A modified Binomial likelihood model for zero/n-inflated count data

James Sweeney\textsuperscript{a}, John Haslett\textsuperscript{b}

\textsuperscript{a}School of Mathematical Sciences and Statistics, University College Dublin
\textsuperscript{b}Department of Statistics, Trinity College Dublin

Abstract

A statistical inconsistency of a zero-inflated Binomial likelihood model for count data is identified. This issue occurs when the response, \( y \), is equal to the sum constraint \( n \), and results in statistically inconsistent and erroneous parameter inferences being drawn from the data. The zero-modified likelihood is amended to address this issue of \( n \)-inflation, resulting in a fully symmetric Binomial likelihood model for both zero and \( n \)-inflated counts. An ecological regression problem is presented which details the superiority, through consistency of inference, of the new likelihood model.

Keywords: zero-inflation, \( n \)-inflation, Binomial distribution

1. Introduction

In many research areas of interest including ecology, epidemiology and engineering, the analysis of count data is of primary concern. However, a relatively frequent feature of such data sets is their tendency to contain many zero counts; see Haslett et al (2006) for an example in the context of an ecological regression problem, or Lambert (1992), who considers models for defects in manufacturing. Data sets which contain a higher proportion of zeroes over that expected by the standard statistical families can be modelled through the use of zero-modified distributions.

In a zero-modified distribution, as detailed in Hall (2000), the observed data are assumed to arise from one of two distinct states, a zero state from which only zero counts are observed, and an alternative state from which all of the non-zero counts and a few of the zero counts arise - the alternative state can be modelled through the use of standard statistical families such...
as the Binomial. However, the use of such zero-modified distributions, in the context of data which are subject to sum constraints, will result in erroneous and inconsistent parameter inferences. The source of this inconsistency is that existing zero-modified likelihoods for Binomial data do not account for possible zero-inflation in both variables which comprise the sum constraint.

In Section 2 we introduce the zero-modified Binomial distribution, as presented in Hall (2000), and detail a statistical inconsistency of this model. In Section 3 we present our solution to this inconsistency, namely a modified Binomial likelihood model which addresses both zero and n-inflation in observed counts. A sample application involving the new likelihood is presented in Section 4 and we conclude the article with a short summary.

2. A statistical inconsistency of the zero-modified Binomial likelihood model

Consider the set of Binomial count observations, $Y = \{y_1, \ldots, y_m\}$, which are subject to the sum constraints $N = \{n_1, \ldots, n_m\}$. If the set of $y_i$ exhibit an excess of zeroes over that expected by the standard Binomial distribution, Hall (2000) proposes the following zero-modified Binomial likelihood model:

$$y_i \sim \begin{cases} 
0 \text{ with probability } (1 - \pi_i) \\
\text{Binomial}(n_i, p_i) \text{ with probability } \pi_i
\end{cases}$$

Thus, with probability $\pi_i$, an observed zero count arises from a Binomial distribution with parameters $n_i$ and $p_i$. Alternatively, with probability $(1 - \pi_i)$, the count arises from a distribution with a point mass at zero. The likelihood for a count of zero is thus $Pr(y_i = 0) = (1 - \pi_i) + \pi_i(1 - p_i)^{n_i}$.

However, consider the setting where $z_i = n_i - y_i$ is designated as the response. The sum constraint for $z_i$ is $n_i$, with probability of success equal to $1 - p_i$. If $y_i = 0$ then $z_i = n_i$ and the corresponding likelihood is $Pr(z_i = n_i) = \pi_i(1 - p_i)^{n_i}$. Being explicit, if zero-inflation is present in the data then $Pr(y_i = 0) \neq Pr(z_i = n_i)$ and $Pr(y_i = n_i) \neq Pr(z_i = 0)$; parameter inferences obtained given the zero-modified likelihood model in Hall (2000) will thus be statistically inconsistent. Differing parameter inferences will be drawn from the data depending on whether $y_i$ or $z_i$ is assigned as the response; in Section 4 we present a worked example which details this result explicitly.

An important point to note is that if the $z_i$ are “zero-inflated” then the $y_i = n_i$ may suitably be described as “n-inflated” and vice versa. This terminology is adopted for the remainder of the article.
3. The zero/n-modified Binomial distribution

We propose an extension to the existing zero-inflated Binomial likelihood model which addresses possible n-inflation in the observed data, specifically,

\[
y_i \sim \begin{cases} 
0 & \text{with probability } \frac{(1-\pi_1)\pi_2}{\pi_1 + \pi_2 - \pi_1\pi_2} \\
n_i & \text{with probability } \frac{\pi_1(1-\pi_2)}{\pi_1 + \pi_2 - \pi_1\pi_2} \\
\text{Binomial}(n_i, p_i) & \text{with probability } \frac{\pi_1\pi_2}{\pi_1 + \pi_2 - \pi_1\pi_2}
\end{cases}
\] (2)

Thus, with probability \( \frac{\pi_1\pi_2}{\pi_1 + \pi_2 - \pi_1\pi_2} \), a count \( y_i \) arises from a Binomial distribution with parameters \( p_i \) and \( n_i \). With probability \( \frac{(1-\pi_1)\pi_2}{\pi_1 + \pi_2 - \pi_1\pi_2} \), the count \( y_i = 0 \) arises from a distribution with a point mass at zero. Alternatively, with probability \( \frac{\pi_1(1-\pi_2)}{\pi_1 + \pi_2 - \pi_1\pi_2} \), a count of \( n_i \) arises from a distribution with a point mass at \( n_i \). Here \((1-\pi_1)\) and \((1-\pi_2)\) refer to the probability of zero-inflation for \( y_i \) and \( z_i = n_i - y_i \) respectively. The probability correction factor \( \pi_1 + \pi_2 - \pi_1\pi_2 \) is due to the constraint that the \( n_i \) must be non-zero.

The mean and variance of the distribution are:

\[
E(y) = \frac{\pi_1(1-\pi_2)}{\pi_1 + \pi_2 - \pi_1\pi_2} n_i + \frac{\pi_1\pi_2}{\pi_1 + \pi_2 - \pi_1\pi_2} n_i p_i
\]

\[
Var(y) = \frac{\pi_1(1-\pi_2)n_i^2 + \pi_1\pi_2(n_i p_i)(1 - p_i + n_i p_i)}{\pi_1 + \pi_2 - \pi_1\pi_2} - E(y)^2
\]

It is simple to show that \( Pr(y_i = Y) = Pr(z_i = n_i - Y) \) for the model in (2) by plugging in the relevant probability values, i.e. the model is fully symmetric regardless of either \( y_i \) or \( z_i \) being designated as the response; parameter inferences drawn from the model will thus be consistent regardless of which variable we choose to model.

4. Application

We utilise an example from the ecological regression literature to detail the superiority, through consistency of inference, of the new zero/n-inflated Binomial likelihood model. The data set consists of pollen counts, obtained from samples of lake sediment, for each of 28 different plant taxa at 7671...
site locations (see Huntley et al (1993) for further details); a measure of
local climate, \(GDD_5\) (Growing degree days above 5\(^\circ\)C), is available for each
site. The primary interest lies in the construction of a model which relates
the pollen abundance at each location to the local measurement of \(GDD_5\).

We divide the plant taxa into two distinct categories, those preferring
either a warmer or cooler type climate. \textit{Abies, Larix, Picea & Salix} are ex-
amples of plant taxa which prefer cooler climates (i.e. lower \(GDD_5\)) whilst
those preferring warmer type climates include (amongst others) \textit{Alnus, Cory-
lus, Fagus & Olea}. Let \(Y = \{y_1, \ldots, y_{7671}\}\), represent the pollen counts of
the “cooler” type and \(Z = \{z_1, \ldots, z_{7671}\}\) the counts of the “warmer” type.
The sum constraint at each location is naturally the sum of the respective
pollen types, i.e. \(N = \{n_1, \ldots, n_{7671}\} = Y + Z\). Note that the \(n_i\) are variable
due to the differing number of pollen samples counted at each specific site.

In Figure 1(a) we present a plot of the “cooler” pollen proportions (i.e.
\(y_i/n_i\)) versus \(GDD_5\). The pollen of plants preferring cooler type climates
 dominate the pollen assemblage at sites with low \(GDD_5\) with this dominance
reversing as \(GDD_5\) increases. Both the \(Y\) and \(Z\) counts exhibit significant
signs of zero-inflation; in Figure 1(b) we detail this visually through a his-
togram of the cooler pollen proportions for \(GDD_5\)’s greater than 7000. The
\(y_i\)’s appear to be both zero and n-inflated as evidenced by the excess of
observed proportions at 0 and 1.

As the primary focus of this article in on the likelihood, we assume a
simple model for the cooler pollen-\(GDD_5\) interaction, namely that the pro-
portion of pollen (\(y_i/n_i\)) observed at a given site is a logistic-linear function
of the \(GDD_5\) measurement (\(c_i\) ) at that location, logit(\(p_i\)) = \(\beta_0 + \beta_i c_i = \beta C_i\).

We also model the zero and n-inflation probabilities as a function of the
underlying response, scaled by the parameters \(\alpha_1\) and \(\alpha_2\); a justification
of this modelling choice is presented in Salter-Townshend and Haslett (2012),
This modelling approach is successful in substantially reducing the number
of parameters that must be inferred.

In the following let \(\alpha = \{\alpha_1, \alpha_2\}\)

\[
\begin{align*}
y_i & \sim \text{zero/n-inflated \text{Binomial}}(n_i, p_i, \alpha) \\
\text{logit}(p_i) & = \beta C_i \\
\pi_{1i} & = (p_i)^{\alpha_1} \\
\pi_{2i} & = (1 - p_i)^{\alpha_2}
\end{align*}
\]

(3)

The log-likelihood for the model in (3) is, up to a constant,
Figure 1: (a) Scatterplot of the “cooler” pollen counts where the local GDD5 measurement is plotted against Y/N. (b) Histogram of the “cooler” pollen counts (scaled by sum totals) with an associated GDD5 greater than 7000. We note that the counts exhibit clear signs of both zero and n-inflation as evidenced by the excess of observations at 0 and 1.

This log-likelihood is unaffected by the choice of either the cooler or warmer pollen counts as the response, i.e. $L(\alpha, \beta; Y, N) = L(\alpha, \beta; Z, N)$; as a result inferences on model parameters will be consistent. Conversely, for a zero-inflated model, the log-likelihoods formulated in terms of either the cooler or warmer pollen types will not be equivalent. Note that the
zero-inflated model log-likelihoods can be obtained from the zero/n-inflated versions by setting the relevant n-inflation parameter equal to zero.

As the log-likelihood in (4) is simple to maximise due to the small number of model parameters ($\alpha_1, \alpha_2, \beta_0, \beta_1$), we proceed to do so via the optim function in the R statistical package [5], which is based on the Newton Raphson method. The resulting maximum likelihood estimates are presented in Table 1 as well as the estimates produced for the zero-modified model of Hall (2000) where both $Y$ and $Z$ are separately modelled as the response. Standard errors are also presented for each variable; these are obtained from the hessian matrix.

| Model | zero/n inflated | zero-inflated (cooler) | zero-inflated (warmer) |
|-------|----------------|------------------------|------------------------|
| $\log(\alpha_1)$ | $-2.236 \pm 0.039$ | $-2.239 \pm 0.039$ | - |
| $\log(\alpha_2)$ | $-2.937 \pm 0.057$ | - | $-2.973 \pm 0.057$ |
| $\beta_0$ | $1.958 \pm 0.0025$ | $2.015 \pm 0.0024$ | $-2.051 \pm 0.0024$ |
| $\beta_1$ | $-0.00117 \pm 0.000002$ | $-0.00119 \pm 0.000002$ | $0.00127 \pm 0.000002$ |
| AIC | 476513.9 | 513868.5 | 519762.1 |

Table 1: Maximum likelihood estimates for model parameters ± standard errors obtained from the hessian matrix (the response choice is placed in brackets).

In Figure 1(a) we plot the predicted proportions ($P$) of cooler type pollen produced by the zero/n-inflated Binomial likelihood model. Note that the same predictions are returned regardless of the choice of response variable; inferences produced by this model are statistically consistent. Conversely, in Figure 1(b) we plot the predicted proportions of the zero-inflated Binomial model with the response set as either $Y$ or $Z$. It is immediately apparent that the inferences obtained by the zero-inflated model are statistically inconsistent; when the warmer pollen counts ($Z$) are chosen as the response, the predicted proportions are close to those of the zero/n-inflated model, however, the predicted proportions when the cooler pollen proportions are modelled significantly differ due to the large number of observations equal to $n_i$ for high $GDD5$. These $n_i$’s are clearly as a result of zero-inflation in the $Z$ counts but this feature is not captured by the zero-inflated model, resulting in erroneous inferences being drawn.
Figure 2: Comparison of the predicted proportions ($P$) of cooler type pollen produced by the (a) zero/n-inflated Binomial model and the (b) zero-inflated Binomial model of Hall (2000). The use of the zero-inflated likelihood results in statistically inconsistent estimates of $P$ which depend on the response variable modelled.

5. Summary

In studies where the collected data consist of zero-inflated count observations subject to a sum constraint, the choice of response variable to model will have a dramatic effect on the conclusions drawn; as existing likelihood models do not account for possible zero-inflation in both variables comprising the sum constraint, inferences drawn from the data will be statistically inconsistent and erroneous.

The specific purpose of this article is to address this issue; we have presented a modified Binomial likelihood model which addresses zero-inflation of both response variables simultaneously, resulting in consistent parameter inferences regardless of response variable choice. The superiority, in terms of consistency of inference, is clearly displayed in the ecological regression example presented herein.

6. Acknowledgement

Support of Science Foundation Ireland [11/PI/1027] is gratefully acknowledged.
References

[1] Hall, D. B. (2000), ‘Zero-inflated Poisson and Binomial Regression with Random Effects: A Case Study’, Biometrics 56, 1030-1039.

[2] Haslett, J., Whiley, M., Bhattacharya, S., Salter-Townshend, M., Wilson, S. P., Allen, J. R. M., Huntley, B. & Mitchell, F. J. G. (2006), ‘Bayesian palaeoclimate reconstruction’, Journal of the Royal Statistical Society: Series A 169(3), 1-36.

[3] Huntley, B. (1993), ‘The Use of Climate Response Surfaces to Reconstruct Paleoclimate from Quartenary Pollen and Plant Macrofossil Data’, Philosophical Transactions of the Royal Statistical Society 341, 215-224.

[4] Lambert, D. (1992), ‘Zero-inflated Poisson Regression, With an Application to Defects in Manufacturing’, Technometrics 34(1), 1-14.

[5] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

[6] Salter-Townshend, M, Haslett, J (2012) ‘Fast Inversion of a Flexible Regression Model for Multivariate, Zero-Inflated Pollen Counts’. Environmetrics 23(7), 595-605.