Zero-Inflated Model: Is it Sufficient for Estimating Excess Zeros?

Charles E. Rose[*]

Abstract
Count data frequently exhibit overdispersion due to an excess of zeros, unexplained heterogeneity, and/or temporal dependency. Zero-inflated (ZI) models have gained in popularity, especially during the past decade, as the modeling preference for count data with excess zeros, and these models have become standard in most statistical software. Conceptually, zeros in ZI models are assumed to be a result of two types: structural and sampling. For example, in response to the question "How often did you drink alcohol during the last 30 days?" there will be individuals who drink alcohol but chose not to drink during the last 30 days (sampling zeros) and individuals who never drink alcohol (structural zeros). I address how adequately ZI models estimate the structural and sampling zeros and the resulting impact on inference. I simulate ZI data using several scenarios for the proportion of structural and sampling zeros and non-zeros. The scenarios demonstrate that the estimated structural and sampling zeros may be biased, and I discuss the impact of ZI model bias on inference.

Key Words: zero-inflated Poisson, ZIP, excess zeros, structural zeros, sampling zeros

1. Introduction

Poisson and negative binomial (NB) regression models are commonly used in epidemiology to model count and rate data. The Poisson model assumes the variance is equal to the mean and this assumption is often violated. The NB model has a dispersion parameter that can account for variance greater than the mean (over-dispersion) that is due to unobserved heterogeneity and/or temporal dependency (Chin and Quddus). Over-dispersion in count data occurs frequently due to excess zeros, unexplained heterogeneity, and/or temporal dependency (Cameron and Trivedi). The term excess zeros has become synonymous with more zeros than can be explained by the Poisson or NB model. Zero-inflated models have been extensively used in other fields, such as econometrics, since Lambert's publication outlining the zero inflated Poisson (ZIP) model. In general, excess zeros may be categorized as sampling or structural zeros. For example, in response to the question How often did you drink alcohol during the last 30 days? there will be individuals who drink alcohol but chose not to drink during the last 30 days (sampling zeros) and individuals who never drink alcohol (structural zeros).

Zero-inflated and hurdle models are generally used to model excess zeros. Zero-inflated models are typically used if the data contains excess structural and sampling zeros, whereas hurdle models are generally used when there are only excess sampling zeros. Zero-inflated and hurdle models are often indistinguishable using goodness of fit statistics (Gray) but one model type may be more applicable based on study objectives. Cheung suggested that the zero-inflated model is useful if there are two underlying disease processes, one which puts the subject at risk and the other which influences the outcome in the at-risk population.

For brevity, my focus is on the ZIP model because it is often used when modeling excess zeros. The ZIP model simultaneously fits two regression models. The logistic model is often used for estimating excess zeros and the Poisson model is used to model non-excess

[*] Office of The Director, National Center on Birth Defects & Developmental Disabilities, Office of Noncommunicable Diseases, Injury & Environmental Health, Centers for Disease Control and Prevention, 4770 Buford Hwy. Mailstop E-87, Atlanta, Georgia 30341 Office: 404-639-3028, cvr7@cdc.gov

zeros and positive counts. There are two situations that generally arise when using ZIP models in epidemiology. The first situation is when we know that there are two processes that generate the zeros (structural and sampling zeros are both present). We are usually interested in 1) estimating the risk (or odds) of being an excess zero and factors that influence the risk of being an excess zero and 2) estimating the risk of having more events in light of not being an excess zero and the factors that influence the risk. However, it is important to note that the available data usually lacks information about the individual zeros; i.e., we lack information about which zeros are structural and which are sampling zeros. The second situation is to use the ZIP model to account for excess zeros even though all zeros arise as the result of only one process, which is usually a result of sampling. In this situation the focus is often on estimating the risk from the Poisson model conditional on not being an excess zero.

My purpose is to develop several count data scenarios with excess zeros and model using ZIP regression. I investigate the adequacy of the ZIP model to estimate the structural and sampling zeros and the resulting impact on inference.

2. Zero-Inflated Poisson (ZIP) Model

The ZIP model has the flexibility to use a variety of models to estimate the excess zeros. Here, I use the logistic model to estimate excess zeros and the ZIP model is defined by letting the independent response count variable be $Y = (Y_1, ..., Y_n)$ and has the following form:

$$f(y) = \begin{cases} p + (1-p)e^{-\mu} & y = 0 \\ (1-p)\dfrac{e^{-\mu}\mu^y}{y!} & y \geq 1 \end{cases} \tag{1}$$

Here I define p using a logit model as,

$$p = \frac{1}{1 + e^{-\eta}} \tag{2}$$

where μ and η are a function of predictor variables for the Poisson and logistic models, respectively.

3. Methods

I develop three scenarios to investigate the impact of excess zeros on the ZIP model. The first scenario assumes there is only one group and that zeros are a result of structural and sampling. The second scenario assumes there is only one process generating zeros, sampling, and that all subjects are at risk. For this scenario I will estimate the excess zeros, discuss interpretation of the two model components, and illustrate the pitfalls of using a ZIP model given only one process governs the zeros. The third scenario assumes two intervention groups and my purpose is to estimate the risk ratio (RR) for group A versus B. I investigate the impact on the RR as the proportion of excess zeros is allowed to vary among the groups. The three scenarios are as follows:

1. Scenario I

   (a) Data generated using a Poisson distribution with μ = 3.0

   (b) Incorporate excess zeros

2. Scenario II
   (a) Zeros result from sampling and structural
   (b) Data generated for Poisson portion using μ = 3.0
   (c) Excess zeros are added beyond what is expected by Poisson distribution
   (d) Allow the proportion of excess (structural) zeros as a proportion of all zeros to vary (0.25, 0.50, 0.75)
3. Scenario III
   (a) Two intervention groups (A, B)
   (b) Poisson distribution for each group with μ = 4.0 and 2.0 for groups A and B, respectively
   (c) Risk ratio equals 2.0 for group A versus B
   (d) Allow excess zeros in both groups to investigate impact on the RR

4. Results

The results for scenario I are presented in Figure 1. Figure 1A illustrates the data generated using a Poisson distribution with μ = 3.0. For scenario I, I included excess zeros and results for the ZIP model fit are presented in Figure 1B. As Figure 1B illustrates the trend of the counts > 0 is unchanged when allowing Figure 1A to have excess zeros. Hence, the parameter estimate and standard error are unchanged for the Poisson portion of the ZIP model in Figure 1B compared to Figure 1A as all excess zeros are estimated by the logistic portion of the ZIP model. Figure 1C illustrates that if I shift the Poisson distribution so that it accounts for some of the excess zeros there would be a lack of fit for the Poisson portion of the ZIP model.

Scenario II results are presented in Figure 2. This scenario assumes the Poisson μ = 3.0 and there are structural zeros. The proportion of structural zeros (orange), as a proportion of all zeros, is 0.25, 0.50, and 0.75 in Figures 2A, 2B, and 2C, respectively. The ZIP model fit for the Poisson portion is represented by the dashed line. Although the proportion of structural zeros varies the parameter estimates for the excess zeros (logistic) and Poisson portion of the ZIP model remains unchanged. Hence, there is no impact on the ZIP model parameter estimates by varying the proportion of structural zeros.

Scenario III focuses on estimating the RR for intervention group A versus B and results are presented in Figure 3. Figure 3A illustrates that the RR = 2.0 for group A (μ = 4.0) versus B ($\mu = 2.0$). Figure 3B illustrates that when I allow for excess zeros in groups A and B that the RR remains 2.0 even though there are excess zeros estimated by the logistic portion of the ZIP model. Figure 3C illustrates that the estimated RR = 2.0 when a high proportion of excess zeros is allowed in group A but few excess zeros in group B. However, if I calculate the figure 3B expected count in groups A and B and then estimate μ, I obtain 1.86, which would imply an underlying RR of 1.0 when including the excess zeros. It is important to note that the Poisson parameters and standard errors for groups A and B are identical for figures 3A, B, and C.

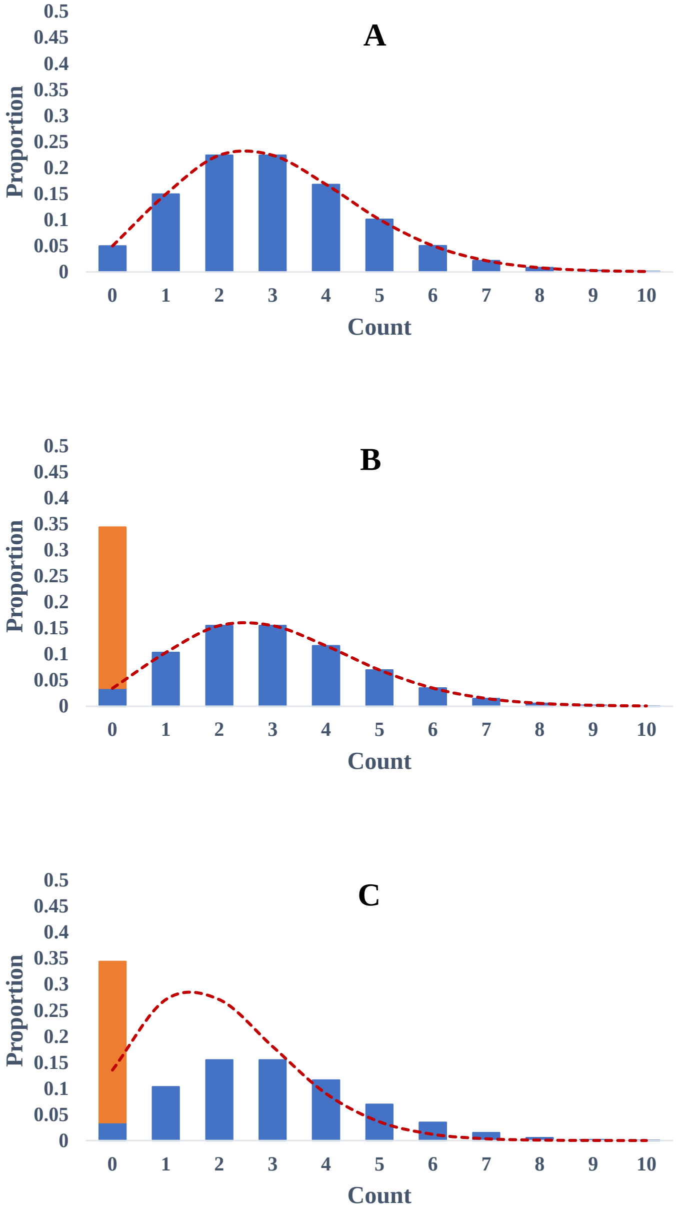

Figure 1: Scenario I assumes an underlying Poisson distribution, μ = 3.0, as illustrated in A. Figure 1B illustrates the inclusion of excess zeros (orange) and the estimated Poisson portion of the zero-inflated Poisson (ZIP) model. Note that the estimated parameter for figure 1B from the ZIP Poisson model remains unchanged from 1A. Figure 1C illustrates the lack of fit if I attempt to alter the Poisson fit of the ZIP model.

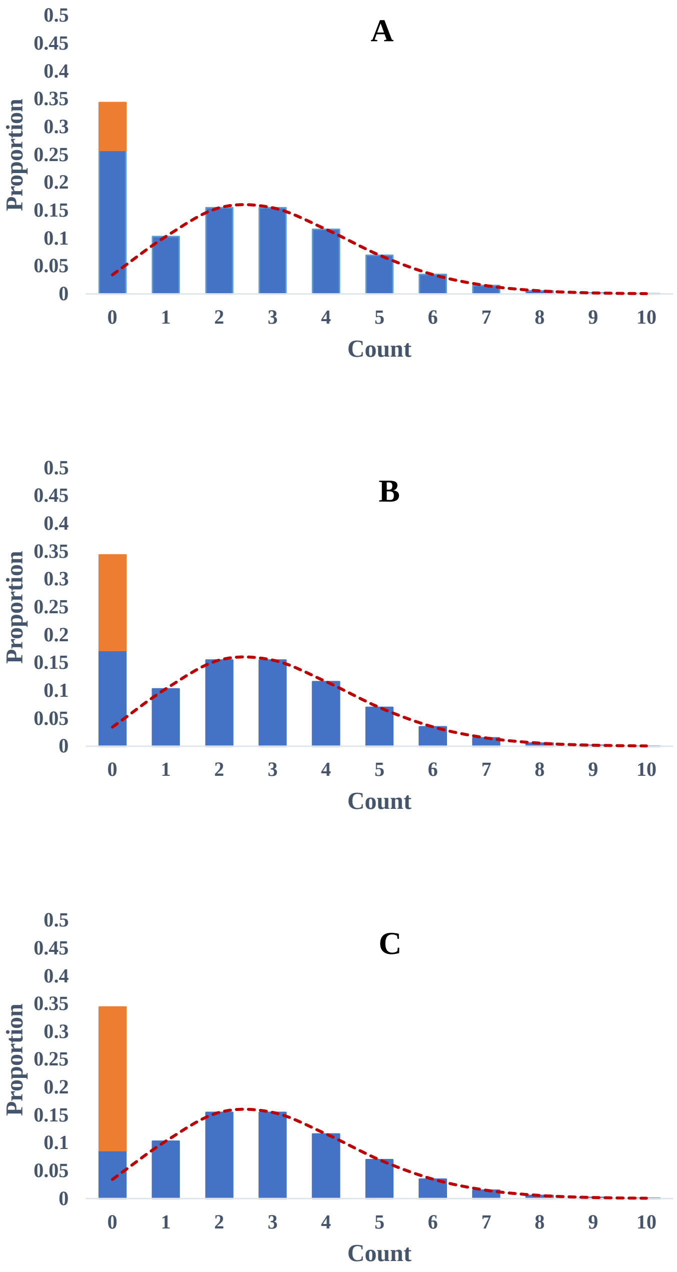

Figure 2: Results for scenario II, which assumes an underlying Poisson distribution of μ = 3.0 and allows proportion of structural zeros (orange) to vary as a proportion of the total zeros. The proportion of structural zeros is 0.25, 0.50, and 0.75 in figures 2A, B, and C, respectively. The estimated proportion of structural and sampling zeros remains unchanged when using the ZIP model.

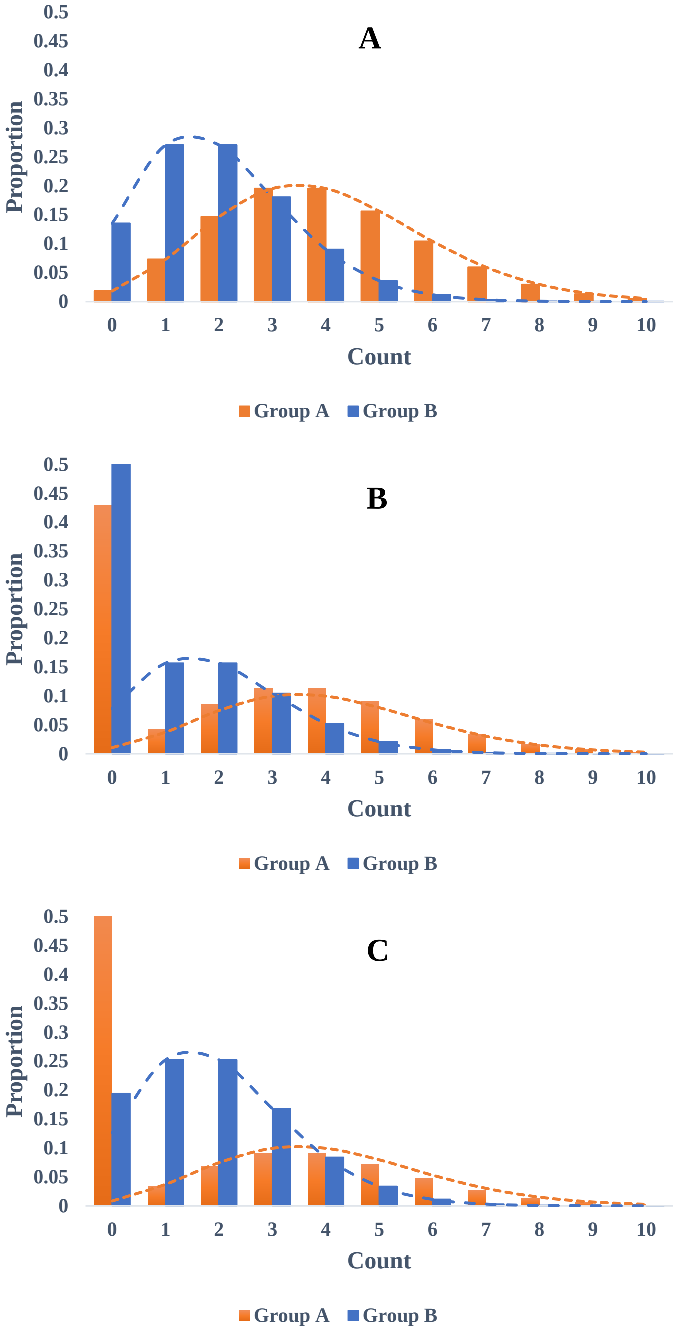

Figure 3: Results for scenario III, which assumes two intervention groups (A, B). Figure 3A illustrates the underlying Poisson distribution for groups A (μ = 4.0) and B (μ = 2.0) with a risk ratio (RR) of 2.0. In figures 3B and C, I allow for excess zeros in groups A and B. Figures 3B and C illustrate that the estimated RR remains unchanged from the RR of 2.0 in figure 3A, which has no excess zeros.

5. Discussion

The ZIP model is commonly used to model count data with excess zeros. Studies generally lack information on the individual zeros and as a result we lack information about which zeros are structural and which result from sampling. However, it is often known for a study if there is only one type of zeros. Studies that have only one type of zeros typically have sampling zeros. My scenario I illustrates that when excess zeros arise from one process that the parameter estimates for the Poisson portion of the ZIP model remains unchanged. This is true regardless of how many excess zeros occur in the data once there are more zeros than to be expected under the Poisson model, which is also true if excess zeros arise from both structural and sampling.

It is usually assumed that the ZIP model sufficiently allows us to estimate the proportion of structural and sampling zeros. I demonstrated that the ZIP model estimated proportion of structural zeros from the logistic model is not sufficient for estimating the proportion of structural zeros (scenario II). Furthermore, we are unlikely to know the proportion of structural zeros in the data and as a result we do not know if the structural zero estimate is reasonable. Hence, if the purpose of the study is to estimate the proportion of structural zeros and the non-structural zeros or counts and estimate risk factors associated with these processes, then the ZIP model should be used with caution.

The ZIP model is often used to estimate the RR for risk factors or intervention groups. I showed in scenario III when comparing two groups that the RR remains unchanged as the proportion of zeros is altered to include excess zeros above the expected Poisson distribution. In particular note that the underlying ZIP Poisson model parameters and standard errors remain unchanged and as a result the RR remains unchanged regardless of the inclusion of additional excess zeros. Hence, our inference may not accurately reflect the difference among groups when using a ZIP (or other zero-inflated) model and conclusions may be erroneous. Recent research has started to address some of these shortcoming (Long et al.) of the zero-inflated modeling framework, but additional research needs to be conducted to understand the impact when the zero-inflated model focus is on identifying risk factors or comparing groups.

REFERENCES

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

Cheung, Yi-Bin. (2002), “Zero-inflated models for regression analysis of count data: a study of growth and development,” *Stat Med.*, 21:14611469.

Chin, H. C., Quddus, M. A. (2003). “Modeling count data with excess zeroes - An empirical application to traffic accidents,” *Sociol. Meth. Res.* 32(1):90-116.

Lambert, D. (1992), “Zero-inflated Poisson regression with an application to defects in manufacturing,” *Technometrics*, 34:114.

Gray, B. R. (2005), “Selecting a distributional assumption for modelling relative densities of benthic macroinvertebrates,” *Ecological Modelling*, 185:1-12.

Long D. L., Preisser J. S., Herring A. H., Golin C. E. (2014), “A marginalized zero-inflated Poisson regression model with overall exposure effects,” *Stat Med.*, 33(29):5151-5165.