Generalized linear models for zero-truncated counts

Daniel G. Kehler and **Ransom A Myers**

Department of Biology,
Dalhousie University,
Halifax, Nova Scotia,
CANADA B3H 4J1

Abstract

We present the framework to model zero-truncated negative binomial and Poisson data via generalized linear mixed models. Zero-truncated data appear in a variety of applied contexts, including situations with large data sets where having a rapid estimation procedure is useful. Generalized linear models are a natural analysis tool for such problems, and the use of random effects a useful tool for modeling unoberservable heterogeneity. Using bycatch data from the U.S. pelagic longline fishery, we demonstrate the application of the GLM context for the truncated negative binomial and Poisson distributions. We illustrate how to overcome several potential pitfalls, including the need for an approximate method for converting truncated into corresponding untruncated fitted values.

Introduction

Truncated observations are routinely encountered in applied contexts such as economics and biology, and often it is the zeros that are missing. For example, only non-zero litter sizes in animal breeding experiments are observed (Foulley et al. 1987), or in on-site sampling of a recreational activity (e.g. sport fishing), data is only recorded on those observed engaged in the activity. A third example, and one we will develop further, is in recording the bycatch of a fishing operation. Commonly, only presence of an organism in a net, or on a line, is recorded, and thus there is no way to distinguish between zeros and missing values. This becomes critical in the evaluation of a temporal trend if the ratio of missing values to zeros has also changed over time.

Given the ubiquity of zero-truncated data, there is need for simple and general analysis tools for inference. The obvious choice for modeling truncated data is to use truncated distributions (Tobin 1958; Grogger and Carson 1991). Several authors have discussed the inherent biases in using non-truncated distributions to model truncated data (Creel and Loomis 1990). Inference, however, is often not confined to the truncated distributions, but to parameters of the distributions that include the unrecorded values. Direct maximization of the likelihood using a gradient search algorithm (e.g. Newton-Raphson (Terza 1985)) or quasi-likelihood methods (Grogger and Carson 1991), have been proposed to obtain consistent parameter estimates from truncated data. These methods have been incorporated into two specialized statistical software packages, GAUSS (Aptech Systems 1989) and LIMDEP (Econometric Software Inc. 2003). Several authors have noted that for the truncated negative binomial (and its special case, the truncated Poisson), generalized linear models (GLMs (McCullagh and Nelder 1989)) and the iteratively reweighted least squares (IRLS) algorithm can also be used. There is a real advantage to using a GLM framework for the analysis of truncated count data. The framework is widely used as it is accessible to a broad range of users familiar with linear models, and is a standard feature of statistical analysis software.

In this paper, we show how GLMs can be used as a fast and reliable framework for the analysis of truncated count data. We extend the GLM analysis by including random effects to account for unobservable variabilty that would otherwise result in violations of the expected variance.

Truncated count distributions

Earlier work has proven the usefulness of Poisson and negative binomial regression models (e.g. Lawton 1987). Modeling truncated count data in a similar fashion is possible, as both the truncated Poisson and negative binomial distribution (with known scale paramter) are one-parameter exponential distribution families. For a discretely distributed random variable, Y, the zero-truncated distribution is of the form

$$P(Y_t = y_t) = \frac{P(Y = y_t)}{1 - P(Y = 0)} \quad \text{for} \quad y_t = 1, 2, 3, \cdots \tag{1}$$

The Poisson is the most commonly used distribution to model counts. If $Z \sim Pois(\mu)$, the zero-truncated Poisson distribution is

$$f_{Z_t}(z_t; \mu) = \frac{\mu^{z_t} e^{-\mu}}{z_t!(1 - e^{-\mu})} \quad \text{for} \quad z_t = 1, 2, 3, ...,$$

with the first two moments

$$E[Z_t] = \frac{\mu}{1 - e^{-\mu}}, \quad V(Z_t) = \frac{\mu + \mu^2}{1 - e^{-\mu}} - \left(\frac{\mu}{1 - e^{-\mu}}\right)^2.$$

We consider the negative binomial as arising from a gamma mixture of Poisson distributions. If the distribution of the unobserved random variable, Z, is gamma with mean 1 and variance $1/\theta$, and the distribution of $Y \mid Z$ is Poisson with mean μZ, then the marginal distribution of Y

is

$$(2) \qquad f_Y(y;\theta,\mu) \quad = \quad \frac{\Gamma(\theta+y)}{\Gamma(\theta)y!}\frac{\mu^y\theta^\theta}{(\mu+\theta)^{\theta+y}}, \quad \text{for} \quad y_t = 0,1,2,...$$

and the zero-truncated distribution is

$$(3) \qquad f_{Y_t}(y_t;\theta,\mu) \quad = \quad \frac{\Gamma(\theta+y_t)}{\Gamma(\theta)y_t!}\frac{\mu^{y_t}\theta^\theta}{(\mu+\theta)^{\theta+y_t}}\left(\frac{1}{1-\left(\frac{\theta}{\theta+\mu}\right)^\theta}\right) \quad \text{for} \quad y_t = 1,2,3,...$$

with mean and variance given by

$$(4) \qquad E[\mathrm{Y}_t] \quad = \quad \frac{E[\mathrm{Y}]}{1-P(\mathrm{Y}=0)} = \frac{\mu}{1-\left(\frac{\theta}{\theta+\mu}\right)^\theta} = \mu_t$$

$$(5) \qquad V(Y_t) \quad = \quad \frac{\mu+\frac{\mu^2}{\theta}+\mu^2}{1-\left(\frac{\theta}{\theta+\mu}\right)^\theta} - \left(\frac{\mu}{1-\left(\frac{\theta}{\theta+\mu}\right)^\theta}\right)^2$$

The Poisson distribution is obtained from the negative binomial by allowing $\theta \to \infty$.

The truncated Poisson distribution can be easily rewritten in exponential form, as can the negative binomial, if θ is treated as fixed.

Specifying the GLM

The use of a GLM requires specifying a link, that describes the relationship betwen the observation scale and the linear predictor (η) scale, and a variance function. The variance functions arise naturally from the distributions above, but leeway exists in the choice of the link function. The canonical link for the zero-truncated negative binomial, $\log\left(\frac{\mu}{\mu+\theta}\right)$ and zero-truncated Poisson, $\log(\mu)$, are both expressed in terms of the untruncated means. This introduces a problem since the fitting algortithm, IRLS, involves minimization of

$$(6) \qquad (y_t-\mu_t)\frac{\partial\eta}{\partial\mu_t}$$

where y_t is the truncated observation, η the linear predictor and μ_t the conditional expectation of Y_t. It is clear that the deviations between the truncated data and the truncated mean are being minimized. Thus, the link function need to be parameterized in terms of the truncated means. We suggest the link $\log(\mu_t-1)$ as an obvious choice, as this maintains the multiplicative error structure, and ensures the proper range for the truncated means ($1 \leq \mu_T \leq \infty$).

A second problem appears, however, since the variance functions are parameterized in terms of the untruncated means. For the Poisson, the functional relationship between the truncated and untruncated means is

(7)
$$\mu_t = \frac{\mu}{1 - e^{-\mu}},$$

which has no analytical solution. For the negative binomial, the functional relationship between the truncated and untruncated means is

(8)
$$\mu_t = \frac{\mu}{1 - \left(\frac{\theta}{\theta+\mu}\right)^{\theta}},$$

and has no explicit solution except when θ =1, and the relationship simplifies to $\mu_t = \mu + 1$. Thus, a numerical approximation is needed. Once this obsctacle is surmounted, estimation proceeds quickly using IRLS. Figure 1 gives the transformation for the Poison and negative binomial distributions.

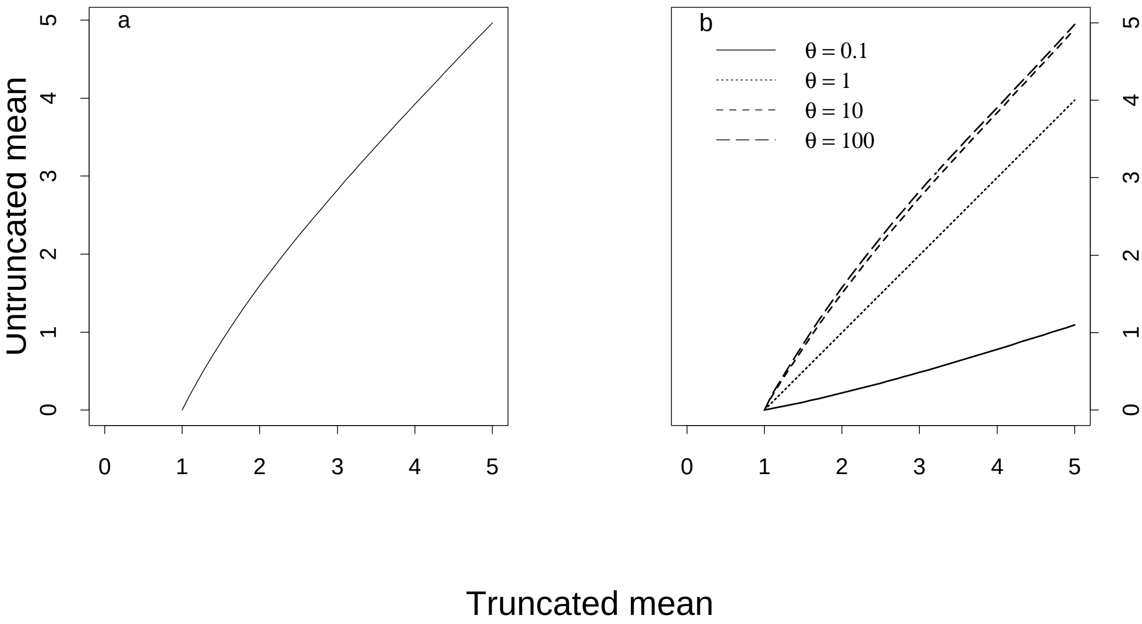

Figure 1: The transformation between the truncated and untruncated mean is shown for the negative binomial (a) and Poisson (b) distributions. For the negative binomial distribution, the transformation is shown for different values of the θ parameter. Remember that the transformation is $\mu_t = \mu/(1 - \left(\frac{\theta}{\theta+\mu}\right)^{\theta})$. For the Poisson distribution, the transformation is $\mu_t = \mu/(1 - e^{-\mu})$, and is well approximated by the negative binomial transformation with larger θ values.

Estimating θ

Note that although θ is fixed in order for the negative binomial to be usable in the GLM context, in practice, θ must be estimated. We follow the advice of Venables and Ripley (1999) and use an iterative approach, alternating fixing θ and the fitted means from the GLM. Alternately, a likelihood profile for θ can be constructed.

Appendix A contains examples of code for S-Plus and SAS to implement both truncated Poisson and negative binmomial GLMs.

Inference on the untruncated scale

Although parameter estimates are easily obtained using the IRLS algorithm implemented in common statistical packages (e.g. Splus, SAS, SYSTAT, SPSS), the inferences drawn from these estimates only apply to the truncated scale, or more specifically, the $\log(\mu_t - 1)$ scale. While this may be adequate for model building, many applications will require inferences to be drawn on the original, untruncated scale (μ scale). What is needed is a way to translate the parameter estimates from the $\log(\mu_t - 1)$ scale to the μ scale. This can be done fairly easily for continuous covariates, by remembering that parameters describe rates and by use of the chain rule for differentiation. For example, in the case of analyzing counts over time (t), interest focuses on the rate of change: $\frac{\partial\mu}{\partial t}$, or on the log scale: $\frac{\partial\log(\mu)}{\partial t}$. The quantity estimated in the truncated GLM is $\frac{\partial\log(\mu_t-1)}{\partial t}$. The relationship between these quantities can be written as :

$$\frac{\partial\log(\mu)}{\partial t} = \left(\frac{\partial\log(\mu_t - 1)}{\partial t}\right)\left(\frac{\partial\log(\mu)}{\partial\log(\mu_t)}\right)\left(\frac{\partial\log(\mu_t)}{\partial\mu_t}\right)\left(\frac{\partial\mu_t}{\partial\log(\mu_t - 1)}\right) \tag{9}$$

The latter three bracketed term represent a correction factor that allows a parameter estimate on the truncated scale to be translated to an estimate on the original scale. Figure 2 gives the relevant correction factor for the zero-truncated Poisson and negative binomial distributions. The remaining step is to choose the appropriate value for μ_t and μ. The use of the mean of the truncated data is a sensible choice for μ_t, and the transformation in Figure 1 gives the corresponding μ value. Appendix B gives the correction factor and its derivation for the Poisson and negative binomial distributions.

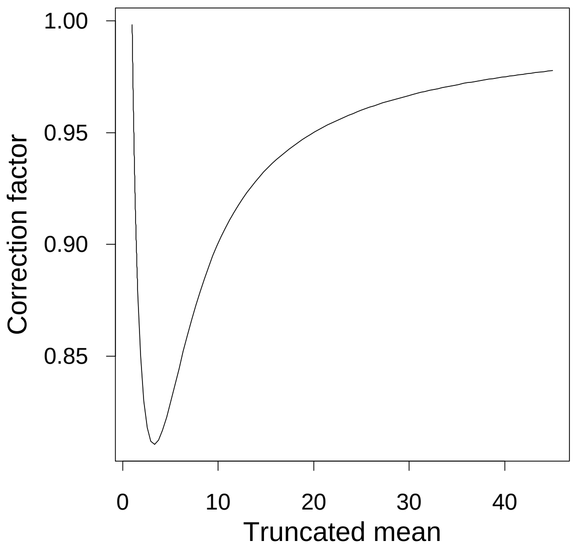

Figure 2: Shown is the correction factor $\frac{\partial \log \mu_T}{\partial \log \mu}$, for the Poisson (a) and negative binomial (b) distributions. In the case of the negative binomial, a contour plot is given, showing the correction factor for different values of μ and θ.

Example - bycatch in the U.S. pelagic longline fishery

Pelagic longlines are a free-floating fishing gear used in open waters to target high valued large pelagic species, like swordfish and tunas. In addition to targeted species, over xxx species are hooked incidentally, and are considered bycatch. There is considerable interest in elucidating any temporal trend in the catch rates for many of these species. Since 1986, U.S. longline boats have been federally mandated to keep logbooks of fishing activity. Detailed information about the positive catch for each longline set is thus available (214234 sets between 1986 and 2000), but absence of catch is not recorded, thus confounding missing values with true zeros. As the data are self-reported their reliability may be questionable, particularly for infrequently caught species. One option is to treat all non-zero entries as zeros, by either analyzing the positive and zero components separately (delta-lognormal method (Lo, Jacobsen, and Squire 1992)), or by accounting for an unexpectedly large proportion of zeros (zero-inflated Poisson (Lambert 1992)). However, if the reporting rate has changed over time, obtaining accurate temporal trends in catch rates is problematic. A second option is to restrict attention to the positives and treat them as a zero-truncated sample. This approach is reasonable if there is little likelihood of the reporting rate of the positive catches changing over time. This is arguably a more realistic assumption in many cases.

As means of illustration, we present results of temporal trends in bycatch for hammerhead sharks using both the Poisson and negative binomial distributions where 1) all non-positives are inferred to be zeros, 2) only positive values are used. We restrict our analysis to one reporting area and one season for simplicity.

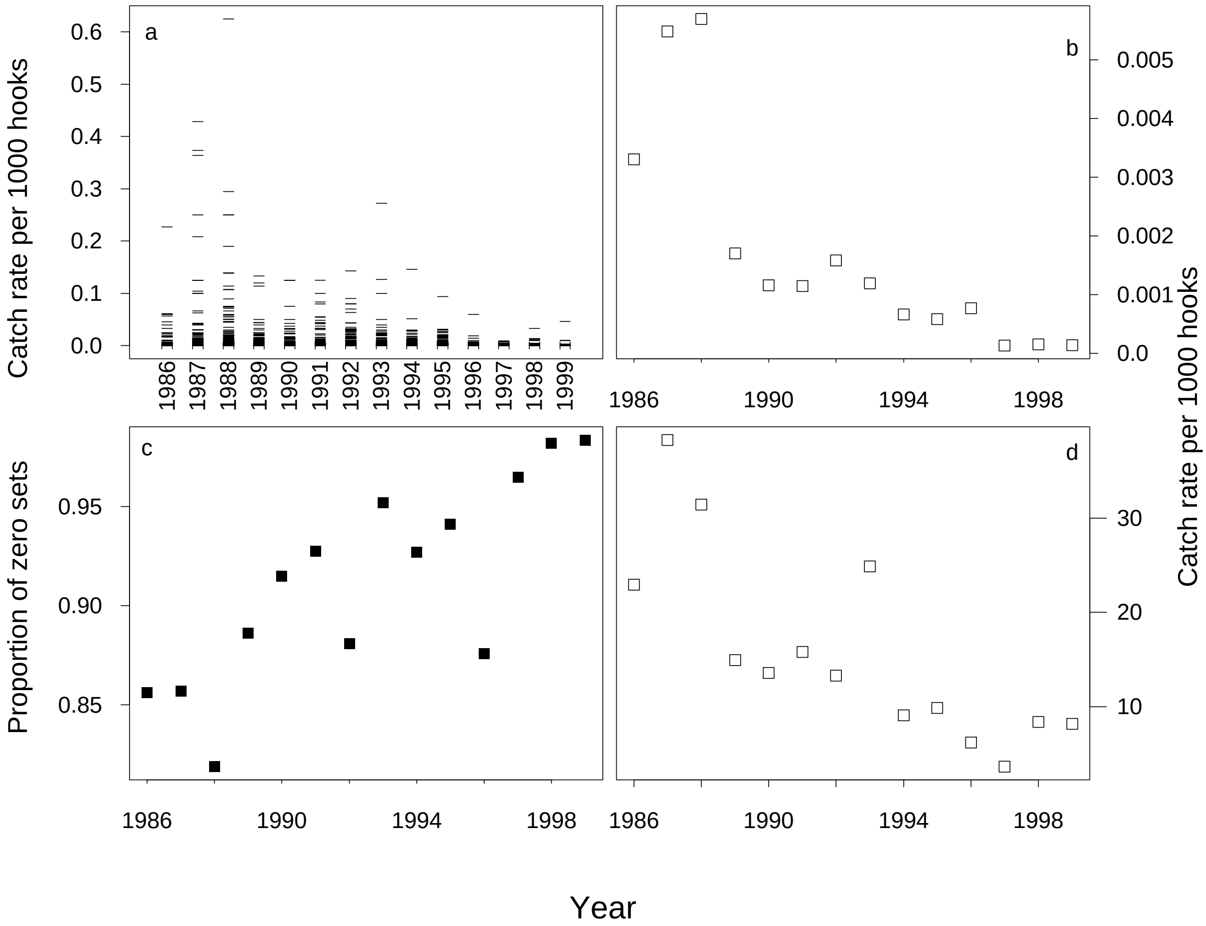

Figure 3: The data from the US logbook programs are shown for one set of species: hammerhead sharks in one reporting area (5) and one season (4). In (a) yearly box plots of the catch rate are presented for all sets ($n = 9683$). In (b), the yearly mean catch rate is shown. In (c) the proportion of sets where no hammerheads were recored are plotted againt year. In (d) the mean catch rate of all non-zero sets ($n = 793$) is plotted against year.

Figure 3 displays the data for the examples. There appears to be a declining trend in the mean catch rate, but at the same time an increase in the proportion of zero sets. Hence, it may be difficult to accurately estimate the trend in catch rate, if the rate at which positive catches are reported is decreasing (The increasing proportion of zero sets). To test the methods outlined above, we fit a series of models to the example data set. The results are presented in Table 1.

Table 1. The parameter estimates of the trend over time in the catch rate per set from various model fits to the hammerhead shark longline data. The model included year as a continuous variable, two additional variables (sea surface temperature and the presence of lightsticks), and the number of hooks per set was used as an offset.

Model	Parameter estimate	SE
Poisson	-0.375	0.0050
Truncated Poisson (glm)	-0.208	0.034
Truncated Poisson (glm - corrected)	-0.183	0.030
Truncated Poisson (ml)	-0.191	
Negative binomial	-0.3	0.034
Truncated negative binomial (glm)	-0.208	0.031
Truncated negative binomial (glm - corrected)	-0.246	0.036
Truncated negative binomial (ml)	-0.241	

It is clear that the inferences from the entire data set (including zero sets) and the zero-truncated data set are quite different. The trend in the proportion of zero sets is not the same as the trend of the positive catches. The corrected estimates match the maximum likelihood estimates quite closely.

1 Acknowledgements

Many thanks to E. Susko, A. Edwards, J. Baum, and W. Blanchard for ideas and helpful comments on the manuscript, and to NSERCC, and the Killam Foundation for funding to DGK and RAM.

Appendix B - Derivation of the correction factor for translating parameter estimates on the truncated scale to the original scale

We know (from the truncated glm slope estimate of the year effect)

$$\frac{\partial \log \mu_T - 1}{\partial t}$$

The (-1) part comes from the link function we are using, $(\log(\mu_t - 1) = \mathbf{X}\beta)$ as this link ensures that all values are ≥ 1 on the η scale. what we really want want to know is how the catches are changing on the untruncated scale:

$$\frac{\partial \log \mu}{\partial t}$$

To obtain this quantity, we use several applications of the chain rule.

$$\begin{aligned}
\frac{\partial \log(\mu)}{\partial t} &= \Big(\frac{\partial \log(\mu_t - 1)}{\partial t}\Big)\Big(\frac{\partial \mu}{\partial \log(\mu_t - 1)}\Big) \\
&= \Big(\frac{\partial \log(\mu_t - 1)}{\partial t}\Big)\Big(\frac{\partial \log(\mu)}{\partial \log(\mu_t)}\Big)\Big(\frac{\partial \log(\mu_t)}{\partial \log(\mu_t - 1)}\Big) \\
&= \Big(\frac{\partial \log(\mu_t - 1)}{\partial t}\Big)\Big(\frac{\partial \log(\ mu)}{\partial \log(\mu_t)}\Big)\Big(\frac{\partial \log(\mu_t)}{\partial \mu_t}\Big)\Big(\frac{\partial \mu_t}{\partial \log(\mu_t - 1)}\Big)
\end{aligned}$$

(10)

The first term is estimated by the GLM on the truncated scale. The second term can be derived by knowing the relationship between the truncated and untruncated means, which we obtain from the appropriate distribution.

- Note 1: we know $\mu_t = f(\mu)$, but we need the relationship between the log of the truncated and untruncated means $\log(\mu_t) = f(\log(\mu))$. This is easily obtained by taking logs of both sides of the equation.

- Note 2: we know $\log(\mu_t) = f(\log(\mu))$ and hence we can obtain the derivative $\frac{\partial \log(\mu_t)}{\partial \log(\mu)}$. However, what we need is actually the inverse: $\frac{\partial \log(\mu)}{\partial \log(\mu_t)}$. This can be obtained by literally taken the inverse of the first derivative.

The remaining terms are easily obtained.

$$\frac{\partial \log(\mu_t)}{\partial \mu_t} = \frac{1}{\mu_t}$$

and

$$\begin{aligned}
\frac{\partial \mu_t}{\partial \log(\mu_t - 1)} &= 1/\frac{\partial \log(\mu_t - 1)}{\partial \mu_t} \\
&= 1/\frac{1}{\mu_t - 1} \\
&= \mu_t - 1
\end{aligned}$$

Negative binomial

For the negative binomial distribtuion, the relationship between the truncated and untruncated means is:

$$\mu_T = \frac{\mu}{1 - \left(\frac{\theta}{\theta+\mu}\right)^{\theta}}$$

and on the log scale:

$$\log(\mu_t) = \log(\mu) - \log\left(1 - \left(\frac{\theta}{\theta+\mu}\right)^{\theta}\right)$$

The derivative of $\log(\mu_T)$ with respect to $\log(\mu)$ is

$$1 - \left(\frac{\mu\theta}{\theta+\mu}\right)\left(\frac{\theta}{\theta+\mu}\right)^{\theta} \frac{1}{1 - \left(\frac{\theta}{\theta+\mu}\right)^{\theta}}$$

The correction factor is thus:

$$\left(\frac{\mu_t - 1}{\mu_t}\right)\left[1 - \left(\frac{\mu\theta}{\theta+\mu}\right)\left(\frac{\theta}{\theta+\mu}\right)^{\theta} \frac{1}{1 - \left(\frac{\theta}{\theta+\mu}\right)^{\theta}}\right]^{-1}$$

Poisson

For the Poisson distribution, the relationship between the truncated and untruncated means is given by:

$$\mu_t = \frac{\mu}{1 - e^{-\mu}}$$

and on the log scale:

$$\log(\mu_t) = \log(\mu) - \log(1 - e^{-\mu})$$

Taking the derivative with respect to $\log \mu$ gives

$$\frac{\partial \log(\mu_t)}{\partial \log(\mu)} = 1 - \frac{\mu e^{-\mu}}{1 - e^{-\mu}}$$

which simplifies to

$$\frac{\partial \log(\mu_t)}{\partial \log(\mu)} = 1 - \frac{\mu}{e^{\mu} - 1}$$

The correction factor is thus:

$$\left(\frac{\mu_t - 1}{\mu_t}\right)\left[1 - \frac{\mu}{e^{\mu} - 1}\right]^{-1} \tag{11}$$

References

Aptech Systems 1989. COUNT module reference list. GAUSS Newsletter **5**: 4–6.

Creel, M. D., and Loomis, J. B. 1990. Theoretical and empirical advantages of truncated count data estimators for analysis of deer hunting in california. American Journal of Agricultural Economics **72**: 434–461.

Econometric Software Inc. 2003. LIMDEP Version 8.0. Econometric Software, Plainview, New York.

Foulley, J. L., Gianola., D., and Im, S. 1987. Genetic evaluation of traits distributed as Poisson-binomial with reference to reproductive characters. Theor. Appl. Genet. **73**: 870–877.

Grogger, J., and Carson, R. 1991. Models for truncated counts. J. of Appl. Econ. **6**: 225–238.

Lambert, D. 1992. Zero-inflated Poisson regression, with an application to defectsin manufacturing. Technometrics **34**: 1–14.

Lo, N. C., Jacobsen, L. D., and Squire, J. L. 1992. Indices of relative abundance for fish spotter data based on delta-lognormal models. Can. J. Fish. Aquat. Sci. **49**: 2515–2526.

McCullagh, P., and Nelder, J. A. 1989. Generalized Linear Models. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Terza, J. V. 1985. A Tobit-style estimator for the censored Poisson regression model. Econ. Lett. **18**: 361–365.

Tobin, J. 1958. Estimation of relationships for limited dependent variables. Econometrica **26**: 24–36.

Venables, W. N. W. N., and Ripley, B. D. 1999. Modern applied statistics with S-PLUS: Volume 1: Data Analysis (Third ed.). Statistics and computing. From the Web site: “The book is also useful with R, a freely-available open-source statistical system ‘not unlike S’. We have tried where possible to use code that works in all versions of S-PLUS and in R.”.