Simulation Study for the Zero-Inflated Negative Binomial

Jelani Wiltshire* David Oakes†

Abstract
In count data there are numerous applications in which the number of observed zeros are substantially larger than the predicted number of zeros under the traditional count data probability models. A popular approach used to model this type of data has been the application of the zero-inflated Poisson (ZIP) model. We applied a variant of the ZIP model, the zero-inflated negative binomial mixed model (ZINBMM) with random effects, to a data set from a longitudinal study. We found that the variance estimates of the some coefficients in the model were unstable. The variance estimate of the intercept parameter in the logistic part of the model was unusually large. There were also several covariates in the model for which the variance of the coefficients were underestimated. In this paper we seek to understand the properties of the variance estimates under this model through the forms of parameter estimation which include maximum likelihood via Gaussian quadrature. This was done through a simulation study which reflected the complex structure of the data in the longitudinal study.

Key Words: Count data, structured zeros, zero inflation,

1 Introduction

In longitudinal data analysis there are times in which the response or outcome of interest takes the form of count data. In a lot of applications of this form of count data the number of observed zeros are much larger than the number of observed zeros that can be predicted by traditional count data models. This is often referred to as "zero-inflation" or "zero-inflated data". Lambert (1992) introduced the zero-inflated Poisson hierarchical regression model to handle zero-inflated cross-sectional count data. This model treats the count data response as a mixture of a Bernoulli random variable and a Poisson random variable. Lambert used the EM algorithm to estimate this model. Hall (2000) extended this model to longitudinal data by including random subjects effect, and again using the EM algorithm. In 2004 Hall and Zhang introduced a marginal models approach to estimating the zero-inflated model by using GEE to model the first two moments of the zero-inflated Poisson response variable. Qu et. al (2013) developed a broader class of models called functional response models (FRMs) and used weighted GEE (WGEE) to estimate the model. The approach used by Qu et. al (2013) relaxes the assumption of missing completely at random and allows for missing at random in the response vector.

*Department of Biostatistics and Computational Biology, URMC, 265 Crittenden Blvd., Rochester, NY 14642
†Department of Biostatistics and Computational Biology, URMC, 265 Crittenden Blvd., Rochester, NY 14642

In a longitudinal dataset that included 75 subjects with 20 repeated measurements per subject, there were an usually large number of zeros (over 30%) in the count response. An application of the ZINBMM model with a random intercept to this data resulted in a very large standard error of the estimate of the intercept. We conducted a simulation study investigate the properties of the intercept and slope parameters of the logistic part of the hierarchical model introduced by Hall (2000). Gaussian quadrature was used to estimate the random effects to estimate the random effects as opposed to the EM algorithm used by Hall (2000). We also compared the predictive probabilities of the this model to two alternatives models on a simulated dataset.

1.1 Primary Goal

The primary research goal for this simulation study was to determine the number of parameters for which the model can provide consistent estimates based on a data structure similar to the longitudinal dataset. Also of interest was determining if the standard error of the estimate of the intercept parameter in the logistic part of the model could be estimated on a data set with a similar structure as the real data.

2 The Model

Below we displayed the zero inflated Poisson regression regression model with random intercept introduced by Hall (2000),

$$\begin{aligned} logit(\mathbf{p_i}) &= \mathbf{X_i}\gamma \\ \log(\mu_i) &= \mathbf{X_i}\beta + Z_i\mathbf{b_i},\, b_i \sim \text{Normal}(0, \sigma_z^2) \end{aligned} \tag{1}$$

where i=1,..,n. The link function that we used looked exactly the same as the above model. The difference lied in the distributional assumptions in the model for $\log(\mu_\mathbf{i})$. We assumed that this was the mean for the negative binomial distribution.

We simulated a model based on a sample size of 75 with 20 repeated measurements per subject. Parameter estimates of the simulation were evaluated at three sets of simulation runs: 500, 1000, and 2000 samples. On the real data the intercept and the slope were unstable therefore we wanted to investigate their performance under simulation. The model for the mean contained an intercept, 19 indicator variables for the repeated measurements and a single covariate. The parameters used for the intercept and slope were -1.1 and 0.38 in the logistic part of the model and 0.42 and -0.01 in the negative binomial part of the model. Random subject effects were generated using a zero mean normal random variable with variance 3.4599 in the negative binomial part of the model.

3 Results

Tables 1 and 2 showed how the intercept and slope of the logistic part of the model perform across the 3 sets of simulations: 500, 1000, and 2000. In Table 1 we see that the intercept estimate fell within one standard error of the true underlying intercept parameter at as little as 500 iterations, with a sample power of 0.638. Though the power was low $(range : 0.638 - 0.66)$, the sample variance of the estimate was also easily estimated while maintaining the properties of the sample covariance matrix. This would suggest that the structure the structure of the data was sufficient to estimate a intercept parameter.

Iterations	Power	Estimate	Variance
500	0.638	-1.17	0.428
1000	0.64	-1.18	0.528
2000	0.6555	-1.17	0.569

Table 1: Table of simulation results for the intercept parameter in logistic part of the model.

Iterations	Power	Estimate	Variance
500	0.846	0.411	0.038
1000	0.855	0.410	0.031
2000	0.851	0.405	0.250

Table 2: Table of simulation results for the slope parameter in logistic part of the model.

The slope parameter (Table 2), which is usually the parameter of primary importance, had large power (> 0.8) at 500 samples. Similarly as in the Table 1, Table 2 shows that slope estimate fell within one standard error of the true underlying slope parameter.

3.1 Additional Results

Alternatives to the ZINBMM include the negative binomial mixed model (NBMM) and the zero-inflated Poisson model (ZIP). In this section we compared the performance of each of these three models to the empirical distribution from the simulated data. This comparison was made by estimating the parameters of each model with a randomly selected data set from our simulation study. We then estimated the difference in the predicted probability and the empirical probability. Hall (2000) and Lambert (1992) also used a similar approach to compare methods for analyzing zero inflated count data.

Figure 1 displays a histogram of a random sample from the simulation study that we used to compare the methods. The sample contains a large number of observed zeros (approximately 47%).

Displayed in figure 2 are differences in the empirical probability and the estimated probabilities at each value (count) of the response the three model, NBMM, ZIP, and ZINBMM, for the sample used in figure 1. It showed the ZINBMM estimated zero probability was closest to the empirical probability and it outperformed the other two methods at all other values of the distribution. The ZIP model was much closer the empirical distribution at zero but the NBMM performed better than ZIP at all other counts of the sample distribution.

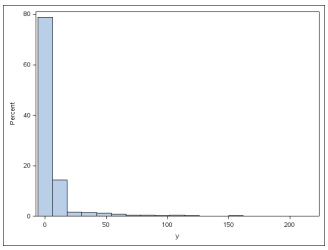

Figure 1: Sample zero-inflated negative binomial distribution.

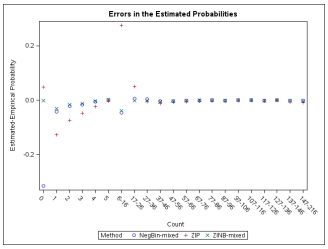

Figure 2: Plot of estimated probability errors($\sum_{i=1}^{N} \hat{P}(Y_i = k)/N - \#(Y_i = k)/N$ where k= count) for three different models: negative binomial mixed model, zero-inflated Poisson, and zero-inflated negative binomial mixed model.

4 Conclusions

Our analysis shows that we should be able to estimate ZINBMM with at least 22 parameters in each part of the model. It may prove to be useful to compare the results of ZINBMM to results from simpler models such as the NBMM or ZIP. In the sample data ZINBMM perform performed better than the alternative approaches. This was expected since the simulation came directly from the ZINBMM.

References

[1] Daniel Hall. Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, 56(4):1030–1039, December 2000.

[2] Daniel Hall and Zhengang Zhang. Marginal models for zero inflated clustered data. *Statistical Modelling*, 4(3):1030–1039, October 2004.

[3] Diane Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, February 1992.

[4] Q. Yu, R. Chen, W. Tang, H. He, R. Gallop, P. Crits-Christoph, J. Hua, and X. M. Tu. Distribution-free models for longitudinal count responses with overdispersion and structural zeros. *Statistics in Medicine*, 32(14):23902405, June 2013.

Acknowledgements

This work was supported by the NIH Training grant 5 T32 ES007271-17 in Environmental Health Biostatistics at the University of Rochester.