Negative Binomial Regression

Second edition, Cambridge University Press

Joseph M Hilbe

ERRATA & COMMENTS

Second Printing: May 2012

errata updated: 30 December, 2012

The second printing of the text was printed and published at Cambridge, UK in May 2012. Distribution began in early May. The errata reported here were identified too late for correction in this printing. Note: Data sets and software code can be downloaded from:

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

I also suggest downloading the PDF document, *Negative Binomial Regression Extensions*, located on the same site. Code to produce all tables and figures in Stata and R are given.

ERRATA

P 70: Table 5.1. The leftmost "<" signs should all read ">" instead.

P 102: line 3. Spelling, "internals" should be "intervals"

p 132: replace R code at bottom p 132 and top 133 with

```
# average discrete effects for outwork
================================================================
bout = coef(rwm1)[2]
mu = fitted.values(rwm1)
xb = rwm1$linear.predictors
pe_out = 0
pe_out = ifelse(rwm5yr$outwork == 0, exp(xb + bout)-exp(xb), NA)
pe_out = ifelse(rwm5yr$outwork == 1, exp(xb)-exp(xb-bout),pe_out)
mean(pe_out) #1.069990
================================================================
```

p 171: Table 7.14 The code in the table has an error. The "ffit" term in the third to last line should be a "t". The corrected code is given below.

Table 7.14 R; Bootstrap standard errors

```
=========================================================
library(COUNT)
data(medpar)
library(boot)
poi <- glm(los ~ hmo + white + factor(type), family=poisson, data=medpar)
summary(poi)
t <- function (x, i) {
xx <- x[i,]
```

```
 bsglm <- glm( los ~ hmo + white + factor(type), family=poisson, data=medpar)
 return(sqrt(diag(vcov(bsglm))))
 }
bse <- boot(medpar, t, R=1000)
sqrt(diag(vcov(poi)))
apply(bse$t,2, mean)
```

==========

p 218: Table 8.7. The standard errors are not as 'accurate' as we can make them given the manner in which they are calculated. Put the code in table 8.7 into a text editor and type in the following additions or amendments (in red)

==========

```
ml.nb2 <- function (formula, data, offset = 0,
              start = NULL, verbose = FALSE) {

xb.hat <- X %*% b.hat[-1] + offset

          y = y,
          method = "BFGS",
          control = list(fnscale = -1,
                    reltol = 1e-16,
                    maxit = 100000),
```

==========

P 225: P 2, line 4, "therefor" should be "therefore"

P 236: Section 9.3, line 2, "was" should be "way".

P 243: P 4, line 2. Repetition of "and". Delete one of them.

P 254: Table 9.27, final line. The R function should be `nb2.obs.pred()`. Replace the *x* in *obx* so that it reads *obs*.

p 306-307: Table 10.8. The standard errors are not as 'accurate' as we can make them given the manner in which they are calculated. Put the code in Table 10.8 into a text editor and type in the following additions or amendments (in red)

==========

```
ml.nb1 <- function (formula, data, offset = 0,
              start = NULL, verbose = FALSE) {

xb.hat <- X %*% b.hat[-1] + offset

          y = y,
          method = "BFGS",
          control = list(fnscale = -1,
                    reltol = 1e-16,
                    maxit = 100000),
```

==========

p 316-317: Table 10.12. The standard errors are not as 'accurate' as we can make them given the manner in which they are calculated. Put the code in Table 10.12 into a text editor and type in the following additions or amendments (in red)

```
=====================================
ml.nbc <- function (formula, data, offset = 0,
            start = NULL, verbose = FALSE) {

xb.hat <- X %*% b.hat[-1] + offset

          y = y,
          method = "BFGS",
          control = list(fnscale = -1,
                    reltol = 1e-16,
                    maxit = 100000),
=====================================
```

p 373: Eq $\partial\mathcal{L}/\partial\beta$ Delete *n* over summation symbol

p 377 Section 11.3.5 Comparative Fit
Amend the entire section to read as follows (between page wide ====== lines):

==========

11.3.5 Tests of comparative fit

The standard fit test for ZINB models is the Vuong test. This test is a comparison of ZINB and the negative binomial, NB2. Essentially, the Vuong test is a comparison of predicted fit values of ZINB and NB2, assessing if there is a significant difference between the two. Given that $P_{NB2}(y|x)$ is the probability of observing y on the basis of x in a NB2 model, and $P_{ZINB}(y|x)$ is the probability of observing the same y on the basis of the same x using a ZINB model, the Vuong test formula may be expressed as:

$$V = \frac{\sqrt{n}\,\bar{u}}{SD(u_i)} \tag{11.16}$$

where *u* is the log ratio of the sum of probabilities, given as

$$u_i = ln\left(\frac{\sum_i P_{ZINB}(y_i|x_i)}{\sum_i P_{NB2}(y_i|x_i)}\right)$$

or

$$u_i = loglikelihood(ZINB) - loglikelihood(NB2)$$

Note that $\bar{u}$ is the mean of *u* and SD(u) is its standard deviation. The Vuong test uses a normal distribution to assess comparative worth. At a 95% confidence level, values of V greater than +1.96 indicate that ZINB is the preferred model. Values lower than -1.96 indicate that NB2 is the preferred model. Values between these to critical points indicate that neither model is preferred over the other.

The Vuong test compares the probabilities of the numerator and denominator, with values greater than 1.96 favoring the probabilities from the numerator. If the ZINB and NB2 models are reversed in the above formula, the interpretation will be reversed as well. Most tests use the formula expressed in equation 11.13. Be certain that you know the which formula is being used with your software prior to reporting a conclusion.

For ease of reading table 11.11 displays how the Vuong test may be calculated by hand for the ZIP model using the *mdvis* data above:

Table 11.11 *Vuong test – ZIP and Poisson*

```
============================================================
use mdvis, clear
global xvars badh age3 educ3

poisson numvisit $xvars, nolog
_predict double xbp, eq(#1) xb                      // xbp = X*beta_p
gen double mup = exp(xbp)                           // mup = exp(X*beta_p)
zip numvisit $xvars, inflate($xvars) vuong nolog
_predict double xbz, eq(#1) xb                      // xbz = X*beta_z
gen double muz = exp(xbz)                           // muz = exp(X*beta_z)
_predict double zgz, eq(#2) xb                      // zgz = Z*gamma_z
* muz = exp(Z*gamma_z) / (1+exp(Z*gamma_z)) assuming logistic
* Loglikelihood: (numvisit=0) log[mug + (1-mug) * exp(-mug)]
* (numvisit>0) log[(1-mug) * exp(-mug) * mug^numvisit / numvisit!]
gen double mug = exp(zgz)/(1+exp(zgz))
gen double u1 = log( mug + (1-mug) * exp(-muz) * muz^numvisit /
exp(lngamma(numvisit+1)) )
replace u1 = log(1-mug) - muz + numvisit*log(muz)-lngamma(numvisit+1) if numvisit >0
* Loglikelihood: log[exp(-mup) * mup^numvisit / numvisit!]
gen double u2 = -mup + numvisit*log(mup) - lngamma(numvisit+1)
* log(likelihood of ZIP model) - log(likelihood of POISSON model)
gen double u = u1-u2
summ u
di "By hand : " %20.0g (sqrt(`r(N)'/`r(Var)') * `r(mean)')
di "From ZIP: " %20.0g e(vuong)
di 1-normprob(e(vuong))
============================================================

From ZIP:    10.59187777147485
. di 1-normprob(e(vuong))
0
```

The result of the ZIP Vuong statistic displayed in the model output is:

```
Vuong test of zip vs. standard Poisson:   z =    10.59  Pr>z = 0.0000
```

A border likelihood ratio test may be calculated for a ZINB model, which is a comparison of the ZINB and ZIP. The model output is on age 375 above. The method and formulae are identical to those that were used to compare NB2 with Poisson models. Recall that the logic of the comparison is based on the distance that the value of heterogeneity or scale parameter, α, is from 0. Is α significantly greater than 0 such that the model is NB2 rather than Poisson (α=0)? Again, the test is given as:

$$LR = -2(\mathcal{L}_P - \mathcal{L}_{NB})$$

or

$$LR = -2(\mathcal{L}_{ZIP} - \mathcal{L}_{ZINB})$$

A modified *Chi2* test is used to evaluate the significance, with one degree of freedom and with the test statistic divided by 2. For the models used in this section, we have

R

```
=======================
LLp <- loglik(modelp)
LLnb <- loglik(model1)
LRtest <- -2*(LLp - LLnb)
dchisq(1, LRtest)/2
=======================

. di -2*(-5394.77 - (-4561.673))
1666.194

. di chiprob(1, 1666.194)/2
0
```

These values are consistent with model output.

==========

P 378: line 3. Replace "to" with "two".

P 378: Table 11.11. For clarity, add line as first in the code: `library(pscl)`

P 382: line 11, replace "probabilites" with "probabilities"

p 389: in the mid-page table of probabilities, "y!" should replace the number "1", which is displayed for each denominator. Therefore, the formulae for each probability should read

$[\mu^{y}e^{-\mu}]/y!$, not $[\mu^{y}e^{-\mu}]/1$

p 454: Table 14.2, should be `corstr="independence"`. Add quotations.

`id=rwm$id` should be `id=rwm5yr$id`.

p 522: Interpretations of interactions at bottom part of page: #'s 1 and 3 need to be corrected to read:

1: The incidence rate ratio of unemployed females to employed females is
exp(.6678495 + (-.4919359 *1)) = 1.192335

Unemployed females have 19% more visits to the doctor during the year than to employed females.

3: The incidence rate ratio of an employed females to unemployed males is
exp(.3826608 + (-.4919359*1)) = .89648376

Unemployed females have 10% fewer visits to the doctor than do unemployed males

COMMENTS

p. 19/20 – comment. The relative risk ratios displayed in the Poisson output under Table 2.4 are identical to the values we created by hand on page 19. If there were interactions in the model, or

if there were more than a single predictor, regardless if it is a categorical variable, the identity would no longer hold. This is implied in the book, but probably should be made explicit.

P 326-328: NB-P. Stata code for constructing a NB-P model is in Hardin & Hilbe (2011). See comment for pages 341-343 below for details. I created synthetic NB2 and NB1 data and models using the algorithms described in this book. I then ran Limdep's NB-P facility on each data, expecting that the NB-P procedure would produce values of P equal to 2 and 1 respectively. It did not, although the other parameter estimates were fine. It appears that Limdep has used Q=P-2 rather than Q=2-P in its estimating algorithm, giving incorrect values for P. I can replicate Limdep output if I use P-2, but doing so results in failure to obtain the correct values of P for "true" synthetically produced data. A new zero-truncated NB-P Stata command (ztnbp) has been authored by Helmut Farbmacher of the University of Munich, Germany.

P 250-255. In the model used for the example, the first 2 levels of "years married", ie, *yrsmarr1* and *yrsmarr2*, are the combined reference level. The *yrsmarr2* level does not differ from *yrsmarr1*.

P 338-339: Fully working Stata commands for generalized Poisson and zero-inflated generalized Poisson models are in Hardin & Hilbe (2012). See below for details.

P 341-343: Fully working Stata commands for PIG and zero-inflated PIG are available in the data sets file for Hardin, JW and JM Hilbe (2012), *Generalized Linear Models and Extensions, 3rd edition*, Stata Press. Chapman & Hall/CRC, owned by Taylor & Francis, markets Stata Press books, as does Stata Corp. The new book will be in print in late May 2012.

Chapters 11-12 (p 346-386). The 3rd edition of James Hardin and Joseph Hilbe, *Generalized Linear Models and Extensions* (Stata Press) was published in May, 2012. We developed complete Stata commands for a new version of generalized Poisson (*gpoisson*), and new commands for zero-inflated generalized Poisson (*zigp*), Poisson inverse-Gaussian (*pigreg*), zero-inflated Poisson inverse-Gaussian (*zipig*), generalized negative binomial or NB-P (*nbregp*), and general censored Poisson and censored negative binomial models, which provide the traditional cut point parameterization, the observation-based censoring and interval censoring, all within a single set of models. These will be available on the book's web site, as well as on a partition for the book in *http://works.bepress.com/joseph_hilbe/*. Other Stata count model commands that are also used in the book will be posted as well, although they may already be posted on the web site for this book. Synthetic models are also displayed in the book, including the following synthetic count models: Poisson, NB2, zero-inflated Poisson, Poisson-logit hurdle, NB-P, 2-component Poisson finite mixture, and a 3-component NB2 finite mixture model.

In addition, Helmut Farbmacher, University of Munich, has authored a zero-truncated Poisson-lognormal (*ztpnm*) command as well as a zero truncated model that provides more a flexible parameter than the lognormal that is associated with *ztpnm*.

p 365-366 Comment: I replaced the Stata command *hnblogit.ado* in the zip file containing the Stata ado and do files used in the book on June 1, 2011. I also added *hnblogit_p.ado* to the list at the same time. An older version of *hnblogit* was inadvertently placed in the initial list of

commands. The changes have been made to the list located on my BePress *Selected Works* site. *http://works.bepress.com/joseph_hilbe/* Changes to other sites may take may take longer.

p 370: Comment: Subsequent to writing the book I developed code to create synthetic zero-inflated Poisson and zero--inflated negative binomial models. The code is given in the COUNT package and displayed in *Negative Binomial Regression Extensions*, which can be downloaded from my Selected Works web site**, http://works.bepress.com/joseph_hilbe/**

==========

Aside from those already acknowledged in the book Preface, I thank the following:
Andreas Krause, Director and Lead Scientist for Modeling and Simulation, Dept. of Clinical Pharmacology, Actelion Pharmaceuticals Ltd, Allschwil, Switzerland, for identifying several of the typo, amendment, and comment items reflected in the above "Errata and Comments".
Wouter Vahl of the Dept of Biosciences, Helsinki University, for identifying typos in Chapter 2
Reginald Jordan and **Timothe'e Vergne**, who each reported several typos throughout the book
Garry Anderson of the University of Melbourne for his continued identification of passages that need clarification.
Ingemar Sjöström, of Ing-Stat, Sweden, for spotting the sign error in table 5.1
Zhenming Su of the Institute for Fisheries Research, Ann Arbor, MI for submitting corrected R code at bottom page 132. He has also pointed out several other items that needed correcting.
Vincent Arel-Bundock. of the Univ. of Michigan for identifying the descrepancy of *ml.nb1* and *ml.nb2* SEs and Stata SEs for the same model
Remzi Gozubuyuk of the faculty of Economics and Administrative Sciences at Ozyegin University, Istanbul, Turkey for identifying a discrepancy in the interpretation of interactions on page 522.

I want to specially acknowledge the contribution of Prof **Andrew Robinson** of the University of Melbourne in constructing and updating the COUNT package, posted on CRAN, which provides the reader with immediate access to R functions, scripts, and data sets written for the text. It is updated on occasion as we add new functions to enhance the R modeling capability. Readers should check for the MSME package which will be posted to CRAN in the near future. It has more functions and scripts for count models, and other models as well, that are discussed in our forthcoming book,
Hilbe, JM and AP Robinson, *Methods of Statistical Model Estimation*, Chapman & Hall/CRC
Note that we amended the original publication date for the book in order to provide more discussion and scope to it. We expect that the book will be available by January 2013.
I again wish to express my thanks to **Tad Hogg** (Institute of Molecular Manufacturing) and **Chris Dorger** (Intel) for their very helpful review of many of the derivations I made throughout the book, **James Hardin** (Univ of South Carolina) who reviewed chapter 14, and **Robert Muenchen** (Univ of Tennessee), who helped resolve various quirks in R. Prof Hardin is also co-author with me of 3 editions of *Generalized Linear Models and Extensions* (2001, 2007, 2012, Stata Press) and two editions of *Generalized Estimating Equations* (2002, 2012, Chapman & Hall/CRC); Prof Muenchen is co-author with me of *R for Stata Users* (2010, Springer).

I encourage readers to send me any typos, suspected errors, or passages that may be in need of clarification or re-wording that they may discover in the book. Send to: **hilbe@asu.edu**

Page 401, Eq (12.13) : the subscripts of the incomplete beta and gamma should be "I". Also in the 3rd line of the equation, y should have asubscript "I". For the two beta_I, how does it defined? do you use the definition given by Abramowitz and Stegun 6.6.1: B_x(a,b) = integral_0^x t^(a-1) (1-t)^(b-1) dt?

(2) Page 372, eq 11.12 and 11.13: should there be parentheses for all the terms after the "ln" function?

(3) Page 373, botton equations: a) there should be no upper limit “n” on the sigma symbols; b) the 3rd line of the equations should be the gradient for gamma, and the 2nd “+” should be “-“?; c) the 2nd line of the gradient for alpha: the scope for I should be “{i: y_i >0}.

(4) Pages 372-373: do you need to change beta1 to gamma?