Practical Guide to Logistic Regression

Chapman & Hall/CRC: 17 July, 2015

Joseph M. Hilbe
hilbe@asu.edu : j.m.hilbe@gmail.com

ERRATA & COMMENTS

as of 15 March, 2016

NOTICE: The LOGIT package may be downloaded from CRAN. After installation, be sure to type

```
> library(LOGIT)
```

to load it into memory. All of the data sets and functions in *Practical Guide to Logistic Regression* will automatically be available to you. My thanks to Dr. Rafael de Souza, Eötvös Loránd University (Budapest, Hungary) for preparing LOGIT for CRAN. We are the listed co-authors of the package.

The next printing of the book will be corrected, having the changes shown in this document made to the new printing.

ERRATA & CHANGES

Added acknowledgements for assistance with the book will be in the Preface of second printing.

p 4 Eq 1.1 Prefer to have *f*(y;p) for left side equation

p 5. Equation 1.3 Prefer to have L(p;y) for left side equation.

P 5 Equation 1.4. Put product sign in front or exp{ …}. First y within braces needs index, i. Left side terms of likelihood have no subscripts.

P 5. Equation 1.5

$$\mathcal{L}(p;y) = \sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{p_i}{1-p_i}\right) + \ln(1-p_i) \right\} \tag{1.5}$$

p.6 For clarity purposes I have reversed the β and *x* terms in the listed equations. xb, or x′b will be reserved for the linear predictor. β*x* indicates a regression coefficient-predictor term. xβ is fine, but may be confusing.

Equation 1.6 should read (this is an equation for linear regression) without summation symbol. I am also changing the index term for the ending regression term as *p* in place of *j*. *p* is more commonly used now, although most any letter is ok if defined as such.

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \quad (1.6)$$

p.7 Equation 1.7 should read same as 1.6 above, except $x_i b$ in place of $\hat{y}_i$.

p. 7 Equation 1.8 should read without the summation symbol.

p 31: mid-page, second of the bullet points. Change to read: ·
"Urgent patients have about a three quarters of the odds of dying in the hospital than do emergency patients."

p. 41 Figure 2.2, caption. The caption should read:
"Predicted probability of death by length of stay"

p. 49. Equation 3.1 should read with the summation symbol

p 50 Equation 3.2 . Delete the summation sign in the exponential.

P 50, Equation 3.3. Revise as:

$$\frac{\mu_i}{1-\mu_i} = exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) \quad (3.3)$$

p. 84 First line of code on page. The word "`modeltest`" should be "`mymod`".

P 84 Add a sentence following the final sentence on page 84. This is not an error or a correction, but rather a clarification or caveat. Add the following:

"Note that the code used in 4.2.1 must be used prior to using the code for ROC analysis. Statistics required for ROC analysis are calculated in the S-S plot test. "

p 88 Bottom line of code on page. Amend to read (new code is in red):

```
> summary(mymod <- glm(died ~ white + los + hmo + factor(type),
          family=binomial, data=medpar))
```

p 89. The top line of code in the middle of the page should read:

```
> library(LOGIT)     #  or source("HLTest.R")
```

At the end of the output table, the test code should not read "> HLChi2", but instead:

```
> HLChi10
```

p. 106 Stata code, section 4.3. Should read as (change is in red):

```
. estat gof, table group(10)
. estat gof, table group(12)
```

Bottom of page 106:
4.6 "`. use phmydata2`" should read "`. use pgmydata2`"

p. 107, Eq 5.2: The last term in the binomial PDF should be logged:

Read as $... + ln\binom{n}{y}$

p. 108, Eq 5.3: The last term in the binomial log-likelihood function should be logged:

Read as $... + ln\binom{n}{y}$

P 119 Eq 5.14 The π term should be μ

p. 131, First sentence on page and code underneath it should read:
"The data have 3874 observations and 15 variables."

```
 > dim(R84)
[1] 3874    15
```

MINOR TYPO'S & REWORDING

In the *medpar* data, the variable *died* is defined as "patient dies within 48 hours of admission". In fact, however, *died* should be defined as: "patient died during hospitalization". This occurs at a number of places in the book. They are listed below:

==========

p. 8 4th line from bottom of the page. Change the "died" inset to read:
died: 1 = patient dies while hospitalized; 0= did not die during this period"

p. 28 last line on page. Read as:
"… *died*(1: died while hospitalized) and …"

p. 30 both bullet points at the bottom of the page. Read as:
"… odds of dying in the hospital than do elective admissions."

p. 31 both bullet points in the middle of the page. read as
"… odds of dying in the hospital than do emergency patients"

p. 31 the last line of the paragraph following the above bullet points. Read as:
" … to death while hospitalized."

p. 38 mid-page following R code, Reword as:
"White patients have a 35% greater odds of death while hospitalized than do nonwhite patients."

p. 38 last block of text on page, above R code at bottom of page. Read as:
"1368 *white* patients have… dying within the hospital. Nonwhite patients …"

p.39 1st line of 1st full paragraph, between sections of R code. Change to read:
" If we wish to … of death while hospitalized for a patient …"

p. 39, last sentence above the heading for Section 2.6.2. Change to read:
"… 27% probability of dying while hospitalized ---- given a specific…"

p. 80, 5th line from the bottom of R code (bottom of page) in line beginning with *plot;*
Change the section on that line beginning with *main=* to read:
" plot …. main= 'P(death) while hospitilized',"

p. 81, Figure 4.4, Title should be changed to read:
"P[Death] while hospitalized"

p. 86, Second line from top of page. Change to read (last of line 1 to first of line 3):
…"Given that died indicates that a patient died while hospitalized, the AUC…"
==========

CONTINUATION OF MINOR ERRATA AND AMENDMENTS

p ix: Preface. End of first paragraph: "Sigma Six" should be "Six Sigma"

p 5, last line on page. "it" should read "them". Read as
" worry about them in this discussion".

p 7. Added words need on line 3 below Equation 1.8 (in red). Read as follows:
"It can also be thought of as the probability of the presence or occurrence of some characteristic, while 1-p can be thought of as the probability of the absence of that characteristic."

p. 7 Reword the paragraph below Eq 1.9 to read in its entirety as follows:

"The equations in (1.9) above are very important, and will be frequently used in our later discussion. Once a logistic model is solved, we may calculate the linear predictor, xb, and then apply either equation to determine the predicted value, μ for each observation in the model."

p. 8 Third paragraph, line 4. "of" should be "or". Read as:
"… click on "Run line or selection". This places…"

p. 10, First line in second block of R code. The predictor name "hite" should be "white".

p 16/17 The "." dot in front of the code at the bottom of page 16 and top of page 27 is a Stata marker. For R, the marker is ">". The lines should read:

p. 16
Odds x=0
```
>  3/1
[1] 3
```

p 16, second full paragraph on page. First sentence, starting from “The odds pf x”
Change to read.
“The odds of y given x= 1 is exp(−1.504077) or 0.22222…”

p.17
Odds Ratio x=1 to x=0

```
>  (2/3)/(3/1)
[1] 0.2222222
```

p 20: 3 lines from top of page. Delete Hilbe(2016). Read as:
“See the PDF document, “Calculating …”

p 20: Directly following the section 2.3 heading. Second line. Delete the word “important”. Read as
“a display of basic model statistics as well as several statistics that”

p 21 first line of code on page. The word “bin” should be “binomial”. Read as:

```
> summary(logit2 <- glm(y~x, family=binomial, data=xdta)))
```

p. 22: Align the column titles over the statistics, shown below as:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.099      1.309   0.839    0.429
x             -1.504      1.669  -0.901    0.397
```

p. 22: First paragraph after R output about ¼ down page. Last sentence of the paragraph should instead read:
“However, the scaled standard error for *x* in the above model *logitsc* is calculated from model *logit2* by”

p 23. Section 2.3.3, 2 lines under R code. For clarity, change to the following:
“More exactly, *p* is the probability of obtaining a coefficient value at least as extreme as the observed coefficient given the assumption that β=0. The smaller the *p*-value, the more likely $\beta \neq 0$.”

p. 23, sec 2.3.3, 4th line after code beginning with “`> pvalue <- 2*norm…`”
Delete first “that” in the line (2nd word in line)

p. 25, 5th line of text on page. Add word (in red) to sentence to read as:
“… do just that.”

p 27, bottom paragraph on page: 4 lines from bottom of page: Delete the sentence:
“A range of values for a coefficient, β are produced for which the null hypothesis of β=0 would not be rejected.”

p. 30 line following computer output in mid page. Change the first sentence to read: (the change in in red)
“Note how the *factor* function excluded factor *type*1 (elective) from the output.”

p. 32, Sec 2.5.1, 2nd paragraph, last sentence. Add "predictors" in the sentence, to read as:
"… to continuous predictors because they …."

p. 33, 4th paragraph on page, last word on top line. Change "logit" to "logic".

p. 40, 1st full paragraph under top code on page. delete the comma in the middle of the sentence.

p. 56, second line from top. Reword sentence, adding (double under-score!). Read as:
" … *glm* estimation. Called *P__disp* (double under-score!), it is`a function …"

p. 57, 3rd line of text from top of page. Delete entire sentence, "These may be calculated for the *ed1* model using *ht* code."
p. 57, bottom paragraph on page. Line 3, "use" should be "used".

p. 64, last sentence on page. Change the word, "expanded" to "expanding".

p. 64 last sentence on page, delete "to" after "standard errors"

p. 66, Paragraph above Eq 3.19. 2nd sentence beginning with, "Only the interaction…" Delete the comma in the middle of the sentence.

p. 85.
A) 8th line from top of page, delete the word "degree".
B) 9th line from top of page, delete the first "the". The sentence should read:
" The model is a better classifier with greater values of the ROC…"
C) 10th line for top of page. The words "Be aware" should read, "Beware".
D) 2nd line of the 2nd paragraph. Add the word "as" between "to" and "Harrell's".

p. 88. 2nd paragraph of Section 4.3. Line 2. The word "softer" should be "software".

P 119. Add the following to the sentence following Eq 5.14:
"… mean, $\mu=a/(a+b)$, and variance, σ^2, as $sqrt(ab/((a+b)^2 (a+b+1)))$.

p.128 4th line on page, "addition" should read "additional"

p 128: add a new # 1 for listings. Add sentence to 6th listing.

There are six foremost characteristic features that distinguish Bayesian…

1: *Regression models have slope, intercept, and sigma parameters*: Each parameter has an associated prior.

. . .

6: *Additional or Prior Information*: The distribution used as the basis of a parameter estimate (likelihood) can be mixed with additional information – information that we know about the variable or parameter that is independent of the data being used in the model. This is called a prior distribution. Priors are PDFs that add information from outside the data into the model.

---

p 132: Delete: `> source("c://Rfiles/toOR.R")`

p 132 bottom of page. Amend sentences to read (change in in red)
For our example I shall employ the default *multivariate normal priors* on
all of the **predictor parameters**. It is used because we have more than one **parameter**, all of

p. 134 Change the last two paragraphs on page. Changes are in bold red.

Although the interpretations differ, the **posterior** mean values are analogous
to maximum likelihood coefficients, **Bayesian** standard **deviations** are like MLE
standard errors, …
Remember that each **parameter** is considered …
… distribution for each **posterior parameter,** the mean of each is the Bayesian logistic
beta. The plots on the right-hand side of Figure 6.1 display the distributions

p. 136. Top 2 lines on page. Change "predictor" to "parameter" and "predictor's" to "parameter's"

p. 136. Mid page. The comment on the line beginning with "`> geweke.diag(mymc)`
should be deleted. That is, delete this text: `# Creates Figure 6.1"`

p. 140, mid page, directly under sentence ending LOGIT.txt. Amend the lines of text beginning with "Within the defining…" and ending with "…multivariately normal for all.".With changed text in red, read as:

Priors and the likelihood function are defined within the *model* parentheses:

```
model{
```

We start by defining the priors. The prior betas are all defined as multivariately normal.

p. 142, 6th line from top of page. Reword sentence for better clarity. With changes in red, the line should read:
" The initial values can vary widely, and skew the results. If all of the"

p. 142. Add an initial sentence to the paragraph under the J0 code in upper mid page, and change the current first sentence in the paragraph to read as:

"After running the *jags* function, which we have called J0, typing J0 on the R command-line will provide raw model results. The final code in Table 6.1 provides nicer looking output. "

p. 143, 2nd line from top. Delete word "the"; amend word "prior" to "priors"
'
p. 143, 1st full paragraph on page, second line. Change the word "produce" to "reveal".

P 143, Section 6.2.3, first paragraph starting at bottom of page, over to page 144. Change to read:

In a regression model the focus is on placing priors on parameters in order to develop adjusted posterior parameter values. For example, we could set a prior …
…negative binomial. The same prior may be set on one or more parameters, and different priors may be set for separate parameters. Each software package specifies how this should be coded.

P 144, 5th and 6th lines from top. Changes in red.
it is assumed to be applied to all model **parameters**. Otherwise, each **parameter**,
including the **parameter of the** intercept, may have a prior.

p. 144, 1st full paragraph, Changes in red
The example below employs **a** Cauchy prior on all three parameters; that
is, **the coefficients on** *intercept*, *cdoc*, and *cage*.

p. 144, mid page, line of text beginning with "the intercept,". Amend line to read more clearly, "the intercept; the reader may want to check if this is indeed the case (Table 6.2). The"

p. 144, mid page text, 3rd line from bottom. Change "to" to "from". Read as:
"code, presented … different manner from Table 6.1 can be used for"

p. 146, First line of text on page. Change "prior" to "priors".

p. 146, 6th line in paragraph at bottom part of page. Change in red.
prior that is distributionally compatible with the distribution of the **parameter** having the prior.

p.151, Second reference book from bottom should be amended to:
Hilbe, J.M., de Souza, R.S., and Ishida, E. (2016), *Bayesian Models for Astrophysical Data: Using R/JAGS and Python/Stan*, Cambridge, UK: Cambridge University Press

COMMENTS

Note: All errata identified in the book will be corrected in the second printing.

DATA and FUNCTIONS/COMMANDS

All data used in the text, as well as the code for all user-created functions and commands are available on my BePress site. The URL is:

http://works.bepress.com/joseph_hilbe/

Scroll down to the first section on Practical Guide for Logistic Regression and download errata, data and functions. Data and functions in R, Stata, and SAS format is available.

ADDITION TO CHAPTER 6 IN BOOK

I have posted "Addition-to-PGLR-Chap _Sept2015" to my Bepress site for readers of PGLR who have an interest in Bayesian logistic modeling using Stata. I prepared a 10 page addition to Ch 6 that provides complete Bayesian code and log-likelihood evaluators for Bayesian logistic regression and Bayesian grouped logistic regression. Examples and explanations are included.

ACKNOWLEDGEMENTS

My thanks to Patricia McKinley for spotting many of the needed changes to the book. I very much appreciate her fine help, and ability to detect needed amendments. The forthcoming second printing also has benefited from the assistance of Ulrike Gr*ö*mping and James Hardin, who offered valuable suggestions.