August 13, 2014

Extensions to Modeling Count Data

Joseph M Hilbe, *Arizona State University*

EXTENSIONS TO *MODELING COUNT DATA*

An e-book extension to
Modeling Count Data
Cambridge University Press (2014)

Joseph M. Hilbe

Arizona State University
&
Jet Propulsion Laboratory,
California Institute of Technology

hilbe@asu.edu ; j.m.hilbe@gmail.com
http://works.bepress.com/joseph_hilbe/

THIS EBOOK WILL BE CONTINUALLY UPDATED. CHECK FOR UPDATES ON A REGULAR BASIS.
13 August, 2014

Preface

Extensions to Modeling Count Data is an ebook that provides additional information regarding the topics discussed in my book, *Modeling Count Data (MCD)*, Cambridge University Press (published 17 July, 2014). Mathematical derivations and equations related to various of the count models discussed in *Modeling Count Data* may be found in *Negative Binomial Regression*, 2nd edition (NR2) ,Cambridge University Press, 2011. What is not discussed in NBR2 or in MCD, I hope to address here. This is particularly the case for new count models.

If there is a particular topic you wish to be discussed, please let me know. hilbe@asu.edu or j.m.hibe@gmail.com. All data files and code in MCD may be found in my personal BePress website, listed under my email address displayed on the front page.

I will be adding material to this *ebook* after 8 August, 2014.

Preface

Loading the COUNT package automatically loads the *MASS*, *msme*, and *sandwich* packages. You then have direct access to the data and functions that are included in these packages. The *sandwich* package is required when employing sandwich or robust standard errors, which I recommend to be used with most count models.

Chapter 1 Variety of Count Models

Chapter 2 Poisson Regression

Chapter 3 Testing Overdispersion

I want to make the distinction of the dispersion parameter and dispersion statistic more clear.

DISPERSION PARAMETER

The negative binomial, PIG, and generalized Poisson are examples of count models having two parameters. The primary parameter is μ, or the estimated mean of the response. It is also called the *location* or shape parameter. The second parameter is used to adjust for Poisson overdispersion. We call this the dispersion parameter. It is also termed the ancillary or heterogeneity parameter. At times statisticians refer to it as a scale parameter. Noting that basic binomial and count models assumed to have a scale parameter set at 1.0, other statisticians refer to the dispersion parameter as a second shape parameter, adjusting the mean parameter. Although it is likely less confusing to simply refer to this second parameter as the dispersion parameter and leave it at that, it is likely more accurate to refer to it as a second shape parameter than a scale parameter, as I have at times in the book. On the other hand, the Gaussian scale parameter, σ^2, adjusts for the variability in the Gaussian or normal distribution. In the sense that the normal scale parameter adjusts for variability in the Gaussian PDF, it is similar in function to the negative binomial dispersion parameter. For this reason referring to it as a scale parameter does have a rationale. I suggest that analysts drop referring to the dispersion as a shape or as a scale parameter, and simply refer to it as the dispersion parameter.

Keep in mind that the dispersion parameter, which is many times symbolized as α in statistical model output, has a single value.

DISPERSION STATISTIC

The dispersion statistic is defined as the Pearson Chi2 statistic of a count model divided by the residual degrees of freedom. Simulation studies demonstrate that "true" Poisson models have a dispersion statistic of 1.0. When the dispersion statistic is one, the mean and variance of the response variable, y, are equal. Poisson models with a dispersion>1 are called "overdispersed". Poisson models with a dispersion <1 are called "underdispersed". When the mean and variance are equal, the model is said to be "equi-dispersed."

Do not use the deviance statistic divided by the residual degrees of freedom as a dispersion statistic appropriate to assessing possible extra-dispersion in the data. Simulation studies demonstrate that the 'deviance dispersion' is biased, indicating more variability in the data than there is.

Be sure to keep in mind that the dispersion parameter and dispersion statistic are different. The Poisson dispersion statistic is an indicator that the model may be extra-dispersed, which is a violation of Poisson distributional assumptions. The negative binomial dispersion parameter is a measure of the adjustment made by the model to accommodate Poisson overdispersion in the data being modeled. Negative binomial models can only adjust for Poisson overdispersion.

Three parameter count models add a second dispersion statistic to adjust for extra variability in the two parameter count model.

Chapter 4 Assessment of Fit

I have amended the *modelfit* function in Table 4.4 so that it works properly with the *nbinomial* function. It works fine with other R *glm* and *glm.nb* functions, but *nbinomial* does not save rank as a post-estimation statistic. This should be fixed in the near future. I have amended the *modelfit* function though so that it works for *nbinomial* as well. The *xvars* line needed amendment.

The corrected version of *modelfit* is:

```
# Table 4.4 R: Version of Stata User Command, abic
# =======================================================
modelfit <- function(x) {
obs <- x$df.null + 1
aic <- x$aic
xvars <- x$df.null – x$df.residual + 1
rdof <- x$df.residual
aic_n <- aic/obs
ll <- xvars - aic/2
bic_r <- x$deviance - (rdof * log(obs))
bic_l <- -2*ll + xvars * log(obs)
bic_qh <- -2*(ll - xvars * log(xvars))/obs
c(AICn=aic_n, AIC=aic, BICqh=bic_qh, BICl=bic_l)
}
modelfit(x) # substitute fitted model name for x
```

Chapter 5 Negative Binomial Regression

Chapter 6 Poisson inverse Gaussian Regression

Chapter 7 Problem with Zeros

Chapter 8 Modeling Underdispersed Count Data

Chapter 9 Complex Data: More Advanced Models