International Statistical Review

International Statistical Review (2021) doi: 10.1111/insr.12479

Modelling Excess Zeros in Count Data: A New Perspective on Modelling Approaches

John Haslett[1], Andrew C. Parnell[2], John Hinde[3] and Rafael de Andrade Moral[4]

[1]*School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland*
[2]*Hamilton Institute, Insight Centre for Data Analytics, Maynooth University, Maynooth, Ireland*
[3]*School of Mathematics, Statistics and Applied Mathematics, NUI Galway, Galway, Ireland*
[4]*Department of Mathematics and Statistics, Maynooth University, Maynooth, Ireland*
Correspondence *Rafael de Andrade Moral, Department of Mathematics and Statistics, Maynooth University, Maynooth, Ireland. Email: rafael.deandrademoral@mu.ie*

Summary

We consider the analysis of count data in which the observed frequency of zero counts is unusually large, typically with respect to the Poisson distribution. We focus on two alternative modelling approaches: over-dispersion (OD) models and zero-inflation (ZI) models, both of which can be seen as generalisations of the Poisson distribution; we refer to these as implicit and explicit ZI models, respectively. Although sometimes seen as competing approaches, they can be complementary; OD is a consequence of ZI modelling, and ZI is a by-product of OD modelling. The central objective in such analyses is often concerned with inference on the effect of covariates on the mean, in light of the apparent excess of zeros in the counts. Typically, the modelling of the excess zeros *per se* is a secondary objective, and there are choices to be made between, and within, the OD and ZI approaches. The contribution of this paper is primarily conceptual. We contrast, descriptively, the impact on zeros of the two approaches. We further offer a novel descriptive characterisation of alternative ZI models, including the classic hurdle and mixture models, by providing a unifying theoretical framework for their comparison. This in turn leads to a novel and technically simpler ZI model. We develop the underlying theory for univariate counts and touch on its implication for multivariate count data.

Key words: hurdle; over-dispersion; zero-altered; zero-deflation; zero-inflation.

1 Introduction

Regression analysis of count data arises in many fields, including agriculture (Blasco-Moreno *et al*., 2019), ecology (Mcmahon *et al*., 2017), climatology (Salter-Townshend & Haslett 2012), finance (Benson & Friel, 2017) and pharma (Min & Agresti, 2005). The simplest modelling framework for such analysis is generalised linear modelling using the Poisson and negative binomial (NB) families. In regression, the mean parameter is related, via a link function, to a linear combination of covariates. In the NB context, parameters may also be related to the dispersion parameter.

It is common in such data to encounter an apparent excess of zeros, often with respect to the Poisson and even with respect to the NB. A generic challenge is thus regression of count data in

the presence of excess zeros. The mechanism by which these zeros arise may not always be a key focus of the analysis, yet the choice of model for this mechanism may have repercussions on other parameters that are of concern. Our focus is on the alternative models. In common with much of the literature, we use the term zero-inflated (ZI) rather broadly to refer to several alternatives, although some models necessarily include zero deflation.

The seminal papers are those of Mullahy (1986), who introduced the 'hurdle' models, and of Lambert (1992), who proposed a mixture model for the zeros. Both can be seen as one-parameter extensions of a simpler base distribution, in which the Poisson (or NB) probability of zero is 'altered'. Over the following three decades, a very large literature has developed (as of 25 August 2021, searching the terms 'zero-inflated' or 'hurdle' in the title of papers in Google Scholar yielded about 6 700 results, and more than 42 800 papers included the term 'zero-inflated' somewhere in the text). The issue of dealing with an excess of zeros has become deeply embedded in the methodology of count data regression.

A complicated nomenclature has developed. Very many authors regard the term ZI to be coterminous with the specific model proposed by Lambert; but other terms used are zero modified (Dietz & Böhning, 2000), zero altered (Rigby *et al*., 2005; Yee 2015), two-part (Pohlmeier & Ulrich 1995) and conditional (Welsh *et al*., 1996). Various re-parameterisations exist, such as marginal ZI models in which the ZI and Poisson components are merged into a single mean parameter (Long *et al*., 2014; Martin & Hall, 2017). Generically, we refer to these as explicit ZI models. Surprisingly, however, there have been relatively few review papers on the topic in the statistical literature. Examples include Ridout *et al*. (1998), Warton (2005), Deng & Paul (2005), Hilbe & Greene (2007), Perumean-Chaney *et al*. (2013) and Farewell *et al*. (2017).

A second approach is via distributions that are over-dispersed (OD) with respect to the Poisson, the best known example of which is the NB. This also is a one-parameter extension of the Poisson. This approach we describe as implicit ZI, for the ZI is a by-product of a wider focus on the mean/variance relationship; in the Poisson distribution, the mean and variance are equal. They also can be said to 'alter' the probability of a zero from its Poisson 'base'. For brevity, below, we sometimes refer to the implicit and explicit approaches as OD and ZI, respectively.

These two approaches may be combined; zero-inflating the NB is then a two-parameter extension of the Poisson. We do not pursue in any detail the many other OD distributions that have been proposed in recent years, nor their ZI variants, but we remark that this is a rapidly growing literature.

But others see them as competitors. In particular, Warton (2005) provided comparisons between different implicit and explicit ZI models fitted to a total of 1 672 abundance variables across 20 multivariate datasets. He argued there is little or no evidence for the need to explicitly model excess zeros; the wide class of OD models is sufficient. However, while his views have not been explicitly rebutted, there are some examples in the literature where explicit ZI is preferred to over-dispersion (see, e.g. Welsh *et al*., 1996; Hall 2000) and others where ZI and over-dispersion are used together, typically in a zero-inflated negative binomial (ZINB) model (Jansakul & Hinde, 2008).

This paper examines both approaches with a view to gaining new insights on their different properties. We show that the link between ZI and OD models is complex; both approaches induce specific versions of each other. These insights are primarily theoretical. In particular, they suggest reasons why it will often in practice be very difficult to distinguish, from data, which of the alternative models will be 'best' in any useful sense. These insights lead to a new form of ZI which we term 'logistic ZI'.

Some argue that the essential difference between these approaches is that there are situations where at least some of the zeros (and *only* the zeros) arise from a process that differs essentially

from that which generates the counts, including counts of zero. This *prescribes* an explicit ZI approach, in which some of the observed zeros are deemed to reflect an unreported structural variable; they may even be deemed 'false'. But, although not strictly necessary, this concept of true and false zeros may impose a burden on some users, for an OD distribution can often achieve the same descriptive effect; see Blasco-Moreno *et al*. (2019) and Martin *et al*. (2005). By contrast, OD distributions can lead to excess low counts *such as* zero. Zeros are no longer special: *low* counts may be inflated with respect to the Poisson. The key distinction is the upper tail of OD distributions is also inflated with respect to the Poisson.

Our fundamental concern in this review is with concepts, rather than with, say, algorithms, power and tests. We defer consideration of vector counts, which introduce new challenges, but not new concepts. We motivate the paper in Section 2, by referring to a concrete example. We then introduce our general approach including our new type of ZI, in Section 3. Section 4 examines, from the same perspective, the implicit ZI induced by the NB. Our concluding thoughts are in Section 5.

2 Motivating Example

In this section, we use a simple dataset to illustrate a brief and conventional overview of some models that deal explicitly or implicitly with excess zeros in the context of regression. Specifically, we adopt the classic nomenclature and motivation of the models but defer a discussion of theoretical aspects until Sections 3 and 4.

We consider the Trajan apple data of Ridout *et al*. (1998) (Table 1). The response variable is the number of roots produced by 270 micropropagated shoots of the apple cultivar Trajan under two different photoperiods (8 and 16 h) and four different concentrations of the growth hormone cytokinin BAP in a completely randomised design with multiple replication at each of the settings. Here, we treat both explanatory variables as factors and restrict attention to the full interaction model for the eight different photoperiod by hormone settings, so in essence fitting the frequency distribution of counts for the replicates at each setting. We refer to these eight

Table 1. *Frequency distributions of the number of roots produced by 270 shoots of the apple cultivar Trajan, under different experimental conditions (four BAP levels and two photoperiods).*

	Photoperiod							
	8				*16*			
BAP (μM)	*2.2*	*4.4*	*8.8*	*17.6*	*2.2*	*4.4*	*8.8*	*17.6*
No. of roots								
0	0	0	0	2	15	16	12	19
1	3	0	0	0	0	2	3	2
2	2	3	1	0	2	1	2	2
3	3	0	2	2	2	1	1	4
4	6	1	4	2	1	2	2	3
5	3	0	4	5	2	1	2	1
6	2	3	4	5	1	2	3	4
7	2	7	4	4	0	0	1	3
8	3	3	7	8	1	1	0	0
9	1	5	5	3	3	0	2	2
10	2	3	4	4	1	3	0	0
11	1	4	1	4	1	0	1	0
12	0	0	2	0	1	1	1	0
>12	13,17	13	14,14	14				

The table shows the number of shoots that produced 0, 1, …, 12 roots, and counts that exceeded 12 roots are shown individually in the last row. This table was adapted from Ridout *et al*. (1998)

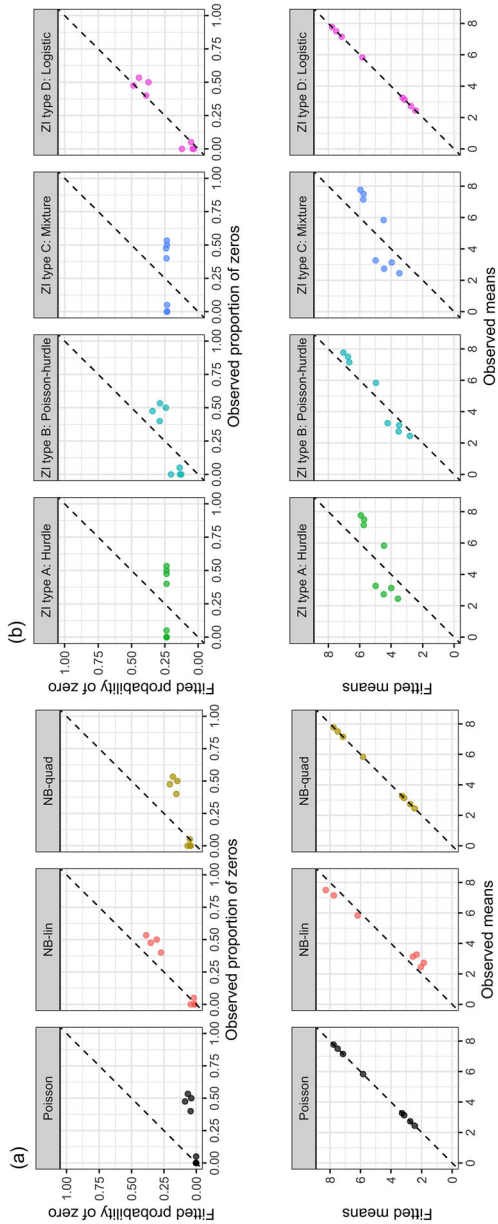

Figure 1. *Fitted versus observed proportion of zeros and means of each combination between photoperiod and hormone concentration in the Trajan data for (a) the Poisson, NB-lin and NB-quad models, and (b) ZI models types A, B, C and D*

distinct settings as $C^{(k)}$, $k \in \{1, \ldots, 8\}$ and index any associated model parameters and data summaries in the same way; writing $\overline{y}^{(k)}$ and $p_0^{(k)}$, $k \in \{1, \ldots, 8\}$, for the cell means and zero proportions. Our attention here focusses on the fit of the various models in terms of both the overall means and the zero proportions for each of the eight combinations by comparing the model fitted values with the sample values $\{\overline{y}^{(k)}, p_0^{(k)}\}$; these are displayed in Figure 1. Each of the seven different models shown in Figure 1 are explained below, and we encourage the reader to read this section while repeatedly returning to Figure 1 for reference.

The datatset is highly structured, permitting the fitting of saturated models for the mean structure. It also has, perhaps somewhat atypically, extensive replication allowing the calculation of natural and non-parametric estimates of, not just, the means but also of the probability of zeros. This allows us to see clearly that a basic Poisson regression model deals inadequately with the incidence of zero counts. It also allows us to evaluate the different explicit approaches to the handling of the excess zeros in terms of both the fitting of the probabilities of zeros and the fitting of the means. Both of these are challenging in more general regression settings. But we argue in this paper that the opaqueness, most especially of explicit ZI modelling, is a contributory factor. In the same way, we can study implicit ZI models, and here, we focus on two versions of the NB regression model as examples of OD count models.

We start with the basic Poisson model with $\pi_y(\lambda) = \lambda^y e^{-\lambda}/y!$, $y \in \mathbb{N} = \{0, 1, 2, \ldots\}$. Fitting a Poisson log-linear full interaction model corresponds to estimating a separate parameter $\lambda^{(k)}$ for each of the settings $C^{(k)}$ with the cell means $\overline{y}^{(k)}$ being the maximum likelihood estimators. Effectively, the data are treated as independent and identically distributed (iid), within each cell. So here the observed cell means are reproduced by the fitted model [Figure 1(a), bottom left panel], which can be viewed as a consequence of the simple exponential family form of the Poisson. For the probabilities of zero, the maximum likelihood estimators are $\hat{\pi}_0^{(k)} = \exp(-\hat{\lambda}^{(k)}) = \exp(-\overline{y}^{(k)})$ are not unbiased, unlike the cell zero proportions $p_0^{(k)}$ that are non-parametric and unbiased estimators of $\pi_0^{(k)}$. The comparison of these in Figure 1(a) shows the *a priori* evidence of excess zeros compared with a Poisson model, especially for the 16-h photoperiod where there are many observed zeros along with some large non-zero counts. This motivates the need to explore ZI extensions of the basic Poisson model.

2.1 Explicit Zero-inflated Extensions

Here, we revisit the two classical approaches: the hurdle model of Mullahy, in two variants, and the mixture model of Lambert. We also introduce a new approach, which we term 'logistic ZI'. Results from all four are presented in Figure 1(b) where comparisons are made with the aid of the non-parametric estimates $\{\overline{y}^{(k)}, p_0^{(k)}\}$.

We denote these explicit ZI extensions by $\tilde{\pi}_y(\lambda, \gamma)$, where here we focus on one-parameter extensions of the underlying Poisson base model $\pi_y(\lambda)$. In the brief overview below of the classical treatment of hurdle and mixture models, γ implicitly parameterises terms specific to these models. The key step is that $\tilde{\pi}_y$ arises from a specific alteration of the Poisson π_0 to $\tilde{\pi}_0$; and that, for $y > 0$, $\tilde{\pi}_y = \rho\pi_y$ where ρ re-normalises to ensure that $\sum_y \tilde{\pi}_y = 1$. So for $y > 0$, the modified distribution retains the same structure as the Poisson for successive terms, that is, $\tilde{\pi}_{y+1}/\tilde{\pi}_y = \pi_{y+1}/\pi_y$.

The first class of ZI extensions that we consider are the hurdle models of Mullahy (1986). The basic idea of the hurdle model is that the zeros and non-zeros arise from separate processes, allowing the zeros to be modelled as binary outcomes and the non-zero counts as outcomes

from a zero-truncated count distribution. The hurdle name comes from the idea that the binary process for the zero/non-zero values essentially models the probability of crossing over the hurdle from zero to non-zero counts. Then conditional on having crossed over this hurdle, the model is a standard count data model for the non-zero counts, typically taken to be a zero-truncated count model. The simplest hurdle model is one in which the zeros are modelled as coming from a Bernoulli distribution with a constant probability $\tilde{\pi}_0 = \tilde{\pi}_0(\gamma)$ and the non-zero counts coming from a zero-truncated Poisson distribution. The zero-truncated Poisson distribution has mean $E[Y|Y > 0] = \lambda/(1 - e^{-\lambda})$. In our notation, this hurdle model has the form

$$\tilde{\pi}_y(\lambda, \gamma) = \begin{cases} \tilde{\pi}_0 & \text{if } y = 0 \\ \dfrac{(1 - \tilde{\pi}_0)}{(1 - e^{-\lambda})}\pi_y(\lambda) & \text{if } y > 0 \end{cases}$$

with the re-normalising $\rho = (1 - \tilde{\pi}_0)/(1 - e^{-\lambda})$ and $E[Y] = (1 - \tilde{\pi}_0)/(1 - e^{-\lambda})\lambda = \rho\lambda$.

Fitting this simple model to the Trajan data with a single additional ZI parameter, the common $\tilde{\pi}_0$, with separate parameters $\lambda^{(k)}$ for each cell $C^{(k)}$, the overall proportion of observed zeros (0.237) is necessarily reproduced by the fitted model. For under this model, the indicator variable $E[\mathbb{I}\{Y = 0\}] = \tilde{\pi}_0$ (where $\mathbb{I}$ is an indicator function, i.e. $\mathbb{I}\{Y = 0\} = 1$, when $Y = 0$, and zero otherwise) irrespective of cell and is estimated by the overall proportion of zeros p_0. But as the eight individual cells have different numbers of zero observations, $\hat{\tilde{\pi}}_0 \neq p_0^{(k)}$ in general. As the zero-truncated Poisson distribution is in the exponential family, the fitted model reproduces the means of the zero-truncated data; thus, $E[Y^{(k)}|Y^{(k)} > 0]$ is estimated by the truncated mean $\overline{y}^{(k)}_{+}$ for each cell, and hence, the overall fitted values, $\left(1 - \hat{\tilde{\pi}}_0\right)\overline{y}^{(k)}_{+} = (1 - p_0)\overline{y}^{(k)}_{+}$, differ from the sample values $\overline{y}^{(k)}$ as seen in Figure 1(b).

More general hurdle models, in particular where $\tilde{\pi}_0$ is itself modelled via covariates in a regression context, involve specifying a particular form for the binary model. In addition to any specific linear predictor, this specification includes the choice of a link function for the zero probability $\tilde{\pi}_0$, or equivalently for the probability of non-zero counts $\tilde{\pi}_+ = 1 - \tilde{\pi}_0$, where $\tilde{\pi}_+$ can be thought of as the probability of crossing the hurdle. The link function can be taken as any standard binary/binomial model link but can also be motivated by considering the zeros to have come from a particular (truncated) count distribution; for example, Mullahy (1986) considers the use of the Poisson distribution leading to a complementary log–log link for the hurdle crossing probability $\tilde{\pi}_+$.

In the general form of Mullahy's Poisson hurdle, the zero and non-zero counts are both assumed to come from Poisson distributions but with different (modelled) parameters. A simplified version of this, with a single additional ZI parameter, takes the Poisson parameters as proportional with $\tilde{\pi}_0(\lambda, \alpha) = \pi_0(\alpha\lambda)$ with $\alpha < 1$ for zero-inflation and where $\alpha = 1$ reduces to a common Poisson model for zero and non-zero-counts. In this shared parameter version of the hurdle model, we no longer have the separation of the zero and non-zero processes and consequent simplification of the estimation. Here, both zero and non-zero counts contribute to the estimation of the regression parameters, and not surprisingly, this model does better than the hurdle at reproducing the overall means and also manages to pick up some of the differences in zero proportions.

The other broad class of explicit ZI approaches are the ZI mixture models, exemplified by the zero-inflated Poisson (ZIP) model of Lambert (1992). Here, the zero probability is extended to a mixture of the base zero probability and degenerate point mass at zero, where the mixing probability provides the additional parameter and corresponds to the specific inflation of the

zero probability. In the usual mixture-like notation, we have $\tilde{\pi}_0(\lambda, q) = q + (1 - q)\pi_0$ for $q \in [0, 1]$, with $\rho = 1 - q$. This can be written in our generic extended form $\tilde{\pi}_0(\lambda, \gamma)$ by parameterising q as a function of γ. For the Trajan data, we use a constant mixing probability q (equivalently constant γ) and separate $\lambda^{(k)}$ values for each of the eight classes $C^{(k)}$. From Figure 1(b), we see that here the results are very similar to those for the hurdle model; this is because the cell means are generally large and so the base Poisson model zero probabilities, $\exp(-\hat{\lambda}^{(k)})$, are small and contribute little to the fitted zero probabilities that are dominated by the estimated constant zero-inflation $\hat{q}$.

As a further novel explicit ZI model we introduce the 'logistic ZI' model, where the alteration of the zero probability is made through its odds ratio. The resulting extended distribution is in the exponential family, and for the Trajan data interaction model with the eight distinct $\lambda^{(k)}$ and a single ZI parameter, the sufficient statistics are the individual cell means, $\overline{y}^{(k)}$, and the overall proportion of zero counts. Consequently, the maximum likelihood fitted model reproduces the eight cell means and the overall proportion of zeros and, here, also seems to recover much of the individual cell zero probability structure, all with a single additional parameter.

While we would not claim that the pattern of behaviour across the four ZI types seen with the Trajan data is going to hold in general, we do feel that there are some important messages. First, the choice of ZI model may have unforeseen consequences on the overall fit, although the specific ways in which this may happen are rather subtle. Second, mis-specification of the ZI parameter (as here in treating it as constant when then is a large difference across the two photoperiod settings) can severely impact the estimation of the mean component of the model and covariate parameters of interest.

Note that for the Trajan data if a full interaction model is also used for the additional ZI parameter, then all four types are equivalent, as the likelihood reduces to one based on eight independent samples with their own parameters. These comments suggest that in practice if interest is in the mean model, then it would be sensible to use a rich model for the ZI parameter, which will provide more robust inferences. Of course, if there is also explicit interest in the covariates affecting the additional zero part, then some care may be needed in choosing an appropriate model and the development of some targeted diagnostics to help in this would be valuable.

2.2 Implicit Zero-inflated Extensions

We now consider the alternative approach of using OD models with apparently ZI data. In general, many OD models are one-parameter extensions of the basic Poisson model, and we write their probability function as $\pi_y(\lambda, \phi)$ with an additional dispersion parameter ϕ, where often $\phi = 0$ corresponds to the Poisson. Here, we restrict our attention to fitting two different forms of NB model. The canonical NB model arises from a Poisson-gamma mixture taking $Y \sim \text{Poisson}(\lambda V)$ where V is a $\text{Gamma}(1/\phi, 1/\phi)$ distributed random variable with mean 1 and variance ϕ. The resulting NB distribution, which we refer to as NB-quad, has mean $\mu = \lambda$ and a quadratic variance function $\mu + \phi\mu^2$ and for a fixed value of ϕ it is in the exponential family. The probability of a zero observation is $\pi_0(\lambda, \phi) = (1 + \phi\lambda)^{-\phi^{-1}} \geq e^{-\lambda}$ with equality for $\phi = 0$, showing that, for $\phi > 0$, this model does indeed inflate the probability of a zero observation compared with the Poisson distribution. For the Trajan data fitting, the NB-quad with the full eight-parameter interaction model and a single dispersion parameter ϕ, we see from Figure 1(a) that this model also reproduces the cell means and hence has some robustness for inference on regression parameters. However, while the single additional over-dispersion parameter accounts for some of the zero-inflation, it fails to model the zero

proportions well. Of course, as with the explicit ZI models, by including a model for the over-dispersion parameter, we can substantially improve the fit of the zeros, at the price of added complexity.

For comparison, we also consider a different parameterisation of the NB model (arising from a different version of the Poisson-gamma mixture) that has a linear variance function, $\mu(1+\phi)$ and which we refer to as NB-lin. As with all Poisson mixture models, as well as over-dispersion, the probability of zero is inflated compared with the Poisson with now $\pi_0(\lambda, \phi) = 1/\{(1+\phi)^{1/\phi}\}^{\lambda} \geq e^{-\lambda}$, again with equality for $\phi = 0$. This model is not in the exponential family, and even when fitting the full eight-parameter interaction model for the $\lambda^{(k)}$, it does not reproduce the cell means (Figure 1). However, it does seem to recover more of the structure in the zero proportions. The point here is that different OD count models may perform better, or worse, in particular examples. Of course, there are many other different over-dispersion models that we could consider, but none are going to be a panacea for fitting all ZI data.

The intention here is not to present a definitive analysis but rather to consider a simple common model where differences are apparent between the different implicit and explicit ZI models. In particular, we restrict our attention to a single ZI or over-dispersion parameter. In practice, as suggested above, we may want this parameter to depend on covariates and then many of the ZI models can be very similar, or indeed identical, and over-dispersion models can also give similar fits, as noted by Warton (2005).

Here, for explicit ZI models, we have taken the base model to be Poisson, but this is not necessary and we could use any suitable count distribution that could itself be a one (or more)-parameter extension of the Poisson. Such combinations of explicit and implicit approaches lead to two-parameter extensions $\tilde{\pi}_y(\lambda, \phi, \gamma)$ of the Poisson model and offer yet more flexibility. Additionally, in principle, it is also possible to have regression models for each of the three parameters. However, such flexibility and complexity comes at a price, and it would require a very rich dataset to distinguish between different models.

3 Explicit Models for Excess Zeros

Here, we consider formally the explicit ZI modelling of excess zeros in count data regression. We revisit and generalise the analysis of the Trajan data. Our objective in this is to put extant models into a theoretical framework that seems to be novel. This emphasises the functional form of what was described above as the specific alteration of the Poisson π_0 to $\tilde{\pi}_0$. This apparently new perspective may facilitate the (under-developed) constructive criticism of model fit, generally, using the Trajan analysis as a primitive example. Further, a new ZI type emerges naturally from the theory, which can inherit the attractive properties of exponential family probability distributions.

Formally, we discuss a wide family of one-parameter extensions of the count distribution $\pi_y(\lambda)$ underlying a generalised linear model (GLM), leading via coefficients to new estimates of the mean and thus of the probability of zero counts. The extensions are based on a new model $\tilde{\pi}_y$ differing from the *base* π_y via a parameter, here termed γ. This itself could be modelled as depending on covariates via further coefficients. This is indeed routine, but is not our immediate concern. Here, by studying a very wide range of functional options for $\tilde{\pi}_y(\lambda, \gamma)$, we hope to assist in the search for parsimonious ZI alternatives to the base probability mass function (pmf).

We first formalise the modelling in the previous section. The immediate focus is with models that *explicitly* address the apparent excess of zeros, but only zeros. The base could in fact be provided by one of many two (or more)-parameter OD pmfs, denoted here by $\pi_y(\lambda, \phi)$; often these are themselves generalisations of the Poisson. But for simplicity of notation, we typically

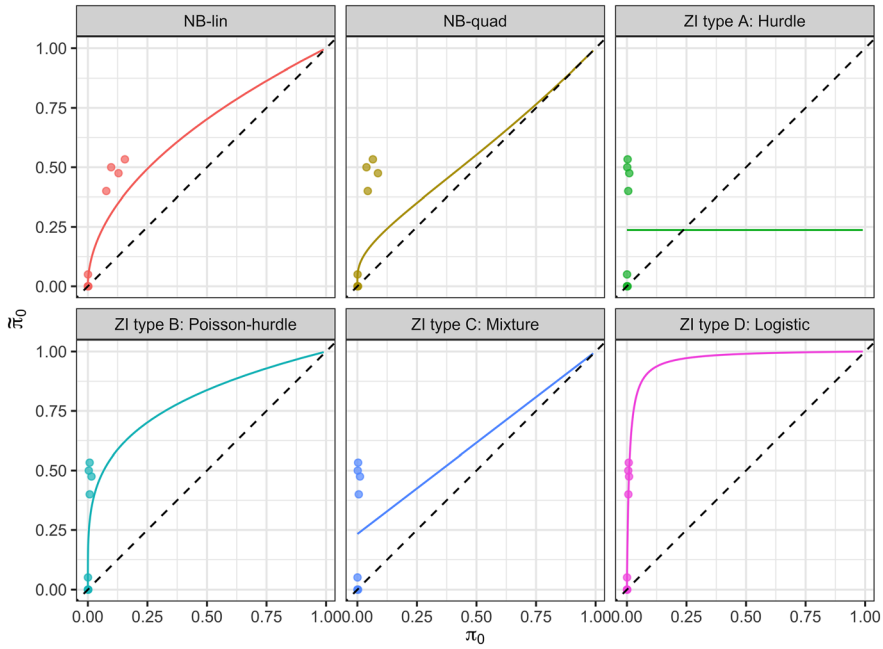

Figure 2. *Altered probability of zero $\tilde{\pi}_0$ versus probability of a zero based on the Poisson distribution π_0 for the NB-lin, NB-quad and ZI models types A, B, C and D, fitted to the Trajan data. Points correspond to the pairs $\left(\hat{\pi}_0^{(k)}, p_0^{(k)}\right)$, $k \in \{1, \ldots, 8\}$, that is, the eight combinations between photoperiod and hormone concentration. The dashed identity line is the Poisson base*

suppress the second parameter. Here, we refer to a zero-altered distribution $\tilde{\pi}_y$ as defined by a function $\tilde{\pi}_0(\pi_0, \gamma)$ with consequent implications for $\tilde{\pi}_y$, $y \neq 0$. As we develop below, the functional form defines a type of ZI, parameterised in the notation below by γ. Three such types (A, B, C) were discussed in Section 2 and a fourth (type D) mentioned.

The *a priori* need for such an extension was established there by a comparison of proportions $p_0^{(k)}$ with fitted $\hat{\pi}_0^{(k)} = \pi_0(\hat{\lambda}^{(k)})$, where π_y denoted a Poisson GLM, leading to the maximum likelihood estimate for each $\lambda^{(k)}$. These plots re-appear in Figure 2. In each of the panels, the Poisson GLM is represented by the diagonal; the points are pairs $(\hat{\pi}_0^{(k)}, p_0^{(k)})$. Note now that the horizontal axis in Figure 1 denotes $\tilde{\pi}_0$ for the given model whereas in Figure 2 it denotes the (Poisson) base π_0. Note also the full unit square is used for the plots to allow for examination of the functions underpinning explicit ZI modelling. As in the previous section, a full explanation of these functional forms is provided below and should be read in conjunction with Figure 2. Furthermore, the OD models, which are also presented in Figure 2, should be read in conjunction with Section 4.

The details of these generalisations are the subject of this section, as we now develop, firstly focussing on the theory of the ZI types $\tilde{\pi}_y$, and subsequently by considering the implications for likelihood inference.

3.1 Explicit Zero-inflated Types

Formally, recalling that $\pi_y(\lambda)$ refers to an arbitrary base pmf, with $E_\pi[Y] = \lambda$ and possibly other parameters ϕ, we write

Table 2. *A typology of zero-inflation and over-dispersion.*

Type: Common names	*Explicit ZI function*
A: Basic hurdle, zero-altered, two-stage	$\text{logit}(\tilde{\pi}_0) = \gamma$
B: Poisson hurdle, zero-altered, two-stage	$\log(-\log(\tilde{\pi}_0)) = \gamma + \log(-\log(\pi_0))$
C: ZIP, mixture, Lambert's mixture	$\log(1 - \tilde{\pi}_0) = \gamma + \log(1 - \pi_0),\ \gamma \leq 0$
D: Logistic ZI (new—this paper)	$\text{logit}(\tilde{\pi}_0) = \gamma + \text{logit}(\pi_0)$
OD dist	*Implicit ZI function*
NB-lin $\text{Var}(Y) = \mu + \phi\mu$	$\log(\tilde{\pi}_0) = \phi^{-1}\log(1+\phi)\log(\pi_0^P)$
NB-quad $\text{Var}(Y) = \mu + \phi\mu^2$	$\log(\tilde{\pi}_0) = -\phi^{-1}\log\left[1 - \phi\log(\pi_0^P)\right]$

Our four types and their common names are shown alongside their ZI behaviour with respect to the Poisson base. NB, negative binomial; OD, over-dispersed; ZI, zero-inflated; ZIP, zero-inflated Poisson.

$$\tilde{\pi}_y(\lambda, \gamma) = \begin{cases} \tilde{\pi}_0(\pi_0, \gamma) & \text{if } y = 0 \\ \rho(\pi_0, \gamma)\pi_y(\lambda) & \text{if } y > 0 \end{cases} \tag{1}$$

where ρ re-normalises. The requirement that $\sum_y \tilde{\pi}_y = \sum \pi_y = 1$ leads to $\rho = (1 - \tilde{\pi}_0)/(1 - \pi_0)$. Note that, under $\tilde{\pi}_y$, $E_{\tilde{\pi}}[Y] = \mu = \rho\lambda$ and so the mean is also altered.

The specific function $\tilde{\pi}_0(\pi_0, \gamma)$ of π_0 characterises the *type* of ZI and is parameterised by γ; it is such functions that are illustrated in Figure 2. Typically, but not necessarily, $\gamma = 0$ denotes the null case $\tilde{\pi}_0 = \pi_0$. Observe that parameters (λ, ϕ) enter $\tilde{\pi}_0$ and ρ only *via* π_0. Note also that the *relative* magnitudes of $\tilde{\pi}_y$ for non-zero y are the same as those for the base π_y. Equivalently, the conditional pmfs, given $y > 0$ (i.e. the zero-truncated distributions), are the same, whether based on $\tilde{\pi}_y$ or the base π_y. In particular, $\mu^+ = E_{\tilde{\pi}}[Y|Y > 0] = E_{\pi}[Y|Y > 0] = \lambda^+$. It is such conditional pmfs that distinguish explicit and implicit approaches to the modelling of excess zeros.

We now consider special cases. Three are relatively common. We will see below that many of the interesting cases can be written as

$$g(\tilde{\pi}_0) = \gamma + g(\pi_0).$$

They are all summarised in Table 2.

3.1.1 Zero-inflation type A

The simplest model has constant $\tilde{\pi}_0 \in [0, 1]$. With this restriction, we could write $\tilde{\pi}_0 = \gamma$. But here, as for other types, γ may of course depend on covariates; we thus write $g(\tilde{\pi}_0) = \gamma$ for some link function $g(.)$. In this form, γ is typically unconstrained.

This form is the hurdle model discussed in Section 2; it could be described as the *constant* type of ZI. But its use in this very basic form (constant $\tilde{\pi}_0$) is almost degenerate. Its inadequacy for the Trajan data is apparent (Figure 2 top right panel). Note that here, and equivalently in the other panels, the value of γ used for the function is the MLE $\hat{\gamma}$ for the Trajan data under type A, as discussed in Section 2. In this case, it is that value of γ that renders $\tilde{\pi}_0 = p_0$, the overall proportion. Further, for this and other panels, the 'data points' are pairs $(\pi_0(\hat{\lambda}^{(k)}), p_0^{(k)})$ under each ZI type. They thus correspond to the plotted functions $\tilde{\pi}_0(\pi_0(\lambda), \hat{\gamma})$, where the value of

$\hat{\gamma}$ is specific to the ZI type. For type A, the MLE $\hat{\lambda}^{(k)}$ is derived from the truncated mean $\overline{y}^{(k)}_{+}$, as previously discussed. We formalise such inference in the next subsection.

Observe that this ZI type is zero-inflationary for small π_0, but zero deflationary for large π_0. This is a necessary consequence of its motivation, where zero counts are generated by a process quite independent from positive counts. Note also that the neutral case of $\tilde{\pi}_y = \pi_y$ is *not* a special case of this model.

3.1.2 Zero-inflation type B

Type B is that referred to in Section 2 as the Poisson hurdle. It can be expressed as $\tilde{\pi}_0 = \pi_0(\alpha\lambda) = (\pi_0)^{\alpha}$ for positive α; for the Poisson base model, we have $\tilde{\pi}_0 = e^{-\alpha\lambda}$, that is, coming from a Poisson model with mean $\alpha\lambda$. An alternative form is through the complementary log–log link function with unconstrained additional parameter $\gamma = \log(\alpha)$ and

$$\log(-\log(\tilde{\pi}_0)) = \gamma + \log(-\log(\pi_0)),$$

giving $\tilde{\pi}_0 = (\pi_0)^{e^{\gamma}}$. (Here and throughout the paper, log refers to the natural logarithm, i.e. $\log_{e^{\cdot}}$.) Observe that $\gamma = 0$ defines the neutral case, with positive and negative γ corresponding to under-inflation and over-inflation. The ZI type B panel in Figure 2 (bottom left panel) makes it quite clear that this more clearly reflects the frequency of zeros observed in the Trajan data.

3.1.3 Zero-inflation type C

In the notation of Section 2, this can be written as $(1 - \tilde{\pi}_0) = (1 - q)(1 - \pi_0)$ for $q \in [0, 1]$; zero-inflation is thus non-zero deflation by a constant factor. It could be characterised as the 'linear model'. It is very widely used; indeed for very many authors, this model is coterminous with ZI. For the Trajan data, like type A, it is clearly a poor model of the observed zeros (Figure 2 middle bottom panel).

Its simplest link function expression is via the complementary log function for $\gamma \leq 0$.

$$\log(1 - \tilde{\pi}_0) = \gamma + \log(1 - \pi_0)$$

It is apparent that for type C $\log(\rho) = \gamma$. When $\gamma = 0$, $\tilde{\pi}_0 = \pi_0$ for all π_0; here, as in type B, the neutral $\tilde{\pi}_0 = \pi_0$ is a special case.

Like types B and D, and in contrast to A, $\tilde{\pi}_0 = 1$ when $\pi_0 = 1$. But like type A, when $\pi_0 = 0$, $\tilde{\pi}_0 > 0$; this distinguishes itself qualitatively from type B, and also from D below. This is the feature that marks both types A and C as inappropriate single-parameter extensions for the Trajan data. Note also that the model is also defined for $\gamma > 0$, subject to $\tilde{\pi}_0 > 0$. Thus, type C includes under-inflation; but now γ is constrained by a function of λ. More formally, when $\gamma > 0$, we must write $\tilde{\pi}_0 = \max\{0, 1 - e^{\gamma}(1 - \pi_0)\}$.

3.1.4 Zero-inflation type D

Type D, which is new, is most simply expressed as

$$\text{logit}(\tilde{\pi}_0) = \gamma + \text{logit}(\pi_0)$$

for unconstrained γ. Equivalently, we have $\tilde{\pi}_0/(1 - \tilde{\pi}_0) = e^{\gamma}\pi_0/(1 - \pi_0)$. The alteration may now be seen as multiplicatively altering the odds ratio of a zero count. In closed form, it may be written as $\tilde{\pi}_0 = e^{\gamma}\pi_0/(1 + (e^{\gamma} - 1)\pi_0)$.

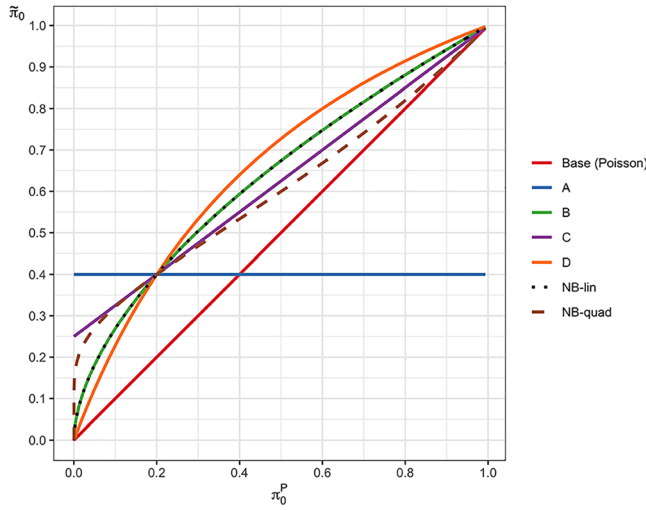

Figure 3. *Theoretical zero-inflated (ZI) relationships. We plot* π_0^P *as Poisson against* $\tilde{\pi}_0$ *for the ZI and over-dispersed types defined in column 1 of Table 2. The parameters of each distribution are matched so that each line (except for Poisson) travels through the point (0.2, 0.4) for a range of means. This yields differing ZI parameters for each distribution:* $\gamma_A = -0.405$, $\gamma_B = -0.563$, $\gamma_C = -0.288$ *and* $\gamma_D = 0.981$, *all provided to match the scale of Table 2. For the negative binomial (NB) models* $\phi_{\text{NB-lin}} = 1.82$ *and* $\phi_{\text{NB-quad}} = 1.13$, *again with reference to the parameterisation in Table 2*

Although its performance is similar to type B, arguably, it is the best of the four ZI types in describing the observed frequency of zeros in the Trajan data (Figure 2 bottom right panel). But it also differs from the others in a more theoretical sense, as we elaborate below. For, in many cases, pmfs involving type D extensions are often within the exponential family. It is for this reason that D is, for the Trajan analysis, the only one of the four ZI types for which $\hat{\mu}^{(k)} = \overline{y}^{(k)}$, as seen in Figure 1, sharing this property with the Poisson and NB-quad distributions.

We conclude this subsection by making two remarks. Firstly, the use of the unit square for the plots of $\tilde{\pi}_0(\pi_0, \gamma)$ emphasises a peculiar feature of the design of the Trajan data. It is lacking any samples corresponding to small means λ and thus large π_0. It is this that renders it difficult to distinguish which is the 'best' model, despite its other peculiar feature—replication—that gives access to conditionally iid data and thus to the proportions p_0 of observed zeros. An even more extreme example would be fully iid data. There we would see that all the four ZI types would in fact be re-parameterisations of each other and would thus be impossible to distinguish. All four functions would intersect at $\tilde{\pi}_0 = p_0$.

In Figure 3, we use this to provide a different contrast of the functions. There it is seen that types B and C can be remarkably similar for middle to large values of π_0 and thus might be difficult to distinguish in a data design that did not include small values of π_0. Further, type D distinguishes itself from type B in two ways. Firstly, it has a symmetry (around the diagonal $\tilde{\pi}_0 = 1 - \pi_0$) that the others lack; but this may in fact be disadvantageous in some cases. For example, it can be steeper than B for very small π_0; but then, necessarily, it will be much flatter for small $1 - \pi_0$.

Secondly, we observe that there are many more options in the design of even the one-parameter functions $\tilde{\pi}_0 = \tilde{\pi}_0(\pi_0, \gamma)$ than the literature might suggest. These clearly extend beyond the four types we have used for illustration. They typically involve classic link functions $g(p)$, such as complementary log–log and logit. Interestingly, however, the particular

characteristic of the overwhelmingly popular type C—namely, that $\tilde{\pi}_0 > 0$ when $\pi_0 = 0$—seems not to be a property associated with any popular link function other than complementary log–log or close but uncommon relatives such as $g(p) = 1/(1 - p)$. Furthermore, they may all be used with OD base pmfs $\pi_y(\lambda, \phi)$.

We now address issues in inference.

3.2 Inference

We focus on likelihood inference and illustrate with the analysis of the Trajan data in Section 2, which provides some new insight. We do not dwell on algorithms or second-order issues; but it may be that our perspective will stimulate ideas on the critical evaluation of explicit ZI modelling. The theory thus focusses on the score equations, being differentials of $\log L = \sum_i \log(\tilde{\pi}_{y_i}) = \sum_i \ell_{y_i}(\mu(\lambda, \gamma), \gamma)$, wrt the parameters γ, λ_i or functions of these such as $\tilde{\pi}_0$, μ_i, and/or the coefficients β underlying the λ terms. This extends of course to cases of π_y where the ϕ parameter is non-null; but we do not dwell on this. Recall that, for the Trajan model, there are nine parameters: one common γ term and eight values of $\lambda^{(k)}$, or equivalently of $\mu^{(k)}$. Particular interest lies in useful decompositions of the likelihood, as we now discuss.

The full distribution $\tilde{\pi}_y(\lambda, \gamma)$ can be written in various forms, some of which lead to different decompositions of $\log L$. Although all forms apply to all ZI types, these are most simply interpreted for types A and D. Equation (2) decomposes into components of the likelihood focussing on zero and non-zero counts. It sheds light on the fitted functions in Figure 2 for all types and on the (λ, γ) parameters in type A. Equation (3) sheds light on type D, for this is seen often to lie within the exponential family, with consequent other decompositions.

We begin by noting alternative versions of (1), which are particularly insightful

$$\begin{aligned}\tilde{\pi}_y &= (\tilde{\pi}_0)^{\mathbb{I}\{y=0\}} \left(\rho\pi_y\right)^{1 - \mathbb{I}\{y=0\}} = (\tilde{\pi}_0)^{\mathbb{I}\{y=0\}} \left(\frac{1 - \tilde{\pi}_0}{1 - \pi_0}\pi_y\right)^{1 - \mathbb{I}\{y=0\}} \\ &= \left((1 - \tilde{\pi}_0)^{1 - \mathbb{I}\{y=0\}} \tilde{\pi}_0^{\mathbb{I}\{y=0\}}\right) \left(\pi_y^+\right)^{1 - \mathbb{I}\{y=0\}} \qquad (2) \\ &= \left(\frac{\tilde{\pi}_0}{\rho\pi_0}\right)^{\mathbb{I}\{y=0\}} \rho\pi_y = e^{\gamma \mathbb{I}\{y=0\}} \left(\frac{\pi_y}{\pi_0}\right) e^{-\gamma} \tilde{\pi}_0, \qquad (3)\end{aligned}$$

where $\pi_y^+ = \pi_y/(1 - \pi_0)$ is the zero-truncated form of π_y and e^γ denotes the odds ratio for $(\tilde{\pi}_0, \pi_0)$. From (2), with an obvious notation $\log\tilde{\pi}_y(\lambda, \gamma) = \ell_y^0(\tilde{\pi}_0(\pi_0, \gamma)) + \ell_y^+(\lambda)$, the log-likelihood decomposes as

$$\log L = \sum_i \ell_{y_i}^0(\pi_0, \gamma) + \sum_i \ell_{y_i}^+(\lambda_i). \qquad (4)$$

This result applies to any ZI type, including A–D above. Importantly, the second term is independent of γ. The first term depends on $\mathbb{I}\{y_i = 0\}$ and supplies the only information on γ through $\tilde{\pi}_0$; but recall that, apart from type A, $\tilde{\pi}_0$ is also a function of λ through π_0, and thus, the first term generally supplies some information on λ. The usual calculus leads to score functions whose solution is the MLE for (λ_i, γ) (and for any other parameters, ϕ). In general regression, the λ_i are dictated by coefficients, these and γ being the true target of the inference.

In the case of type A, of course, $\tilde{\pi}_{0i}$ is independent of $\hat{\pi}_{0i}$, this being $e^{-\hat{\lambda}_i}$ for the Poisson base. In the case of iid observations, as within a single cell of the Trajan design, $\hat{\gamma}$ is such that $\tilde{\pi}_0(\hat{\gamma}) =$

p_0 the observed proportion of zeros in the data. Recall that this remark applies to *all* two-parameter (λ, γ) types of ZI in the iid case. However, in the Trajan case with common γ across all cells, results will differ with ZI type, as we now illustrate.

It is useful first to consider in more detail the estimation of λ under type A, first for the iid case and then for the Trajan design. For type A, from (2), the estimation of λ focusses solely on $\ell_y^+(\lambda)$ and thus on the zero-truncated distribution $\pi_y^+(\lambda)$. The expected value of the zero-truncated distribution is $(1-\pi_0)^{-1}\lambda$ for which the MLE is $\overline{y}^+$, this being the case for all pmfs in the exponential family, such as the Poisson and NB-quad, and truncated versions thereof. Thus, $\hat{\lambda} = (1 - \hat{\pi}_0)\overline{y}^+$. But $E_{\tilde{\pi}}[Y] = \mu = \rho\lambda$; it is thus estimated, for type A, by

$$\hat{\mu} = \hat{\rho}\hat{\lambda} = \frac{1 - p_0}{1 - \hat{\pi}_0}(1 - \hat{\pi}_0)\overline{y}^+ = \overline{y}.$$

Again, because all two-parameter ZI types are equivalent in the iid case, this result applies also to them, as is well known. We revisit this for type D below, where this demonstration is very simple.

But in the Trajan model, there are eight cells with $\lambda = \lambda^{(k)}$ in each but sharing a *common* ZI parameter γ, defining for all ZI types, a common $\tilde{\pi}_0$ estimated by the overall p_0. The term ℓ_y^+ decomposes into sums over each of these cells, each leading, separately, to estimates of each $\lambda^{(k)}$. In the case of type A, these are zero-truncated means and thus lead to estimates

$$\hat{\mu}^{(k)} = \hat{\rho}^{(k)}\lambda^{(k)} = (1 - p_0)\overline{y}^{+(k)} \neq (1 - p_0^{(k)})\overline{y}^{+(k)} = \overline{y}^{(k)}.$$

It is this that leads, in Figure 1, to the fact that ZI type A leads to values $\hat{\mu}^{(k)}$ other than $\overline{y}^{(k)}$, unlike the simple Poisson. Similar but technically more difficult arguments apply to other ZI types. The exception is type D, which does lead to $\hat{\mu}^{(k)} = \overline{y}^{(k)}$.

The second form for $\tilde{\pi}_0$ (Equation 3) is of particular importance for type D when the base distribution $\pi_y(\lambda)$ is, like the Poisson, within the exponential family. For the Trajan analysis, it highlights an important distinction between the models corresponding to (here similar) types B and D. In particular, it explains why, in Figure 1, pmfs for the Poisson, NB-quad and ZI type D with Poisson base both lead to the same estimators of the cell expected values. ZI type D may thus be a useful addition to the ZI types.

The key point in Equation (3) is that, for type D, the odds ratio for $\tilde{\pi}_0$ and π_0 is taken to be constant, denoted e^{γ}. As a consequence, the parameter γ can have a particular significance, when $\pi_y(\lambda)$ is in the (one-parameter) exponential family and can thus be written as $\log(\pi_y) = y\eta(\lambda) - A(\lambda) + c(y)$; for the Poisson, the natural parameter is $\eta = \log(\lambda)$. Firstly, defining $c(y)$ such that $c(0) = 0$, it is clear that the cumulant function $A(\lambda) = -\log(\pi_0)$, being λ for the Poisson; and further, from the properties of the exponential family, we have $E[Y] = A'_\eta(\lambda, \gamma)$, being λ in the Poisson case. From (3), we thus have

$$\log(\tilde{\pi}_y) = y\eta(\lambda) + \mathbb{I}\{y = 0\}\gamma - \tilde{A}(\lambda, \gamma) + c(y), \tag{5}$$

where $\tilde{A}(\lambda, \gamma) = \gamma - \log(\tilde{\pi}_0)$, which plays the same role as $A(\lambda)$ for π_0. The implication is that $\tilde{\pi}_y$ is in the two-parameter exponential family, inheriting this property from the π_y. The natural parameters are (γ, η).

Further, the log-likelihood in Equation (4) includes a decomposition into the sum of two terms, being $\ell_y^0(\gamma)$ and $\ell_y^+(\lambda)$. And indeed, if $\pi(y)$ is in the m-parameter family, then $\tilde{\pi}_y$ is in

the $(m+1)$-parameter family. It follows immediately that sufficient statistics for an iid sample $y_1, \ldots, y_n$ from $\tilde{\pi}_y$ are $(\sum y_i, \sum \mathbb{I}\{y_i = 0\})$, being equivalent to $(\bar{y}, p_0)$ and having expected values already known here to be $(\mu, \tilde{\pi}_0)$, recalling that $\mu = \rho\lambda$. This result is also available in general from differentiating $\tilde{A}(\lambda, \gamma)$. And, following the observation above on the equivalence, in the iid case, of all two-parameter ZI types, $\bar{y}$ and p_0 are unbiased MLE estimators of μ and $\tilde{\pi}_0$, as already shown above, via the more cumbersome type A arguments. This is the reason why ZI type D (with Poisson base) shares this property with the usual Poisson and NB-quad models. And further, we can immediately assert that ZI type D with NB-quad as base will also share this property.

4 Implicit Zero-inflation

As we have already noted, over-dispersion models, such as the NB, have often been used for the analysis of count data with excess zeros. Indeed, some authors have suggested that, in practice, such OD models may be sufficient with no need to explicitly model the excess zeros. In Section 2, we considered the use of particular NB models, and here, we provide a little more detail. Of course the NB family is just one form of OD extension of the Poisson distribution, albeit the most widely studied and used. General mixing of a Poisson with any distribution (also often referred to as compounding) leads to OD count distributions. Other generalisations include weighted Poisson models (Del Castillo & Pérez-Casany 1998) extending the original idea of size weighting of counts in Rao (1965). This general family of models includes the currently popular COM-Poisson distribution (Shmueli *et al*., 2005; Sellers & Shmueli 2010; Ribeiro *et al*., 2020), and unlike the mixture-based over-dispersion models, this (and other members of the family) can also accommodate under-dispersion. Other classes of extended Poisson models that can handle both over-dispersion and under-dispersion include the generalised Poisson (Consul & Jain, 1973), the gamma count (Zeviani *et al*., 2014) based on using gamma-distributed inter-arrival times rather than the exponential times as in a Poisson process and the discrete Weibull distribution (Luyts *et al*., 2019) as an example of the general approach of discretisation of a continuous random variable. While OD versions of these can be used to model excess zeros, the combination of these models with explicit ZI also provides the possibility of modelling excess zeros where the non-zero counts are under-dispersed.

In many OD models, not only do we have $\mathrm{Var}(Y) > \mathrm{E}[Y]$ but also, often as a consequence, $\pi_0(\mu, \phi) > \pi_0^P$. Simply stated, in general, over-dispersion puts more weight in the tails of a distribution, but as a count random variable Y is constrained below by zero, this extra weight in the lower tail accumulates on low values of Y and, in particular, on $Y = 0$. As such, these models can provide an alternative approach to the apparent problem of excess zeros in data. As this is an indirect consequence of the OD model, we call this *implicit* ZI. Puig & Valero (2006) provide a theoretical treatment of the circumstances under which OD induces ZI, characterising the conditions under which this is necessary. They use it to also contrast several count distributions; one of these is the NB, which here we use as an exemplar of OD models. Interestingly, as we develop below, the details depend not only on the OD distribution but also on its parameterisation. In Figure 2, we show the implied ZI behaviour of these models for comparison with that of the explicit ZI models.

4.1 Negative Binomial

We focus on two re-parameterisations of the NB with pmf

$$\pi_y(\mu, k) = \frac{\Gamma(k+y)}{y!\Gamma(k)}\left(\frac{k}{k+\mu}\right)^k\left(\frac{\mu}{k+\mu}\right)^y$$

which has $\mathrm{E}[Y] = \mu$ and $\mathrm{Var}(Y) = \mu + \mu^2/k$ and $\pi_0 = (k/(k+\mu))^k$. The most frequently used, to which we refer as NB-quad, has $k = \phi^{-1}$, $\mathrm{Var}(Y) = \mu + \phi\mu^2$ and $\pi_0 = (1+\phi\mu)^{-\phi^{-1}}$. An alternative, NB-lin, has $k = \mu\phi^{-1}$, with $\mathrm{Var}(Y) = \mu(1+\phi)$ and $\pi_0 = 1/\{(1+\phi)^{\phi^{-1}}\}^{\mu}$. The Poisson is a special case of both, corresponding to $\phi \to 0$. There are in fact other versions of the NB, generically NB-P (Gurmu & Trivedi, 1996), having $\mathrm{Var}(Y) = \mu(1+\phi\mu^{p-1})$ with a power relationship for the over-dispersion index. Effectively, this introduces a third parameter; but we do not pursue this option.

These two versions of the NB are included in Figure 2, where, as discussed in Section 3, the extended model zero probabilities $\tilde{\pi}_0$ are plotted against the Poisson base π_0, augmented by the observed and model zero probabilities for the Trajan data. Here, we have expressed $\tilde{\pi}_0(\mu, \phi)$ as a function of π_0^P via $\mu = -\log(\pi_0^P)$. Figure 3 also illustrates this for the two different parameterisations of the NB distribution. In constructing this theoretical plot, we have fixed the dispersion parameter ϕ so that the curves pass through the same common $(\pi_0, \tilde{\pi}_0)$ point (0.2, 0.4). For both, the induced ZI, characterised by $\tilde{\pi}_0(\pi_0^P, \phi)$, is such that $\tilde{\pi}_0 \to 0$ as $\pi_0^P \to 0$ or equivalently as $\mu \to \infty$, or equivalently as $\pi_0^P \to 0$, as with the ZI of types B and D. More specifically, for NB-lin, from $\log(\tilde{\pi}_0(\pi_0^P, \phi)) = \phi^{-1}\log(1+\phi)\log(\pi_0^P)$, if we set $e^{\gamma} = \phi^{-1}\log(1+\phi)$, the ZI induced by NB-lin is *exactly* as type B for all π_0^P and thus for all μ. Also, except for large μ (small π_0^P), this induced ZI bears some resemblance to type C, the classic mixture model of ZI. Further, we see that, apart from the behaviour for very small π_0^P (i.e. for π_0^P above the 'elbow'), the ZI function $\pi_0(\pi_0^P, \phi)$ for NB-quad is even more similar to that for type C. This is essentially because here this function has unit slope both at the elbow and, unlike NB-lin, as $\pi_0^P \to 1$ (equivalently $\mu \to 0$).

But recall that, despite ZI type B and NB-lin having parameters γ and ϕ such that they have the same probabilities of zero for all μ, they are quite different distributions and have different variance functions. Although there are similarities in behaviour between the NB-quad and type C models, their probability structures are different and, for example, $\tilde{\pi}_y^C$ can be bi-modal. The point is that the over-dispersion in the type C model is centred solely on the zero probability, whereas in the NB-quad, the extra dispersion is smoothly spread across the whole range of y-values. The central challenge is that with only two parameters available, correspondence cannot be obtained on three important aspects: mean, variance and the probability of zero.

4.2 Simple Inference for the Negative Binomial

Inference is well established for the NB model. But there is a subtlety, even in the iid case. Moments-based estimators of (μ, ϕ) will necessarily match the sample mean and variance, and thus not match p_0, the observed frequency of zeros, especially if this is large. But other, more widely used, estimators will, especially for large p_0, result in a $\hat{\phi}$ that (approximately) matches the fitted $\pi_0(\hat{\mu}, \hat{\phi})$ by overestimating the variance. The theory is more straightforward for NB-quad.

NB-quad is a member of the exponential dispersion family. Here, in the iid case, the sample mean is the MLE for μ, while the MLE for the dispersion parameter ϕ requires an iterative solution of the score equation. However, the joint ML estimators do have the nice property of being asymptotically independent (Lawless, 1987). Other estimators used for ϕ include a

moment estimator (this can be obtained for more general models by equating the generalised Pearson chi-square statistic to the degrees of freedom). Historically, a zero-frequency-based estimator has also been used and, as for the explicit ZI models, it results in the zero-frequency being fitted perfectly; however, for data generated by an explicit ZI Poisson process, it does this by overestimating the variance with a larger fitted ϕ value. Interestingly, in this ZI setting, the MLE lies between the moment and the zero-frequency estimators, giving both a fitted variance $\text{Var}_{\pi_0(\hat{\mu},\,\hat{\phi})}$ that is larger than the observed sample variance, and a fitted $\pi_0(\hat{\mu},\,\hat{\phi})$ that is closer to the observed p_0 than might be expected, at the price of an extended fitted upper tail. However, in applications to data with excess zeros, there is typically also some OD in the non-zero counts and so this inflated estimated value of ϕ may simultaneously capture both excess zeros and additional dispersion and lead to an adequate and parsimonious fit.

This behaviour perhaps explains the findings in Warton (2005), who provides comparisons between different implicit and explicit ZI models fitted to multivariate abundance data from 20 datasets. His results show that the NB-quad model is superior to ZIP and ZINB in terms of AIC, when looking at an average over 1 672 count variables. He comments that these abundances do not have extra zeros when compared with the NB-quad distribution and are likely to have arisen from NB-quad distributions with small means.

The NB-lin is not in the exponential family, even for a known value of ϕ, and is inferentially less convenient. Of course, in the iid case, the NB-lin is simply a re-parameterisation of the NB-quad, but for fixed values of the dispersion parameters, they behave differently as μ varies (Table 2 and Figure 3), and these differences potentially become apparent in the regression model setting.

5 R Packages for Fitting Zero-inflated Models

In this section, we describe the use and background of various commonly used R packages for modelling zero-inflation. There are many packages listed on the R package web page that purport to perform ZI regression, but many of these apply to specific data types or constraints (e.g. hidden Markov models, monotonic zero-inflation and compositional data), apply to a particular application area or deal in continuous rather than count data (the focus of this paper). Instead, we focus on four main packages: `zic`, `pscl`, `VGAM` and `gamlss`. A further popular package is `COZIGAM`, but this package has been archived from CRAN and has not been updated since 2012.

Table 3. *A list of* R *packages we use to demonstrate zero-inflated regression modelling.*

Name	*Approach*	*Likelihoods supported*	*Covariate types supported*
`zic`	Bayesian	Poisson (and Poisson log-normal)	Restricted to be the same in both regular and zero-inflated components. Also includes variable selection
`pscl`	Frequentist	Poisson, negative binomial and geometric (count distribution), and binomial, Poisson, negative binomial and geometric (zero-inflated distributions; hurdle only)	Restricted to be the same in both regular and zero-inflated components
`VGAM`	Frequentist	Poisson, negative binomial and geometric (zero-inflated and hurdle)	Zero-inflation components do not vary by covariate, but formulae allow for spline and other GAM-type relationships
`gamlss`	Frequentist	At least 12 different types (zero-inflated and adjusted)	Has the ability to model each parameter in a distribution separately using, for example, the GAM framework

The purpose of this section is to showcase the wide array of possibilities currently available in R for ZI/zero-adjusted modelling and provide brief examples of their use on our example data to enable those new to the field to pick up on the salient features. Table 3 shows a summary of these packages and an overview of their features and differences.

All of the aforementioned packages are Frequentist in their inferential approach with the exception of `zic`. A number of packages exist that potentially allow for Bayesian ZI and zero-adjusted models to be fitted, such as JAGS (Plummer *et al*., 2003), Stan (Stan Development Team 2014), INLA (Rue *et al*., 2009) and Nimble (de Valpine *et al*., 2017). However, these require considerable extra coding experience (and often many more lines of code), so we do not review them here. A particular key issue is that many of these packages support only a small set of probability distributions, so that extra distributions have to be added by hand (or hack) and often lack vectorisation speed-ups. For those wishing to persist in learning one or more of these packages (which we would recommend), we suggest first consulting the manual to determine whether the chosen custom probability distributions are included by default.

The `zic` package (Jochmann, 2013) uses Markov chain Monte Carlo (MCMC) to produce posterior distributions of ZI (though not hurdle) models with covariates. The covariates are assumed to be common across both the Poisson likelihood and the binary zero-inflation component. Only a Poisson likelihood is supported, although the model includes latent effects to account for over-dispersion to create, marginally, a Poisson log-normal likelihood. In addition, the package has methods to run a stochastic search variable selection (SVSS), a method that removes the need to choose between model structures.

R package `pscl` (Zeileis *et al*., 2008) contains a very wide array of both hurdle and ZI models. A key feature is the possibility of selecting different distribution types for π_y and π_0. Three different distributions are permitted for the counts (Poisson, NB and geometric) and three types of zero-inflation (binomial, Poisson, NB and geometric). These latter distributions are censored at value 1 to produce the required hurdle effect. For standard zero-inflation models, only binomial (i.e. Bernoulli) inflation is permitted. Link functions can also be changed (for both hurdle and ZI) although again some restrictions apply.

The `VGAM` package interface (Yee 2015) is very similar to that of the standard R base `lm` and `glm` functions while adding in multivariate components (via constraint matrices), which do not apply to the ZI models implemented. Exactly as with `glm`, the `vglm` and `vgam` functions take a `formula` and `family` argument, the latter of which has a variety of ZI options, including ZIP, NB and geometric, and similarly zero-altered (two-stage/hurdle) versions. We found the output of these models to be somewhat confusing, because the tabular output displays multiple intercepts (corresponding seemingly to the different linear predictors).

The `gamlss` package contains the richest set of ZI and zero-altered (two-stage/hurdle) models, including a large number of ZI continuous (and bounded) likelihood distributions. The count distributions supported include binomial, beta-binomial, NB, negative beta-binomial, logarithmic, Poisson and Zipf. The `gamlss` package in particular allows for the modelling of multiple different parameters in a distribution, each with their own parametric relationship, but none of the examples we found in the package used this feature.

Finally, we provide unified code to fit ZI types A, B, C and D (and reproduce all results obtained from analysing the Trajan data) at www.github.com/andrewcparnell/ZI_review.

6 Conclusions

In this paper, we have only examined the basic underpinnings of regression in the presence of excess zeros in univariate count data. Our first contribution to this is a novel perspective on the modelling of excess zeros in count data regression. This is primarily in the presentation of an

approach that we have referred as explicit ZI. The modelling issue, as presented here, is simply the choice of link function, and the consequent estimation of the sole parameter. Four such simple options are presented; others and extensions to two-parameter variations are possible. The two approaches discussed—explicit ZI and implicit ZI via the use of distributions such as the NB—may of course be combined, and all parameters may be modelled via covariates.

It could be argued that the real difficulty is the embarrassment of choice. The estimation of the parameters is no longer a challenge (for univariate counts), given modern computing algorithms; nor is the identification of the *best*, given a metric such as the AIC, a metric that may not be natural for some users. The real challenge is in fact the identification of the most *useful* model given the vagaries of that term, and the ubiquity of potential outliers within all datasets.

Our approach to the current paradigm of explicit ZI models contrasts with that of data generating mechanisms that involve latent variables. Of course, as pointed out by Cameron & Trivedi (2013), p. 147: 'The latent class interpretation is not essential …. As such the approach is an alternative to non-parametric estimation.' However, from the almost invariable discussion in papers in the applied literature, it does seem that the user community feels obliged to explain the mixture interpretation. This can be useful, of course; but sometimes it can involve shoehorning. Nevertheless, mixtures defined by latent variables can be a fruitful theoretical avenue for defining models for use with multivariate counts. Indeed, this was the purpose of Salter-Townshend & Haslett (2012).

The practical benefit of the new perspective in univariate regression may not in fact be new models. The non-parametric perspective may yield new diagnostics. It may even help to establish *a priori* evidence of excess zeros. Current practice usually involves marginalisation over all covariates. An empirical approach would involve, as a first step, a comparison between observed $\mathbb{I}\{y_i = 0\}$ and the fitted probabilities of zero under a default 'base' model, generically $\hat{\pi}_{0i} = \pi_0\left(\hat{\mu}_i,\ \hat{\phi}_i\right)$. This comparison could simply regress the binary $\mathbb{I}\{y_i = 0\}$—non-parametrically, using splines, for example—on functions including $\log(\hat{\pi}_{0i})$ and $\log(1 - \hat{\pi}_{0i})$, yielding values $\hat{\tilde{\pi}}_{0i}$. Any departure from the unit line $\tilde{\pi}_{0i} = \pi_{0i}$ would constitute *a priori* evidence of excess zeros with respect to the fitted base model.

Furthermore, such a plot might serve, *post hoc*, as a diagnostic for a fitted ZI model, by regressing $\mathbb{I}\{y_i = 0\}$ on functions of fitted $\tilde{\pi}_{0i}$. We remark that the current literature on count data regression in general—and on excess zeros in particular—seems sadly lacking in criticism that most users would deem to be constructive. Portmanteau statistics based on information criteria (IC) do not always have such a constructive interpretation. Model details can be dominated by extreme values; and this is particularly so for models that admit very large values for the variance. But IC statistics do not readily identify such data points.

In circumstances where it is deemed to be important to choose amongst alternative models for the excess zeros, such regressions highlight the importance of the design. For without data corresponding to very large μ_i (and thus small π_{0i}), it will be impossible to distinguish ZI models of type B, C or D and difficult to distinguish from the OD approach. In practice, of course, this will be further complicated if—as is common—the parameters γ and/or ϕ are themselves modelled via covariates. But the fact that this is common may reflect poor diagnostics (or indeed over-enthusiasm).

Our second contribution is to make more explicit the parallels between the OD and ZI approaches to modelling excess zeros. But we have not resolved the choice between the explicit (ZI) and implicit (OD) approaches, when the base is the Poisson. On the contrary, we see that, from the point of view of the excess zeros, one ZI model—type B—behaves exactly like the NB-lin; and NB-quad is not unlike the (dominant) ZI type C. The difference between the latter

two is only apparent for very large μ, and only carefully designed data will differentiate them. They *do* differ as regard the mean-variance relationship, but only in detail for both exhibit a quadratic relationship; and this too will only be apparent at very large μ. In addition, for iid data, technical details of the usual inference procedures for NB-quad do little to help differentiate the very different models. Further, for the more typical case of regression, these details can only become more challenging. What is clear is that any attempt to discriminate between explicit and implicit ZI models will require a rich and varied dataset and diagnostics focussing jointly on both fitted zero probability and upper tail behaviour. This discussion suggests that the burgeoning literature on other OD generalisations of the Poisson—and on zero-inflating them—may itself be premature. As a final contribution, Table 2 offers an alternative to the confusing nomenclature that has developed.

ACKNOWLEDGEMENTS

Andrew Parnell's work was supported by a Science Foundation Ireland Career Development Award (17/CDA/4695); an investigator award (16/IA/4520); a Marine Research Programme funded by the Irish Government, co-financed by the European Regional Development Fund (Grant-Aid Agreement No. PBA/CC/18/01); European Union's Horizon 2020 research and innovation programme under grant agreement no. 818144; and SFI Research Centre awards 16/RC/3872 and 12/RC/2289_P2. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Benson, A. & Friel, N. (2017). Bayesian inference, model selection and likelihood estimation using fast rejection sampling: The Conway-Maxwell-Poisson distribution. Arxiv, 1709.03471.

Blasco-Moreno, A., Pérez-Casany, M., Puig, P., Morante, M., & Castells, E. (2019). What does a zero mean? Understanding false, random and structural zeros in ecology. *Methods in Ecology and Evolution*, **10**(7), 949–959.

Cameron, A. C. & Trivedi, P. K. (2013). *Regression Analysis of Count Data.* , Second edition. Cambridge University Press.

Consul, P. C. & Jain, G. C. (1973). A generalization of the Poisson distribution. *Technometrics*, **15**(4), 791–799.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*

Del Castillo, J. & Pérez-Casany, M. (1998). Weighted Poisson distributions for overdispersion and underdispersion situations. *Annals of the Institute of Statistical Mathematics*, **50**, 567–585.

Deng, D. & Paul, S. R. (2005). Score tests for zero-inflation and over-dispersion in generalized linear models. *Statistica Sinica*, **15**(1), 257–276.

Dietz, E. & Böhning, D. (2000). On estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics and Data Analysis*, **34**(4), 441–459.

Farewell, V. T., Long, D. L., Tom, B. D., Yiu, S., & Su, L. (2017). Two-part and related regression models for longitudinal data. *Annual Review of Statistics and Its Applications*, **4**, 283–315.

Gurmu, S. & Trivedi, P. K. (1996). Excess zeros in count models for recreational trips. *Journal of Business and Economic Statistics*, **14**(4), 469–477.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics*, **56**(4), 1030–1039.

Hilbe, J. M. & Greene, W. H. (2007). 7 count response regression models. *Handbook of Statistics*, **27**, 210–252.

Jansakul, N. & Hinde, J. P. (2008). Score tests for extra-zero models in zero-inflated negative binomial models. *Communications in Statistics: Simulation and Computation*, **38**(1), 92–108.

Jochmann, M. (2013). What belongs where? Variable selection for zero-inflated count models with an application to the demand for health care. *Computational Statistics*, **28**(5), 1947–1964.

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

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, **15**(3), 209–225.

Long, D. L., Preisser, J. S., Herring, A. H., & Golin, C. E. (2014). A marginalized zero-inflated Poisson regression model with overall exposure effects. *Statistics in Medicine*, **33**(29).

Luyts, M., Molenberghs, G., Verbeke, G., Matthijs, K., Ribeiro Jr, E. E., Demétrio, C. G., & Hinde, J. (2019). A Weibull-count approach for handling under- and overdispersed longitudinal/clustered data structures. *Statistical Modelling*, **19**(5), 569–589.

Martin, J. & Hall, D. B. (2017). Marginal zero-inflated regression models for count data. *Journal of Applied Statistics*, **44**(10), 1807–1826.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., & Possingham, H. P. (2005). Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations. *Ecology Letters*, **8**(11), 1235–1246.

Mcmahon, B. J., Purvis, G., Sheridan, H., Siriwardena, G. M., & Parnell, A. C. (2017). A novel method for quantifying overdispersion in count data and its application to farmland birds. *Ibis*, **159**(2), 406–414.

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

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, **33**(3), 341–365.

Perumean-Chaney, S. E., Morgan, C., McDowall, D., & Aban, I. (2013). Zero-inflated and overdispersed: What's one to do? *Journal of Statistical Computation and Simulation*, **83**(9), 1671–1683.

Plummer, M., Hornik, K., Leisch, F., & Zeileis, A. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proc 3rd Int Work Distrib Stat Comput*.

Pohlmeier, W. & Ulrich, V. (1995). An econometric model of the two-part decisionmaking process in the demand for healthcare. *Journal of Human Resources*, **30**(2), 339–361.

Puig, P. & Valero, J. (2006). Count data distributions: Some characterizations with applications. *Journal of the American Statistical Association*, **101**(473), 332–340.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. *Sankhy□: The Indian Journal of Statistics, Series A*, **27**, 311–324.

Ribeiro Jr, E. E., Zeviani, W. M., Bonat, W. H., Demétrio, C. G. B.,, & Hinde, J. (2020). Reparametrization of COM–Poisson regression models with applications in the analysis of experimental data. *Statistical Modelling*, **20**, 443–466.

Ridout, M., Demetrio, C. G., & Hinde, J. (1998). Models for count data with many zeros. In *Proceedings of the XIXth international biometric conference* (179–192).

Rigby, R. A., Stasinopoulos, D. M., & Lane, P. W. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, **54**(3), 507–554.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **71**(2), 319–392.

Salter-Townshend, M. & Haslett, J. (2012). Fast inversion of a flexible regression model for multivariate pollen counts data. *Environmetrics*, **23**(7), 595–605.

Sellers, K. F. & Shmueli, G. (2010). Predicting censored count data with COM-Poisson regression. *Robert H. Smith School Research Paper No. RHS-06-129*, **28**.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., & Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society: Series C: Applied Statistics*. **54**(1), 127-142

Stan Development Team. (2014). RStan: The R interface to Stan, Version 2.3.

Warton, D. I. (2005). Many zeros does not mean zero inflation: Comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics*, **16**(3), 275–289.

Welsh, A. H., Cunningham, R. B., Donnelly, C. F., & Lindenmayer, D. B. (1996). Modelling the abundance of rare species: Statistical models for counts with extra zeros. *Ecological Modelling*, **88**(1–3), 297–308.

Yee, T. W. (2015). *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer.

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, **27**(8), 1–25.

Zeviani, W. M., Ribeiro, P. J., Bonat, W. H., Shimakura, S. E., & Muniz, J. A. (2014). The Gamma-count distribution in the analysis of experimental underdispersed data. *Journal of Applied Statistics*, **41**(12), 2616–2626.

[*Received October 2020; accepted October 2021*]