Logistic Regression Models

Chapman & Hall/CRC Press

Joseph M. Hilbe

Hilbe@asu.edu

27Mar2010

ERRATA & COMMENTS/AMENDMENTS DUE TO UPDATED STATISTICAL CODE USED FOR EXAMPLES

CHANGES TO 2ND PRINTING : NO CHANGES WERE MADE TO 3RD PRINTING

PREFACE

End of the Preface, page xvii. .

“Note: The preferred method of converting Stata data files into R differs from what is shown in the text when using Stata versions 10 and later. Using the **medpar.dta** Stata file, stored in c:/data, as an example, I suggest using the following code:

```
  library(“Hmisc”)
  medpar <- stata.get(“c://data/medpar.dta”)”
```

CHAPTER 2

Page 35: Add to paragraph directly above Equation 2.20:

“Although we do not use it in our calculations here, the standard error of an odds ratio is determined by using the **delta method**; i.e. exp(β)*se(β).”

CHAPTER 3

Page 52: Equation 3.1: second “+” sign should be deleted.

Page 52: Equation 3.6, missing = sign.

Page 53: Amend everything from directly underneath **Solving for ∂L – the Gradient**
At the top of the page through Equation 3.12 to read as follows between double lines:

==

In exponential family form, the log-likelihood function is expressed as:

$$L(\theta, \phi; \mathrm{y}) = \sum \frac{y\theta - b(\theta)}{\alpha(\phi)} + \sum C(y, \phi) \tag{3.10}$$

Solving for *L* with respect to β by using the chain rule, we have

$$\frac{\partial \mathrm{L}}{\partial \beta} = \sum \frac{\partial \mathrm{L}}{\partial \theta} \times \frac{\partial \theta}{\partial \mu} \times \frac{\partial \mu}{\partial \eta} \times \frac{\partial \eta}{\partial \beta} \tag{3.11}$$

Solving for each term:

$$\frac{\partial \mathrm{L}}{\partial \theta} = \sum \frac{y - b'(\theta)}{\alpha(\phi)} = \sum \frac{y - \mu}{\alpha(\phi)} \tag{3.12}$$

==

Page 53: Equation 3.17, far left term should read $\partial\mu/\partial\eta$, not the inverse.
Page 54: top line should read:
"where y and μ are the response and fitted values respectively, x is …"

Page 54: 3rd line on page: change to: "**Solving for $\partial^2 L$ − Fisher Scoring**"
Page 56: Equation 3.38, replace $\sum$ to $\prod$, and have subscripts, to now appear as:

$$f(y;\theta,\phi) = \prod_{I=1}^{N} \left\{ \frac{y_i\theta_i - b(\theta_i)}{\alpha(\phi)} + C(y_i;\phi) \right\} \tag{3.38}$$

CHAPTER 4

Page 63: The final sentence before Equation 4.1 should read:
"Given these terms, the Bernoulli PDF can be expressed as:"

Page 65: Add another formula to equation 4.15 for added information:

$$L(y=1) = \Sigma\{\ln(\mu/(1-\mu)) + \ln(1-\mu)\} \quad = \quad \Sigma \ln(\mu) \tag{3.15}$$

Page 66: Expand Equation 4.24 to read as:

$$\frac{\partial(L)}{\partial\,\mu} = \frac{y}{\mu} - (1-y)(1-\mu)^{-1} = \frac{y-\mu}{\mu(1-\mu)} \tag{4.24}$$

Page 69: Table 4.2: The logit link function should read ln(μ/(1-μ)).

CHAPTER 5

Page 77: Top line on page: Substitute "three" for "two".

Page 100: Equation 5.9: Should read as

$$Std(\beta_i) = \beta_i \frac{SD(X_i)}{\sqrt{\frac{\pi^2}{3}}}$$

Page 107
4th line of 1st full paragraph: Amend to read:
"… For instance, consider a response, e.g. *death*, that we are attempting to …"

Pages 115 (top) and 121 (mid)
The tables should appear as:

```
           |        0          1
     death |  Inferior   Anterior |     Total
-----------+----------------------+----------
         0 |     2,504      2,005 |     4,509
         1 |        67        120 |       187
-----------+----------------------+----------
     Total |     2,571      2,125 |     4,696
```

Page 125: Equation 5.31: The denominator of the rightmost term should read c(a+b)

Page 133:
CREATE TWO RANDOM VARIATES
The terms “invnorm(uniform())” may now read with Stata 11 as “rnormal()”. Amend to read as

```
. gen x1 = abs(rnormal())
. gen x2 = abs(rnormal())
```

Page 133
Change from after “`. gen xb = 1 + .25*x1 - 1.5*x2`” in mid page up to “GLM COMMAND BINOMIAL LOGISTIC MODEL” a few lines down on same page to read as (between double lines)

========== ==

CREATE BINOMIAL LOGISTIC RESPONSE WITH DEFINED DATA

```
. gen d = 100
. gen exb = 1/(1+exp(-xb))
. gen by = rbinomial(d, exb)
```

==

Page 154: 1st full paragraph, change from Long and Freese (2006a) to:
Long and Freese (2006)

CHAPTER 6

Page 193: 3rd line from bottom. Coefficient of β_3 should be negative; ie −1.846994

Page 218: Logit command ¾ down on page. There is an extra comma in the command.

CHAPTER 7

Page 270: 3rd line under Eq 7.25
The y^ should be ŷ.

Page 272: Eq 7.33 and Eq 7.34 mistaken. Should read as:

$$d = \mathrm{sqrt}[2\textstyle\sum\{\ln(1/\mu)\}] \qquad \text{if } y = 1 \tag{7.33}$$

$$d = \mathrm{sqrt}[2\textstyle\sum\{\ln(1/(1-\mu)\}] \qquad \text{if } y = 0 \tag{7.34}$$

Page 279: 3rd line from top (under 7.4.1.6 Likelihood Residuals), and line directly above Eq 7.47, should read:

“…deviance residuals, and is defined as:”

Page 279: bottom of page, starting with equation 7.49, add new material until end of page. Everything between double lines is to replace what is currently given. Text resumes as is now in book starting at the top of page 280.

===

$$Bernoulli\text{: }\{A(y) - A(\mu)\}/\{\mu(1-\mu)\}^{1/6} \tag{7.49}$$

with A(z) = Beta(2/3, 2/3) * [Incomplete Beta(2/3, 2/3, z)]
A(z) = 2.05339 * [Incomplete Beta(2/3, 2/3, z)]

where z takes the value of y or μ as appropriate. The constant value of the two-term *beta* function with both parameters at 2/3, is 2.05339.

```
. di exp(lngamma(.666667)+lngamma(.666667)-lngamma(.666667+.666667))
2.0533894
```

Code for calculating the residual can be given as:

```
((2.0533902*ibeta(2/3,2/3,y))-(2.0533902*ibeta(2/3,2/3,mu)))/(mu*(1-mu))^(1/6)
```

Anscombe residuals for the binomial family appear as

$$\frac{\left(A\left(\frac{2}{3},\frac{2}{3},\frac{y}{m}\right) - A\left(\frac{2}{3},\frac{2}{3},\mu\right)\right)}{\mu^{1/6}(1-\mu)^{1/6}} \tag{7.50}$$

```
((2.0533902*ibeta(2/3,2/3,(y/m)))-(2.0533902*ibeta(2/3,2/3,(mu))))/((mu*(1-mu))^(1/6))
```

Some statisticians multiply $\sqrt{1-h}/m$ to the denominator, where h is the hat matrix diagonal.

==

CHAPTER 9 [update to new code if using Stata version 11]

Page 323 near top
DELETE=>

```
. gen xb = .5+1*x1-1.25*x2+.25*x3
. genbinomial y, xbeta(xb) de(d)
```

REPLACE WITH=>

```
. gen y =rbinomial(d, 1/(1+exp(-(.5+1*x1-1.25*x2+.25*x3))))
```

Page 324 near bottom
DELETE=>

```
. gen xbi = .5 + 1*x1 - 1.25*x2 + .25*x3 + .2*x23
. genbinomial yi, xbeta(xbi) de(d)
```

REPLACE WITH=>

```
. gen yi =rbinomial(d, 1/(1+exp(-(.5+1*x1-1.25*x2+.25*x3+.2*x23))))
```

Page 326 near top
DELETE=>

```
. xbsq = .5 + .5*x1sq -1.25*x2 + .25*x3
. genbinomial ysq, xbeta(xbsq) de(d)
```

REPLACE WITH=>

```
. gen ysq =rbinomial(d, 1/(1+exp(-(.5+.5*x1sq-1.25*x2+.25*x3))))
```

Page 327 near bottom
DELETE=>

We can use the same data as in the previous models, but generate a random binomial probit response. The data are still based on what was seeded at the model setup

```
. genbinomial yp, xbeta(xb) de(d)
```

REPLACE WITH=> Between double lines:

==

Random probit data is best generated using a pseudo-random uniform generator. Keeping the same seed, we create three pseudo-random variates, and a linear predictor with the same values used for the logit models above.

```
. gen xx1 = runiform()  // create xx2 and xx3 in the same manner
. gen yp = rbinomial(d, normprob(.5+1*xx1-1.25*xx2+.25*xx3))
. glm yp xx1 xx2 xx3, nolog fam(bin d) link(probit)

Generalized linear models              No. of obs      =     10000
Optimization     : ML                  Residual df     =      9996
                                       Scale parameter =         1
Deviance         =  9920.985699        (1/df) Deviance =  .9924956
Pearson          =  9803.934329        (1/df) Pearson  =  .9807857

Variance function: V(u) = u*(1-u/d)    [Binomial]
Link function    : g(u) = invnorm(u/d) [Probit]

                                         AIC           =  5.392163
Log likelihood   = -26956.81434          BIC           = -82145.58
------------------------------------------------------------------
      |               OIM
   yp |     Coef.  Std. Err.     z    P>|z|  [95% Conf. Interval]
------+-----------------------------------------------------------
  xx1 |  .9981278  .0055718   179.14  0.000   .9872073   1.009048
  xx2 | -1.247718   .005676  -219.82  0.000  -1.258843  -1.236594
  xx3 |  .2470518  .0055181    44.77  0.000   .2362365   .2578671
_cons |  .5017432  .0049456   101.45  0.000   .4920499   .5114365
------------------------------------------------------------------
```

```
. abic
AIC Statistic   =   5.392163           AIC*n      = 53921.629
BIC Statistic   =   5.392472           BIC(Stata) = 53950.469
```

The simulated binomial probit model has parameter estimates very close to those assigned, the dispersion is .98, only two one-thousandths from unity, and the AIC statistic is 53922, higher than the 51172 value for the logit model.

We have no knowledge that the true model for these data is a binomial probit. First, attempt a binomial logistic model on the same data.

MODEL PROBIT DATA WITH A BINOMIAL LOGISTIC MODEL

```
. glm yp xx1 xx2 xx3, nolog fam(bin d)

Generalized linear models             No. of obs      =     10000
Optimization     : ML                 Residual df     =      9996
                                      Scale parameter =         1
Deviance         =   9989.42777       (1/df) Deviance =  .9993425
Pearson          =  9838.864611       (1/df) Pearson  =  .9842802

Variance function: V(u) = u*(1-u/d)   [Binomial]
Link function    : g(u) = ln(u/(d-u)) [Logit]

                                        AIC           =  5.399007
Log likelihood   = -26991.03538         BIC           = -82077.13
-----------------------------------------------------------------
      |              OIM
   yp |     Coef.  Std. Err.      z    P>|z|  [95% Conf. Interval]
------+----------------------------------------------------------
  xx1 |  1.665472  .0094186   176.83  0.000   1.647012   1.683932
  xx2 | -2.083331  .0096725  -215.39  0.000  -2.102288  -2.064373
  xx3 |  .4147304  .0092279    44.94  0.000    .396644   .4328168
_cons |  .8288387  .0082547   100.41  0.000   .8126598   .8450176
-----------------------------------------------------------------

. abic
AIC Statistic   =   5.399007           AIC*n      = 53990.07
BIC Statistic   =   5.399316           BIC(Stata) = 54018.914
```

The logistic model is different, with an AIC of 53990, nearly 70 higher than the probit model on the same data. The logistic model is not overdispersed, but the AIC and BIC values are substantially higher than the probit.

Table 9.3 displays a comparison of the parameter estimates. The binomial logistic parameter estimates are nearly two times greater than the true probit estimates. Without prior knowledge of the true model, the only indication of mis-specification are the AIC and BIC statistics. Monte Carlo simulation, however, shows that with repeated estimation without seeds, a logistic model of true probit data will be overdispersed.

TABLE 9.3
Probit Data: True, Probit, Logistic Models

	True	Probit	Logistic
xx1	1.000	0 .998	1.665
xx2	-1.250	-1.248	-2.083
xx3	0.250	0.247	0.415
Cons	0.500	0.502	0.829
AIC		53922	53990
DISP	1.00	53950	54019

Page 335 middle of page
DELETE=>

```
. gen xb = .5 + 1*x1 - 1.25*x2 + .25*x3
. genbinomial y, xbeta(xb) de(d)
```

REPLACE WITH=>

```
. gen y =rbinomial(d, 1/(1+exp(-(.5+1*x1-1.25*x2+.25*x3))))
```

Page 337 Top full paragraph plus text through `glm y x1 x2 x3, <…>`

CHANGE FROM THIS (CURRENTLY IN TEXT)=>

The model below is supplied with the negative binomial heterogeneous or ancillary parameter value (.0357547). This value was previously obtained by modeling a maximum likelihood negative binomial, which estimated the ancillary parameter. In Stata, maximum likelihood negative binomial estimates are obtained using the **nbreg** command. The SAS GENMOD procedure estimates the ancillary parameter. See Hilbe (2007a) for a thorough discussion of this subject.

RATE NEGATIVE BINOMIAL

```
[. nbreg y x1 x2 x3, nolog exp(d)] /// obtain ML estimate of
                                       ancillary parameter

. glm y x1 x2 x3, fam(nb .0357547) lnoffset(d) nolog
```

CHANGE TO THIS=>

The model below obtains the negative binomial heterogeneity or ancillary parameter value of .0357547 from a maximum likelihood negative binomial algorithm called from within the **glm** program. It supplies the value and employs it as a constant to the **glm** estimating equations. See Hilbe (2007a) for a comprehensive discussion of this subject.

RATE NEGATIVE BINOMIAL

```
. glm y x1 x2 x3, fam(nb ml) lnoffset(d) nolog
------------------------------------------------------------------

<rest of page the same>
```

CHAPTER 10

Page 357 : formula under CATEGORY OR LEVEL 3 should read
Logit = ln[($p_1 + p_2 + p_3$)/(1 - $p_1 + p_2 + p_3$)]

Page 364: Add sentence below to last sentence on page (before **distinct** command):
The **distinct** command below is from Longton and Cox
(http://fmwww.bc.edu/repec/bocode/d/distinct.ado).

CHAPTER 11

Page 391: Change final sentence of text and add another sentence to read as:
We use the **prtab** command to do our work (Long, 1997). **prtab** is in **spost9_ado**
(http://www.indiana.edu/~jslsoc/stata)

Page 409: Replace the current text in the R code section 11.2 to read:
Available after chapter written: library("mlogit"); hmftest()

CHAPTER 12

Page 414, top most programming code: The second line of the "recode" command should read (5 42/52=57), not (4 42/52=57) as in the book.

Page 419, Under the top-most statistical output, and over section 12.3, change text to read:

Interpretation of the odds ratios follows the same logic as the ordered logistic model. Predicted levels may be accessed using the **ocrpred** command, as done for **ologit**. Note that the number of observations in the **ocratio** model above has been inflated to 997 from 601. The reason is based on how levels are compared: Level 1 vs Levels 2,3, and Level 2 vs Level 3.
This results in [205+(204+192)] + [204+192] = 997.

Page 423: 2nd line of text from bottom. Change from Long and Freese (2006a) to:
Long and Freese (2006)

CHAPTER 13

Page 518: Section 13.4, first two library() functions:
Change from library(nlme) to library(nlme4)

CHAPTER 15

Page 548: 4th line from top. The word "converge" should be "convergence"

Page 558: Exercise 15.2, start of second line. Change Exercise 12.3 to 5.3.

APPENDIX A

Page 586 Final paragraph on page and genbinomial command, change to read:

The **genbinomial** random number generator (Roberto Gutierrez) was used to create synthetic models in earlier printings of this book. It is based on the logic of the **rnd** commands—in this case on **rndbinx** (Hilbe). The commands are on this book's web site. For example, given *x*1, *x*2, and *xb*,

```
. genbinomial y, xbeta(xb) n(1)
```

<the rest of the page is OK as is>

Page 587: Paragraph above Section A.6, Change to read:

Note: After this section was written, Stata enhanced is random number capabilities by including a new suite of random number generators with the official software. These have been used in this printing, whereas the user authored **genbinomial** (Gutierrez) was employed in the first two printings. All lead to similar results.

APPENDIX G

Page 601: right column, item : prtab. The author is Long, not Williams.

APPENDIX H

Page 611: First paragraph, add date to Long and Freese.
Long and Freese (2006) have …

REFERENCE

Page 619: Only the second reference to Long and Freese should be given, and only with the date (2006). In other words,
delete the reference to Long, J.S. and J. Freese (2006a), *Regression* …."
In the next reference, change to "Long, J.S. and J. Freese (2006), *Regression* …"

Page 621
Add to the end of the reference: Shults,J., S Ratcliffe, M Leonard (2007) …
(http://www.cceb.upenn.edu/~sratclif/QLSproject.html)

Page 621
Add to the end of the reference: Shults, J., W. Sun, X. Tu, J. Amsterdam (2006)…
(http://biostats.bepress.com/upennbiostat/papers/art8/