October 30, 2016

HILBE MCD ERRATA 03Nov2016 update

Joseph M Hilbe

Modeling Count Data

Cambridge University Press: 17 Jul, 2014

Joseph M. Hilbe

ERRATA AND ADDITIONS as of 13Jul, 2015 updated
(red type sometimes given to show new additions)

Page 3, 3rd line from the bottom regular text on the page. (eta) should be (epsilon)

Page 12 bottom paragraphs. Delete the two following paragraphs:
Delete paragraph starting with "Note how sharply peaked the PIG distribution…"

Delete paragraph starting with "The Stata graphic of the same…" The paragraph ends under Figure 1.2B on page 15 where the top 4 lines are deleted as carry over from page 12.

The paragraph starting "A sixth type of …" is where the text resumes after deleting the previous two paragraphs.

Table 1.2a (p13)
Add as top line: `library(gamlss)`
Replace "`y <- (0:140)/10;`" with "`y=seq(0:(obs-1));`"
In the middle of Table 1.2a, replace the 2 lines for *ypig* with the following one line of code

```
ypig = dPIG(y-1, mu, 1/alpha)
```

Fig 1.2a (p14)
Replace Figure 1.2a with the Figure below.

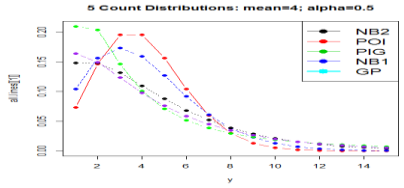
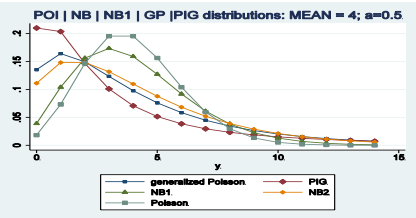


Figure 1.2a. R - Five count distributions mean=4; alpha=0.5

Table 1.2b (p 14)
Delete the line in the middle of the table beginning with "`gen ypig =`". Replace with:

```
PigPr y, m(2) mu(4) pr(ypig)
```

Fig 1.2b (p15)
Replace the Figure in the book with the following:



1.2b Stata graphic of the same distributions as in Figure 1.2a

Section 2.1 (P 37)
Second complete sentence on page (not in table). Amend as displayed below
"In order to properly model zero-excluded the count data having a low mean value, a zero-truncated distribution should normally be used for modeling."

Table 2.6 (p47)
The final line of the code should read as follows: `apply(matrix(unlist(B[1,]),4,100),1,mean)`

Page 50; type directly under the Stata code, " . center age, pre(c). . ."
R
```
> cage <- scale(rwm1984$age, center=TRUE, scale=FALSE)
```

p 68 top of page, delete the R code now in book. Substitute the following in its place. Graph code added since it was supposed to originally be in the book
R
```
library(COUNT); data(rwm5yr); rwm1984 <- subset(rwm5yr, year==1984);
myglm <- glm(docvis ~ outwork + age, family=poisson, data=rwm1984);
lpred <- predict(myglm, newdata=rwm1984, type="link", se.fit=TRUE);
up <- lpred$fit + (1.96 * lpred$se.fit);lo <- lpred$fit - (1.96 * lpred$se.fit);
eta <- lpred$fit; upci <- myglm$family$linkinv(up); mu <- myglm$family$linkinv(eta);
loci <- myglm$family$linkinv(lo); summary(loci); summary(mu); summary(upci);
layout(1); plot(eta, mu); lines(eta, loci, col=2, type='p');
lines(eta,upci, col=3, type='p')
```

p 50. Add following Stata code, ". Center age, pre(c)"

R
> cage <- scale(rwm1984$age, center=TRUE, scale=FALSE)

p 76, 12th line from top. The sentence should read:"If the resulting Chi2 *p*-value is greater than 0.05, …"

Table 3.1 (p77)
In the second half of the table, the word *mymodel* should be substituted for *Model* in the line beginning with "> dev" The "print" line also needs to be amended . The 4 lines in question should be given as follows.
```
dev<-deviance(mymod); df<-df.residual(mymod)
p_value<-1-pchisq(dev,df)
print(matrix(c("Deviance GOF","D","df","p-value", " ",
   round(dev,4),df, p_value), ncol=2))
```

Also, the second paragraph on page and inset, the greater/less than sign should be reversed. That is, "a Chi2 p<0.05" should read "Chi2 p>0.05" In the inset, "p<0.05" should be "p>0.05"

Section 3.4.1 (p 94) R: quasipoisson table. Add a first line to read: data(medpar)

Table 3.8 (p95) The final line in the table should read: `modelfit(poi)`

Page 107 2nd paragraph on page, top line. "discusssed" should be "discussed".

Section 4.1 (p 110) Line under Eq 4.6. Read as $A(\cdot) = \int V^{-1/3}$

Section 4.1 (p 112) Top of page – R code. Add line and amend final plot.

```
> mu <- predict(rwm)
> grd <- par(mfrow = c(2,2))
> respon <- residuals(rwm, type="response")
> plot(x=mu, y= rwm$docvis, main = "Response residuals")
> plot(x=mu, y= respon, main = "Pearson residuals")
```

Table 4.4 (p120)

The *modelfit.R* function has been amended. I amended the *xvars* line in *modelfit.r*, changing the line

`xvars<- x$rank` to read `xvars <- x$df.null - x$df.residual + 1`

Page 127, 3rd paragraph, line 6. "somewhere in the middle" should be changed to that the line now reads: "to the left, right, or to both sides of a distribution of counts."

Section 5.1 (p 128) Amend second to final sentence in section 5.1 Read from second sentence above Section 5.2 as

"Also, see the same source for a discussion on the canonical negative binomial and why it is in general not suitable for modeling count data. It is to my knowledge …"

Section 5.2 (p 131)

8-15th lines in first full paragraph. Amend the lines by adding the text in red.

"as possible. This value is the correct value of the dispersion parameter if the NB2 model is not extradispersed, but the method is tedious. Stata's **glm** and **Genmod** provide an option that calculates maximum likelihood estimates of α in a subroutine, returning the value to the main estimating algorithm as a constant. The result is that both GLM procedures can produce full maximum likelihood estimates of the coefficients and of α. The developers of R's **glm** function did not provide such capability, but a function called **nb.glm**, was authored which calculates the dispersion parameter in the same manner as Stata's **glm** and **Genmod**. The caveat, though, is"

Section 5.3 (p 136)

4th line above section 5.3.1. The word "Poisson" should be "negative binomial". Read as

"model data that are underdispersed modeled using a negative binomial regression, the software"

Table 5.3 (p 143)

Final line in table. Read as: `Physician visits || lowess obpr count, bwidth(.3)`

Section 5.4.3 (p 158)

Third > on page. R code under "NEGATIVE BINOMIAL" near top of page. Add additional ")" at end of line.

```
> summary(NB <- nbinomial(cones ~ sntrees + sheight + scover, data=nut))
```

Section 5.5 (p 161)

Top line on page. Insert word "to" as shown below:

"predictors contribute to model extra-dispersion. Both are valuable diagnostic"

Page 163 5th line under Table 6.1. The world "function" should be "functions."

Equation 6.1 (p164) and associated text
Directly following the inset on page 164 starting with "PIG regression is used…" , please substitute the following text between the lines below for what is now in the book. The correct texts begins again with the paragraph "Again, a value of the model …"

The PIG probability distribution, as a variety of Sichel distribution, can be given as a mixture of the Poisson and inverse Gaussian

$$f(y;\mu,\tau) = \left(\frac{2z}{\pi}\right)^{.5} \frac{\mu^y e^{1/\tau} K_{y/2}(z)}{z t^y y!} \qquad (6.1)$$

with $z = \sqrt{\frac{1}{\tau^2} + \frac{2\mu}{\tau}}$ and where K() is the Bessel function of the third kind. Note that the dispersion parameter, α, is 1/τ in Eq 6.1, and that {μ,α}>0 and y>=0.

Page 164. Bottom two lines on page. Delete the phrase "as we did for the Poisson and NB2 distributions." The period should be placed after "PIG probabilities."

Section 6.1 (P165)
The code at the top of the page should be amended to read, as well as first sentence following the code:
----
probability of

```
> library(gamlss)
> dPIG(2,.5,1)
[1] 0.07050989
```

Where the first term in the parenthesis is the number, the second is the mean and third the value of alpha. We can place this formula …

P 165, top line following mid-page inset. Change to read:
The first two commands listed were written for Hardin & Hilbe (2012), whereas…"

Table 6.3 (p166) and Table 6.5 (p170) Delete the last two lines in each table.

P 166 Statistical output following < not displayed > on page 166 should read.

```
. abic
AIC Statistic   =    4.30344          AIC*n      = 16671.525
BIC Statistic   =   4.303592          BIC(Stata) = 16690.311

. predict munb
(option n assumed; predicted number of events)

. linktest
  _hatsq |  1.059762   .7525146   1.41   0.159   -.4151395    2.534664
```

P 167: mid page, directly under "_hatsq | .3137388 . . ."
Add the following line:
"Note: p=0.021 indicates a violation of linearity assumption"

Section 6.3 (p171)
Rewrite all of page 171 to appear as:
==========

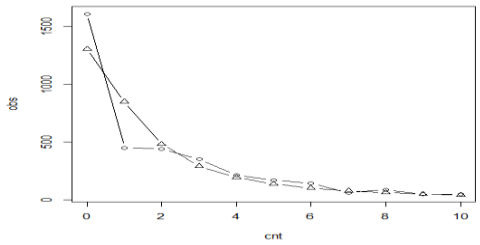


Fig 6.1. PIG model: *docvis* observed vs predicted counts

test model for the **medpar** data is a zero-truncated model. We address both of these models in the following chapter.

Using the **rwm1984** data, we can produce a table of observed vs predicted PIG counts from 0-10 using the following R code. *pigrng.r* is located in the COUNT package and author's website.

```
library(COUNT); library(gamlss); data(rwm1984)
summary(pigmod <- gamlss(docvis ~ outwork + age + married + female +
             edlevel3 + edlevel3 + edlevel4, data=rwm1984, family=PIG))
# PIG dispersion parameter
exp(1.323)
# Predicted probabilities
yp <- pigrng(mean(pigmod$mu.fv),exp(pigmod$sigma.coefficient), 11)
ypig <- yp[ ,2]; ypig
# predicted & observed counts
pigexp <- dim(rwm1984)[1]*ypig ; pigexp
obs=table(rwm1984$docvis)[1:11] ; obs
# table of observed and predicted counts from 0-10
rbind(obs, pigexp[1:11])
chisq.test(obs, pigexp[1:11])
# Figure 6.1
pigpred <- pigexp[1:11]; cnt <- 0:10; plot(cnt, obs)
lines(cnt, obs, type="b"); lines(cnt, pigpred, type="b", pch=24)
```

Note that the line in Figure 6.1 having 0's are the observed counts; the line with triangles's at the count indicator are the predicted counts.
==========

P. 176 mid page small caps. Change to read:
PREDICTED PROBABILITY OF 0 COUNTS FOR MEAN=9.854181

Table 7.1 (p 176) bottom of page
Table 7.1, line 6, change word "ztp" to "pois"
Tablr 7.1, line 7, amend as shown below

```
    gen.trun(0, "PO", type="left", name = "lefttr")
```

Section 7.1.2 (P 178)
Table 7.2, line 5. Change word "ztnb" to "nb2"

2 lines above `ztnb` Stata code, Read as: … the term "1 – Pr(0)" must be …"

Equation 7.5 (p181) The dividing line in the last term under the radical (2/αμ) should be deleted. Read as

$$\Pr(Y = 0) = \exp\left\{\frac{1}{\alpha}\left(1 - \sqrt{1 + 2\alpha\mu}\right)\right\}$$

Section 7.1.4 (p 182) Fourth line from bottom, change word "excessive" to "structurally no"

P 187

Table 7.4: *R Components to Poisson-Logit Hurdle*

```
=================================================
visit <- ifelse(rwm1984$docvis >0, 1, 0)
table(visit)
logis <- glm(visit ~ outwork + age, data=rwm1984,
                     family=binomial(link="logit"))
summary(logis)
library(gamlss); library(gamlss.tr)
pltvis<-subset(rwm1984, rwm1984$docvis>0)
summary(ltpo <- gamlss(docvis~outwork+age,
        family=trun(0, "PO", "left"), data=pltvis))
library(pscl)
hpl2 <- hurdle(docvis ~ outwork + age, data=rwm1984,
    dist = "poisson", zero.dist="binomial", link="logit")
summary(hpl2)
=================================================
```

Section 7.2.1 (p 188)

Top line on page: Change to read as: "It is the same as for the binary component of the hurdle model. Next the count component is modeled:"

P 188 Final two sentences in last full paragraph on page. Change to read as:

"A positive coefficient in the logit frame is interpreted in such a manner that a one-unit change in a coefficient increases the odds ratio of visits to the doctor by exp(β). For example, the logistic coefficient in the hurdle model for *outwork* is 0.5247121; therefore the odds of visiting a physician in 1984 were some 69% greater for out-of-work patients than for working patients [exp(0.5247121) = 1.6899722]"

Table 7.5 (P 191)

This is a NB-logit hurdle model. Therefore `dist="negbin"`, not `dist="poisson"`.

Section 7.2.1 (p 191)

Amend last sentence in middle paragraph on page to read as:

"For example, to obtain marginal effects at means for the negative binomial component of the hurdle model we just reviewed, we do as follows:"

Section 7.2.1 (P 192)

Two lines above Section 7.2.2. "outwork" should be "outwork1" in *tab* command

```
. tab outwork, gen(outwork1)
```

Section 7.3.1 (p 196) Fourth line from the bottom of the first paragraph on page: change word "msme" to "COUNT"

P 196 First "." Inset at bottom of page, change the last part of the second to last line on page to read:
"have a theory as to why there are a class of observations having excess zeros for both observed and expected zero counts. "

Second to last line on page, amend word "supposed" to "supposedly"

Section 7.3.7 (p207) Amend the following line of R code in the middle of the page.

```
summary(zpig <- gamlss(docvis ~ outwork + age, sigma.fo= ~ -1,
    family="ZIPIG", data=rwm1984))
```

Section 7.4 (p 208, 209)
Final sentence on page 208, and top two on page 209, are changed to read as:
"models. We have also discussed the ZI-PIG model. I have placed additional Stata zero-inflated models on the books website, including ZI generalized Poisson, ZI NB-P, ZI heterogeneous NB, ZI generalized Waring and ZI generalized Famoye regressions. Each allows a variety of binary component models."

Chapter 8, p 211: Amend sentence before Eq 8.1 to read:
"The generalized Poisson probability function is based on Consul (1989) and Famoye (1993), and this parameterization and Stata code on Harris, Yang and Hardin (2012).

Add a space and then the following 2 lines under Eq. 8.1

$$\mathcal{L}(\mu; y, \delta) = log(\mu) + (y-1)log(\mu + \alpha\mu) - \mu - \alpha y - \log\big(\text{gamma}(y+1)\big)$$

with μ=θ.

Chapter 8 (p 214)
First paragraph on page under code: In several places words *delta* and *tandelta* should be in italics.
Equation 8.3: *tanhdelta* should be in italics

Final paragraph on page, *delta* in italics. Also amend second and third lines of paragraph, substituting "excess-zeros" for "extradispersed" to read as:
"eling any type of excess-zeros count data. Negative values of *delta* adjust for … Poisson overdispersion. Using a Stata ZIGP command from Hardin & Hilbe (2012), I will revert to …:"

Ch 8, page 216. Add sentences at end of main discussion, just above "Summary".
The *vglm* function is the VGAM package can be used to model the data using R. *vglm* output is the same as Stata's *gpoisson* output, except that the *vglm* dispersion parameter value (intercept:1) appears different from *gpoisson*. But by symbolizing the *vglm* dispersion as V, the value can be converted to Stata's *delta* value by using the formula: delta = (exp(V)-1) / (exp(V) +1). The results are the same. See Hilbe (2011) for details.

To predict GP probabilities when delta<0, use

```
. qui gen `newvar' = min(1,mu*(mu+delta*`l')^(`l'-1) *   ///
    exp(-mu-delta*`l') / exp(lngamma(`l'+1))))
```

To predict GP probabilities when delta>0, use

```
.  qui replace `newvar' = mu*(mu+delta*`l') * (`l'-1) * ///
    exp(-mu-delta*`l') / exp(lngamma(`l'+1))) if `newvar'==.
```

Section 9.1 (p. 218) 3rd line from bottom of page. Change to read as:
Response: los = length (in days) of hospital …

Section 9.1 (p. 222)
R code in unnamed table a little under mid page. Amend as shown below:

```
gen.trun(0,"PO", type="left", name="leftr")
summary(c91c <- gamlss(los~ procedure+type, data=azcabgptca, family=POleftr))
```

Section 9.2.1 (p 225)
4 lines down from Poisson PDF equation which is three lines into section 9.2.1. Replace the equation $(1-\alpha\mu)^{-1/2}$with $(1+\alpha\mu)^{-1/\alpha}$.

The PIG equation needs amending in the following line. Read as: `exp((1/α)*(1-sqrt(1+2αμ)))`

Section 9.2.1 (p 226)
3rd line following equation 9.3: 2014b should read 2015; ie, (Hardin & Hilbe, 2015)

Last two lines on page 226. Amend these two lines to read as:
3-parameter generalized Waring and Famoye count regressions. The latter five models have no prior statistical software implementations. The truncated PIG

p. 226: The Stata command **`treg`** is mentioned twice on the page (the third time has been dropped). The name **`treg`** should be changed to **`trncregress.`**

p. 227: bottom page Stata command. Substitute **`trncregress`** for **`treg`**
p. 228: bottom page Stata command. Substitute **`trncregress`** for **`treg`**

Section 9.2.1 (p228)
Amend final line of R code in table as shown below `family=trun(10, "PO", type="right")))`

Section 9.2.1 (p228)
The Stata *treg* command at bottom page. The first *docvis* term should read "if docvis>0". Read:

```
. treg docvis outwork age if docvis>0 & docvis<19, disp(nbp) ltrunc(0) rtrunc(19
vce(robust) eform
```

Section 9.2.2 (p 229)
Change the slight insets near the bottom of the page. The last sentence in each needs amending.
Left censoring: Left: P(Y ≤ C),
If C=3, 3 is the smallest value in the model. Values that may have been lower are revalued to C. Any response in the data that is less than 3 is considered to be equal to 3.
Right censoring: Right: P(Y ≥C),
If C=15, 15 is the highest observed value in the model. Values that may have been greater in the data are revalued to the value at C. Any response in the data that is greater than 15 is considered to be equal to 15.

Section 9.2.2 (p 230) Italics needed for word visit and complete Stata command near bottom of page. Amend the second and third line from bottom of page to read as:
where counts begin at greater than 2. That is, we created a *visit* variable at "*gen visit* = *docvis*>0" where counts 1 and greater are included. We could

Table 9.4 (p231) Amend lines 4 and the second to last line in table as shown below:

```
cy <- with(lcvis, ifelse(docvis<3, 3, docvis))
lcat30<-gamlss(Surv(cy, ci, type="left") ~ outwork + age,
    data=lcmdvis, family=POlc)
```

Table 9.5 (p231)
Amend last line in table as shown below

```
summary(rcat30<-gamlss(Surv(cy, ci) ~ outwork + age,
   data=rcvis, family=PORc, n.cyc=100))
```

Table 9.6 (p233) New top line in table to read:

```
# flexmix only allows  "gaussian", "binomial", "poisson", and "Gamma" families
```

Amend lines

```
model=list(FLXMRglm(totabund~., family="poisson"),
           FLXMRglm(totabund~., family="poisson")))
```

Section 9.4 (p 236) Amend line directly following equation 9.4 to read as:
"It is important to remember that the original purpose of using GAM was to determine the appropriate transformation needed by a continuous predictor in order to affect linearity, although many analysts now use GAM as a model in its own right. A GAM employs…"

Section 9.4 (p 237) Final sentence on page. Change "…, but SAS does." to "… but SAS and R do."
Section 9.7 (p 246)
Stata commands at bottom of page. Add `. use rwm1984`
above `". global xvar "outwork age female married"`

Section 9.7 (p 246) Equation 9.6. Change to read:

$$f(y) = \frac{\Gamma(\alpha+\rho)\Gamma(k+\rho)}{\Gamma(\alpha+k+\rho)\Gamma(\rho)} \frac{\alpha_y k_y}{(\alpha+k+\rho)_y} \frac{1}{y!}$$

with $\alpha,\rho,k \geq 0$; $y = 0, 1, 2, \ldots$

Section 9.7, (p248)
5th line higher than "Section 9.8". Second "Famoye" should read "Faddy". Change line to:
"discussion. The generalized Famoye negative binomial model, Faddy and "

Section 9.8 (p 249) 2nd paragraph, 8th line from top. Sentence begins with a "For". Change "For": to "Given".
"described by a gamma distribution. Given a binary variable, for instance, we"

Third line from bottom p 249. Insert "informative" between words "no" and "priors".
Section 9.9 (p 253)
Amend bottom lines of middle paragraph beginning with "Truncated and.." Read as:

"Stata's new **`trncregress`** command (Hardin and Hilbe (2015) provides truncated models for Poisson, NB2, PIG, generalized Poisson, NB-P, NB-W, and NB-F. Censored models for these distributions will be available in early 2016."

Appendix (p271) Correct the following entries,

and J.M Hilbe (2012), *Generalized Linear Models and Extensions, third edition*, College Station, TX: Stata Press

and J.M Hilbe (2013), *Generalized Estimating Equations, Second edition*, Boca Raton, FL: Chapman and Hall/CRC.

and J.M. Hilbe, (2014a), "Regression Models for Count Data Based on the Negative Binomial(p) Distribution", *Stata Journal*, 14, 2:180-191.

and J.M. Hilbe, 2015. "Regression models for count models from truncated distributions", *Stata Journal* 15, 1:226-246.

and J.M. Hilbe. 2014b. "Truncation Regression Models for Count Data.", *Stata Journal* 14. 2:280-291.

Harris, T., J.M. Hilbe, and J.W. Hardin 2014, "Modeling Count Data with Generalized Distributions." *Stata Journal*. 14, 3:562-579.