Zero Augmentation: A method for fitting zero-modified count models that allows both zero-inflation and zero-deflation

Paul Wilson[1]

[1] School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway.

Paul.Wilson@nuigalway.ie

Abstract: The concept of zero-inflation is now well established, and several software packages exist that fit zero-inflated models. Whilst there is some mention of zero-*deflation* in the literature, there is very little published evidence of research into zero-*modified* models that allow for both zero-inflation and zero-deflation within the one model. We present a very simple method that enables the fitting of zero-modified models via *any* software that fits zero-inflated models, and investigate the benefits of fitting zero-modified models.

Keywords: Zero-modification, Zero-inflation, Zero-deflation

1 Introduction

Whilst the concept of zero-inflation is now well established, and several software packages exist (e.g. the R-packages PSCL, ZIGP, VGAM and GAMLSS) that fit zero-inflated models, i.e. models of the form:

$$f(y;\Theta) = \begin{cases} \gamma + (1-\gamma)f(0;\Theta) & y = 0 \\ \\ (1-\gamma)f(y;\Theta) & y = 1,2,3,\ldots \end{cases} \tag{1}$$

where $\gamma > 0$, there is a near absence of work concerning zero–*modified* models that allow for both zero-inflation and zero-deflation within the same dataset, and thus permit negative (as well as positive) values of γ. Zero-deflation may arise as a consequence of under-reporting of zero counts. For example, say a study was concerned with the distribution of the number of eggs laid in bird's nest dependent on various covariates. Certain species of birds, for instance the Whooping Crane (Grus Americana), make nests by making shallow depressions in marshy ground. Clearly it may be difficult to distinguish a natural shallow depression in the ground from one made by a whooping crane as a nest, but in which no eggs were laid. Over-classification of empty nests as natural depressions will lead to under-reporting of the

number of zero counts of the number of eggs laid in such nests, and hence zero-deflation, whereas over-classification will lead to zero-inflation.
Even if theoretically zero-deflation does not make sense in the context of the data being analysed, zero deflation may occur in the observed data for certain combinations of covariates. If a model that is constrained to return positive values of γ is fitted to observed data that is zero-deflated for certain combinations of covariate values, a model fitting mechanism that employs an EM algorithm to estimate the proportion of the data that arises from a perfect-zero distribution will return a value of zero, or possibly a small positive value. This will both influence the estimates of positive values of γ for other covariate combinations, and, perhaps more importantly will affect the estimation of the mean: to compensate for the increase to zero of a negative γ estimate a greater value of the estimate of the mean will occur.
Whilst a Bayesian approach to zero-modified models has been proposed by Angers and Biswas (2003), the only frequentist approach would appear to be that of Dietz and Böhning (2000). Their method incorporates the unusual link-function:

$$\eta = \eta(\gamma) = \log\left(\frac{1-\gamma}{e^{\mu}/(e^{u}-1)-(1-\gamma)}\right) \tag{2}$$

where $\gamma = 1 - \frac{e^{X\beta}}{1+e^{X\beta}}\frac{e^{\mu}}{e^{\mu}-1}$
for the zero-modification parameter. This technique has the undesirable property that the link function varies according to the Poisson mean, and is not readily implementable using standard software packages for fitting zero-inflated models. The method of zero-augmentation introduced here enables *any* zero-modified models to be fitted using *any* software, and furthermore, *any* link functions available for fitting such models may be utilised.

2 Zero Augmentation

We define *Zero-Augmentation* of a dataset to be the artificial addition of zeros to the response variable of that dataset. The proposed system of zero-augmentation is to replicate the data $\kappa - 1$ times to form a "κ-augmented dataset", whilst the values of the various covariates remain the same in the replicated data, the values of the response variable are replaced by zeros. For instance if the original data (where the response variable is in the final column) is:

1	0	2	1	0
2	1	3	1	5
3	1	1	2	3

then the 2–augmented data is:

1	0	2	1	0
2	1	3	1	5
3	1	1	2	3
1	0	2	1	0
2	1	3	1	0
3	1	1	2	0

The idea behind zero-augmentation is extremely simple: the zeros added to the response variable by zero-augmentation are by definition extra-zeros. If the model fitting mechanism were to classify all such augmented zeros as extra-zeros, then zero-inflation parameter estimates for the original, non-augmented data may be deduced from those of the augmented data, (for example, for 2-augmented data a parameter value of $\gamma^\star$ for the zero-modification parameter may be shown to correspond to an estimate of

$$\hat{\gamma} = 2\gamma^\star - 1 \tag{3}$$

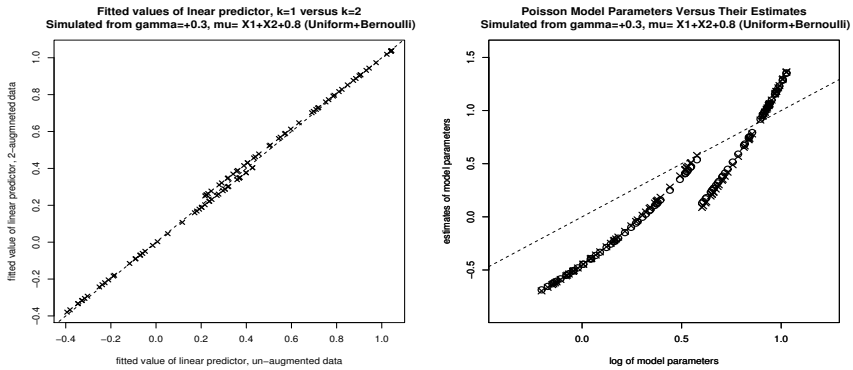
after 2-augmentation). Hence a positive γ parameter in the data of, say, 0.3 will return an estimate of approximately 0.65 when the model is fitted to the 2-augmented data which may then be "converted back" to an estimate of approximately $(2 \times 0.65 - 1) = 0.3$ for the original data. However if fitting the model to the 2-augmented data returns an estimate of $\gamma^\star = 0.2$, this corresponds to an estimate of $\hat{\gamma} = (2 \times 0.2) - 1 = 0.6$ for the original data, indicating zero-deflation. Other model parameter estimates should remain unchanged by zero-augmentation, the structure of the non–extra zeros count data has not been altered. In practice the model fitting procedure may not classify all augmented zeros as extra zeros, but any resulting bias is not serious. In the vast majority of cases, 2-augmentation is sufficient to detect and model any zero-deflation that exists in the data, for the remainder of this paper we present examples of the technique restricting ourselves to non-augmentation and 2-augmentation.

2.1 Example 1

To illustrate that the estimates of the linear predictors of the mean are extremely similar under non-augmentation and 2-augmentation, 100 data were simulated from a zero-inflated Poisson model where $\gamma \sim 0.3$ and $\mu \sim X_1 + X_2 + 0.8$, where $X_1 \sim U(0,1)$ and $X_2 \sim bernoulli(0,1)$, and zero-inflated models fitted (using log and cloglog links). The left hand diagram of Figure 1 plots the values of the estimated values of the linear predictors of the Poisson means against each other, whilst the right-hand diagram plots the values of the estimates under non-augmentation ($\times$) and 2-augmentation ($\circ$) against the parameters of the model from which the data was simulated. We see that these estimates are very similar. The

values of $\hat{\gamma}$ obtained from fitting the model to the original data and the 2-augmented data (after applying formula of equation (3) was 0.29 in both cases. The log-likelihood of the models fitted to the data calculated using the parameter estimates from the non-augmented and augmented data respectively were almost identical at -134.63 and -134.59 respectively.

FIGURE 1. Comparative fits of the Poisson parameters of zero-modified Poisson data under non-augmentation and 2-augmentation



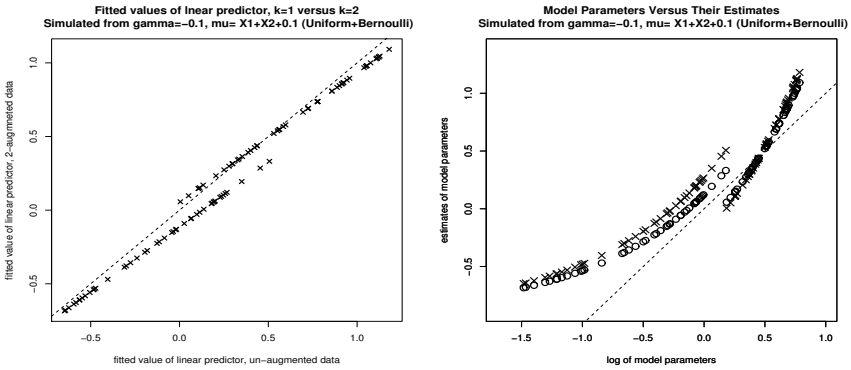
2.2 Example 2

The diagrams of Figure 2 pertain to 100 data simulated from zero-deflated Poisson data where $\gamma \sim -0.1$ and $\mu \sim X_1 + X_2 + 0.1$, where $X_1 \sim U(0,1)$ and $X_2 \sim bernoulli(0,1)$. We see from the left hand diagram that here the estimates are not similar, indicating zero-deflation. When the zero-inflated model was fitted to the non-augmented data a value of zero was returned for $\hat{\gamma}$, whereas, after using the formula of equation (3) an estimate of $\hat{\gamma} = -0.08$ is returned when the model is fitted using the 2-augmented data. Note that, as a consequence of this more accurate estimation of γ, in general the estimates of the values of the Poisson means ($\circ$) are closer to the values of the means of the model from which the data was simulated. Here the log-likelihood of the data calculated using the parameters estimated from the augmented data was -133.67, whereas that obtained from the non-augmented data was slightly poorer at -134.40

2.3 Example 3

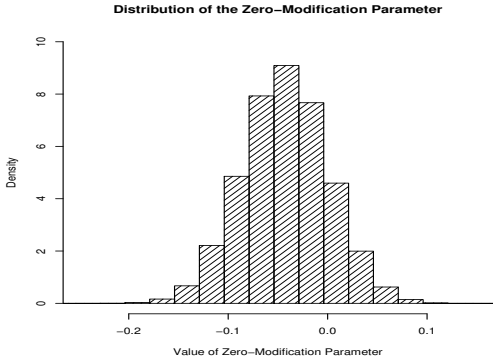
For this example, 300 data were simulated from the zero-modified Poisson model $\gamma \sim \frac{\mu-3}{10}$ where $\mu \sim 0.3X_1 + 0.2X_2 + 0.5X_3 + 0.3X_4 + 0.5$, and $X_1 \sim N(0,1)$, $X_2 \sim U(0.5,2)$, $X_3 \sim N(3,0.5)$ and $X_4 \sim U(0.2,2)$. Figure

FIGURE 2. Comparative fits of the Poisson parameters of zero-modified Poisson data under non-augmentation and 2-augmentation



3 illustrates the distribution of the resultant zero-modification parameters. As is apparent both zero-inflation and deflation are present in the model, the values of the zero inflation parameter being approximately normally distributed with mean -0.042 and standard deviation 0.042.

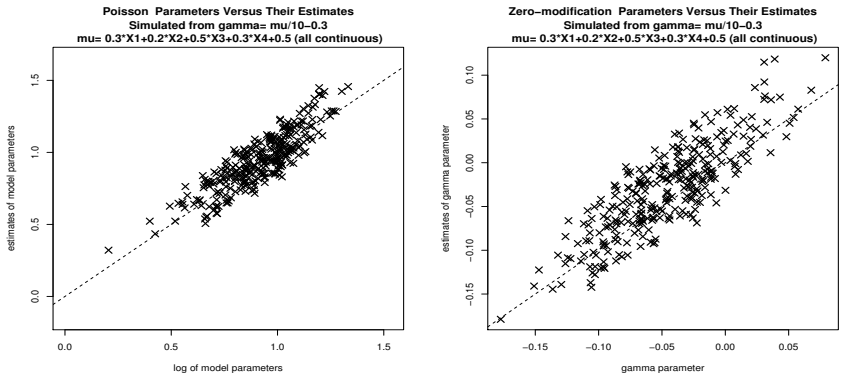
FIGURE 3. Distribution of the Zero-Modification Parameter, Example 3



When it was attempted to fit a zero-inflated model to the non-augmented data, (using logit and log links), the fitting algorithm failed due to the considerable zero-deflation. The fitting of the 2-augmented data was successful however. The left hand diagram of Figure 4 plots the estimates of the linear predictors of the Poisson means against the parameters of the model from which the data was simulated. We see that this diagram indicates that the fitted means correspond to what is expected from data simulated

from such a model. The right hand diagram plots the estimates of the fitted values of the of the zero-inflation parameters (adjusted using the formula of equation (3)) against the parameters of the model from which the data was simulated. Again, these indicate a good degree of fit.

FIGURE 4. Fits of the parameters of zero-modified Poisson data under 2-augmentation



3 Conclusion

Zero-augmentation is a very simple technique that allows data that contains both zero-inflation and zero-deflation to be modelled using standard software. Whilst the discussion above concentrates on the estimation of the means of zero-modified Poisson data, the technique may be utilised to estimate any parameter of any zero-modified model. The development of this technique is in the early stages, future research areas include quantification of the amount if bias introduced by the augmentation procedure, the refinement of the technique to compensate for such bias and the development of associated tests to determine whether zero-deflation may be present in the data being analysed.

References

Angers JF, Biswas A (2003) A Bayesian analysis of zero-inflated generalized Poisson model. *Computational Statistics and data Analysis* Vol. 42, pp. 37–46

Dietz D, Böhning D (2000) On the estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics and data Analysis* Vol. 34, pp. 441–459