B}'\mathbf{X}_1'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}_1\mathbf{B}| = -\log|\det\mathbf{B}| - \frac{1}{2}\log|\mathbf{X}_1'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}_1|$$

and, similarly, that

$$\frac{1}{2}\log|\mathbf{X}_2'\mathbf{X}_2| = \log|\det\mathbf{B}| + \frac{1}{2}\log|\mathbf{X}_1'\mathbf{X}_1|,$$

so that the only effect on expression (9.38) of a change in the choice of $\mathbf{X}_*$ from $\mathbf{X}_1$ to $\mathbf{X}_2$ is to add a constant to the third term and to subtract the same constant from the fifth term. Thus, the choice of $\mathbf{R}$ and the choice of $\mathbf{X}_*$ are immaterial.

The last term of expression (9.38) can be reexpressed in terms of an arbitrary solution, say $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$, to the linear system

$$\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\mathbf{b} = \mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\underline{\mathbf{y}} \tag{9.39}$$

(in the $P \times 1$ vector $\mathbf{b}$)—recall (from Subsection a) that this linear system is consistent, that $\mathbf{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ does not depend on the choice of $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$, and that the choices for $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ include the vector $(\{\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\}^{-})'\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\underline{\mathbf{y}}$. We find that

$$\begin{aligned}\underline{\mathbf{y}}'\big([\mathbf{V}(\boldsymbol{\theta})]^{-1} - [\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\{\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\}^{-}\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\big)\underline{\mathbf{y}}& \\ = \underline{\mathbf{y}}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\underline{\mathbf{y}} - [\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})]'\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\underline{\mathbf{y}} & \qquad (9.40)\\ = [\underline{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})]'[\mathbf{V}(\boldsymbol{\theta})]^{-1}[\underline{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})]. & \qquad (9.41)\end{aligned}$$

It is informative to compare expression (9.38) for $\ell(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ with expression (9.15) for the profile log-likelihood function $\ell_*(\boldsymbol{\theta}; \underline{\mathbf{y}})$. Aside from the terms that do not depend on $\boldsymbol{\theta}$ [the 1st term of expression (9.15) and the first 3 terms of expression (9.38)], the only difference between the two expressions is the inclusion in expression (9.38) of the term $-\frac{1}{2}\log|\mathbf{X}_*'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}_*|$. This term depends on $\boldsymbol{\theta}$, but not on $\underline{\mathbf{y}}$. Its inclusion serves to adjust the profile log-likelihood function $\ell_*(\boldsymbol{\theta}; \underline{\mathbf{y}})$ so as to compensate for the failure of ordinary ML (in estimating functions of $\boldsymbol{\theta}$) to account for the estimation of $\boldsymbol{\beta}$. Unlike the profile log-likelihood function, $\ell(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ is the logarithm of an actual likelihood function and, consequently, has the properties thereof—it is the logarithm of the likelihood function $L(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ obtained by regarding the observed value $\mathbf{R}'\underline{\mathbf{y}}$ of $\mathbf{z}$ as the data vector.

If the form of the $N \times N$ matrix $\mathbf{V}(\boldsymbol{\theta})$ is such that $\mathbf{V}(\boldsymbol{\theta})$ is relatively easy to invert (as is often the case in practice), then expression (9.38) for $\ell(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ is likely to be much more useful for computational purposes than expression (9.26). Expression (9.38) [along with expression (9.40) or (9.41)] serves to relate the numerical evaluation of $\ell(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ for any particular value of $\boldsymbol{\theta}$ to the solution of the linear system (9.39), comprising P equations in P "unknowns."

Special case: Aitken model. Let us now specialize to the case where $\mathbf{y}$ follows an Aitken model (and where $\mathbf{H}$ is nonsingular). As in Subsection a, this case is to be regarded as the special case of the general linear model where $T = 1$, where $\boldsymbol{\theta} = (\sigma)$, and where $\mathbf{V}(\boldsymbol{\theta}) = \sigma^2\mathbf{H}$. In this special case, linear system (9.39) is equivalent to (i.e., has the same solutions as) the linear system

$$\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{H}^{-1}\underline{\mathbf{y}}, \tag{9.42}$$

comprising the Aitken equations. And taking $\tilde{\boldsymbol{\beta}}$ to be any solution to linear system (9.42), we find

[in light of results (9.38) and (9.41)] that the log-likelihood function $\ell(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ is expressible as

$$\ell(\sigma; \mathbf{R}'\underline{\mathbf{y}}) = -\frac{N-\operatorname{rank}\mathbf{X}}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{R}'\mathbf{R}| + \frac{1}{2}\log|\mathbf{X}_*'\mathbf{X}_*| - \frac{1}{2}\log|\mathbf{H}| - \frac{1}{2}\log|\mathbf{X}_*'\mathbf{H}^{-1}\mathbf{X}_*| - \frac{N-\operatorname{rank}\mathbf{X}}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\underline{\mathbf{y}}-\mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{H}^{-1}(\underline{\mathbf{y}}-\mathbf{X}\tilde{\boldsymbol{\beta}}). \tag{9.43}$$

Unless $\underline{\mathbf{y}}-\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{0}$ (which is an event of probability 0), $\ell(\sigma; \mathbf{R}'\underline{\mathbf{y}})$ is of the form of the function $g(\sigma)$ defined (in Part 1 of Subsection a) by equality (9.4); upon setting $a = -[(N-\operatorname{rank}\mathbf{X})/2]\log(2\pi) - (1/2)\log|\mathbf{R}'\mathbf{R}| + (1/2)\log|\mathbf{X}_*'\mathbf{X}_*| - (1/2)\log|\mathbf{H}| - (1/2)\log|\mathbf{X}_*'\mathbf{H}^{-1}\mathbf{X}_*|$, $c = (\underline{\mathbf{y}}-\mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{H}^{-1}(\underline{\mathbf{y}}-\mathbf{X}\tilde{\boldsymbol{\beta}})$, and $K = N-\operatorname{rank}\mathbf{X}$, $g(\sigma) = \ell(\sigma; \mathbf{R}'\underline{\mathbf{y}})$. Accordingly, it follows from the results of Part 1 of Subsection a that $\ell(\sigma; \mathbf{R}'\underline{\mathbf{y}})$ attains its maximum value when σ^2 equals

$$\frac{(\underline{\mathbf{y}}-\mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{H}^{-1}(\underline{\mathbf{y}}-\mathbf{X}\tilde{\boldsymbol{\beta}})}{N-\operatorname{rank}\mathbf{X}}. \tag{9.44}$$

The quantity (9.44) is the REML estimate of σ^2; it is the estimate obtained by dividing $(\underline{\mathbf{y}}-\mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{H}^{-1}(\underline{\mathbf{y}}-\mathbf{X}\tilde{\boldsymbol{\beta}})$ by $N-\operatorname{rank}\mathbf{X}$. It differs from the (ordinary) ML estimate of σ^2; which (as is evident from the results of Subsection a) is obtained by dividing $(\underline{\mathbf{y}}-\mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{H}^{-1}(\underline{\mathbf{y}}-\mathbf{X}\tilde{\boldsymbol{\beta}})$ by N. Note that in the further special case of the G–M model (i.e., the further special case where $\mathbf{H} = \mathbf{I}$), the Aitken equations simplify to the normal equations $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\underline{\mathbf{y}}$ and expression (9.44) (for the REML estimate) is (upon setting $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\underline{\mathbf{y}}$) reexpressible as $[(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\underline{\mathbf{y}}]'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\underline{\mathbf{y}}/(N-\operatorname{rank}\mathbf{X})$, in agreement with the expression for the REML estimator [expression (9.24)] derived in a previous part of the present subsection.

c. Elliptical distributions

The results of Subsections a and b (on the ML and REML estimation of functions of the parameters of a G–M, Aitken, or general linear model) were obtained under the assumption that the distribution of the vector $\mathbf{e}$ of residual effects is MVN. Some of the properties of the MVN distribution extend (in a relatively straightforward way) to a broader class of distributions called elliptical distributions (or elliptically contoured or elliptically symmetric distributions). Elliptical distributions are introduced (and some of their basic properties described) in the present subsection—this follows the presentation (in Part 1 of the present subsection) of a useful result on orthogonal matrices. Then, in Subsection d, the results of Subsections a and b are revisited with the intent of obtaining extensions suitable for G–M, Aitken, or general linear models when the form of the distribution of the vector $\mathbf{e}$ of residual effects is taken to be that of an elliptical distribution other than a multivariate normal distribution.

A matrix lemma.

Lemma 5.9.9. For any two M-dimensional column vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, $\mathbf{x}_2'\mathbf{x}_2 = \mathbf{x}_1'\mathbf{x}_1$ if and only if there exists an $M \times M$ orthogonal matrix $\mathbf{O}$ such that $\mathbf{x}_2 = \mathbf{O}\mathbf{x}_1$.

Proof. If there exists an orthogonal matrix $\mathbf{O}$ such that $\mathbf{x}_2 = \mathbf{O}\mathbf{x}_1$, then, clearly,

$$\mathbf{x}_2'\mathbf{x}_2 = (\mathbf{O}\mathbf{x}_1)'\mathbf{O}\mathbf{x}_1 = \mathbf{x}_1'\mathbf{O}'\mathbf{O}\mathbf{x}_1 = \mathbf{x}_1'\mathbf{x}_1.$$

For purposes of establishing the converse, take $\mathbf{u} = (1, 0, 0, \ldots, 0)'$ to be the first column of $\mathbf{I}_M$, and assume that both $\mathbf{x}_1$ and $\mathbf{x}_2$ are nonnull—if $\mathbf{x}_2'\mathbf{x}_2 = \mathbf{x}_1'\mathbf{x}_1$ and either $\mathbf{x}_1$ or $\mathbf{x}_2$ is null, then both $\mathbf{x}_1$ and $\mathbf{x}_2$ are null, in which case $\mathbf{x}_2 = \mathbf{O}\mathbf{x}_1$ for any $M \times M$ orthogonal matrix $\mathbf{O}$. And for $i = 1, 2$, define

$$\mathbf{P}_i = \mathbf{I} - 2(\mathbf{v}_i'\mathbf{v}_i)^{-1}\mathbf{v}_i\mathbf{v}_i',$$

where $\mathbf{v}_i = \mathbf{x}_i - (\mathbf{x}_i'\mathbf{x}_i)^{1/2}\mathbf{u}$—if $\mathbf{v}_i = \mathbf{0}$, take $\mathbf{P}_i = \mathbf{I}$. The two matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ are Householder matrices; they are orthogonal and are such that, for $i = 1, 2$, $\mathbf{P}_i\mathbf{x}_i = (\mathbf{x}_i'\mathbf{x}_i)^{1/2}\mathbf{u}$—refer, e.g., to Golub and Van Loan (2013, sec. 5.1.2). Thus, if $\mathbf{x}_2'\mathbf{x}_2 = \mathbf{x}_1'\mathbf{x}_1$, then

$$\mathbf{P}_2\mathbf{x}_2 = (\mathbf{x}_2'\mathbf{x}_2)^{1/2}\mathbf{u} = (\mathbf{x}_1'\mathbf{x}_1)^{1/2}\mathbf{u} = \mathbf{P}_1\mathbf{x}_1,$$

implying that

$$\mathbf{x}_2 = \mathbf{P}_2'\mathbf{P}_1\mathbf{x}_1$$

and hence (since $\mathbf{P}_2'\mathbf{P}_1$ is orthogonal) that there exists an orthogonal matrix $\mathbf{O}$ such that $\mathbf{x}_2 = \mathbf{O}\mathbf{x}_1$. Q.E.D.

Spherical distributions. Elliptical distributions are defined in terms of spherical distributions (which are themselves elliptical distributions, albeit of a relatively simple kind). An $M \times 1$ random vector $\mathbf{z}$ is said to have a *spherical* (or spherically symmetric) *distribution* if, for every $M \times M$ orthogonal matrix $\mathbf{O}$, the distribution of $\mathbf{O}\mathbf{z}$ is the same as that of $\mathbf{z}$. For example, the $N(\mathbf{0}, \sigma^2\mathbf{I})$ distribution (where σ is any nonnegative scalar) is a spherical distribution.

Suppose that the distribution of the M-dimensional random vector $\mathbf{z} = (z_1, z_2, \ldots, z_M)'$ is spherical. Then, upon observing that $-\mathbf{I}_M$ is an orthogonal matrix, we find that

$$-\mathbf{z} = -\mathbf{I}_M\mathbf{z} \sim \mathbf{z}. \tag{9.45}$$

Thus, a spherical distribution is symmetric. And, it follows, in particular, that if $\mathrm{E}(\mathbf{z})$ exists, then

$$\mathrm{E}(\mathbf{z}) = \mathbf{0}. \tag{9.46}$$

Further, if the second-order moments of the distribution of $\mathbf{z}$ exist, then

$$\mathrm{var}(\mathbf{z}) = c\mathbf{I} \tag{9.47}$$

for some nonnegative scalar c.

To verify result (9.47), take $\mathbf{O}_i$ to be the $M \times M$ orthogonal matrix obtained by interchanging the first and ith rows of $\mathbf{I}_M$, and take $\mathbf{P}_i$ to be the $M \times M$ orthogonal matrix obtained by multiplying the ith row of $\mathbf{I}_M$ by -1. Then, upon observing that $\mathbf{O}_i\mathbf{z} \sim \mathbf{z}$ and that z_i is the first element of $\mathbf{O}_i\mathbf{z}$, we find that

$$z_i \sim z_1. \tag{9.48}$$

And upon observing that $\mathbf{P}_i\mathbf{z} \sim \mathbf{z}$ and that $-z_i$ is the ith element of $\mathbf{P}_i\mathbf{z}$, we find that (for $j > i$)

$$\begin{pmatrix} -z_i \\ z_j \end{pmatrix} \sim \begin{pmatrix} z_i \\ z_j \end{pmatrix}$$

and hence that

$$-z_i z_j \sim z_i z_j. \tag{9.49}$$

It follows from equality (9.48) that the diagonal elements of $\mathrm{var}(\mathbf{z})$ have a common value c and from equality (9.49) that the off-diagonal elements of $\mathrm{var}(\mathbf{z})$ [the ijth of which equals $\mathrm{E}(z_i z_j)$] are 0.

According to result (9.47), the M elements $z_1, z_2, \ldots, z_M$ of the spherically distributed random vector $\mathbf{z}$ are uncorrelated. However, it is only in the special case where the distribution of $\mathbf{z}$ is MVN that $z_1, z_2, \ldots, z_M$ are statistically independent—refer, e.g., to Kollo and von Rosen (2005, sec. 2.3) or to Fang, Kotz, and Ng (1990, sec. 4.3) for a proof.

The variance-covariance matrix of the spherically distributed random vector $\mathbf{z}$ is a scalar multiple $c\mathbf{I}$ of $\mathbf{I}$. Note that [aside from the degenerate special case where $\mathrm{var}(\mathbf{z}) = \mathbf{0}$ or, equivalently, where $\mathbf{z} = \mathbf{0}$ with probability 1] the elements of $\mathbf{z}$ can be rescaled by dividing each of them by $\sqrt{c}$, the effect of which is to transform $\mathbf{z}$ into the vector $c^{-1/2}\mathbf{z}$ whose variance-covariance matrix is $\mathbf{I}$. Note also that, like $\mathbf{z}$, the transformed vector $c^{-1/2}\mathbf{z}$ has a spherical distribution.

Pdf of a spherical distribution. Take $\mathbf{z} = (z_1, z_2, \ldots, z_M)'$ to be an M-dimensional random (column) vector that has an absolutely continuous distribution with pdf $f(\cdot)$. Clearly, whether or not this distribution is spherical depends on the nature of the pdf.

Define $\mathbf{u} = \mathbf{O}\mathbf{z}$, where $\mathbf{O}$ is an arbitrary $M \times M$ orthogonal matrix, and denote by u_i the ith element of $\mathbf{u}$. Then, the distribution of $\mathbf{u}$ has as a pdf the function $h(\cdot)$ obtained by taking (for every value of $\mathbf{u}$)

$$h(\mathbf{u}) = |\det \mathbf{J}|\, f(\mathbf{O}'\mathbf{u}),$$

where $\mathbf{J}$ is the $M \times M$ matrix with ijth element $\partial z_i / \partial u_j$ (e.g., Bickel and Doksum 2001, sec. B.2). Moreover, $\mathbf{J} = \mathbf{O}'$, implying (in light of Corollary 2.14.19) that $\det \mathbf{J} = \pm 1$. Thus,

$$h(\mathbf{u}) = f(\mathbf{O}'\mathbf{u}) \qquad \text{or, equivalently,} \qquad h(\mathbf{O}\mathbf{z}) = f(\mathbf{z}).$$

And upon observing [in light of the fundamental theorem of (integral) calculus (e.g., Billingsley 1995)] that $\mathbf{u} \sim \mathbf{z}$ if and only if $h(\mathbf{O}\mathbf{z}) = f(\mathbf{O}\mathbf{z})$ (with probability 1), it follows that $\mathbf{u} \sim \mathbf{z}$ if and only if

$$f(\mathbf{O}\mathbf{z}) = f(\mathbf{z}) \quad \text{(with probability 1)}. \tag{9.50}$$

In effect, we have established that $\mathbf{z}$ has a spherical distribution if and only if, for every orthogonal matrix $\mathbf{O}$, the pdf $f(\cdot)$ satisfies condition (9.50). Now, suppose that $f(\mathbf{z})$ depends on the value of $\mathbf{z}$ only through $\mathbf{z}'\mathbf{z}$ or, equivalently, that there exists a (nonnegative) function $g(\cdot)$ (of a single nonnegative variable) such that

$$f(\mathbf{z}) = g(\mathbf{z}'\mathbf{z}) \quad \text{(for every value of } \mathbf{z}\text{)}. \tag{9.51}$$

Clearly, if $f(\cdot)$ is of the form (9.51), then, for every orthogonal matrix $\mathbf{O}$, $f(\cdot)$ satisfies condition (9.50) and, in fact, satisfies the more stringent condition

$$f(\mathbf{O}\mathbf{z}) = f(\mathbf{z}) \quad \text{(for every value of } \mathbf{z}\text{)}. \tag{9.52}$$

Thus, if $f(\cdot)$ is of the form (9.51), then the distribution of $\mathbf{z}$ is spherical.

Consider the converse. Suppose that the distribution of $\mathbf{z}$ is spherical and hence that, for every orthogonal matrix $\mathbf{O}$, $f(\cdot)$ satisfies condition (9.50). Is $f(\cdot)$ necessarily of the form (9.51)? If for every orthogonal matrix $\mathbf{O}$, $f(\cdot)$ satisfies condition (9.52), then the answer is yes.

To see this, suppose that (for every orthogonal matrix $\mathbf{O}$) $f(\cdot)$ satisfies condition (9.52). Then, for any $M \times 1$ vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ such that $\mathbf{z}_2'\mathbf{z}_2 = \mathbf{z}_1'\mathbf{z}_1$, we find [upon observing (in light of Lemma 5.9.9) that $\mathbf{z}_2 = \mathbf{O}\mathbf{z}_1$ for some orthogonal matrix $\mathbf{O}$] that $f(\mathbf{z}_2) = f(\mathbf{z}_1)$. Thus, for any particular (nonnegative) constant c, $f(\mathbf{z})$ has the same value for every $\mathbf{z}$ for which $\mathbf{z}'\mathbf{z} = c$. And it follows that there exists a function $g(\cdot)$ for which $f(\cdot)$ is expressible in the form (9.51).

Subsequently, there will be occasion to refer to the distribution of an M-dimensional random column vector that is absolutely continuous with a pdf $f(\cdot)$ of the form (9.51). Accordingly, as a matter of convenience, let us interpret any reference to an absolutely continuous spherical distribution as a reference to a distribution with those characteristics.

Let $g(\cdot)$ represent a nonnegative function whose domain is the interval $[0, \infty)$. And let $\mathbf{z} = (z_1, z_2, \ldots, z_M)'$ represent an $M \times 1$ vector of (unrestricted) variables, and suppose that

$$0 < \int_{\mathbb{R}^M} g(\mathbf{z}'\mathbf{z})\, d\mathbf{z} < \infty. \tag{9.53}$$

Further, take $f(\mathbf{z})$ to be the (nonnegative) function of $\mathbf{z}$ defined by

$$f(\mathbf{z}) = c^{-1} g(\mathbf{z}'\mathbf{z}), \tag{9.54}$$

where $c = \int_{\mathbb{R}^M} g(\mathbf{z}'\mathbf{z})\, d\mathbf{z}$ (and observe that $\int_{\mathbb{R}^M} f(\mathbf{z})\, d\mathbf{z} = 1$). Then, there is an absolutely continuous distribution (of an $M \times 1$ random vector) having $f(\cdot)$ as a pdf, and [since $f(\cdot)$ is of the form (9.51)] that distribution is spherical.

The M-dimensional integral $\int_{\mathcal{R}^M} g(\mathbf{z}'\mathbf{z})\, d\mathbf{z}$ can be simplified. Clearly,

$$\int_{\mathcal{R}^M} g(\mathbf{z}'\mathbf{z})\, d\mathbf{z} = 2^M \int_0^\infty \int_0^\infty \cdots \int_0^\infty g\left(\textstyle\sum_{i=1}^M z_i^2\right) dz_1 dz_2 \cdots dz_M.$$

Upon making the change of variables $u_i = z_i^2$ $(i = 1, 2, \ldots, M)$ and observing that $\partial z_i/\partial u_i = (1/2)u_i^{-1/2}$, we find that

$$\begin{aligned}\int_{\mathcal{R}^M} g(\mathbf{z}'\mathbf{z})\, d\mathbf{z} &= 2^M \int_0^\infty \int_0^\infty \cdots \int_0^\infty g\left(\textstyle\sum_{i=1}^M u_i\right) \left(\tfrac{1}{2}\right)^M \textstyle\prod_{i=1}^M u_i^{-1/2}\, du_1 du_2 \cdots du_M \\ &= \int_0^\infty \int_0^\infty \cdots \int_0^\infty g\left(\textstyle\sum_{i=1}^M u_i\right) \textstyle\prod_{i=1}^M u_i^{-1/2}\, du_1 du_2 \cdots du_M.\end{aligned}$$

And upon making the further change of variables $y_i = u_i$ $(i = 1, 2, \ldots, M-1)$, $y_M = \sum_{i=1}^M u_i$ and observing that the $M \times M$ matrix with ijth element $\partial u_i/\partial y_j$ equals $\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{1}' & 1 \end{pmatrix}$ (the determinant of which equals 1), we find that

$$\int_{\mathcal{R}^M} g(\mathbf{z}'\mathbf{z})\, d\mathbf{z} = \int_D g(y_M) \textstyle\prod_{i=1}^{M-1} y_i^{-1/2} \left(y_M - \sum_{i=1}^{M-1} y_i\right)^{-1/2} dy_1 dy_2 \cdots dy_M,$$

where $D = \{y_1, y_2, \ldots, y_M : y_i \geq 0 \ (i = 1, 2, \ldots, M-1),\ y_M \geq \sum_{i=1}^{M-1} y_i\}$. Moreover, upon making yet another change of variables $w_i = y_i/y_M$ $(i = 1, 2, \ldots, M-1)$, $w_M = y_M$ and observing that the $M \times M$ matrix with ijth element $\partial y_i/\partial w_j$ equals $\begin{bmatrix} w_M \mathbf{I} & (w_1, w_2, \ldots, w_{M-1})' \\ \mathbf{0} & 1 \end{bmatrix}$ (the determinant of which equals w_M^{M-1}), we find that

$$\begin{aligned}\int_{\mathcal{R}^M} g(\mathbf{z}'\mathbf{z})\, d\mathbf{z} = \int_{D_*} \textstyle\prod_{i=1}^{M-1} w_i^{-1/2} \left(1 - \sum_{i=1}^{M-1} w_i\right)^{-1/2} dw_1 dw_2 \cdots dw_{M-1} \\ \times \int_0^\infty w_M^{(M/2)-1} g(w_M)\, dw_M, \qquad (9.55)\end{aligned}$$

where $D_* = \{w_1, w_2, \ldots, w_{M-1} : w_i \geq 0 \ (i = 1, 2, \ldots, M-1),\ \sum_{i=1}^{M-1} w_i \leq 1\}$.

According to a basic result on the normalizing constant for the pdf of a Dirichlet distribution—the Dirichlet distribution is the subject of Section 6.1e—

$$\int_{D_*} \textstyle\prod_{i=1}^{M-1} w_i^{-1/2} \left(1 - \sum_{i=1}^{M-1} w_i\right)^{-1/2} dw_1 dw_2 \cdots dw_{M-1} = \frac{[\Gamma(1/2)]^M}{\Gamma(M/2)} = \frac{\pi^{M/2}}{\Gamma(M/2)}.$$

Thus,

$$\int_{\mathcal{R}^M} g(\mathbf{z}'\mathbf{z})\, d\mathbf{z} = \frac{\pi^{M/2}}{\Gamma(M/2)} \int_0^\infty w_M^{(M/2)-1} g(w_M)\, dw_M; \qquad (9.56)$$

and upon introducing the change of variable $s = w_M^{1/2}$ and observing that $dw_M/ds = 2s$, we find that

$$\int_{\mathcal{R}^M} g(\mathbf{z}'\mathbf{z})\, d\mathbf{z} = \frac{2\pi^{M/2}}{\Gamma(M/2)} \int_0^\infty s^{M-1} g(s^2)\, ds. \qquad (9.57)$$

In light of result (9.57), the function $g(\cdot)$ satisfies condition (9.53) if and only if

$$0 < \int_0^\infty s^{M-1} g(s^2)\, ds < \infty,$$

in which case the constant c in expression (9.54) is expressible in the form (9.57).

Moment generating function of a spherical distribution. Spherical distributions can be characterized in terms of their moment generating functions (or, more generally, their characteristic functions) as

well as in terms of their pdfs. Take $\mathbf{z} = (z_1, z_2, \ldots, z_M)'$ to be an M-dimensional random (column) vector, and suppose that the distribution of $\mathbf{z}$ has a moment generating function, say $\psi(\cdot)$. Then, for the distribution of $\mathbf{z}$ to be spherical, it is necessary and sufficient that

$$\psi(\mathbf{Ot}) = \psi(\mathbf{t}) \quad \text{for every } M \times M \text{ orthogonal matrix } \mathbf{O} \quad (\text{and for every } M \times 1 \text{ vector } \mathbf{t} \text{ in a neighborhood of } \mathbf{0}). \tag{9.58}$$

To see this, let $\mathbf{O}$ represent an arbitrary $M \times M$ matrix, and observe that (for any $M \times 1$ vector $\mathbf{t}$)

$$\psi(\mathbf{Ot}) = \mathrm{E}\big[e^{(\mathbf{Ot})'\mathbf{z}}\big] = \mathrm{E}\big[e^{\mathbf{t}'(\mathbf{O}'\mathbf{z})}\big]$$

and hence that $\psi(\mathbf{Ot}) = \psi(\mathbf{t})$ (for every $M \times 1$ vector $\mathbf{t}$ in a neighborhood of $\mathbf{0}$) if and only if $\psi(\cdot)$ is the moment generating function of the distribution of $\mathbf{O}'\mathbf{z}$ (as well as that of the distribution of $\mathbf{z}$), or equivalently—refer, e.g., to Casella and Berger (2002, p. 65)—if and only if $\mathbf{O}'\mathbf{z}$ and $\mathbf{z}$ have the same distribution.

For the distribution of $\mathbf{z}$ to be spherical, it is necessary and sufficient that $\psi(\mathbf{t})$ depend on the $M \times 1$ vector $\mathbf{t}$ only through the value of $\mathbf{t}'\mathbf{t}$ or, equivalently, that there exists a function $\phi(\cdot)$ (of a single nonnegative variable) such that

$$\psi(\mathbf{t}) = \phi(\mathbf{t}'\mathbf{t}) \quad (\text{for every } M \times 1 \text{ vector } \mathbf{t} \text{ in a neighborhood of } \mathbf{0}). \tag{9.59}$$

Let us verify the necessity and sufficiency of the existence of a function $\phi(\cdot)$ that satisfies condition (9.59). If there exists a function $\phi(\cdot)$ that satisfies condition (9.59), then for every $M \times M$ orthogonal matrix $\mathbf{O}$ (and for every $M \times 1$ vector $\mathbf{t}$ in a neighborhood of $\mathbf{0}$),

$$\psi(\mathbf{Ot}) = \phi[(\mathbf{Ot})'\mathbf{Ot}] = \phi(\mathbf{t}'\mathbf{O}'\mathbf{Ot}) = \phi(\mathbf{t}'\mathbf{t}) = \psi(\mathbf{t}),$$

so that condition (9.58) is satisfied and, consequently, the distribution of $\mathbf{z}$ is spherical.

Conversely, suppose that the distribution of $\mathbf{z}$ is spherical and hence that condition (9.58) is satisfied. Then, for "any" $M \times 1$ vectors $\mathbf{t}_1$ and $\mathbf{t}_2$ such that $\mathbf{t}_2'\mathbf{t}_2 = \mathbf{t}_1'\mathbf{t}_1$, we find [upon observing (in light of Lemma 5.9.9) that $\mathbf{t}_2 = \mathbf{Ot}_1$ for some orthogonal matrix $\mathbf{O}$] that $\psi(\mathbf{t}_2) = \psi(\mathbf{t}_1)$. Thus, for any sufficiently small nonnegative constant c, $\psi(\mathbf{t})$ has the same value for every $M \times 1$ vector $\mathbf{t}$ for which $\mathbf{t}'\mathbf{t} = c$. And it follows that there exists a function $\phi(\cdot)$ that satisfies condition (9.59).

What can be said about the nature of the function $\phi(\cdot)$? Clearly, $\phi(0) = 1$. Moreover, $\phi(\cdot)$ is a strictly increasing function. To see this, take $\mathbf{t}$ to be any $M \times 1$ vector (of constants) such that $\mathbf{t}'\mathbf{t} = 1$, and observe that for any nonnegative scalar k,

$$\phi(k) = \phi(k\mathbf{t}'\mathbf{t}) = \tfrac{1}{2}\phi(k\mathbf{t}'\mathbf{t}) + \tfrac{1}{2}\phi[k(-\mathbf{t})'(-\mathbf{t})] = \tfrac{1}{2}\,\mathrm{E}\Big[e^{\sqrt{k}\mathbf{t}'\mathbf{z}} + e^{-\sqrt{k}\mathbf{t}'\mathbf{z}}\Big].$$

Observe also that (for $k > 0$)

$$\frac{d\big[e^{\sqrt{k}\mathbf{t}'\mathbf{z}} + e^{-\sqrt{k}\mathbf{t}'\mathbf{z}}\big]}{dk} = (1/2)k^{-1/2}\mathbf{t}'\mathbf{z}\Big(e^{\sqrt{k}\mathbf{t}'\mathbf{z}} - e^{-\sqrt{k}\mathbf{t}'\mathbf{z}}\Big) > 0 \quad \text{if } \mathbf{t}'\mathbf{z} \neq 0.$$

Thus (for $k > 0$)

$$\frac{d\,\phi(k)}{dk} = \tfrac{1}{2}\,\mathrm{E}\left\{\frac{d\big[e^{\sqrt{k}\mathbf{t}'\mathbf{z}} + e^{-\sqrt{k}\mathbf{t}'\mathbf{z}}\big]}{dk}\right\} > 0,$$

which confirms that $\phi(\cdot)$ is a strictly increasing function.

Linear transformation of a spherically distributed random vector. Let M and N represent arbitrary positive integers. And define

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Gamma}'\mathbf{z}, \tag{9.60}$$

where $\boldsymbol{\mu}$ is an arbitrary M-dimensional nonrandom column vector, $\boldsymbol{\Gamma}$ is an arbitrary $N \times M$ nonrandom matrix, and $\mathbf{z}$ is an N-dimensional spherically distributed random column vector. Further, let $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}'\boldsymbol{\Gamma}$.

If $\mathrm{E}(\mathbf{z})$ exists, then $\mathrm{E}(\mathbf{x})$ exists and [in light of result (9.46)]

$$\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}. \tag{9.61}$$

And if the second-order moments of the distribution of $\mathbf{z}$ exist, then so do those of the distribution of $\mathbf{x}$ and [in light of result (9.47)]

$$\mathrm{var}(\mathbf{x}) = c\boldsymbol{\Sigma}, \tag{9.62}$$

where c is the variance of any element of $\mathbf{z}$—every element of $\mathbf{z}$ has the same variance.

If the distribution of $\mathbf{z}$ has a moment generating function, say $\omega(\cdot)$, then there exists a (nonnegative) function $\phi(\cdot)$ (of a single nonnegative variable) such that (for every $N \times 1$ vector $\mathbf{s}$ in a neighborhood of $\mathbf{0}$) $\omega(\mathbf{s}) = \phi(\mathbf{s}'\mathbf{s})$, and the distribution of $\mathbf{x}$ has the moment generating function $\psi(\cdot)$, where (for every $M \times 1$ vector $\mathbf{t}$ in a neighborhood of $\mathbf{0}$)

$$\psi(\mathbf{t}) = \mathrm{E}\big(e^{\mathbf{t}'\mathbf{x}}\big) = \mathrm{E}\big[e^{\mathbf{t}'(\boldsymbol{\mu}+\boldsymbol{\Gamma}'\mathbf{z})}\big] = e^{\mathbf{t}'\boldsymbol{\mu}}\mathrm{E}\big[e^{(\boldsymbol{\Gamma}\mathbf{t})'\mathbf{z}}\big] = e^{\mathbf{t}'\boldsymbol{\mu}}\omega(\boldsymbol{\Gamma}\mathbf{t}) = e^{\mathbf{t}'\boldsymbol{\mu}}\phi(\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}). \tag{9.63}$$

Note that the moment generating function of the distribution of $\mathbf{x}$ and hence the distribution itself depend on the value of the $N \times M$ matrix $\boldsymbol{\Gamma}$ only through the value of the $M \times M$ matrix $\boldsymbol{\Sigma}$.

Marginal distributions (of spherically distributed random vectors). Let $\mathbf{z}$ represent an N-dimensional spherically distributed random column vector. And take $\mathbf{z}_*$ to be an M-dimensional subvector of $\mathbf{z}$ (where $M < N$), say the subvector obtained by striking out all of the elements of $\mathbf{z}$ save the $i_1, i_2, \ldots, i_M$th elements.

Suppose that the distribution of $\mathbf{z}$ has a moment generating function, say $\psi(\cdot)$. Then, necessarily, there exists a (nonnegative) function $\phi(\cdot)$ (of a single nonnegative variable) such that $\psi(\mathbf{s}) = \phi(\mathbf{s}'\mathbf{s})$ (for every $N \times 1$ vector $\mathbf{s}$ in a neighborhood of $\mathbf{0}$). Clearly, the subvector $\mathbf{z}_*$ can be regarded as a special case of the random column vector $\mathbf{x}$ defined by expression (9.60); it is the special case obtained by setting $\boldsymbol{\mu} = \mathbf{0}$ and taking $\boldsymbol{\Gamma}$ to be the $N \times M$ matrix whose first, second, ..., Mth columns are, respectively, the $i_1, i_2, \ldots, i_M$th columns of $\mathbf{I}_N$. And (in light of the results of the preceding part of the present subsection) it follows that the distribution of $\mathbf{z}_*$ has a moment generating function, say $\psi_*(\cdot)$, and that (for every $M \times 1$ vector $\mathbf{t}$ in some neighborhood of $\mathbf{0}$)

$$\psi_*(\mathbf{t}) = \phi(\mathbf{t}'\mathbf{t}). \tag{9.64}$$

Thus, the moment generating function of the distribution of the subvector $\mathbf{z}_*$ is characterized by the same function $\phi(\cdot)$ as that of the distribution of $\mathbf{z}$ itself.

Suppose now that $\mathbf{u}$ is an M-dimensional random column vector whose distribution has a moment generating function, say $\omega(\cdot)$, and that (for every $M \times 1$ vector $\mathbf{t}$ in a neighborhood of $\mathbf{0}$) $\omega(\mathbf{t}) = \phi(\mathbf{t}'\mathbf{t})$. Then, the distribution of $\mathbf{u}$ is spherical. Moreover, it has the same moment generating function as the distribution of $\mathbf{z}_*$ (and, consequently, $\mathbf{u} \sim \mathbf{z}_*$). There is an implication that the elements of $\mathbf{u}$, like those of $\mathbf{z}_*$, have the same variance as the elements of $\mathbf{z}$.

The moment generating function of a marginal distribution of $\mathbf{z}$ (i.e., of the distribution of a subvector of $\mathbf{z}$) is characterized by the same function $\phi(\cdot)$ as that of the distribution of $\mathbf{z}$ itself. In the case of pdfs, the relationship is more complex.

Suppose that the distribution of the N-dimensional spherically distributed random column vector $\mathbf{z}$ is an absolutely continuous spherical distribution. Then, the distribution of $\mathbf{z}$ has a pdf $f(\cdot)$, where $f(\mathbf{z}) = g(\mathbf{z}'\mathbf{z})$ for some (nonnegative) function $g(\cdot)$ of a single nonnegative variable (and for every value of $\mathbf{z}$). Accordingly, the distribution of the M-dimensional subvector $\mathbf{z}_*$ is the absolutely continuous distribution with pdf $f_*(\cdot)$ defined (for every value of $\mathbf{z}_*$) by

$$f_*(\mathbf{z}_*) = \int_{\mathbb{R}^{N-M}} g(\mathbf{z}_*'\mathbf{z}_* + \bar{\mathbf{z}}_*'\bar{\mathbf{z}}_*)\, d\bar{\mathbf{z}}_*,$$

where $\bar{\mathbf{z}}_*$ is the $(N-M)$-dimensional subvector of $\mathbf{z}$ obtained by striking out the $i_1, i_2, \ldots, i_M$th

elements. And upon regarding $g(\mathbf{z}_*'\mathbf{z}_* + w)$ as a function of a nonnegative variable w and applying result (9.57), we find that (for every value of $\mathbf{z}_*$)

$$f_*(\mathbf{z}_*) = \frac{2\pi^{(N-M)/2}}{\Gamma[(N-M)/2]} \int_0^\infty s^{N-M-1} g(\mathbf{z}_*'\mathbf{z}_* + s^2)\, ds. \tag{9.65}$$

Clearly, $f(\mathbf{z}_*)$ depends on the value of $\mathbf{z}_*$ only through $\mathbf{z}_*'\mathbf{z}_*$, so that (as could have been anticipated from our results on the moment generating function of the distribution of a subvector of a spherically distributed random vector) the distribution of $\mathbf{z}_*$ is spherical. Further, upon introducing the changes of variable $w = s^2$ and $u = \mathbf{z}_*'\mathbf{z}_* + w$, we obtain the following variations on expression (9.65):

$$f_*(\mathbf{z}_*) = \frac{\pi^{(N-M)/2}}{\Gamma[(N-M)/2]} \int_0^\infty w^{[(N-M)/2]-1} g(\mathbf{z}_*'\mathbf{z}_* + w)\, dw \tag{9.66}$$

$$= \frac{\pi^{(N-M)/2}}{\Gamma[(N-M)/2]} \int_{\mathbf{z}_*'\mathbf{z}_*}^\infty (u - \mathbf{z}_*'\mathbf{z}_*)^{[(N-M)/2]-1} g(u)\, du. \tag{9.67}$$

Elliptical distributions: definition. The distribution of a random column vector of the form of the vector $\mathbf{x}$ of equality (9.60) is said to be *elliptical*. And a random column vector whose distribution is that of the vector $\mathbf{x}$ of equality (9.60) may be referred to as being distributed elliptically about $\boldsymbol{\mu}$ or, in the special case where $\boldsymbol{\Gamma} = \mathbf{I}$ (or where $\boldsymbol{\Gamma}$ is orthogonal), as being distributed spherically about $\boldsymbol{\mu}$. Clearly, a random column vector $\mathbf{x}$ is distributed elliptically about $\boldsymbol{\mu}$ if and only if $\mathbf{x} - \boldsymbol{\mu}$ is distributed elliptically about $\mathbf{0}$ and is distributed spherically about $\boldsymbol{\mu}$ if and only if $\mathbf{x} - \boldsymbol{\mu}$ is distributed spherically about $\mathbf{0}$. Let us consider the definition of an elliptical distribution as applied to distributions whose second-order moments exist.

For any $M \times 1$ vector $\boldsymbol{\mu}$ and any $M \times M$ nonnegative definite matrix $\boldsymbol{\Sigma}$, an $M \times 1$ random vector $\mathbf{x}$ has an elliptical distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ if and only if

$$\mathbf{x} \sim \boldsymbol{\mu} + \boldsymbol{\Gamma}'\mathbf{z} \tag{9.68}$$

for some matrix $\boldsymbol{\Gamma}$ such that $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}'\boldsymbol{\Gamma}$ and some random (column) vector $\mathbf{z}$ (of compatible dimension) having a spherical distribution with variance-covariance matrix $\mathbf{I}$—recall that if a random column vector $\mathbf{z}$ has a spherical distribution with a variance-covariance matrix that is a nonzero scalar multiple $c\mathbf{I}$ of $\mathbf{I}$, then the rescaled vector $c^{-1/2}\mathbf{z}$ has a spherical distribution with variance-covariance matrix $\mathbf{I}$. In connection with condition (9.68), define $K = \text{rank}\, \boldsymbol{\Sigma}$, and denote by N the number of rows in the matrix $\boldsymbol{\Gamma}$ (or, equivalently, the number of elements in $\mathbf{z}$)—necessarily, $N \geq K$. For any particular N, the distribution of $\boldsymbol{\mu} + \boldsymbol{\Gamma}'\mathbf{z}$ does not depend on the choice of $\boldsymbol{\Gamma}$ [as is evident (for the case where the distribution of $\mathbf{z}$ has a moment generating function) from result (9.63)]; rather, it depends only on $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and the distribution of $\mathbf{z}$.

Now, consider the distribution of $\boldsymbol{\mu} + \boldsymbol{\Gamma}'\mathbf{z}$ for different choices of N. Assume that $K \geq 1$—if $K = 0$, then (for any choice of N) $\boldsymbol{\Gamma} = \mathbf{0}$ and hence $\boldsymbol{\mu} + \boldsymbol{\Gamma}'\mathbf{z} = \boldsymbol{\mu}$. And take $\boldsymbol{\Gamma}_*$ to be a $K \times M$ matrix such that $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}_*'\boldsymbol{\Gamma}_*$, and take $\mathbf{z}_*$ to be a $K \times 1$ random vector having a spherical distribution with variance-covariance matrix $\mathbf{I}_K$.

Suppose that the distribution of $\mathbf{z}_*$ has a moment generating function, say $\omega_*(\cdot)$. Then, because the distribution of $\mathbf{z}_*$ is spherical, there exists a (nonnegative) function $\phi(\cdot)$ (of a single nonnegative variable) such that (for every $K \times 1$ vector $\mathbf{t}_*$ in a neighborhood of $\mathbf{0}$) $\omega_*(\mathbf{t}_*) = \phi(\mathbf{t}_*'\mathbf{t}_*)$.

Take $\omega(\mathbf{t})$ to be the function of an $N \times 1$ vector $\mathbf{t}$ defined (for every value of $\mathbf{t}$ in some neighborhood of $\mathbf{0}$) by $\omega(\mathbf{t}) = \phi(\mathbf{t}'\mathbf{t})$. There may or may not exist an (N-dimensional) distribution having $\omega(\cdot)$ as a moment generating function. If such a distribution exists, then that distribution is spherical, and for any random vector, say $\mathbf{w}$, having that distribution, the distribution of $\mathbf{z}_*$ is a marginal distribution of $\mathbf{w}$ and $\text{var}(\mathbf{w}) = \mathbf{I}_N$. Accordingly, if there exists a distribution having $\omega(\cdot)$ as a moment generating function, then the distribution of the random vector $\mathbf{z}$ [in expression (9.68)] could be taken to be

that distribution, in which case the distribution of $\boldsymbol{\mu} + \boldsymbol{\Gamma}'\mathbf{z}$ would have the same moment generating function as the distribution of $\boldsymbol{\mu} + \boldsymbol{\Gamma}_*'\mathbf{z}_*$ [as is evident from result (9.63)] and it would follow that $\boldsymbol{\mu} + \boldsymbol{\Gamma}'\mathbf{z} \sim \boldsymbol{\mu} + \boldsymbol{\Gamma}_*'\mathbf{z}_*$.

Thus, as long as there exists an (N-dimensional) distribution having $\omega(\cdot)$ as a moment generating function [where $\omega(\mathbf{t}) = \phi(\mathbf{t}'\mathbf{t})$] and as long as the distribution of $\mathbf{z}$ is taken to be that distribution, the distribution of $\boldsymbol{\mu} + \boldsymbol{\Gamma}'\mathbf{z}$ is invariant to the choice of N. This invariance extends to every N for which there exists an (N-dimensional) distribution having $\omega(\cdot)$ as a moment generating function.

Let us refer to the function $\phi(\cdot)$ as the *mgf generator* of the distribution of the M-dimensional random vector $\boldsymbol{\mu} + \boldsymbol{\Gamma}_*'\mathbf{z}_*$ (with mgf being regarded as an acronym for moment generating function). The moment generating function of the distribution of $\boldsymbol{\mu} + \boldsymbol{\Gamma}_*'\mathbf{z}_*$ is the function $\psi(\cdot)$ defined (for every $M \times 1$ vector $\mathbf{t}$ in a neighborhood of $\mathbf{0}$) by

$$\psi(\mathbf{t}) = e^{\mathbf{t}'\boldsymbol{\mu}}\phi(\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}) \tag{9.69}$$

[as is evident from result (9.63)]. The distribution of $\boldsymbol{\mu} + \boldsymbol{\Gamma}_*'\mathbf{z}_*$ is completely determined by the mean vector $\boldsymbol{\mu}$, the variance-covariance matrix $\boldsymbol{\Sigma}$, and the mgf generator $\phi(\cdot)$. Accordingly, we may refer to this distribution as an (M-dimensional) elliptical distribution with mean $\boldsymbol{\mu}$, variance-covariance matrix $\boldsymbol{\Sigma}$, and mgf generator $\phi(\cdot)$. The mgf generator $\phi(\cdot)$ serves to identify the applicable distribution of $\mathbf{z}_*$; alternatively, some other characteristic of the distribution of $\mathbf{z}_*$ could be used for that purpose (e.g., the pdf). Note that the $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution is an elliptical distribution with mean $\boldsymbol{\mu}$, variance-covariance matrix $\boldsymbol{\Sigma}$, and mgf generator $\phi_*(\cdot)$, where (for every nonnegative scalar u) $\phi_*(u) = \exp(u/2)$.

Pdf of an elliptical distribution. Let $\mathbf{x} = (x_1, x_2, \ldots, x_M)'$ represent an $M \times 1$ random vector, and suppose that for some $M \times 1$ (nonrandom) vector $\boldsymbol{\mu}$ and some $M \times M$ (nonrandom) positive definite matrix $\boldsymbol{\Sigma}$,

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Gamma}'\mathbf{z},$$

where $\boldsymbol{\Gamma}$ is an $M \times M$ (nonsingular) matrix such that $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}'\boldsymbol{\Gamma}$ and where $\mathbf{z} = (z_1, z_2, \ldots, z_M)'$ is an $M \times 1$ spherically distributed random vector with variance-covariance matrix $\mathbf{I}$. Then, $\mathbf{x}$ has an elliptical distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

Now, suppose that the distribution of $\mathbf{z}$ is an absolutely continuous spherical distribution. Then, the distribution of $\mathbf{z}$ is absolutely continuous with a pdf $h(\cdot)$ defined as follows in terms of some (nonnegative) function $g(\cdot)$ (of a single nonnegative variable) for which $\int_0^\infty s^{M-1}g(s^2)\,ds < \infty$:

$$h(\mathbf{z}) = c^{-1}g(\mathbf{z}'\mathbf{z}),$$

where $c = [2\pi^{M/2}/\Gamma(M/2)]\int_0^\infty s^{M-1}g(s^2)\,ds$. And the distribution of $\mathbf{x}$ is absolutely continuous with a pdf, say $f(\cdot)$, that is derivable from the pdf of the distribution of $\mathbf{z}$.

Let us derive an expression for $f(\mathbf{x})$. Clearly, $\mathbf{z} = (\boldsymbol{\Gamma}')^{-1}(\mathbf{x} - \boldsymbol{\mu})$, and the $M \times M$ matrix with ijth element $\partial z_i/\partial x_j$ equals $(\boldsymbol{\Gamma}')^{-1}$. Moreover,

$$\begin{aligned}|\det(\boldsymbol{\Gamma}')^{-1}| = |\det\boldsymbol{\Gamma}'|^{-1} = [(\det\boldsymbol{\Gamma}')^2]^{-1/2} &= [(\det\boldsymbol{\Gamma}')\det\boldsymbol{\Gamma}]^{-1/2} \\ &= [\det(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})]^{-1/2} = (\det\boldsymbol{\Sigma})^{-1/2}.\end{aligned}$$

Thus, making use of standard results on a change of variables (e.g., Bickel and Doksum 2001, sec. B.2) and observing that $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\Gamma}')^{-1} = [(\boldsymbol{\Gamma}')^{-1}]'(\boldsymbol{\Gamma}')^{-1}$, we find that

$$f(\mathbf{x}) = c^{-1}|\boldsymbol{\Sigma}|^{-1/2}g[(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})]. \tag{9.70}$$

Linear transformation of an elliptically distributed random vector. Let $\mathbf{x}$ represent an $N \times 1$ random vector that has an (N-dimensional) elliptical distribution with mean $\boldsymbol{\mu}$, variance-covariance matrix $\boldsymbol{\Sigma}$, and (if $\boldsymbol{\Sigma} \neq \mathbf{0}$) mgf generator $\phi(\cdot)$. And take $\mathbf{y}$ to be the $M \times 1$ random vector obtained by transforming $\mathbf{x}$ as follows:

$$\mathbf{y} = \mathbf{c} + \mathbf{A}\mathbf{x}, \tag{9.71}$$

where $\mathbf{c}$ is an $M \times 1$ (nonrandom) vector and $\mathbf{A}$ an $M \times N$ (nonrandom) matrix. Then, $\mathbf{y}$ has an (M-dimensional) elliptical distribution with mean $\mathbf{c} + \mathbf{A}\boldsymbol{\mu}$, variance-covariance matrix $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$, and (if $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' \neq \mathbf{0}$) mgf generator $\phi(\cdot)$ (identical to the mgf generator of the distribution of $\mathbf{x}$).

Let us verify that $\mathbf{y}$ has this distribution. Define $K = \text{rank}\,\boldsymbol{\Sigma}$. And suppose that $K > 0$ (or, equivalently, that $\boldsymbol{\Sigma} \neq \mathbf{0}$), in which case

$$\mathbf{x} \sim \boldsymbol{\mu} + \boldsymbol{\Gamma}'\mathbf{z},$$

where $\boldsymbol{\Gamma}$ is any $K \times N$ (nonrandom) matrix such that $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}'\boldsymbol{\Gamma}$ and where $\mathbf{z}$ is a $K \times 1$ random vector having a spherical distribution with a moment generating function $\omega(\cdot)$ defined (for every $K \times 1$ vector $\mathbf{s}$ in a neighborhood of $\mathbf{0}$) by $\omega(\mathbf{s}) = \phi(\mathbf{s}'\mathbf{s})$. Then,

$$\mathbf{y} \sim \mathbf{c} + \mathbf{A}(\boldsymbol{\mu} + \boldsymbol{\Gamma}'\mathbf{z}) = \mathbf{c} + \mathbf{A}\boldsymbol{\mu} + (\boldsymbol{\Gamma}\mathbf{A}')'\mathbf{z}. \tag{9.72}$$

Now, let $K_* = \text{rank}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$, and observe that $K_* \leq K$ and that $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' = (\boldsymbol{\Gamma}\mathbf{A}')'\boldsymbol{\Gamma}\mathbf{A}'$. Further, suppose that $K_* > 0$ (or, equivalently, that $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' \neq \mathbf{0}$), take $\boldsymbol{\Gamma}_*$ to be any $K_* \times M$ (nonrandom) matrix such that $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' = \boldsymbol{\Gamma}_*'\boldsymbol{\Gamma}_*$, and take $\mathbf{z}_*$ to be a $K_* \times 1$ random vector having a distribution that is a marginal distribution of $\mathbf{z}$ and that, consequently, has a moment generating function $\omega_*(\cdot)$ defined (for every $K_* \times 1$ vector $\mathbf{s}_*$ in a neighborhood of $\mathbf{0}$) by $\omega_*(\mathbf{s}_*) = \phi(\mathbf{s}_*'\mathbf{s}_*)$. Then, it follows from what was established earlier (in defining elliptical distributions) that

$$\mathbf{c} + \mathbf{A}\boldsymbol{\mu} + (\boldsymbol{\Gamma}\mathbf{A}')'\mathbf{z} \sim \mathbf{c} + \mathbf{A}\boldsymbol{\mu} + \boldsymbol{\Gamma}_*'\mathbf{z}_*,$$

which [in combination with result (9.72)] implies that $\mathbf{y}$ has an elliptical distribution with mean $\mathbf{c} + \mathbf{A}\boldsymbol{\mu}$, variance-covariance matrix $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$, and mgf generator $\phi(\cdot)$. It remains only to observe that even in the "degenerate" case where $\boldsymbol{\Sigma} = \mathbf{0}$ or, more generally, where $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' = \mathbf{0}$, $\text{E}(\mathbf{y}) = \mathbf{c} + \mathbf{A}\boldsymbol{\mu}$ and $\text{var}(\mathbf{y}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ (and to observe that the distribution of a random vector whose variance-covariance matrix equals a null matrix qualifies as an elliptical distribution).

Marginal distributions (of elliptically distributed random vectors). Let $\mathbf{x}$ represent an $N \times 1$ random vector that has an (N-dimensional) elliptical distribution with mean $\boldsymbol{\mu}$, nonnull variance-covariance matrix $\boldsymbol{\Sigma}$, and mgf generator $\phi(\cdot)$. And take $\mathbf{x}_*$ to be an M-dimensional subvector of $\mathbf{x}$ (where $M < N$), say the subvector obtained by striking out all of the elements of $\mathbf{x}$ save the $i_1, i_2, \ldots, i_M$th elements. Further, take $\boldsymbol{\mu}_*$ to be the M-dimensional subvector of $\boldsymbol{\mu}$ obtained by striking out all of the elements of $\boldsymbol{\mu}$ save the $i_1, i_2, \ldots, i_M$th elements and $\boldsymbol{\Sigma}_*$ to be the $M \times M$ submatrix of $\boldsymbol{\Sigma}$ obtained by striking out all of the rows and columns of $\boldsymbol{\Sigma}$ save the $i_1, i_2, \ldots, i_M$th rows and columns.

Consider the distribution of $\mathbf{x}_*$. Clearly, $\mathbf{x}_* = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is the $M \times N$ matrix whose first, second, ..., Mth rows are, respectively, the $i_1, i_2, \ldots, i_M$th rows of $\mathbf{I}_N$. Thus, upon observing that $\mathbf{A}\boldsymbol{\mu} = \boldsymbol{\mu}_*$ and that $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' = \boldsymbol{\Sigma}_*$, it follows from the result of the preceding subsection (the subsection pertaining to linear transformation of elliptically distributed random vectors) that $\mathbf{x}_*$ has an elliptical distribution with mean $\boldsymbol{\mu}_*$, variance-covariance matrix $\boldsymbol{\Sigma}_*$, and (if $\boldsymbol{\Sigma}_* \neq \mathbf{0}$) mgf generator $\phi(\cdot)$ (identical to the mgf generator of $\mathbf{x}$).

d. Maximum likelihood as applied to elliptical distributions (besides the MVN distribution)

Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a general linear model. Suppose further that the variance-covariance matrix $\mathbf{V}(\boldsymbol{\theta})$ of the vector $\mathbf{e}$ of residual effects is nonsingular and that

$$\mathbf{e} \sim [\boldsymbol{\Gamma}(\boldsymbol{\theta})]'\mathbf{u}, \tag{9.73}$$

where $\boldsymbol{\Gamma}(\boldsymbol{\theta})$ is an $N \times N$ (nonsingular) matrix (whose elements may be functionally dependent on $\boldsymbol{\theta}$) such that $\mathbf{V}(\boldsymbol{\theta}) = [\boldsymbol{\Gamma}(\boldsymbol{\theta})]'\boldsymbol{\Gamma}(\boldsymbol{\theta})$ and where $\mathbf{u}$ is an $N \times 1$ random vector having an absolutely

continuous spherical distribution with variance-covariance matrix $\mathbf{I}$. The distribution of $\mathbf{u}$ has a pdf $h(\cdot)$, where (for every value of $\mathbf{u}$)

$$h(\mathbf{u}) = c^{-1} g(\mathbf{u}'\mathbf{u}).$$

Here, $g(\cdot)$ is a nonnegative function (of a single nonnegative variable) such that $\int_0^\infty s^{N-1} g(s^2)\, ds < \infty$, and $c = [2\pi^{N/2}/\Gamma(N/2)] \int_0^\infty s^{N-1} g(s^2)\, ds$. As a consequence of supposition (9.73), $\mathbf{y}$ has an elliptical distribution.

Let us consider the ML estimation of functions of the parameters of the general linear model (i.e., functions of the elements $\beta_1, \beta_2, \ldots, \beta_P$ of the vector $\boldsymbol{\beta}$ and the elements $\theta_1, \theta_2, \ldots, \theta_T$ of the vector $\boldsymbol{\theta}$). That topic was considered earlier (in Subsection a) in the special case where $\mathbf{e} \sim N[\mathbf{0}, \mathbf{V}(\boldsymbol{\theta})]$—when $g(s^2) = \exp(-s^2/2)$, $h(\mathbf{u}) = (2\pi)^{-N/2} \exp(-\frac{1}{2}\mathbf{u}'\mathbf{u})$, which is the pdf of the $N(\mathbf{0}, \mathbf{I}_N)$ distribution.

Let $f(\cdot\,; \boldsymbol{\beta}, \boldsymbol{\theta})$ represent the pdf of the distribution of $\mathbf{y}$, and denote by $\underline{\mathbf{y}}$ the observed value of $\mathbf{y}$. Then, the likelihood function is the function, say $L(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}})$ of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ defined by $L(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}}) = f(\underline{\mathbf{y}}; \boldsymbol{\beta}, \boldsymbol{\theta})$. Accordingly, it follows from result (9.70) that

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}}) = c^{-1} |\mathbf{V}(\boldsymbol{\theta})|^{-1/2} g\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' [\mathbf{V}(\boldsymbol{\theta})]^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\}. \tag{9.74}$$

Maximum likelihood estimates of functions of $\boldsymbol{\beta}$ and/or $\boldsymbol{\theta}$ are obtained from values, say $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\theta}}$, of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ at which $L(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}})$ attains its maximum value: a maximum likelihood estimate of a function, say $r(\boldsymbol{\beta}, \boldsymbol{\theta})$, of $\boldsymbol{\beta}$ and/or $\boldsymbol{\theta}$ is provided by the quantity $r(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})$ obtained by substituting $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\theta}}$ for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

Profile likelihood function. Now, suppose that the function $g(\cdot)$ is a strictly decreasing function (as in the special case where the distribution of $\mathbf{e}$ is MVN). Then, for any particular value of $\boldsymbol{\theta}$, the maximization of $L(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}})$ with respect to $\boldsymbol{\beta}$ is equivalent to the minimization of $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' [\mathbf{V}(\boldsymbol{\theta})]^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. Thus, upon regarding the value of $\boldsymbol{\theta}$ as "fixed," upon recalling (from Part 3 of Subsection a) that the linear system

$$\mathbf{X}' [\mathbf{V}(\boldsymbol{\theta})]^{-1} \mathbf{X}\mathbf{b} = \mathbf{X}' [\mathbf{V}(\boldsymbol{\theta})]^{-1} \underline{\mathbf{y}} \tag{9.75}$$

(in the $P \times 1$ vector $\mathbf{b}$) is consistent, and upon employing the same line of reasoning as in Part 3 of Subsection a, we find that $L(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}})$ attains its maximum value at a value $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ of $\boldsymbol{\beta}$ if and only if $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ is a solution to linear system (9.75) or, equivalently, if and only if

$$\mathbf{X}' [\mathbf{V}(\boldsymbol{\theta})]^{-1} \mathbf{X} \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \mathbf{X}' [\mathbf{V}(\boldsymbol{\theta})]^{-1} \underline{\mathbf{y}},$$

in which case

$$L[\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}; \underline{\mathbf{y}}] = c^{-1} |\mathbf{V}(\boldsymbol{\theta})|^{-1/2} g\{[\underline{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})]' [\mathbf{V}(\boldsymbol{\theta})]^{-1} [\underline{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})]\} \tag{9.76}$$

$$= c^{-1} |\mathbf{V}(\boldsymbol{\theta})|^{-1/2} g\{\underline{\mathbf{y}}' [\mathbf{V}(\boldsymbol{\theta})]^{-1} \underline{\mathbf{y}} - [\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})]' \mathbf{X}' [\mathbf{V}(\boldsymbol{\theta})]^{-1} \underline{\mathbf{y}}\} \tag{9.77}$$

$$= c^{-1} |\mathbf{V}(\boldsymbol{\theta})|^{-1/2} g\{\underline{\mathbf{y}}' ([\mathbf{V}(\boldsymbol{\theta})]^{-1} - [\mathbf{V}(\boldsymbol{\theta})]^{-1} \mathbf{X}\{\mathbf{X}' [\mathbf{V}(\boldsymbol{\theta})]^{-1} \mathbf{X}\}^{-} \mathbf{X}' [\mathbf{V}(\boldsymbol{\theta})]^{-1}) \underline{\mathbf{y}}\}. \tag{9.78}$$

Accordingly, the function $L_*(\boldsymbol{\theta}; \underline{\mathbf{y}})$ of $\boldsymbol{\theta}$ defined by $L_*(\boldsymbol{\theta}; \underline{\mathbf{y}}) = L[\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}; \underline{\mathbf{y}}]$ is a profile likelihood function.

Values, say $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\theta}}$ (of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively), at which $L(\boldsymbol{\beta}, \boldsymbol{\theta}; \underline{\mathbf{y}})$ attains its maximum value can be obtained by taking $\tilde{\boldsymbol{\theta}}$ to be a value at which the profile likelihood function $L_*(\boldsymbol{\theta}; \underline{\mathbf{y}})$ attains its maximum value and by then taking $\tilde{\boldsymbol{\beta}}$ to be a solution to the linear system

$$\mathbf{X}' [\mathbf{V}(\tilde{\boldsymbol{\theta}})]^{-1} \mathbf{X}\mathbf{b} = \mathbf{X}' [\mathbf{V}(\tilde{\boldsymbol{\theta}})]^{-1} \underline{\mathbf{y}}.$$

Except for relatively simple special cases, the maximization of $L_*(\boldsymbol{\theta}; \underline{\mathbf{y}})$ must be accomplished numerically via an iterative procedure.

REML variant. REML is a variant of ML in which inferences about functions of $\boldsymbol{\theta}$ are based on the likelihood function associated with a vector of so-called error contrasts. REML was introduced and discussed in an earlier subsection (Subsection b) under the assumption that the distribution of $\mathbf{e}$ is MVN. Let us consider REML in the present, more general context (where the distribution of $\mathbf{e}$ is taken to be elliptical).

Let $\mathbf{R}$ represent an $N \times (N - \operatorname{rank}\mathbf{X})$ matrix (of constants) of full column rank $N - \operatorname{rank}\mathbf{X}$ such that $\mathbf{X}'\mathbf{R} = \mathbf{0}$, and take $\mathbf{z}$ to be the $(N - \operatorname{rank}\mathbf{X}) \times 1$ vector defined by $\mathbf{z} = \mathbf{R}'\mathbf{y}$. Note that $\mathbf{z} = \mathbf{R}'\mathbf{e}$ and hence that the distribution of $\mathbf{z}$ does not depend on $\boldsymbol{\beta}$. Further, let $k(\cdot\,;\boldsymbol{\theta})$ represent the pdf of the distribution of $\mathbf{z}$, and take $L(\boldsymbol{\theta};\mathbf{R}'\underline{\mathbf{y}})$ to be the function of $\boldsymbol{\theta}$ defined (for $\boldsymbol{\theta} \in \Theta$) by $L(\boldsymbol{\theta};\mathbf{R}'\underline{\mathbf{y}}) = k(\mathbf{R}'\underline{\mathbf{y}};\boldsymbol{\theta})$. The function $L(\boldsymbol{\theta};\mathbf{R}'\underline{\mathbf{y}})$ is a likelihood function; it is the likelihood function obtained by regarding the value of $\mathbf{z}$ as the data vector.

Now, suppose that the (N-dimensional spherical) distribution of the random vector $\mathbf{u}$ [in expression (9.73)] has a moment generating function, say $\psi(\cdot)$. Then, necessarily, there exists a (nonnegative) function $\phi(\cdot)$ (of a single nonnegative variable) such that (for every $N \times 1$ vector $\mathbf{t}$ in a neighborhood of $\mathbf{0}$) $\psi(\mathbf{t}) = \phi(\mathbf{t}'\mathbf{t})$. And in light of the results of Subsection c, it follows that $\mathbf{z}$ has an [$(N - \operatorname{rank}\mathbf{X})$-dimensional] elliptical distribution with mean $\mathbf{0}$, variance-covariance matrix $\mathbf{R}'\mathbf{V}(\boldsymbol{\theta})\mathbf{R}$, and mgf generator $\phi(\cdot)$. Further,

$$\mathbf{z} \sim [\boldsymbol{\Gamma}_*(\boldsymbol{\theta})]'\mathbf{u}_*,$$

where $\boldsymbol{\Gamma}_*(\boldsymbol{\theta})$ is any $(N - \operatorname{rank}\mathbf{X}) \times (N - \operatorname{rank}\mathbf{X})$ matrix such that $\mathbf{R}'\mathbf{V}(\boldsymbol{\theta})\mathbf{R} = [\boldsymbol{\Gamma}_*(\boldsymbol{\theta})]'\boldsymbol{\Gamma}_*(\boldsymbol{\theta})$ and where $\mathbf{u}_*$ is an $(N - \operatorname{rank}\mathbf{X}) \times 1$ random vector whose distribution is spherical with variance-covariance matrix $\mathbf{I}$ and with moment generating function $\psi_*(\cdot)$ defined [for every $(N - \operatorname{rank}\mathbf{X}) \times 1$ vector $\mathbf{t}_*$ in a neighborhood of $\mathbf{0}$] by $\psi_*(\mathbf{t}_*) = \phi(\mathbf{t}_*'\mathbf{t}_*)$—the distribution of $\mathbf{u}_*$ is a marginal distribution of $\mathbf{u}$.

The distribution of $\mathbf{u}_*$ is absolutely continuous with a pdf $h_*(\cdot)$ that (at least in principle) is determinable from the pdf of the distribution of $\mathbf{u}$ and that is expressible in the form

$$h_*(\mathbf{u}_*) = c_*^{-1} g_*(\mathbf{u}_*'\mathbf{u}_*),$$

where $g_*(\cdot)$ is a nonnegative function (of a single nonnegative variable) such that $\int_0^\infty s^{N-\operatorname{rank}(\mathbf{X})-1} g_*(s^2)\, ds < \infty$ and where c_* is a strictly positive constant. Necessarily,

$$c_* = \frac{2\pi^{(N-\operatorname{rank}\mathbf{X})/2}}{\Gamma[(N-\operatorname{rank}\mathbf{X})/2]} \int_0^\infty s^{N-\operatorname{rank}(\mathbf{X})-1} g_*(s^2)\, ds.$$

Thus, in light of result (9.70), the pdf of the distribution of $\mathbf{z}$ is absolutely continuous with a pdf $k(\cdot\,;\boldsymbol{\theta})$ that is expressible as

$$k(\mathbf{z};\boldsymbol{\theta}) = c_*^{-1}|\mathbf{R}'\mathbf{V}(\boldsymbol{\theta})\mathbf{R}|^{-1/2} g_*\{\mathbf{z}'[\mathbf{R}'\mathbf{V}(\boldsymbol{\theta})\mathbf{R}]^{-1}\mathbf{z}\}.$$

And it follows that the REML likelihood function is expressible as

$$L(\boldsymbol{\theta};\mathbf{R}'\underline{\mathbf{y}}) = c_*^{-1}|\mathbf{R}'\mathbf{V}(\boldsymbol{\theta})\mathbf{R}|^{-1/2} g_*\{\underline{\mathbf{y}}'\mathbf{R}[\mathbf{R}'\mathbf{V}(\boldsymbol{\theta})\mathbf{R}]^{-1}\mathbf{R}'\underline{\mathbf{y}}\}. \tag{9.79}$$

As in the special case where the distribution of $\mathbf{e}$ is MVN, an alternative expression for $L(\boldsymbol{\theta};\mathbf{R}'\underline{\mathbf{y}})$ can be obtained by taking advantage of identities (9.32) and (9.37). Taking $\mathbf{X}_*$ to be any $N \times (\operatorname{rank}\mathbf{X})$ matrix whose columns form a basis for $\mathcal{C}(\mathbf{X})$, we find that

$$\begin{aligned} L(\boldsymbol{\theta};\mathbf{R}'\underline{\mathbf{y}}) = {} & c_*^{-1}|\mathbf{R}'\mathbf{R}|^{-1/2}|\mathbf{X}_*'\mathbf{X}_*|^{1/2}|\mathbf{V}(\boldsymbol{\theta})|^{-1/2}|\mathbf{X}_*'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}_*|^{-1/2} \\ & \times g_*\{\underline{\mathbf{y}}'([\mathbf{V}(\boldsymbol{\theta})]^{-1} - [\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\{\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\}^{-}\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1})\underline{\mathbf{y}}\}. \end{aligned} \tag{9.80}$$

Alternative versions of this expression can be obtained by replacing the argument of the function $g_*(\cdot)$ with expression (9.40) or expression (9.41).

As in the special case where the distribution of $\mathbf{e}$ is MVN, $L(\boldsymbol{\theta};\mathbf{R}'\underline{\mathbf{y}})$ depends on the choice of the matrix $\mathbf{R}$ only through the multiplicative constant $|\mathbf{R}'\mathbf{R}|^{-1/2}$. In some special cases including that where the distribution of $\mathbf{e}$ is MVN, the function $g_*(\cdot)$ differs from the function $g(\cdot)$ by no more than a multiplicative constant. However, in general, the relationship between $g_*(\cdot)$ and $g(\cdot)$ is more complex.

5.10 Prediction

a. Some general results

Let $\mathbf{y}$ represent an $N \times 1$ observable random vector. And consider the use of $\mathbf{y}$ in predicting an unobservable random variable or, more generally, an unobservable random vector, say an $M \times 1$ unobservable random vector $\mathbf{w} = (w_1, w_2, \ldots, w_M)'$. That is, consider the use of the observed value of $\mathbf{y}$ (the so-called data vector) in making inferences about an unobservable quantity that can be regarded as a realization (i.e., sample value) of $\mathbf{w}$. Here, an unobservable quantity is a quantity that is unobservable at the time the inferences are to be made; it may become observable at some future time (as suggested by the use of the word prediction). In the present section, the focus is on obtaining a point estimate of the unobservable quantity; that is, on what might be deemed a point prediction.

Suppose that the second-order moments of the joint distribution of $\mathbf{w}$ and $\mathbf{y}$ exist. And adopt the following notation: $\boldsymbol{\mu}_y = \mathrm{E}(\mathbf{y})$, $\boldsymbol{\mu}_w = \mathrm{E}(\mathbf{w})$, $\mathbf{V}_y = \mathrm{var}(\mathbf{y})$, $\mathbf{V}_{yw} = \mathrm{cov}(\mathbf{y}, \mathbf{w})$, and $\mathbf{V}_w = \mathrm{var}(\mathbf{w})$. Further, in considering the special case $M = 1$, let us write w, μ_w, $\mathbf{v}_{yw}$, and v_w for $\mathbf{w}$, $\boldsymbol{\mu}_w$, $\mathbf{V}_{yw}$, and $\mathbf{V}_w$, respectively.

It is informative to consider the prediction of $\mathbf{w}$ under each of the following states of knowledge: (1) the joint distribution of $\mathbf{y}$ and $\mathbf{w}$ is known; (2) only $\boldsymbol{\mu}_y$, $\boldsymbol{\mu}_w$, $\mathbf{V}_y$, $\mathbf{V}_{yw}$, and $\mathbf{V}_w$ are known; and (3) only $\mathbf{V}_y$, $\mathbf{V}_{yw}$, and $\mathbf{V}_w$ are known.

Let $\tilde{\mathbf{w}}(\mathbf{y})$ represent an $(M \times 1)$-dimensional vector-valued function of $\mathbf{y}$ that qualifies as a (point) predictor of $\mathbf{w}$—in considering the special case where $M = 1$, let us write $\tilde{w}(\mathbf{y})$ for $\tilde{\mathbf{w}}(\mathbf{y})$. That $\tilde{\mathbf{w}}(\mathbf{y})$ qualifies as a predictor implies that the vector-valued function $\tilde{\mathbf{w}}(\cdot)$ depends on the joint distribution of $\mathbf{y}$ and $\mathbf{w}$ (if at all) only through characteristics of the joint distribution that are known.

The difference $\tilde{\mathbf{w}}(\mathbf{y}) - \mathbf{w}$ is referred to as the *prediction error*. The predictor $\tilde{\mathbf{w}}(\mathbf{y})$ is said to be *unbiased* if $\mathrm{E}[\tilde{\mathbf{w}}(\mathbf{y}) - \mathbf{w}] = \mathbf{0}$, that is, if the expected value of the prediction error equals $\mathbf{0}$, or, equivalently, if $\mathrm{E}[\tilde{\mathbf{w}}(\mathbf{y})] = \boldsymbol{\mu}_w$, that is, if the expected value of the predictor is the same as that of the random vector $\mathbf{w}$ whose realization is being predicted.

Attention is sometimes restricted to linear predictors. An $(M \times 1)$-dimensional vector-valued function $\mathbf{t}(\mathbf{y})$ of $\mathbf{y}$ is said to be *linear* if it is expressible in the form

$$\mathbf{t}(\mathbf{y}) = \mathbf{c} + \mathbf{A}'\mathbf{y}, \tag{10.1}$$

where $\mathbf{c}$ is an $M \times 1$ vector of constants and $\mathbf{A}$ is an $M \times N$ matrix of constants. A vector-valued function $\mathbf{t}(\mathbf{y})$ that is expressible in the form (10.1) is regarded as linear even if the vector $\mathbf{c}$ and the matrix $\mathbf{A}$ depend on the joint distribution of $\mathbf{y}$ and $\mathbf{w}$—the linearity reflects the nature of the dependence on the value of $\mathbf{y}$, not the nature of any dependence on the joint distribution. And it qualifies as a predictor if any dependence on the joint distribution of $\mathbf{y}$ and $\mathbf{w}$ is confined to characteristics of the joint distribution that are known.

The $M \times M$ matrix $\mathrm{E}\{[\tilde{\mathbf{w}}(\mathbf{y}) - \mathbf{w}][\tilde{\mathbf{w}}(\mathbf{y}) - \mathbf{w}]'\}$ is referred to as the *mean-squared-error* (MSE) *matrix* of the predictor $\tilde{\mathbf{w}}(\mathbf{y})$. If $\tilde{\mathbf{w}}(\mathbf{y})$ is an unbiased predictor (of $\mathbf{w}$), then

$$\mathrm{E}\{[\tilde{\mathbf{w}}(\mathbf{y}) - \mathbf{w}][\tilde{\mathbf{w}}(\mathbf{y}) - \mathbf{w}]'\} = \mathrm{var}[\tilde{\mathbf{w}}(\mathbf{y}) - \mathbf{w}].$$

That is, the MSE matrix of an unbiased predictor equals the variance-covariance matrix of its prediction error (not the variance-covariance matrix of the predictor itself). Note that in the special case where $M = 1$, the MSE matrix has only one element, which is expressible as $\mathrm{E}\{[\tilde{w}(\mathbf{y}) - w]^2\}$ and which is referred to simply as the *mean squared error* (MSE) of the (scalar-valued) predictor $\tilde{w}(\mathbf{y})$.

State (1): joint distribution known. Suppose that the joint distribution of $\mathbf{y}$ and $\mathbf{w}$ is known or that, at the very least, enough is known about the joint distribution to determine the conditional expected

value $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ of $\mathbf{w}$ given $\mathbf{y}$. And observe that

$$\mathrm{E}[\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w} \mid \mathbf{y}] = \mathbf{0} \quad \text{(with probability 1)}. \tag{10.2}$$

Observe also that, for "any" column vector $\mathbf{h}(\mathbf{y})$ of functions of $\mathbf{y}$,

$$\mathrm{E}\{\mathbf{h}(\mathbf{y})[\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}]' \mid \mathbf{y}\} = \mathbf{0} \quad \text{(with probability 1)}. \tag{10.3}$$

Now, let $\mathbf{t}(\mathbf{y})$ represent "any" $(M \times 1)$-dimensional vector-valued function of $\mathbf{y}$—in the special case where $M = 1$, let us write $t(\mathbf{y})$ for $\mathbf{t}(\mathbf{y})$. Then, upon observing that

$$\begin{aligned}[\mathbf{t}(\mathbf{y}) - \mathbf{w}][\mathbf{t}(\mathbf{y}) - \mathbf{w}]' &= \{\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y}) + [\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}]\}\{\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y}) + [\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}]\}' \\ &= [\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})][\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})]' + [\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}][\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}]' \\ &\quad + [\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})][\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}]' + \{[\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})][\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}]'\}'\end{aligned}$$

and [in light of result (10.3)] that

$$\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})][\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}]' \mid \mathbf{y}\} = \mathbf{0} \quad \text{(with probability 1)}, \tag{10.4}$$

we find that

$$\begin{aligned}&\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \mathbf{w}][\mathbf{t}(\mathbf{y}) - \mathbf{w}]' \mid \mathbf{y}\} \\ &\quad = [\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})][\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})]' \\ &\qquad + \mathrm{E}\{[\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}][\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}]' \mid \mathbf{y}\} \quad \text{(with probability 1)}.\end{aligned} \tag{10.5}$$

Result (10.5) implies that $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ is an optimal predictor of $\mathbf{w}$. It is optimal in the sense that the difference $\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \mathbf{w}][\mathbf{t}(\mathbf{y}) - \mathbf{w}]' \mid \mathbf{y}\} - \mathrm{E}\{[\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}][\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}]' \mid \mathbf{y}\}$ between the conditional (given $\mathbf{y}$) MSE matrix of an arbitrary predictor $\mathbf{t}(\mathbf{y})$ and that of $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ equals (with probability 1) the matrix $[\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})][\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})]'$, which is nonnegative definite and which equals $\mathbf{0}$ if and only if $\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y}) = \mathbf{0}$ or, equivalently, if and only if $\mathbf{t}(\mathbf{y}) = \mathrm{E}(\mathbf{w} \mid \mathbf{y})$. In the special case where $M = 1$, we have that [for an arbitrary predictor $t(\mathbf{y})$]

$$\mathrm{E}\{[t(\mathbf{y}) - w]^2 \mid \mathbf{y}\} \geq \mathrm{E}\{[\mathrm{E}(w \mid \mathbf{y}) - w]^2 \mid \mathbf{y}\} \quad \text{(with probability 1)}.$$

It is worth noting that

$$\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \mathbf{w}][\mathbf{t}(\mathbf{y}) - \mathbf{w}]'\} = \mathrm{E}\big(\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \mathbf{w}][\mathbf{t}(\mathbf{y}) - \mathbf{w}]' \mid \mathbf{y}\}\big),$$

so that $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ is optimal when the various predictors are compared on the basis of their unconditional MSE matrices as well as when they are compared on the basis of their conditional MSE matrices. The conditional MSE matrix of the optimal predictor $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ is

$$\mathrm{E}\{[\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}][\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}]' \mid \mathbf{y}\} = \mathrm{var}(\mathbf{w} \mid \mathbf{y}),$$

and the (unconditional) MSE matrix of $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ or, equivalently, the (unconditional) variance-covariance matrix of $\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}$ is

$$\mathrm{var}[\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}] = \mathrm{E}\{[\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}][\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}]'\} = \mathrm{E}[\mathrm{var}(\mathbf{w} \mid \mathbf{y})].$$

Clearly, $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ is an unbiased predictor; in fact, the expected value of its prediction error equals $\mathbf{0}$ conditionally on $\mathbf{y}$ (albeit with probability 1) as well as unconditionally [as is evident from result (10.2)]. Whether or not $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ is a linear predictor (or, more generally, equal to a linear predictor with probability 1) depends on the form of the joint distribution of $\mathbf{y}$ and $\mathbf{w}$; a sufficient (but not a necessary) condition for $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ to be linear (or, at least, "linear with probability 1") is that the joint distribution of $\mathbf{y}$ and $\mathbf{w}$ be MVN.

State (2): only the means and the variances and covariances are known. Suppose that $\boldsymbol{\mu}_y$, $\boldsymbol{\mu}_w$, $\mathbf{V}_y$, $\mathbf{V}_{yw}$, and $\mathbf{V}_w$ are known, but that nothing else is known about the joint distribution of $\mathbf{y}$ and $\mathbf{w}$. Then, $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ is not determinable from what is known, forcing us to look elsewhere for a predictor of $\mathbf{w}$.

Assume (for the sake of simplicity) that $\mathbf{V}_y$ is nonsingular. And consider the predictor

$$\boldsymbol{\eta}(\mathbf{y}) = \boldsymbol{\mu}_w + \mathbf{V}_{yw}'\mathbf{V}_y^{-1}(\mathbf{y} - \boldsymbol{\mu}_y) = \boldsymbol{\tau} + \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{y},$$

where $\boldsymbol{\tau} = \boldsymbol{\mu}_w - \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\boldsymbol{\mu}_y$—in the special case where $M = 1$, let us write $\eta(\mathbf{y})$ for $\boldsymbol{\eta}(\mathbf{y})$.

Clearly, $\boldsymbol{\eta}(\mathbf{y})$ is linear; it is also unbiased. Now, consider its MSE matrix $\mathrm{E}\{[\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}][\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]'\}$ or, equivalently, the variance-covariance matrix $\mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]$ of its prediction error. Let us compare the MSE matrix of $\boldsymbol{\eta}(\mathbf{y})$ with the MSE matrices of other linear predictors.

Let $\mathbf{t}(\mathbf{y})$ represent an $(M \times 1)$-dimensional vector-valued function of $\mathbf{y}$ of the form $\mathbf{t}(\mathbf{y}) = \mathbf{c} + \mathbf{A}'\mathbf{y}$, where $\mathbf{c}$ is an $M \times 1$ vector of constants and $\mathbf{A}$ an $N \times M$ matrix of constants—in the special case where $M = 1$, let us write $t(\mathbf{y})$ for $\mathbf{t}(\mathbf{y})$. Further, decompose the difference between $\mathbf{t}(\mathbf{y})$ and $\mathbf{w}$ into two components as follows:

$$\mathbf{t}(\mathbf{y}) - \mathbf{w} = [\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})] + [\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]. \tag{10.6}$$

And observe that

$$\mathrm{cov}[\mathbf{y},\ \boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}] = \mathrm{cov}(\mathbf{y},\ \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{y} - \mathbf{w}) = \mathbf{V}_y[\mathbf{V}_{yw}'\mathbf{V}_y^{-1}]' - \mathbf{V}_{yw} = \mathbf{0}. \tag{10.7}$$

Then, because $\mathrm{E}[\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}] = \mathbf{0}$ and because $\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y}) = \mathbf{c} - \boldsymbol{\tau} + (\mathbf{A}' - \mathbf{V}_{yw}'\mathbf{V}_y^{-1})\mathbf{y}$, it follows that

$$\begin{aligned}\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})][\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]'\} &= \mathrm{cov}[\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y}),\ \boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}] \\ &= (\mathbf{A}' - \mathbf{V}_{yw}'\mathbf{V}_y^{-1})\,\mathrm{cov}[\mathbf{y},\ \boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}] = \mathbf{0}.\end{aligned} \tag{10.8}$$

Thus,

$$\begin{aligned}&\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \mathbf{w}][\mathbf{t}(\mathbf{y}) - \mathbf{w}]'\} \\ &\quad = \mathrm{E}\big(\{[\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})] + [\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]\}\{[\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})]' + [\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]'\}\big) \\ &\quad = \mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})][\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})]'\} + \mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}].\end{aligned} \tag{10.9}$$

Any linear predictor of $\mathbf{w}$ is expressible in the form $[\mathbf{t}(\mathbf{y}) = \mathbf{c} + \mathbf{A}'\mathbf{y}]$ of the vector-valued function $\mathbf{t}(\mathbf{y})$. Accordingly, result (10.9) implies that $\boldsymbol{\eta}(\mathbf{y})$ is the best linear predictor of $\mathbf{w}$. It is the best linear predictor in the sense that the difference between the MSE matrix $\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \mathbf{w}][\mathbf{t}(\mathbf{y}) - \mathbf{w}]'\}$ of an arbitrary linear predictor $\mathbf{t}(\mathbf{y})$ and the matrix $\mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]$ [which is the MSE matrix of $\boldsymbol{\eta}(\mathbf{y})$] equals the matrix $\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})][\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})]'\}$, which is nonnegative definite and which equals $\mathbf{0}$ if and only if $\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y}) = \mathbf{0}$ or, equivalently, if and only if $\mathbf{t}(\mathbf{y}) = \boldsymbol{\eta}(\mathbf{y})$. (To see that $\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})][\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})]'\} = \mathbf{0}$ implies that $\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y}) = \mathbf{0}$, observe that (for $j = 1, 2, \ldots, M$) the jth element of $\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})$ equals $k_j + \boldsymbol{\ell}_j'\mathbf{y}$, where k_j is the jth element of $\mathbf{c} - \boldsymbol{\tau}$ and $\boldsymbol{\ell}_j$ the jth column of $\mathbf{A} - \mathbf{V}_y^{-1}\mathbf{V}_{yw}$, that the jth diagonal element of $\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})][\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})]'\}$ equals $\mathrm{E}[(k_j + \boldsymbol{\ell}_j'\mathbf{y})^2]$, and that $\mathrm{E}[(k_j + \boldsymbol{\ell}_j'\mathbf{y})^2] = 0$ implies that $\mathrm{E}(k_j + \boldsymbol{\ell}_j'\mathbf{y}) = 0$ and $\mathrm{var}(k_j + \boldsymbol{\ell}_j'\mathbf{y}) = 0$ and hence that $\boldsymbol{\ell}_j = \mathbf{0}$ and $k_j = 0$.) In the special case where $M = 1$, we have [for an arbitrary linear predictor $t(\mathbf{y})$] that

$$\mathrm{E}\{[t(\mathbf{y}) - w]^2\} \geq \mathrm{var}[\eta(\mathbf{y}) - w] \quad \big(= \mathrm{E}\{[\eta(\mathbf{y}) - w]^2\}\big), \tag{10.10}$$

with equality holding in inequality (10.10) if and only if $t(\mathbf{y}) = \eta(\mathbf{y})$.

The prediction error of the best linear predictor $\boldsymbol{\eta}(\mathbf{y})$ can be decomposed into two components on the basis of the following identity:

$$\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w} = [\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})] + [\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}]. \tag{10.11}$$

The second component $\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}$ of this decomposition has an expected value of $\mathbf{0}$ [conditionally on $\mathbf{y}$ (albeit with probability 1) as well as unconditionally], and because $\mathrm{E}[\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}] = \mathbf{0}$, the first

component $\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})$ also has an expected value of $\mathbf{0}$. Moreover, it follows from result (10.3) that

$$\mathrm{E}\{[\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})][\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}]' \mid \mathbf{y}\} = \mathbf{0} \quad \text{(with probability 1)}, \tag{10.12}$$

implying that

$$\mathrm{E}\{[\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})][\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}]'\} = \mathbf{0} \tag{10.13}$$

and hence that the two components $\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})$ and $\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}$ of decomposition (10.11) are uncorrelated. And upon applying result (10.5) [with $\mathbf{t}(\mathbf{y}) = \boldsymbol{\eta}(\mathbf{y})$], we find that

$$\begin{aligned}\mathrm{E}\{[\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}][\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]' \mid \mathbf{y}\} = {} & [\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})][\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})]' \\ & + \mathrm{var}(\mathbf{w} \mid \mathbf{y}) \quad \text{(with probability 1)}. \end{aligned} \tag{10.14}$$

In the special case where $M = 1$, result (10.14) is reexpressible as

$$\mathrm{E}\{[\eta(\mathbf{y}) - w]^2 \mid \mathbf{y}\} = [\eta(\mathbf{y}) - \mathrm{E}(w \mid \mathbf{y})]^2 + \mathrm{var}(w \mid \mathbf{y}) \quad \text{(with probability 1)}.$$

Equality (10.14) serves to decompose the conditional (on $\mathbf{y}$) MSE matrix of the best linear predictor $\boldsymbol{\eta}(\mathbf{y})$ into two components, corresponding to the two components of the decomposition (10.11) of the prediction error of $\boldsymbol{\eta}(\mathbf{y})$. The (unconditional) MSE matrix of $\boldsymbol{\eta}(\mathbf{y})$ or, equivalently, the (unconditional) variance-covariance matrix of $\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}$ lends itself to a similar decomposition. We find that

$$\begin{aligned}\mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}] &= \mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})] + \mathrm{var}[\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}] \\ &= \mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})] + \mathrm{E}[\mathrm{var}(\mathbf{w} \mid \mathbf{y})]. \end{aligned} \tag{10.15}$$

Of the two components of the prediction error of $\boldsymbol{\eta}(\mathbf{y})$, the second component $\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}$ can be regarded as an "inherent" component. It is inherent in the sense that it is an error that would be incurred even if enough were known about the joint distribution of $\mathbf{y}$ and $\mathbf{w}$ that $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ were determinable and were employed as the predictor. The first component $\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})$ of the prediction error can be regarded as a "nonlinearity" component; it equals $\mathbf{0}$ if and only if $\mathrm{E}(\mathbf{w} \mid \mathbf{y}) = \mathbf{c} + \mathbf{A}'\mathbf{y}$ for some vector $\mathbf{c}$ of constants and some matrix $\mathbf{A}$ of constants.

The variance-covariance matrix of the prediction error of $\boldsymbol{\eta}(\mathbf{y})$ is expressible as

$$\begin{aligned}\mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}] &= \mathrm{var}(\mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{y} - \mathbf{w}) \\ &= \mathbf{V}_w + \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{V}_y(\mathbf{V}_{yw}'\mathbf{V}_y^{-1})' - \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{V}_{yw} - [\mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{V}_{yw}]' \\ &= \mathbf{V}_w - \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{V}_{yw}. \end{aligned} \tag{10.16}$$

It differs from the variance-covariance matrix of $\boldsymbol{\eta}(\mathbf{y})$; the latter variance-covariance matrix is expressible as

$$\mathrm{var}[\boldsymbol{\eta}(\mathbf{y})] = \mathrm{var}(\mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{y}) = \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{V}_y[\mathbf{V}_{yw}'\mathbf{V}_y^{-1}]' = \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{V}_{yw}.$$

In fact, $\mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]$ and $\mathrm{var}[\boldsymbol{\eta}(\mathbf{y})]$ are the first and second components in the following decomposition of $\mathrm{var}(\mathbf{w})$:

$$\mathrm{var}(\mathbf{w}) = \mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}] + \mathrm{var}[\boldsymbol{\eta}(\mathbf{y})].$$

The best linear predictor $\boldsymbol{\eta}(\mathbf{y})$ can be regarded as an approximation to $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$. The expected value $\mathrm{E}[\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})]$ of the error of this approximation equals $\mathbf{0}$. Note that $\mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})] = \mathrm{E}\{[\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})][\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})]'\}$. Further, $\boldsymbol{\eta}(\mathbf{y})$ is the best linear approximation to $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ in the sense that, for any $(M \times 1)$-dimensional vector-valued function $\mathbf{t}(\mathbf{y})$ of the form $\mathbf{t}(\mathbf{y}) = \mathbf{c} + \mathbf{A}'\mathbf{y}$, the difference between the matrix $\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})][\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})]'\}$ and the matrix $\mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})]$ equals the matrix $\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})][\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})]'\}$, which is nonnegative definite

and which equals $\mathbf{0}$ if and only if $\mathbf{t}(\mathbf{y}) = \boldsymbol{\eta}(\mathbf{y})$. This result follows from what has already been established (in regard to the best linear prediction of $\mathbf{w}$) upon observing [in light of result (10.5)] that

$$\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \mathbf{w}][\mathbf{t}(\mathbf{y}) - \mathbf{w}]'\} = \mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})][\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})]'\} + \mathrm{E}[\mathrm{var}(\mathbf{w} \mid \mathbf{y})],$$

which in combination with result (10.15) implies that the difference between the two matrices $\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})][\mathbf{t}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})]'\}$ and $\mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})]$ is the same as that between the two matrices $\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \mathbf{w}][\mathbf{t}(\mathbf{y}) - \mathbf{w}]'\}$ and $\mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]$. In the special case where $M = 1$, we have [for any function $t(\mathbf{y})$ of $\mathbf{y}$ of the form $t(\mathbf{y}) = c + \mathbf{a}'\mathbf{y}$ (where c is a constant and $\mathbf{a}$ an $N \times 1$ vector of constants)] that

$$\mathrm{E}\{[t(\mathbf{y}) - \mathrm{E}(w \mid \mathbf{y})]^2\} \geq \mathrm{var}[\eta(\mathbf{y}) - \mathrm{E}(w \mid \mathbf{y})] \quad \big(= \mathrm{E}\{[\eta(\mathbf{y}) - \mathrm{E}(w \mid \mathbf{y})]^2\}\big), \tag{10.17}$$

with equality holding in inequality (10.17) if and only if $t(\mathbf{y}) = \eta(\mathbf{y})$.

Hartigan (1969) refers to $\boldsymbol{\eta}(\mathbf{y})$ as the *linear expectation of* $\mathbf{w}$ *given* $\mathbf{y}$. And in the special case where $M = 1$, he refers to $\mathrm{var}[\mathbf{w} - \boldsymbol{\eta}(\mathbf{y})]$ $(= \mathbf{V}_w - \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{V}_{yw})$ as the *linear variance of* $\mathbf{w}$ *given* $\mathbf{y}$—in the general case (where M can exceed 1), this quantity could be referred to as the *linear variance-covariance matrix of* $\mathbf{w}$ *given* $\mathbf{y}$. It is only in special cases, such as that where the joint distribution of $\mathbf{y}$ and $\mathbf{w}$ is MVN, that the linear expectation and linear variance-covariance matrix of $\mathbf{w}$ given $\mathbf{y}$ coincide with the conditional expectation $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ and conditional variance-covariance matrix $\mathrm{var}(\mathbf{w} \mid \mathbf{y})$ of $\mathbf{w}$ given $\mathbf{y}$.

Note that for the vector-valued function $\boldsymbol{\eta}(\cdot)$ to be determinable from what is known about the joint distribution of $\mathbf{y}$ and $\mathbf{w}$, the supposition that $\boldsymbol{\mu}_y$, $\boldsymbol{\mu}_w$, $\mathbf{V}_y$, $\mathbf{V}_{yw}$, and $\mathbf{V}_w$ are known is stronger than necessary. It suffices to know the vector $\boldsymbol{\tau} = \boldsymbol{\mu}_w - \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\boldsymbol{\mu}_y$ and the matrix $\mathbf{V}_{yw}'\mathbf{V}_y^{-1}$.

State (3): only the variances and covariances are known. Suppose that $\mathbf{V}_y$, $\mathbf{V}_{yw}$, and $\mathbf{V}_w$ are known (and that $\mathbf{V}_y$ is nonsingular), but that nothing else is known about the joint distribution of $\mathbf{y}$ and $\mathbf{w}$. Then, $\boldsymbol{\eta}(\cdot)$ is not determinable from what is known, and consequently $\boldsymbol{\eta}(\mathbf{y})$ does not qualify as a predictor. Thus, we are forced to look elsewhere for a predictor of $\mathbf{w}$.

Corresponding to any estimator $\tilde{\boldsymbol{\tau}}(\mathbf{y})$ of the vector $\boldsymbol{\tau}$ $(= \boldsymbol{\mu}_w - \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\boldsymbol{\mu}_y)$ is the predictor $\tilde{\boldsymbol{\eta}}(\mathbf{y})$ of $\mathbf{w}$ obtained from $\boldsymbol{\eta}(\mathbf{y})$ by substituting $\tilde{\boldsymbol{\tau}}(\mathbf{y})$ for $\boldsymbol{\tau}$. That is, corresponding to $\tilde{\boldsymbol{\tau}}(\mathbf{y})$ is the predictor $\tilde{\boldsymbol{\eta}}(\mathbf{y})$ defined as follows:

$$\tilde{\boldsymbol{\eta}}(\mathbf{y}) = \tilde{\boldsymbol{\tau}}(\mathbf{y}) + \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{y}. \tag{10.18}$$

Equality (10.18) serves to establish a one-to-one correspondence between estimators of $\boldsymbol{\tau}$ and predictors of $\mathbf{w}$—corresponding to any predictor $\tilde{\boldsymbol{\eta}}(\mathbf{y})$ of $\mathbf{w}$ is a unique estimator $\tilde{\boldsymbol{\tau}}(\mathbf{y})$ of $\boldsymbol{\tau}$ that satisfies equality (10.18), namely, the estimator $\tilde{\boldsymbol{\tau}}(\mathbf{y})$ defined by $\tilde{\boldsymbol{\tau}}(\mathbf{y}) = \tilde{\boldsymbol{\eta}}(\mathbf{y}) - \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{y}$.

Clearly, the predictor $\tilde{\boldsymbol{\eta}}(\mathbf{y})$ is linear if and only if the corresponding estimator $\tilde{\boldsymbol{\tau}}(\mathbf{y})$ is linear. Moreover,

$$\mathrm{E}[\tilde{\boldsymbol{\eta}}(\mathbf{y})] = \mathrm{E}[\tilde{\boldsymbol{\tau}}(\mathbf{y})] + \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\boldsymbol{\mu}_y \tag{10.19}$$

and, consequently,

$$\mathrm{E}[\tilde{\boldsymbol{\eta}}(\mathbf{y})] = \boldsymbol{\mu}_w \quad \Leftrightarrow \quad \mathrm{E}[\tilde{\boldsymbol{\tau}}(\mathbf{y})] = \boldsymbol{\tau}. \tag{10.20}$$

Thus, $\tilde{\boldsymbol{\eta}}(\mathbf{y})$ is an unbiased predictor of $\mathbf{w}$ if and only if the corresponding estimator $\tilde{\boldsymbol{\tau}}(\mathbf{y})$ is an unbiased estimator of $\boldsymbol{\tau}$.

The following identity serves to decompose the prediction error of the predictor $\tilde{\boldsymbol{\eta}}(\mathbf{y})$ into two components:

$$\tilde{\boldsymbol{\eta}}(\mathbf{y}) - \mathbf{w} = [\tilde{\boldsymbol{\eta}}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})] + [\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]. \tag{10.21}$$

Clearly,

$$\tilde{\boldsymbol{\eta}}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y}) = \tilde{\boldsymbol{\tau}}(\mathbf{y}) - \boldsymbol{\tau}. \tag{10.22}$$

Thus, decomposition (10.21) can be reexpressed as follows:

$$\tilde{\boldsymbol{\eta}}(\mathbf{y}) - \mathbf{w} = [\tilde{\boldsymbol{\tau}}(\mathbf{y}) - \boldsymbol{\tau}] + [\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]. \tag{10.23}$$

Let us now specialize to linear predictors. Let us write $\tilde{\boldsymbol{\eta}}_L(\mathbf{y})$ for a linear predictor of $\mathbf{w}$ and $\tilde{\boldsymbol{\tau}}_L(\mathbf{y})$ for the corresponding estimator of $\boldsymbol{\tau}$ [which, like $\tilde{\boldsymbol{\eta}}_L(\mathbf{y})$, is linear]. Then, in light of results (10.22) and (10.8),

$$\mathrm{E}\{[\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}][\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]'\} = \mathrm{E}\{[\tilde{\boldsymbol{\eta}}_L(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})][\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]'\} = \mathbf{0}. \tag{10.24}$$

And making use of results (10.23) and (10.16), it follows that

$$\begin{aligned}
&\mathrm{E}\{[\tilde{\boldsymbol{\eta}}_L(\mathbf{y}) - \mathbf{w}][\tilde{\boldsymbol{\eta}}_L(\mathbf{y}) - \mathbf{w}]'\} \\
&\qquad = \mathrm{E}\big(\{[\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}] + [\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]\}\{[\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}]' + [\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]'\}\big) \\
&\qquad = \mathrm{E}\{[\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}][\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}]'\} + \mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}] \\
&\qquad = \mathrm{E}\{[\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}][\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}]'\} + \mathbf{V}_w - \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{V}_{yw}.
\end{aligned} \tag{10.25}$$

Let $\mathfrak{L}_p$ represent a collection of linear predictors of $\mathbf{w}$. And let $\mathfrak{L}_e$ represent the collection of (linear) estimators of $\boldsymbol{\tau}$ that correspond to the predictors in $\mathfrak{L}_p$. Then, for a predictor, say $\hat{\boldsymbol{\eta}}_L(\mathbf{y})$, in the collection $\mathfrak{L}_p$ to be best in the sense that, for every predictor $\tilde{\boldsymbol{\eta}}_L(\mathbf{y})$ in $\mathfrak{L}_p$, the matrix $\mathrm{E}\{[\tilde{\boldsymbol{\eta}}_L(\mathbf{y}) - \mathbf{w}][\tilde{\boldsymbol{\eta}}_L(\mathbf{y}) - \mathbf{w}]'\} - \mathrm{E}\{[\hat{\boldsymbol{\eta}}_L(\mathbf{y}) - \mathbf{w}][\hat{\boldsymbol{\eta}}_L(\mathbf{y}) - \mathbf{w}]'\}$ is nonnegative definite, it is necessary and sufficient that

$$\hat{\boldsymbol{\eta}}_L(\mathbf{y}) = \hat{\boldsymbol{\tau}}_L(\mathbf{y}) + \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{y}$$

for some estimator $\hat{\boldsymbol{\tau}}_L(\mathbf{y})$ in $\mathfrak{L}_e$ that is best in the sense that, for every estimator $\tilde{\boldsymbol{\tau}}_L(\mathbf{y})$ in $\mathfrak{L}_e$, the matrix $\mathrm{E}\{[\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}][\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}]'\} - \mathrm{E}\{[\hat{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}][\hat{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}]'\}$ is nonnegative definite. In general, there may or may not be an estimator that is best in such a sense; the existence of such an estimator depends on the nature of the collection $\mathfrak{L}_e$ and on any assumptions that may be made about $\boldsymbol{\mu}_y$ and $\boldsymbol{\mu}_w$.

If $\mathfrak{L}_p$ is the collection of all linear unbiased predictors of $\mathbf{w}$, then $\mathfrak{L}_e$ is the collection of all linear unbiased estimators of $\boldsymbol{\tau}$. As previously indicated (in Section 5.5a), it is customary to refer to an estimator that is best among linear unbiased estimators as a BLUE (an acronym for best linear unbiased estimator or estimation). Similarly, a predictor that is best among linear unbiased predictors is customarily referred to as a BLUP (an acronym for best linear unbiased predictor or prediction).

The prediction error of the predictor $\tilde{\boldsymbol{\eta}}(\mathbf{y})$ can be decomposed into three components by starting with decomposition (10.23) and by expanding the component $\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}$ into two components on the basis of decomposition (10.11). As specialized to the linear predictor $\tilde{\boldsymbol{\eta}}_L(\mathbf{y})$, the resultant decomposition is

$$\tilde{\boldsymbol{\eta}}_L(\mathbf{y}) - \mathbf{w} = [\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}] + [\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})] + [\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}]. \tag{10.26}$$

Recall (from the preceding part of the present subsection) that $\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})$ and $\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}$ [which are the 2nd and 3rd components of decomposition (10.26)] are uncorrelated and that each has an expected value of $\mathbf{0}$. Moreover, it follows from result (10.3) that $\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}$ is uncorrelated with $\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}$ and from result (10.7) that it is uncorrelated with $\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}$ and hence uncorrelated with $\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})$ [which is expressible as the difference between $\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}$ and $\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}$]. Thus, all three components of decomposition (10.26) are uncorrelated. Expanding on the terminology introduced in the preceding part of the present subsection, the first, second, and third components of decomposition (10.26) can be regarded, respectively, as an "unknown-means" component, a "nonlinearity" component, and an "inherent" component.

Corresponding to decomposition (10.26) of the prediction error of $\tilde{\boldsymbol{\eta}}_L(\mathbf{y})$ is the following decomposition of the MSE matrix of $\tilde{\boldsymbol{\eta}}_L(\mathbf{y})$:

$$\begin{aligned}
&\mathrm{E}\{[\tilde{\boldsymbol{\eta}}_L(\mathbf{y}) - \mathbf{w}][\tilde{\boldsymbol{\eta}}_L(\mathbf{y}) - \mathbf{w}]'\} \\
&\qquad = \mathrm{E}\{[\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}][\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}]'\} + \mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathrm{E}(\mathbf{w} \mid \mathbf{y})] + \mathrm{var}[\mathrm{E}(\mathbf{w} \mid \mathbf{y}) - \mathbf{w}].
\end{aligned} \tag{10.27}$$

In the special case where $M = 1$, this decomposition can [upon writing $\tilde{\eta}_L(\mathbf{y})$ for $\tilde{\boldsymbol{\eta}}_L(\mathbf{y})$, $\tilde{\tau}_L(\mathbf{y})$ for $\tilde{\boldsymbol{\tau}}_L(\mathbf{y})$, and τ for $\boldsymbol{\tau}$, as well as $\eta(\mathbf{y})$ for $\boldsymbol{\eta}(\mathbf{y})$ and w for $\mathbf{w}$] be reexpressed as follows:

$$\mathrm{E}\{[\tilde{\eta}_L(\mathbf{y}) - w]^2\} = \mathrm{E}\{[\tilde{\tau}_L(\mathbf{y}) - \tau]^2\} + \mathrm{var}[\eta(\mathbf{y}) - \mathrm{E}(w \mid \mathbf{y})] + \mathrm{var}[\mathrm{E}(w \mid \mathbf{y}) - w].$$

In taking $\tilde{\boldsymbol{\tau}}(\mathbf{y})$ to be an estimator of $\boldsymbol{\tau}$ and regarding $\tilde{\boldsymbol{\eta}}(\mathbf{y})$ as a predictor of $\mathbf{w}$, it is implicitly assumed that the functions $\tilde{\boldsymbol{\tau}}(\cdot)$ and $\tilde{\boldsymbol{\eta}}(\cdot)$ depend on the joint distribution of $\mathbf{y}$ and $\mathbf{w}$ only through $\mathbf{V}_y$, $\mathbf{V}_{yw}$, and $\mathbf{V}_w$. In practice, the dependence may only be through the elements of the matrix $\mathbf{V}_{yw}'\mathbf{V}_y^{-1}$ and through various functions of the elements of $\mathbf{V}_y^{-1}$, in which case $\tilde{\boldsymbol{\tau}}(\mathbf{y})$ may qualify as an estimator and $\tilde{\boldsymbol{\eta}}(\mathbf{y})$ as a predictor even in the absence of complete knowledge of $\mathbf{V}_y$, $\mathbf{V}_{yw}$, and $\mathbf{V}_w$.

b. Prediction on the basis of a G–M, Aitken, or general linear model

Suppose that the value of an $N \times 1$ observable random vector $\mathbf{y}$ is to be used to predict the realization of an $M \times 1$ unobservable random vector $\mathbf{w} = (w_1, w_2, \ldots, w_M)'$. How might we proceed? As is evident from the results of Subsection a, the answer depends on what is "known" about the joint distribution of $\mathbf{y}$ and $\mathbf{w}$.

We could refer to whatever assumptions are made about the joint distribution of $\mathbf{y}$ and $\mathbf{w}$ as a (statistical) model. However, while doing so might be logical, it would be unconventional and hence potentially confusing. It is customary to restrict the use of the word model to the assumptions made about the distribution of the observable random vector $\mathbf{y}$.

Irrespective of the terminology, the assumptions made about the distribution of $\mathbf{y}$ do not in and of themselves provide an adequate basis for prediction. The prediction of $\mathbf{w}$ requires the larger set of assumptions that apply to the joint distribution of $\mathbf{y}$ and $\mathbf{w}$. It is this larger set of assumptions that establishes a statistical relationship between $\mathbf{y}$ and $\mathbf{w}$.

Now, assume that $\mathbf{y}$ follows a general linear model. And for purposes of predicting the realization of $\mathbf{w}$ from the value of $\mathbf{y}$, let us augment that assumption with an assumption that

$$\mathrm{E}(\mathbf{w}) = \boldsymbol{\Lambda}'\boldsymbol{\beta} \tag{10.28}$$

for some $(P \times M)$ matrix $\boldsymbol{\Lambda}$ of (known) constants and an assumption that

$$\mathrm{cov}(\mathbf{y}, \mathbf{w}) = \mathbf{V}_{yw}(\boldsymbol{\theta}) \quad \text{and} \quad \mathrm{var}(\mathbf{w}) = \mathbf{V}_w(\boldsymbol{\theta}) \tag{10.29}$$

for some matrices $\mathbf{V}_{yw}(\boldsymbol{\theta})$ and $\mathbf{V}_w(\boldsymbol{\theta})$ whose elements are known functions of the parametric vector $\boldsymbol{\theta}$—it is assumed that $\mathrm{cov}(\mathbf{y}, \mathbf{w})$ and $\mathrm{var}(\mathbf{w})$, like $\mathrm{var}(\mathbf{y})$, do not depend on $\boldsymbol{\beta}$. Further, let us (in the present context) write $\mathbf{V}_y(\boldsymbol{\theta})$ for $\mathbf{V}(\boldsymbol{\theta})$. Note that the $(N+M) \times (N+M)$ matrix $\begin{bmatrix} \mathbf{V}_y(\boldsymbol{\theta}) & \mathbf{V}_{yw}(\boldsymbol{\theta}) \\ [\mathbf{V}_{yw}(\boldsymbol{\theta})]' & \mathbf{V}_w(\boldsymbol{\theta}) \end{bmatrix}$—which is the variance-covariance matrix of the $(N+M)$-dimensional vector $\begin{pmatrix} \mathbf{y} \\ \mathbf{w} \end{pmatrix}$—is inherently nonnegative definite.

Note that the assumption that $\mathbf{w}$ satisfies condition (10.28) is consistent with taking $\mathbf{w}$ to be of the form

$$\mathbf{w} = \boldsymbol{\Lambda}'\boldsymbol{\beta} + \mathbf{d}, \tag{10.30}$$

where $\mathbf{d}$ is an M-dimensional random column vector with $\mathrm{E}(\mathbf{d}) = \mathbf{0}$. The vector $\mathbf{d}$ can be regarded as the counterpart of the vector $\mathbf{e}$ of residual effects in the model equation (1.14). Upon taking $\mathbf{w}$ to be of the form (10.30), assumption (10.29) is reexpressible as

$$\mathrm{cov}(\mathbf{e}, \mathbf{d}) = \mathbf{V}_{yw}(\boldsymbol{\theta}) \quad \text{and} \quad \mathrm{var}(\mathbf{d}) = \mathbf{V}_w(\boldsymbol{\theta}). \tag{10.31}$$

In the present context, a predictor, say $\tilde{\mathbf{w}}(\mathbf{y})$, of $\mathbf{w}$ is unbiased if and only if $\mathrm{E}[\tilde{\mathbf{w}}(\mathbf{y})] = \boldsymbol{\Lambda}'\boldsymbol{\beta}$. If there exists a linear unbiased predictor of $\mathbf{w}$, let us refer to $\mathbf{w}$ as *predictable*; otherwise, let us refer to $\mathbf{w}$ as *unpredictable*. Clearly, $\tilde{\mathbf{w}}(\mathbf{y})$ is an unbiased predictor of $\mathbf{w}$ if and only if it is an unbiased estimator of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$. And $\mathbf{w}$ is predictable if and only if $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is estimable, that is, if and only if all M of the elements of the vector $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ are estimable linear combinations of the P elements of the parametric vector $\boldsymbol{\beta}$.

As defined and discussed in Sections 5.2 and 5.6b, translation equivariance is a criterion that is applicable to estimators of a linear combination of the elements of $\boldsymbol{\beta}$ or, more generally, to estimators of a vector of such linear combinations. This criterion can also be applied to predictors. A predictor $\tilde{\mathbf{w}}(\mathbf{y})$ of the random vector $\mathbf{w}$ (the expected value of which is $\boldsymbol{\Lambda}'\boldsymbol{\beta}$) is said to be *translation equivariant* if $\tilde{\mathbf{w}}(\mathbf{y}+\mathbf{X}\mathbf{k}) = \tilde{\mathbf{w}}(\mathbf{y}) + \boldsymbol{\Lambda}'\mathbf{k}$ for every $P \times 1$ vector $\mathbf{k}$ (and for every value of $\mathbf{y}$). Clearly, $\tilde{\mathbf{w}}(\mathbf{y})$ is a translation-equivariant predictor of $\mathbf{w}$ if and only if it is a translation-equivariant estimator of the expected value $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ of $\mathbf{w}$.

Special case: Aitken and G–M models. Let us now specialize to the case where $\mathbf{y}$ follows an Aitken model. Under the Aitken model, $\text{var}(\mathbf{y})$ is an unknown scalar multiple of a known (nonnegative definite) matrix $\mathbf{H}$. It is convenient (and potentially useful) to consider the prediction of $\mathbf{w}$ under the assumption that $\text{var}\begin{pmatrix}\mathbf{y}\\ \mathbf{w}\end{pmatrix}$ is also an unknown scalar multiple of a known (nonnegative definite) matrix. Accordingly, it is supposed that $\text{cov}(\mathbf{y}, \mathbf{w})$ and $\text{var}(\mathbf{w})$ are of the form

$$\text{cov}(\mathbf{y}, \mathbf{w}) = \sigma^2\mathbf{H}_{yw} \quad \text{and} \quad \text{var}(\mathbf{w}) = \sigma^2\mathbf{H}_w,$$

where $\mathbf{H}_{yw}$ and $\mathbf{H}_w$ are known matrices. Thus, writing $\mathbf{H}_y$ for $\mathbf{H}$, the setup is such that

$$\text{var}\begin{pmatrix}\mathbf{y}\\ \mathbf{w}\end{pmatrix} = \sigma^2\begin{pmatrix}\mathbf{H}_y & \mathbf{H}_{yw}\\ \mathbf{H}'_{yw} & \mathbf{H}_w\end{pmatrix}.$$

As in the general case, it is supposed that

$$\text{E}(\mathbf{w}) = \boldsymbol{\Lambda}'\boldsymbol{\beta}$$

(where $\boldsymbol{\Lambda}$ is a known matrix).

The setup can be regarded as a special case of the more general setup where $\mathbf{y}$ follows a general linear model and where $\text{E}(\mathbf{w})$ is of the form (10.28) and $\text{cov}(\mathbf{y}, \mathbf{w})$ and $\text{var}(\mathbf{w})$ of the form (10.29). Specifically, it can be regarded as the special case where $\boldsymbol{\theta}$ is the one-dimensional vector whose only element is $\theta_1 = \sigma$, where $\Theta = \{\boldsymbol{\theta} \mid \theta_1 > 0\}$, and where $\mathbf{V}_y(\boldsymbol{\theta}) = \theta_1^2\mathbf{H}_y$, $\mathbf{V}_{yw}(\boldsymbol{\theta}) = \theta_1^2\mathbf{H}_{yw}$, and $\mathbf{V}_w(\boldsymbol{\theta}) = \theta_1^2\mathbf{H}_w$. Clearly, in this special case, $\text{var}\begin{pmatrix}\mathbf{y}\\ \mathbf{w}\end{pmatrix}$ is known up to the value of the unknown scalar multiple $\theta_1^2 = \sigma^2$.

In the further special case where $\mathbf{y}$ follows a G–M model [i.e., where $\text{var}(\mathbf{y}) = \sigma^2\mathbf{I}$], $\mathbf{H}_y = \mathbf{I}$. When $\mathbf{H}_y = \mathbf{I}$, the case where $\mathbf{H}_{yw} = \mathbf{0}$ and $\mathbf{H}_w = \mathbf{I}$ is often singled out for special attention. The case where $\mathbf{H}_y = \mathbf{I}$, $\mathbf{H}_{yw} = \mathbf{0}$, and $\mathbf{H}_w = \mathbf{I}$ is encountered in applications where the realization of $\mathbf{w}$ corresponds to a vector of future data points and where the augmented vector $\begin{pmatrix}\mathbf{y}\\ \mathbf{w}\end{pmatrix}$ is assumed to follow a G–M model, the model matrix of which is $\begin{pmatrix}\mathbf{X}\\ \boldsymbol{\Lambda}'\end{pmatrix}$.

Best linear unbiased prediction (under a G–M model). Suppose that the $N \times 1$ observable random vector $\mathbf{y}$ follows a G–M model, in which case $\text{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{var}(\mathbf{y}) = \sigma^2\mathbf{I}$. And consider the prediction of the $M \times 1$ unobservable random vector $\mathbf{w}$ whose expected value is of the form

$$\text{E}(\mathbf{w}) = \boldsymbol{\Lambda}'\boldsymbol{\beta} \tag{10.32}$$

(where $\boldsymbol{\Lambda}$ is a matrix of known constants). Assume that $\text{cov}(\mathbf{y}, \mathbf{w})$ and $\text{var}(\mathbf{w})$ are of the form

$$\text{cov}(\mathbf{y}, \mathbf{w}) = \sigma^2\mathbf{H}_{yw} \quad \text{and} \quad \text{var}(\mathbf{w}) = \sigma^2\mathbf{H}_w \tag{10.33}$$

(where $\mathbf{H}_{yw}$ and $\mathbf{H}_w$ are known matrices). Assume also that $\mathbf{w}$ is predictable or, equivalently, that $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is estimable.

For purposes of applying the results of the final part of the preceding subsection (Subsection a), take $\boldsymbol{\tau}$ to be the $M \times 1$ vector (of linear combinations of the elements of $\boldsymbol{\beta}$) defined as follows:

$$\boldsymbol{\tau} = \text{E}(\mathbf{w}) - [\text{cov}(\mathbf{y}, \mathbf{w})]'[\text{var}(\mathbf{y})]^{-1}\,\text{E}(\mathbf{y}) = \boldsymbol{\Lambda}'\boldsymbol{\beta} - \mathbf{H}'_{yw}\mathbf{X}\boldsymbol{\beta} = (\boldsymbol{\Lambda}' - \mathbf{H}'_{yw}\mathbf{X})\boldsymbol{\beta}. \tag{10.34}$$

Clearly, $\boldsymbol{\tau}$ is estimable, and its least squares estimator is the vector $\hat{\boldsymbol{\tau}}_L(\mathbf{y})$ defined as follows:

$$\hat{\boldsymbol{\tau}}_L(\mathbf{y}) = (\boldsymbol{\Lambda}' - \mathbf{H}'_{yw}\mathbf{X})(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y} = \boldsymbol{\Lambda}'(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y} - \mathbf{H}'_{yw}\mathbf{P}_{\mathbf{X}}\mathbf{y} \tag{10.35}$$

[where $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$]. Moreover, according to Theorem 5.6.1, $\hat{\boldsymbol{\tau}}_L(\mathbf{y})$ is a linear unbiased estimator of $\boldsymbol{\tau}$ and, in fact, is the BLUE (best linear unbiased estimator) of $\boldsymbol{\tau}$. It is the BLUE in the sense that the difference between the MSE matrix $\mathrm{E}\{[\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}][\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}]'\} = \mathrm{var}[\tilde{\boldsymbol{\tau}}_L(\mathbf{y})]$ of an arbitrary linear unbiased estimator $\tilde{\boldsymbol{\tau}}_L(\mathbf{y})$ of $\boldsymbol{\tau}$ and the MSE matrix $\mathrm{E}\{[\hat{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}][\hat{\boldsymbol{\tau}}_L(\mathbf{y}) - \boldsymbol{\tau}]'\} = \mathrm{var}[\hat{\boldsymbol{\tau}}_L(\mathbf{y})]$ of the least squares estimator $\hat{\boldsymbol{\tau}}_L(\mathbf{y})$ is nonnegative definite [and is equal to $\mathbf{0}$ if and only if $\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) = \hat{\boldsymbol{\tau}}_L(\mathbf{y})$].

Now, let

$$\hat{\mathbf{w}}_L(\mathbf{y}) = \hat{\boldsymbol{\tau}}_L(\mathbf{y}) + [\mathrm{cov}(\mathbf{y}, \mathbf{w})]'[\mathrm{var}(\mathbf{y})]^{-1}\mathbf{y} = \boldsymbol{\Lambda}'(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y} + \mathbf{H}'_{yw}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}. \tag{10.36}$$

Then, it follows from the results of the final part of Subsection a that $\hat{\mathbf{w}}_L(\mathbf{y})$ is a linear unbiased predictor of $\mathbf{w}$ and, in fact, is the BLUP (best linear unbiased predictor) of $\mathbf{w}$. It is the BLUP in the sense that the difference between the MSE matrix $\mathrm{E}\{[\tilde{\mathbf{w}}_L(\mathbf{y}) - \mathbf{w}][\tilde{\mathbf{w}}_L(\mathbf{y}) - \mathbf{w}]'\} = \mathrm{var}[\tilde{\mathbf{w}}_L(\mathbf{y}) - \mathbf{w}]$ of an arbitrary linear unbiased predictor $\tilde{\mathbf{w}}_L(\mathbf{y})$ of $\mathbf{w}$ and the MSE matrix $\mathrm{E}\{[\hat{\mathbf{w}}_L(\mathbf{y}) - \mathbf{w}][\hat{\mathbf{w}}_L(\mathbf{y}) - \mathbf{w}]'\} = \mathrm{var}[\hat{\mathbf{w}}_L(\mathbf{y}) - \mathbf{w}]$ of $\hat{\mathbf{w}}_L(\mathbf{y})$ is nonnegative definite [and is equal to $\mathbf{0}$ if and only if $\tilde{\mathbf{w}}_L(\mathbf{y}) = \hat{\mathbf{w}}_L(\mathbf{y})$]. In the special case where $M = 1$, the sense in which $\hat{\mathbf{w}}_L(\mathbf{y})$ is the BLUP can [upon writing $\hat{w}_L(\mathbf{y})$ for $\hat{\mathbf{w}}_L(\mathbf{y})$ and w for $\mathbf{w}$] be restated as follows: the MSE of $\hat{w}_L(\mathbf{y})$ [or, equivalently, the variance of the prediction error of $\hat{w}_L(\mathbf{y})$] is smaller than that of any other linear unbiased predictor of w.

In light of result (6.7), the variance-covariance matrix of the least squares estimator of $\boldsymbol{\tau}$ is

$$\mathrm{var}[\hat{\boldsymbol{\tau}}_L(\mathbf{y})] = \sigma^2(\boldsymbol{\Lambda}' - \mathbf{H}'_{yw}\mathbf{X})(\mathbf{X}'\mathbf{X})^-(\boldsymbol{\Lambda} - \mathbf{X}'\mathbf{H}_{yw}). \tag{10.37}$$

Accordingly, it follows from result (10.25) that the MSE matrix of the BLUP of $\mathbf{w}$ or, equivalently, the variance-covariance matrix of the prediction error of the BLUP is

$$\mathrm{var}[\hat{\mathbf{w}}_L(\mathbf{y}) - \mathbf{w}] = \sigma^2(\boldsymbol{\Lambda}' - \mathbf{H}'_{yw}\mathbf{X})(\mathbf{X}'\mathbf{X})^-(\boldsymbol{\Lambda} - \mathbf{X}'\mathbf{H}_{yw}) + \sigma^2(\mathbf{H}_w - \mathbf{H}'_{yw}\mathbf{H}_{yw}). \tag{10.38}$$

In the special case where $\mathbf{H}_{yw} = \mathbf{0}$, we find that $\boldsymbol{\tau} = \boldsymbol{\Lambda}'\boldsymbol{\beta}$, $\hat{\mathbf{w}}_L(\mathbf{y}) = \hat{\boldsymbol{\tau}}_L(\mathbf{y})$, $\mathrm{var}[\hat{\tau}_L(\mathbf{y})] = \sigma^2\boldsymbol{\Lambda}'(\mathbf{X}'\mathbf{X})^-\boldsymbol{\Lambda}$, and

$$\mathrm{var}[\hat{\mathbf{w}}_L(\mathbf{y}) - \mathbf{w}] = \sigma^2\boldsymbol{\Lambda}'(\mathbf{X}'\mathbf{X})^-\boldsymbol{\Lambda} + \sigma^2\mathbf{H}_w.$$

Note that even in this special case [where the BLUP of $\mathbf{w}$ equals the BLUE of $\boldsymbol{\tau}$ and where $\boldsymbol{\tau} = \mathrm{E}(\mathbf{w})$], the MSE matrix of the BLUP typically differs from that of the BLUE. The difference between the two MSE matrices [$\sigma^2\mathbf{H}_w$ in the special case and $\sigma^2(\mathbf{H}_w - \mathbf{H}'_{yw}\mathbf{H}_{yw})$ in the general case] is nonnegative definite. This difference is attributable to the variability of $\mathbf{w}$, which contributes to the variability of the prediction error $\hat{\mathbf{w}}(\mathbf{y}) - \mathbf{w}$ but not to the variability of $\hat{\boldsymbol{\tau}}(\mathbf{y}) - \boldsymbol{\tau}$.

Best linear translation-equivariant prediction (under a G–M model). Let us continue to consider the prediction of the $M \times 1$ unobservable random vector $\mathbf{w}$ on the basis of the $N \times 1$ observable random vector $\mathbf{y}$, doing so under the same conditions as in the preceding part of the present subsection. Thus, it is supposed that $\mathbf{y}$ follows a G–M model, that $\mathrm{E}(\mathbf{w})$ is of the form (10.32), that $\mathrm{cov}(\mathbf{y}, \mathbf{w})$ and $\mathrm{var}(\mathbf{w})$ are of the form (10.33), and that $\mathbf{w}$ is predictable. Further, define $\boldsymbol{\tau}$, $\hat{\boldsymbol{\tau}}_L(\mathbf{y})$, and $\hat{\mathbf{w}}_L(\mathbf{y})$ as in equations (10.34), (10.35), and (10.36) [so that $\hat{\boldsymbol{\tau}}_L(\mathbf{y})$ is the BLUE of $\boldsymbol{\tau}$ and $\hat{\mathbf{w}}_L(\mathbf{y})$ the BLUP of $\mathbf{w}$].

Let us consider the translation-equivariant prediction of $\mathbf{w}$. Denote by $\tilde{\mathbf{w}}(\mathbf{y})$ an arbitrary predictor of $\mathbf{w}$, and take

$$\tilde{\boldsymbol{\tau}}(\mathbf{y}) = \tilde{\mathbf{w}}(\mathbf{y}) - [\mathrm{cov}(\mathbf{y}, \mathbf{w})]'[\mathrm{var}(\mathbf{y})]^{-1}\mathbf{y} = \tilde{\mathbf{w}}(\mathbf{y}) - \mathbf{H}'_{yw}\mathbf{y}$$

to be the corresponding estimator of $\boldsymbol{\tau}$—refer to the final part of Subsection a. Then, $\tilde{\mathbf{w}}(\mathbf{y})$ is a translation-equivariant predictor (of $\mathbf{w}$) if and only if $\tilde{\boldsymbol{\tau}}(\mathbf{y})$ is a translation-equivariant estimator (of $\boldsymbol{\tau}$), as can be readily verified. Further, $\tilde{\mathbf{w}}(\mathbf{y})$ is a linear translation-equivariant predictor (of $\mathbf{w}$) if and only if $\tilde{\boldsymbol{\tau}}(\mathbf{y})$ is a linear translation-equivariant estimator (of $\boldsymbol{\tau}$).

In light of Corollary 5.6.4, the estimator $\hat{\boldsymbol{\tau}}_L(\mathbf{y})$ is a linear translation-equivariant estimator of $\boldsymbol{\tau}$ and, in fact, is the best linear translation-equivariant estimator of $\boldsymbol{\tau}$. It is the best linear translation-equivariant estimator in the sense that the difference between the MSE matrix $\mathrm{E}\{[\tilde{\boldsymbol{\tau}}_L(\mathbf{y})-\boldsymbol{\tau}][\tilde{\boldsymbol{\tau}}_L(\mathbf{y})-\boldsymbol{\tau}]'\}$ of an arbitrary linear translation-equivariant estimator $\tilde{\boldsymbol{\tau}}_L(\mathbf{y})$ of $\boldsymbol{\tau}$ and the MSE matrix $\mathrm{E}\{[\hat{\boldsymbol{\tau}}_L(\mathbf{y})-\boldsymbol{\tau}][\hat{\boldsymbol{\tau}}_L(\mathbf{y})-\boldsymbol{\tau}]'\}$ of $\hat{\boldsymbol{\tau}}_L(\mathbf{y})$ is nonnegative definite [and is equal to $\mathbf{0}$ if and only if $\tilde{\boldsymbol{\tau}}_L(\mathbf{y}) = \hat{\boldsymbol{\tau}}_L(\mathbf{y})$]. And upon recalling the results of the final part of Subsection a, it follows that the predictor $\hat{\mathbf{w}}_L(\mathbf{y})$ is a linear translation-equivariant predictor of $\mathbf{w}$ and, in fact, is the best linear translation-equivariant predictor of $\mathbf{w}$. It is the best linear translation-equivariant predictor in the sense that the difference between the MSE matrix $\mathrm{E}\{[\tilde{\mathbf{w}}_L(\mathbf{y})-\mathbf{w}][\tilde{\mathbf{w}}_L(\mathbf{y})-\mathbf{w}]'\}$ of an arbitrary linear translation-equivariant predictor $\tilde{\mathbf{w}}_L(\mathbf{y})$ of $\mathbf{w}$ and the MSE matrix $\mathrm{E}\{[\hat{\mathbf{w}}_L(\mathbf{y})-\mathbf{w}][\hat{\mathbf{w}}_L(\mathbf{y})-\mathbf{w}]'\} = \mathrm{var}[\hat{\mathbf{w}}_L(\mathbf{y})-\mathbf{w}]$ of $\hat{\mathbf{w}}_L(\mathbf{y})$ is nonnegative definite [and is equal to $\mathbf{0}$ if and only if $\tilde{\mathbf{w}}_L(\mathbf{y}) = \hat{\mathbf{w}}_L(\mathbf{y})$]. In the special case where $M = 1$, the sense in which $\hat{\mathbf{w}}_L(\mathbf{y})$ is the best linear translation-equivariant predictor can [upon writing $\hat{w}_L(\mathbf{y})$ for $\hat{\mathbf{w}}_L(\mathbf{y})$ and w for $\mathbf{w}$] be restated as follows: the MSE of $\hat{w}_L(\mathbf{y})$ is smaller than that of any other linear translation-equivariant predictor of w.

c. Conditional expected values: elliptical distributions

Let $\mathbf{w}$ represent an $M \times 1$ random vector and $\mathbf{y}$ an $N \times 1$ random vector. Suppose that the second-order moments of the joint distribution of $\mathbf{w}$ and $\mathbf{y}$ exist. And adopt the following notation: $\boldsymbol{\mu}_y = \mathrm{E}(\mathbf{y})$, $\boldsymbol{\mu}_w = \mathrm{E}(\mathbf{w})$, $\mathbf{V}_y = \mathrm{var}(\mathbf{y})$, $\mathbf{V}_{yw} = \mathrm{cov}(\mathbf{y}, \mathbf{w})$, and $\mathbf{V}_w = \mathrm{var}(\mathbf{w})$. Further, suppose that $\mathbf{V}_y$ is nonsingular.

Let $\boldsymbol{\eta}(\mathbf{y}) = \boldsymbol{\mu}_w + \mathbf{V}_{yw}'\mathbf{V}_y^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)$. If $\mathbf{y}$ is observable but $\mathbf{w}$ is unobservable, we might wish to use the value of $\mathbf{y}$ to predict the realization of $\mathbf{w}$. If $\boldsymbol{\eta}(\cdot)$ is determinable from what is known about the joint distribution of $\mathbf{w}$ and $\mathbf{y}$, we could use $\boldsymbol{\eta}(\mathbf{y})$ to make the prediction; it would be the best linear predictor in the sense described in Part 2 of Subsection a. If enough more is known about the joint distribution of $\mathbf{w}$ and $\mathbf{y}$ that $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ is determinable, we might prefer to use $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ to make the prediction; it would be the best predictor in the sense described in Part 1 of Subsection a.

Under what circumstances is $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ equal to $\boldsymbol{\eta}(\mathbf{y})$ (at least with probability 1) or, equivalently, under what circumstances is $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ linear (or at least "linear with probability 1"). As previously indicated (in Part 1 of Subsection a), one such circumstance is that where the joint distribution of $\mathbf{w}$ and $\mathbf{y}$ is MVN. More generally, $\mathrm{E}(\mathbf{w} \mid \mathbf{y})$ equals $\boldsymbol{\eta}(\mathbf{y})$ (at least with probability 1) if the joint distribution of $\mathbf{w}$ and $\mathbf{y}$ is elliptical, as will now be shown.

Let $\mathbf{e} = \mathbf{w} - \boldsymbol{\eta}(\mathbf{y})$, and observe that

$$\begin{pmatrix}\mathbf{e}\\ \mathbf{y}\end{pmatrix} = \begin{pmatrix}-\boldsymbol{\mu}_w + \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\boldsymbol{\mu}_y\\ \mathbf{0}\end{pmatrix} + \begin{pmatrix}\mathbf{I} & -\mathbf{V}_{yw}'\mathbf{V}_y^{-1}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix}.$$

Observe also [in light of Part (2) of Theorem 3.5.9] that

$$\mathrm{E}(\mathbf{e}) = \mathbf{0}, \qquad \mathrm{var}(\mathbf{e}) = \mathbf{V}_w - \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{V}_{yw}, \quad \text{and} \quad \mathrm{cov}(\mathbf{e}, \mathbf{y}) = \mathbf{0}.$$

Now, suppose that the distribution of the vector $\begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix}$ is elliptical with mgf generator $\phi(\cdot)$. Then, it follows from the next-to-last part of Section 5.9c that the vector $\begin{pmatrix}\mathbf{e}\\ \mathbf{y}\end{pmatrix}$ has an elliptical distribution with mean $\begin{pmatrix}\mathbf{0}\\ \boldsymbol{\mu}_y\end{pmatrix}$, variance-covariance matrix $\begin{pmatrix}\mathbf{V}_w - \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{V}_{yw} & \mathbf{0}\\ \mathbf{0} & \mathbf{V}_y\end{pmatrix}$, and mgf generator $\phi(\cdot)$ and that the vector $\begin{pmatrix}-\mathbf{e}\\ \mathbf{y}\end{pmatrix}$ has this same distribution. Thus, the conditional distribution of $-\mathbf{e}$ given $\mathbf{y}$ is the same as that of $\mathbf{e}$ given $\mathbf{y}$, so that the conditional distribution of $\mathbf{e}$ is symmetrical about $\mathbf{0}$ and hence $\mathrm{E}(\mathbf{e} \mid \mathbf{y}) = \mathbf{0}$ (with probability 1). And since $\mathbf{w} = \boldsymbol{\eta}(\mathbf{y}) + \mathbf{e}$, we conclude that $\mathrm{E}(\mathbf{w} \mid \mathbf{y}) = \boldsymbol{\eta}(\mathbf{y})$ (with probability 1).

Exercises

Exercise 1. Take the context to be that of estimating parametric functions of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ from an $N \times 1$ observable random vector $\mathbf{y}$ that follows a G–M, Aitken, or general linear model. Verify (1) that linear combinations of estimable functions are estimable and (2) that linear combinations of nonestimable functions are not necessarily nonestimable.

Exercise 2. Take the context to be that of estimating parametric functions of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ from an $N \times 1$ observable random vector $\mathbf{y}$ that follows a G–M, Aitken, or general linear model. And let $R = \text{rank}(\mathbf{X})$.

(a) Verify (1) that there exists a set of R linearly independent estimable functions; (2) that no set of estimable functions contains more than R linearly independent estimable functions; and (3) that if the model is not of full rank (i.e., if $R < P$), then at least one and, in fact, at least $P - R$ of the individual parameters $\beta_1, \beta_2, \ldots, \beta_P$ are nonestimable.

(b) Show that the jth of the individual parameters $\beta_1, \beta_2, \ldots, \beta_P$ is estimable if and only if the jth element of every vector in $\mathfrak{N}(\mathbf{X})$ equals 0 $(j = 1, 2, \ldots, P)$.

Exercise 3. Show that for a parametric function of the form $\boldsymbol{\lambda}'\boldsymbol{\beta}$ to be estimable from an $N \times 1$ observable random vector $\mathbf{y}$ that follows a G–M, Aitken, or general linear model, it is necessary and sufficient that

$$\text{rank}(\mathbf{X}', \boldsymbol{\lambda}) = \text{rank}(\mathbf{X}).$$

Exercise 4. Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M, Aitken, or general linear model. Further, take $\underline{\mathbf{y}}$ to be any value of $\mathbf{y}$, and consider the quantity $\boldsymbol{\lambda}'\tilde{\mathbf{b}}$, where $\boldsymbol{\lambda}$ is an arbitrary $P \times 1$ vector of constants and $\tilde{\mathbf{b}}$ is any solution to the linear system $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\underline{\mathbf{y}}$ (in the $P \times 1$ vector $\mathbf{b}$). Show that if $\boldsymbol{\lambda}'\tilde{\mathbf{b}}$ is invariant to the choice of the solution $\tilde{\mathbf{b}}$, then $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is an estimable function. And discuss the implications of this result.

Exercise 5. Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M, Aitken, or general linear model. And let $\mathbf{a}$ represent an arbitrary $N \times 1$ vector of constants. Show that $\mathbf{a}'\mathbf{y}$ is the least squares estimator of its expected value $\text{E}(\mathbf{a}'\mathbf{y})$ (i.e., of the parametric function $\mathbf{a}'\mathbf{X}\boldsymbol{\beta}$) if and only if $\mathbf{a} \in \mathcal{C}(\mathbf{X})$.

Exercise 6. Let $\mathcal{U}$ represent a subspace of the linear space $\mathcal{R}^M$ of all M-dimensional column vectors. Verify that the set $\mathcal{U}^\perp$ (comprising all M-dimensional column vectors that are orthogonal to $\mathcal{U}$) is a linear space.

Exercise 7. Let $\mathbf{X}$ represent an $N \times P$ matrix. A $P \times N$ matrix $\mathbf{G}$ is said to be a *least squares generalized inverse* of $\mathbf{X}$ if it is a generalized inverse of $\mathbf{X}$ (i.e., if $\mathbf{X}\mathbf{G}\mathbf{X} = \mathbf{X}$) and if, in addition, $(\mathbf{X}\mathbf{G})' = \mathbf{X}\mathbf{G}$ (i.e., $\mathbf{X}\mathbf{G}$ is symmetric).

(a) Show that $\mathbf{G}$ is a least squares generalized inverse of $\mathbf{X}$ if and only if $\mathbf{X}'\mathbf{X}\mathbf{G} = \mathbf{X}'$.

(b) Using Part (a) (or otherwise), establish the existence of a least squares generalized inverse of $\mathbf{X}$.

(c) Show that if $\mathbf{G}$ is a least squares generalized inverse of $\mathbf{X}$, then, for any $N \times Q$ matrix $\mathbf{Y}$, the matrix $\mathbf{G}\mathbf{Y}$ is a solution to the linear system $\mathbf{X}'\mathbf{X}\mathbf{B} = \mathbf{X}'\mathbf{Y}$ (in the $P \times Q$ matrix $\mathbf{B}$).

Exercise 8. Let $\mathbf{A}$ represent an $M \times N$ matrix. An $N \times M$ matrix $\mathbf{H}$ is said to be a *minimum norm generalized inverse* of $\mathbf{A}$ if it is a generalized inverse of $\mathbf{A}$ (i.e., if $\mathbf{A}\mathbf{H}\mathbf{A} = \mathbf{A}$) and if, in addition, $(\mathbf{H}\mathbf{A})' = \mathbf{H}\mathbf{A}$ (i.e., $\mathbf{H}\mathbf{A}$ is symmetric).

(a) Show that $\mathbf{H}$ is a minimum norm generalized inverse of $\mathbf{A}$ if and only if $\mathbf{H}'$ is a least squares generalized inverse of $\mathbf{A}'$ (where least squares generalized inverse is as defined in Exercise 7).

(b) Using the results of Exercise 7 (or otherwise), establish the existence of a minimum norm generalized inverse of $\mathbf{A}$.

(c) Show that if $\mathbf{H}$ is a minimum norm generalized inverse of $\mathbf{A}$, then, for any vector $\mathbf{b} \in \mathcal{C}(\mathbf{A})$, $\|\mathbf{x}\|$ attains its minimum value over the set $\{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}\}$ [comprising all solutions to the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ (in $\mathbf{x}$)] uniquely at $\mathbf{x} = \mathbf{H}\mathbf{b}$ (where $\|\cdot\|$ denotes the usual norm).

Exercise 9. Let $\mathbf{X}$ represent an $N \times P$ matrix, and let $\mathbf{G}$ represent a $P \times N$ matrix that is subject to the following four conditions: (1) $\mathbf{X}\mathbf{G}\mathbf{X} = \mathbf{X}$; (2) $\mathbf{G}\mathbf{X}\mathbf{G} = \mathbf{G}$; (3) $(\mathbf{X}\mathbf{G})' = \mathbf{X}\mathbf{G}$; and (4) $(\mathbf{G}\mathbf{X})' = \mathbf{G}\mathbf{X}$.

(a) Show that if a $P \times P$ matrix $\mathbf{H}$ is a minimum norm generalized inverse of $\mathbf{X}'\mathbf{X}$, then conditions (1)–(4) can be satisfied by taking $\mathbf{G} = \mathbf{H}\mathbf{X}'$.

(b) Use Part (a) and the result of Part (b) of Exercise 8 (or other means) to establish the existence of a $P \times N$ matrix $\mathbf{G}$ that satisfies conditions (1)–(4) and show that there is only one such matrix.

(c) Let $\mathbf{X}^+$ represent the unique $P \times N$ matrix $\mathbf{G}$ that satisfies conditions (1)–(4)—this matrix is customarily referred to as the *Moore–Penrose inverse*, and conditions (1)–(4) are customarily referred to as the *Moore–Penrose conditions*. Using Parts (a) and (b) and the results of Part (c) of Exercise 7 and Part (c) of Exercise 8 (or otherwise), show that $\mathbf{X}^+\mathbf{y}$ is a solution to the linear system $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$ (in $\mathbf{b}$) and that $\|\mathbf{b}\|$ attains its minimum value over the set $\{\mathbf{b} : \mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}\}$ (comprising all solutions to the linear system) uniquely at $\mathbf{b} = \mathbf{X}^+\mathbf{y}$ (where $\|\cdot\|$ denotes the usual norm).

Exercise 10. Consider further the alternative approach to the least squares computations, taking the formulation and the notation to be those of the final part of Section 5.4e.

(a) Let $\tilde{\mathbf{b}} = \mathbf{L}_1\tilde{\mathbf{h}}_1 + \mathbf{L}_2\tilde{\mathbf{h}}_2$, where $\tilde{\mathbf{h}}_2$ is an arbitrary $(P-K)$-dimensional column vector and $\tilde{\mathbf{h}}_1$ is the solution to the linear system $\mathbf{R}_1\mathbf{h}_1 = \mathbf{z}_1 - \mathbf{R}_2\tilde{\mathbf{h}}_2$. Show that $\|\tilde{\mathbf{b}}\|$ is minimized by taking

$$\tilde{\mathbf{h}}_2 = [\mathbf{I} + (\mathbf{R}_1^{-1}\mathbf{R}_2)'\mathbf{R}_1^{-1}\mathbf{R}_2]^{-1}(\mathbf{R}_1^{-1}\mathbf{R}_2)'\mathbf{R}_1^{-1}\mathbf{z}_1.$$

Do so by formulating this minimization problem as a least squares problem in which the role of $\mathbf{y}$ is played by the vector $\begin{pmatrix}\mathbf{R}_1^{-1}\mathbf{z}_1 \\ \mathbf{0}\end{pmatrix}$, the role of $\mathbf{X}$ is played by the matrix $\begin{pmatrix}\mathbf{R}_1^{-1}\mathbf{R}_2 \\ \mathbf{I}\end{pmatrix}$, and the role of $\mathbf{b}$ is played by $\tilde{\mathbf{h}}_2$.

(b) Let $\mathbf{O}_1$ represent a $P \times K$ matrix with orthonormal columns and $\mathbf{T}_1$ a $K \times K$ upper triangular matrix such that $\begin{pmatrix}\mathbf{R}_1' \\ \mathbf{R}_2'\end{pmatrix} = \mathbf{O}_1\mathbf{T}_1'$—the existence of a decomposition of this form can be established in much the same way as the existence of the QR decomposition (in which $\mathbf{T}_1$ would be lower triangular rather than upper triangular). Further, take $\mathbf{O}_2$ to be any $P \times (P-K)$ matrix such that the $P \times P$ matrix $\mathbf{O}$ defined by $\mathbf{O} = (\mathbf{O}_1, \mathbf{O}_2)$ is orthogonal.

(1) Show that $\mathbf{X} = \mathbf{Q}\mathbf{T}(\mathbf{L}\mathbf{O})'$, where $\mathbf{T} = \begin{pmatrix}\mathbf{T}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{pmatrix}$.

(2) Show that $\mathbf{y} - \mathbf{X}\mathbf{b} = \mathbf{Q}_1(\mathbf{z}_1 - \mathbf{T}_1\mathbf{d}_1) + \mathbf{Q}_2\mathbf{z}_2$, where $\mathbf{d} = (\mathbf{L}\mathbf{O})'\mathbf{b}$ and $\mathbf{d}$ is partitioned as $\mathbf{d} = \begin{pmatrix}\mathbf{d}_1 \\ \mathbf{d}_2\end{pmatrix}$.

(3) Show that $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) = (\mathbf{z}_1 - \mathbf{T}_1\mathbf{d}_1)'(\mathbf{z}_1 - \mathbf{T}_1\mathbf{d}_1) + \mathbf{z}_2'\mathbf{z}_2$.

(4) Taking $\tilde{\mathbf{d}}_1$ to be the solution to the linear system $\mathbf{T}_1\mathbf{d}_1 = \mathbf{z}_1$ (in $\mathbf{d}_1$), show that $(\mathbf{y} -$

$\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})$ attains a minimum value of $\mathbf{z}_2'\mathbf{z}_2$ and that it does so at a value $\tilde{\mathbf{b}}$ of $\mathbf{b}$ if and only if $\tilde{\mathbf{b}} = \mathbf{L}\mathbf{O}\begin{pmatrix}\tilde{\mathbf{d}}_1\\ \tilde{\mathbf{d}}_2\end{pmatrix}$ for some $(P-K)\times 1$ vector $\tilde{\mathbf{d}}_2$.

(5) Letting $\tilde{\mathbf{b}}$ represent an arbitrary one of the values of $\mathbf{b}$ at which $(\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})$ attains a minimum value [and, as in Part (4), taking $\tilde{\mathbf{d}}_1$ to be the solution to $\mathbf{T}_1\mathbf{d}_1 = \mathbf{z}_1$], show that $\|\tilde{\mathbf{b}}\|^2$ (where $\|\cdot\|$ denotes the usual norm) attains a minimum value of $\tilde{\mathbf{d}}_1'\tilde{\mathbf{d}}_1$ and that it does so uniquely at $\tilde{\mathbf{b}} = \mathbf{L}\mathbf{O}\begin{pmatrix}\tilde{\mathbf{d}}_1\\ \mathbf{0}\end{pmatrix}$.

Exercise 11. Verify that the difference (6.14) is a nonnegative definite matrix and that it equals $\mathbf{0}$ if and only if $\mathbf{c}+\mathbf{A}'\mathbf{y} = \boldsymbol{\ell}(\mathbf{y})$.

Exercise 12. Suppose that $\mathbf{y}$ is an $N\times 1$ observable random vector that follows a G–M, Aitken, or general linear model. And let $s(\mathbf{y})$ represent any particular translation-equivariant estimator of an estimable linear combination $\boldsymbol{\lambda}'\boldsymbol{\beta}$ of the elements of the parametric vector $\boldsymbol{\beta}$—e.g., $s(\mathbf{y})$ could be the least squares estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$. Show that an estimator $t(\mathbf{y})$ of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is translation equivariant if and only if

$$t(\mathbf{y}) = s(\mathbf{y}) + d(\mathbf{y})$$

for some translation-invariant statistic $d(\mathbf{y})$.

Exercise 13. Suppose that $\mathbf{y}$ is an $N\times 1$ observable random vector that follows a G–M model. And let $\mathbf{y}'\mathbf{A}\mathbf{y}$ represent a quadratic unbiased nonnegative-definite estimator of σ^2, that is, a quadratic form in $\mathbf{y}$ whose matrix $\mathbf{A}$ is a symmetric nonnegative definite matrix of constants and whose expected value is σ^2.

(a) Show that $\mathbf{y}'\mathbf{A}\mathbf{y}$ is translation invariant.

(b) Suppose that the fourth-order moments of the distribution of the vector $\mathbf{e} = (e_1, e_2, \ldots, e_N)'$ are such that (for $i, j, k, m = 1, 2, \ldots, N$) $\mathrm{E}(e_i e_j e_k e_m)$ satisfies condition (7.38). For what choice of $\mathbf{A}$ is the variance of the quadratic unbiased nonnegative-definite estimator $\mathbf{y}'\mathbf{A}\mathbf{y}$ a minimum? Describe your reasoning.

Exercise 14. Suppose that $\mathbf{y}$ is an $N\times 1$ observable random vector that follows a G–M model. Suppose further that the distribution of the vector $\mathbf{e} = (e_1, e_2, \ldots, e_N)'$ has third-order moments $\lambda_{jkm} = \mathrm{E}(e_j e_k e_m)$ $(j, k, m = 1, 2, \ldots, N)$ and fourth-order moments $\gamma_{ijkm} = \mathrm{E}(e_i e_j e_k e_m)$ $(i, j, k, m = 1, 2, \ldots, N)$. And let $\mathbf{A} = \{a_{ij}\}$ represent an $N\times N$ symmetric matrix of constants.

(a) Show that in the special case where the elements $e_1, e_2, \ldots, e_N$ of $\mathbf{e}$ are statistically independent,

$$\mathrm{var}(\mathbf{y}'\mathbf{A}\mathbf{y}) = \mathbf{a}'\boldsymbol{\Omega}^*\mathbf{a} + 4\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\boldsymbol{\Lambda}^*\mathbf{a} + 2\sigma^4\,\mathrm{tr}(\mathbf{A}^2) + 4\sigma^2\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}^2\mathbf{X}\boldsymbol{\beta}, \tag{E.1}$$

where $\boldsymbol{\Omega}^*$ is the $N\times N$ diagonal matrix whose ith diagonal element is $\gamma_{iiii} - 3\sigma^4$, where $\boldsymbol{\Lambda}^*$ is the $N\times N$ diagonal matrix whose ith diagonal element is λ_{iii}, and where $\mathbf{a}$ is the $N\times 1$ vector whose elements are the diagonal elements $a_{11}, a_{22}, \ldots, a_{NN}$ of $\mathbf{A}$.

(b) Suppose that the elements $e_1, e_2, \ldots, e_N$ of $\mathbf{e}$ are statistically independent, that (for $i = 1, 2, \ldots, N$) $\gamma_{iiii} = \gamma$ (for some scalar γ), and that all N of the diagonal elements of the $\mathbf{P}_{\mathbf{X}}$ matrix are equal to each other. Show that the estimator $\hat{\sigma}^2 = \tilde{\mathbf{e}}'\tilde{\mathbf{e}}/(N-\mathrm{rank}\,\mathbf{X})$ [where $\tilde{\mathbf{e}} = (\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{y}$] has minimum variance among all quadratic unbiased translation-invariant estimators of σ^2.

Exercise 15. Suppose that $\mathbf{y}$ is an $N\times 1$ observable random vector that follows a G–M model, and assume that the distribution of the vector $\mathbf{e}$ of residual effects is MVN.

(a) Letting $\boldsymbol{\lambda}'\boldsymbol{\beta}$ represent an estimable linear combination of the elements of the parametric vector $\boldsymbol{\beta}$, find a minimum-variance unbiased estimator of $(\boldsymbol{\lambda}'\boldsymbol{\beta})^2$.

(b) Find a minimum-variance unbiased estimator of σ^4.

Exercise 16. Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M model, and assume that the distribution of the vector $\mathbf{e}$ of residual effects is MVN. Show that if σ^2 were known, $\mathbf{X}'\mathbf{y}$ would be a complete sufficient statistic.

Exercise 17. Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a general linear model. Suppose further that the distribution of the vector $\mathbf{e}$ of residual effects is MVN or, more generally, that the distribution of $\mathbf{e}$ is known up to the value of the vector $\boldsymbol{\theta}$. And take $\mathbf{h}(\mathbf{y})$ to be any (possibly vector-valued) translation-invariant statistic.

(a) Show that if $\boldsymbol{\theta}$ were known, $\mathbf{h}(\mathbf{y})$ would be an ancillary statistic—for a definition of ancillarity, refer, e.g., to Casella and Berger (2002, def. 6.2.16) or to Lehmann and Casella (1998, p. 41).

(b) Suppose that $\mathbf{X}'\mathbf{y}$ would be a complete sufficient statistic if $\boldsymbol{\theta}$ were known. Show (1) that the least squares estimator of any estimable linear combination $\boldsymbol{\lambda}'\boldsymbol{\beta}$ of the elements of the parametric vector $\boldsymbol{\beta}$ has minimum variance among all unbiased estimators, (2) that any vector of least squares estimators of estimable linear combinations (of the elements of $\boldsymbol{\beta}$) is distributed independently of $\mathbf{h}(\mathbf{y})$, and (3) (using the result of Exercise 12 or otherwise) that the least squares estimator of any estimable linear combination $\boldsymbol{\lambda}'\boldsymbol{\beta}$ has minimum mean squared error among all translation-equivariant estimators. {*Hint* [for Part (2)]. Make use of Basu's theorem—refer, e.g., to Lehmann and Casella (1998, p. 42) for a statement of Basu's theorem.}

Exercise 18. Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M model. Suppose further that the distribution of the vector $\mathbf{e}$ of residual effects is MVN. And, letting $\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{P}_{\mathbf{X}}\mathbf{y}$, take $\tilde{\sigma}^2 = \tilde{\mathbf{e}}'\tilde{\mathbf{e}}/N$ to be the ML estimator of σ^2 and $\hat{\sigma}^2 = \tilde{\mathbf{e}}'\tilde{\mathbf{e}}/(N - \operatorname{rank}\mathbf{X})$ to be the unbiased estimator.

(a) Find the bias and the MSE of the ML estimator $\tilde{\sigma}^2$.

(b) Compare the MSE of the ML estimator $\tilde{\sigma}^2$ with that of the unbiased estimator $\hat{\sigma}^2$: for which values of N and of $\operatorname{rank}\mathbf{X}$ is the MSE of the ML estimator smaller than that of the unbiased estimator and for which values is it larger?

Exercise 19. Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a general linear model, that the distribution of the vector $\mathbf{e}$ of residual effects is MVN, and that the variance-covariance matrix $\mathbf{V}(\boldsymbol{\theta})$ of $\mathbf{e}$ is nonsingular (for all $\boldsymbol{\theta} \in \Theta$). And, letting $K = N - \operatorname{rank}\mathbf{X}$, take $\mathbf{R}$ to be any $N \times K$ matrix (of constants) of full column rank K such that $\mathbf{X}'\mathbf{R} = \mathbf{0}$, and (as in Section 5.9b) define $\mathbf{z} = \mathbf{R}'\mathbf{y}$. Further, let $\mathbf{w} = \mathbf{s}(\mathbf{z})$, where $\mathbf{s}(\cdot)$ is a $K \times 1$ vector of real-valued functions that defines a one-to-one mapping of $\mathcal{R}^K$ onto some set $\mathcal{W}$.

(a) Show that $\mathbf{w}$ is a maximal invariant.

(b) Let $f_1(\cdot\,;\boldsymbol{\theta})$ represent the pdf of the distribution of $\mathbf{z}$, and assume that $\mathbf{s}(\cdot)$ is such that the distribution of $\mathbf{w}$ has a pdf, say $f_2(\cdot\,;\boldsymbol{\theta})$, that is obtainable from $f_1(\cdot\,;\boldsymbol{\theta})$ via an application of the basic formula (e.g., Bickel and Doksum 2001, sec. B.2) for a change of variables. And, taking $L_1(\boldsymbol{\theta};\, \mathbf{R}'\underline{\mathbf{y}})$ and $L_2[\boldsymbol{\theta};\, \mathbf{s}(\mathbf{R}'\underline{\mathbf{y}})]$ (where $\underline{\mathbf{y}}$ denotes the observed value of $\mathbf{y}$) to be the likelihood functions defined by $L_1(\boldsymbol{\theta};\, \mathbf{R}'\underline{\mathbf{y}}) = f_1(\mathbf{R}'\underline{\mathbf{y}};\, \boldsymbol{\theta})$ and $L_2[\boldsymbol{\theta};\, \mathbf{s}(\mathbf{R}'\underline{\mathbf{y}})] = f_2[\mathbf{s}(\mathbf{R}'\underline{\mathbf{y}});\, \boldsymbol{\theta}]$, show that $L_1(\boldsymbol{\theta};\, \mathbf{R}'\underline{\mathbf{y}})$ and $L_2[\boldsymbol{\theta};\, \mathbf{s}(\mathbf{R}'\underline{\mathbf{y}})]$ differ from each other by no more than a multiplicative constant.

Exercise 20. Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a general linear model, that the distribution of the vector $\mathbf{e}$ of residual effects is MVN, and that the variance-covariance matrix $\mathbf{V}(\boldsymbol{\theta})$ of $\mathbf{e}$ is nonsingular (for all $\boldsymbol{\theta} \in \Theta$). Further, let $\mathbf{z} = \mathbf{R}'\mathbf{y}$, where $\mathbf{R}$ is any $N \times (N - \operatorname{rank}\mathbf{X})$ matrix (of constants) of full column rank $N - \operatorname{rank}\mathbf{X}$ such that $\mathbf{X}'\mathbf{R} = \mathbf{0}$; and let $\mathbf{u} = \mathbf{X}_*'\mathbf{y}$, where

$\mathbf{X}_*$ is any $N \times (\text{rank}\,\mathbf{X})$ matrix (of constants) whose columns form a basis for $\mathcal{C}(\mathbf{X})$. And denote by $\underline{\mathbf{y}}$ the observed value of $\mathbf{y}$.

(a) Verify that the likelihood function that would result from regarding the observed value $(\mathbf{X}_*, \mathbf{R})'\underline{\mathbf{y}}$ of $\begin{pmatrix}\mathbf{u}\\ \mathbf{z}\end{pmatrix}$ as the data vector differs by no more than a multiplicative constant from that obtained by regarding the observed value $\underline{\mathbf{y}}$ of $\mathbf{y}$ as the data vector.

(b) Let $f_0(\cdot\,|\,\cdot\,;\boldsymbol{\beta},\boldsymbol{\theta})$ represent the pdf of the conditional distribution of $\mathbf{u}$ given $\mathbf{z}$. And take $L_0[\boldsymbol{\beta},\boldsymbol{\theta};\,(\mathbf{X}_*,\,\mathbf{R})'\underline{\mathbf{y}}]$ to be the function of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ defined by $L_0[\boldsymbol{\beta},\boldsymbol{\theta};\,(\mathbf{X}_*,\,\mathbf{R})'\underline{\mathbf{y}}] = f_0(\mathbf{X}_*'\underline{\mathbf{y}}\,|\,\mathbf{R}'\underline{\mathbf{y}};\boldsymbol{\beta},\boldsymbol{\theta})$. Show that

$$L_0[\boldsymbol{\beta},\boldsymbol{\theta};\,(\mathbf{X}_*,\,\mathbf{R})'\underline{\mathbf{y}}] = (2\pi)^{-(\text{rank}\,\mathbf{X})/2}\,|\mathbf{X}_*'\mathbf{X}_*|^{-1}\,|\mathbf{X}_*'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}_*|^{1/2} \times \exp\{-\tfrac{1}{2}[\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})-\boldsymbol{\beta}]'\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}[\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})-\boldsymbol{\beta}]\},$$

where $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ is any solution to the linear system $\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\mathbf{b} = \mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\underline{\mathbf{y}}$ (in the $P \times 1$ vector $\mathbf{b}$).

(c) In connection with Part (b), show (1) that

$$[\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})-\boldsymbol{\beta}]'\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}[\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})-\boldsymbol{\beta}] = (\underline{\mathbf{y}}-\mathbf{X}\boldsymbol{\beta})'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\{\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\}^{-}\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}(\underline{\mathbf{y}}-\mathbf{X}\boldsymbol{\beta})$$

and (2) that the distribution of the random variable s defined by

$$s = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\{\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}\mathbf{X}\}^{-}\mathbf{X}'[\mathbf{V}(\boldsymbol{\theta})]^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$

does not depend on $\boldsymbol{\beta}$.

Exercise 21. Suppose that $\mathbf{z}$ is an $S \times 1$ observable random vector and that $\mathbf{z} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, where σ is a (strictly) positive unknown parameter.

(a) Show that $\mathbf{z}'\mathbf{z}$ is a complete sufficient statistic.

(b) Take $\mathbf{w}(\mathbf{z})$ to be the S-dimensional vector-valued statistic defined by $\mathbf{w}(\mathbf{z}) = (\mathbf{z}'\mathbf{z})^{-1/2}\mathbf{z}$—$\mathbf{w}(\mathbf{z})$ is defined for $\mathbf{z} \neq \mathbf{0}$ and hence with probability 1. Show that $\mathbf{z}'\mathbf{z}$ and $\mathbf{w}(\mathbf{z})$ are statistically independent. (*Hint*. Make use of Basu's theorem.)

(c) Show that any estimator of σ^2 of the form $\mathbf{z}'\mathbf{z}/k$ (where k is a nonzero constant) is scale equivariant—an estimator, say $t(\mathbf{z})$, of σ^2 is to be regarded as *scale equivariant* if for every (strictly) positive scalar c (and for every nonnull value of $\mathbf{z}$) $t(c\mathbf{z}) = c^2 t(\mathbf{z})$.

(d) Let $t_0(\mathbf{z})$ represent any particular scale-equivariant estimator of σ^2 such that $t_0(\mathbf{z}) \neq 0$ for $\mathbf{z} \neq \mathbf{0}$. Show that an estimator $t(\mathbf{z})$ of σ^2 is scale equivariant if and only if, for some function $u(\cdot)$ such that $u(c\mathbf{z}) = u(\mathbf{z})$ (for every strictly positive constant c and every nonnull value of $\mathbf{z}$),

$$t(\mathbf{z}) = u(\mathbf{z})t_0(\mathbf{z}) \quad \text{for} \quad \mathbf{z} \neq \mathbf{0}. \tag{E.2}$$

(e) Show that a function $u(\mathbf{z})$ of $\mathbf{z}$ is such that $u(c\mathbf{z}) = u(\mathbf{z})$ (for every strictly positive constant c and every nonnull value of $\mathbf{z}$) if and only if $u(\mathbf{z})$ depends on the value of $\mathbf{z}$ only through $\mathbf{w}(\mathbf{z})$ [where $\mathbf{w}(\mathbf{z})$ is as defined in Part (b)].

(f) Show that the estimator $\mathbf{z}'\mathbf{z}/(S+2)$ has minimum MSE among all scale-equivariant estimators of σ^2.

Exercise 22. Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M model and that the distribution of the vector $\mathbf{e}$ of residual effects is MVN. Using the result of Part (f) of Exercise 21 (or otherwise), show that the Hodges–Lehmann estimator $\mathbf{y}'(\mathbf{I}-\mathbf{P}_\mathbf{X})\mathbf{y}/[N-\text{rank}(\mathbf{X})+2]$ has minimum MSE among all translation-invariant estimators of σ^2 that are scale equivariant—a

translation-invariant estimator, say $t(\mathbf{y})$, of σ^2 is to be regarded as *scale equivariant* if $t(c\mathbf{y}) = c^2 t(\mathbf{y})$ for every (strictly) positive scalar c and for every nonnull value of $\mathbf{y}$ in $\mathfrak{N}(\mathbf{X}')$.

Exercise 23. Let $\mathbf{z} = (z_1, z_2, \ldots, z_M)'$ represent an M-dimensional random (column) vector that has an absolutely continuous distribution with a pdf $f(\cdot)$. And suppose that for some (nonnegative) function $g(\cdot)$ (of a single nonnegative variable), $f(\mathbf{z}) \propto g(\mathbf{z}'\mathbf{z})$ (in which case the distribution of $\mathbf{z}$ is spherical). Show (for $i = 1, 2, \ldots, M$) that $\mathrm{E}(z_i^2)$ exists if and only if $\int_0^\infty s^{M+1} g(s^2)\, ds < \infty$, in which case

$$\mathrm{var}(z_i) = \mathrm{E}(z_i^2) = \frac{1}{M} \frac{\int_0^\infty s^{M+1} g(s^2)\, ds}{\int_0^\infty s^{M-1} g(s^2)\, ds}.$$

Exercise 24. Let $\mathbf{z}$ represent an N-dimensional random column vector, and let $\mathbf{z}_*$ represent an M-dimensional subvector of $\mathbf{z}$ (where $M < N$). And suppose that the distributions of $\mathbf{z}$ and $\mathbf{z}_*$ are absolutely continuous with pdfs $f(\cdot)$ and $f_*(\cdot)$, respectively. Suppose also that there exist (nonnegative) functions $g(\cdot)$ and $g_*(\cdot)$ (of a single nonnegative variable) such that (for every value of $\mathbf{z}$) $f(\mathbf{z}) = g(\mathbf{z}'\mathbf{z})$ and (for every value of $\mathbf{z}_*$) $f_*(\mathbf{z}_*) = g_*(\mathbf{z}_*'\mathbf{z}_*)$ (in which case the distributions of $\mathbf{z}$ and $\mathbf{z}_*$ are spherical).

(a) Show that (for $v \geq 0$)

$$g_*(v) = \frac{\pi^{(N-M)/2}}{\Gamma[(N-M)/2]} \int_v^\infty (u - v)^{[(N-M)/2]-1} g(u)\, du.$$

(b) Show that if $N - M = 2$, then (for $v > 0$)

$$g(v) = -\frac{1}{\pi} g_*'(v),$$

where $g_*'(\cdot)$ is the derivative of $g_*(\cdot)$.

Exercise 25. Let $\mathbf{y}$ represent an $N \times 1$ random vector and $\mathbf{w}$ an $M \times 1$ random vector. Suppose that the second-order moments of the joint distribution of $\mathbf{y}$ and $\mathbf{w}$ exist, and adopt the following notation: $\boldsymbol{\mu}_y = \mathrm{E}(\mathbf{y})$, $\boldsymbol{\mu}_w = \mathrm{E}(\mathbf{w})$, $\mathbf{V}_y = \mathrm{var}(\mathbf{y})$, $\mathbf{V}_{yw} = \mathrm{cov}(\mathbf{y}, \mathbf{w})$, and $\mathbf{V}_w = \mathrm{var}(\mathbf{w})$. Further, assume that $\mathbf{V}_y$ is nonsingular.

(a) Show that the matrix $\mathbf{V}_w - \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{V}_{yw} - \mathrm{E}[\mathrm{var}(\mathbf{w} \mid \mathbf{y})]$ is nonnegative definite and that it equals $\mathbf{0}$ if and only if (for some nonrandom vector $\mathbf{c}$ and some nonrandom matrix $\mathbf{A}$) $\mathrm{E}(\mathbf{w} \mid \mathbf{y}) = \mathbf{c} + \mathbf{A}'\mathbf{y}$ (with probability 1).

(b) Show that the matrix $\mathrm{var}[\mathrm{E}(\mathbf{w} \mid \mathbf{y})] - \mathbf{V}_{yw}'\mathbf{V}_y^{-1}\mathbf{V}_{yw}$ is nonnegative definite and that it equals $\mathbf{0}$ if and only if (for some nonrandom vector $\mathbf{c}$ and some nonrandom matrix $\mathbf{A}$) $\mathrm{E}(\mathbf{w} \mid \mathbf{y}) = \mathbf{c} + \mathbf{A}'\mathbf{y}$ (with probability 1).

Exercise 26. Let $\mathbf{y}$ represent an $N \times 1$ observable random vector and $\mathbf{w}$ an $M \times 1$ unobservable random vector. Suppose that the second-order moments of the joint distribution of $\mathbf{y}$ and $\mathbf{w}$ exist, and adopt the following notation: $\boldsymbol{\mu}_y = \mathrm{E}(\mathbf{y})$, $\boldsymbol{\mu}_w = \mathrm{E}(\mathbf{w})$, $\mathbf{V}_y = \mathrm{var}(\mathbf{y})$, $\mathbf{V}_{yw} = \mathrm{cov}(\mathbf{y}, \mathbf{w})$, and $\mathbf{V}_w = \mathrm{var}(\mathbf{w})$. Assume that $\boldsymbol{\mu}_y$, $\boldsymbol{\mu}_w$, $\mathbf{V}_y$, $\mathbf{V}_{yw}$, and $\mathbf{V}_w$ are known. Further, define $\boldsymbol{\eta}(\mathbf{y}) = \boldsymbol{\mu}_w + \mathbf{V}_{yw}'\mathbf{V}_y^{-}(\mathbf{y} - \boldsymbol{\mu}_y)$, and take $\mathbf{t}(\mathbf{y})$ to be an $(M \times 1)$-dimensional vector-valued function of the form $\mathbf{t}(\mathbf{y}) = \mathbf{c} + \mathbf{A}'\mathbf{y}$, where $\mathbf{c}$ is a vector of constants and $\mathbf{A}$ is an $N \times M$ matrix of constants. Extend various of the results of Section 5.10a (to the case where $\mathbf{V}_y$ may be singular) by using Theorem 3.5.11 to show (1) that $\boldsymbol{\eta}(\mathbf{y})$ is the best linear predictor of $\mathbf{w}$ in the sense that the difference between the matrix $\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \mathbf{w}][\mathbf{t}(\mathbf{y}) - \mathbf{w}]'\}$ and the matrix $\mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}]$ [which is the MSE matrix of $\boldsymbol{\eta}(\mathbf{y})$] equals the matrix $\mathrm{E}\{[\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})][\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\mathbf{y})]'\}$, which is nonnegative definite and which equals $\mathbf{0}$ if and only if $\mathbf{t}(\mathbf{y}) = \boldsymbol{\eta}(\mathbf{y})$ for every value of $\mathbf{y}$ such that $\mathbf{y} - \boldsymbol{\mu}_y \in \mathcal{C}(\mathbf{V}_y)$, (2) that $\Pr[\mathbf{y} - \boldsymbol{\mu}_y \in \mathcal{C}(\mathbf{V}_y)] = 1$, and (3) that $\mathrm{var}[\boldsymbol{\eta}(\mathbf{y}) - \mathbf{w}] = \mathbf{V}_w - \mathbf{V}_{yw}'\mathbf{V}_y^{-}\mathbf{V}_{yw}$.

Exercise 27. Suppose that $\mathbf{y}$ is an $N \times 1$ observable random vector that follows a G–M model, and take $\mathbf{w}$ to be an $M \times 1$ unobservable random vector whose value is to be predicted. Suppose further that $\mathrm{E}(\mathbf{w})$ is of the form $\mathrm{E}(\mathbf{w}) = \boldsymbol{\Lambda}'\boldsymbol{\beta}$ (where $\boldsymbol{\Lambda}$ is a matrix of known constants) and that $\mathrm{cov}(\mathbf{y}, \mathbf{w})$ is of the form $\mathrm{cov}(\mathbf{y}, \mathbf{w}) = \sigma^2\mathbf{H}_{yw}$ (where $\mathbf{H}_{yw}$ is a known matrix). Let $\boldsymbol{\tau} = (\boldsymbol{\Lambda}' - \mathbf{H}'_{yw}\mathbf{X})\boldsymbol{\beta}$, denote by $\tilde{\mathbf{w}}(\mathbf{y})$ an arbitrary predictor (of $\mathbf{w}$), and define $\tilde{\boldsymbol{\tau}}(\mathbf{y}) = \tilde{\mathbf{w}}(\mathbf{y}) - \mathbf{H}'_{yw}\mathbf{y}$. Verify that $\tilde{\mathbf{w}}(\mathbf{y})$ is a translation-equivariant predictor (of $\mathbf{w}$) if and only if $\tilde{\boldsymbol{\tau}}(\mathbf{y})$ is a translation-equivariant estimator of $\boldsymbol{\tau}$.

Bibliographic and Supplementary Notes

§2. In some presentations, the use of the term translation (or location) invariance is extended to include what is herein referred to as translation equivariance.

§3e. For an extensive (book-length) discussion of mixture data, refer to Cornell (2002).

§4c. My acquaintance with the term conjugate normal equations came through some class notes authored by Oscar Kempthorne.

§4d. For a discussion of projections that is considerably more extensive and at a somewhat more general level than that provided herein, refer, for example, to Harville (1997, chaps. 12 and 17).

§7a. For a relatively extensive discussion of the vec and vech operations and of Kronecker products, refer, for example, to Chapter 16 of Harville's (1997) book and to the references cited therein.

§7c. Justification for referring to the estimator (7.44) as the Hodges–Lehmann estimator is provided by results presented by Hodges and Lehmann in their 1951 paper. Refer to the expository note by David (2009) for some discussion of a historical nature that relates to the statistical independence of the residual sum of squares and a least squares estimator.

§7d. The results in this subsection that pertain to the minimum-variance quadratic unbiased translation-invariant estimation of σ^2 (and the related results that are the subject of Exercise 14) are variations on the results of Atiqullah (1962) on the minimum-variance quadratic unbiased nonnegative-definite estimation of σ^2, which are covered by Ravishanker and Dey (2002) in their Section 4.4—Exercise 13 serves to relate the minimum-variance quadratic unbiased nonnegative-definite estimation of σ^2 to the minimum-variance quadratic unbiased translation-invariant estimation of σ^2.

§9b. REML originated with the work of Patterson and R. Thompson (1971)—while related ideas can be found in earlier work by others (e.g., W. A. Thompson, Jr. 1962), it was Patterson and R. Thompson who provided the kind of substantive development that was needed for REML to become a viable alternative to ordinary ML. The discussion of maximal invariants is based on results presented (in a more general context) in Section 6.2 of Lehmann and Romano's (2005b) book. Refer to Verbyla (1990) and to LaMotte (2007) for discussion of various matters pertaining to the derivation of expression (9.38) (and related expressions) for the log-likelihood function $\ell(\boldsymbol{\theta}; \mathbf{R}'\underline{\mathbf{y}})$ employed in REML in making inferences about functions of $\boldsymbol{\theta}$.

§9c. Refer, e.g., to Kollo and von Rosen (2005, table 2.3.1) or Fang, Kotz, and Ng (1990, table 3.1) for a table [originating with Jensen (1985)] that characterizes (in terms of the pdf or the characteristic function) various subclasses of multidimensional spherical distributions.

§10a. The approach (to point prediction) taken in this subsection is essentially the same as that taken in Harville's (1985) paper.

Exercises 7, 8, and 9. For a relatively extensive discussion of generalized inverses that satisfy one or more of Moore–Penrose conditions (2)–(4) [as well as Moore–Penrose condition (1)], including least squares generalized inverses, minimum norm generalized inverses, and the Moore–Penrose inverse itself, refer, e.g., to Harville (1997, chap. 20).

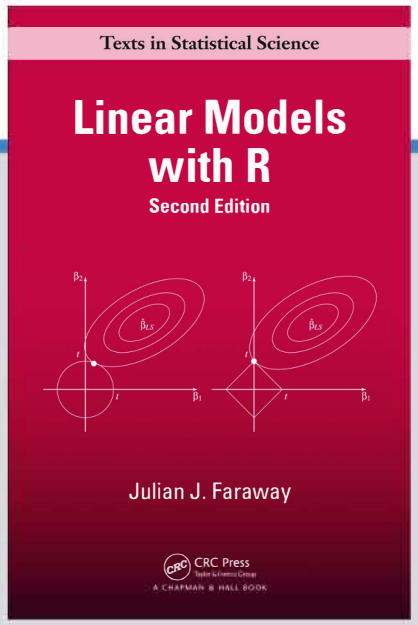
Exercise 20. For some general discussion bearing on the implications of Parts (b) and (c) of Exercise 20, refer to Sprott (1975).

Exercise 25. Exercise 25 is based on results presented by Harville (2003a).

CHAPTER

3

INSURANCE REDLINING - A COMPLETE EXAMPLE



Chapter 4

Introducing GAMs

4.1 Introduction

A generalized additive model (Hastie and Tibshirani, 1986, 1990) is a generalized linear model with a linear predictor involving a sum of smooth functions of covariates. In general the model has a structure something like

$$g(\mu_i) = \mathbf{A}_i\boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots \tag{4.1}$$

where $\mu_i \equiv \mathbb{E}(Y_i)$ and $Y_i \sim \text{EF}(\mu_i, \phi)$. Y_i is a response variable, $\text{EF}(\mu_i, \phi)$ denotes an exponential family distribution with mean μ_i and scale parameter, ϕ, $\mathbf{A}_i$ is a row of the model matrix for any strictly parametric model components, $\boldsymbol{\theta}$ is the corresponding parameter vector, and the f_j are smooth functions of the covariates, x_k. The model allows for flexible specification of the dependence of the response on the covariates, but by specifying the model only in terms of 'smooth functions', rather than detailed parametric relationships, it is possible to avoid the sort of cumbersome and unwieldy models seen in section 3.3.5, for example. This flexibility and convenience comes at the cost of two new theoretical problems. It is necessary both to represent the smooth functions in some way and to choose how smooth they should be.

This chapter illustrates how GAMs can be represented using basis expansions for each smooth, each with an associated penalty controlling function smoothness. Estimation can then be carried out by penalized regression methods, and the appropriate degree of smoothness for the f_j can be estimated from data using cross validation or marginal likelihood maximization. To avoid obscuring the basic simplicity of the approach with a mass of technical detail, the most complicated model considered here will be a simple GAM with two univariate smooth components. Furthermore, the methods presented will not be those that are most suitable for general practical use, being rather the methods that enable the basic framework to be explained simply. The ideal way to read this chapter is sitting at a computer, working through the statistics, and its implementation in R, side by side. If adopting this approach recall that the help files for R functions can be accessed by typing `?` followed by the function name, at the command line (e.g., `?lm`, for help on the linear modelling function).

4.2 Univariate smoothing

The representation and estimation of component functions of a model is best introduced by considering a model containing one function of one covariate,

$$y_i = f(x_i) + \epsilon_i, \tag{4.2}$$

where y_i is a response variable, x_i a covariate, f a smooth function and the ϵ_i are independent $N(0, \sigma^2)$ random variables.

4.2.1 Representing a function with basis expansions

To estimate f, using the methods covered in chapters 1 to 3, requires that f be represented in such a way that (4.2) becomes a linear model. This can be done by choosing a *basis*, defining the space of functions of which f (or a close approximation to it) is an element. Choosing a basis amounts to choosing some *basis functions*, which will be treated as completely known: if $b_j(x)$ is the j^{th} such basis function, then f is assumed to have a representation

$$f(x) = \sum_{j=1}^{k} b_j(x)\beta_j, \tag{4.3}$$

for some values of the unknown parameters, β_j. Substituting (4.3) into (4.2) clearly yields a linear model.

A very simple basis: Polynomials

As a simple example, suppose that f is believed to be a 4^{th} order polynomial, so that the space of polynomials of order 4 and below contains f. A basis for this space is $b_1(x) = 1$, $b_2(x) = x$, $b_3(x) = x^2$, $b_4(x) = x^3$ and $b_5(x) = x^4$, so that (4.3) becomes

$$f(x) = \beta_1 + x\beta_2 + x^2\beta_3 + x^3\beta_4 + x^4\beta_5,$$

and (4.2) becomes the simple model

$$y_i = \beta_1 + x_i\beta_2 + x_i^2\beta_3 + x_i^3\beta_4 + x_i^4\beta_5 + \epsilon_i.$$

Figures 4.1 and 4.2 illustrate a basis function representation of a function, f, using a polynomial basis.

The problem with polynomials

Taylor's theorem implies that polynomial bases will be useful for situations in which interest focuses on properties of f in the vicinity of a single specified point. But when the questions of interest relate to f over its whole domain, polynomial bases are problematic (see exercise 1).

The difficulties are most easily illustrated in the context of interpolation. The middle panel of figure 4.3 illustrates an attempt to approximate the function shown in

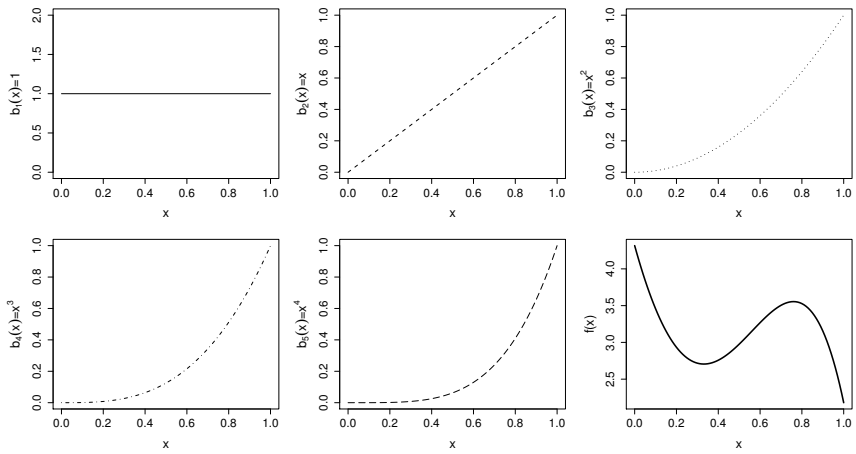


Figure 4.1 *Illustration of the idea of representing a function in terms of basis functions, using a polynomial basis. The first 5 panels (starting from top left) illustrate the 5 basis functions, $b_j(x)$, for a 4th order polynomial basis. The basis functions are each multiplied by a real valued parameter, β_j, and are then summed to give the final curve $f(x)$, an example of which is shown in the bottom right panel. By varying the β_j, we can vary the form of $f(x)$, to produce any polynomial function of order 4 or lower. See also figure 4.2*

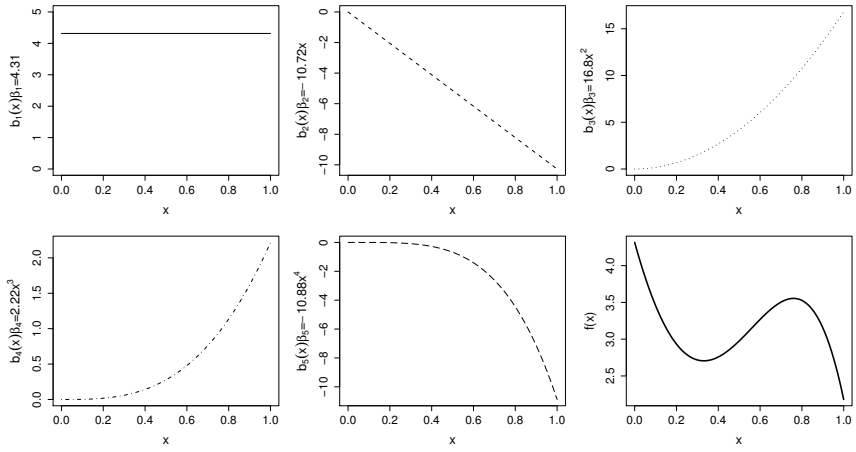


Figure 4.2 *An alternative illustration of how a function is represented in terms of basis functions. As in figure 4.1, a 4th order polynomial basis is illustrated. In this case the 5 basis functions, $b_j(x)$, each multiplied by its coefficient β_j, are shown in the first five figures (starting at top left). Simply summing these 5 curves yields the function, $f(x)$, shown at bottom right.*

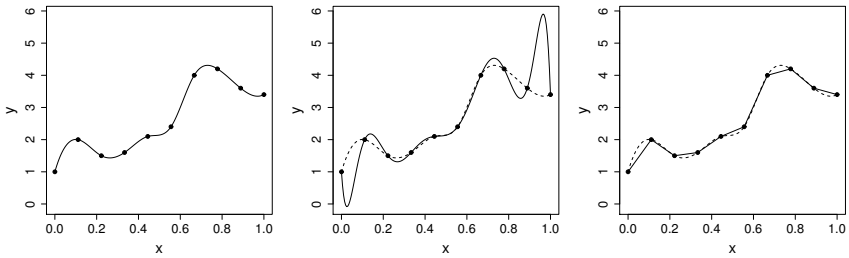


Figure 4.3 *The left panel shows a smooth function sampled at the points shown as black dots. The middle panel shows an attempt to reconstruct the function (dashed curve) by polynomial interpolation (black curve) of the black dots. The right panel shows the equivalent piecewise linear interpolant. The condition that the polynomial should interpolate the data and have all its derivatives continuous leads to quite wild oscillation.*

the left panel of figure 4.3, by polynomial interpolation of the points shown as black dots. The polynomial oscillates wildly in places, in order to accommodate the twin requirements to interpolate the data *and* to have all derivatives w.r.t. x continuous. If we relax the requirement for continuity of derivatives, and simply use a piecewise linear interpolant, then a much better approximation is obtained, as the right hand panel of figure 4.3 illustrates.

It clearly makes sense to use bases that are good at approximating known functions in order to represent unknown functions. Similarly, bases that perform well for interpolating exact observations of a function are also a good starting point for the closely related task of smoothing noisy observations of a function. In subsequent chapters we will see that piecewise linear bases can be improved upon by spline bases having continuity of just a few derivatives, but the piecewise linear case provides such a convenient illustration that we will stick with it for this chapter.

The piecewise linear basis

A basis for piecewise linear functions of a univariate variable x is determined entirely by the locations of the function's derivative discontinuities, that is by the locations at which the linear pieces join up. Let these *knots* be denoted $\{x_j^* : j = 1, \cdots k\}$, and suppose that $x_j^* > x_{j-1}^*$. Then for $j = 2, \ldots, k-1$

$$b_j(x) = \begin{cases} (x - x_{j-1}^*)/(x_j^* - x_{j-1}^*) & x_{j-1}^* < x \leq x_j^* \\ (x_{i+j}^* - x)/(x_{j+1}^* - x_j^*) & x_j^* < x < x_{j+1}^* \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

while

$$b_1(x) = \begin{cases} (x_2^* - x)/(x_2^* - x_1^*) & x < x_2^* \\ 0 & \text{otherwise} \end{cases}$$

and

$$b_k(x) = \begin{cases} (x - x_{k-1}^*)/(x_k^* - x_{k-1}^*) & x > x_{k-1}^* \\ 0 & \text{otherwise} \end{cases}$$

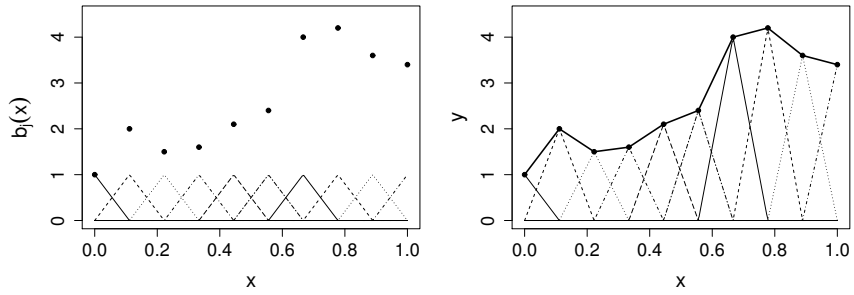


Figure 4.4 *The left panel shows an example tent function basis for interpolating the data shown as black dots. The continuous lines show the tent function basis functions, each of which peaks with value 1 at the x-axis value of one of the data points. The right panel illustrates how the basis functions are each multiplied by a coefficient, before being summed to give the interpolant, shown as the thick black line.*

So $b_j(x)$ is zero everywhere, except over the interval between the knots immediately to either side of x_j^*. $b_j(x)$ increases linearly from 0 at x_{j-1}^* to 1 at x_j^*, and then decreases linearly to 0 at x_{j+1}^*. Basis functions like this, that are non zero only over some finite intervals, are said to have *compact support*. Because of their shape the b_j are often known as *tent functions*. See figure 4.4.

Notice that an exactly equivalent way of defining $b_j(x)$ is as the linear interpolant of the data $\{x_i^*, \delta_i^j : i = 1, \ldots, k\}$ where $\delta_i^j = 1$ if $i = j$ and zero otherwise. This definition makes for very easy coding of the basis in R.

Using this basis to represent $f(x)$, (4.2) now becomes the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $X_{ij} = b_j(x_i)$.

Using the piecewise linear basis

Now consider an illustrative example. It is often claimed, at least by people with little actual knowledge of engines, that a car engine with a larger cylinder capacity will wear out less quickly than a smaller capacity engine. Figure 4.5 shows some data for 19 Volvo engines. The pattern of variation is not entirely clear, so (4.2) might be an appropriate model.

First read the data into R.

```
require(gamair); data(engine); attach(engine)
plot(size,wear,xlab="Engine capacity",ylab="Wear index")
```

Now write an R function defining $b_j(x)$

```
tf <- function(x,xj,j) {
## generate jth tent function from set defined by knots xj
  dj <- xj*0;dj[j] <- 1
  approx(xj,dj,x)$y
}
```

and use it to write an R function that will take a sequence of knots and an array of x values to produce a model matrix for the piecewise linear function.

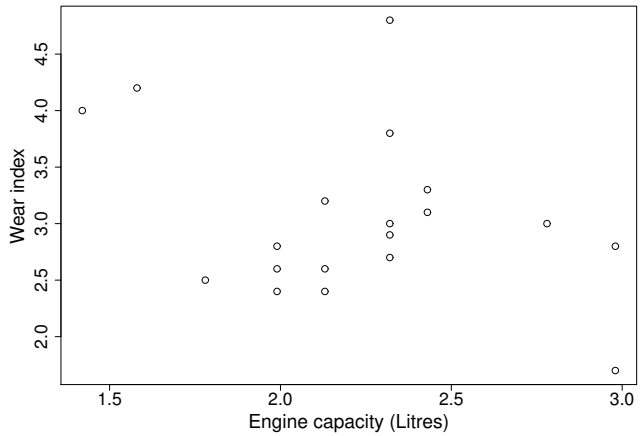


Figure 4.5 *Data on engine wear index versus engine capacity for 19 Volvo car engines, obtained from* `http://www3.bc.sympatico.ca/Volvo_Books/engine3.html`.

```
tf.X <- function(x,xj) {
## tent function basis matrix given data x
## and knot sequence xj
  nk <- length(xj); n <- length(x)
  X <- matrix(NA,n,nk)
  for (j in 1:nk) X[,j] <- tf(x,xj,j)
  X
}
```

All that is required now is to select a set of knots, x_j^*, and the model can be fitted. In the following a rank $k = 6$ basis is used, with the knots spread evenly over the range of the `size` data.

```
sj <- seq(min(size),max(size),length=6) ## generate knots
X <- tf.X(size,sj)                      ## get model matrix
b <- lm(wear ~ X - 1)                   ## fit model
s <- seq(min(size),max(size),length=200)## prediction data
Xp <- tf.X(s,sj)                        ## prediction matrix
plot(size,wear)                         ## plot data
lines(s,Xp %*% coef(b))                 ## overlay estimated f
```

The model fit looks quite plausible (figure 4.6), but the choice of degree of model smoothness, controlled here by the basis dimension, k, was essentially arbitrary. This issue must be addressed if a satisfactory theory for modelling with unknown functions is to be developed.

4.2.2 Controlling smoothness by penalizing wiggliness

One obvious possibility for choosing the degree of smoothing is to try to make use of the hypothesis testing methods from chapter 1, to select k by backwards selection.

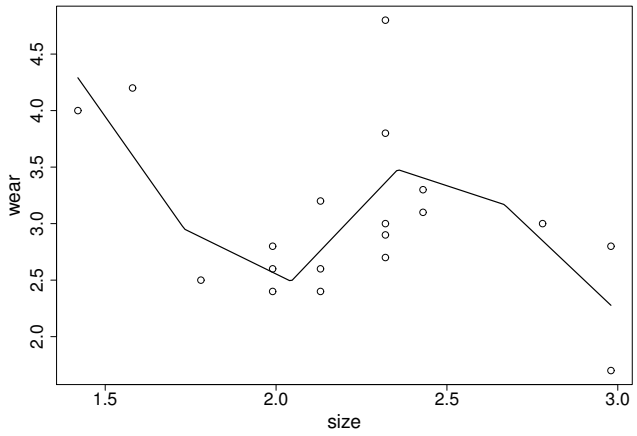


Figure 4.6 *Piecewise linear regression fit (continuous line) to data (∘) on engine wear index versus engine capacity (size) for 19 Volvo car engines.*

However, such an approach is problematic, since a model based on $k-1$ evenly spaced knots will not generally be nested within a model based on k evenly spaced knots. It is possible to start with a fine grid of knots and simply drop knots sequentially, as part of backward selection, but the resulting uneven knot spacing can itself lead to poor model performance. Furthermore, the fit of such regression models tends to depend quite strongly on the locations chosen for the knots.

An alternative is to keep the basis dimension fixed at a size a little larger than it is believed could reasonably be necessary, but to control the model's smoothness by adding a 'wiggliness' penalty to the least squares fitting objective. For example, rather than fitting the model by minimizing

$$\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2,$$

it could be fitted by minimizing

$$\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda\sum_{j=2}^{k-1}\{f(x^*_{j-1})-2f(x^*_j)+f(x^*_{j+1})\}^2,$$

where the summation term measures wiggliness as a sum of squared second differences of the function at the knots (which crudely approximates the integrated squared second derivative penalty used in cubic spline smoothing: see section 5.1.2, p. 198). When f is very wiggly the penalty will take high values and when f is 'smooth' the penalty will be low.* If f is a straight line then the penalty is actually zero. So the penalty has a *null space* of functions that are un-penalized: the straight lines in this

*Note that even knot spacing has been assumed: uneven knot spacing would usually require some re-weighting of the penalty terms.

case. The dimension of the penalty null space is 2, since the basis for straight lines is 2-dimensional.

The *smoothing parameter*, λ, controls the trade-off between smoothness of the estimated f and fidelity to the data. $\lambda \to \infty$ leads to a straight line estimate for f, while $\lambda = 0$ results in an un-penalized piecewise linear regression estimate.

For the basis of tent functions, it is easy to see that the coefficients of f are simply the function values at the knots, i.e., $\beta_j = f(x_j^*)$. This makes it particularly straightforward to express the penalty as a quadratic form, $\boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta}$, in the basis coefficients (although in fact linearity of f in the basis coefficients is all that is required for this). Firstly note that

$$\begin{bmatrix} \beta_1 - 2\beta_2 + \beta_3 \\ \beta_2 - 2\beta_3 + \beta_4 \\ \beta_3 - 2\beta_4 + \beta_5 \\ . \\ . \end{bmatrix} = \begin{bmatrix} 1 & -2 & 1 & 0 & . & . & . \\ 0 & 1 & -2 & 1 & 0 & . & . \\ 0 & 0 & 1 & -2 & 1 & 0 & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ . \\ . \end{bmatrix}$$

so that writing the right hand side as $\mathbf{D}\boldsymbol{\beta}$, by definition of $(k-2) \times k$ matrix $\mathbf{D}$, the penalty becomes

$$\sum_{j=2}^{k-1} (\beta_{j-1} - 2\beta_j + \beta_{j+1})^2 = \boldsymbol{\beta}^\mathsf{T}\mathbf{D}^\mathsf{T}\mathbf{D}\boldsymbol{\beta} = \boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta} \tag{4.5}$$

where $\mathbf{S} = \mathbf{D}^\mathsf{T}\mathbf{D}$ ($\mathbf{S}$ is obviously rank deficient by the dimension of the penalty null space). Hence the penalized regression fitting problem is to minimize

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta} \tag{4.6}$$

w.r.t. $\boldsymbol{\beta}$. The problem of estimating the degree of smoothness for the model is now the problem of estimating the smoothing parameter λ. But before addressing λ estimation, consider $\boldsymbol{\beta}$ estimation given λ.

It is fairly straightforward to show (see exercise 3) that the formal expression for the minimizer of (4.6), the penalized least squares estimator of $\boldsymbol{\beta}$, is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{S}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}. \tag{4.7}$$

Similarly the influence (hat) matrix, $\mathbf{A}$, for the model can be written

$$\mathbf{A} = \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{S}\right)^{-1}\mathbf{X}^\mathsf{T}.$$

Recall that $\hat{\boldsymbol{\mu}} = \mathbf{A}\mathbf{y}$. As with the un-penalized linear model, these expressions are not the ones to use for computation, for which the greater numerical stability offered by orthogonal matrix methods is to be preferred. For practical computation, therefore, note that

$$\left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} \boldsymbol{\beta} \right\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta}.$$

Obviously any matrix square root such that $\mathbf{D}^{\mathsf{T}}\mathbf{D} = \mathbf{S}$ could be substituted for the original $\mathbf{D}$ here, but at the moment there is no reason to use an alternative. The sum of squares term, on the left hand side, is just a least squares objective for a model in which the model matrix has been augmented by a square root of the penalty matrix, while the response data vector has been augmented with $k - 2$ zeros. Fitting the augmented linear model will therefore yield $\hat{\boldsymbol{\beta}}$.

To see a penalized regression spline in action, first note that $\mathbf{D}$ can be obtained in R using `diff(diag(k),differences=2)`, which applies second order differencing to each column of the rank k identity matrix. Now it is easy to write a simple function for fitting a penalized piecewise linear smoother.

```
prs.fit <- function(y,x,xj,sp) {
  X <- tf.X(x,xj)       ## model matrix
  D <- diff(diag(length(xj)),differences=2) ## sqrt penalty
  X <- rbind(X,sqrt(sp)*D) ## augmented model matrix
  y <- c(y,rep(0,nrow(D))) ## augmented data
  lm(y ~ X - 1)    ## penalized least squares fit
}
```

To use this function, we need to choose the basis dimension, k, the (evenly spaced) knot locations, x_j^*, and a value for the smoothing parameter, λ. Provided that k is large enough that the basis is more flexible than we expect to *need* to represent $f(x)$, then neither the exact choice of k, nor the precise selection of knot locations, has a great deal of influence on the model fit. Rather it is the choice of λ that now plays the crucial role in determining model flexibility, and ultimately the shape of $\hat{f}(x)$. In the following example $k = 20$ and the knots are evenly spread out over the range of observed engine sizes. It is the smoothing parameter, $\lambda = 2$, which really controls the behaviour of the fitted model.

```
sj <- seq(min(size),max(size),length=20) ## knots
b <- prs.fit(wear,size,sj,2) ## penalized fit
plot(size,wear)   ## plot data
Xp <- tf.X(s,sj)  ## prediction matrix
lines(s,Xp %*% coef(b)) ## plot the smooth
```

By changing the value of the smoothing parameter, λ, a variety of models of different smoothness can be obtained. Figure 4.7 illustrates this, but begs the question, which value of λ is 'best'?

4.2.3 Choosing the smoothing parameter, λ, by cross validation

If λ is too high then the data will be over-smoothed, and if it is too low then the data will be under-smoothed: in both cases this will mean that the estimate $\hat{f}$ will not be close to the true function f. Ideally, it would be good to choose λ so that $\hat{f}$ is as close as possible to f. A suitable criterion might be to choose λ to minimize

$$M = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i - f_i)^2,$$

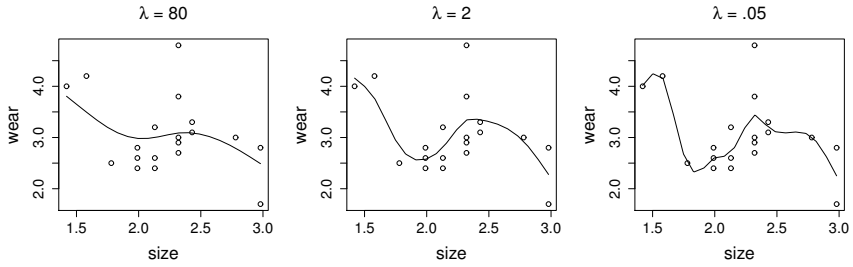


Figure 4.7 *Penalized piecewise linear fits to the engine wear versus capacity data, using three different values for the smoothing parameter, λ. Notice how penalization produces quite smooth estimates, despite the piecewise linear basis.*

where the notation $\hat{f}_i \equiv \hat{f}(x_i)$ and $f_i \equiv f(x_i)$ has been adopted for conciseness. Since f is unknown, M cannot be used directly, but it is possible to derive an estimate of $\mathbb{E}(M) + \sigma^2$, which is the expected squared error in predicting a new variable. Let $\hat{f}^{[-i]}$ be the model fitted to all data except y_i, and define the *ordinary cross validation* score

$$\mathcal{V}_o = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - y_i)^2.$$

This score results from leaving out each datum in turn, fitting the model to the remaining data and calculating the squared difference between the missing datum and its predicted value: these squared differences are then averaged over all the data. Substituting $y_i = f_i + \epsilon_i$,

$$\begin{aligned}\mathcal{V}_o &= \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i - \epsilon_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i)^2 - 2(\hat{f}_i^{[-i]} - f_i)\epsilon_i + \epsilon_i^2.\end{aligned}$$

Since $\mathbb{E}(\epsilon_i) = 0$, and ϵ_i and $\hat{f}_i^{[-i]}$ are independent, the second term in the summation vanishes if expectations are taken:

$$\mathbb{E}(\mathcal{V}_o) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i)^2\right) + \sigma^2.$$

Now, $\hat{f}^{[-i]} \approx \hat{f}$ with equality in the large sample limit, so $\mathbb{E}(\mathcal{V}_o) \approx \mathbb{E}(M) + \sigma^2$ also with equality in the large sample limit. Hence choosing λ in order to minimize $\mathcal{V}_o$ is a reasonable approach if the ideal would be to minimize M. Choosing λ to minimize $\mathcal{V}_o$ is known as ordinary cross validation.

Ordinary cross validation is a reasonable approach, in its own right, even without a mean square (prediction) error justification. If models are judged only by their ability to fit the data from which they were estimated, then complicated models are

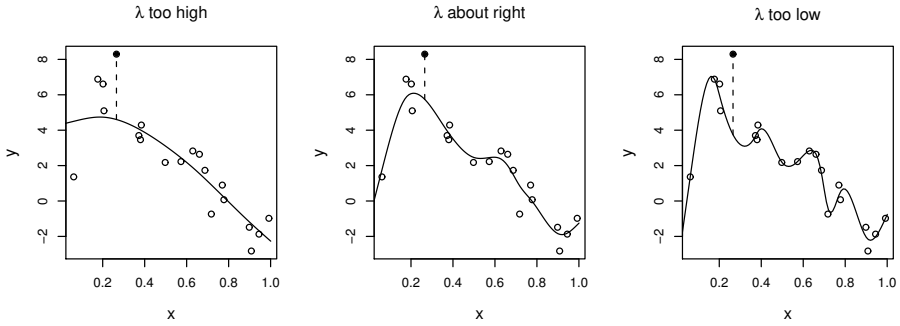


Figure 4.8 *Illustration of the principle behind cross validation. The fifth datum (•) has been omitted from fitting and the continuous curve shows a penalized regression spline fitted to the remaining data (○). When the smoothing parameter is too high the spline fits many of the data poorly and does no better with the missing point. When λ is too low the spline fits the noise as well as the signal and the consequent extra variability causes it to predict the missing datum poorly. For intermediate λ the spline is fitting the underlying signal quite well, but smoothing through the noise: hence, the missing datum is reasonably well predicted. Cross validation leaves out each datum from the data in turn and considers the average ability of models fitted to the remaining data to predict the omitted data.*

always selected over simpler ones. Choosing a model in order to maximize the ability to predict data to which the model was not fitted, does not suffer from this problem, as figure 4.8 illustrates.

It is computationally costly to calculate $\mathcal{V}_o$ by leaving out one datum at a time, refitting the model to each of the n resulting data sets, but it can be shown that

$$\mathcal{V}_o = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_i)^2/(1 - A_{ii})^2,$$

where $\hat{f}$ is the estimate from fitting to all the data, and $\mathbf{A}$ is the corresponding influence matrix (see section 6.2.2, p. 256). In practice the A_{ii} are often replaced by their mean, $\mathrm{tr}\,(\mathbf{A})/n$, resulting in the *generalized cross validation* score

$$\mathcal{V}_g = \frac{n\sum_{i=1}^{n}(y_i - \hat{f}_i)^2}{[n - \mathrm{tr}\,(\mathbf{A})]^2}.$$

GCV has computational advantages over OCV, and it also has advantages in terms of invariance (see Wahba, 1990, p.53 or sections 6.2.2 and 6.2.3, p. 258). In any case, it can also be shown to minimize $\mathbb{E}(M)$ in the large sample limit.

Returning to the engine wear example, a simple direct search for the GCV optimal smoothing parameter can be made as follows.

```
rho = seq(-9,11,length=90)
n <- length(wear)
V <- rep(NA,90)
```

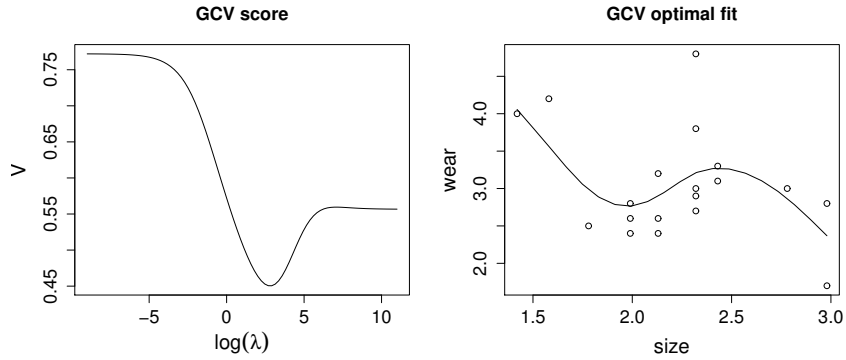


Figure 4.9 *Left panel: the GCV function for the engine wear example against log smoothing parameter. Right panel: the fitted model which minimizes the GCV score.*

```
for (i in 1:90) { ## loop through smoothing params
  b <- prs.fit(wear,size,sj,exp(rho[i])) ## fit model
  trF <- sum(influence(b)$hat[1:n])      ## extract EDF
  rss <- sum((wear-fitted(b)[1:n])^2)    ## residual SS
  V[i] <- n*rss/(n-trF)^2                ## GCV score
}
```

Note that the `influence()` function returns a list of diagnostics including `hat`, an array of the elements on the leading diagonal of the influence/hat matrix of the augmented model. The first n of these are the leading diagonal of the influence matrix of the penalized model (see exercise 4).

For the example, `V[54]` is the lowest GCV score, so that the optimal smoothing parameter, from those tried, is $\hat{\lambda} \approx 18$. Plots of the GCV score and the optimal model are easily produced

```
plot(rho,V,type="l",xlab=expression(log(lambda)),
                    main="GCV score")
sp <- exp(rho[V==min(V)])     ## extract optimal sp
b <- prs.fit(wear,size,sj,sp) ## re-fit
plot(size,wear,main="GCV optimal fit")
lines(s,Xp %*% coef(b))
```

The results are shown in figure 4.9.

4.2.4 The Bayesian/mixed model alternative

At some level we introduce smoothing penalties because we believe that the truth is more likely to be smooth than wiggly. We might as well formalise this belief in a Bayesian way, and specify a prior distribution on function wiggliness. Perhaps the simplest choice is an exponential prior

$$\propto \exp(-\lambda \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S} \boldsymbol{\beta} / \sigma^2)$$

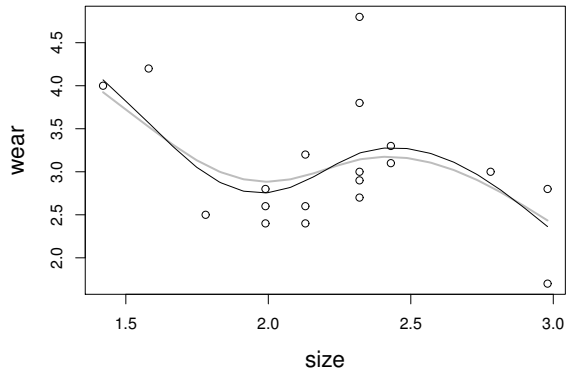


Figure 4.10 *Smooth model fits to the engine wear data with smoothing parameters estimated using marginal likelihood maximization (grey) or REML (black).*

(where scaling by σ^2 is introduced merely for later convenience), but this is immediately recognisable as being equivalent to an improper multivariate normal prior $\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2 \mathbf{S}^-/\lambda)$. That is, the prior precision matrix is proportional to $\mathbf{S}$: because $\mathbf{S}$ is rank deficient by the dimension of the penalty null space, the prior covariance matrix is proportional to the pseudo-inverse[†] $\mathbf{S}^-$.

This Bayesian interpretation of the smoothing penalty gives the model the structure of a linear mixed model as discussed in chapter 2, and in consequence the MAP estimate of $\boldsymbol{\beta}$ is the solution (4.7) to (4.6), while

$$\boldsymbol{\beta}|\mathbf{y} \sim N(\hat{\boldsymbol{\beta}}, (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{S})^{-1}\sigma^2)$$

— the Bayesian posterior distribution of $\boldsymbol{\beta}$ (this is equivalent to (2.17), p. 80). Also, having given the model this extra structure opens up the possibility of estimating σ^2 and λ using marginal likelihood maximization or REML.

In this section we will re-parameterize slightly to get the smooth model into a form such that its marginal likelihood can be evaluated using the simple routine `llm` from section 2.4.4 (p. 81). R routine `optim` can be used to fit the model. The same re-parameterization allows the model to be easily estimated using `lme` (see section 2.5, p. 86). As we will see in chapter 6, this re-parameterization is not necessary: it just simplifies matters for the moment, and perhaps makes the relationship between fixed effects and the penalty null space clearer than might otherwise be the case.

The re-parameterization is to re-write the model in terms of parameters, $\boldsymbol{\beta}' = \mathbf{D}_+\boldsymbol{\beta}$ where

$$\mathbf{D}_+ = \begin{bmatrix} \mathbf{I}_2 & \mathbf{0} \\ & \mathbf{D} \end{bmatrix}.$$

So we now have $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\mathbf{D}_+^{-1}\boldsymbol{\beta}'$ and $\boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta} = \sum_{i=3}^{k} \beta_i'^2$. If we write the first 2

[†]Consider eigen-decomposition $\mathbf{S} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\mathsf{T}$. Let $\boldsymbol{\Lambda}^-$ denote the diagonal matrix of the inverse of the non-zero eigenvalues, with zeroes in place of the inverse for any zero eigenvalues. Then $\mathbf{S}^- = \mathbf{U}\boldsymbol{\Lambda}^-\mathbf{U}^\mathsf{T}$.

elements of $\boldsymbol{\beta}'$ as $\boldsymbol{\beta}^*$ and the remainder as $\mathbf{b}$, the Bayesian smoothing prior becomes $\mathbf{b} \sim N(\mathbf{0}, \mathbf{I}\sigma^2/\lambda)$ (which is proper). $\boldsymbol{\beta}^*$ is completely unpenalized, so we treat this as a vector of fixed effects. To make the connection to a standard mixed model completely clear, let $\mathbf{X}^*$ now denote the first 2 columns of $\mathbf{X}\mathbf{D}_+^{-1}$, while $\mathbf{Z}$ is the matrix of the remaining columns. Then the smooth model has become

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \quad \mathbf{b} \sim N(\mathbf{0}, \mathbf{I}\sigma^2/\lambda), \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$$

which is self-evidently in the standard form of a linear mixed model given in section 2.3, (p.77). Here is the code to re-parameterize the model and estimate it using `optim` and `llm` from section 2.4.4 (p. 81):

```
X0 <- tf.X(size,sj)          ## X in original parameterization
D <- rbind(0,0,diff(diag(20),difference=2))
diag(D) <- 1                    ## augmented D
X <- t(backsolve(t(D),t(X0)))  ## re-parameterized X
Z <- X[,-c(1,2)]; X <- X[,1:2] ## mixed model matrices
## estimate smoothing and variance parameters...
m <- optim(c(0,0),llm,method="BFGS",X=X,Z=Z,y=wear)
b <- attr(llm(m$par,X,Z,wear),"b") ## extract coefficients
## plot results...
plot(size,wear)
Xp1 <- t(backsolve(t(D),t(Xp))) ## re-parameterized pred. mat.
lines(s,Xp1 %*% as.numeric(b),col="grey",lwd=2)
```

The resulting plot is shown in figure 4.10.

Estimation using REML via `lme` is also easy. In `lme` terms all the data belong to a single group, so to use `lme` we must create a dummy grouping variable enforcing this. A covariance matrix proportional to the identity matrix is then specified via the `pdIdent` function.

```
library(nlme)
g <- factor(rep(1,nrow(X)))         ## dummy factor
m <- lme(wear ~ X - 1, random=list(g = pdIdent(~ Z-1)))
lines(s,Xp1 %*% as.numeric(coef(m))) ## and to plot
```

The curve of the estimated smooth is shown in black in figure 4.10. Notice how the REML based estimate (black) is more variable than the ML based estimate (grey), as expected from section 2.4.5 (p. 83).

4.3 Additive models

Now suppose that two explanatory variables, x and v, are available for a response variable, y, and that a simple additive model structure,

$$y_i = \alpha + f_1(x_i) + f_2(v_i) + \epsilon_i, \tag{4.8}$$

is appropriate. α is an intercept parameter, the f_j are smooth functions, and the ϵ_i are independent $N(0, \sigma^2)$ random variables.

There are two points to note about this model. Firstly, the assumption of additive effects is a fairly strong one: $f_1(x) + f_2(v)$ is a quite restrictive special case of

the general smooth function of two variables $f(x, v)$. Secondly, the fact that the model now contains more than one function introduces an identifiability problem: f_1 and f_2 are each only estimable to within an additive constant. To see this, note that any constant could be simultaneously added to f_1 and subtracted from f_2, without changing the model predictions. Hence identifiability constraints have to be imposed on the model before fitting.

Provided that the identifiability issue is addressed, the additive model can be represented using penalized regression splines, estimated by penalized least squares and the degree of smoothing selected by cross validation or (RE)ML, in the same way as for the simple univariate model.

4.3.1 Penalized piecewise regression representation of an additive model

Each smooth function in (4.8) can be represented using a penalized piecewise linear basis. Specifically, let

$$f_1(x) = \sum_{j=1}^{k_1} b_j(x)\delta_j$$

where the δ_j are unknown coefficients, while the $b_j(x)$ are basis functions of the form (4.4), defined using a sequence of k_1 knots, x_j^*, evenly spaced over the range of x. Similarly

$$f_2(v) = \sum_{j=1}^{k_2} \mathcal{B}_j(v)\gamma_j$$

where the γ_j are the unknown coefficients and the $\mathcal{B}_j(v)$ are basis functions of the form (4.4), defined using a sequence of k_2 knots, v_j^*, evenly spaced over the range of v. Defining n-vector $\mathbf{f}_1 = [f_1(x_1), \ldots, f_1(x_n)]^\mathsf{T}$, we have $\mathbf{f}_1 = \mathbf{X}_1\boldsymbol{\delta}$ where $b_j(x_i)$ is element i, j of $\mathbf{X}_1$. Similarly, $\mathbf{f}_2 = \mathbf{X}_2\boldsymbol{\gamma}$, where $\mathcal{B}_j(v_i)$ is element i, j of $\mathbf{X}_2$.

A penalty of the form (4.5) is also associated with each function: $\boldsymbol{\delta}^\mathsf{T}\mathbf{D}_1^\mathsf{T}\mathbf{D}_1\boldsymbol{\delta} = \boldsymbol{\delta}^\mathsf{T}\bar{\mathbf{S}}_1\boldsymbol{\delta}$ for f_1 and $\boldsymbol{\gamma}^\mathsf{T}\mathbf{D}_2^\mathsf{T}\mathbf{D}_2\boldsymbol{\gamma} = \boldsymbol{\gamma}^\mathsf{T}\bar{\mathbf{S}}_2\boldsymbol{\gamma}$ for f_2.

Now it is necessary to deal with the identifiability problem. For estimation purposes, almost any linear constraint that removed the problem could be used, but most choices lead to uselessly wide confidence intervals for the constrained functions. The best constraints from this viewpoint are sum-to-zero constraints, such as

$$\sum_{i=1}^{n} f_1(x_i) = 0, \text{ or equivalently } \mathbf{1}^\mathsf{T}\mathbf{f}_1 = 0,$$

where $\mathbf{1}$ is an n vector of 1's. Notice how this constraint still allows f_1 to have exactly the same shape as before constraint, with exactly the same penalty value. The constraint's only effect is to shift f_1, vertically, so that its mean value is zero.

To apply the constraint, note that we require $\mathbf{1}^\mathsf{T}\mathbf{X}_1\boldsymbol{\delta} = 0$ for all $\boldsymbol{\delta}$, which implies that $\mathbf{1}^\mathsf{T}\mathbf{X}_1 = \mathbf{0}$. To achieve this latter condition the column mean can be subtracted from each column of $\mathbf{X}_1$. That is, we define a column centred matrix

$$\tilde{\mathbf{X}}_1 = \mathbf{X}_1 - \mathbf{1}\mathbf{1}^\mathsf{T}\mathbf{X}_1/n$$

and set $\tilde{\mathbf{f}}_1 = \tilde{\mathbf{X}}_1\boldsymbol{\delta}$. It's easy to check that this constraint imposes no more than a shift in the level of $\mathbf{f}_1$:

$$\tilde{\mathbf{f}}_1 = \tilde{\mathbf{X}}_1\boldsymbol{\delta} = \mathbf{X}_1\boldsymbol{\delta} - \mathbf{1}\mathbf{1}^\mathsf{T}\mathbf{X}_1\boldsymbol{\delta}/n = \mathbf{X}_1\boldsymbol{\delta} - \mathbf{1}c = \mathbf{f}_1 - c$$

by definition of the scalar $c = \mathbf{1}^\mathsf{T}\mathbf{X}_1\boldsymbol{\delta}/n$. Finally note that the column centring reduces the rank of $\tilde{\mathbf{X}}_1$ to $k_1 - 1$, so that only $k_1 - 1$ elements of the k_1 vector $\boldsymbol{\delta}$ can be uniquely estimated. A simple identifiability constraint deals with this problem: a single element of $\boldsymbol{\delta}$ is set to zero, and the corresponding column of $\tilde{\mathbf{X}}_1$ and $\mathbf{D}$ is deleted.[‡] The column centred rank reduced basis will automatically satisfy the identifiability constraint. In what follows the tildes will be dropped, and it is assumed that the $\mathbf{X}_j$, $\mathbf{D}_j$, etc. are the constrained versions.

Here is an R function which produces constrained versions of $\mathbf{X}_j$ and $\mathbf{D}_j$.

```
tf.XD <- function(x,xk,cmx=NULL,m=2) {
## get X and D subject to constraint
  nk <- length(xk)
  X <- tf.X(x,xk)[,-nk]                    ## basis matrix
  D <- diff(diag(nk),differences=m)[,-nk] ## root penalty
  if (is.null(cmx)) cmx <- colMeans(X)
  X <- sweep(X,2,cmx)        ## subtract cmx from columns
  list(X=X,D=D,cmx=cmx)
}
```

`tf.XD` calls the functions producing the unconstrained basis and square root penalty matrices, given knot sequence `xk` and covariate values `x`. It drops a column of each resulting matrix and centres the remaining columns of the basis matrix. `cmx` is the vector of values to subtract from the columns of the `X`. For setting up a basis `cmx` should be `NULL`, in which case it is set to the column means of the basis matrix `X`. However, when using `tf.XD` to produce a basis matrix for *predicting* at new covariate values, it is essential that the basis matrix columns are centred using the same constants used for the *original* basis setup, so these must be supplied. Later code will clarify this.

Having set up constrained bases for the f_j it is now straightforward to re-express (4.8) as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{X} = (\mathbf{1}, \mathbf{X}_1, \mathbf{X}_2)$ and $\boldsymbol{\beta}^\mathsf{T} = (\alpha, \boldsymbol{\delta}^\mathsf{T}, \boldsymbol{\gamma}^\mathsf{T})$. Largely for later notational convenience it is useful to express the penalties as quadratic forms in the full coefficient vector $\boldsymbol{\beta}$, which is easily done by simply padding out $\bar{\mathbf{S}}_j$ with zeroes, as appropriate. For example,

$$\boldsymbol{\beta}^\mathsf{T}\mathbf{S}_1\boldsymbol{\beta} = (\alpha, \boldsymbol{\delta}^\mathsf{T}, \boldsymbol{\gamma}^\mathsf{T}) \begin{bmatrix} 0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{S}}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \alpha \\ \boldsymbol{\delta} \\ \boldsymbol{\gamma} \end{bmatrix} = \boldsymbol{\delta}^\mathsf{T}\bar{\mathbf{S}}_1\boldsymbol{\delta}.$$

[‡]The recipe given here is applicable to any basis which includes the constant function in its span, and has a penalty that is zero for constant functions. However, for bases that explicitly include a constant function, it is not sufficient to set any coefficient to zero: the coefficient for the constant is the one to constrain.

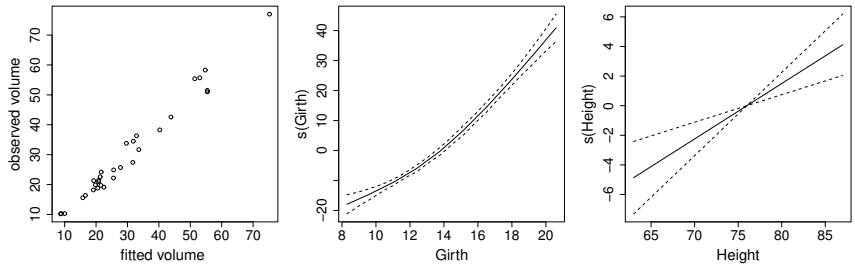


Figure 4.11 *The best fit two term additive model for the* `tree` *data. The left panel shows actual versus predicted tree volumes. The middle panel is the estimate of the smooth function of girth. The right panel is the estimate of the smooth function of height.*

4.3.2 Fitting additive models by penalized least squares

The coefficient estimates $\hat{\boldsymbol{\beta}}$ of the model (4.8) are obtained by minimization of the penalized least squares objective

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_1 \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_1 \boldsymbol{\beta} + \lambda_2 \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_2 \boldsymbol{\beta},$$

where the smoothing parameters λ_1 and λ_2 control the weight to be given to the objective of making f_1 and f_2 smooth, relative to the objective of closely fitting the response data. For the moment, assume that these smoothing parameters are given.

Similarly to the single smooth case we have

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda_1 \mathbf{S}_1 + \lambda_2 \mathbf{S}_2\right)^{-1} \mathbf{X}^{\mathsf{T}} \mathbf{y} \text{ and } \mathbf{A} = \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda_1 \mathbf{S}_1 + \lambda_2 \mathbf{S}_2\right)^{-1} \mathbf{X}^{\mathsf{T}},$$

but again these expressions are sub-optimal with regard to computational stability and it is better to re-write the objective as

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_1 \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_1 \boldsymbol{\beta} + \lambda_2 \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_2 \boldsymbol{\beta} = \left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \mathbf{B} \end{bmatrix} \boldsymbol{\beta} \right\|^2, \tag{4.9}$$

where

$$\mathbf{B} = \begin{bmatrix} \mathbf{0} & \sqrt{\lambda_1}\mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sqrt{\lambda_2}\mathbf{D}_2 \end{bmatrix}$$

(or any other matrix such that $\mathbf{B}^{\mathsf{T}}\mathbf{B} = \lambda_1 \mathbf{S}_1 + \lambda_2 \mathbf{S}_2$).

As in the single smooth case, the right hand side of (4.9) is simply the unpenalized least squares objective for an augmented version of the model and corresponding response data. Hence, the model can be fitted by standard linear regression using stable orthogonal matrix based methods.

Here is a function to set up and fit a simple two term additive model, assuming the same number of knots for each smooth.

```
am.fit <- function(y,x,v,sp,k=10) {
  ## setup bases and penalties...
  xk <- seq(min(x),max(x),length=k)
  xdx <- tf.XD(x,xk)
  vk <- seq(min(v),max(v),length=k)
  xdv <- tf.XD(v,vk)
  ## create augmented model matrix and response...
  nD <- nrow(xdx$D)*2
  sp <- sqrt(sp)
  X <- cbind(c(rep(1,nrow(xdx$X)),rep(0,nD)),
             rbind(xdx$X,sp[1]*xdx$D,xdv$D*0),
             rbind(xdv$X,xdx$D*0,sp[2]*xdv$D))
  y1 <- c(y,rep(0,nD))
  ## fit model..
  b <- lm(y1 ~ X - 1)
  ## compute some useful quantities...
  n <- length(y)
  trA <- sum(influence(b)$hat[1:n]) ## EDF
  rsd <- y - fitted(b)[1:n] ## residuals
  rss <- sum(rsd^2)         ## residual SS
  sig.hat <- rss/(n-trA)       ## residual variance
  gcv <- sig.hat*n/(n-trA)     ## GCV score
  Vb <- vcov(b)*sig.hat/summary(b)$sigma^2 ## coeff cov matrix
  ## return fitted model...
  list(b=coef(b),Vb=Vb,edf=trA,gcv=gcv,fitted=fitted(b)[1:n],
       rsd=rsd,xk=list(xk,vk),cmx=list(xdx$cmx,xdv$cmx))
}
```

In addition to the quantities that we met in the single smooth case, `am.fit` also returns an estimate of the Bayesian covariance matrix for the model coefficients:

$$\hat{\mathbf{V}}_\beta = (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda_1\mathbf{S}_1 + \lambda_2\mathbf{S}_2)^{-1}\hat{\sigma}^2$$

where $\hat{\sigma}^2$ is taken as the residual sum of squares for the fitted model, divided by the effective residual degrees of freedom. Following section 4.2.4 the posterior distribution for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}|\mathbf{y} \sim N(\hat{\boldsymbol{\beta}}, \mathbf{V}_\beta), \tag{4.10}$$

and this result can be used for further inference about $\boldsymbol{\beta}$ (see section 6.10, p. 293).

Let us use the routine to estimate an additive model for the data in R data frame `trees`. The data are `Volume`, `Girth` and `Height` for 31 felled cherry trees. Interest lies in predicting `Volume`, and we can try estimating the model

$$\texttt{Volume}_i = \alpha + f_1(\texttt{Girth}_i) + f_2(\texttt{Height}_i) + \epsilon_i.$$

Now that we have two smoothing parameters, grid searching for the GCV optimal values starts to become inefficient. Instead R function `optim` can be used to minimize the GCV score. The function to be optimised has to be in a particular form for use with `optim`: the optimization parameter vector must be the first argument, and the function must be real valued. A simple wrapper for `am.fit` suffices:

```
am.gcv <- function(lsp,y,x,v,k) {
## function suitable for GCV optimization by optim
  am.fit(y,x,v,exp(lsp),k)$gcv
}
```

Using log smoothing parameters for optimization ensures that the estimated smoothing parameters are non-negative. Fitting the model is now straightforward

```
## find GCV optimal smoothing parameters...
fit <- optim(c(0,0), am.gcv, y=trees$Volume, x=trees$Girth,
             v=trees$Height,k=10)
sp <- exp(fit$par) ## best fit smoothing parameters
## Get fit at GCV optimal smoothing parameters...
fit <- am.fit(trees$Volume,trees$Girth,trees$Height,sp,k=10)
```

Now let's plot the smooth effects. The following function will do this.

```
am.plot <- function(fit,xlab,ylab) {
## produces effect plots for simple 2 term
## additive model
  start <- 2 ## where smooth coeffs start in beta
  for (i in 1:2) {
    ## sequence of values at which to predict...
    x <- seq(min(fit$xk[[i]]), max(fit$xk[[i]]), length=200)
    ## get prediction matrix for this smooth...
    Xp <- tf.XD(x, fit$xk[[i]], fit$cmx[[i]])$X
    ## extract coefficients and cov matrix for this smooth
    stop <- start + ncol(Xp)-1; ind <- start:stop
    b <- fit$b[ind];Vb <- fit$Vb[ind,ind]
    ## values for smooth at x...
    fv <- Xp %*% b
    ## standard errors of smooth at x....
    se <- rowSums((Xp %*% Vb) * Xp)^.5
    ## 2 s.e. limits for smooth...
    ul <- fv + 2 * se; ll <- fv - 2 * se
    ## plot smooth and limits...
    plot(x, fv, type="l", ylim=range(c(ul,ll)), xlab=xlab[i],
         ylab=ylab[i])
    lines(x, ul, lty=2); lines(x, ll, lty=2)
    start <- stop + 1
  }
}
```

Calling it with the fitted tree model

```
par(mfrow=c(1,3))
plot(fit$fitted,trees$Vol,xlab="fitted volume ",
     ylab="observed volume")
am.plot(fit,xlab=c("Girth","Height"),
        ylab=c("s(Girth)","s(Height)"))
```

gives the result in figure 4.11. Notice that the smooth of `Height` is estimated to be a straight line, and as a result its confidence interval has zero width at some point.

The zero width point in the interval occurs because the sum to zero constraint exactly determines where the straight line must pass through zero.

As with the one dimensional smooth, the additive model could also be estimated as a linear mixed model, but let us move on.

4.4 Generalized additive models

Generalized additive models (GAMs) follow from additive models, as generalized linear models follow from linear models. That is, the linear predictor now predicts some known smooth monotonic function of the expected value of the response, and the response may follow any exponential family distribution, or simply have a known mean variance relationship, permitting the use of a quasi-likelihood approach. The resulting model has a general form something like (4.1) in section 4.1.

As an illustration, suppose that we would like to model the `trees` data using a GAM of the form:

$$\log\{\mathbb{E}(\texttt{Volume}_i)\} = f_1(\texttt{Girth}_i) + f_2(\texttt{Height}_i), \quad \texttt{Volume}_i \sim \text{ gamma.}$$

This model is perhaps more natural than the additive model, as we might expect volume to be the product of some function of girth and some function of height, and it is reasonable to expect the variance in volume to increase with mean volume.

Whereas the additive model was estimated by penalized least squares, the GAM will be fitted by penalized likelihood maximization, and in practice this will be achieved by penalized iterative least squares (PIRLS).[§] For given smoothing parameters, the following steps are iterated to convergence.

1. Given the current linear predictor estimate, $\hat{\boldsymbol{\eta}}$, and corresponding estimated mean response vector, $\hat{\boldsymbol{\mu}}$, calculate:

$$w_i = \frac{1}{V(\hat{\mu}_i)g'(\hat{\mu}_i)^2} \quad \text{and} \quad z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i) + \hat{\eta}_i$$

   where $\text{var}(Y_i) = V(\mu_i)\phi$, as in section 3.1.2, and g is the link function.

2. Defining $\mathbf{W}$ as the diagonal matrix such that $W_{ii} = w_i$, minimize

$$\|\sqrt{\mathbf{W}}\mathbf{z} - \sqrt{\mathbf{W}}\mathbf{X}\boldsymbol{\beta})\|^2 + \lambda_1\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_1\boldsymbol{\beta} + \lambda_2\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_2\boldsymbol{\beta}$$

   w.r.t. $\boldsymbol{\beta}$ to obtain new estimate $\hat{\boldsymbol{\beta}}$, and hence updated estimates $\hat{\boldsymbol{\eta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$.

The penalized least squares problem at step 2 is exactly the problem already solved for the simple additive model. Note the link to section 3.4.1 (p. 148).

For the trees GAM, the link function, g, is the $\log$ function, so $g'(\mu_i) = \mu_i^{-1}$, while for the gamma distribution, $V(\mu_i) = \mu_i^2$ (see table 3.1, p. 104). Hence, for the log-link gamma model, we have:

$$w_i = 1 \quad \text{and} \quad z_i = (y_i - \hat{\mu}_i)/\hat{\mu}_i + \hat{\eta}_i.$$

[§] There is no simple trick to produce an unpenalized GLM whose likelihood is equivalent to the penalized likelihood of the GAM that we wish to fit.

So, given λ_1 and λ_2 it will be straightforward to obtain $\hat{\boldsymbol{\beta}}$, but what should be used as the GCV score for this model? A natural choice is to use the GCV score for the final linear model in the PIRLS iteration (although this choice is poor for binary data: see section 6.2, p. 255 for better performing alternatives). It is easy to show that this GCV score is equivalent to the usual GCV score, but with the Pearson statistic replacing the residual sum of squares. Obviously we could also estimate the smoothing parameters by exploiting the Bayesian/mixed model connection of section 4.2.4, and estimating the model as a generalized linear mixed model using the methods of section 3.4 (p. 147).

The following function implements the PIRLS loop for the log-gamma model, and returns the required GCV score in its return list.

```
gam.fit <- function(y,x,v,sp,k=10) {
## gamma error log link 2 term gam fit...
  eta <- log(y) ## get initial eta
  not.converged <- TRUE
  old.gcv <- -100 ## don't converge immediately
  while (not.converged) {
    mu <- exp(eta)  ## current mu estimate
    z <- (y - mu)/mu + eta ## pseudodata
    fit <- am.fit(z,x,v,sp,k) ## penalized least squares
    if (abs(fit$gcv-old.gcv)<1e-5*fit$gcv) {
      not.converged <- FALSE
    }
    old.gcv <- fit$gcv
    eta <- fit$fitted ## updated eta
  }
  fit$fitted <- exp(fit$fitted) ## mu
  fit
}
```

Again a simple wrapper is needed in order to optimize the GCV score using `optim`

```
gam.gcv <- function(lsp,y,x,v,k=10) {
  gam.fit(y,x,v,exp(lsp),k=k)$gcv
}
```

Now fitting and plotting proceeds exactly as in the simple additive case.

```
fit <- optim(c(0,0),gam.gcv,y=trees$Volume,x=trees$Girth,
             v=trees$Height,k=10)
sp <- exp(fit$par)
fit <- gam.fit(trees$Volume,trees$Girth,trees$Height,sp)
par(mfrow=c(1,3))
plot(fit$fitted,trees$Vol,xlab="fitted volume ",
     ylab="observed volume")
am.plot(fit,xlab=c("Girth","Height"),
        ylab=c("s(Girth)","s(Height)"))
```

The resulting plots are shown in figure 4.12.

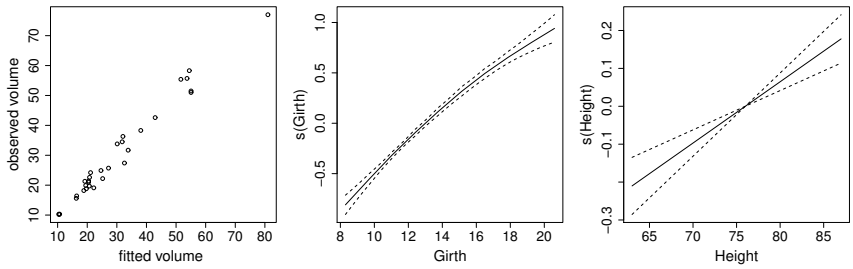


Figure 4.12 *The best fit two term generalized additive model for the* `tree` *data. The left panel shows actual versus predicted tree volumes. The middle panel is the estimate of the smooth function of girth. The right panel is the estimate of the smooth function of height.*

4.5 Summary

The preceding sections have illustrated how models based on smooth functions of predictor variables can be represented, and estimated, once a basis and wiggliness penalty have been chosen for each smooth in the model. Estimation is by penalized versions of the least squares and maximum-likelihood methods used for linear models and GLMs. Indeed technically GAMs are simply GLMs estimated subject to smoothing penalties. The most substantial difficulty introduced by this penalization is the need to select the degree of penalization, that is, to estimate the smoothing parameters. As we have seen, GCV provides one quite reasonable solution, and marginal likelihood provides an alternative.

The rest of this book will stick to the basic framework presented here, simply adding refinements to it. We will consider a variety of basis-penalty smoothers somewhat preferable to the piecewise linear basis given here, and some alternatives to GCV for smoothness estimation. More efficient and reliable computational methods will be developed, and the theoretical basis for inference will be more fully expounded. The link between smooths and random effects will also be developed, as will models based on linear functionals of smooths. However, throughout, functions are represented using penalized basis expansions, estimation of coefficients is by penalized likelihood maximisation and estimation of smoothing parameters uses a separate criterion, such as GCV or REML.

4.6 Introducing package `mgcv`

Before considering smoothers and GAM theory in more detail, it is worth briefly introducing the `mgcv` package. The `gam` function from `mgcv` is very much like the `glm` function covered in chapter 3. The main difference is that the `gam` model formula can include smooth terms, `s()` and `te()` (as well as the `te` variants `ti` and `t2`). Also there are a number of options available for controlling automatic smoothing parameter estimation, or for directly controlling model smoothness (summarized in table 4.1).

The cherry tree data provide a simple example with which to introduce the modelling functions available in R package mgcv.

```
library(mgcv)    ## load the package
data(trees)
ct1 <- gam(Volume ~ s(Height) + s(Girth),
           family=Gamma(link=log),data=trees)
```

This fits the generalized additive model

$$\log(\mathbb{E}[\texttt{Volume}_i]) = f_1(\texttt{Height}_i) + f_2(\texttt{Girth}_i) \quad \text{where } \texttt{Volume}_i \sim \text{gamma}$$

and the f_j are smooth functions. By default, the degree of smoothness of the f_j (within certain limits) is estimated by GCV. The results can be checked by typing the name of the fitted model object to invoke the `print.gam` print method, and by plotting the fitted model object. For example

```
> ct1

Family: Gamma
Link function: log

Formula:
Volume ~ s(Height) + s(Girth)

Estimated degrees of freedom:
1.00 2.42  total = 4.42

GCV score: 0.008082356
> plot(ct1,residuals=TRUE)
```

The resulting plot is displayed in the upper two panels of figure 4.13. Notice that the default print method reports the model distribution family, link function and formula, before displaying the effective degrees of freedom for each term (in the order that the terms appear in the model formula) and the whole model: in this case a nearly straight line, corresponding to about one degree of freedom, is estimated for the effect of height, while the effect of girth is estimated as a smooth curve with 2.4 degrees of freedom; the total degrees of freedom is the sum of these two, plus one degree of freedom for the model intercept. Finally, the GCV score for the fitted model is reported.

The plots show the estimated effects as solid lines/curves, with 95% confidence limits (strictly Bayesian credible intervals; see section 6.10, p. 293), based on (4.10), shown as dashed lines. The coincidence of the confidence limits and the estimated straight line, at the point where the line passes through zero on the vertical axis, is a result of the identifiability constraints applied to the smooth terms.[¶] The points shown on the plots are *partial residuals*. These are simply the Pearson residuals

[¶]The identifiability constraint is that the sum of the values of each curve, at the observed covariate values, must be zero: for a straight line, this condition exactly determines where the line must pass through zero, so there can be no uncertainty about this point.

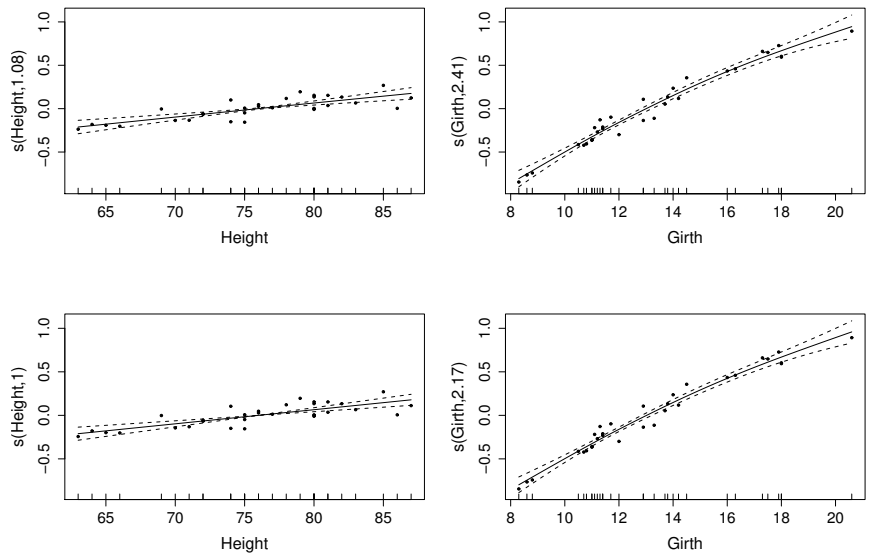


Figure 4.13 *Components of GAM model fits to the cherry tree data. The upper two panels are from* `ct1` *and the lower 2 from* `ct4`.

added to the smooth terms evaluated at the appropriate covariate values. For example, the residuals plotted in the top left panel of figure 4.13 are given by

$$\hat{\epsilon}_{1i}^{\text{partial}} = f_1(\texttt{Height}_\texttt{i}) + \hat{\epsilon}_i^p$$

plotted against $\texttt{Height}_i$. For a well fitting model the partial residuals should be evenly scattered around the curve to which they relate. The 'rug plots', along the bottom of each plot, show the values of the covariates of each smooth. The number in each y-axis caption is the effective degrees of freedom of the term being plotted.

4.6.1 Finer control of `gam`

The simple form of the `gam` call producing `ct1` hides a number of options that have been set to default values. The first of these is the choice of basis used to represent the smooth terms. The default is to use thin plate regression splines (section 5.5.1, p. 215), which have some appealing properties, but can be somewhat computationally costly for large data sets. The full range of available smoothers is covered in chapter 5. In the following, penalized cubic regression splines are selected using `s(...,bs="cr")`.

```
> ct2 <- gam(Volume ~ s(Height,bs="cr") + s(Girth,bs="cr"),
+            family=Gamma(link=log),data=trees)
> ct2
```

scale	The value of the scale parameter, or a negative value if it is to be estimated. For method="GCV.Cp" then scale > 0 implies Mallows' C_p/UBRE/AIC is used. scale $< 0 \Rightarrow$ implies GCV is used. scale $= 0 \Rightarrow$ UBRE/AIC for Poisson or binomial, otherwise GCV.
gamma	This multiplies the model degrees of freedom in the GCV or UBRE/AIC criteria. Hence as gamma is increased from 1 the 'penalty' per degree of freedom increases in the GCV or UBRE/AIC criterion and increasingly smooth models are produced. Increasing gamma to around 1.5 can usually reduce over-fitting, without much degradation in prediction error performance.
sp	An array of supplied smoothing parameters. When this array is non-null, a negative element signals that a smoothing parameter should be estimated, while a non-negative value is used as the smoothing parameter for the corresponding term. This is useful for directly controlling the smoothness of some terms.
method	Selects the smoothing parameter selection criterion: GCV.Cp, GACV, ML or REML.

Table 4.1 *Main arguments to* gam *for controlling the smoothness estimation process.*

```
Family: Gamma
Link function: log

Formula:
Volume ~ s(Height, bs = "cr") + s(Girth, bs = "cr")

Estimated degrees of freedom:
 1.000126 2.418591   total =  4.418718

GCV score:  0.008080546
```

As you can see, the change in basis has made very little difference to the fit. Plots are almost indistinguishable to those for ct1. This is re-assuring: it would be unfortunate if the model depended very strongly on details like the exact choice of basis. However, larger changes to the basis, such as using P-splines (section 5.3.3, p. 204), can make an appreciable difference.

Another choice, hidden in the previous two model fits, is the *dimension*, k, of the basis used to represent smooth terms. In the previous two fits, the (arbitrary) default, $k = 10$, was used. The choice of basis dimensions amounts to setting the *maximum* possible degrees of freedom allowed for each model term. The actual effective degrees of freedom for each term will usually be estimated from the data, by GCV or another smoothness selection criterion, but the upper limit on this estimate is $k - 1$: the basis dimension minus one degree of freedom due to the identifiability constraint on each smooth term. The following example sets k to 20 for the smooth of Girth (and illustrates, by the way, that there is no problem in mixing different bases).

```
> ct3 <- gam(Volume ~ s(Height) + s(Girth,bs="cr",k=20),
+            family=Gamma(link=log),data=trees)
> ct3

Family: Gamma
Link function: log

Formula:
Volume ~ s(Height) + s(Girth, bs = "cr", k = 20)

Estimated degrees of freedom:
 1.000003 2.424226   total =  4.424229

GCV score:  0.00808297
```

Again, this change makes boringly little difference in this case, and the plots (not shown) are indistinguishable from those for `ct1`. This insensitivity to basis dimension is not universal, of course, and checking of this choice is covered in section 5.9 (p. 242). One quite subtle point is worth being aware of. This is that a space of functions of dimension 20 will contain a larger subspace of functions with effective degrees of freedom 5, than will a function space of dimension 10 (the particular numbers being arbitrary here). Hence it is often the case that increasing k will change the effective degrees of freedom estimated for a term, even though both old and new estimated degrees of freedom are lower than the original $k - 1$.

Another choice is the parameter `gamma` which can be used to multiply the model effective degrees of freedom in the GCV or UBRE scores in order to (usually) increase the amount of smoothing selected. The default value is 1, but GCV is known to have some tendency to overfitting on occasion, and it has been suggested that using $\gamma \approx 1.5$ can somewhat correct this without compromising model fit (e.g., Kim and Gu, 2004). See section 6.2.4 for one justification. Applying this idea to the current model results in the bottom row of figure 4.13 and the following output.

```
> ct4 <- gam(Volume ~ s(Height) + s(Girth),
+            family=Gamma(link=log),data=trees,gamma=1.4)
> ct4

Family: Gamma
Link function: log

Formula:
Volume ~ s(Height) + s(Girth)

Estimated degrees of freedom:
 1.00011 2.169248   total =  4.169358

GCV score:  0.00922805

> plot(ct4,residuals=TRUE)
```

The heavier penalty on each degree of freedom in the GCV score has resulted in

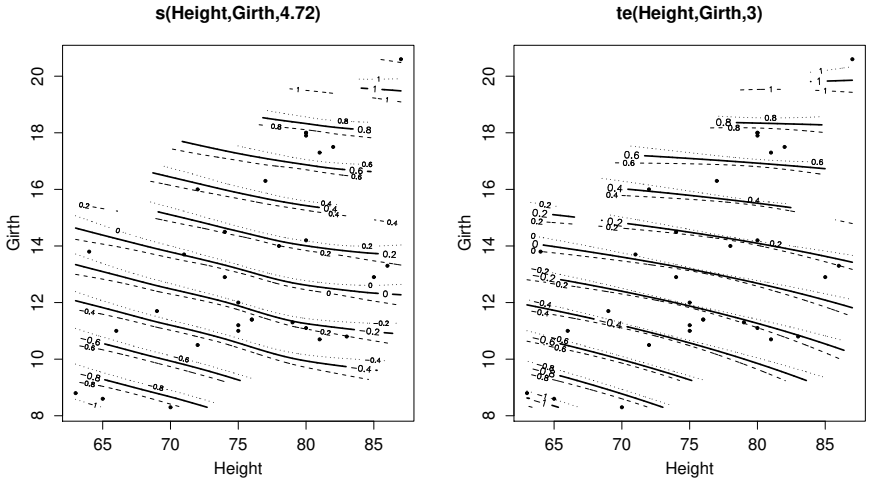


Figure 4.14 *Smooth functions of height and girth fitted to the cherry tree data, with degree of smoothing chosen by GCV. The left hand panel shows a thin plate regression spline fit (*`ct5`*), while the right panel shows a tensor product spline fit (*`ct6`*). For both plots the bold contours show the estimate of the smooth; the dashed contours show the smooth plus the standard error of the smooth and the dotted contours show the smooth less its standard error. The symbols show the locations of the covariate values on the height–girth plane. Parts of the smooths that are far away from covariate values have been excluded from the plots using the* `too.far` *argument to* `plot.gam`*.*

a model with fewer degrees of freedom, but the figure indicates that the change in estimates that this produces is barely perceptible.

4.6.2 Smooths of several variables

`gam` is not restricted to models containing only smooths of one predictor. In principle, smooths of any number of predictors are possible via two types of smooth. Within a model formula, `s()` terms, using the `"tp"`, `"ds"` or `"gp"` bases,[‖] produce isotropic smooths of multiple predictors, while `te()` terms produce smooths of multiple predictors from tensor products of *any* singly penalized bases available for use with `s()` (including mixtures of different bases). The tensor product smooths are invariant to linear rescaling of covariates, and can be quite computationally efficient. Alternative versions `t2()` and `ti()` are available for different sorts of functional ANOVA decomposition. Section 5.7 (p. 237) compares isotropic and tensor product smoothers.

[‖] Or indeed `"sos"` or `"so"` bases.

By way of illustration, the following code fragments both fit the model

$$\log(\mathbb{E}[\texttt{Volume}_i]) = f(\texttt{Height}_i, \texttt{Girth}_i) \text{ where } \texttt{Volume}_i \sim \text{gamma},$$

and f is a smooth function. Firstly an isotropic thin plate regression spline is used:

```
> ct5 <- gam(Volume ~ s(Height,Girth,k=25),
+            family=Gamma(link=log),data=trees)
> ct5

Family: Gamma
Link function: log

Formula:
Volume ~ s(Height, Girth, k = 25)

Estimated degrees of freedom:
 4.668129   total =  5.668129

GCV score:  0.009358786

> plot(ct5,too.far=0.15)
```

yielding the left hand panel of figure 4.14. Secondly a tensor product smooth is used. Note that the `k` argument to `te` specifies the dimension for each marginal basis: if different dimensions are required for the marginal bases then `k` can also be supplied as an array. The basis dimension of the tensor product smooth is the product of the dimensions of the marginal bases.

```
> ct6 <- gam(Volume ~ te(Height,Girth,k=5),
+            family=Gamma(link=log),data=trees)
> ct6

Family: Gamma
Link function: log

Formula:
Volume ~ te(Height, Girth, k = 5)

Estimated degrees of freedom:
 3.000175   total =  4.000175

GCV score:  0.008197151

> plot(ct6,too.far=0.15)
```

Notice how the tensor product model has fewer degrees of freedom and a lower GCV score than the TPRS smooth. In fact, with just 3 degrees of freedom, the tensor product smooth model amounts to

$$\log(\mathbb{E}[\texttt{Volume}_i]) = \beta_0 + \beta_1\texttt{Height}_i + \beta_2\texttt{Girth}_i + \beta_3\texttt{Height}_i\texttt{Girth}_i,$$

the 'wiggly' components of the model having been penalized away altogether.

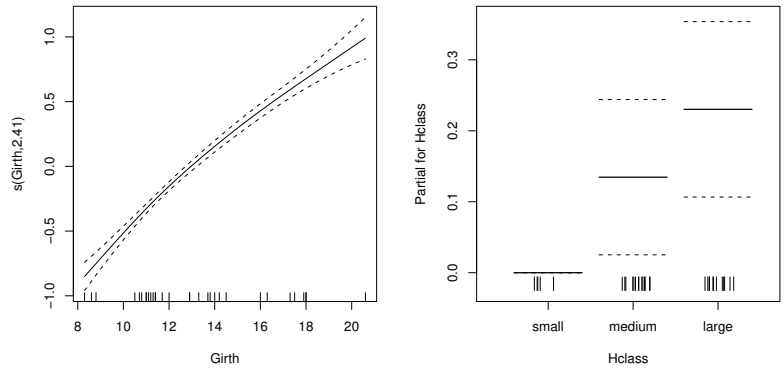


Figure 4.15 *Plot of model* `ct7`*, a* semi-parametric *model of cherry tree volume, with a factor for height and a smooth term for the dependence on girth. The left plot shows the smooth of girth, with 95% confidence interval, while the right panel shows the estimated effect, for each level of factor* `Hclass`*. The effect of being in the* `small` *height class is shown as zero, because the default contrasts have been used here, which set the parameter for the first level of each factor to zero.*

4.6.3 Parametric model terms

So far, only models consisting of smooth terms have been considered, but there is no difficulty in mixing smooth and parametric model components. For example, given that the model `ct1` smooth of height is estimated to be a straight line, we might as well fit the model:

```
gam(Volume ~ Height+s(Girth),family=Gamma(link=log),data=trees)
```

but to make the example more informative, let us instead suppose that the `Height` is actually only measured as a categorical variable. This can easily be arranged, by creating a factor variable which simply labels each tree as small, medium or large:

```
trees$Hclass <- factor(floor(trees$Height/10)-5,
                labels=c("small","medium","large"))
```

Now we can fit a generalized additive model to these data, using the `Hclass` variable as a factor variable, and plot the result (figure 4.15).

```
ct7 <- gam(Volume ~ Hclass + s(Girth),
           family=Gamma(link=log), data=trees)
par(mfrow=c(1,2)); plot(ct7,all.terms=TRUE)
```

Often, more information about a fitted model is required than is supplied by plots or the default print method, and various utility functions exist to provide this. For example the `anova` function can be used to investigate the approximate significance of model terms.

```
> anova(ct7)

Family: Gamma
```

```
Link function: log

Formula:
Volume ~ Hclass + s(Girth)

Parametric Terms:
       df     F p-value
Hclass  2 7.076 0.00358

Approximate significance of smooth terms:
           edf Est.rank     F  p-value
s(Girth) 2.414    9.000 54.43 1.98e-14
```

Clearly there is quite strong evidence that both height and girth matter (see section 6.12, for information on the p-value calculations for the smooth terms). Similarly, an approximate AIC value can be obtained for the model (see section 6.11, p. 301):

```
> AIC(ct7)
[1] 154.9411
```

The `summary` method provides considerable detail.

```
> summary(ct7)

Family: Gamma
Link function: log

Formula:
Volume ~ Hclass + s(Girth)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.12693    0.04814  64.949  < 2e-16 ***
Hclassmedium  0.13459    0.05428   2.479 0.020085 *
Hclasslarge   0.23024    0.06137   3.752 0.000908 ***

Approximate significance of smooth terms:
           edf Est.rank     F  p-value
s(Girth) 2.414    9.000 54.43 1.98e-14 ***

R-sq.(adj) =  0.967   Deviance explained = 96.9%
GCV score = 0.012076   Scale est. = 0.0099671  n = 31
```

Notice that, in this case, the significance of individual parameters of the parametric terms is given, rather than whole term significance. Other measures of fit are also reported, such as the adjusted r^2 and percentage deviance explained, along with the GCV score, an estimate of the scale parameter of the model, and the number of data fitted.

4.6.4 The mgcv *help pages*

mgcv has quite extensive help pages, both documenting functions and attempting to provide overviews of a topic. The easiest way to access the pages is via the HTML versions, by typing `help.start()` in R, then navigating to the mgcv pages and browsing. Several pages are well worth knowing about:

- `mgcv-package` offers an overview of the package and what it offers.
- `family.mgcv` gives an overview of the distributions available.
- `smooth.terms` gives an overview of the smooths types available.
- `random.effects` is an overview of random effects in mgcv.
- `gam.models` reviews some aspects of model specification; `gam.selection` covers model selection options.
- `gam`, `bam`, `gamm` and `jagam` cover the main modelling functions.

4.7 Exercises

1. This question is about illustrating the problems with polynomial bases. First run

   ```
   set.seed(1)
   x<-sort(runif(40)*10)^.5
   y<-sort(runif(40))^0.1
   ```

   to simulate some apparently innocuous x, y data.

   (a) Fit 5^{th} and 10^{th} order polynomials to the simulated data using, e.g., `lm(y~poly(x,5))`.

   (b) Plot the x, y data, and overlay the fitted polynomials. (Use the `predict` function to obtain predictions on a fine grid over the range of the x data: only predicting at the data fails to illustrate the polynomial behaviour adequately).

   (c) One particularly simple basis for a cubic regression spline is $b_1(x) = 1$, $b_2(x) = x$ and $b_{j+2}(x) = |x - x_j^*|^3$ for $j = 1 \ldots q-2$, where q is the basis dimension, and the x_j^* are knot locations. Use this basis to fit a rank 11 cubic regression spline to the x, y data (using `lm` and evenly spaced knots).

   (d) Overlay the predicted curve according to the spline model, onto the existing x, y plot, and consider which basis you would rather use.

2. Polynomial models of the data from question 1 can also provide an illustration of why orthogonal matrix methods are preferable to fitting models by solution of the 'normal equations' $\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\mathsf{T}\mathbf{y}$. The bases produced by `poly` are actually orthogonal polynomial bases, which are a numerically stable way of representing polynomial models, but if a naïve basis is used then a numerically badly behaved model can be created.

   ```
   form<-paste("I(x^",1:10,")",sep="",collapse="+")
   form <- as.formula(paste("y~",form))
   ```

   produces the model formula for a suitably ill-behaved model. Fit this model using `lm`, extract the model matrix from the fitted model object using `model.matrix`, and re-estimate the model parameters by solving the 'normal equations' given

above (see `?solve`). Compare the estimated coefficients in both cases, along with the fits. It is also instructive to increase the order of the polynomial by one or two and examine the results (and to decrease it to 5, say, in order to confirm that the QR and normal equations approaches agree if everything is 'well behaved'). Finally, note that the singular value decomposition (see B.10) provides a reliable way of diagnosing the linear dependencies that can cause problems when model fitting. `svd(X)$d` obtains the singular values of a matrix `X`. The largest divided by the smallest gives the 'condition number' of the matrix — a measure of how ill-conditioned computations with the matrix are likely to be.

3. Show that the $\boldsymbol{\beta}$ minimizing (4.6), in section 4.2.2, is given by (4.7).
4. Let $\mathbf{X}$ be an $n \times p$ model matrix, $\mathbf{S}$ a $p \times p$ penalty matrix, and $\mathbf{B}$ any matrix such that $\mathbf{B}^{\mathsf{T}}\mathbf{B} = \mathbf{S}$. If
$$\tilde{\mathbf{X}} = \left[\begin{array}{c} \mathbf{X} \\ \mathbf{B} \end{array} \right]$$
is an augmented model matrix, show that the sum of the first n elements on the leading diagonal of $\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^{\mathsf{T}}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\mathsf{T}}$ is $\mathrm{tr}\left(\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X}+\mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}\right)$.
5. The 'obvious' way to estimate smoothing parameters is by treating smoothing parameters just like the other model parameters, $\boldsymbol{\beta}$, and to choose λ to minimize the residual sum of squares for the fitted model. What estimate of λ will such an approach always produce?
6. Show that for any function f, which has a basis expansion
$$f(x) = \sum_j \beta_j b_j(x),$$
it is possible to write
$$\int f''(x)^2 dx = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}\boldsymbol{\beta},$$
where the coefficient matrix $\mathbf{S}$ can be expressed in terms of the known basis functions b_j (assuming that these possess at least two (integrable) derivatives). As usual $\boldsymbol{\beta}$ is a parameter vector with β_j in its j^{th} element.
7. Show that for any function f which has a basis expansion
$$f(x,z) = \sum_j \beta_j b_j(x,z),$$
it is possible to write
$$\int \left(\frac{\partial^2 f}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x \partial z}\right)^2 + \left(\frac{\partial f^2}{\partial z^2}\right)^2 dxdz = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}\boldsymbol{\beta},$$
where the coefficient matrix $\mathbf{S}$ can be expressed in terms of the known basis functions b_j (assuming that these possess at least two (integrable) derivatives w.r.t. x and z). Again, $\boldsymbol{\beta}$ is a parameter vector with β_j in its j^{th} element.
8. The additive model of section 4.3 can equally well be estimated as a mixed model.

(a) Write a function which converts the model matrix and penalty returned by `tf.XD` into mixed model form. Hint: because of the constraints the penalty null space is of dimension 1 now, leading to a slight modification of $\mathbf{D}_+$.

(b) Using your function from part (a) obtain the model matrices required to fit the two term additive tree model, and estimate it using `lme`. Because there are now two smooths, two `pdIdent` terms will be needed in the `random` list supplied to `lme`, which will involve two dummy grouping variables (which can just be differently named copies of the same variable).

(c) Produce residual versus fitted volume and raw volume against fitted volume plots.

(d) Produce plots of the two smooth effect estimates with partial residuals.

9. Following on from question 8, we can also estimate a GAM as a GLMM. This is particularly easy to implement using the PQL method of section 3.4.2 (p. 149).

(a) Modify the function `gam.fit` from section 4.4, so that in place of the call to `am.fit` there is an appropriate call to `lme` to estimate the coefficients and smoothing parameters of a working linear mixed model. The modified function should take a response vector and the model matrices from the previous question as inputs, and return the `lme` fitted model object for the working model at convergence.

(b) Use your function to fit the Gamma additive model of section 4.4 to the `trees` data.

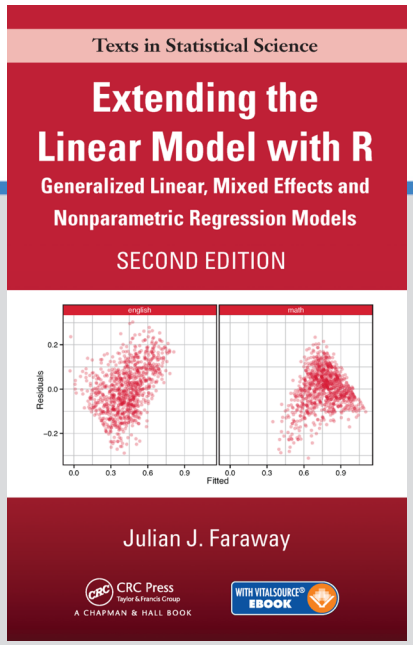
(c) Produce plots of measured volume against predicted volume, and of residuals against the linear predictor of the model.

(d) Plot the smooth effects with partial residuals.

CHAPTER

4

RANDOM EFFECTS



This chapter is excerpted from

Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second Edition

by Julian J. Faraway

Learn more

Chapter 10

Random Effects

Grouped data arise in almost all areas of statistical application. Sometimes the grouping structure is simple, where each case belongs to a single group and there is only one grouping factor. More complex datasets have a hierarchical or nested structure or include longitudinal or spatial elements. Sometimes the grouping arises because the same individual is measured repeatedly or sometimes each individual is measured once only but these individuals have some form of group structure. We defer examination of the repeated measurement of individuals to the next chapter, although the statistical methodology used is the same.

All such data share the common feature of correlation of observations within the same group and so analyses that assume independence of the observations will be inappropriate. The use of random effects is one common and convenient way to model such grouping structure.

A *fixed effect* is an unknown constant that we try to estimate from the data. Almost all the parameters used in the linear and generalized linear models we have presented earlier in this book are fixed effects. In contrast, a *random effect* is a random variable. It does not make sense to estimate a random effect; instead, we try to estimate the parameters that describe the distribution of this random effect.

Consider an experiment to investigate the effect of several drug treatments on a sample of patients. Typically, we are interested in specific drug treatments and so we would treat the drug effects as fixed. However, it makes most sense to treat the patient effects as random. It is often reasonable to treat the patients as being randomly selected from a larger collection of patients whose characteristics we would like to estimate. Furthermore, we are not particularly interested in these specific patients, but in the whole population of patients. A random effects approach to modeling effects is more ambitious in the sense that it attempts to say something about the wider population beyond the particular sample. Blocking factors can often be viewed as random effects, because these often arise as a random sample of those blocks potentially available.

There is some judgment required in deciding when to use fixed and when to use random effects. Sometimes the choice is clear, but in other cases, reasonable statisticians may differ. In some analyses, random effects are used simply to induce a certain correlation structure in the data and there is sense in which the chosen levels represent a sample from a population. Gelman (2005) remarks on the variety of definitions for random effects and proposes a particular straightforward solution to the dilemma of whether to use fixed or random effects — he recommends always using random effects.

A *mixed effects* model has both fixed and random effects. A simple example of such a model would be a two-way analysis of variance (ANOVA):

$$y_{ijk} = \mu + \tau_i + v_j + \varepsilon_{ijk}$$

where the μ and τ_i are fixed effects and the error ε_{ijk} and the random effects v_j are independent and identically distributed $N(0,\sigma^2)$ and $N(0,\sigma_v^2)$, respectively.

We would want to estimate the τ_i and test the hypothesis $H_0 : \tau_i = 0\ \forall i$ while we would estimate σ_v^2 and might test $H_0 : \sigma_v^2 = 0$. Notice the difference: we need to estimate and test several fixed effect parameters while we need only estimate and test a single random effect parameter.

In the following sections, we consider estimation and inference for mixed effects models and then illustrate the application to several common designs.

10.1 Estimation

This is not as simple as it was for fixed effects models, where least squares is an easily applied method with many good properties. Let's start with the simplest possible random effects model: a one-way ANOVA design with a factor at a levels:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad i = 1,\ldots,a \qquad j = 1,\ldots,n$$

where the αs and εs have mean zero, but variances σ_α^2 and σ_ε^2, respectively. These variances are known as the variance components. Notice that this induces a correlation between observations at the same level equal to:

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}$$

This is known as the *intraclass correlation coefficient* (ICC). In the limiting case when there is no variation between the levels, $\sigma_\alpha = 0$ so then $\rho = 0$. Alternatively, when the variation between the levels is much larger than that within the levels, the value of ρ will approach 1. This illustrates how random effects generate correlations between observations.

For simplicity, let there be an equal number of observations n per level. We can decompose the variation as follows (where dot in the subscript indicates the average over that index):

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

or $SST = SSE + SSA$, respectively. SSE is the residual sum of squares, SST is the total sum of squares (corrected for the mean) and SSA is the sum of squares due to α. These quantities are often displayed in an ANOVA table along with the degrees of freedom associated with each sum of squares. Dividing through by the respective degrees of freedom, we obtain the mean squares, MSE and MSA. Now we find that:

$$E(SSE) = a(n-1)\sigma_\varepsilon^2, \qquad E(SSA) = (a-1)(n\sigma_\alpha^2 + \sigma_\varepsilon^2)$$

which suggests using the estimators:

$$\hat{\sigma}^2_\varepsilon = SSE/(a(n-1)) = MSE, \qquad \hat{\sigma}^2_\alpha = \frac{SSA/(a-1) - \hat{\sigma}^2_\varepsilon}{n} = \frac{MSA - MSE}{n}$$

This method of estimating variance components can be used for more complex designs. The ANOVA table is constructed, the expected mean squares calculated and the variance components obtained by solving the resulting equations. These estimators are known as *ANOVA estimators*. These were the first estimators developed for variance components. They have the advantage of taking explicit forms suitable for hand calculation which was important in precomputing days, but they have a number of disadvantages:

1. The estimates can take negative values. For example, in our situation above, if $MSA < MSE$ then $\hat{\sigma}^2_\alpha < 0$. This is rather embarrassing since variances cannot be negative. Various fixes have been proposed, but these all take away from the original simplicity of the estimation method.
2. A balanced design has an equal number of observations per cell, where cell is defined as the finest subdivision of the data according to the factors. In such circumstances, the ANOVA decomposition into sums of squares is unique. For unbalanced data, this is not true and we must choose which ANOVA decomposition to use, which will in turn affect the estimation of the variance components. Various rules have been suggested about how the decomposition should be done, but none of these have universal appeal.
3. The need for complicated algebraic calculations. Formulae for the simpler models are easy to find and coded in software. More complex models will require difficult and opaque constructions.

We would like a method that would avoid negative variances, work unambiguously for unbalanced data and that can be applied in a transparent and straightforward manner. Maximum likelihood (ML) estimation satisfies these requirements. This does require that we assume some distribution for the errors and the random effects. The usual assumption is normality; ML would work for other distributions, but these are rarely considered in this context.

For a fixed effect model with normal errors, we can write:

$$y = X\beta + \varepsilon \qquad \text{or} \qquad y \sim N(X\beta, \sigma^2 I)$$

where X is an $n \times p$ model matrix and β is a vector of length p. We can generalize to a mixed effect model with a vector γ of q random effects with associated model matrix Z which has dimension $n \times q$. Then we can model the response y, given the value of the random effects as:

$$y = X\beta + Z\gamma + \varepsilon \qquad \text{or} \qquad y|\gamma \sim N(X\beta + Z\gamma, \sigma^2 I)$$

If we further assume that the random effects $\gamma \sim N(0, \sigma^2 D)$ then $\text{var}\, y = \text{var}\, Z\gamma + \text{var}\, \varepsilon = \sigma^2 ZDZ^T + \sigma^2 I$ and we can write the unconditional distribution of y as:

$$y \sim N(X\beta, \sigma^2(I + ZDZ^T))$$

If we knew D, then we could estimate β using generalized least squares; see, for example, Chapter 8 in Faraway (2014). However, the estimation of the variance components, D, is often one purpose of the analysis. Standard maximum likelihood is one method of estimation that can be used here. If we let $V = I + ZDZ^T$, then the joint density for the response is:

$$\frac{1}{2\pi^{n/2}|\sigma^2 V|^{1/2}} e^{-\frac{1}{2\sigma^2}(y-X\beta)^T V^{-1}(y-X\beta)}$$

so that the log-likelihood for the data is:

$$l(\beta,\sigma,D|y) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\sigma^2 V| - \frac{1}{2\sigma^2}(y-X\beta)^T V^{-1}(y-X\beta)$$

This can be optimized to find maximum likelihood estimates of β, σ^2 and D. This is straightforward in principle, but there may be difficulties in practice. More complex models involving larger numbers of random effects parameters can be difficult to estimate. Sometimes the MLE of a variance parameter may be zero which occurs on the boundary of its domain. The derivative of the likelihood may not be zero in this boundary state which causes problems for many optimization methods.

Standard errors can be obtained using the usual large sample theory for maximum likelihood estimates. The variance can be estimated using the inverse of the negative second derivative of the log-likelihood evaluated at the MLE.

MLEs have some drawbacks. One particular problem is that they are biased. For example, consider an i.i.d. sample of normal data $x_1, \ldots, x_n$, then the MLE is:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

A denominator of $n-1$ is needed for an unbiased estimator. Similar problems occur with the estimation of variance components. Given that the number of levels of a factor may not be large, the bias of the MLE of the variance component associated with that factor may be quite large. *Restricted maximum likelihood* (REML) estimators are an attempt to get round this problem. The idea is to find all independent linear combinations of the response, k, such that $k^T X = 0$. Form matrix K with columns k, so that:

$$K^T y \sim N(0, K^T V K)$$

We can then proceed to maximize the likelihood based on $K^T y$ which does not involve any of the fixed effect parameters. Once the random effect parameters have been estimated, it is simple enough to obtain the fixed effect parameter estimates. REML generally produces less biased estimates. For balanced data, the REML estimates are usually the same as the ANOVA estimates.

We illustrate the fitting methods using some data from an experiment to test the paper brightness depending on a shift operator described in Sheldon (1960). We start with a fixed effects one-way ANOVA:

```
data(pulp, package="faraway")
op <- options(contrasts=c("contr.sum", "contr.poly"))
lmod <- aov(bright ~ operator, pulp)
summary(lmod)
```

```
           Df Sum Sq Mean Sq F value Pr(>F)
operator    3  1.340   0.447     4.2  0.023
Residuals  16  1.700   0.106
```

We can plot the data as seen in Figure 10.1.

```
library(ggplot2)
ggplot(pulp, aes(x=operator, y=bright))+geom_point(position = position
    ↪ _jitter(width=0.1, height=0.0))
```

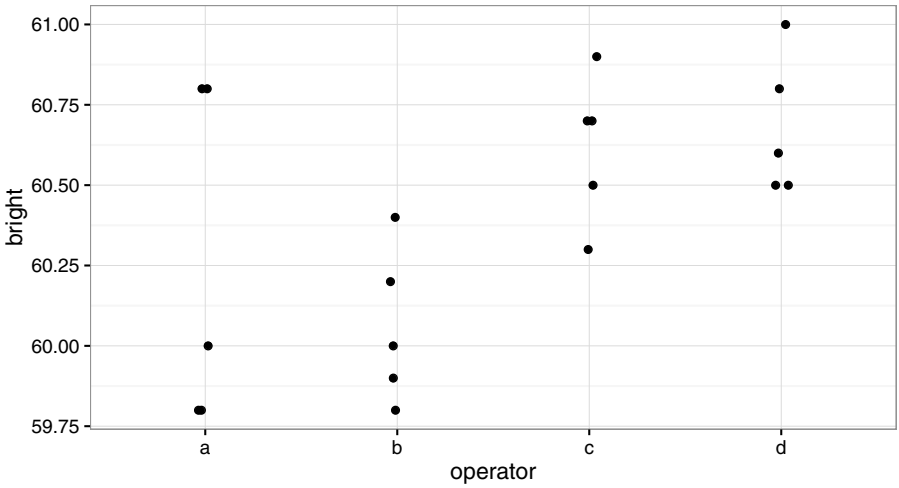


Figure 10.1 *Paper brightness varying by operator. Some jittering has been used to make coincident points apparent.*

```
coef(lmod)
```

```
(Intercept)   operator1   operator2   operator3
      60.40       -0.16       -0.34        0.22
```

```
options(op)
```

We have specified sum contrasts here instead of the default treatment contrasts to make the later connection to the corresponding random effects clearer. The `aov` function is just a wrapper for the standard `lm` function that produces results more appropriate for ANOVA models. We see that the operator effect is significant with a p-value of 0.023. The estimate of σ^2 is 0.106 and the estimated overall mean is 60.4. For sum contrasts, $\sum \alpha_i = 0$, so we can calculate the effect for the fourth operator as $0.16 + 0.34 - 0.22 = 0.28$.

Turning to the random effects model, we can compute the variance of the operator effects, σ^2_α, using the formula above as:

```
(0.447-0.106)/5
```

```
[1] 0.0682
```

Now we demonstrate the maximum likelihood estimators. The original R package for fitting mixed effects models was `nlme` as described in Pinheiro and Bates (2000). Subsequently Bates (2005) introduced the package `lme4`. The syntax for these two packages is somewhat different. The `lme4` package is generally more capable especially for larger datasets. The estimates these two packages produce for smaller, simpler datasets as considered in this chapter will generally be the same. However,

there are some crucial differences in the approach to inference that we will discuss later. We use the `lme4` packages in preference to `nlme`:

```
library(lme4)
mmod <- lmer(bright ~ 1+(1|operator), pulp)
summary(mmod)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: bright ~ 1 + (1 | operator)
   Data: pulp

REML criterion at convergence: 18.6

Scaled residuals:
   Min     1Q Median     3Q    Max
-1.467 -0.759 -0.124  0.628  1.601

Random effects:
 Groups   Name        Variance Std.Dev.
 operator (Intercept) 0.0681   0.261
 Residual             0.1062   0.326
Number of obs: 20, groups:  operator, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept)   60.400      0.149     404
```

The model has fixed and random effects components. The fixed effect here is just the intercept represented by the first `1` in the model formula. The random effect is represented by `(1|operator)` indicating that the data is grouped by `operator` and the `1` indicating that the random effect is constant within each group. The parentheses are necessary to ensure that expression is parsed in the correct order.

The default fitting method is REML. We see that this gives identical estimates to the ANOVA method above — $\hat{\sigma}^2 = 0.106$, $\hat{\sigma}^2_\alpha = 0.068$ and $\hat{\mu} = 60.4$. For unbalanced designs, the REML and ANOVA estimators are not necessarily identical. The standard deviations are simply the square roots of the variances and not estimates of the uncertainty in the variances.

As with the GLM summary output, we find it rather verbose and prefer our own abbreviated version (which is adapted from the `display()` function in the `arm` package of Gelman and Su (2013)):

```
sumary(mmod)
```

```
Fixed Effects:
coef.est  coef.se
   60.40     0.15

Random Effects:
 Groups   Name        Std.Dev.
 operator (Intercept) 0.26
 Residual             0.33
---
number of obs: 20, groups: operator, 4
AIC = 24.6, DIC = 14.4
deviance = 16.5
```

This output contains just the information we need. It is better to use standard devi-

ations rather than variances as the former are measured in the units of the response and so much easier to interpret.

The maximum likelihood estimates may also be computed:

```
smod <- lmer(bright ~ 1+(1|operator), pulp, REML=FALSE)
sumary(smod)
Fixed Effects:
coef.est  coef.se
   60.40     0.13

Random Effects:
 Groups   Name        Std.Dev.
 operator (Intercept) 0.21
 Residual             0.33
---
number of obs: 20, groups: operator, 4
AIC = 22.5, DIC = 16.5
deviance = 16.5
```

The between-subjects SD, 0.21, is smaller than with the REML method as the ML method biases the estimates towards zero. The fixed effects are unchanged.

10.2 Inference

Test Statistic: We follow a general procedure. Decide which component(s) of the model you wish to test. These can be fixed and/or random effects. Specify two models: a null H_0 which does not contain your specified component(s) and an alternative H_1 which does include your component(s). The other terms in the models must be the same. These other terms (usually) make a difference to the result and must be chosen with care.

Using standard likelihood theory, we may derive a test to compare two nested hypotheses, H_0 and H_1, by computing the likelihood ratio test statistic:

$$2(l(\hat{\beta}_1, \hat{\sigma}_1, \hat{D}_1|y) - l(\hat{\beta}_0, \hat{\sigma}_0, \hat{D}_0|y))$$

where $\hat{\beta}_0, \hat{\sigma}_0, \hat{D}_0$ are the MLEs of the parameters under the null hypothesis and $\hat{\beta}_1, \hat{\sigma}_1, \hat{D}_1$ are the MLEs of the parameters under the alternative hypothesis.

If you plan to use the likelihood ratio test to compare two nested models that differ only in their fixed effects, you cannot use the REML estimation method. The reason is that REML estimates the random effects by considering linear combinations of the data that remove the fixed effects. If these fixed effects are changed, the likelihoods of the two models will not be directly comparable. Use ordinary maximum likelihood in this situation if you also wish to use the likelihood ratio test.

Approximate Null Distribution: This test statistic is approximately chi-squared with degrees of freedom equal to the difference in the dimensions of the two parameters spaces (the difference in the number of parameters when the models are identifiable). Unfortunately, this test is not exact and also requires several assumptions — see a text such as Cox and Hinkley (1974) for more details. Serious problems can arise with this approximation.

One crucial assumption is that the parameters under the null are not on the boundary of the parameter space. Since we are often interested in testing hypotheses about

the random effects that take the form $H_0 : \hat{\sigma}^2 = 0$, this is a common problem which makes the asymptotic inference invalid. If you do use the χ^2 distribution with the usual degrees of freedom, then the test will tend to be conservative — the p-values will tend to be larger than they should be. This means that if you observe a significant effect using the χ^2 approximation, you can be fairly confident that it is actually significant. The p-values generated by the likelihood ratio test for fixed effects are also approximate and unfortunately tend to be too small, thereby sometimes overstating the importance of some effects.

Regrettably the p-value based on the χ^2 approximation can either be entirely or just somewhat wrong. Perhaps with sufficient data and favorable models, the approximation may be satisfactory but it is difficult to say exactly when such propitious conditions may arise. Hence the safest advice is to not use this approximation.

Expected mean squares: Another method of hypothesis testing is based on the sums of squares found in the ANOVA decompositions. These tests are sometimes more powerful than their likelihood ratio test equivalents. However, the correct derivation of these tests usually requires extensive tedious algebra that must be recalculated for each type of model. Furthermore, the tests cannot be used (at least without complex and unsatisfactory adjustments) when the experiment is unbalanced. This method only works for simple models and balanced data.

***F*-tests for fixed effects:** We might try to use the F-test used in standard linear models to perform hypothesis tests regarding the fixed effects. The F-statistic is based on residual sums of squares and degrees of freedom as described in Chapter 3 of Faraway (2014). This is the method used in the `nlme` package. In the standard linear model setting, provided the normality assumption is correct, the null distribution has an exact F-distribution. Unfortunately, problems arise in transferring this method to mixed effect models. Firstly, the definition of degrees of freedom becomes murky in the presence of random effect parameters. Secondly, the test statistic is not necessarily F-distributed.

For some simple models with balanced data, the F-test is correct but in other cases with more complex models or unbalanced data, the p-values can be substantially incorrect. It is difficult to specify exactly when this test may be relied upon. For this reason, the `lme4` now declines to state p-values. Furthermore, the t-statistics that one might generate to test or form a confidence interval for a single fixed effect parameter also rely on the same problematic approximations.

Strategies for inference: We have good test statistics in the likelihood ratio test (LRT) or F-statistic but as yet no universally reliable way to obtain a null distribution. One solution would be to ignore the possible problem and use either the `nlme` package or the `lmerTest` package (which restores the questionable p-values to `lme4`). In certain known simple models with balanced data, this will produce accurate results but it would be speculative to report such results in other situations without at least verifying the results using other methods. A number of alternatives exist.

The standard degrees of freedom for the F-statistic in mixed models are not always reliable. Various researchers have developed methods for adjusting these degrees of freedom. One popular method is due to Kenward and Roger (1997). We will illustrate the use of this method later in this chapter. Even if the adjustment is opti-

mal, there remains the problem that the null distribution may not be F. Furthermore, the method is relevant only for the testing of fixed effects.

We can use bootstrap methods to find more accurate p-values for the likelihood ratio test. The usual bootstrap approach is nonparametric in that no distribution is assumed. Since we are willing to assume normality for the errors and the random effects, we can use a technique called the *parametric bootstrap*. We generate data under the null model using the fitted parameter estimates. We compute the likelihood ratio statistic for this generated data. We repeat this many times and use this to judge the significance of the observed test statistic. This approach will be demonstrated below. The problem may also be addressed by using Bayesian methods to fit the models. We discuss these in Chapter 12.

Model Selection: For comparing larger numbers of models, it is unwise to take a testing-based approach to selection. The problems are similar to those encountered in model selection for standard linear models. When the number of models considered becomes more than a handful, the issue of multiple testing arises and p-values lose their normal meaning. Instead it is better to take a criterion-based approach to model selection. Although we can develop the ideas of model selection of linear models and extend them to linear mixed models, there are some important additional difficulties which means that this extension is not straightforward. Firstly, the dependent response means that effective sample size is less than the total number of cases. Secondly, we have two kinds of parameters, some for the fixed effects and some for the random effects. It is not clear how these two types of parameters should be counted together. Thirdly, most criteria are based on the likelihood which does not behave well at the boundary of the parameter space as can occur with variance parameters.

The Akaike Information Criterion (AIC) and its variations are the most popular model selection criterion. In the `lme4` package, AIC is defined as:

$$-2(\texttt{max log likelihood}) + 2p$$

where p is the total number of parameters. We can confidently use this criterion to compare models which differ only in their fixed effects, as the number of random effect parameters will be the same for all models considered. If we compare models where the random effects are also varied, then we must think more carefully about how to count the random effect parameters. This is problematic due to the aforementioned boundary problems.

Other criteria can be considered. The Bayes Information Criterion (BIC) replaces the $2p$ in the AIC with $p \log n$ and tends to prefer smaller models to the AIC. Another popular criterion used with mixed effect models is the Deviance Information Criterion (DIC) of Spiegelhalter et al. (2002). This criterion is more suited to the Bayesian models discussed in Chapter 12. For a discussion of model selection criteria, see Section A.3. For the specific application to linear mixed models, see Müller et al. (2013). For most of the examples considered in this chapter, there are only a few variables so we are able to rely on testing methods to choose between just a few models. We defer an example of using these methods to Section 10.10.

Example: Now let's demonstrate these inferential methods on the `pulp` data. The fixed effect analysis shows that the operator effects are statistically significant

with a p-value of 0.023. A random effects analysis using the expected mean squares approach yields exactly the same F-statistic for the one-way ANOVA. This method works exactly for such a simple model.

We can also employ the likelihood ratio approach to test the null hypothesis that the variance between the operators is zero. In the fixed effects model, we tested the hypothesis that the four operators had the same effect. In the mixed effect model where the operators are treated as random, the hypothesis that this variance is zero claims that there is no differences between operators in the population. This is a stronger claim than the fixed effect model hypothesis about just the four chosen operators.

We first fit the null model:

```
nullmod <- lm(bright ~ 1, pulp)
```

As there are no random effects in this model, we must use `lm`. For models of the same class, we could use `anova` to compute the LRT and its p-value. Here, we need to compute this directly:

```
lrtstat <- as.numeric(2*(logLik(smod)-logLik(nullmod)))
pvalue <- pchisq(lrtstat,1,lower=FALSE)
data.frame(lrtstat, pvalue)
```

```
  lrtstat  pvalue
1  2.5684 0.10902
```

The p-value is now well above the 5% significance level. We cannot say that this result is necessarily wrong, but the use of the χ^2 approximation does cause us to doubt the result.

We can use the parametric bootstrap approach to obtain a more accurate p-value. We need to estimate the probability, given that the null hypothesis is true, of observing an LRT of 2.5684 or greater. Under the null hypothesis, $y \sim N(\mu, \sigma^2)$. A simulation approach generates data under this model, fits the null and alternative models and computes the LRT statistic. The process is repeated a large number of times and the proportion of LRT statistics exceeding the observed value of 2.5684 is used to estimate the p-value. In practice, we do not know the true values of μ and σ, but we can use the estimated values; this distinguishes the parametric bootstrap from the purely simulation approach. The `simulate` function makes it simple to generate a sample from a model:

```
y <- simulate(nullmod)
```

Now taking the data we generate, we fit both the null and alternative models and then compute the LRT. We repeat the process 1000 times:

```
lrstat <- numeric(1000)
set.seed(123)
for(i in 1:1000){
   y <- unlist(simulate(nullmod))
   bnull <- lm(y ~ 1)
   balt <- lmer(y ~ 1 + (1|operator), pulp, REML=FALSE)
   lrstat[i] <- as.numeric(2*(logLik(balt)-logLik(bnull)))
  }
```

We have set the random number seed here so that the results will reproduce exactly if you run the same code. You do not need to set a seed for your own data unless you need to achieve the same reproducibility. Be aware that simulation naturally contains

some variation. If this variation might make a difference to your conclusions, you need to use a larger number of bootstrap samples.

We may examine the distribution of the bootstrapped LRTs. We compute the proportion that are close to zero:

```
mean(lrstat < 0.00001)
```
```
[1] 0.703
```

We see there is a 70% chance that the likelihoods for the null and alternatives are virtually identical giving an LRT statistic of practically zero. The LRT clearly does not have a χ^2 distribution. There is some discussion of this matter in Stram and Lee (1994), who propose a 50:50 mixture of a χ^2 and a mass at zero. Unfortunately, as we can see, the relative proportions of these two components vary from case to case. Crainiceanu and Ruppert (2004) give a more complete solution to the one-way ANOVA problem, but there is no general and exact result for this and more complex problems. The parametric bootstrap may be the simplest approach. The method we have used above is transparent and could be computed much more efficiently if speed is an issue.

Our estimated *p*-value is:

```
mean(lrstat > 2.5684)
```
```
[1] 0.019
```

We can compute the standard error for this estimate by:

```
sqrt(0.019*0.981/1000)
```
```
[1] 0.0043173
```

So we can be fairly sure it is under 5%. If in doubt, do some more replications to make sure; this only costs computer time. As it happens, this *p*-value is close to the fixed effects *p*-value.

The `RLRsim` package of Scheipl et al. (2008) can be used to test random effect terms:

```
library(RLRsim)
exactLRT(smod, nullmod)
```
```
No restrictions on fixed effects. REML-based inference preferable.

        simulated finite sample distribution of LRT. (p-value based
        on 10000 simulated values)

data:
LRT = 2.5684, p-value = 0.0213
```

The result is obtained with less computing time than our explicitly worked example. The difference in the outcomes is within the sampling error. As the output points out, it is slightly better to use REML when testing the random effects (although remember that REML would be invalid for testing fixed effects). We can make this computation:

```
exactRLRT(mmod)
```
```
        simulated finite sample distribution of RLRT.

        (p-value based on 10000 simulated values)

data:
RLRT = 3.4701, p-value = 0.021
```

Notice that the testing function is now `exactRLRT` and that only the alternative model needs to be specified as there is only one random effect component. The outcome is very similar to those obtained previously.

The parametric bootstrap can also be used to construct confidence intervals for the parameters. We simulate data from the chosen model and estimate the parameters. We repeat this process many times, storing the results each time. Quantiles of the bootstrapped estimates are then used to compute the intervals. We need to be able to extract the parameter estimates from the model. We can view the estimates of variance parameters using:

```
VarCorr(mmod)
 Groups   Name        Std.Dev.
 operator (Intercept) 0.261
 Residual             0.326
```

A more convenient form for extracting the values can be obtained as:

```
as.data.frame(VarCorr(mmod))
       grp        var1 var2     vcov   sdcor
1 operator (Intercept) <NA> 0.068083 0.26093
2 Residual        <NA> <NA> 0.106250 0.32596
```

Now we are ready to bootstrap:

```
bsd <- numeric(1000)
for(i in 1:1000){
  y <- unlist(simulate(mmod))
  bmod <- refit(mmod, y)
  bsd[i] <- as.data.frame(VarCorr(bmod))$sdcor[1]
}
```

The `refit` function changes only the response in a model we have already fit. This is significantly faster than fitting the model from scratch as the overhead in setting up the model is avoided. The 95% bootstrap confidence interval for σ_α is:

```
quantile(bsd, c(0.025, 0.975))
   2.5%   97.5%
0.00000 0.51335
```

Essentially the same result can be obtained more directly using the `confint` function:

```
confint(mmod, method="boot")
Computing bootstrap confidence intervals ...
                             2.5 %   97.5 %
sd_(Intercept)|operator  0.00000  0.51539
sigma                    0.21347  0.45522
(Intercept)             60.09417 60.69724
```

Nevertheless, it is worth understanding the detailed method of construction to know how it works and to allow one to modify the method if circumstances require it.

In this case, the lower bound is zero. This is not surprising given our earlier uncertainty over whether there really is a difference between the operators. In simpler circumstances, there is a duality between confidence intervals and hypothesis tests in that the outcome of a test can be determined by whether the point null hypothesis lies within the confidence interval. Unfortunately, this duality does not apply in all circumstances, this being a case in point. If you want to do a hypothesis test, use the method described earlier and not the confidence interval.

In this example, the random and fixed effect tests gave similar outcomes. However, the hypotheses in random and fixed effects are intrinsically different. To generalize somewhat, it is easier to conclude there is an effect in a fixed effects model since the conclusion applies only to the levels of the factor used in the experiment, while for random effects, the conclusion extends to levels of the factor not considered.

Since the range of the random effect conclusions is greater, the evidence necessarily has to be stronger.

10.3 Estimating Random Effects

In a fixed effects model, the effects are represented by parameters and it makes sense to estimate them. For example, in the one-way ANOVA model:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

We can calculate $\hat{\alpha}_i$. We do need to resolve the identifiability problem with the αs and the μ, but once we decide on this, the meaning of the $\hat{\alpha}$s is clear enough. We can then proceed to make further inference such as multiple comparisons of these levels.

In a model with random effects, the αs are no longer parameters, but random variables. Using the standard normal assumption:

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

It does not make sense to estimate the αs because they are random variables. So instead, we might think about the expected values. However:

$$E\alpha_i = 0 \qquad \forall i$$

which is clearly not very interesting. If one looks at this from a Bayesian point of view, as described in, for example, Gelman et al. (2013), we have a prior density on the αs. The prior mean is $E\alpha_i = 0$. Let f represent density, then the posterior density for α is given by:

$$f(\alpha_i|y) \propto f(y|\alpha_i) f(\alpha_i)$$

We can then find the posterior mean, denoted by $\hat{\alpha}$ as:

$$E(\alpha_i|y) = \int \alpha_i f(\alpha_i|y) d\alpha_i$$

For the general case, this works out to be:

$$\hat{\alpha} = DZ^T V^{-1}(y - X\beta)$$

Now a purely Bayesian approach would specify the parameters of the prior (or specify priors for these) and compute a posterior distribution for α. Here we take an empirical Bayes point of view and substitute the MLEs into D, V and β to obtain the predicted random effects. These may be computed as:

```
ranef(mmod)$operator
```

```
  (Intercept)
a    -0.12194
b    -0.25912
c     0.16767
d     0.21340
```

The predicted random effects are related to the fixed effects. These fixed effects are:

```
(cc <- model.tables(lmod))
```

```
Tables of effects

 operator
operator
    a     b     c     d
-0.16 -0.34  0.22  0.28
```

Let's compute the ratio to the random effects as:

```
cc[[1]]$operator/ranef(mmod)$operator
  X.Intercept.
a       1.3121
b       1.3121
c       1.3121
d       1.3121
```

We see that the predicted random effects are exactly in proportion to the fixed effects. Typically, the predicted random effects are smaller and could be viewed as a type of *shrinkage* estimate.

The 95% confidence intervals for the random effects can be calculated and displayed as seen in Figure 10.2.

```
library(lattice)
dotplot(ranef(mmod, condVar=TRUE))
```

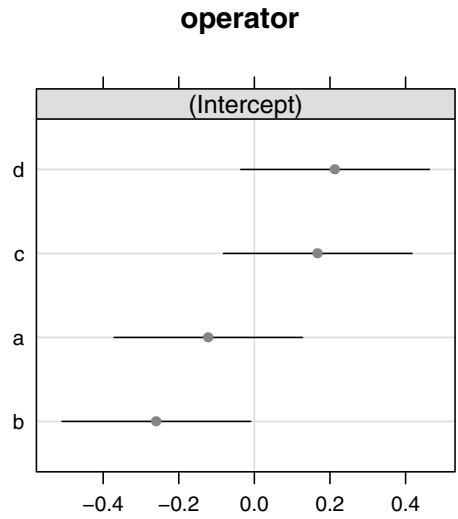


Figure 10.2 *Confidence intervals for the random effects in the pulp data.*

10.4 Prediction

Suppose we wish to predict a new value. If the prediction is to be made for a new operator or unknown operator, the best we can do is give $\hat{\mu} = 60.4$. If we know the operator, then we can combine this with our fixed effects to produce what are known as the *best linear unbiased predictors* (BLUPs) as follows:

```
fixef(mmod)+ranef(mmod)$operator
  (Intercept)
a      60.278
b      60.141
c      60.568
d      60.613
```

We can also use the `predict` function to construct these predictions. First consider the prediction for a new or unknown operator. We specify the random effects part of the prediction as `~0` meaning that this term is not present. In more complex models with more than one random effect, more choices are available. By default this would produce a fitted value for each case in the data but since these are identical we take only the first value:

```
predict(mmod, re.form=~0)[1]
   1
60.4
```

Now we specify that the operator is 'a':

```
predict(mmod, newdata=data.frame(operator="a"))
     1
60.278
```

The `predict` function for mixed model objects does not compute standard errors or prediction intervals. For this simple model, it would be possible to compute these explicitly but for more general models, it becomes much more difficult. For this reason, we present a parametric bootstrap method for computing these as it is clearer how the bands are computed. We start with the unknown operator case:

```
group.sd <- as.data.frame(VarCorr(mmod))$sdcor[1]
resid.sd <- as.data.frame(VarCorr(mmod))$sdcor[2]
pv <- numeric(1000)
for(i in 1:1000){
 y <- unlist(simulate(mmod))
 bmod <- refit(mmod, y)
 pv[i] <- predict(bmod, re.form=~0)[1] + rnorm(n=1,sd=group.sd) +
     ↪ rnorm(n=1,sd=resid.sd)
}
quantile(pv, c(0.025, 0.975))
  2.5%  97.5%
59.535 61.286
```

As in previous bootstraps, the first step is to simulate from the fitted model. We refit the model with the simulated response and generate a predicted value. But there are two additional sources of variation. We have variation due to the new operator and also due to a new observation from that operator. For this reason, we add normal sample values with standard deviations equal to those estimated earlier. If you really want a confidence interval for the mean prediction, you should not add these extra error terms. We repeat this 1000 times and take the appropriate quantiles to get a 95% interval.

Some modification is necessary if we know the operator we are making the prediction interval for. We use the option `use.u=TRUE` in the `simulate` function indicating that we should simulate new values conditional on the estimated random effects. We need to do this because otherwise we would simulate an entirely new 'a' effect in each replication. Instead, we want to preserve the originally generated 'a' effect.

```
for(i in 1:1000){
 y <- unlist(simulate(mmod, use.u=TRUE))
 bmod <- refit(mmod, y)
 pv[i] <- predict(bmod, newdata=data.frame(operator="a")) + rnorm(n=1,
     ↪ sd=resid.sd)
}
quantile(pv, c(0.025, 0.975))
```

```
  2.5%  97.5%
59.606 61.023
```

In a simple model such as this, we could mathematically calculate the standard error formulas and use this to compute these intervals more efficiently. However, the bootstrap is more general and is easier to apply in more complex situations. More bootstrapping functionality can be found in the `lme4::bootMer()` function and also in the `merTools` package. Bootstrapping is fast enough for simple models but greater efficiency is needed in more complex cases.

10.5 Diagnostics

It is important to check the assumptions made in fitting the model. Diagnostic methods available for checking linear mixed models largely mirror those used for linear models but there are some variations. Residuals are commonly defined as the difference between the observed and fitted values. In mixed models, there is more than one kind of fitted (or predicted) value resulting in more than one kind of residual. The default predicted values and residuals use the estimated random effects. This means these residuals can be regarded as estimates of ε which is usually what we want.

As with linear models, this pair of diagnostics plots is most valuable:

```
qqnorm(residuals(mmod),main="")
plot(fitted(mmod),residuals(mmod),xlab="Fitted",ylab="Residuals")
abline(h=0)
```

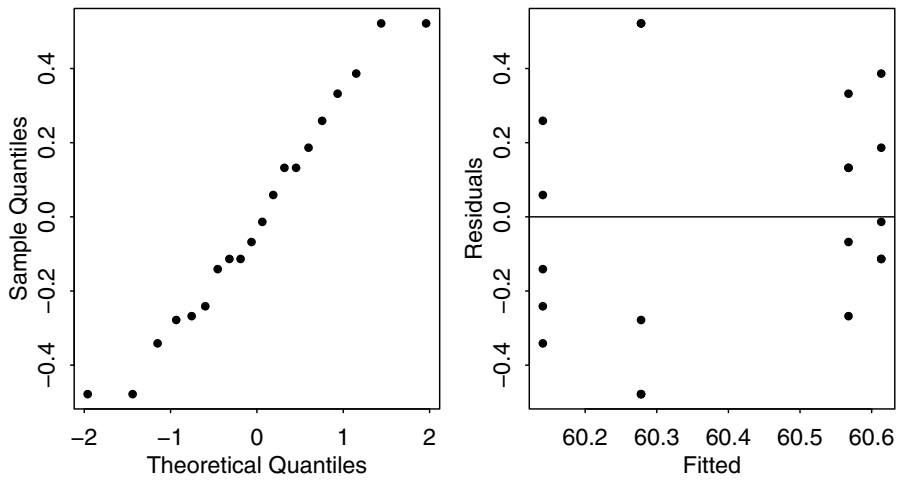


Figure 10.3 *Diagnostic plots for the one-way random effects model.*

The plots are shown in Figure 10.3 and indicate no particular problems. Random effects models are particularly sensitive to outliers, because they depend on variance components that can be substantially inflated by unusual points. The QQ plot is one way to pick out outliers. We also need the normality for the testing. The residual-fitted plot is also important because we made the assumption that the error variance was constant.

If we had more than four groups, we could also look at the normality of the group level effects and check for constant variance also. With so few groups, it is not sensible to do this. Also note that there is no particular reason to think about multiple comparisons. These are for comparing selected levels of a factor. For a random effect, the levels were randomly selected, so such comparisons have less motivation.

10.6 Blocks as Random Effects

Blocks are properties of the experimental units. The blocks are either clearly defined by the conditions of the experiment or they are formed with the judgment of the experimenter. Sometimes, blocks represent groups of runs completed in the same period of time. Typically, we are not interested in the block effects specifically, but must account for their effect. It is therefore natural to treat blocks as random effects.

We illustrate with an experiment to compare four processes, A, B, C and D, for the production of penicillin. These are the treatments. The raw material, corn steep liquor, is quite variable and can only be made in blends sufficient for four runs. Thus a randomized complete block design is suggested by the nature of the experimental units. The data comes from Box et al. (1978). We start with the fixed effects analysis:

```
data(penicillin, package="faraway")
summary(penicillin)
```

```
 treat    blend       yield
 A:5   Blend1:4   Min.   :77
 B:5   Blend2:4   1st Qu.:81
 C:5   Blend3:4   Median :87
 D:5   Blend4:4   Mean   :86
       Blend5:4   3rd Qu.:89
                  Max.   :97
```

We plot the data as seen in Figure 10.4. We create a version of the blend variable to get neater labeling.

```
penicillin$Blend <- gl(5,4)
ggplot(penicillin, aes(y=yield, x=treat, shape=Blend))+geom_point()+
    ↪ xlab("Treatment")
ggplot(penicillin, aes(y=yield, x=Blend, shape=treat))+geom_point()
```

It is convenient to use sum contrasts rather than the default treatment contrasts for the purpose of comparison to the mixed effect modeling to come.

```
op <- options(contrasts=c("contr.sum", "contr.poly"))
lmod <- aov(yield ~ blend + treat, penicillin)
summary(lmod)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
blend        4  264.0    66.0    3.50  0.041
treat        3   70.0    23.3    1.24  0.339
Residuals   12  226.0    18.8
```

```
coef(lmod)
```

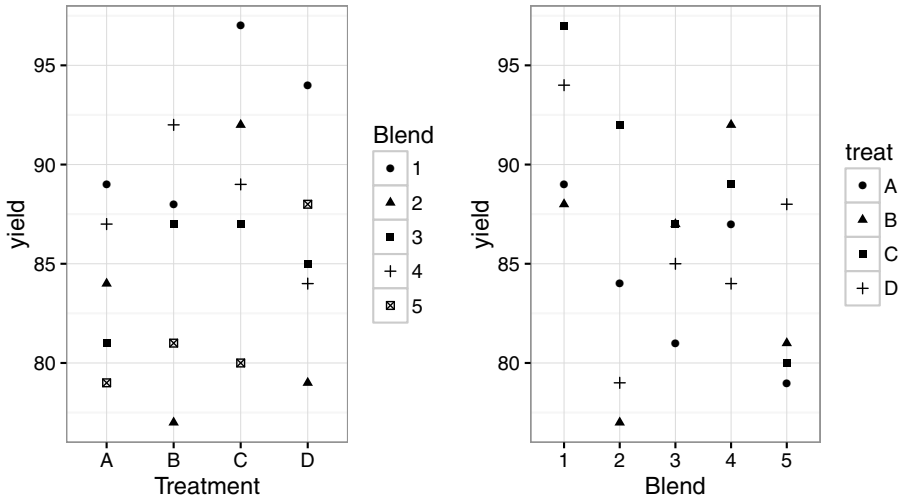


Figure 10.4 *Yield from penicillin blends varying by treatment.*

```
(Intercept)      blend1      blend2      blend3      blend4
         86           6          -3          -1           2
     treat1      treat2      treat3
         -2          -1           3
```

From this we see that there is no significant difference between the treatments, but there is between the blends. Now let's fit the data with a mixed model, where we have fixed treatment effects, but random blend effects. This seems natural since the blends we use can be viewed as having been selected from some notional population of blends.

```
mmod <- lmer(yield ~ treat + (1|blend), penicillin)
sumary(mmod)
Fixed Effects:
            coef.est coef.se
(Intercept) 86.00     1.82
treat1      -2.00     1.68
treat2      -1.00     1.68
treat3       3.00     1.68

Random Effects:
 Groups   Name        Std.Dev.
 blend    (Intercept) 3.43
 Residual             4.34
---
number of obs: 20, groups: blend, 5
AIC = 118.6, DIC = 128
deviance = 117.3
options(op)
```

We notice a few connections. The residual variance is the same in both cases: $18.8 = 4.34^2$. This is because we have a balanced design and so REML is equivalent to the

ANOVA estimator. The treatment effects are also the same as is the overall mean. The BLUPs for the random effects are:

```
ranef(mmod)$blend
       (Intercept)
Blend1     4.28788
Blend2    -2.14394
Blend3    -0.71465
Blend4     1.42929
Blend5    -2.85859
```

which, as with the one-way ANOVA, are a shrunken version of the corresponding fixed effects. The usual diagnostics show nothing amiss.

We have a number of options in testing the fixed effects in this example. For this simple balanced model, the `aov` function can be used:

```
amod <- aov(yield ~ treat + Error(blend), penicillin)
summary(amod)
Error: blend
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  4    264      66

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
treat      3     70    23.3    1.24   0.34
Residuals 12    226    18.8
```

Notice how the random effects term for `blend` is specified. The p-value for testing the treatment effects is 0.34 indicating no significant effect. This test is exact and works well here but only works for simple balanced models. For example, a single missing value would invalidate this test so we have good reason to explore more general methods.

We might try to base a test on the F-statistic which can be obtained like this:

```
anova(mmod)
Analysis of Variance Table
      Df Sum Sq Mean Sq F value
treat  3     70    23.3    1.24
```

For this simple balanced case, it can be shown that this F-statistic has a true F null distribution with usual degrees of freedom from which we could derive a p-value. Unfortunately, as with the `aov` method, this result does not generalize well.

More reliable F-tests can be achieved by using adjusted degrees of freedom. The `pbkrtest` package (Halekoh and Højsgaard (2014)) implements the Kenward-Roger (Kenward and Roger (1997)) method:

```
library(pbkrtest)
amod <- lmer(yield ~ treat + (1|blend), penicillin, REML=FALSE)
nmod <- lmer(yield ~ 1 + (1|blend), penicillin, REML=FALSE)
KRmodcomp(amod, nmod)
F-test with Kenward-Roger approximation; computing time: 0.14 sec.
large : yield ~ treat + (1 | blend)
small : yield ~ 1 + (1 | blend)
       stat   ndf   ddf F.scaling p.value
Ftest  1.24  3.00 12.00         1    0.34
```

It is essential to use the ML method of estimation when testing fixed effects. Since we wish to test the treatment effect, we fit the model with this term and the same model but without this term. As can be seen, it produces an identical result to the `aov`

output with the same degrees of freedom and p-value. The advantage of this method is that it can be generalized to a much wider class of problems.

We can also use the parametric bootstrap. First we compute the LRT statistic:

```
as.numeric(2*(logLik(amod)-logLik(nmod)))
```

```
[1] 4.0474
```

Just for reference, we could use the χ^2 approximation to quickly compute a p-value:

```
1-pchisq(4.0474,3)
```

```
[1] 0.25639
```

This is just an approximation of unknown quality. We aim to do better than this.

We can improve the accuracy with the parametric bootstrap approach. We can generate a response from the null model and use this to compute the LRT. We repeat this 1000 times, saving the LRT each time:

```
lrstat <- numeric(1000)
for(i in 1:1000){
ryield <- unlist(simulate(nmod))
nmodr <- refit(nmod, ryield)
amodr <- refit(amod, ryield)
lrstat[i] <- 2*(logLik(amodr)-logLik(nmodr))
}
```

Notice how we have used `refit` to speed up the computation. Under the standard likelihood theory, the LRT statistic here should have a χ^2_3 distribution. A QQ plot of these simulated LRT values indicates that this is a poor approximation. We can compute our estimated p-value as:

```
mean(lrstat > 4.0474)
```

```
[1] 0.353
```

which is much closer to the F-test result than the χ^2_3-based approximation.

The `pbkrtest` package offers a convenient way to perform the parametric bootstrap for fixed effect terms:

```
pmod <- PBmodcomp(amod, nmod)
summary(pmod)
```

```
Parametric bootstrap test; time: 32.22 sec; samples: 1000 extremes: 333;
large : yield ~ treat + (1 | blend)
small : yield ~ 1 + (1 | blend)
         stat   df  ddf p.value
PBtest   4.05              0.33
Gamma    4.05              0.33
Bartlett 3.42 3.00         0.33
F        1.35 3.00 12.9    0.30
LRT      4.05 3.00         0.26
```

The parametric bootstrap p-value is 0.33, which is similar to our previous results. Remember that bootstrap is based on random sampling so if you repeat this, you will get slightly different results. Since this p-value is not close to significance, we have no worries about this. Notice that the output also produces the χ^2-based LRT result along with three other versions that are explained in the documentation for the `pbkrtest` package. The package also offers the possibility of using the multiple cores available now on most computers. This parallel computing can be helpful as the parametric bootstrap is computationally expensive.

We can also test the significance of the blends. As with a fixed effects analysis, we are typically not directly interested in size of the blocking effects. Once having decided to design the experiment with blocks, we must retain them in the model.

However, we may wish to examine the blocking effects for information useful for the design of future experiments. We can fit the model with and without random effects and compute the LRT:

```
rmod <- lmer(yield ~ treat + (1|blend), penicillin)
nlmod <- lm(yield ~ treat, penicillin)
as.numeric(2*(logLik(rmod)-logLik(nlmod,REML=TRUE)))
```

```
[1] 2.7629
```

We need to specify the nondefault REML option for null model to ensure that the LRT is computed correctly. Now we perform the parametric bootstrap much as before:

```
lrstatf <- numeric(1000)
for(i in 1:1000){
   ryield <-  unlist(simulate(nlmod))
   nlmodr <- lm(ryield ~ treat, penicillin)
   rmodr <- refit(rmod, ryield)
   lrstatf[i] <- 2*(logLik(rmodr)-logLik(nlmodr,REML=TRUE))
  }
```

Again, the distribution is far from χ^2_1 which is clear when we examine the proportion of generated LRTs which are close to zero:

```
mean(lrstatf < 0.00001)
```

```
[1] 0.551
```

We can see from this that the LRT is clearly not χ^2_1 distributed. Even the nonzero values seem to have some other distribution. This makes it clear that asymptotic approximations cannot be relied on these circumstances.

We can compute the estimated p-value as:

```
mean(lrstatf > 2.7629)
```

```
[1] 0.043
```

So we find a significant blend effect. The p-value is close to that observed for the fixed effects analysis. Given that the p-value is close to 5%, we might wish to increase the number of bootstrap samples to increase our confidence in the result.

We can also use `RLRsim` to obtain a p-value.

```
library(RLRsim)
exactRLRT(rmod)
```

```
        simulated finite sample distribution of RLRT.

        (p-value based on 10000 simulated values)

data:
RLRT = 2.7629, p-value = 0.0406
```

In this example, we saw no major advantage in modeling the blocks as random effects, so we might prefer to use the fixed effects analysis as it is simpler to execute. However, in subsequent analyses, we shall see that the use of random effects will be mandatory as equivalent results may not be obtained from a purely fixed effects analysis.

10.7 Split Plots

Split plot designs originated in agriculture, but occur frequently in other settings. As the name implies, main plots are split into several subplots. The main plot is treated with a level of one factor while the levels of some other factor are allowed to vary

with the subplots. The design arises as a result of restrictions on a full randomization. For example, a field may be divided into four subplots. It may be possible to plant different varieties in the subplots, but only one type of irrigation may be used for the whole field. Note the distinction between split plots and blocks. Blocks are features of the experimental units which we have the option to take advantage of in the experimental design. Split plots impose restrictions on what assignments of factors are possible. They impose requirements on the design that prevent a complete randomization. Split plots often arise in nonagricultural settings when one factor is easy to change while another factor takes much more time to change. If the experimenter must do all runs for each level of the hard-to-change factor consecutively, a split-plot design results with the hard-to-change factor representing the whole plot factor.

Consider the following example. In an agricultural field trial, the objective was to determine the effects of two crop varieties and four different irrigation methods. Eight fields were available, but only one type of irrigation may be applied to each field. The fields may be divided into two parts with a different variety planted in each half. The whole plot factor is the method of irrigation, which should be randomly assigned to the fields. Within each field, the variety is randomly assigned. Here is a summary of the data:

```
data(irrigation,package="faraway")
summary(irrigation)
```

```
     field   irrigation variety     yield
 f1     :2   i1:4       v1:8    Min.   :34.8
 f2     :2   i2:4       v2:8    1st Qu.:37.6
 f3     :2   i3:4               Median :40.1
 f4     :2   i4:4               Mean   :40.2
 f5     :2                      3rd Qu.:42.7
 f6     :2                      Max.   :47.6
 (Other):4
```

We can plot the data as seen in Figure 10.5.

```
ggplot(irrigation, aes(y=yield, x=field, shape=irrigation, color=
    ↪ variety)) + geom_point()
```

The irrigation and variety are fixed effects, but the field is clearly a random effect. We must also consider the interaction between field and variety, which is necessarily also a random effect because one of the two components is random. The fullest model that we might consider is:

$$y_{ijk} = \mu + i_i + v_j + (iv)_{ij} + f_k + (vf)_{jk} + \varepsilon_{ijk}$$

$\mu, i_i, v_j, (iv)_{ij}$ are fixed effects; the rest are random having variances σ^2_f, σ^2_{vf} and σ^2_ε. Note that we have no $(if)_{ik}$ term in this model. It would not be possible to estimate such an effect since only one type of irrigation is used on a given field; the factors are not crossed. We would fit such a model using the expression

```
lmer(yield ~ irrigation * variety + (1|field) +(1|field:variety),
    ↪ irrigation)
```

However, if you try to fit such a model, it will fail because it is not possible to distinguish the variety within the field variation. We would need more than one observation per variety within each field for us to separate the two variabilities. We resort to a simpler model that omits the variety by field interaction random effect:

$$y_{ijk} = \mu + i_i + v_j + (iv)_{ij} + f_k + \varepsilon_{ijk}$$

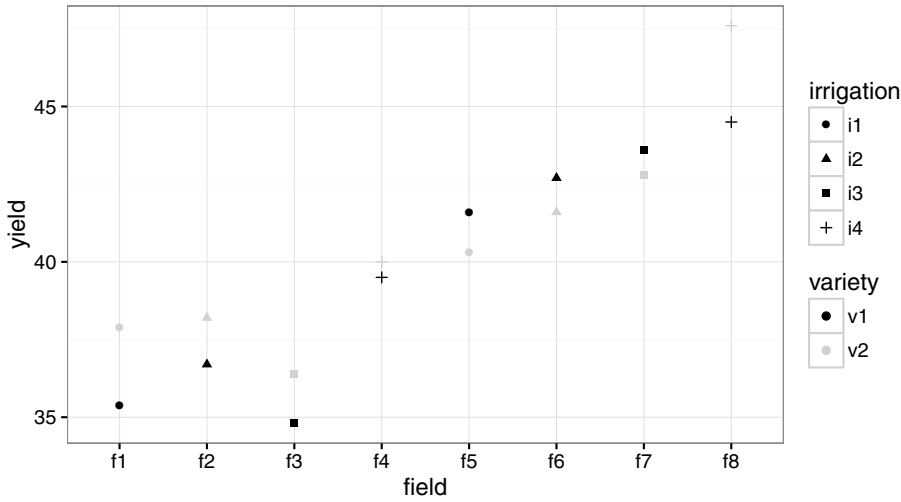


Figure 10.5 *Yield on fields with different irrigation methods.*

```
lmod <- lmer(yield ~ irrigation * variety + (1|field), irrigation)
sumary(lmod)
Fixed Effects:
                       coef.est coef.se
(Intercept)            38.50     3.03
irrigationi2            1.20     4.28
irrigationi3            0.70     4.28
irrigationi4            3.50     4.28
varietyv2               0.60     1.45
irrigationi2:varietyv2 -0.40     2.05
irrigationi3:varietyv2 -0.20     2.05
irrigationi4:varietyv2  1.20     2.05

Random Effects:
 Groups   Name        Std.Dev.
 field    (Intercept) 4.02
 Residual             1.45
---
number of obs: 16, groups: field, 8
AIC = 65.4, DIC = 91.8
deviance = 68.6
```

We can see that the largest variance component is that due to the field effect: $\hat{\sigma}_f = 4.02$ with $\hat{\sigma}_\varepsilon = 1.45$.

The relatively large standard errors compared to the fixed effect estimates suggest that there may be no significant fixed effects. We can check this sequentially using F-tests with adjusted degrees of freedom:

```
library(pbkrtest)
lmoda <- lmer(yield ~ irrigation + variety + (1|field),data=irrigation
    ↪ )
KRmodcomp(lmod, lmoda)
F-test with Kenward-Roger approximation; computing time: 0.07 sec.
```

```
large : yield ~ irrigation * variety + (1 | field)
small : yield ~ irrigation + variety + (1 | field)
      stat  ndf  ddf F.scaling p.value
Ftest 0.25 3.00 4.00         1    0.86
```

We find there is no significant interaction term. We can now test each of the main effects starting with the variety:

```
lmodi <- lmer(yield ~ irrigation + (1|field), irrigation)
KRmodcomp(lmoda, lmodi)
```

```
F-test with Kenward-Roger approximation; computing time: 0.06 sec.
large : yield ~ irrigation + variety + (1 | field)
small : yield ~ irrigation + (1 | field)
      stat  ndf  ddf F.scaling p.value
Ftest 1.58 1.00 7.00         1    0.25
```

Dropping variety from the model seems reasonable since the p-value of 0.25 is large. We can test irrigation in a similar manner:

```
lmodv <- lmer(yield ~  variety + (1|field), irrigation)
KRmodcomp(lmoda, lmodv)
```

```
F-test with Kenward-Roger approximation; computing time: 0.06 sec.
large : yield ~ irrigation + variety + (1 | field)
small : yield ~ variety + (1 | field)
      stat  ndf  ddf F.scaling p.value
Ftest 0.39 3.00 4.00         1    0.77
```

Irrigation also fails to be significant.

We should check the diagnostic plots to make sure there is nothing amiss:

```
plot(fitted(lmod),residuals(lmod),xlab="Fitted",ylab="Residuals")
qqnorm(residuals(lmod),main="")
```

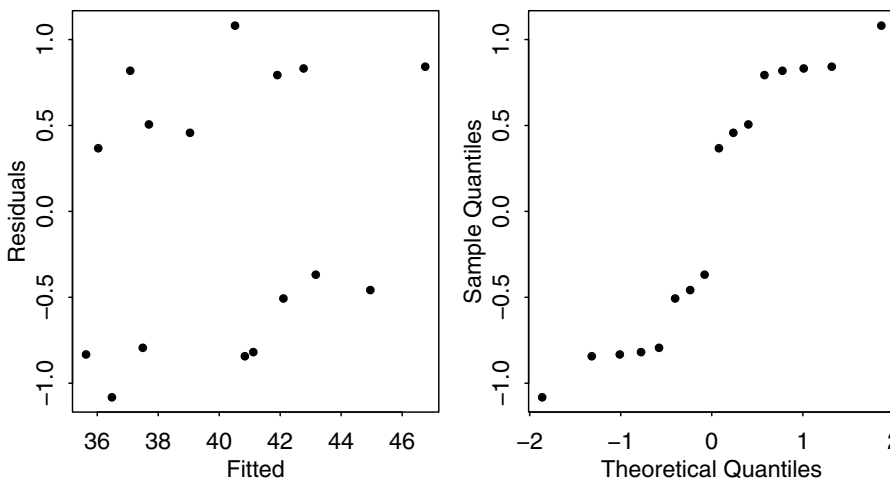


Figure 10.6 *Diagnostic plots for the split plot example.*

We can see in Figure 10.6 that there is no problem with the nonconstant variance, but that the residuals indicate a bimodal distribution caused by the pairs of observations in each field. This type of divergence from normality is unlikely to cause any major problems with the estimation and inference.

We can test the random effects term like this:

```
library(RLRsim)
exactRLRT(lmod)
```

```
        simulated finite sample distribution of RLRT.

        (p-value based on 10000 simulated values)

data:
RLRT = 6.1118, p-value = 0.0098
```

We see that the fields do seem to vary as the result is clearly significant.

Sometimes analysts ignore the split-plot variable as in:

```
mod <- lm(yield ~ irrigation * variety, data=irrigation)
anova(mod)
```

```
Analysis of Variance Table

Response: yield
                  Df Sum Sq Mean Sq F value Pr(>F)
irrigation         3   40.2    13.4    0.73   0.56
variety            1    2.2     2.2    0.12   0.73
irrigation:variety  3    1.6     0.5    0.03   0.99
Residuals          8  146.5    18.3
```

The results will not be the same. This last model is incorrect because it fails to take into account the restrictions on the randomization introduced by the fields and the additional variability thereby induced.

10.8 Nested Effects

When the levels of one factor vary only within the levels of another factor, that factor is said to be *nested*. For example, when measuring the performance of workers at several different job locations, if the workers only work at one location, the workers are nested within the locations. If the workers work at more than one location, then the workers are *crossed* with locations.

Here is an example to illustrate nesting. Consistency between laboratory tests is important and yet the results may depend on who did the test and where the test was performed. In an experiment to test levels of consistency, a large jar of dried egg powder was divided up into a number of samples. Because the powder was homogenized, the fat content of the samples is the same, but this fact is withheld from the laboratories. Four samples were sent to each of six laboratories. Two of the samples were labeled as G and two as H, although in fact they were identical. The laboratories were instructed to give two samples to two different technicians. The technicians were then instructed to divide their samples into two parts and measure the fat content of each. So each laboratory reported eight measures, each technician four measures, that is, two replicated measures on each of two samples. The data comes from Bliss (1967):

```
data(eggs, package="faraway")
summary(eggs)
```

```
      Fat          Lab    Technician Sample
 Min.   :0.060   I  :8   one:24     G:24
 1st Qu.:0.307   II :8   two:24     H:24
 Median :0.370   III:8
```

```
 Mean   :0.388   IV :8
 3rd Qu.:0.430   V  :8
 Max.   :0.800   VI :8
```

We can plot the data as seen in Figure 10.7.

```
library(ggplot2)
ggplot(eggs, aes(y=Fat, x=Lab, color=Technician, shape=Sample)) + geom
    ↪ _point(position = position_jitter(width=0.1, height=0.0))+scale
    ↪ _color_grey()
```

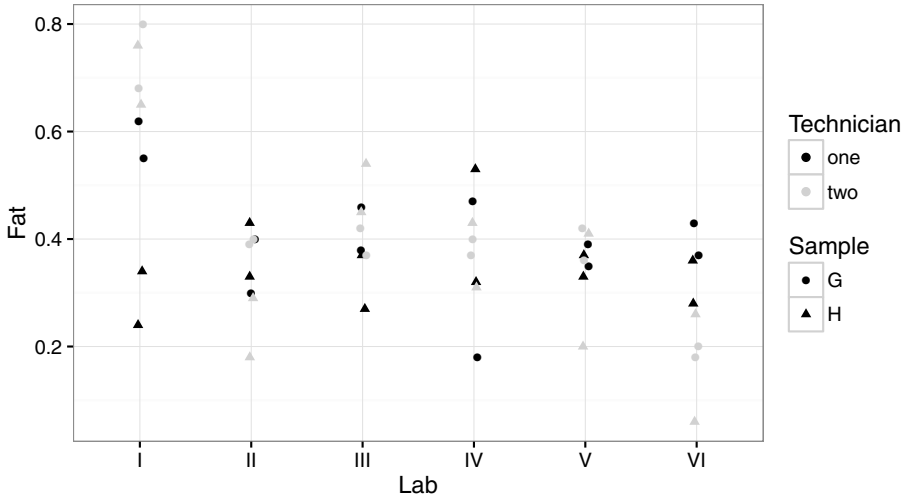


Figure 10.7 *Fat content of homogenous powdered egg as tested by different laboratories, technicians and samples.*

Although the technicians have been labeled "one" and "two," they are two different people in each lab. Thus the technician factor is nested within laboratories. Furthermore, even though the samples are labeled "H" and "G," these are not the same samples across the technicians and the laboratories. Hence we have samples nested within technicians. Technicians and samples should be treated as random effects since we may consider these as randomly sampled. If the labs were specifically selected, then they should be taken as fixed effects. If, however, they were randomly selected from those available, then they should be treated as random effects. If the purpose of the study is to come to some conclusion about consistency across laboratories, the latter approach is advisable.

For the purposes of this analysis, we will treat labs as random. So all our effects (except the grand mean) are random. The model is:

$$y_{ijkl} = \mu + L_i + T_{ij} + S_{ijk} + \varepsilon_{ijkl}$$

This can be fit using:

```
cmod <- lmer(Fat ~ 1 + (1|Lab) + (1|Lab:Technician) +  (1|Lab:
    ↪ Technician:Sample), data=eggs)
sumary(cmod)
```

```
Fixed Effects:
coef.est  coef.se
    0.39     0.04

Random Effects:
 Groups                Name        Std.Dev.
 Lab:Technician:Sample (Intercept) 0.06
 Lab:Technician        (Intercept) 0.08
 Lab                   (Intercept) 0.08
 Residual                          0.08
---
number of obs: 48, groups: Lab:Technician:Sample, 24; Lab:Technician, 12; Lab, 6
AIC = -54.2, DIC = -73.3
deviance = -68.8
```

So we have $\hat{\sigma}_L = 0.08$, $\hat{\sigma}_T = 0.08$, $\hat{\sigma}_S = 0.06$ and $\hat{\sigma}_\varepsilon = 0.08$. So all four variance components are of a similar magnitude. The lack of consistency in measures of fat content can be ascribed to variance between labs, variance between technicians, variance in measurement due to different labeling and just plain measurement error. We can see if the model can be simplified by removing the lowest level of the variance components. Again the parametric bootstrap can be used:

```
cmodr <- lmer(Fat ~ 1 + (1|Lab) + (1|Lab:Technician), data=eggs)
lrstat <- numeric(1000)
for(i in 1:1000){
   rFat <- unlist(simulate(cmodr))
   nmod <- lmer(rFat ~ 1 + (1|Lab) + (1|Lab:Technician), data=eggs)
   amod <- lmer(rFat ~ 1 + (1|Lab) + (1|Lab:Technician) +
   (1|Lab:Technician:Sample), data=eggs)
   lrstat[i] <- 2*(logLik(amod)-logLik(nmod))
  }
mean(lrstat > 2*(logLik(cmod)-logLik(cmodr)))
```

```
[1] 0.092
```

We do not reject $H_0 : \sigma_S^2 = 0$. A similar computation may be made using the `RLRsim` package. This requires us to specify another model where only the tested random effect is included:

```
library(RLRsim)
cmods <- lmer(Fat ~ 1 + (1|Lab:Technician:Sample), data=eggs)
exactRLRT(cmods, cmod, cmodr)
```

```
        simulated finite sample distribution of RLRT.

        (p-value based on 10000 simulated values)

data:
RLRT = 1.6034, p-value = 0.1056
```

An examination of the reduced model is interesting:

```
VarCorr(cmodr)
```

```
 Groups         Name        Std.Dev.
 Lab:Technician (Intercept) 0.0895
 Lab            (Intercept) 0.0769
 Residual                   0.0961
```

The variation due to samples has been absorbed into the other components.

So we can reasonably say that the variation due to samples can be ignored. We may now test the significance of the variation between technicians. Using the same method above, this is found to be significant.

Although the data has a natural hierarchical structure which suggests a particular order of testing, we might reasonably wonder which of the components contribute substantially to the overall variation. Why test the sample effect first? A look at the confidence intervals reveals the problem:

```
confint(cmod, method="boot")
```

```
                                       2.5 %   97.5 %
sd_(Intercept)|Lab:Technician:Sample 0.000000 0.097527
sd_(Intercept)|Lab:Technician        0.000000 0.136021
sd_(Intercept)|Lab                   0.000000 0.152663
sigma                                0.058872 0.107040
(Intercept)                          0.299666 0.473920
```

We might drop any of the three random effect terms but it is not possible to be sure which is best to go. It is safest to conclude there is some variation in the fat measurement coming from all three sources.

10.9 Crossed Effects

Effects are said to be crossed when they are not nested. In full factorial designs, effects are completely crossed because every level of one factor occurs with every level of another factor. However, in some other designs, crossing is not complete. An example of less than complete crossing is a latin square design, where there is one treatment factor and two blocking factors. Although not all combinations of factors occur, the blocking factors are not nested. When at least some crossing occurs, methods for nested designs cannot be used. We consider a latin square example.

In an experiment reported by Davies (1954), four materials, A, B, C and D, were fed into a wear-testing machine. The response is the loss of weight in 0.1 mm over the testing period. The machine could process four samples at a time and past experience indicated that there were some differences due to the position of these four samples. Also some differences were suspected from run to run. A fixed effects analysis of this dataset may be found in Faraway (2014). Four runs were made. The latin square structure of the design may be observed:

```
data(abrasion, package="faraway")
matrix(abrasion$material,4,4)
```

```
     [,1] [,2] [,3] [,4]
[1,] "C"  "A"  "D"  "B"
[2,] "D"  "B"  "C"  "A"
[3,] "B"  "D"  "A"  "C"
[4,] "A"  "C"  "B"  "D"
```

We can plot the data as seen in Figure 10.8.

```
library(ggplot2)
ggplot(abrasion,aes(x=material, y=wear, shape=run, color=position))+
    ↪ geom_point(position = position_jitter(width=0.1, height=0.0))+
    ↪ scale_color_grey()
```

A fixed effects analysis of the data reveals:

```
lmod <- aov(wear ~ material + run + position, abrasion)
summary(lmod)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
material     3   4622    1540   25.15 0.00085
run          3    986     329    5.37 0.03901
```

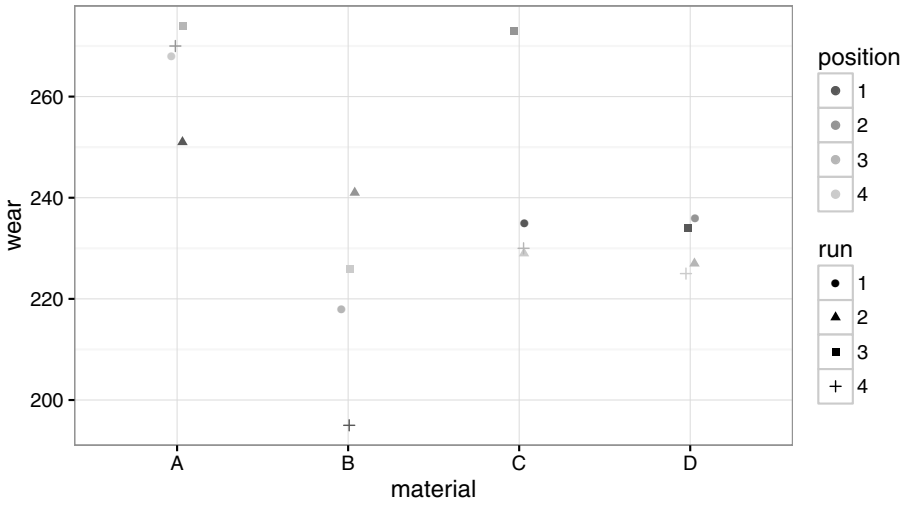


Figure 10.8 *Abrasion wear on material according to run and position.*

```
position     3   1468     489    7.99 0.01617
Residuals    6    367      61
```

All the effects are significant. However, we might regard the run and position as random effects. The appropriate model is then:

```
mmod <- lmer(wear ~ material + (1|run) + (1|position), abrasion)
sumary(mmod)
Fixed Effects:
            coef.est coef.se
(Intercept) 265.75     7.67
materialB   -45.75     5.53
materialC   -24.00     5.53
materialD   -35.25     5.53

Random Effects:
 Groups   Name        Std.Dev.
 run      (Intercept)  8.18
 position (Intercept) 10.35
 Residual              7.83
---
number of obs: 16, groups: run, 4; position, 4
AIC = 114.3, DIC = 140.4
deviance = 120.3
```

The `lmer` function is able to recognize that the run and position effects are crossed and fit the model appropriately. We can test the random effects using the `RLRsim` package. We need to fit both models that use just one random effect:

```
library(RLRsim)
mmodp <- lmer(wear ~ material + (1|position), abrasion)
mmodr <- lmer(wear ~ material + (1|run), abrasion)
exactRLRT(mmodp, mmod, mmodr)
        simulated finite sample distribution of RLRT.
```

```
        (p-value based on 10000 simulated values)

data:
RLRT = 4.5931, p-value = 0.0139
```

This first comparison tests the significance of the `position` term. The first model in the `exactRLRT` specifies the model with only that random effect term being tested. The second and third terms specify the alternative and null models under the hypothesis being tested. We see that the position variance is statistically significant. We can also test the run term:

```
exactRLRT(mmodr, mmod, mmodp)
        simulated finite sample distribution of RLRT.

        (p-value based on 10000 simulated values)

data:
RLRT = 3.0459, p-value = 0.0345
```

We see that the run variation is also statistically significant. Since the design of this experiment has already restricted the randomization to allow for these effects, we would keep these terms in the model even if they were found not to be significant. This information would only be valuable for future experiments.

The fixed effect term can be tested using the `pbkrtest` package. Given the small balanced nature of the experiment, we can feel confident in using the Kenward-Roger adjustment. Note that we need to use ML estimation for the fixed effect comparison.

```
library(pbkrtest)
mmod <- lmer(wear ~ material + (1|run) + (1|position), abrasion,REML=
    ↪ FALSE)
nmod <- lmer(wear ~ 1+ (1|run) + (1|position), abrasion,REML=FALSE)
KRmodcomp(mmod, nmod)
F-test with Kenward-Roger approximation; computing time: 0.15 sec.
large : wear ~ material + (1 | run) + (1 | position)
small : wear ~ 1 + (1 | run) + (1 | position)
      stat  ndf  ddf F.scaling p.value
Ftest 25.1  3.0  6.0         1 0.00085
```

We find that there is a clearly significant difference in the materials.

The fixed effects analysis was somewhat easier to execute, but the random effects analysis has the advantage of producing estimates of the variation in the blocking factors which will be more useful in future studies. Fixed effects estimates of the run effect for this experiment are only useful for the current study.

10.10 Multilevel Models

Multilevel models is a term used for models for data with hierarchical structure. The term is most commonly used in the social sciences. We can use the methodology we have already developed to fit some of these models.

We take as our example some data from the Junior School Project collected from primary (U.S. term is elementary) schools in inner London. The data is described in detail in Mortimore et al. (1988) and a subset is analyzed extensively in Goldstein (1995).

The variables in the data are the `school`, the `class` within the school (up to

four), `gender`, `social` class of the father (I=1; II=2; III nonmanual=3; III manual=4; IV=5; V=6; Long-term unemployed=7; Not currently employed=8; Father absent=9), `raven`'s test in year 1, student `id` number, `english` test score, `math`ematics test score and school `year` (coded 0, 1 and 2 for years one, two and three). So there are up to three measures per student. The data was obtained from the *Multilevel Models project*.

We shall take as our response the math test score result from the final year and try to model this as a function of gender, social class and the Raven's test score from the first year which might be taken as a measure of ability when entering the school. We subset the data to ignore the math scores from the first two years:

```
data(jsp, package="faraway")
jspr <- jsp[jsp$year==2,]
```

We start with two plots of the data. Due to the discreteness of the score results, it is helpful to *jitter* (add small random perturbations) the scores to avoid overprinting. The use of transparency, specified using the `alpha` parameter, also helps with dense data.

```
ggplot(jspr, aes(x=raven, y=math))+xlab("Raven Score")+ylab("Math
    ↪ Score")+geom_point(position = position_jitter(),alpha=0.3)
ggplot(jspr, aes(x=social, y=math))+xlab("Social Class")+ylab("Math
    ↪ Score")+geom_boxplot()
```

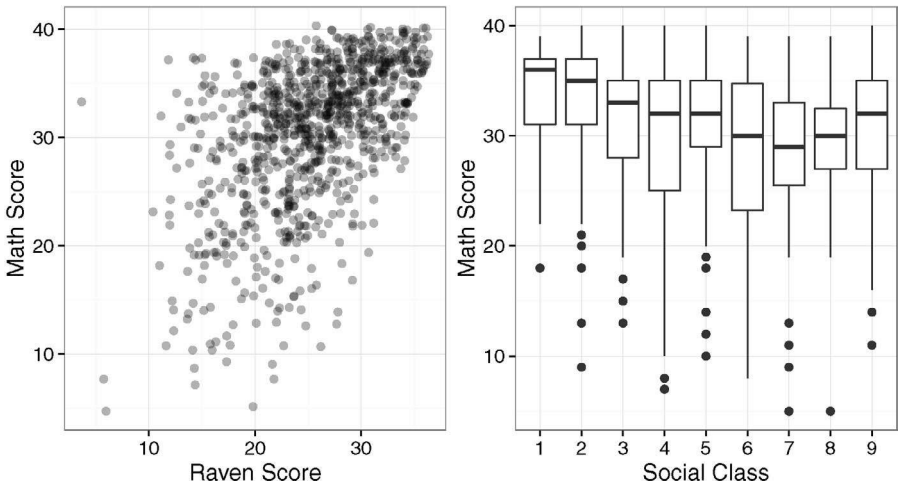


Figure 10.9 *Plots of the Junior School Project data.*

In Figure 10.9, we can see the positive correlation between the Raven's test score and the final math score. The maximum math score was 40, which reduces the variability at the upper end of the scale. We also see how the math scores tend to decline with social class.

One possible approach to analyzing these data is multiple regression. For example, we could fit:

```
glin <- lm(math ~ raven*gender*social,jspr)
anova(glin)
```

```
Analysis of Variance Table

Response: math
                     Df  Sum Sq Mean Sq F value Pr(>F)
raven                 1   11481   11481  368.06 <2e-16
gender                1      44      44    1.41 0.2347
social                8     779      97    3.12 0.0017
raven:gender          1 0.01145 0.01145 0.00037 0.9847
raven:social          8     583      73    2.33 0.0175
gender:social         8     450      56    1.80 0.0727
raven:gender:social   8     235      29    0.94 0.4824
Residuals           917   28603      31
```

It would seem that gender effects can be removed entirely, giving us:

```
glin <- lm(math ~ raven*social,jspr)
anova(glin)
Analysis of Variance Table

Response: math
             Df Sum Sq Mean Sq F value Pr(>F)
raven         1  11481   11481  365.72 <2e-16
social        8    778      97    3.10 0.0019
raven:social  8    564      71    2.25 0.0222
Residuals   935  29351      31
```

This is a fairly large dataset, so even small effects can be significant. Even though the `raven:social` term is significant at the 5% level, we remove it to simplify interpretation:

```
glin <- lm(math ~ raven+social,jspr)
summary(glin)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.0248     1.3745   12.39   <2e-16
raven         0.5804     0.0326   17.83   <2e-16
social2       0.0495     1.1294    0.04    0.965
social3      -0.4289     1.1957   -0.36    0.720
social4      -1.7745     1.0599   -1.67    0.094
social5      -0.7823     1.1892   -0.66    0.511
social6      -2.4937     1.2609   -1.98    0.048
social7      -3.0485     1.2907   -2.36    0.018
social8      -3.1175     1.7749   -1.76    0.079
social9      -0.6328     1.1273   -0.56    0.575

n = 953, p = 10, Residual SE = 5.632, R-Squared = 0.29
```

We see that the final math score is strongly related to the entering Raven score and that the math scores of the lower social classes are lower, even after adjustment for the entering score. Of course, any regression analysis requires more investigation than this; there are diagnostics and transformations to be considered and more. However, even if we were to do this, there would still be a problem with this analysis. We are assuming that the 953 students in the dataset are independent observations. This is not a tenable assumption as the students come from 50 different schools. The number coming from each school varies:

```
table(jspr$school)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
26 11 14 24 26 18 11 27 21  0 11 23 22 13  7 16  6 18 14 13 28
22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
```

```
14 18 21 14 20 22 15 13 27 35 23 44 27 16 28 17 12 14 10 10 41
44 45 46 47 48 49 50
 5 11 15 33 63 22 14
```

It is highly likely that students in the same school (and perhaps class) will show some dependence. So we have somewhat less than 953 independent cases worth of information. Any analysis that pretends these are independent is likely to overstate the significance of the results. Furthermore, the analysis above tells us nothing about the variation between and within schools. People will certainly be interested in this. We could aggregate the results across schools but this would lose information and expose us to the dangers of an ecological regression.

We need an analysis that uses the individual-level information, but also reflects the grouping in the data. Our first model has fixed effects representing all interactions between `raven`, `social` and `gender` with random effects for the school and the class nested within the school:

```
mmod <- lmer(math ~ raven*social*gender+(1|school)+(1|school:class),
    ↪ data=jspr)
```

A look at the summary output from this model suggests that gender may not be significant. We can test this using the Kenward-Roger adjusted F-test from the `pbkrtest` package:

```
mmodr <- lmer(math ~ raven*social+(1|school)+(1|school:class),  data=
    ↪ jspr)
KRmodcomp(mmod, mmodr)
```

```
F-test with Kenward-Roger approximation; computing time: 0.39 sec.
large : math ~ raven * social * gender + (1 | school) + (1 | school:class)
small : math ~ raven * social + (1 | school) + (1 | school:class)
        stat    ndf    ddf F.scaling p.value
Ftest   1.01  18.00 892.94         1    0.44
```

This can be verified using the parametric bootstrap although with a dataset of this size, it does take some time to run. The size of the dataset means that we can be quite confident about the adjusted F-test in any case.

In this example, we have more than a handful of potential models we might consider even if we vary only the fixed effect part of the model. In such circumstances, we might prefer to take a criterion-based approach to model selection. One approach is to specify all the models we wish to consider:

```
all3 <- lmer(math ~ raven*social*gender+(1|school)+(1|school:class),
    ↪ data=jspr, REML=FALSE)
all2 <- update(all3, . ~ . - raven:social:gender)
notrs <- update(all2, . ~ . -raven:social)
notrg <- update(all2, . ~ . -raven:gender)
notsg <- update(all2, . ~ . -social:gender)
onlyrs <- update(all2, . ~ . -social:gender - raven:gender)
all1 <-  update(all2, . ~ . -social:gender - raven:gender - social:
    ↪ raven)
nogen <- update(all1, . ~ . -gender)
```

It is important to use the ML method for constructing the AICs. As explained previously, it is not sensible to use the REML method when comparing models with different fixed effects. We have specified models with a three-way interaction, all two-way interactions, models leaving out each two-way interaction, a model excluding any interaction involving gender, a model with just main effects and finally a

model without gender entirely. Now we can create a table showing the AIC and BIC values:

```
anova(all3, all2, notrs, notrg, notsg, onlyrs, all1, nogen)[,1:4]
```

```
       Df  AIC  BIC logLik
all1   14 5956 6024  -2964
nogen  21 5949 6051  -2954
onlyrs 22 5950 6057  -2953
notrs  23 5962 6073  -2958
notsg  23 5952 6064  -2953
notrg  30 5956 6102  -2948
all2   31 5958 6108  -2948
all3   39 5967 6156  -2944
```

The `anova` output produces chi-squared tests for comparing the models. This is not correct here as the sequence of models is not nested and furthermore, these tests are inaccurate for reasons previously explained. We exclude this part of the output using `[,1:4]`. We can see that the AIC is minimized by the model that removes gender entirely. This confirms our hypothesis-testing based approach to selecting the model but rather more thoroughly by also considering the intermediate models.

The BIC criterion commonly prefers models that are smaller than the AIC. We see that illustrated in this example as BIC picks the model with only the main effects. We might reasonably add other models to the comparison. It becomes tedious to list all the possibilities when there are more variables but it requires some more complex R code to generate these automatically.

Given that we have decided that gender is not important, we simplify to:

```
jspr$craven <- jspr$raven-mean(jspr$raven)
mmod <- lmer(math ~ craven*social+(1|school)+(1|school:class),jspr)
sumary(mmod)
```

```
Fixed Effects:
              coef.est coef.se
(Intercept)    31.91     1.20
craven          0.61     0.19
social2         0.02     1.27
social3        -0.63     1.31
social4        -1.97     1.20
social5        -1.36     1.30
social6        -2.27     1.37
social7        -2.55     1.41
social8        -3.39     1.80
social9        -0.83     1.25
craven:social2 -0.13     0.21
craven:social3 -0.22     0.22
craven:social4  0.04     0.19
craven:social5 -0.15     0.21
craven:social6 -0.04     0.23
craven:social7  0.40     0.23
craven:social8  0.26     0.26
craven:social9 -0.08     0.21

Random Effects:
 Groups       Name        Std.Dev.
 school:class (Intercept) 1.08
 school       (Intercept) 1.77
 Residual                 5.21
```

```
---
number of obs: 953, groups: school:class, 90; school, 48
AIC = 5963.2, DIC = 5893.6
deviance = 5907.4
```

We centered the Raven score about its overall mean. This means that we can interpret the social effects as the predicted differences from social class one at the mean Raven score. If we did not do this, these parameter estimates would represent differences for `raven`=0 which is not very useful. We can see the math score is strongly related to the entering Raven score. We see that for the same entering score, the final math score tends to be lower as social class goes down. Note that class 9 here is when the father is absent and class 8 is not necessarily worse than 7, so this factor is not entirely ordinal. We also see the most substantial variation at the individual level with smaller amounts of variation at the school and class level.

We check the standard diagnostics first:

```
diagd <- fortify(mmod)
ggplot(diagd,aes(sample=.resid))+stat_qq()
ggplot(diagd,aes(x=.fitted,y=.resid)) +geom_point(alpha=0.3) +geom_
    ↪ hline(yintercept=0) +xlab("Fitted") +ylab("Residuals")
```

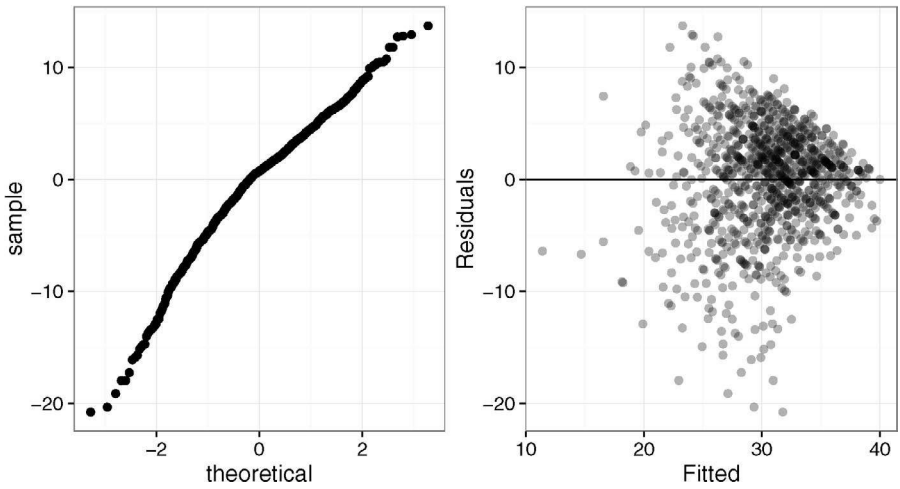


Figure 10.10 *Diagnostic plots for the Junior Schools Project model.*

In Figure 10.10, we see that the residuals are close to normal, but there is a clear decrease in the variance with an increase in the fitted values. This is due to the reduced variation in higher scores already observed. We might consider a transformation of the response to remove this effect.

We can also check the assumption of normally distributed random effects. We can do this at the school and class level:

```
qqnorm(ranef(mmod)$school[[1]],main="School effects")
qqnorm(ranef(mmod)$"school:class"[[1]],main="Class effects")
```

We see in Figure 10.11 that there is approximate normality in both cases with some

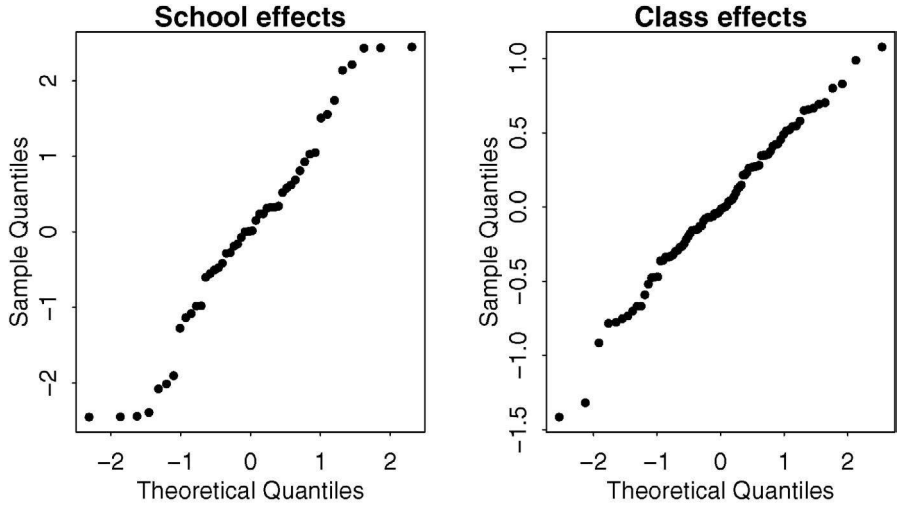


Figure 10.11 *QQ plots of the random effects at the school and class levels.*

evidence of short tails for the school effects. It is interesting to look at the sorted school effects:

```
adjscores <- ranef(mmod)$school[[1]]
```

These represent a ranking of the schools adjusted for the quality of the intake and the social class of the students. The difference between the best and the worst is about five points on the math test. Of course, we must recognize that there is variability in these estimated effects before making any decisions about the relative strengths of these schools. Compare this with an unadjusted ranking that simply takes the average score achieved by the school, centered by the overall average:

```
rawscores <- coef(lm(math ~ school-1,jspr))
rawscores <- rawscores-mean(rawscores)
```

We compare these two measures of school quality in Figure 10.12:

```
plot(rawscores,adjscores)
sint <- c(9,14,29)
text(rawscores[sint],adjscores[sint]+0.2,c("9","15","30"))
```

School 10 is listed but has no students, hence the need to adjust the labeling. There are some interesting differences. School 15 looks best on the raw scores but after adjustment, it drops to 15th place. This is a school that apparently performs well, but when the quality of the incoming students is considered, its performance is not so impressive. School 30 illustrates the other side of the coin. This school looks average on the raw scores, but is doing quite well given the ability of the incoming students. School 9 is actually doing a poor job despite raw scores that look quite good.

It is also worth plotting the residuals and the random effects against the predictors. We would be interested in finding any inhomogeneity or signs of structure that might lead to an improved model.

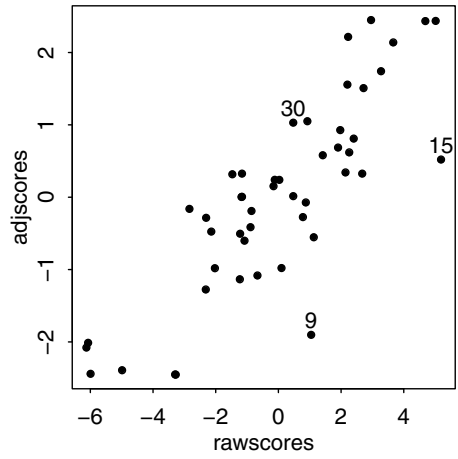


Figure 10.12 *Raw and adjusted school-quality measures. Three selected schools are marked.*

We may also be interested to know whether there really is much variation between schools or classes within schools. We can investigate this by testing the random effect terms using the `RLRsim` package. We need to fit models without each of the random effect terms.

```
library(RLRsim)
mmodc <- lmer(math ~ craven*social+(1|school:class),jspr)
mmods <- lmer(math ~ craven*social+(1|school),jspr)
```

We can test the class effect:

```
exactRLRT(mmodc, mmod, mmods)
```

```
        simulated finite sample distribution of RLRT.

        (p-value based on 10000 simulated values)

data:
RLRT = 2.3903, p-value = 0.0549
```

The evidence for a class effect is quite marginal. We would certainly choose to include it for testing fixed effect terms as we would rather be sure that it had been taken account of. Even so we can see that the class effect may be quite small. In contrast, we can test for a school effect:

```
exactRLRT(mmods, mmod, mmodc)
```

```
        simulated finite sample distribution of RLRT.

        (p-value based on 10000 simulated values)

data:
RLRT = 7.1403, p-value = 0.0033
```

The school effect comes through strongly. It seems schools matter more than specific teachers.

Compositional Effects: Fixed effect predictors in this example so far have been at the lowest level, the student, but it is not improbable that factors at the school or

class level might be important predictors of success in the math test. We can construct some such predictors from the individual-level information; such factors are called *compositional effects*. For example, the average entering score for a school might be an important predictor. The ability of one's fellow students may have an impact on future achievement. We construct this variable:

```
schraven <- lm(raven ~ school, jspr)$fit
```

and insert it into our model:

```
mmodc <- lmer(math ~ craven*social+schraven*social+(1|school)+  (1|
    ↪ school:class),jspr)
KRmodcomp(mmod, mmodc)
```

```
F-test with Kenward-Roger approximation; computing time: 0.16 sec.
large : math ~ craven * social + schraven * social + (1 | school) + (1 |
    school:class)
small : math ~ craven * social + (1 | school) + (1 | school:class)
        stat    ndf    ddf F.scaling p.value
Ftest   0.68   9.00 640.14     0.997    0.73
```

We see that this new effect is not significant. We are not constrained to taking means. We might consider various quantiles or measures of spread as potential compositional variables.

Much remains to be investigated with this dataset. We have only used the simplest of error structures and we should investigate whether the random effects may also depend on some of the other covariates.

Further Reading: The classical approach to random effects can be found in many older books such as Snedecor and Cochran (1989) or Scheffé (1959). More recent books such as Searle et al. (1992) also focus on the ANOVA approach. A wide range of models are explicitly considered in Milliken and Johnson (1992). Multilevel models are covered in Goldstein (1995), Raudenbush and Bryk (2002) and Gelman and Hill (2006). The predecessor to the `lme4` package was `nlme` which is described in Pinheiro and Bates (2000), but the book still contains much general material of interest.

Exercises

1. The `denim` dataset concerns the amount of waste in material cutting for a jeans manufacturer due to five suppliers.
   (a) Plot the data and comment.
   (b) Fit the linear fixed effects model. Is the operator significant?
   (c) Make a useful diagnostic plot for this model and comment.
   (d) Analyze the data with supplier as a random effect. What are the estimated standard deviations of the effects?
   (e) Test the significance of the supplier term.
   (f) Compute confidence intervals for the random effect SDs.
   (g) Locate two outliers and remove them from the data. Repeat the fitting, testing and computation of the confidence intervals, commenting on the differences you see from the complete data.

(h) Estimate the effect of each supplier. If only one supplier will be used, choose the best.

2. The `coagulation` dataset comes from a study of blood coagulation times. Twenty-four animals were randomly assigned to four different diets and the samples were taken in a random order.
   (a) Plot the data and comment.
   (b) Fit a fixed effects model and construct a prediction together with a 95% prediction interval for the response of a new animal assigned to diet D.
   (c) Now fit a random effects model using REML. A new animal is assigned to diet D. Predict the blood coagulation time for this animal along with a 95% prediction interval.
   (d) A new diet is given to a new animal. Predict the blood coagulation time for this animal along with a 95% prediction interval
   (e) A new diet is given to the first animal in the dataset. Predict the blood coagulation time for this animal with a prediction interval. You may assume that the effects of the initial diet for this animal have washed out.
3. The `eggprod` dataset concerns an experiment where six pullets were placed into each of 12 pens. Four blocks were formed from groups of three pens based on location. Three treatments were applied. The number of eggs produced was recorded.
   (a) Make suitable plots of the data and comment.
   (b) Fit a fixed effects model for the number of eggs produced with the treatments and blocks as predictors. Determine the significance of the two predictors and perform a basic diagnostic check.
   (c) Fit a model for the number of eggs produced with the treatments as fixed effects and the blocks as random effects. Which treatment is best in terms of maximizing production according to the model? Are you sure it is better than other two treatments?
   (d) Use the Kenward-Roger approximation for an F-test to check for differences between the treatments. How does the result compare to the fixed effects result?
   (e) Perform the same test but using a bootstrap method. How do the results compare?
   (f) Test for the significance of the blocks. Does the outcome agree with the fixed effects result?
4. Data on the cutoff times of lawnmowers may be found in the dataset `lawn`. Three machines were randomly selected from those produced by manufacturers A and B. Each machine was tested twice at low speed and high speed.
   (a) Make plots of the data and comment.
   (b) Fit a fixed effects model for the cutoff time response using just the main effects of the three predictors. Explain why not all effects can be estimated.

(c) Fit a mixed effects model with manufacturer and speed as main effects along with their interaction and machine as a random effect. If the same machine were tested at the same speed, what would be the SD of the times observed? If different machines were sampled from the same manufacturer and tested at the same speed once only, what would be the SD of the times observed?

(d) Test whether the interaction term of the model can be removed. If so, go on to test the two main fixed effects terms.

(e) Check whether there is any variation between machines.

(f) Fit a model with speed as the only fixed effect and manufacturer as a random effect with machines also as a random effect nested within manufacturer. Compare the variability between machines with the variability between manufacturers.

(g) Construct bootstrap confidence intervals for the terms of the previous model. Discuss whether the variability can be ascribed solely to manufacturers or to machines.

5. A number of growers supply broccoli to a food processing plant. The plant instructs the growers to pack the broccoli into standard-size boxes. There should be 18 clusters of broccoli per box. Because the growers use different varieties and methods of cultivation, there is some variation in the cluster weights. The plant manager selected three growers at random and then four boxes at random supplied by these growers. Three clusters were selected from each box. The data may be found in the `broccoli` dataset. The weight in grams of the cluster is given.

(a) Plot the data and comment on the nature of the variation seen.

(b) Compute the mean weights within growers. Compute the mean weights within boxes.

(c) Fit an appropriate mixed effects model. Comment on how the variation is assigned to the possible sources.

(d) Test whether there may be no variation attributable to growers.

(e) Test whether there may be no variation attributable to boxes.

(f) Compute confidence intervals for the SD components in your full model.

6. An experiment was conducted to select the supplier of raw materials for production of a component. The breaking strength of the component was the objective of interest. Four suppliers were considered. The four operators can only produce one component each per day. A latin square design is used and the data is presented in `breaking`.

(a) Plot the data and interpret.

(b) Fit a fixed effects model for the main effects. Determine which factors are significant.

(c) Fit a mixed effects model with operators and days as random effects but the suppliers as fixed effects. Why is this a natural choice of fixed and random effects? Which supplier results in the highest breaking point? What is the nature of the variation between operators and days?

(d) Test the operator and days effects.

(e) Test the significance of the supplier effect.

(f) For the best choice of supplier, predict the proportion of components produced in the future that will have a breaking strength less than 1000.

7. An experiment was conducted to optimize the manufacture of semiconductors. The `semicond` data has the resistance recorded on the wafer as the response. The experiment was conducted during four different time periods denoted by ET and three different wafers during each period. The position on the wafer is a factor with levels 1 to 4. The `Grp` variable is a combination of ET and wafer. Analyze the data as a split plot experiment where ET and position are considered as fixed effects. Since the wafers are different in experimental time periods, the `Grp` variable should be regarded as the block or group variable.

   (a) Plot the data appropriately and comment.

   (b) Fit a fixed effects model with an interaction between `ET` and `position` (no other predictors). What terms are significant? What is wrong with using this model to make inference about these predictors?

   (c) Fit a model appropriate to the split plot design used here. Comment on the relative variation between and within the groups (`Grp`).

   (d) Test for the effect of position.

   (e) Which level of `ET` results in the highest resistance? Can we be sure that this is really better than the second highest level?

   (f) Make a plot of the residuals and fitted values and interpret. Make a QQ plot and comment.

8. Redo the Junior Schools Project data analysis in the text with the final year English score as the response. Highlight any differences from the analysis of the final year Math scores.

9. An experiment was conducted to determine the effect of recipe and baking temperature on chocolate cake quality. Fifteen batches of cake mix for each recipe were prepared. Each batch was sufficient for six cakes. Each of the six cakes was baked at a different temperature which was randomly assigned. Several measures of cake quality were recorded of which breaking angle was just one. The dataset is presented as `choccake`.

   (a) Plot the data and comment.

   (b) Fit linear model with an interaction between recipe and temperature as fixed effects and no random effects. Which terms are significant? Why is this analysis unreliable?

   (c) Fit a mixed effects model that takes account of the batch structure, identifying the design type. Compare the temperature effect (minimum to maximum) with the likely difference between batches. How do they compare?

   (d) Test for a recipe effect.

   (e) Check the following diagnostic plots and comment.

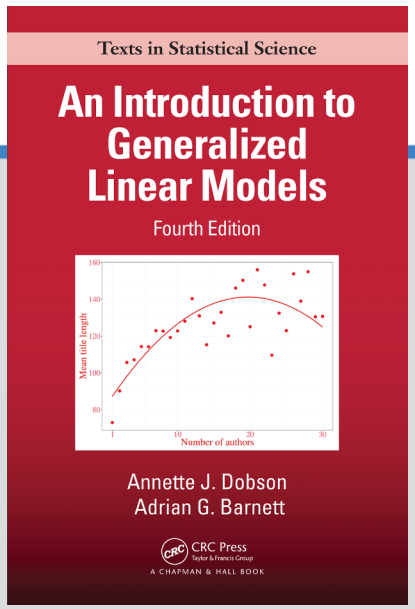
      i. The residuals against fitted values.

ii. A QQ plot of the residuals.
iii. A QQ plot of the batch random effects.

CHAPTER

5

POISSON REGRESSION AND LOG-LINEAR MODELS



This chapter is excerpted from

An Introduction to Generalized Linear Models, Fourth Edition

by Annette J. Dobson, Adrian G. Barnett.

Learn more

Chapter 9

Poisson Regression and Log-Linear Models

9.1 Introduction

The number of times an event occurs is a common form of data. Examples of **count** or **frequency** data include the number of tropical cyclones crossing the North Queensland coast (Section 1.6.5) or the numbers of people in each cell of a contingency table summarizing survey responses (e.g., satisfaction ratings for housing conditions, Exercise 8.2).

The **Poisson distribution** Po(μ) is often used to model count data. If Y is the number of occurrences, its probability distribution can be written as

$$f(y) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, 2, \ldots,$$

where μ is the average number of occurrences. It can be shown that $E(Y) = \mu$ and $\text{var}(Y) = \mu$ (see Exercise 3.4).

The parameter μ requires careful definition. Often it needs to be described as a rate; for example, the average number of customers who buy a particular product out of every 100 customers who enter the store. For motor vehicle crashes, the rate parameter may be defined in many different ways: crashes per 1,000 population, crashes per 1,000 licensed drivers, crashes per 1,000 motor vehicles, or crashes per 100,000 km travelled by motor vehicles. The time scale should be included in the definition; for example, the motor vehicle crash rate is usually specified as the rate per year (e.g., crashes per 100,000 km per year), while the rate of tropical cyclones refers to the cyclone season from November to April in Northeastern Australia. More generally, the rate is specified in terms of units of "exposure"; for instance, customers entering a store are "exposed" to the opportunity to buy the product of interest. For occupational injuries, each worker is exposed for the period he or she is at work, so the rate may be defined in terms of person-years "at risk."

The effect of explanatory variables on the response Y is modelled through the parameter μ. This chapter describes models for two situations.

In the first situation, the events relate to varying amounts of exposure which need to be taken into account when modelling the rate of events. **Poisson regression** is used in this case. The other explanatory variables (in addition to exposure) may be continuous or categorical.

In the second situation, exposure is constant (and therefore not relevant to the model) and the explanatory variables are usually categorical. If there are only a few explanatory variables the data are summarized in a cross-classified table. The response variable is the frequency or count in each cell of the table. The variables used to define the table are all treated as explanatory variables. The study design may mean that there are some constraints on the cell frequencies (for example, the totals for each row of the table may be equal), and these need to be taken into account in the modelling. The term **log-linear model**, which basically describes the role of the link function, is used for the generalized linear models appropriate for this situation.

The next section describes Poisson regression. A numerical example is used to illustrate the concepts and methods, including model checking and inference. Subsequent sections describe relationships between probability distributions for count data, constrained in various ways, and the log-linear models that can be used to analyze the data.

9.2 Poisson regression

Let $Y_1, \ldots, Y_N$ be independent random variables with Y_i denoting the number of events observed from exposure n_i for the ith covariate pattern. The expected value of Y_i can be written as

$$E(Y_i) = \mu_i = n_i \theta_i.$$

For example, suppose Y_i is the number of insurance claims for a particular make and model of car. This will depend on the number of cars of this type that are insured, n_i, and other variables that affect θ_i, such as the age of the cars and the location where they are used. The subscript i is used to denote the different combinations of make and model, age, location and so on.

The dependence of θ_i on the explanatory variables is usually modelled by

$$\theta_i = e^{\mathbf{x}_i^T \boldsymbol{\beta}}. \tag{9.1}$$

Therefore, the generalized linear model is

$$E(Y_i) = \mu_i = n_i e^{\mathbf{x}_i^T \boldsymbol{\beta}}; \quad Y_i \sim \text{Po}(\mu_i). \tag{9.2}$$

The natural link function is the logarithmic function

$$\log \mu_i = \log n_i + \mathbf{x}_i^T \boldsymbol{\beta}. \tag{9.3}$$

Equation (9.3) differs from the usual specification of the linear component due to the inclusion of the term $\log n_i$. This term is called the **offset**. It is a known constant, which is readily incorporated into the estimation procedure. As usual, the terms $\mathbf{x}_i$ and $\boldsymbol{\beta}$ describe the covariate pattern and parameters, respectively.

For a binary explanatory variable denoted by an indictor variable, $x_j = 0$ if the factor is absent and $x_j = 1$ if it is present. The **rate ratio**, *RR*, for presence vs. absence is

$$RR = \frac{\mathrm{E}(Y_i \mid present)}{\mathrm{E}(Y_i \mid absent)} = e^{\beta_j}$$

from (9.1), provided all the other explanatory variables remain the same. Similarly, for a continuous explanatory variable x_k, a one-unit increase will result in a multiplicative effect of e^{β_k} on the rate μ. Therefore, parameter estimates are often interpreted on the exponential scale e^{β} in terms of ratios of rates.

Hypotheses about the parameters β_j can be tested using the Wald, score or likelihood ratio statistics. Confidence intervals can be estimated similarly. For example, for parameter β_j

$$\frac{b_j - \beta_j}{s.e.(b_j)} \sim \mathrm{N}(0,1) \tag{9.4}$$

approximately. Alternatively, hypothesis testing can be performed by comparing the goodness of fit of appropriately defined nested models (see Chapter 4).

The fitted values are given by

$$\widehat{Y}_i = \widehat{\mu}_i = n_i e^{\mathbf{x}_i^T \mathbf{b}}, \quad i = 1, \ldots, N.$$

These are often denoted by e_i because they are estimates of the expected values $\mathrm{E}(Y_i) = \mu_i$. As $\mathrm{var}(Y_i) = \mathrm{E}(Y_i)$ for the Poisson distribution, the standard error of Y_i is estimated by $\sqrt{e_i}$ so the **Pearson residuals** are

$$r_i = \frac{o_i - e_i}{\sqrt{e_i}}, \tag{9.5}$$

where o_i denotes the observed value of Y_i. As outlined in Section 6.2.6, these residuals may be further refined to

$$r_{pi} = \frac{o_i - e_i}{\sqrt{e_i}\sqrt{1 - h_i}},$$

where the leverage, h_i, is the ith element on the diagonal of the hat matrix.

For the Poisson distribution, the residuals given by (9.5) and the chi-squared goodness of fit statistic are related by

$$X^2 = \sum r_i^2 = \sum \frac{(o_i - e_i)^2}{e_i},$$

which is the usual definition of the chi-squared statistic for contingency tables.

The deviance for a Poisson model is given in Section 5.6.3. It can be written in the form

$$D = 2\sum \left[o_i \log(o_i/e_i) - (o_i - e_i) \right]. \tag{9.6}$$

However, for most models $\sum o_i = \sum e_i$ (see Exercise 9.1), so the deviance simplifies to

$$D = 2\sum \left[o_i \log(o_i/e_i) \right]. \tag{9.7}$$

The **deviance residuals** are the components of D in (9.6),

$$d_i = \text{sign}(o_i - e_i)\sqrt{2\left[o_i \log(o_i/e_i) - (o_i - e_i) \right]}, \quad i = 1, \ldots, N \tag{9.8}$$

so that $D = \sum d_i^2$.

The goodness of fit statistics X^2 and D are closely related. Using the Taylor series expansion given in Section 7.5,

$$o \log\left(\frac{o}{e}\right) = (o - e) + \tfrac{1}{2}\frac{(o-e)^2}{e} + \ldots$$

so that, approximately, from (9.6)

$$\begin{aligned} D &= 2\sum_{i=1}^{N} \left[(o_i - e_i) + \tfrac{1}{2}\frac{(o_i - e_i)^2}{e_i} - (o_i - e_i) \right] \\ &= \sum_{i=1}^{N} \frac{(o_i - e_i)^2}{e_i} = X^2. \end{aligned}$$

The statistics D and X^2 can be used directly as measures of goodness of fit, as they can be calculated from the data and the fitted model (because they do not involve any nuisance parameters like σ^2 for the Normal distribution). They can be compared with the central chi-squared distribution with $N - p$ degrees of freedom, where p is the number of parameters that are estimated. The chi-squared distribution is likely to be a better approximation for

the sampling distribution of X^2 than for the sampling distribution of D (see Section 7.5).

Two other summary statistics provided by some software are the likelihood ratio chi-squared statistic and pseudo R^2. These are based on comparisons between the maximum value of the log-likelihood function for a minimal model with the same rate parameter β_1 for all Y_i's and no covariates, $\log \mu_i = \log n_i + \beta_1$, and the maximum value of the log-likelihood function for Model (9.3) with p parameters. The likelihood ratio chi-squared statistic $C = 2[l(\mathbf{b}) - l(\mathbf{b}_{\min})]$ provides an overall test of the hypotheses that $\beta_2 = \ldots = \beta_p = 0$, by comparison with the central chi-squared distribution with $p-1$ degrees of freedom (see Exercise 7.4). Less formally, pseudo $R^2 = [l(\mathbf{b}_{\min}) - l(\mathbf{b})]/l(\mathbf{b}_{\min})$ provides an intuitive measure of fit.

Other diagnostics, such as delta-betas and related statistics, are also available for Poisson models.

9.2.1 Example of Poisson regression: British doctors' smoking and coronary death

The data in Table 9.1 are from a famous study conducted by Sir Richard Doll and colleagues. In 1951, all British doctors were sent a brief questionnaire about whether they smoked tobacco. Since then information about their deaths has been collected. Table 9.1 shows the numbers of deaths from coronary heart disease among male doctors 10 years after the survey. It also shows the total number of person-years of observation at the time of the analysis (Breslow and Day, 1987: Appendix 1A and page 112).

Table 9.1 *Deaths from coronary heart disease after 10 years among British male doctors categorized by age and smoking status in 1951.*

Age	Smokers		Non-smokers	
group	Deaths	Person-years	Deaths	Person-years
35–44	32	52407	2	18790
45–54	104	43248	12	10673
55–64	206	28612	28	5710
65–74	186	12663	28	2585
75–84	102	5317	31	1462

The questions of interest are

1. Is the death rate higher for smokers than non-smokers?
2. If so, by how much?

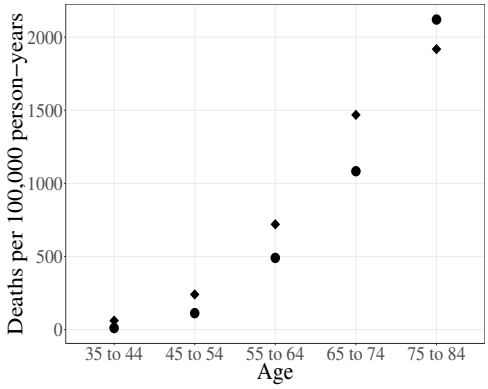


Figure 9.1 *Deaths rates from coronary heart disease per 100,000 person-years for smokers (diamonds) and non-smokers (dots).*

3. Is the differential effect related to age?

Figure 9.1 shows the death rates per 100,000 person-years from coronary heart disease for smokers and non-smokers. It is clear that the rates increase with age but more steeply than in a straight line. Death rates among smokers appear to be generally higher than among non-smokers but they do not rise as rapidly with age. Various models can be specified to describe these data well. One model, in the form of (9.3) is

$$\begin{aligned} \log(deaths_i) &= \log(personyears_i) + \beta_1 + \beta_2 smoke_i + \beta_3 agecat_i \\ &\quad + \beta_4 agesq_i + \beta_5 smkage_i \end{aligned} \tag{9.9}$$

where the subscript i denotes the ith subgroup defined by age group and smoking status ($i = 1,\ldots,5$ for ages 35–44,...,75–84 for smokers and $i = 6,\ldots,10$ for the corresponding age groups for non-smokers). The term $deaths_i$ denotes the expected number of deaths and $personyears_i$ denotes the number of doctors at risk and the observation periods in group i. For the other terms, $smoke_i$ is equal to 1 for smokers and 0 for non-smokers; $agecat_i$ takes the values $1,\ldots,5$ for age groups 35–44,...,75–84; $agesq_i$ is the square of $agecat_i$ to take account of the non-linearity of the rate of increase; and $smkage_i$ is equal to $agecat_i$ for smokers and 0 for non-smokers; thus, describing a differential rate of increase with age.

In Stata, Model (9.9) can be fitted using either of the following commands

Stata code (Poisson regression models)

```
.poisson deaths agecat agesq smoke smkage, exposure(personyears)
```

```
.glm deaths agecat agesq smoke smkage, family(poisson) link(log)
 lnoffset(personyears)
```

The option `irr` can be used to obtain the rate ratios and 95% confidence limits.

The corresponding command for R is

R code (Poisson regression model)

```
>data(doctors)
>res.doc<-glm(deaths~age + agesq + smoking + smoking:age +
 offset(log(personyears)),family=poisson(),data=doctors)
```

Table 9.2 shows the parameter estimates in the form of rate ratios $e^{\widehat{\beta}_j}$. The Wald statistics (9.4) to test $\beta_j = 0$ all have very small p-values and the 95% confidence intervals for e^{β_j} do not contain unity showing that all the terms are needed in the model. The estimates show that the risk of coronary deaths was, on average, about 4 times higher for smokers than non-smokers (based on the rate ratio for *smoke*), after the effect of age is taken into account. However, the effect is attenuated as age increases (coefficient for *smkage*). Table 9.3 shows that the model fits the data very well; the expected numbers of deaths estimated from (9.9) are quite similar to the observed numbers of deaths, and so the Pearson residuals calculated from (9.5) and deviance residuals from (9.8) are very small.

Table 9.2 *Parameter estimates obtained by fitting Model (9.9) to the data in Table 9.1.*

Term	*agecat*	*agesq*	*smoke*	*smkage*
$\widehat{\beta}$	2.376	−0.198	1.441	−0.308
$s.e.(\widehat{\beta})$	0.208	0.027	0.372	0.097
Wald statistic	11.43	−7.22	3.87	−3.17
p-value	< 0.001	< 0.001	< 0.001	0.002
Rate ratio	10.77	0.82	4.22	0.74
95% confidence interval	7.2, 16.2	0.78, 0.87	2.04, 8.76	0.61, 0.89

To obtain these results from Stata after the command `poisson` use `predict fit` and then calculate the residuals, or use `poisgof` to obtain the deviance statistic D or `poisgof, pearson` to obtain the statistic X^2. Alternatively, after `glm` use `predict fit; predict d, deviance;` and `predict c, pearson`.

The R commands are as follows

R code (Poisson model residuals)

Table 9.3 *Observed and estimated expected numbers of deaths and residuals for the model described in Table 9.2.*

Age category	Smoking category	Observed deaths	Expected deaths	Pearson residual	Deviance residual
1	1	32	29.58	0.444	0.438
2	1	104	106.81	−0.272	−0.273
3	1	206	208.20	−0.152	−0.153
4	1	186	182.83	0.235	0.234
5	1	102	102.58	−0.057	−0.057
1	0	2	3.41	−0.766	−0.830
2	0	12	11.54	0.135	0.134
3	0	28	27.74	0.655	0.641
4	0	28	30.23	−0.405	−0.411
5	0	31	31.07	−0.013	−0.013
Sum of squares*				1.550	1.635

* Calculated from residuals correct to more significant figures than shown here.

```
>fit_p=c(fitted(res.doc))
>pearsonresid<-(doctors$deaths-fit_p)/sqrt(fit_p)
>chisq<-sum(pearsonresid*pearsonresid)
>devres<-sign(doctors$deaths-fit_p)*(sqrt(2*(doctors$deaths*
 log(doctors$deaths/fit_p)-(doctors$deaths-fit_p))))
>deviance<-sum(devres*devres)
```

For the minimal model, with only the parameter β_1, the maximum value for the log-likelihood function is $l(b_{\min}) = -495.067$. The corresponding value for Model (9.9) is $l(\mathbf{b}) = -28.352$. Therefore, an overall test of the model (testing $\beta_j = 0$ for $j = 2,\ldots,5$) is $C = 2[l(\mathbf{b}) - l(b_{\min})] = 933.43$ which is highly statistically significant compared with the chi-squared distribution with 4 degrees of freedom. The *pseudo* R^2 value is 0.94, or 94%, which suggests a good fit. More formal tests of the goodness of fit are provided by the statistics $X^2 = 1.550$ and $D = 1.635$ which are small compared with the chi-squared distribution with $N - p = 10 - 5 = 5$ degree of freedom.

9.3 Examples of contingency tables

Before specifying log-linear models for frequency data summarized in contingency tables, it is important to consider how the design of the study may determine constraints on the data. The study design also affects the choice

of probability models to describe the data. These issues are illustrated in the following three examples.

9.3.1 Example: Cross-sectional study of malignant melanoma

These data are from a cross-sectional study of patients with a form of skin cancer called malignant melanoma (Roberts et al. 1981). For a sample of $n = 400$ patients, the site of the tumor and its histological type were recorded. The data, numbers of patients with each combination of tumor type and site, are given in Table 9.4.

Table 9.4 *Malignant melanoma: frequencies for tumor type and site (Roberts et al. 1981).*

	Site			
Tumor type	Head & neck	Trunk	Extrem -ities	Total
Hutchinson's melanotic freckle	22	2	10	34
Superficial spreading melanoma	16	54	115	185
Nodular	19	33	73	125
Indeterminate	11	17	28	56
Total	68	106	226	400

The question of interest is whether there is any association between tumor type and site. Table 9.5 shows the data displayed as percentages of row and column totals. It appears that Hutchinson's melanotic freckle is more common on the head and neck but there is little evidence of association between other tumor types and sites.

Let Y_{jk} denote the frequency for the (j,k)th cell with $j = 1,\ldots,J$ and $k = 1,\ldots,K$. In this example, there are $J = 4$ rows, $K = 3$ columns and the constraint that $\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{jk} = n$, where $n = 400$ is fixed by the design of the study. If the Y_{jk}'s are independent random variables with Poisson distributions with parameters $\mathrm{E}(Y_{jk}) = \mu_{jk}$, then their sum has the Poisson distribution with parameter $\mathrm{E}(n) = \mu = \sum\sum \mu_{jk}$. Hence, the joint probability distribution of the Y_{jk}'s, conditional on their sum n, is the Multinomial distribution

$$f(\mathbf{y}|n) = n! \prod_{j=1}^{J}\prod_{k=1}^{K} \theta_{jk}^{y_{jk}} \big/ y_{jk}!\,,$$

where $\theta_{jk} = \mu_{jk}/\mu$. This result is derived in Section 8.2. The sum of the terms θ_{jk} is unity because $\sum\sum \mu_{jk} = \mu$; also $0 < \theta_k < 1$. Thus, θ_{jk} can be interpreted

Table 9.5 *Malignant melanoma: row and column percentages for tumor type and site.*

	Site			
Tumor type	Head & neck	Trunk	Extrem -ities	Total
Row percentages				
Hutchinson's melanotic freckle	64.7	5.9	29.4	100
Superficial spreading melanoma	8.6	29.2	62.2	100
Nodular	15.2	26.4	58.4	100
Indeterminate	19.6	30.4	50.0	100
All types	17.0	26.5	56.5	100
Column percentages				
Hutchinson's melanotic freckle	32.4	1.9	4.4	8.50
Superficial spreading melanoma	23.5	50.9	50.9	46.25
Nodular	27.9	31.1	32.3	31.25
Indeterminate	16.2	16.0	12.4	14.00
All types	100.0	99.9	100.0	100.0

as the probability of an observation in the (j,k)th cell of the table. Also the expected value of Y_{jk} is

$$\mathrm{E}(Y_{jk}) = \mu_{jk} = n\theta_{jk}.$$

The usual link function for a Poisson model gives

$$\log \mu_{jk} = \log n + \log \theta_{jk},$$

which is like Equation (9.3), except that the term $\log n$ is the same for all the Y_{jk}'s.

9.3.2 Example: Randomized controlled trial of influenza vaccine

In a prospective study of a new living attenuated recombinant vaccine for influenza, patients were randomly allocated to two groups, one of which was given the new vaccine and the other a saline placebo. The responses were titre levels of hemagglutinin inhibiting antibody found in the blood six weeks after vaccination; they were categorized as "small," "medium" or "large." The cell frequencies in the rows of Table 9.6 are constrained to add to the number of subjects in each treatment group (35 and 38, respectively). We want to know if the pattern of responses is the same for each treatment group.

Table 9.6 *Flu vaccine trial.*

	Response			
	Small	Moderate	Large	Total
Placebo	25	8	5	38
Vaccine	6	18	11	35

(Data from R.S. Gillett, personal communication, 1992)

In this example the row totals are fixed. Thus, the joint probability distribution for each row is Multinomial

$$f(y_{j1}, y_{j2}, \ldots, y_{jK} \,|\, y_{j.}) = y_{j.}! \prod_{k=1}^{K} \theta_{jk}^{y_{jk}} / y_{jk}!,$$

where $y_{j.} = \sum_{k=1}^{K} y_{jk}$ is the row total and $\sum_{k=1}^{K} \theta_{jk} = 1$. So the joint probability distribution for all the cells in the table is the **product multinomial distribution**

$$f(\mathbf{y} \,|\, y_{1.}, y_{2.}, \ldots, y_{J.}) = \prod_{j=1}^{J} y_{j.}! \prod_{k=1}^{K} \theta_{jk}^{y_{jk}} / y_{jk}!,$$

where $\sum_{k=1}^{K} \theta_{jk} = 1$ for each row. In this case $E(Y_{jk}) = y_{j.}\theta_{jk}$ so that

$$\log E(Y_{jk}) = \log \mu_{jk} = \log y_{j.} + \log \theta_{jk}.$$

If the response pattern was the same for both groups, then $\theta_{jk} = \theta_{.k}$ for $k = 1, \ldots, K$.

9.3.3 Example: Case–control study of gastric and duodenal ulcers and aspirin use

In this retrospective case–control study, a group of ulcer patients was compared with a group of control patients not known to have peptic ulcer, but who were similar to the ulcer patients with respect to age, sex and socioeconomic status (Duggan et al. 1986). The ulcer patients were classified according to the site of the ulcer: gastric or duodenal. Aspirin use was ascertained for all subjects. The results are shown in Table 9.7.

This is a $2 \times 2 \times 2$ contingency table. Some questions of interest are

1. Is gastric ulcer associated with aspirin use?
2. Is duodenal ulcer associated with aspirin use?

Table 9.7 *Gastric and duodenal ulcers and aspirin use: frequencies (Duggan et al. 1986).*

	Aspirin use		
	Non-user	User	Total
Gastric ulcer			
Control	62	6	68
Cases	39	25	64
Duodenal ulcer			
Control	53	8	61
Cases	49	8	57

Table 9.8 *Gastric and duodenal ulcers and aspirin use: row percentages for the data in Table 9.7.*

	Aspirin use		
	Non-user	User	Total
Gastric ulcer			
Control	91	9	100
Cases	61	39	100
Duodenal ulcer			
Control	87	13	100
Cases	86	14	100

3. Is any association with aspirin use the same for both ulcer sites?

   When the data are presented as percentages of row totals (Table 9.8), it appears that aspirin use is more common among ulcer patients than among controls for gastric ulcer but not for duodenal ulcer.

In this example, the numbers of patients with each type of ulcer and the numbers in each of the groups of controls, that is, the four row totals in Table 9.7, were all fixed.

Let $j = 1$ or 2 denote the controls or cases, respectively; $k = 1$ or 2 denote gastric ulcers or duodenal ulcers, respectively; and $l = 1$ for patients who did not use aspirin and $l = 2$ for those who did. In general, let Y_{jkl} denote the frequency of observations in category (j,k,l) with $j = 1,\ldots,J$, $k = 1,\ldots,K$ and $l = 1,\ldots,L$. If the marginal totals $y_{jk.}$ are fixed, the joint probability distribution for the Y_{jkl}'s is

$$f(\mathbf{y}\,|y_{11.},\ldots,y_{JK.}) = \prod_{j=1}^{J}\prod_{k=1}^{K} y_{jk.}! \prod_{l=1}^{L} \theta_{jkl}^{y_{jkl}} / y_{jkl}!,$$

where $\mathbf{y}$ is the vector of Y_{jkl}'s and $\sum_l \theta_{jkl} = 1$ for $j = 1,\ldots,J$ and $k = 1,\ldots,K$. This is another form of **product multinomial distribution**. In this case, $\mathrm{E}(Y_{jkl}) = \mu_{jkl} = y_{jk.}\theta_{jkl}$, so that

$$\log \mu_{jkl} = \log y_{jk.} + \log \theta_{jkl}.$$

9.4 Probability models for contingency tables

The examples in Section 9.3 illustrate the main probability models for contingency table data. In general, let the vector $\mathbf{y}$ denote the frequencies Y_i in N cells of a cross-classified table.

9.4.1 Poisson model

If there were no constraints on the Y_i's, they could be modelled as independent random variables with the parameters $\mathrm{E}(Y_i) = \mu_i$ and joint probability distribution

$$f(\mathbf{y};\boldsymbol{\mu}) = \prod_{i=1}^{N} \mu_i^{y_i} e^{-\mu_i}/y_i!,$$

where $\boldsymbol{\mu}$ is a vector of μ_i's.

9.4.2 Multinomial model

If the only constraint is that the sum of the Y_i's is n, then the following Multinomial distribution may be used

$$f(\mathbf{y};\boldsymbol{\mu}\,|n) = n!\prod_{i=1}^{N} \theta_i^{y_i}/y_i!,$$

where $\sum_{i=1}^{N} \theta_i = 1$ and $\sum_{i=1}^{N} y_i = n$. In this case, $\mathrm{E}(Y_i) = n\theta_i$.

For a two-dimensional contingency table (such as Table 9.4 for the melanoma data), if j and k denote the rows and columns, then the most commonly considered hypothesis is that the row and column variables are independent so that

$$\theta_{jk} = \theta_{j.}\theta_{.k},$$

where $\theta_{j.}$ and $\theta_{.k}$ denote the marginal probabilities with $\sum_j \theta_{j.} = 1$ and $\sum_k \theta_{.k} = 1$. This hypothesis can be tested by comparing the fit of two linear models for the logarithm of $\mu_{jk} = \mathrm{E}(Y_{jk})$; namely

$$\log \mu_{jk} = \log n + \log \theta_{jk}$$

and

$$\log \mu_{jk} = \log n + \log \theta_{j.} + \log \theta_{.k}.$$

9.4.3 Product multinomial models

If there are more fixed marginal totals than just the overall total n, then appropriate products of multinomial distributions can be used to model the data.

For example, for a three-dimensional table with J rows, K columns and L layers, if the row totals are fixed in each layer, the joint probability for the Y_{jkl}'s is

$$f(\mathbf{y}|y_{j.l},\, j=1,\ldots,J,\, l=1,\ldots,L) = \prod_{j=1}^{J}\prod_{l=1}^{L} y_{j.l}! \prod_{k=1}^{K} \theta_{jkl}^{y_{jkl}} / y_{jkl}!,$$

where $\sum_k \theta_{jkl} = 1$ for each combination of j and l. In this case, $\mathrm{E}(Y_{jkl}) = y_{j.l}\theta_{jkl}$.

If only the layer totals are fixed, then

$$f(\mathbf{y}|y_{..l},\, l=1,\ldots,L) = \prod_{l=1}^{L} y_{..l}! \prod_{j=1}^{J}\prod_{k=1}^{K} \theta_{jkl}^{y_{jkl}} / y_{jkl}!$$

with $\sum_j \sum_k \theta_{jkl} = 1$ for $l = 1,\ldots,L$. Also $\mathrm{E}(Y_{jkl}) = y_{..l}\theta_{jkl}$.

9.5 Log-linear models

All the probability models given in Section 9.4 are based on the Poisson distribution and in all cases $\mathrm{E}(Y_i)$ can be written as a product of parameters and other terms. Thus, the natural link function for the Poisson distribution, the logarithmic function, yields a linear component

$$\log \mathrm{E}(Y_i) = \text{constant} + \mathbf{x}_i^T \boldsymbol{\beta}.$$

The term **log-linear model** is used to describe all these generalized linear models.

For the melanoma Example 9.3.1, if there are no associations between site and type of tumor so that these two variables are independent, their joint probability θ_{jk} is the product of the marginal probabilities

$$\theta_{jk} = \theta_{j.}\theta_{.k}, \quad j = 1,\ldots,J \text{ and } k = 1,\ldots,K.$$

The hypothesis of independence can be tested by comparing the additive model (on the logarithmic scale)

$$\log \mathrm{E}(Y_{jk}) = \log n + \log \theta_{j.} + \log \theta_{.k} \tag{9.10}$$

with the model

$$\log \mathrm{E}(Y_{jk}) = \log n + \log \theta_{jk}. \tag{9.11}$$

This is analogous to analysis of variance for a two-factor experiment without replication (see Section 6.4.2). Equation (9.11) can be written as the saturated model

$$\log \mathrm{E}(Y_{jk}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk},$$

and Equation (9.10) can be written as the additive model

$$\log \mathrm{E}(Y_{jk}) = \mu + \alpha_j + \beta_k.$$

Since the term $\log n$ has to be in all models, the minimal model is

$$\log \mathrm{E}(Y_{jk}) = \mu.$$

For the flu vaccine trial, Example 9.3.2, $\mathrm{E}(Y_{jk}) = y_{j.}\theta_{jk}$ if the distribution of responses described by the θ_{jk}'s differs for the j groups, or $\mathrm{E}(Y_{jk}) = y_{j.}\theta_{.k}$ if it is the same for all groups. So the hypothesis of **homogeneity** of the response distributions can be tested by comparing the model

$$\log \mathrm{E}(Y_{jk}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk},$$

corresponding to $\mathrm{E}(Y_{jk}) = y_{j.}\theta_{jk}$, and the model

$$\log \mathrm{E}(Y_{jk}) = \mu + \alpha_j + \beta_k,$$

corresponding to $\mathrm{E}(Y_{jk}) = y_{j.}\theta_{.k}$. The minimal model for these data is

$$\log \mathrm{E}(Y_{jk}) = \mu + \alpha_j$$

because the row totals, corresponding to the subscript j, are fixed by the design of the study.

More generally, the specification of the linear components for log-linear models bears many resemblances to the specification for ANOVA models. The models are **hierarchical**, meaning that if a higher-order (interaction) term is included in the model, then all the related lower-order terms are also included. Thus, if the two-way (first-order) interaction $(\alpha\beta)_{jk}$ is included, then so are the main effects α_j and β_k and the constant μ. Similarly, if second-order interactions $(\alpha\beta\gamma)_{jkl}$ are included, then so are the first-order interactions $(\alpha\beta)_{jk}$, $(\alpha\gamma)_{jl}$ and $(\beta\gamma)_{kl}$.

If log-linear models are specified analogously to ANOVA models, they

include too many parameters, so sum-to-zero or corner-point constraints are needed. Interpretation of the parameters is usually simpler if reference or corner-point categories are identified so that parameter estimates describe effects for other categories relative to the reference categories.

For contingency tables the main questions almost always relate to associations between variables. Therefore, in log-linear models, the terms of primary interest are the interactions involving two or more variables.

9.6 Inference for log-linear models

Although three types of probability distributions are used to describe contingency table data (see Section 9.4), Birch (1963) showed that for any log-linear model the maximum likelihood estimators are the same for all these distributions provided that the parameters which correspond to the fixed marginal totals are always included in the model. This means that for the purpose of estimation, the Poisson distribution can always be assumed. As the Poisson distribution belongs to the exponential family and the parameter constraints can be incorporated into the linear component, all the standard methods for generalized linear models can be used.

The adequacy of a model can be assessed using the goodness of fit statistics X^2 or D (and sometimes C and pseudo R^2) summarized in Section 9.2 for Poisson regression. More insight into model adequacy can often be obtained by examining the Pearson or deviance residuals given by Equations (9.5) and (9.8), respectively. Hypothesis tests can be conducted by comparing the difference in goodness of fit statistics between a general model corresponding to an alternative hypothesis and a nested, simpler model corresponding to a null hypothesis.

These methods are illustrated in the following examples.

9.7 Numerical examples

9.7.1 Cross-sectional study of malignant melanoma

For the data in Table 9.4 the question of interest is whether there is an association between tumor type and site. This can be examined by testing the null hypothesis that the variables are independent.

The conventional chi-squared test of independence for a two-dimensional table is performed by calculating expected frequencies for each cell based on the marginal totals, $e_{jk} = y_{j.}y_{.k}/n$, calculating the chi-squared statistic $X^2 = \sum_j \sum_k (y_{jk} - e_{jk})^2 / e_{jk}$ and comparing this with the central chi-squared

distribution with $(J-1)(K-1)$ degrees of freedom. The observed and expected frequencies are shown in Table 9.9. These give

$$X^2 = \frac{(22-5.78)^2}{5.78} + \ldots + \frac{(28-31.64)^2}{31.64} = 65.8.$$

The value $X^2 = 65.8$ is very significant compared with the $\chi^2(6)$ distribution. Examination of the observed frequencies y_{jk} and expected frequencies e_{jk} shows that Hutchinson's melanotic freckle is more common on the head and neck than would be expected if site and type were independent.

Table 9.9 *Conventional chi-squared test of independence for melanoma data in Table 9.4; expected frequencies are shown in brackets.*

	Site			
Tumor type	Head & Neck	Trunk	Extrem -ities	Total
Hutchinson's melanotic freckle	22 (5.78)	2 (9.01)	10 (19.21)	34
Superficial spreading melanoma	16 (31.45)	54 (49.03)	115 (104.52)	185
Nodular	19 (21.25)	33 (33.13)	73 (70.62)	125
Indeterminate	11 (9.52)	17 (14.84)	28 (31.64)	56
Total	68	106	226	400

The corresponding analysis using log-linear models involves fitting the additive Model (9.10) corresponding to the hypothesis of independence. The saturated Model (9.11) and the minimal model with only a term for the mean effect are also fitted for illustrative purposes. In Stata the commands for the three models are

Stata code (log-linear models)

```
.xi:glm frequency i.tumor i.site i.tumor*i.site, family(poisson)
 link(log)
.xi:glm frequency i.tumor i.site, family(poisson) link(log)

.glm frequency, family(poisson) link(log)
```

The corresponding commands in R (`site` and `tumor` should be text variables)

R code (log-linear models)

```
>ressat.melanoma<-glm(frequency~site*tumor,family=poisson(),
 data=melanoma)
```

Table 9.10 *Log-linear models for the melanoma data in Table 9.4; coefficients, b, with standard errors in brackets.*

Term*	Saturated Model (9.10)	Additive Model (9.9)	Minimal model
Constant	3.091 (0.213)	1.754 (0.204)	3.507 (0.05)
SSM	−0.318 (0.329)	1.694 (0.187)	
NOD	−0.147 (0.313)	1.302 (0.193)	
IND	−0.693 (0.369)	0.499 (0.217)	
TNK	−2.398 (0.739)	0.444 (0.155)	
EXT	−0.788 (0.381)	1.201 (0.138)	
$SSM*TNK$	3.614 (0.792)		
$SSM*EXT$	2.761 (0.465)		
$NOD*TNK$	2.950 (0.793)		
$NOD*EXT$	2.134 (0.460)		
$IND*TNK$	2.833 (0.834)		
$IND*EXT$	1.723 (0.522)		
log-likelihood	−29.556	−55.453	−177.16
X^2	0.0	65.813	
D	0.0	51.795	

*Reference categories are Hutchinson's melanotic freckle (HMF) and head and neck (HNK). Other categories are for type, superficial spreading melanoma (SSM), nodular (NOD) and indeterminate (IND), and for site, trunk (TNK) and extremities (EXT).

```
>resadd.melanoma<-glm(frequency~site + tumor,family=poisson(),
 data=melanoma)
>resmin.melanoma<-glm(frequency~1, family=poisson(),
data=melanoma)
```

The results for all three models are shown in Table 9.10. For the reference category of Hutchinson's melanotic freckle (HMF) on the head or neck (HNK), the expected frequencies are as follows:

minimal model: $e^{3.507} = 33.35$;

additive model: $e^{1.754} = 5.78$, as in Table 9.9;

saturated model: $e^{3.091} = 22$, equal to observed frequency.

For indeterminate tumors (IND) in the extremities (EXT), the expected frequencies are

minimal model: $e^{3.507} = 33.35$;

additive model: $e^{1.754+0.499+1.201} = 31.64$, as in Table 9.9;

saturated model: $e^{3.091-0.693-0.788+1.723} = 28$, equal to observed frequency.

The saturated model with 12 parameters fits the 12 data points exactly. The additive model corresponds to the conventional analysis. The deviance for the additive model can be calculated from the sum of squares of the deviance residuals given by (9.8), or from twice the difference between the maximum values of the log-likelihood function for this model and the saturated model, $\triangle D = 2[-29.556 - (-55.453)] = 51.79$.

For this example, the conventional chi-squared test for independence and log-linear modelling produce exactly the same results. The advantage of log-linear modelling is that it provides a method for analyzing more complicated cross-tabulated data as illustrated by the next example.

9.7.2 Case–control study of gastric and duodenal ulcer and aspirin use

Preliminary analysis of the 2×2 tables for gastric ulcer and duodenal ulcer separately suggests that aspirin use may be a risk factor for gastric ulcer but not for duodenal ulcer. For analysis of the full data set, Table 9.7, the main effects for case–control status (*CC*), ulcer site (*GD*) and the interaction between these terms ($CC \times GD$) have to be included in all models (as these correspond to the fixed marginal totals). Table 9.11 shows the results of fitting this and several more complex models involving aspirin use (*AP*).

Table 9.11 *Results of log-linear modelling of data in Table 9.7.*

Terms in model	d.f.*	Deviance
$GD + CC + GD \times CC$	4	126.708
$GD + CC + GD \times CC + AP$	3	21.789
$GD + CC + GD \times CC + AP + AP \times CC$	2	10.538
$GD + CC + GD \times CC + AP + AP \times CC + AP \times GD$	1	6.283

*d.f. denotes degrees of freedom = number of observations (8) minus number of parameters

In Stata the commands are

Stata code (log-linear models)

```
.xi:glm frequency i.GD i.CC i.GD*i.CC, family(poisson) link(log)
.xi:glm frequency i.GD i.CC i.GD*i.CC i.AP, family(poisson)
 link(log)
```

```
.xi:glm frequency i.GD i.CC i.GD*i.CC i.AP i.AP*i.CC,
 family(poisson) link(log)
.xi:glm frequency i.GD i.CC i.GD*i.CC i.AP i.AP*i.CC i.AP*i.GD,
 family(poisson) link(log)
```

In R the corresponding commands are

R code (log-linear models)

```
>data(ulcer)
>res1.aspirin<-glm(frequency~GD + CC + GD*CC, family=poisson(),
 data=ulcer)
>res2.aspirin<-glm(frequency~GD + CC + GD*CC + AP,
 family=poisson(), data=ulcer)
>res3.aspirin<-glm(frequency~GD + CC + GD*CC + AP + AP*CC,
 family=poisson(), data=ulcer)
>res4.aspirin<-glm(frequency~GD + CC + GD*CC + AP + AP*CC +
 AP*GD, family=poisson(), data=ulcer)
```

The comparison of aspirin use between cases and controls can be summarized by the deviance difference for the second and third rows of Table 9.11, $\triangle D = 11.25$. This value is statistically significant compared with the $\chi^2(1)$ distribution, suggesting that aspirin is a risk factor for ulcers. Comparison between the third and fourth rows of the table, $\triangle D = 4.26$, provides only weak evidence of a difference between ulcer sites, possibly due to the lack of statistical power (p-value = 0.04 from the distribution $\chi^2(1)$).

The fit of the model with all three two-way interactions is shown in Table 9.12. The goodness of fit statistics for this table are $X^2 = 6.49$ and $D = 6.28$, which suggest that the model is not particularly good (compared with the $\chi^2(1)$ distribution) even though $p = 7$ parameters have been used to describe $N = 8$ data points.

9.8 Remarks

Two issues relevant to the analysis of a count data have not yet been discussed in this chapter.

First, **overdispersion** occurs when $\text{var}(Y_i)$ is greater than $\text{E}(Y_i)$, although $\text{var}(Y_i) = \text{E}(Y_i)$ for the Poisson distribution. The **negative binomial distribution** provides an alternative model with $\text{var}(Y_i) = \phi\ \text{E}(Y_i)$, where $\phi > 1$ is a parameter that can be estimated. Overdispersion can be due to lack of independence between the observations, in which case the methods described in Chapter 11 for correlated data can be used.

Second, contingency tables may include cells which cannot have any observations (e.g., male hysterectomy cases). This phenomenon, termed **structural zeros** may not be easily incorporated in Poisson regression unless the parameters can be specified to accommodate the situation. Alternative approaches are discussed by Agresti (2013) and Hilbe (2014).

9.9 Exercises

9.1 Let $Y_1, \ldots, Y_N$ be independent random variables with $Y_i \sim \text{Po}(\mu_i)$ and $\log \mu_i = \beta_1 + \sum_{j=2}^{J} x_{ij}\beta_j,\ i = 1, \ldots, N$.

a. Show that the score statistic for β_1 is $U_1 = \sum_{i=1}^{N}(Y_i - \mu_i)$.

b. Hence, show that for maximum likelihood estimates $\widehat{\mu}_i$, $\sum \widehat{\mu}_i = \sum y_i$.

c. Deduce that the expression for the deviance in (9.6) simplifies to (9.7) in this case.

9.2 The data in Table 9.13 are numbers of insurance policies, n, and numbers of claims, y, for cars in various insurance categories, *CAR*, tabulated by age of policy holder, *AGE*, and district where the policy holder lived ($DIST = 1$, for London and other major cities, and $DIST = 0$, otherwise). The table is derived from the *CLAIMS* data set in Aitkin et al. (2005) obtained from a paper by Baxter et al. (1980).

a. Calculate the rate of claims y/n for each category and plot the rates by *AGE*, *CAR* and *DIST* to get an idea of the main effects of these factors.

b. Use Poisson regression to estimate the main effects (each treated as categorical and modelled using indicator variables) and interaction terms.

c. Based on the modelling in (b), Aitkin et al. (2005) determined that all

Table 9.12 *Comparison of observed frequencies and expected frequencies obtained from the log-linear model with all two-way interaction terms for the data in Table 9.7; expected frequencies in brackets.*

	Aspirin use		
	Non-user	User	Total
Gastric ulcer			
Controls	62 (58.53)	6 (9.47)	68
Cases	39 (42.47)	25 (21.53)	64
Duodenal ulcer			
Controls	53 (56.47)	8 (4.53)	61
Cases	49 (45.53)	8 (11.47)	57

the interactions were unimportant and decided that *AGE* and *CAR* could be treated as though they were continuous variables. Fit a model incorporating these features and compare it with the best model obtained in (b). What conclusions do you reach?

Table 9.13 *Car insurance claims: based on the CLAIMS data set reported by Aitkin et al. (2005).*

		$DIST = 0$		$DIST = 1$	
CAR	*AGE*	y	n	y	n
1	1	65	317	2	20
1	2	65	476	5	33
1	3	52	486	4	40
1	4	310	3259	36	316
2	1	98	486	7	31
2	2	159	1004	10	81
2	3	175	1355	22	122
2	4	877	7660	102	724
3	1	41	223	5	18
3	2	117	539	7	39
3	3	137	697	16	68
3	4	477	3442	63	344
4	1	11	40	0	3
4	2	35	148	6	16
4	3	39	214	8	25
4	4	167	1019	33	114

9.3 This question relates to the flu vaccine trial data in Table 9.6.

a. Using a conventional chi-squared test and an appropriate log-linear model, test the hypothesis that the distribution of responses is the same for the placebo and vaccine groups.

b. For the model corresponding to the hypothesis of homogeneity of response distributions, calculate the fitted values, the Pearson and deviance residuals, and the goodness of fit statistics X^2 and D. Which of the cells of the table contribute most to X^2 (or D)? Explain and interpret these results.

c. Re-analyze these data using ordinal logistic regression to estimate cutpoints for a latent continuous response variable and to estimate a location shift between the two treatment groups. Sketch a rough diagram to

illustrate the model which forms the conceptual base for this analysis (see Exercise 8.4).

9.4 For a 2×2 contingency table, the maximal log-linear model can be written as

$$\begin{aligned} \eta_{11} &= \mu + \alpha + \beta + (\alpha\beta), \quad \eta_{12} = \mu + \alpha - \beta - (\alpha\beta), \\ \eta_{21} &= \mu - \alpha + \beta - (\alpha\beta), \quad \eta_{22} = \mu - \alpha - \beta + (\alpha\beta), \end{aligned}$$

where $\eta_{jk} = \log\mathrm{E}(Y_{jk}) = \log(n\theta_{jk})$ and $n = \sum\sum Y_{jk}$.
Show that the interaction term $(\alpha\beta)$ is given by

$$(\alpha\beta) = \tfrac{1}{4}\log\phi,$$

where ϕ is the **odds ratio** $(\theta_{11}\theta_{22})/(\theta_{12}\theta_{21})$, and hence that $\phi = 1$ corresponds to no interaction.

9.5 Use log-linear models to examine the housing satisfaction data in Table 8.5. The numbers of people surveyed in each type of housing can be regarded as fixed.

a. First, analyze the associations between level of satisfaction (treated as a nominal categorical variable) and contact with other residents, separately for each type of housing.

b. Next, conduct the analyses in (a) simultaneously for all types of housing.

c. Compare the results from log-linear modelling with those obtained using nominal or ordinal logistic regression (see Exercise 8.2).

9.6 Consider a $2 \times K$ contingency table (Table 9.14) in which the column totals $y_{.k}$ are fixed for $k = 1, \ldots, K$.

Table 9.14 *Contingency table with 2 rows and K columns.*

	1	...	k	...	K
Success	y_{11}		y_{1k}		y_{1K}
Failure	y_{21}		y_{2k}		y_{2K}
Total	$y_{.1}$		$y_{.k}$		$y_{.K}$

a. Show that the product multinomial distribution for this table reduces to

$$f(z_1, \ldots, z_K | n_1, \ldots, n_K) = \sum_{k=1}^{K} \binom{n_k}{z_k} \pi_k^{z_k} (1 - \pi_k)^{n_k - z_k},$$

where $n_k = y_{.k}, z_k = y_{1k}, n_k - z_k = y_{2k}, \pi_k = \theta_{1k}$ and $1 - \pi_k = \theta_{2k}$ for $k = 1, \ldots, K$. This is the **product binomial distribution** and is the joint distribution for Table 7.1 (with appropriate changes in notation).

b. Show that the log-linear model with

$$\eta_{1k} = \log \mathrm{E}(Z_k) = \mathbf{x}_{1k}^T \boldsymbol{\beta}$$

and

$$\eta_{2k} = \log \mathrm{E}(n_k - Z_k) = \mathbf{x}_{2k}^T \boldsymbol{\beta}$$

is equivalent to the logistic model

$$\log\left(\frac{\pi_k}{1-\pi_k}\right) = \mathbf{x}_k^T \boldsymbol{\beta},$$

where $\mathbf{x}_k = \mathbf{x}_{1k} - \mathbf{x}_{2k}$, $k = 1, \ldots, K$.

c. Based on (b), analyze the case–control study data on aspirin use and ulcers using logistic regression and compare the results with those obtained using log-linear models.

9.7 Mittlbock and Heinzl (2001) compare Poisson and logistic regression models for data in which the event rate is small so that the Poisson distribution provides a reasonable approximation to the Binomial distribution. An example is the number of deaths from coronary heart disease among British doctors (Table 9.1). In Section 9.2.1 we fitted the model $Y_i \sim \mathrm{Po}(deaths_i)$ with Equation (9.9)

$$\begin{aligned} \log(deaths_i) &= \log(personyears_i) + \beta_1 + \beta_2 smoke_i + \beta_3 agecat_i + \\ &\quad \beta_4 agesq_i + \beta_5 smkage_i. \end{aligned}$$

An alternative is $Y_i \sim Bin(personyears_i, \pi_i)$ with

$$\mathrm{logit}(\pi_i) = \beta_1 + \beta_2 smoke_i + \beta_3 agecat_i + \beta_4 agesq_i + \beta_5 smkage_i.$$

Another version is based on a Bernoulli distribution $Z_j \sim B(\pi_i)$ for each doctor in group i with

$$Z_j = \begin{cases} 1, & j = 1, \ldots, deaths_i \\ 0, & j = deaths_i + 1, \ldots, personyears_i \end{cases}$$

and

$$\mathrm{logit}(\pi_i) = \beta_1 + \beta_2 smoke_i + \beta_3 agecat_i + \beta_4 agesq_i + \beta_5 smkage_i.$$

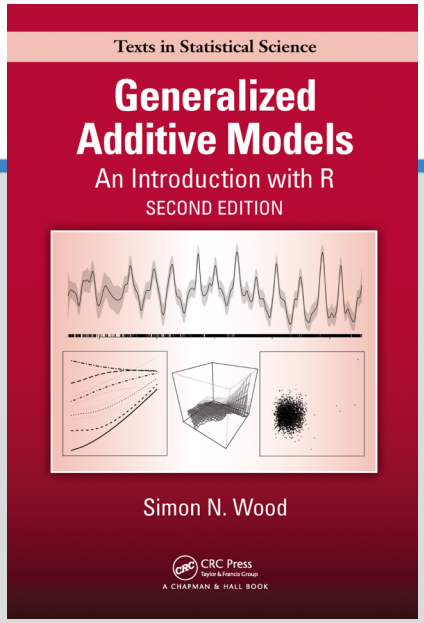
a. Fit all three models (in Stata the Bernoulli model cannot be fitted with `glm`; use `blogit` instead). Verify that the β estimates are very similar.

b. Calculate the statistics D, X^2 and pseudo R^2 for all three models. Notice that the pseudo R^2 is much smaller for the Bernoulli model. As Mittlbock and Heinzl (2001) point out this is because the Poisson and Binomial models are estimating the probability of death for each group (which is relatively easy) whereas the Bernoulli model is estimating the probability of death for an individual (which is much more difficult).

CHAPTER

6

INTRODUCING GAMS



Chapter 12

Insurance Redlining — A Complete Example

In this chapter, we present a relatively complete data analysis. The example is interesting because it illustrates several of the ambiguities and difficulties encountered in statistical practice.

Insurance redlining refers to the practice of refusing to issue insurance to certain types of people or within some geographic area. The name comes from the act of drawing a red line around an area on a map. Now few would quibble with an insurance company refusing to sell auto insurance to a frequent drunk driver, but other forms of discrimination would be unacceptable.

In the late 1970s, the US Commission on Civil Rights examined charges by several Chicago community organizations that insurance companies were redlining their neighborhoods. Because comprehensive information about individuals being refused homeowners insurance was not available, the number of FAIR plan policies written and renewed in Chicago by zip code for the months of December 1977 through May 1978 was recorded. The FAIR plan was offered by the city of Chicago as a default policy to homeowners who had been rejected by the voluntary market. Information on other variables that might affect insurance writing such as fire and theft rates was also collected at the zip code level. The variables are:

`race` racial composition in percentage of minority

`fire` fires per 100 housing units

`theft` thefts per 1000 population

`age` percentage of housing units built before 1939

`involact` new FAIR plan policies and renewals per 100 housing units

`income` median family income in thousands of dollars

`side` north or south side of Chicago

The data come from Andrews and Herzberg (1985) where more details of the variables and the background are provided.

12.1 Ecological Correlation

Notice that we do not know the races of those denied insurance. We only know the racial composition in the corresponding zip code. This is an important difficulty that needs to be considered before starting the analysis.

When data are collected at the group level, we may observe a correlation between two variables. The ecological fallacy is concluding that the same correlation holds

at the individual level. For example, in countries with higher fat intakes in the diet, higher rates of breast cancer have been observed. Does this imply that individuals with high fat intakes are at a higher risk of breast cancer? Not necessarily. Relationships seen in observational data are subject to confounding, but even if this is allowed for, bias is caused by aggregating data. We consider an example taken from US demographic data:

```
> data(eco, package="faraway")
> plot(income ~ usborn, data=eco, xlab="Proportion US born", ylab="
    Mean Annual Income")
```

In the first panel of Figure 12.1, we see the relationship between 1998 per capita income dollars from all sources and the proportion of legal state residents born in the United States in 1990 for each of the 50 states plus the District of Columbia (D.C.). We can see a clear negative correlation. We can fit a regression line and show the

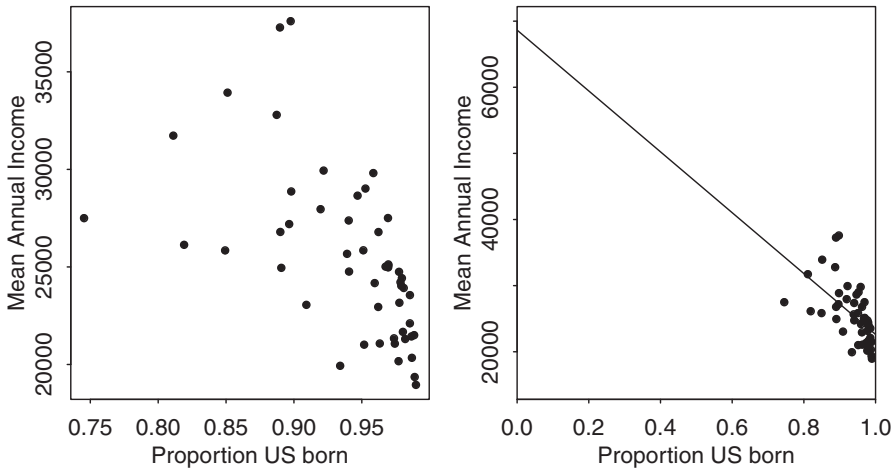


Figure 12.1 1998 annual per capita income and proportion US born for 50 states plus D.C. The plot on the right shows the same data as on the left, but with an extended scale and the least squares fit shown.

fitted line on an extended range:

```
> lmod <- lm(income ~ usborn, eco)
> sumary(lmod)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    68642       8739    7.85  3.2e-10
usborn        -46019       9279   -4.96  8.9e-06

n = 51, p = 2, Residual SE = 3489.541, R-Squared = 0.33
> plot(income ~ usborn, data=eco, xlab="Proportion US born", ylab="
    Mean Annual Income",xlim=c(0,1),ylim=c(15000,70000),xaxs="i")
> abline(coef(lmod))
```

We see that there is a clear statistically significant relationship between the per capita

annual income and the proportion who are US born. What does this say about the average annual income of people who are US born and those who are naturalized citizens? If we substitute `usborn=1` into the regression equation, we get 68642 − 46019 = $22,623, while if we put `usborn=0`, we get $68,642. This suggests that on average, naturalized citizens earn three times more than US born citizens. In truth, information from the US Bureau of the Census indicates that US born citizens have an average income just slightly larger than naturalized citizens. What went wrong with our analysis?

The ecological inference from the aggregate data to the individuals requires an assumption of constancy. Explicitly, the assumption would be that the incomes of the native born do not depend on the proportion of native born within the state (and similarly for naturalized citizens). This assumption is unreasonable for these data because immigrants are naturally attracted to wealthier states.

This assumption is also relevant to the analysis of the Chicago insurance data since we have only aggregate data. We must keep in mind that the results for the aggregated data may not hold true at the individual level.

12.2 Initial Data Analysis

Start by reading the data in and examining it:

```
> data(chredlin, package="faraway")
> head(chredlin)
      race fire theft  age involact income side
60626 10.0  6.2    29 60.4       0.0 11.744    n
60640 22.2  9.5    44 76.5       0.1  9.323    n
60613 19.6 10.5    36 73.5       1.2  9.948    n
60657 17.3  7.7    37 66.9       0.5 10.656    n
60614 24.5  8.6    53 81.4       0.7  9.730    n
60610 54.0 34.1    68 52.6       0.3  8.231    n
```

Summarize:

```
> summary(chredlin)
      race             fire            theft             age
 Min.   : 1.00   Min.   : 2.00   Min.   :  3.0   Min.   : 2.0
 1st Qu.: 3.75   1st Qu.: 5.65   1st Qu.: 22.0   1st Qu.:48.6
 Median :24.50   Median :10.40   Median : 29.0   Median :65.0
 Mean   :34.99   Mean   :12.28   Mean   : 32.4   Mean   :60.3
 3rd Qu.:57.65   3rd Qu.:16.05   3rd Qu.: 38.0   3rd Qu.:77.3
 Max.   :99.70   Max.   :39.70   Max.   :147.0   Max.   :90.1
    involact         income      side
 Min.   :0.000   Min.   : 5.58   n:25
 1st Qu.:0.000   1st Qu.: 8.45   s:22
 Median :0.400   Median :10.69
 Mean   :0.615   Mean   :10.70
 3rd Qu.:0.900   3rd Qu.:11.99
 Max.   :2.200   Max.   :21.48
```

We see that there is a wide range in the `race` variable, with some zip codes almost entirely minority or non-minority. This is good for our analysis since it will reduce the variation in the regression coefficient for race, allowing us to assess this effect

more accurately. If all the zip codes were homogeneous, we would never be able to discover an effect from these aggregated data. We also note some skewness in the `theft` and `income` variables. The response `involact` has a large number of zeros. This is not good for the assumptions of the linear model but we have little choice but to proceed. We will not use the information about north versus south side until later. Now make some graphical summaries:

```
> require(ggplot2)
> ggplot(chredlin,aes(race,involact)) + geom_point() +stat_smooth(
    method="lm")
> ggplot(chredlin,aes(fire,involact)) + geom_point() +stat_smooth(
    method="lm")
> ggplot(chredlin,aes(theft,involact)) + geom_point() +stat_smooth(
    method="lm")
> ggplot(chredlin,aes(age,involact)) + geom_point() +stat_smooth(
    method="lm")
> ggplot(chredlin,aes(income,involact)) + geom_point() +stat_smooth(
    method="lm")
> ggplot(chredlin,aes(side,involact)) + geom_point(position = position
    _jitter(width = .2,height=0))
```

The plots are seen in Figure 12.2. We have superimposed a linear fit to each pair of variables with a 95% confidence band shown in grey. Strong relationships can be seen in several of the plots. We can see some outlier and influential points. We can also see that the fitted line sometimes goes below zero which is problematic since observed values of the response cannot be negative. Jittering has been added in the final plot to avoid overplotting of symbols. Let's focus on the relationship between `involact` and `race`:

```
> sumary(lm(involact ~ race,chredlin))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12922    0.09661    1.34     0.19
race         0.01388    0.00203    6.84  1.8e-08

n = 47, p = 2, Residual SE = 0.449, R-Squared = 0.51
```

We can clearly see that homeowners in zip codes with a high percentage of minorities are taking the default FAIR plan insurance at a higher rate than other zip codes. That is not in doubt. However, can the insurance companies claim that the discrepancy is due to greater risks in some zip codes? The insurance companies could claim that they were denying insurance in neighborhoods where they had sustained large fire-related losses and any discriminatory effect was a by-product of legitimate business practice. We plot some of the variables involved by this question in Figure 12.3.

```
> ggplot(chredlin,aes(race,fire)) + geom_point() +stat_smooth(method="
    lm")
> ggplot(chredlin,aes(race,theft)) + geom_point() +stat_smooth(method
    ="lm")
```

We can see that there is indeed a relationship between the fire rate and the percentage of minorities. We also see that there is large outlier that may have a disproportionate effect on the relationship between the theft rate and the percentage of minorities.

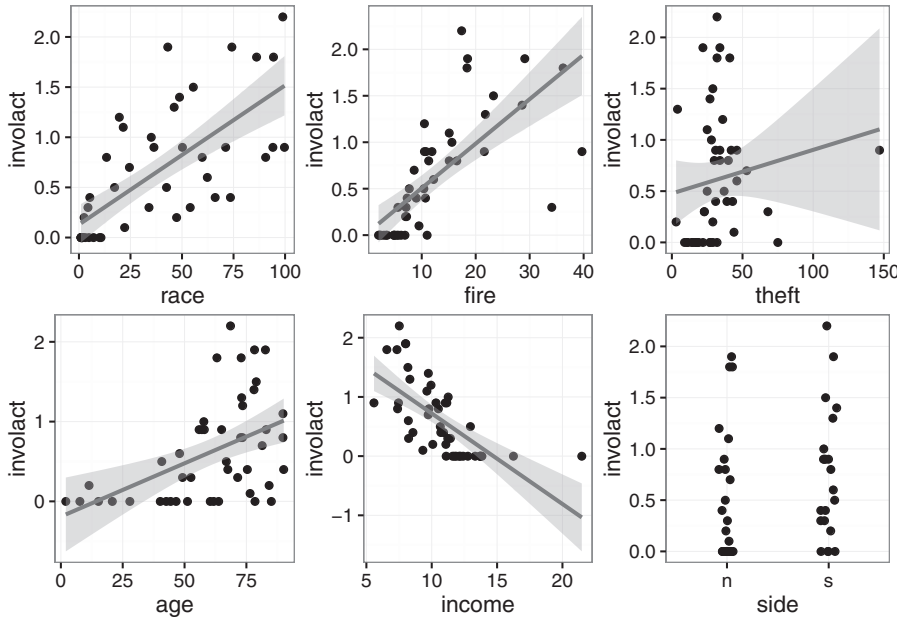


Figure 12.2 Plots of the Chicago insurance data.

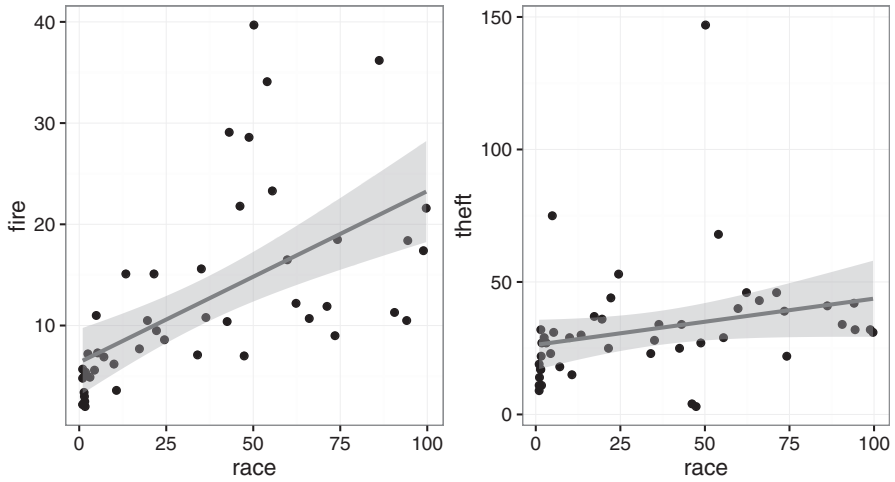


Figure 12.3 Relationship between `fire`, `theft` and `race` in the Chicago data.

The question of which variables should also be included in the regression so that their effect may be adjusted for is difficult. Statistically, we can do it, but the important question is whether it should be done at all. For example, it is known that the incomes of women in the United States and other countries are generally lower than those of men. However, if one adjusts for various predictors such as type of job and length of service, this gender difference is reduced or can even disappear. The controversy is not statistical but political — should these predictors be used to make the adjustment?

For the present data, suppose that the effect of adjusting for income differences was to remove the race effect. This would pose an interesting, but non-statistical question. I have chosen to include the `income` variable in the analysis just to see what happens.

I have decided to use `log(income)` partly because of skewness in this variable, but also because income is better considered on a multiplicative rather than additive scale. In other words, $1,000 is worth a lot more to a poor person than a millionaire because $1,000 is a much greater fraction of the poor person's wealth.

12.3 Full Model and Diagnostics

We start with the full model:

```
> lmod <- lm(involact ~ race + fire + theft + age + log(income),
    chredlin)
> sumary(lmod)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.18554    1.10025   -1.08  0.28755
race         0.00950    0.00249    3.82  0.00045
fire         0.03986    0.00877    4.55 0.000048
theft       -0.01029    0.00282   -3.65  0.00073
age          0.00834    0.00274    3.04  0.00413
log(income)  0.34576    0.40012    0.86  0.39254

n = 47, p = 6, Residual SE = 0.335, R-Squared = 0.75
```

Before leaping to any conclusions, we should check the model assumptions. These two diagnostic plots are seen in Figure 12.4:

```
> plot(lmod,1:2)
```

The diagonal streak in the residual-fitted plot is caused by the large number of zero response values in the data. When $y = 0$, the residual $\hat{\varepsilon} = -\hat{y} = -x^T\hat{\beta}$, hence the line. Turning a blind eye to this feature, we see no particular problem. The Q–Q plot looks fine too. This is reassuring since we know from the form of the response with so many zero values, that it cannot possibly be normally distributed. We'll rely on the central limit theorem, the size of the sample and lack of long-tailed or skewed residuals to be comfortable with the reported p-values.

We now look for transformations. We try some partial residual plots as seen in Figure 12.5:

```
> termplot(lmod, partial.resid=TRUE, terms=1:2)
```

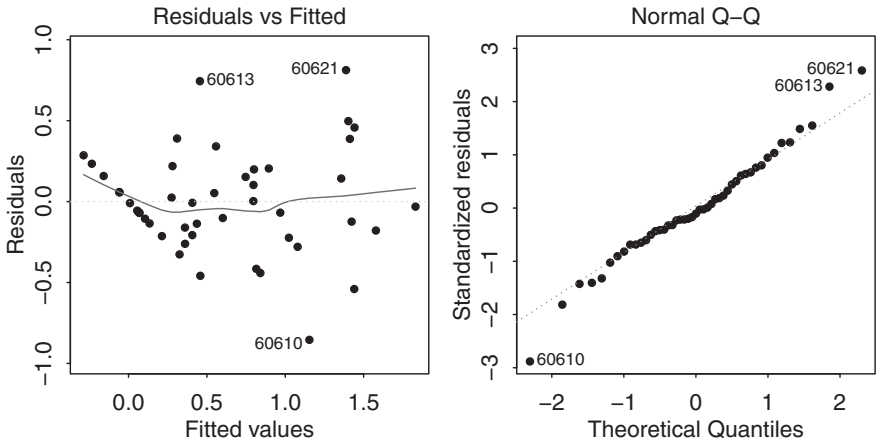


Figure 12.4 Diagnostic plots of the initial model for the Chicago insurance data.

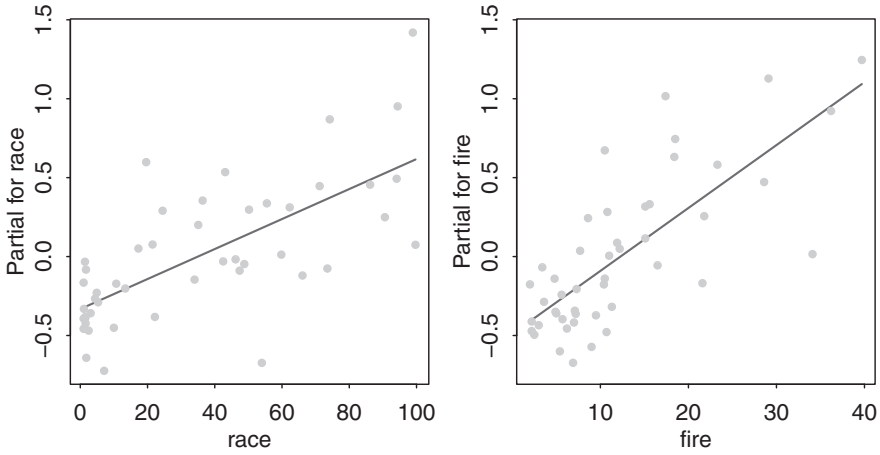


Figure 12.5 Partial residual plots for `race` and `fire`.

These plots indicate no need to transform. It would have been inconvenient to transform the `race` variable since that would have made interpretation more difficult. Fortunately, we do not need to worry about this. We examined the other partial residual plots and experimented with polynomials for the predictors. No transformation of the predictors appears to be worthwhile.

We choose to avoid a transformation of the response. The zeros in the response would have restricted the possibilities and furthermore would have made interpretation more difficult. A square root transformation is possible but whatever slim advantage this might offer, it makes explanation more problematic.

12.4 Sensitivity Analysis

How robust is our conclusion to the choice of covariates used to adjust the response? In the full model used earlier, we used all four covariates but we may wonder how sensitive our findings are to this choice. Certainly, one might question whether we should adjust the response for the average income of the zip code. Other objections or uncertainties might be raised by use of the other covariates also.

We can investigate these concerns by fitting other models that vary the choice of adjusting covariates. In this example, there are four such covariates and so there are only 16 possible combinations in which they may be added to the model. It is practical to fit and examine all these models.

The mechanism for creating all 16 models is rather complex and you may wish to skip to the output. The first line creates all subsets of $(1,2,3,4)$. The second line creates the predictor part of the model formulae by pasting together the chosen variables. We then iterate over all 16 models, saving the terms of interest for the `race` variable:

```
> listcombo <- unlist(sapply(0:4,function(x) combn(4, x, simplify=
    FALSE)),recursive=FALSE)
> predterms <- lapply(listcombo, function(x) paste(c("race",c("fire","
    theft","age","log(income)")[x]),collapse="+"))
> coefm <- matrix(NA,16,2)
> for(i in 1:16){
   lmi <- lm(as.formula(paste("involact ~ ",predterms[[i]])), data=
       chredlin)
   coefm[i,] <- summary(lmi)$coef[2,c(1,4)]
}
> rownames(coefm) <- predterms
> colnames(coefm) <- c("beta","pvalue")
> round(coefm,4)
                                 beta pvalue
race                           0.0139 0.0000
race+fire                      0.0089 0.0002
race+theft                     0.0141 0.0000
race+age                       0.0123 0.0000
race+log(income)               0.0082 0.0087
race+fire+theft                0.0082 0.0002
race+fire+age                  0.0089 0.0001
race+fire+log(income)          0.0070 0.0160
race+theft+age                 0.0128 0.0000
race+theft+log(income)         0.0084 0.0083
race+age+log(income)           0.0099 0.0017
race+fire+theft+age            0.0081 0.0001
race+fire+theft+log(income)    0.0073 0.0078
race+fire+age+log(income)      0.0085 0.0041
race+theft+age+log(income)     0.0106 0.0010
race+fire+theft+age+log(income) 0.0095 0.0004
```

The output shows the $\hat{\beta}_1$ and the associated p-values for all 16 models. We can see see that the value of $\hat{\beta}_1$ varies somewhat with a high value about double the low value. But in no case does the p-value rise above 5%. So although we may have some uncertainty over the magnitude of the effect, we can be sure that the significance of the effect is not sensitive to the choice of adjusters.

Suppose the outcome had not been so clear cut and we were able to find models where the predictor of interest (in this case, race) was not statistically significant. The investigation would then have become more complex because we would need to consider more deeply which covariates should be adjusted for and which not. Such a discussion is beyond the scope of this book, but illustrates why causal inference is a difficult subject.

We should also be concerned whether our conclusions are sensitive to the inclusion or exclusion of a small number of cases. Influence diagnostics are useful for this purpose. We start with a plot of the differences in the coefficient caused by the removal of one point. These can be seen for the race variable in Figure 12.6.

```
> diags <- data.frame(lm.influence(lmod)$coef)
> ggplot(diags,aes(row.names(diags),race)) +geom_linerange(aes(ymax=0,
    ymin=race)) +theme(axis.text.x=element_text(angle=90)) + xlab("
    ZIP") +geom_hline(yintercept=0)
```

The ggplot function requires the data in the form of a data frame. We extract the relevant component from the the lm.influence call for this purpose. We can see that the largest reduction is about 0.001 which would be insufficient to change the statistical significance of this term.

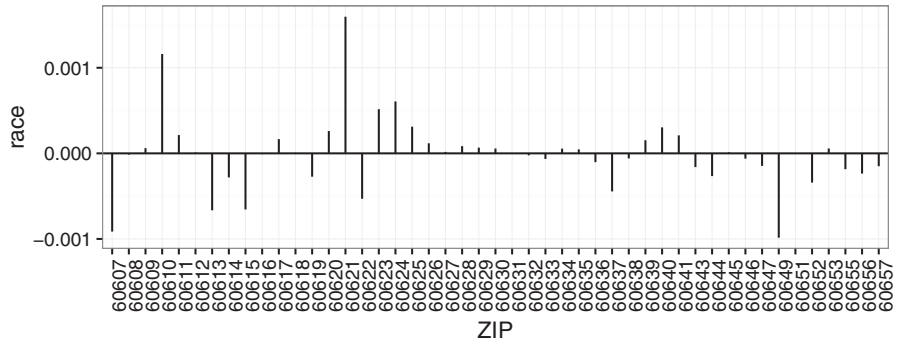


Figure 12.6 Leave-out-one change in coefficient values for $\hat{\beta}_{race}$.

It is also worth considering the influence on the adjustment covariates We plot the leave-out-one differences in $\hat{\beta}$ for theft and fire:

```
> ggplot(diags,aes(x=fire,y=theft))+geom_text(label=row.names(diags))
```

Let's also take a look the standardized residuals and leverage which can be conveniently constructed using the default plot function for a linear model object:

```
> plot(lmod,5)
```

See Figure 12.7 where zip codes 60607 and 60610 stick out. It is worth looking at other leave-out-one coefficient plots also. We also notice that there is no standardized residual extreme enough to call an outlier. Let's take a look at the two cases:

```
> chredlin[c("60607","60610"),]
      race fire theft  age involact income side
```

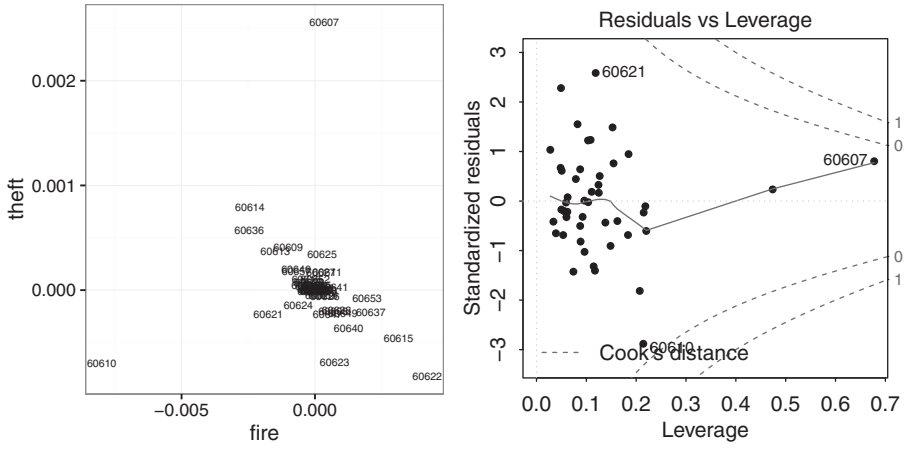


Figure 12.7 Plot of the leave-out one coefficient differences is shown on the left. Plot of the standardized residuals against the leverages is shown on the right

```
60607 50.2 39.7   147 83.0       0.9  7.459    n
60610 54.0 34.1    68 52.6       0.3  8.231    n
```

These are high theft and fire zip codes. See what happens when we exclude these points:

```
> match(c("60607", "60610"),row.names(chredlin))
[1] 24  6
> lmode <- lm(involact ~ race + fire + theft + age + log(income),
    chredlin,subset=-c(6,24))
> sumary(lmode)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.57674    1.08005   -0.53    0.596
race         0.00705    0.00270    2.62    0.013
fire         0.04965    0.00857    5.79    1e-06
theft       -0.00643    0.00435   -1.48    0.147
age          0.00517    0.00289    1.79    0.082
log(income)  0.11570    0.40111    0.29    0.775

n = 45, p = 6, Residual SE = 0.303, R-Squared = 0.8
```

The predictors `theft` and `age` are no longer significant at the 5% level. The coefficient for `race` is reduced compared to the full data fit but remains statistically significant.

So we have verified that our conclusions are also robust to the exclusion of one or perhaps two cases from the data. This is reassuring since a conclusion based on the accuracy of measurement for a single case would be a cause for concern. If this problem did occur, we would need to be particularly sure of these measurements. In some situations, one might wish to drop such influential cases but this would require strong arguments that such points were in some way exceptional. In any case, it would be very important to disclose this choice in the analysis.

Now if we try very hard to poke a hole in our result, we can find this model where two cases have been dropped:

```
> modalt <- lm(involact ~ race+fire+log(income),chredlin,  subset=-c
    (6,24))
> sumary(modalt)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.75326    0.83588    0.90    0.373
race         0.00421    0.00228    1.85    0.072
fire         0.05102    0.00845    6.04  3.8e-07
log(income) -0.36238    0.31916   -1.14    0.263

n = 45, p = 4, Residual SE = 0.309, R-Squared = 0.79
```

In this model, `race` no longer meets the threshold for significance. However, there is no compelling reason to advocate for this model against the large weight of other alternatives we have considered.

This illustrates a wider problem with regression modeling in that the data usually do not unequivocally suggest one particular model. It is easy for independent analysts to apply similar methods but in different orders and in somewhat different ways resulting in different model choices. See Faraway (1994) for some examples. For this reason, the good analyst explores the data thoroughly and considers multiple models. One might settle on one final model but confidence in the conclusions will be enhanced if it can be shown that competing models result in similar conclusions. Our analysis in this chapter demonstrates this concern for alternatives but there is an unavoidable reliance on human judgement. An unscrupulous analyst can explore a large number of models but report only the one that favors a particular conclusion.

A related concept is *model uncertainty*. We surely do not know the true model for this data and somehow our conclusions should reflect this. The regresssion summary outputs provide standard errors and p-values that express our uncertainty about the parameters of the model but they do not reflect the uncertainty about the model itself. This means that we will tend to be more confident about our inferences than is justified. There are several possible ways to mitigate this problem. One simple approach is data splitting as used in the running example on the `meatspec` data in Chapter 11. Another idea is to bootstrap the whole data analysis as demonstrated by Faraway (1992). Alternatively, it may be possible to use *model averaging* as in Raftery, Madigan, and Hoeting (1997).

12.5 Discussion

There is some ambiguity in the conclusion here. These reservations have several sources. There is some doubt because the response is not a perfect measure of people being denied insurance. It is an aggregate measure that raises the problem of ecological correlations. We have implicitly assumed that the probability a minority homeowner would obtain a FAIR plan after adjusting for the effect of the other covariates is constant across zip codes. This is unlikely to be true. If the truth is simply a variation about some constant, then our conclusions will still be reasonable, but if this probability varies in a systematic way, then our conclusions may be off the mark. It would be a very good idea to obtain some individual level data.

We have demonstrated statistical significance for the effect of race on the response. But statistical significance is not the same as practical significance. The largest value of the response is only 2.2% and most other values are much smaller. Using our preferred models, the predicted difference between 0% minority and 100% minority is about 1%. So while we may be confident that some people are affected, there may not be so many of them. We would need to know more about predictors like insurance renewal rates to say much more but the general point is that the size of the p-value does not tell you much about the practical size of the effect.

There is also the problem of a potential latent variable that might be the true cause of the observed relationship. Someone with first-hand knowledge of the insurance business might propose one. This possibility always casts a shadow of doubt on our conclusions.

Another issue that arises in cases of this nature is how much the data should be aggregated. For example, suppose we fit separate models to the two halves of the city. Fit the model to the south of Chicago:

```
> lmod <- lm(involact ~ race+fire+theft+age, subset=(side == "s"),
    chredlin)
> sumary(lmod)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.23441    0.23774   -0.99    0.338
race         0.00595    0.00328    1.81    0.087
fire         0.04839    0.01689    2.87    0.011
theft       -0.00664    0.00844   -0.79    0.442
age          0.00501    0.00505    0.99    0.335

n = 22, p = 5, Residual SE = 0.351, R-Squared = 0.74
```

and now to the north:

```
> lmod <- lm(involact ~ race+fire+theft+age, subset=(side == "n"),
    chredlin)
> sumary(lmod)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.31857    0.22702   -1.40    0.176
race         0.01256    0.00448    2.81    0.011
fire         0.02313    0.01398    1.65    0.114
theft       -0.00758    0.00366   -2.07    0.052
age          0.00820    0.00346    2.37    0.028

n = 25, p = 5, Residual SE = 0.343, R-Squared = 0.76
```

We see that `race` is significant in the north, but not in the south. By dividing the data into smaller and smaller subsets it is possible to dilute the significance of any predictor. On the other hand, it is important not to aggregate all data without regard to whether it is reasonable. Clearly a judgment has to be made and this can be a point of contention in legal cases.

There are some special difficulties in presenting this during a court case. With scientific inquiries, there is always room for uncertainty and subtlety in presenting the results, particularly if the subject matter is not contentious. In an adversarial proceeding, it is difficult to present statistical evidence when the outcome is not clear-

cut, as in this example. There are particular difficulties in explaining such evidence to non-mathematically trained people.

After all this analysis, the reader may be feeling somewhat dissatisfied. It seems we are unable to come to any truly definite conclusions and everything we say has been hedged with "ifs" and "buts." Winston Churchill once said:

> Indeed, it has been said that democracy is the worst form of Government except all those other forms that have been tried from time to time.

We might say the same about statistics with respect to how it helps us reason in the face of uncertainty. It is not entirely satisfying but the alternatives are worse.

Exercises

In all the following questions, a full answer requires you to perform a complete analysis of the data including an initial data anaysis, regression diagnostics, a search for possible transformations and a consideration of model selection. A report on your analysis needs to be selective in its content. You should include enough information for the steps leading to your selection of model to be clear and reproducible by the reader. But you should not include everything you tried. Dead ends can be reported in passing but do not need to be described in full detail unless they contain some message of interest. Above all your analysis should have a clear statement of the conclusion of your analysis.

1. Reliable records of temperature taken using thermometers are only available back to the 1850s, but it would be interesting to estimate global temperatures in the pre-industrial era. It is possible to obtain various *proxy* measures of temperature. Trees grow faster in warmer years so the width of tree rings (seen in tree cross-sections) provides some evidence of past temperatures. Other natural sources of proxies include coral and ice cores. Such information can go back for a thousand years or more. The dataset `globwarm` contains information on eight proxy measures and northern hemisphere temperatures back to 1856. Build a model and predict temperatures back to 1000 AD. State the uncertainty in your predictions. Comment on your findings. (Note: that this data has been modified and simplified for the purposes of this exercise — see the R help page for the data to find out about the source).
2. The `happy` dataset contains data from 39 MBA students on predictors affecting happiness. Analyze the data to provide an interpretation for the relationship between happiness and money, sex, love and work.
3. The `mammalsleep` dataset contains data on 62 mammals. Focus your interest on the effect of predation on sleep, making proper adjustments for the effects of other predictors.