Statistics
in Medicine

Research Article

Received 9 December 2011, Accepted 11 June 2012 Published online 24 July 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5510

Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field

Anne Buu,[a*†] Runze Li,[b] Xianming Tan[c] and Robert A. Zucker[a]

This study fills in the current knowledge gaps in statistical analysis of longitudinal zero-inflated count data by providing a comprehensive review and comparison of the hurdle and zero-inflated Poisson models in terms of the conceptual framework, computational advantage, and performance under different real data situations. The design of simulations represents the special features of a well-known longitudinal study of alcoholism so that the results can be generalizable to the substance abuse field. When the hurdle model is more natural under the conceptual framework of the data, the zero-inflated Poisson model tends to produce inaccurate estimates. Model performance improves with larger sample sizes, lower proportions of missing data, and lower correlations between covariates. The simulation also shows that the computational strength of the hurdle model disappears when random effects are included. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: hurdle model; zero-inflated Poisson model; random effect; regression spline

1. Introduction

Measuring symptomatology of a target disease longitudinally can provide useful data for assessing disease progression or evaluating long-term effects of treatment or intervention. Substance use disorders are relapsing–remitting in nature [1]. The disease course manifests itself clinically by nondeterministic fluctuations between periods of worsening symptoms and periods of improvement. Such longitudinal trajectories can hardly be delineated by the popular growth curve modeling that employs polynomial functions of time because (i) they have all orders of derivatives everywhere, (ii) polynomial degree cannot be controlled continuously, and (iii) further individual observations can have a large influence on remote parts of the curve [2]. Another common feature of this type of data is that the symptom count measure tends to have excess zero values beyond what would be expected by a classical Poisson model, especially when the sample is drawn from the general population or a community [3]. Moreover, the large individual differences in developmental trajectories commonly observed in the substance abuse field (e.g., [4]) certainly increase the difficulty level of the data analysis.

Longitudinal zero-inflated count data frequently occur in not only the substance abuse field but also in other fields such as healthcare utilization [5], pharmaceutical research [6], and vaccination safety [7]. Thus, statistical models designed for this kind of data have many practical applications. In recent years, the applications of regression splines to longitudinal data analysis have increased dramatically because of their flexibility to different developmental patterns and robust model assumptions. The aim of this paper is to review and compare two competing models for zero-inflated count data, the hurdle model, and the zero-inflated Poisson (ZIP) model, in the setting of longitudinal data analysis with regression splines for modeling longitudinal trajectories as well as random effects for handling within-subject correlation and between-subject heterogeneity. We demonstrate the programming in SAS [8] and conduct

[a]*Department of Psychiatry, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[b]*Department of Statistics and The Methodology Center, Pennsylvania State University, University Park, PA 16802, U.S.A.*
[c]*The Methodology Center, Pennsylvania State University, University Park, PA 16802, U.S.A.*
**Correspondence to: Anne Buu, Department of Psychiatry, University of Michigan, 4250 Plymouth Road, Ann Arbor, MI 48109, U.S.A.*
[†]*E-mail: buu@umich.edu*

simulations to evaluate the performance of the competing models based on the data features of the Michigan Longitudinal Study (MLS), which is an ongoing multiwave prospective study of youth at high risk for alcoholism. The alcohol use disorder (AUD) symptom counts collected from this sample from childhood to adulthood provide a typical example of longitudinal zero-inflated count data.

We organize this paper as follows. In Section 2, we introduce and compare the hurdle model and the ZIP model with an emphasis on the conceptual and computational differences between the models. We also discuss the issues of applications in the substance abuse field. In Section 3, we present a motivational example using the MLS data. In Section 4, we conduct simulation studies to assess the performance of the statistical models under different levels of the sample size, proportion of missing data, and correlation between covariates. We also compare the computational time consumed by the competing models. We present discussion and concluding remarks in Section 5. We provide a SAS program example in Appendix A.

2. The statistical models

2.1. The hurdle model

The hurdle model [9] has mostly been adopted to conduct economic analysis of healthcare utilization. The model postulates a two-stage decision structure in the demand process: the first stage involves the decision to seek care, and the second stage determines how much care is demanded among the subgroup of users for whom the hurdle is crossed. One major strength of the hurdle model is that it can *simultaneously* accommodate two sets of factors that contribute to separate stages.

In a longitudinal setting, suppose that Y_{ij} is the response such as the symptom count for subject i at time t_{ij} $(i = 1, \ldots, n,\ j = 1, \ldots, m_i)$ and it is from a finite mixture:

$$Y_{ij} \sim \begin{cases} 0, & \text{with probability } \phi_{ij} \\ \text{truncated Poisson}(\mu_{ij}), & \text{with probability } 1-\phi_{ij} \end{cases}$$

The probability distribution is thus written as

$$P(Y_{ij} = 0) = \phi_{ij}$$
$$P(Y_{ij} = y_{ij}) = (1-\phi_{ij})\frac{e^{-\mu_{ij}}\mu_{ij}^{y_{ij}}/y_{ij}!}{1-e^{-\mu_{ij}}}, \quad y_{ij} = 1, 2, \ldots$$

Let $\boldsymbol{x}_{ij}$ be the set of factors that contribute to the severity of symptomatology for the people who have developed some symptoms and $\boldsymbol{z}_{ij}$ be another set of factors that account for the probability of being symptom free. The parameters can be modeled by

$$\log(\mu_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + a_i + g(t_{ij}) \tag{1}$$

$$\text{logit}(\phi_{ij}) = \boldsymbol{z}_{ij}'\boldsymbol{\gamma} + b_i + h(t_{ij}), \tag{2}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are fixed effects for covariates $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$, respectively; a_i and b_i are random effects accounting for within-subject correlation and between-subject heterogeneity; $g(t_{ij})$ and $h(t_{ij})$ are baseline functions for the nonlinear time effects. It is assumed that

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim N(\mathbf{0}, \Sigma),$$

where

$$\Sigma = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}.$$

We fit the functions $g(\cdot)$ and $h(\cdot)$ with the following piecewise quadratic polynomials (for the ease of presentation, we drop ij from t):

$$\alpha_0 + \alpha_1 t + \alpha_2 t^2 + \sum_{k=1}^{K} \theta_k (t - t_k)_+^2, \tag{3}$$

where the + indicates the positive value from the expression inside the parentheses; $t_1, \ldots, t_K$ are the knots that divide the range of t into segments. The number of knots controls the amount of smoothing and can be chosen by goodness-of-fit statistics such as BIC. We can use SAS PROC TRANSREG [8] to estimate these regression splines. We include a program example in Appendix A.

By assuming that Y_{ij}, $j = 1, \ldots, m_i$ are conditionally independent given the random effects (a_i, b_i), the likelihood function for data from subject i can be written as

$$L_i = \int \prod_{j=1}^{m_i} P(Y_{ij} = 0)^{I(Y_{ij}=0)} P(Y_{ij} = y_{ij})^{1-I(Y_{ij}=0)} \mathrm{d}F(a_i, b_i; \Sigma), \tag{4}$$

where $I(\cdot)$ is an indicator function; $F(\cdot)$ is the joint distribution function of (a_i, b_i). To obtain the maximum likelihood estimators for the unknown parameters $\boldsymbol{\beta}, \boldsymbol{\gamma}$, and Σ, we need to optimize the likelihood function

$$L = \prod_{i=1}^{n} L_i$$

that involves numerical integration. SAS PROC NLMIXED [8] is equipped with both the numerical integration and optimization methods. Appendix A shows a program example.

2.2. *The zero-inflated Poisson model*

An alternative model for zero-inflated count data, the ZIP model, was originally proposed to model the number of defects on an item in a manufacturing process that is assumed to move randomly back and forth between a perfect state and an imperfect state [10]. Like the hurdle model, the ZIP model can *simultaneously* accommodate one set of factors that make the perfect state more likely and another set of factors that contribute to fewer defects in the imperfect state. The model has been applied in many fields including health science where the population consists of the people who are at risk for a disease and others who are not at risk [7]. The original model has also been extended to accommodate an upper-bounded count situation as well as a repeated measures design [11].

Like the hurdle model, the ZIP model is a finite mixture model:

$$Y_{ij} \sim \begin{cases} 0, & \text{with probability } \pi_{ij} \\ \text{Poisson}(\lambda_{ij}), & \text{with probability } 1 - \pi_{ij} \end{cases}$$

The probability distribution is thus written as

$$P(Y_{ij} = 0) = \pi_{ij} + (1 - \pi_{ij}) e^{-\lambda_{ij}}$$
$$P(Y_{ij} = y_{ij}) = (1 - \pi_{ij}) \frac{e^{-\lambda_{ij}} \lambda_{ij}^{y_{ij}}}{y_{ij}!}, \quad y_{ij} = 1, 2, \ldots$$

We can model the parameters (λ_{ij}, π_{ij}) as Equations (1) and (2) in the hurdle model, respectively. The likelihood function also has the same form as Equation (4) in the hurdle model.

2.3. *Properties of estimators and goodness-of-fit statistic*

We can view both the hurdle and ZIP models with random effects as nonlinear mixed effects models. Vonesh and Chinchilli have provided a systematic account on this class of models [12]. The estimation procedure used in SAS PROC NLMIXED is based on the marginal likelihood. Following the theory of maximum likelihood estimation, the resulting estimators follow an asymptotic normal distribution. Furthermore, the standard errors of the estimates can be computed using the inverse of Fisher information matrix.

SAS PROC NLMIXED uses twice the negative of the log likelihood to measure goodness of fit. This is similar to the deviance used in generalized linear models. Specifically, for the ZIP model with random effects, define

$$d_{ij} = -2 \begin{cases} \log[\hat{\pi}_{ij} + (1 - \hat{\pi}_{ij}) \exp(-\hat{\lambda}_{ij})] & y_{ij} = 0 \\ \log(1 - \hat{\pi}_{ij}) - \hat{\lambda}_{ij} + y_{ij} \log(\hat{\lambda}_{ij}) - \log(y_{ij}!) & y_{ij} > 0 \end{cases}$$

for each observation y_{ij} and

$$D = \sum_{i=1}^{n} \sum_{j=1}^{m_i} d_{ij}.$$

To calculate d_{ij}, we need to have predicted values of the random variables, $\hat{\lambda}_{ij}$ and $\hat{\pi}_{ij}$, which can be estimated using the empirical Bayes method. We can then use the Chi-square test (degrees of freedom = number of observations−number of parameters) to examine goodness of fit. We can conduct a goodness-of-fit test for the hurdle model with random effects in a similar way.

2.4. Applications of models in substance abuse research

Conceptually, the ZIP model is more intuitive when the population consists of a group of people who are at risk for a disease and another group who are not at risk (e.g., women are not at risk for prostate cancer), whereas the hurdle model is more appropriate when all people in the population are considered at risk of an event and the realization of the event represents a hurdle that has been crossed [7]. In substance abuse research, either way of conceptualization makes sense. For example, the ZIP model is applicable when we divide the population into nondrinkers who can *only* have zero alcohol-related symptoms and drinkers who *may* have zero symptoms. For another example, the hurdle model is appropriate if every person in the population is considered at risk for alcohol dependence but only some people meet at least one symptom criterion. For longitudinal studies such as the one described in the next section, people are assumed to be at different levels of risk for alcoholism-related symptoms at different ages. People also change from nondrinkers to drinkers or the other way around across time. The hurdle model tends to have reasonable grounds for such settings.

As shown in previous sections, the hurdle and ZIP models are both finite mixture models – the hurdle mixes 0 and a truncated Poisson, whereas the ZIP mixes 0 and a regular Poisson. The two models can be shown to be mathematically equivalent – one is just a reparameterization of the other when there are no covariates involved [13]. In the case when covariates are included in the models, it is not clear how the two models relate to each other, although it was shown that these two models produce similar estimates and have indistinguishable goodness-of-fit measures in empirical data analysis [7]. Simulation studies are needed to investigate the consequences of using one model when the other model is more natural under the conceptual framework of the data.

The major strength of the hurdle model is that it can handle not only zero-inflated data but also zero-deflated data, whereas the ZIP model can only deal with zero-inflated data [6]. Although this strength makes the hurdle model more applicable in general settings, it is less relevant for substance abuse data because zero-inflated count data are typically the norm, whereas zero-deflated count data are extremely rare in the field. Symptom count data collected from the general population or a community sample tend to be zero-inflated because many participants are nondrinkers or drinkers who have not yet developed symptomatology. Even when we collect symptom count data from a treatment sample such as patients with alcohol dependence, we are likely to observe many participants having only the minimum number of symptoms (e.g., the DSM-IV [14] requires at least three of seven possible symptoms to meet an alcohol dependence diagnosis). The resultant data can again be analyzed by the hurdle or ZIP model with a shift on the location.

Another strength of the hurdle model is its computational simplicity. When there is no random effect, it is shown that the log-likelihood function of the hurdle model can be factored into two terms with one involving $\boldsymbol{\beta}$ (in Equation (1)) and another involving $\boldsymbol{\gamma}$ (in Equation (2)), so one can obtain maximum likelihood estimates by *separately* maximizing the two terms [6]. For the ZIP model, on the other hand, the model components must be fit *simultaneously* and therefore is more complex computationally. Nevertheless, when random effects are included such as Equations (1) and (2), this strength may disappear as the model components must be fit simultaneously for both models because the random effects have to be integrated out in the optimization process (Equation (4)). To the best of our knowledge, the computational time of the two models has never been compared systematically, especially when random effects are involved.

In this study, we aim to extend the applications of the hurdle and ZIP models to analyze longitudinal zero-inflated count data in the substance abuse field. We fill in the current knowledge gaps by conducting simulations on the basis of the data features of a well-known longitudinal study on alcoholism to (i) investigate the consequences of using one model when the other model is more natural under the

conceptual framework of the data; (ii) compare the computational time consumed by the two models; and (iii) evaluate the performance under different sample sizes, proportions of missing data, and correlations between covariates.

3. A motivating example: the Michigan Longitudinal Study

The MLS is an ongoing multiwave prospective study of people at high risk for substance use disorders [15, 16]. The study recruited participant families using fathers' drunk driving conviction records and door-to-door community canvassing in a four-county area in mid-Michigan. All participants received extensive in-home assessments of their psychiatric symptoms including alcoholism-related symptoms at baseline, and thereafter at 3-year intervals. In this study, we use longitudinal data from a sample of 635 children (71% male) for analysis. Their mean age at the latest assessment wave was 20 years.

The following is a brief list of the 11 DSM-IV symptom criteria for AUD [14]:

Abuse symptom 1:	Failure to fulfill major role obligations
Abuse symptom 2:	Hazardous use
Abuse symptom 3:	Legal problems
Abuse symptom 4:	Social or interpersonal problems
Dependence symptom 1:	Tolerance
Dependence symptom 2:	Withdrawal
Dependence symptom 3:	Taken in larger amounts or over a longer period
Dependence symptom 4:	Persistent desire or unsuccessful efforts to cut down
Dependence symptom 5:	A great deal of time spent
Dependence symptom 6:	Important activities given up or reduced
Dependence symptom 7:	Physical or psychological problems

The symptom count (range of 0–11) serves as an important indicator for AUD severity. The zero values in the data from this community sample are more than what would be expected from a classical Poisson regression model (83% across waves). Thus, statistical models for zero-inflated data such as the hurdle and ZIP models are needed.

The substance abuse literature has shown that children of alcoholics, early onset drinkers, men, or youth with high internalizing or externalizing behavior are at a higher risk for progression into AUD [17, 18]. We applied both the hurdle model and the ZIP model to estimate the effects of these risk factors on AUD symptom counts using the MLS data. Table I shows the estimated regression coefficients, standard errors, p-values for t-tests, and goodness-of-fit test results. Most of the estimates in the Poisson component produced by the two models are close, whereas the ones in the zero component are very different. Given the significance level at .05, both models identify externalizing behavior in youth as an important risk factor that contributes to more AUD symptoms and a lower likelihood of being nondrinkers (ZIP) or symptom free (hurdle). Early onset of drinking is found by both models to be a significant factor in the zero component but is only significant in the Poisson component under the hurdle model. Moreover, children of alcoholics are shown by the hurdle model to be less likely to have no AUD symptoms. Although there are significant individual differences (i.e., the random effect) in the Poisson component under both models, the random effect term in the zero component is only found to be necessary under the hurdle model. As shown on the bottom of Table I, both models fit as well as the saturated model.

Figures 1–2 show the spline functions of time. Like the estimates of fixed effects in the Poisson component (shown in Table I), the spline functions in the Poisson component generated by the two models look very similar – the severity level of AUD symptomatology grows rapidly from age 10 to 13 years and stays at the same high level throughout early adulthood (Figure 1). On the other hand, Figure 2 shows that the two models produce very different spline functions in the zero component during early adolescence. The ZIP model indicates that the probability of being a nondrinker increases from age 10 to 13 years, whereas the hurdle model delineates that the probability of being symptom free decreases rapidly during the same period. Both models demonstrate that the corresponding probability gradually decreases year by year from age 13 to early 20s and afterwards stays relatively flat.

The developmental trajectories of AUD symptomatology under the hurdle model are more legitimate as the youth in this high risk sample tend to start drinking earlier and thus are more likely to develop alcohol-related symptoms quickly in early adolescence. Furthermore, for longitudinal studies such as the MLS, people are assumed to be at different levels of risk for alcoholism-related symptoms at

Table I. The fitted ZIP and hurdle models on the MLS data.

	ZIP model			Hurdle model		
Parameter	Estimate	Std. error	p-value	Estimate	Std. error	p-value
The Poisson component						
β_1 Children of alcoholics	0.0663	0.1360	0.6260	0.0809	0.1056	0.4437
β_2 Early onset drinking	−0.0904	0.1619	0.5768	0.2341	0.1060	0.0276
β_3 Male	0.1421	0.1289	0.2705	0.1292	0.1029	0.2096
β_4 Internalizing	0.0789	0.0630	0.2110	0.0918	0.0501	0.0676
β_5 Externalizing	0.1712	0.0673	0.0112	0.1893	0.0578	0.0011
σ_a^2	0.3856	0.0880	0.0001	0.1997	0.0503	0.0001
The zero component						
γ_1 Children of alcoholics	−1.1087	0.7656	0.1481	−0.6016	0.2147	0.0052
γ_2 Early onset drinking	−4.4751	1.9402	0.0214	−1.3614	0.2115	0.0001
γ_3 Male	0.9166	0.8271	0.2682	0.2317	0.2142	0.2798
γ_4 Internalizing	0.2297	0.3417	0.5017	0.1002	0.0983	0.3084
γ_5 Externalizing	−1.7627	0.5236	0.0008	−0.9785	0.1131	0.0001
σ_b^2	15.4720	10.3614	0.1359	2.2824	0.4173	0.0001
σ_{ab}	0.6702	0.6152	0.2764	−0.2760	0.1213	0.0232
Goodness of fit	$\chi^2 = 2310.79$, $df = 3133$, $p \approx 1$			$\chi^2 = 2685.85$, $df = 3133$, $p \approx 1$		

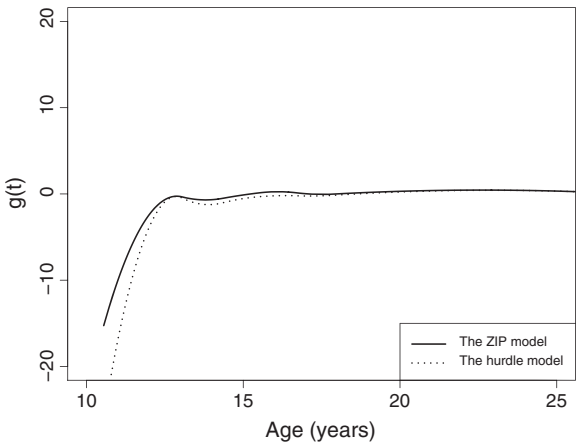

Figure 1. Regression splines: the Poisson component.

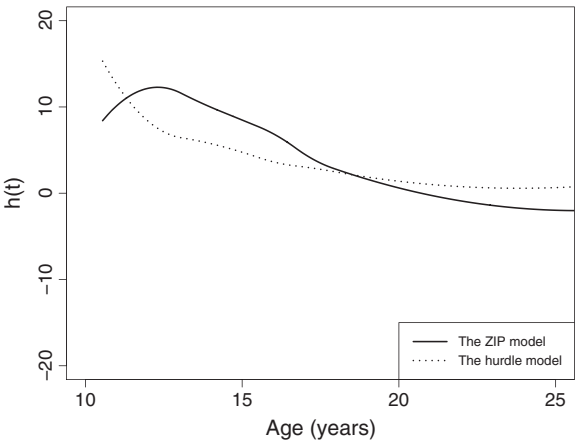

Figure 2. Regression splines: the zero component.

different ages. People also change from nondrinkers to drinkers or the other way around across time. The logic behind the hurdle model thus tends to have reasonable grounds in such a setting. For these reasons, we designate the hurdle model as the *true model* in the simulation study.

4. Simulation study

We adopted the fitted hurdle model in the motivating example as the *true model* to generate simulated data that closely represent the structure of real data so that the results can be more generalizable to the substance abuse field [19]. To generate the five covariates ($\boldsymbol{x} = z$), we drew random samples from a multivariate normal distribution $N_5(\mathbf{0}, \Psi)$, where the diagonal element $\psi_{ii} = 1$ and the off-diagonal element $\psi_{ij} = r^{|i-j|}$ $(i, j = 1, \ldots, 5)$. We employed three levels of correlation in this experiment: small ($r = 0.00$), medium ($r = 0.25$), and large ($r = 0.50$). We also manipulated the sample size at three levels: small ($N = 100$), medium ($N = 200$), and large ($N = 400$). For each subject, we randomly generated a covariate set at each of 20 predetermined assessment waves with a random shift added to each wave so that the subjects may not follow exactly the same assessment schedule such as real data. In addition, we varied the proportion of missing data commonly observed in real data at three levels: small ($p = 0.20$), medium ($p = 0.30$), and large ($p = 0.40$). In summary, we manipulated three factors: the sample size (N), the proportion of missing data (p), and the correlation between covariates (r), with three levels for each factor. Thus, there were in total nine situations. For each situation, we conducted 200 replications.

In each situation, we evaluated the performance of three alternative models: the Poisson regression model, the ZIP model, and the hurdle model. All the models involve both fixed and random effects (i.e., mixed effects). We used the Poisson regression model as a *control model* to examine the consequences of ignoring zero inflation in the data. We compared the performance of the ZIP model with the one of the hurdle model to evaluate the consequences of using the ZIP model when the hurdle model was more natural under the conceptual framework of the data. For the fixed effects $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and the variance–covariance of the random effects Σ, we used the mean squared error (MSE) summarizing the deviation of the 200 estimates from the parameter as the criterion for performance evaluation. For the nonlinear function of time $g(t)$, we calculated the integrated mean squared error (IMSE) for each replication:

$$\text{IMSE} = \int (\hat{g}(t) - g(t))^2 \mathrm{d}t.$$

We used the average of the 200 IMSEs to measure the deviation from the true function. We applied the same computation to the other spline function $h(t)$. To save space, we only show those results featuring the effects of one factor holding the other two factors at their medium values in Tables II–IV. Interested readers may request the tables for other situations from the first author.

Table II shows the effects of varied sample sizes on performance of the three alternative models holding the correlation between covariates and the proportion of missing data at the medium values ($r = 0.25$, $p = 0.30$). For any model, the performance improves (i.e., smaller MSE) as the sample size increases. When the levels of the three factors are fixed, the hurdle model performs better than the ZIP model, especially on estimating the parameters in the zero component. The Poisson regression model performs much worse than the other two models because it ignores the zero inflation in the data.

In Table III, the effects of varied proportions of missing data are depicted with the correlation between covariates and the sample size fixed at the medium values ($r = 0.25, N = 200$). As the proportion of missing data increases, the performance of every model becomes worse (i.e., larger MSE). For any combinations of levels in the three factors, the hurdle model outperforms the ZIP model, especially on estimating the parameters in the zero component. Both the hurdle model and the ZIP model perform much better than the Poisson regression model that imposes an incorrect assumption on the count data.

Table IV summarizes the simulation results with the correlation between covariates manipulated while holding the other two factors constant ($p = 0.30$, $N = 200$). Like the results in Tables II–III, the Poisson regression model tends to produce much larger MSE's than the other two models across different levels of correlation. The performance of the hurdle model tends to decline as the correlation becomes higher. The other two models, on the other hand, do not have a clear pattern of changes with varied correlations.

One objective of this study is to compare the two models in terms of their computational time, particularly when random effects are involved. Our simulation shows that the average computational time

Table II. Mean squared error with varied sample sizes ($r = 0.25$, $p = 0.30$).

	Poisson regression model			ZIP model			Hurdle model		
Parameter	$N = 100$	$N = 200$	$N = 400$	$N = 100$	$N = 200$	$N = 400$	$N = 100$	$N = 200$	$N = 400$
The Poisson component									
β_1	0.0720	0.0675	0.0554	0.0138	0.0042	0.0020	0.0089	0.0039	0.0019
β_2	0.3266	0.2848	0.2813	0.0172	0.0057	0.0022	0.0134	0.0044	0.0023
β_3	0.0355	0.0252	0.0200	0.0150	0.0043	0.0016	0.0090	0.0033	0.0016
β_4	0.0141	0.0118	0.0081	0.0110	0.0050	0.0018	0.0081	0.0038	0.0016
β_5	0.1577	0.1529	0.1384	0.0158	0.0082	0.0020	0.0100	0.0050	0.0019
$g(t)^*$	197.06	147.67	61.58	24.07	22.56	19.86	23.79	21.53	20.34
σ_a^2	0.2247	0.1709	0.1780	0.7660	0.7663	0.7282	0.8870	0.7708	0.6959
The zero component									
γ_1	N/A	N/A	N/A	0.9796	0.1446	0.0305	0.0486	0.0145	0.0068
γ_2	N/A	N/A	N/A	3.1837	0.3632	0.0820	0.0674	0.0245	0.0114
γ_3	N/A	N/A	N/A	0.6031	0.1361	0.0447	0.0294	0.0143	0.0067
γ_4	N/A	N/A	N/A	0.3763	0.0721	0.0295	0.0228	0.0136	0.0064
γ_5	N/A	N/A	N/A	1.5438	0.2396	0.0429	0.0512	0.0187	0.0094
$h(t)^*$	N/A	N/A	N/A	20.08	18.42	13.98	13.83	13.12	11.14
σ_b^2	N/A	N/A	N/A	3.1340	1.7310	1.4238	1.0860	1.0362	1.0361
σ_{ab}	N/A	N/A	N/A	2.1606	0.3449	0.1214	0.0979	0.0534	0.0179

*The average of the integrated mean squared error was calculated for the spline function.

Table III. Mean squared error with varied proportions of missing data ($r = 0.25$, $N = 200$).

	Poisson regression model			ZIP model			Hurdle model		
Parameter	$p = 0.20$	$p = 0.30$	$p = 0.40$	$p = 0.20$	$p = 0.30$	$p = 0.40$	$p = 0.20$	$p = 0.30$	$p = 0.40$
The Poisson component									
β_1	0.0601	0.0675	0.0635	0.0038	0.0042	0.0058	0.0036	0.0039	0.0043
β_2	0.2720	0.2848	0.3036	0.0047	0.0057	0.0085	0.0044	0.0044	0.0069
β_3	0.0221	0.0252	0.0269	0.0028	0.0043	0.0075	0.0027	0.0033	0.0052
β_4	0.0104	0.0118	0.0114	0.0035	0.0050	0.0066	0.0033	0.0038	0.0056
β_5	0.1403	0.1529	0.1512	0.0038	0.0082	0.0095	0.0033	0.0050	0.0050
$g(t)^*$	114.49	147.67	160.37	21.30	22.56	21.83	25.41	21.53	21.45
σ_a^2	0.1822	0.1709	0.1365	0.7517	0.7663	0.7422	0.7330	0.7708	0.7795
The zero component									
γ_1	N/A	N/A	N/A	0.0785	0.1446	0.2256	0.0130	0.0145	0.0199
γ_2	N/A	N/A	N/A	0.1759	0.3632	0.5452	0.0186	0.0245	0.0280
γ_3	N/A	N/A	N/A	0.0776	0.1361	0.1664	0.0109	0.0143	0.0183
γ_4	N/A	N/A	N/A	0.0538	0.0721	0.1156	0.0110	0.0136	0.0192
γ_5	N/A	N/A	N/A	0.1133	0.2396	0.3388	0.0157	0.0187	0.0221
$h(t)^*$	N/A	N/A	N/A	15.16	18.42	23.57	12.18	13.12	13.57
σ_b^2	N/A	N/A	N/A	1.6315	1.7310	1.9227	1.0753	1.0362	1.1108
σ_{ab}	N/A	N/A	N/A	0.2874	0.3449	0.5669	0.0350	0.0534	0.0649

*The average of the integrated mean squared error was calculated for the spline function.

Table IV. Mean squared error with varied correlations between covariates ($p = 0.30$, $N = 200$).

	Poisson regression model			ZIP model			Hurdle model		
Parameter	$r = 0.00$	$r = 0.25$	$r = 0.50$	$r = 0.00$	$r = 0.25$	$r = 0.50$	$r = 0.00$	$r = 0.25$	$r = 0.50$
The Poisson component									
β_1	0.0631	0.0675	0.0587	0.0043	0.0042	0.0047	0.0037	0.0039	0.0046
β_2	0.3025	0.2848	0.2679	0.0058	0.0057	0.0064	0.0060	0.0044	0.0061
β_3	0.0270	0.0252	0.0231	0.0042	0.0043	0.0062	0.0038	0.0033	0.0049
β_4	0.0095	0.0118	0.0127	0.0054	0.0050	0.0060	0.0036	0.0038	0.0056
β_5	0.1454	0.1529	0.1401	0.0050	0.0082	0.0055	0.0041	0.0050	0.0050
$g(t)^*$	127.31	147.67	125.03	20.42	22.56	21.74	22.55	21.53	22.48
σ_a^2	0.1626	0.1709	0.1660	0.7695	0.7663	0.7255	0.7394	0.7708	0.7539
The zero component									
γ_1	N/A	N/A	N/A	0.1328	0.1446	0.1010	0.0132	0.0145	0.0212
γ_2	N/A	N/A	N/A	0.3293	0.3632	0.2694	0.0231	0.0245	0.0358
γ_3	N/A	N/A	N/A	0.1118	0.1361	0.1128	0.0096	0.0143	0.0192
γ_4	N/A	N/A	N/A	0.0792	0.0721	0.0698	0.0127	0.0136	0.0185
γ_5	N/A	N/A	N/A	0.2657	0.2396	0.1241	0.0192	0.0187	0.0271
$h(t)^*$	N/A	N/A	N/A	16.10	18.42	15.73	12.62	13.12	12.07
σ_b^2	N/A	N/A	N/A	1.7645	1.7310	1.6049	1.0649	1.0362	1.0481
σ_{ab}	N/A	N/A	N/A	0.4622	0.3449	0.2564	0.0386	0.0534	0.0558

*The average of the integrated mean squared error was calculated for the spline function.

per replication is 15 min and 45 s for the ZIP model; it is 17 min and 24 s for the hurdle model. Therefore, this confirms our hypothesis that the computational strength of the hurdle model disappears when random effects are included in the model.

5. Discussion

This study has filled in the current knowledge gaps in statistical analysis of longitudinal zero-inflated count data by providing a comprehensive review and comparison of the ZIP and hurdle models in terms of the conceptual framework, computational advantage, and performance under different real-data situations. The design of our simulation study is unique because it represents the special features of a well-known longitudinal study on alcoholism risk so that the results can be generalizable to the substance abuse field.

Our simulation results demonstrate the danger of using the Poisson regression model to analyze longitudinal count data when excess zeros exist in the data: the model tends to produce much higher MSEs than the hurdle and ZIP models. On the basis of both the simulations and empirical analysis of real data, when the hurdle model is more natural under the conceptual framework of the data, the ZIP model tends to result in relatively high MSEs, particularly in the zero component. Moreover, the performance of the models improves with a larger sample size, lower proportion of missing data, and lower correlation between covariates. The simulation also shows that the computational strength of the hurdle model disappears when random effects are included in the model.

According to our comparison between the hurdle and ZIP models in Section 2.4 and the simulation results, we would like to provide a general guideline on applications of these models for practitioners. If there exist a group of subjects who are at risk for an event and another group who are not at risk, the ZIP model captures the conceptual framework better than the hurdle model that is more appropriate when all the subjects are considered at risk for an event. Both models are designed to address the issue of zero-inflated data. However, only the hurdle model can handle zero-deflated data. When there is no random effect (or the random effect is ignorable), the hurdle model has computational advantage. But such advantage should play a minimal role in choosing a model because the conceptual framework and features of the data are more important.

In this study, we only consider the Poisson link function for both the hurdle and ZIP models because it fits well with the real data example. In some applications where overdispersion is common, such as analyzing vaccine adverse event count data [7], the negative binomial link should be considered.

In our analysis, we implicitly assumed that the mechanism of missing data is missing completely at random, so we only needed to model the response process with the use of available observations. It would be an interesting research topic to consider other patterns of missing data such as missing at random, where the missing process is correlated with the response process. For the case of missing at random, to obtain consistent and unbiased estimates of the parameters for the response process, we need to take into account the missing process appropriately (e.g., inverse probability approach), and we should model both the missing process and the response process. This may be a topic for future research.

As shown by the simulations, the average computational time for fitting either model is 15–17 min because it involves both numerical integration and optimization procedures. Future methodological work is needed to improve the computational efficiency in real-life applications.

Appendix A. SAS program for hurdle and ZIP models with random effects and regression splines

We carried out all programming in SAS version 9.2. We employed PROC TRANSREG to construct piecewise quadratic polynomials in Equation (3) with the use of degree two B-splines (`DEGREE=2`) that have nice numerical properties and are easy to manipulate [20]. We chose five knots (`NKNOTS=5`) because the corresponding model fits the MLS data well. This results in 5 + 3 terms, `t_0-t_7`, that were stored in the data set `basis`.

```
PROC TRANSREG DATA= mls;
     MODEL IDENTITY(y) = BSPLINE(t / DEGREE=2 NKNOTS = 5 );
     OUTPUT OUT = basis PREDICTED;
RUN;
```

We merged the resulting transformations of t with the original data set `mls`, which contains the ID number (`target`), the outcome (`y`), and the covariates (`coa`, `earlyos`, `male`, `inttscr`, `exttscr`).

```
DATA final;
     MERGE mls basis;
     KEEP target y coa earlyos male inttscr exttscr t_0-t_7;
RUN;
```

We used PROC NLMIXED to fit the hurdle model as described in Section 2.1. The variables created in the program are defined as follows:

`infprob`: The probability of being symptom-free ϕ_{ij}
`linpinfl`: The logit(ϕ_{ij}) in Equation (2)
`mu`: The mean for truncated Poisson μ_{ij}
`sp_eff`: The random effect b_i in Equation (2)
`varsp`: The variance of b_i
`sl_eff`: The random effect a_i in Equation (1)
`varsl`: The variance of a_i
`cors`: The correlation between a_i and b_i
`ll`: The log-likelihood function

```
PROC NLMIXED DATA=final;
     varsp = exp(2*logsp);
     varsl = exp(2*logsl);
     linpinfl = sp_eff+a1*coa+a2*earlyos+a3*male+a4*inttscr
                +a5*exttscr+a6*t_0+a7*t_1+a8*t_2+a9*t_3+a10*t_4
                +a11*t_5+a12*t_6+a13*t_7;
     infprob = 1/(1+exp(-linpinfl));
     mu = exp(sl_eff+b1*coa+b2*earlyos+b3*male+b4*inttscr+b5*exttscr
              +b6*t_0+b7*t_1+b8*t_2+b9*t_3+b10*t_4+b11*t_5
              +b12*t_6+ b13*t_7);
     IF y=0 THEN
          ll = log(infprob);
     ELSE ll = log((1-infprob))-mu+y*log(mu)-lgamma(y+1)
               -log(1-exp(-mu));
     MODEL y ~ general(ll);
     RANDOM sp_eff sl_eff
            ~ NORMAL([0,0],
                     [varsp,
                      cors, varsl])  SUBJECT=target;
     ESTIMATE 'Varsp     ' varsp;
     ESTIMATE 'Varsl     ' varsl;
     ESTIMATE 'cors      ' cors;
 RUN;
```

We used PROC NLMIXED to fit the ZIP model as described in Section 2.2. All the codes are identical to the codes for the hurdle model except for the log-likelihood function `ll`.

```
PROC NLMIXED DATA=final;
     varsp = exp(2*logsp);
     varsl = exp(2*logsl);
     linpinfl = sp_eff+a1*coa+a2*earlyos+a3*male+a4*inttscr
                +a5*exttscr+a6*t_0+a7*t_1+a8*t_2+a9*t_3+a10*t_4
                +a11*t_5+a12*t_6+a13*t_7;
     infprob = 1/(1+exp(-linpinfl));
     mu = exp(sl_eff+b1*coa+b2*earlyos+b3*male+b4*inttscr+b5*exttscr
              +b6*t_0+b7*t_1+b8*t_2+b9*t_3+b10*t_4+b11*t_5
              +b12*t_6+b13*t_7);
```

```
      IF y=0 THEN
           ll = log(infprob+(1-infprob)*exp(-mu));
      ELSE ll = log((1-infprob))-mu + y*log(mu)-lgamma(y+1);
      MODEL y ~ general(ll);
      RANDOM sp_eff sl_eff
             ~ NORMAL([0,0],
                      [varsp,
                       cors, varsl])  SUBJECT=target;
      ESTIMATE 'Varsp     ' varsp;
      ESTIMATE 'Varsl     ' varsl;
      ESTIMATE 'cors      ' cors;
RUN;
```

Acknowledgements

Buu's research was supported by a National Institutes of Health (NIH) grant, K01 AA16591; Li's research was supported by NIH grants, R21 DA024260 & P50 DA10075, and a National Science Foundation (NSF) grant DMS 0348869; Tan's research was supported by an NIH grant P50 DA10075; and Zucker's research was supported by an NIH grant R37 AA07065. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the NSF.

References

1. McLellan AT, Lewis DC, O'Brien CP, Kleber HD. Drug dependence, a chronic medical illness: implications for treatment, insurance, and outcomes evaluation. *Journal of The American Medical Association* 2000; **284**:1689–1695. DOI: 10.1001/jama.284.13.1689.
2. Fan J, Gijbels I. *Local Polynomial Modeling and Its Applications*. Chapman and Hall: London, UK, 1996.
3. Buu A, Johnson NJ, Li R, Tan X. New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine* 2011; **30**:2326–2340. DOI: 10.1002/sim.4268.
4. Hill KG, White HR, Chung IJ, Hawkins JD, Catalano RF. Early adult outcomes of adolescent binge drinking: person- and variable-centered analyses of binge drinking trajectories. *Alcoholism: Clinical and Experimental Research* 2000; **24**:892–901. DOI: 10.1111/j.1530-0277.2000.tb02071.x.
5. Alfo M, Maruotti A. Two-part regression models for longitudinal zero-inflated count data. *The Canadian Journal of Statistics* 2010; **38**:197–216. DOI: 10.1002/cjs.
6. Min Y, Agresti A. Random effect models for repeated measures of zero-inflated count data. *Statistical Modeling* 2005; **5**:1–19. DOI: 10.1191/1471082X05st084oa.
7. Rose CE, Martin SW, Wannemuehler KA, Plikaytis BD. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics* 2006; **16**:463–481. DOI: 10.1080/10543400600719384.
8. SAS Institute Inc. *SAS/STAT 9.2 User's Guide*. SAS Institute Inc: Cary, NC, 2008.
9. Mullahy J. Specification and testing of some modified count data models. *Journal of Econometrics* 1986; **33**:341–365. DOI: 10.1016/0304-4076(86)90002-3.
10. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**:1–13. DOI: 10.2307/1269547.
11. Hall DB. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* 2000; **56**:1030–1039. DOI: 10.1111/j.0006-341X.2000.01030.x.
12. Vonesh E, Chinchilli VM. *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. Marcel Dekker: New York, 1997.
13. Baughman AL. Mixture model framework facilitates understanding of zero-inflated and hurdle models for count data. *Journal of Biopharmaceutical Statistics* 2007; **17**:943–946. DOI: 10.1080/10543400701514098.
14. American Psychiatric Association. *Diagnostic and Statistical Manual. 4th edition*. American Psychiatric Association: Washington DC, 1994.
15. Zucker RA, Ellis DA, Fitzgerald HE, Bingham CR, Sanford K. Other evidence for at least two alcoholisms II: life course variation in antisociality and heterogeneity of alcoholic outcome. *Development and Psychopathology* 1996; **8**:831–848.
16. Zucker RA, Fitzgerald HE, Refior SK, Puttler LI, Pallas DM, Ellis DA. The clinical and social ecology of childhood for children of alcoholics: description of a study and implications for a differentiated social policy. In *Children of Addiction*, Fitzgerald HE, Lester BM, Zuckerman BS (eds). Garland Press: New York, 2000; 109-141.
17. Hussong A, Bauer D, Chassin L. Telescoped trajectories from alcohol initiation to disorder in children of alcoholic parents. *Journal of Abnormal Psychology* 2008; **117**:63–78. DOI: 10.1037/0021-843x.117.1.63.
18. Guo J, Hawkins JD, Hill KG, Abbott RD. Childhood and adolescent predictors of alcohol abuse and dependence in young adulthood. *Journal of Studies on Alcohol* 2001; **62**:754–762.
19. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine* 2006; **25**:4279–4292. DOI: 10.1002/sim.2673.
20. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer-Verlag: New York, 2001.