Modeling Correlated Counts with Excess Zeros and Time-Dependent Covariates: A Comparison of ZIP and Hurdle Mixed Models

Trent L. Lalonde*

Abstract

Count responses often show an excess of zeros under the assumption of a Poisson distribution. Common modeling solutions include the zero-inflated Poisson model and the hurdle mixture model (Hu et el (2011); Mullahy (1986); Lambert (1992)). Recently researchers have begun to consider the modeling options for clustered or correlated count responses with excess zeros (Kassahun et al, preprint). However, it has yet to be considered whether certain models are preferred for correlated count responses with excess zeros in the presence of time-dependent covariates. Time-dependent covariates have been shown to affect parameter estimate bias and efficiency in longitudinal analyses (Pepe and Anderson (1994); Fitzmaurice (1995); Lai and Small (2007)). In this paper a comparison is made between the zero-inflated Poisson and the hurdle model for correlated count data with time-dependent covariates. Consideration is given to parameter estimate bias and hypothesis test results. An example data set is analyzed, using a longitudinal measure of the number of times of drug use as response.

Key Words: Count Regression, Excess Zeros, Longitudinal, Time-Dependent Covariates

1. Introduction

Count response data are common in applications, and are often modeled using generalized linear models with assumptions of a Poisson response distribution or a Negative Binomial response distribution Cameron and Trivedi (1998); McCullagh and Nelder (1989). However, it is common in practice to see a greater proportion of "zero" values than would be expected by either a Poisson or a Negative Binomial response distribution. If these zero responses are not explained, the model can suffer from underdispersion and struggle to identify significance due to reduced power of standard hypothesis tests.

To account for an excess of zeros in count response data, researchers often turn to joint generalized linear models that combine two components: one that models the probability of observing a zero, and a second that models the positive count process (Hu et al. (2011); Mullahy (1986); Lambert (1992); Welsh et al. (1996)). One option is the Zero-Inflated Poisson (ZIP) model, that includes a logistic regression model for prediction of a "certain zero," and an ordinary Poisson count regression model for prediction of counts. The distribution can be written,

$$f_{ZIP}(y;\pi,\lambda) = \begin{cases} \pi + (1-\pi)f_P(0;\lambda) & y = 0 \\ (1-\pi)f_P(y;\lambda) & y > 0 \end{cases}$$

Here f_P indicates the ordinary Poisson pdf. In this case there are two sources of zeros in the data. One source is a zero produced as part of the normal Poisson distribution that describes the count responses. The second source is a "certain zero," which is produced with probability π.

*Applied Statistics and Research Methods, University of Northern Colorado, Greeley, CO

A second option for modeling excess-zero count data is the Hurdle Poisson (HurP) model, that includes a logistic regression model for prediction of a "certain zero," and a zero-truncated Poisson count regression model for prediction of counts (Hu et al., 2011; Gurmu, 1998). The distribution can be written,

$$f_{HurP}(y;\pi,\lambda) = \begin{cases} \pi & y = 0 \\ (1-\pi)\frac{f_P(y;\lambda)}{1-f_P(0;\lambda)} & y > 0 \end{cases}$$

With the HurP model there is only a single source of zeros in the data, as zeros are only produced as "certain zeros" with probability π.

Both ZIP and HurP models include two systematic components and link functions corresponding to the two joint models,

$$\text{logit}(\pi) = \mathbf{X}_l\boldsymbol{\beta}_l,$$

$$\ln(\lambda) = \mathbf{X}_c\boldsymbol{\beta}_c,$$

where $\mathbf{X}_l$ and $\boldsymbol{\beta}_l$ are the design matrix and parameter vector, respectively, corresponding to the logistic component, and $\mathbf{X}_p$ and $\boldsymbol{\beta}_p$ are the design matrix and parameter vector, respectively, corresponding to the count component. Parameters can be estimated by maximizing the sum of the log-likelihoods of the Bernoulli and Poisson components of each model (Cameron and Trivedi, 1998; Lambert, 1992).

There is a difference in data assumptions implied by using the ZIP model versus the HurP model. The ZIP model inherently assumes there are two processes generating the zeros in the data: there are both "certain zeros" for individuals with specific characteristics, and there are zeros generated as part of the count process for all individuals. On the other hand, the HurP model separately describes zero observations and positive counts. This imposes an assumption that the "certain zeros" are the only zeros possible. If an individual does not have a certain zero, that individual will have a positive count, described by the positive Poisson distribution.

In many data situations, responses are clustered according to physical groups (as in nested or hierarchical data) or according to subject (as in longitudinal or repeated-measures data). In this case the dependence in the responses violates the assumption of independence inherent in the likelihood-based estimation for standard ZIP and HurP models. When data are clustered or longitudinal, models and estimation for excess-zero count responses should reflect this property of the data.

Correlated counts with excess zeros can be modeled using either conditional (or mixed) models, or using marginal estimation methods (Min and Agresti, 2005; Dobbie and Welsh, 2001). In the following sections, mixed ZIP and mixed HurP models will be presented and compared in terms of power and type I error rate; ZIP GEE will be presented and briefly compared to the mixed model options; concluding remarks and recommendations will be presented.

2. Conditional Models for Counts with Excess Zeros

2.1 Mixed ZIP Models

With longitudinal data it is common to apply mixed models to account for the extra-variation caused by repeated observation of individuals. Similarly, correlated subjects are often accounted for using random effects to account for the auto-correlation inherent in homogenous groups of individuals. For both the ZIP and HurP models, the approach of including a random effect to account for auto-correlation has been explored (Min and Agresti, 2005).

The mixed ZIP model can be written as a joint generalized linear mixed model. This model is simile rot the ordinary ZIP model, which is a joint GLM, except that the response distribution is conditional on the random grouping effect. The random component can be written,

$$Y_{ij}|u_i \sim ZIP(\pi, \lambda),$$
$$u_i \sim \mathcal{N}(0, \sigma_u^2),$$

where the conditional pdf of the response is as given in the previous section. The systematic components of the two joint models are similarly updated to include random effects,

$$\text{logit}(\pi) = \mathbf{X}_l\boldsymbol{\beta}_l + \mathbf{Zu},$$
$$\ln(\lambda) = \mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{Zu},$$

where **Z** is he random effects design matrix and **u** is the random effect parameter vector. It is common for the random effects in both the logistic and count components of the mixed ZIP model to be identical. This is because the subject repetition or homogenous grouping in the data should affect both the logistic and count components. The mixed HurP model can be written similarly to the mixed ZIP model, with random component,

$$Y_{ij}|u_i \sim HurP(\pi, \lambda).$$
$$u_i \sim \mathcal{N}(0, \sigma_u^2),$$

where the conditional pdf of the response is as given in the previous section. Systematic components and link functions are identical to those of the mixed ZIP model.

2.1.1 Estimation for Mixed Excess-Zero Models

Estimation of model parameters for both the mixed ZIP and mixed HurP models is typically performed using likelihood-based methods (Min and Agresti, 2005). For this paper and the included simulation, the R package *MCMCglmm* was applied for all model fitting (Hadfield, 2010). This package uses Markov chain Monte Carlo methods, which address the integration of the random effects distribution by estimating the marginal mean through a Markov chain based on an appropriate target distribution.

Alternatively, parameters for correlated ZIP and correlated HurP models can be estimated using marginal methods. For example, a ZIP GEE method was introduced by Dobbie and Welsh (2001). These methods will not be pursued in the current paper.

Table 1: Results of Analyzing EMA Pilot Data

Parameter	**Mixed ZIP**	**Mixed HurP**
$\beta_{l,0}$	-0.41304	0.51807
$\beta_{l,1}$	-0.46604	-0.21967
$\beta_{c,0}$	0.06884	-0.42188
$\beta_{c,1}$	0.03532	0.07880

Table 2: Results of Fitting ZIP-Generated Responses

	Subjects S	Relative Bias β_l	Relative Bias β_c	Error Rate $\hat{\alpha}_c$	Power $\beta_{l,1}$	Power $\beta_{c,1}$
	25	0.2289	0.2523	0.3560	1	0.884
Mixed ZIP	50	0.2211	0.1666	0.3720	1	0.988
	100	0.2209	0.1530	0.3680	1	1
	25	0.7928	0.2307	0.3800	0.904	0.688
Mixed HurP	50	0.7962	0.1475	0.4760	1	0.932
	100	0.7989	0.1122	0.3880	1	1

3. Data Analysis and Simulation Study

A simulation study was performed to compare the performance of the mixed ZIP and mixed HurP models with respect to correlated count responses with time-dependent covariates in terms of bias, power, and Type I error rates. Analysis of the pilot EMA Marijuana data set was used to determine parameters for the simulation study.

The interest in the EMA data is to model Marijuana usage using craving and motivation as predictors. Craving is of primary interest, motivation is treated as a nuisance predictor, and both are time-dependent covariates. Based on the pilot EMA data, both mixed ZIP and mixed HurP models were fit using craving as the only predictor in both the logistic and count components of the model. The following parameter estimates were obtained, as shown in Table 1.

In addition, the variation in craving was found to be 0.2500, and the subject variation from the model was estimated as 0.1886. Using these initial values, data were simulated according to a mixed ZIP model, and data were simulated according to a mixed HurP model. For each type (ZIP or HurP) of simulated data, both the mixed ZIP and mixed HurP models were applied in order to evaluate the consequences of selecting an inappropriate model for correlated counts with excess zeros. The data were generated as balanced, with complete $T = 10$ replicates for each subject, and subject counts of $S = 25$, $S = 50$, and $S = 100$. A total of $M = 250$ replicates was used for each combination of sample size and type of data simulated. For model fitting, the *MCMCglmm* R Package was applied (Hadfield, 2010), with prior covariance structures for the random effects taken to be diagonal with variance 0.002.

The results of fitting both the mixed ZIP and the mixed HurP models to the simulated correlated ZIP responses are shown in Table 2. The Type I error rate is reported only for the count component of the model.

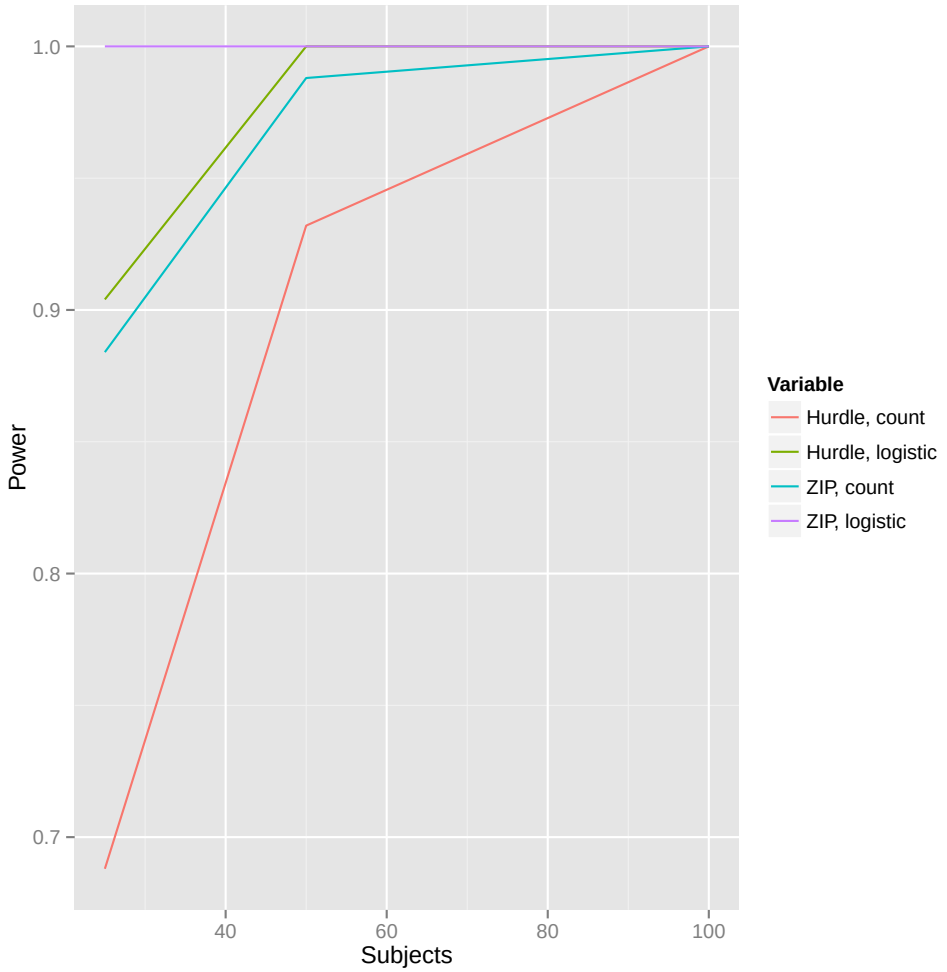

Figure 1: Results for ZIP-Generated Data

Based on these values, a few conclusions are evident. The relative bias for the logistic component is much larger for the incorrect model, and does not improve with increased sample size. Second, the relative bias for the count component decreases with sample size for both the correct and the incorrect models. The type I error rates are extremely large. Finally, both correct and incorrect models have sufficient power, and are probably over-powered for the logistic component of the model. A plot of the power by sample size is shown in Figure 1. While the incorrect model shows the lowest power for the count component, power appears to be sufficient after reaching $S = 50$ subjects.

The results of fitting both the mixed ZIP and the mixed HurP models to the simulated correlated HurP responses are shown in Table 3. The Type I error rate is reported only for the count component of the model.

Based on the values generated, a few conclusions can be made. The relative bias for both components of the model is quite large for the incorrect mixed ZIP model and does not decrease with sample size, while the corresponding smaller relative bias for the correct model does decrease with sample size. The type I error rates are again extremely large, and the logistic component appears to be overpowered. A plot of the power by sample size is shown in Figure 2. The count component does not show sufficient power for smaller sample sizes when using the incorrect model. Overall the incorrect model is much worse

Table 3: Results of Fitting Hurdle-Generated Responses

	Subjects S	Relative Bias β_l	Relative Bias β_c	Error Rate $\hat{\alpha}_c$	Power $\beta_{l,1}$	Power $\beta_{c,1}$
	25	0.2324	0.0365	0.4120	1	0.876
Mixed HurP	50	0.2196	0.0300	0.3400	1	0.980
	100	0.2054	0.0093	0.3240	1	0.996
	25	2.8636	0.6887	0.4160	0.904	0.580
Mixed ZIP	50	2.8146	0.6501	0.5120	0.991	0.808
	100	2.8388	0.6512	0.6000	1	0.952

for mixed HurP data than for mixed ZIP data. In other words, fitting the mixed ZIP model comes with more significant consequences if that model is incorrect.

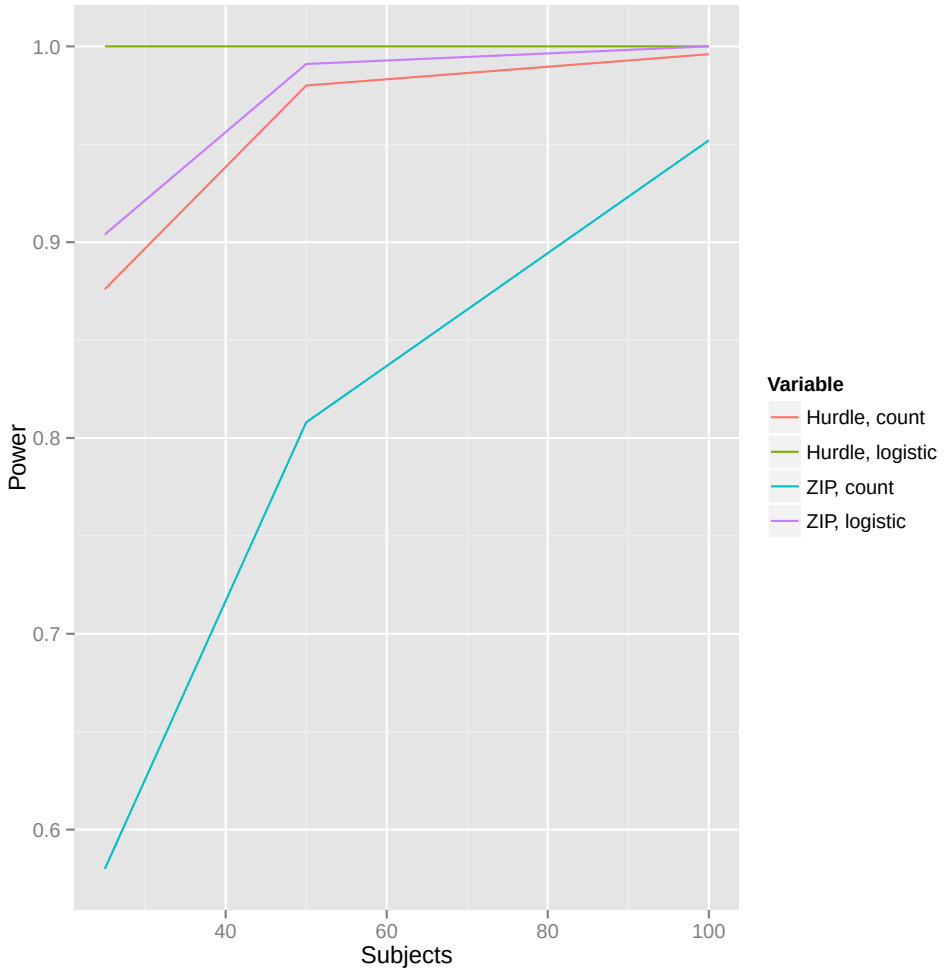

Figure 2: Results for Hurdle-Generated Data

4. Concluding Remarks

Considering the results of the limited simulation study in the previous section, a number of conclusions can be made. When modeling correlated counts with excess zeros described according to the EMA pilot data, both the mixed ZIP and mixed HurP models are extremely liberal when performing hypothesis tests on time-dependent covariates. Specifically, the logistic component of the model is over-powered, rarely failing to reject the null hypothesis. This may be compounded by the omission of a confounding variable in this component of the simulated data. In addition, the type I error rates for the count component of the model were inflated greatly, with unacceptable levels of incorrect rejection. It seem the models struggle to detect variables that are not significant. This suggests there may still be an over dispersion problem in these models when time-dependent covariates are present, as a reduction in parameter estimator standard errors could explain the power and type I error rate issues.

In terms of selecting an appropriate model according the data-generating process, it seems that the mixed HurP model is a "safer" choice than the mixed ZIP model. While the mixed HurP model did suffer some effects from being applied to correlated ZIP data, the consequences of applying the mixed ZIP model to correlated HurP data were much more dra-

matic. Incorrectly applying the mixed ZIP model leads to great increases in parameter estimate bias in both components of the model, along with reduced power.

One consideration in selecting between a ZIP and HurP model is the intentions of the investigator. The ZIP and HurP models are constructed under different assumptions about the response distribution, and consequently are associated with slightly different parameter interpretations. Within any Hurdle model, it is inherently assumed that zero counts are produced only as certain zeros. If a response is *not* to be a zero (and the "hurdle" is cleared) then the response will be strictly positive. Thus the parameters in the logistic component of a Hurdle model represent effects on the probability of an observed zero, and are the only components of the model directly related to observed zeros. The parameters in the count component of a Hurdle model represent effects on the positive counts observed. The Hurdle model also allows the researcher to model a *deflation* of zeros in the response, while the ZIP model does not.

On the other hand, any ZIP model allows zeros to be generated as certain zeros and also as part of the Poisson distribution that also describes the positive observations. While the parameters of the logistic component of a ZIP model represent effects on the probability of an observed zero, similar to the Hurdle model, the parameters of the count component of a ZIP model represent effects on any count. Philosophically, the Hurdle models assume that zero counts are associated with different processes and populations than positive counts. The ZIP model allows that zero counts do not necessarily define a separate population.

A final consideration when selecting between the ZIP and HurP models is software availability. The mixed ZIP and mixed HurP models can both be fit using either SAS or R. In SAS, *PROC NLMIXED* can be used to estimate parameters for both ZIP and HurP and also mixed ZIP and HurP models. However, expressions for the log-likelihoods are required and must be entered. Correspondingly, in R, both the *MCMCglmm* Package and the *glmmACMB* Package can be used to fit mixed ZIP and HurP models. These packages have the ZIP and Hurdle distributions built in, and allow for specification of prior distributions on the dispersion components of the models.

References

Cameron, A. and Trivedi, P. (1998). *Regression Analysis of Count Data*. Cambridge UK. Cambridge University Press.

Dobbie, M. J. and Welsh, A. H. (2001). Modelling correlated zero-inflated count data. *Australian-New Zealand Journal of Statistics*, 43(4):431–444.

Gurmu, S. (1998). Generalized hurdle count data regression models. *Economics Letters*, 58:263–268.

Hadfield, J. D. (2010). Mcmc methods for multi-response generalised linear mixed models: The mcmcglmm r package. *Journal of Statistical Software*, 33(2):1–22.

Hu, M. C., Pavlicova, M., and Nunes, E. V. (2011). Zero-inflated and hurdle models of count data with extra zeros: Example from an hiv-risk reducation intervention trial. *The American Journal of Drug and Alcohol Abuse*, 37:367–375.

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

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models.* London: Chapman & Hall, second edition.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, 5:1–19.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33:341–365.

Welsh, A. H., Cunningham, R. B., Donnelly, C. F., and Lindenmayer, D. B. (1996). Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, 88:297–308.