One-inflated zero-truncated count regression models

Ryan T. Godwin[*]

University of Manitoba

Summary

We find that in zero-truncated count data $(y = 1, 2, \dots)$, individuals often gain information at first observation, leading to a common but unaddressed phenomenon of "one-inflation". The current standard, the zero-truncated negative binomial (ZTNB) model, is misspecified under one-inflation, causing bias and inconsistency. To address this, we introduce the one-inflated zero-truncated negative binomial (OIZTNB) regression model. The importance of our model is highlighted through simulation studies, and through the discovery of one-inflation in four datasets that have traditionally championed ZTNB. We recommended OIZTNB over ZTNB for most data, and provide estimation, marginal effects, and testing in the accompanying R package `oneinfl`.

Keywords: count data, one-inflation, Poisson, negative binomial, regression model

1   Introduction

The zero-truncated negative binomial (ZTNB) model is the workhorse model for when the outcome variable is a non-zero count $(y = 1, 2, ...)$. Yet, we find that ZTNB is misspecified in nearly every instance, due to one-inflation, which leads to biased and inconsistent estimators for the parameters of the model. It is common to find a preponderance of 1 counts in data, but distributions that allow for one-inflation have only recently been developed. While gaining popularity in biometrics, criminology and epidemiology, one-inflation has not yet extended into economics, partly due to a lack

---

[*]email: ryan.godwin@umanitoba.ca; address: 646 Fletcher Argue Building, Winnipeg, MB, R3T5V5

of an available regression model. To rectify this we develop the one-inflated zero-truncated negative binomial (OIZTNB) and one-inflated positive Poisson (OIPP) regression models. To illustrate their importance, we apply them to four common datasets traditionally used to highlight and exemplify the standard ZTNB, finding the presence of one-inflation in each case. We will argue that one-inflation should be the norm when data is collected administratively, and therefore recommend the estimation of OIZTNB instead of ZTNB, for most truncated data.

Truncated count models are used in several areas of economics. In health economics, individual health care usage is often measured through doctor visits or days in a hospital. Patients do not enter the administrative sample until they make their first visit, so 0 counts are not observed. In tourism economics and recreation demand, the number of times an individual visits a site is often collected on-site making 0 counts impossible. In labour economics, the number of times a worker becomes unemployed or changes jobs is often zero-truncated when the data are collected administratively. Further examples of interest to economists include the number of arrests or incarcerations, the number of patents filed by a firm, or the number of insurance claims.

Truncation occurs if the observational apparatus only becomes active with the occurrence of one event (Johnson and Kotz, 1969). One-inflation occurs if the observational apparatus has an effect on that which is being observed, or on some associated party. It is usually not only the researcher that gains information at the first event. Revelations made at first visit disrupt the probability of a 1 count, but subsequent visits do not garner additional information, and the distortion to the counting distribution is confined to the 1s. A doctor may learn that a patient is healthy and need not return after the first visit. A consumer may learn that they hate the experience and only do it once. A worker may have lower utility than they expected due to unemployment, and learn to avoid subsequent unemployment spells. In situations where information is learned at first visit, but not from subsequent visits, one-inflation (or deflation) is likely to occur.

The recent attention to one-inflated distributions in biology, epidemiology, and criminology, is due to population size estimation. A truncated count model is fit to "capture-recapture" data, in order to infer the number of missing 0s, thus obtaining an estimate for the entire population. One-inflation has been found for many such data. For example, consider the number of times a prostitute is arrested. Rossmo and Routledge (1990) write: "prudent ones seem to learn from the experience, not to avoid street prostitution, but to avoid being arrested... This would lead to a relatively large number of single arrests." A portion of individuals gain a desire or ability to avoid recapture; this portion is the amount of one-inflation in the data. The recognition of

the one-inflating phenomenon has led to the development of several one-inflated distributions; the one-inflated positive Poisson (Godwin and Böhning, 2017), the one-inflated zero-truncated negative binomial (Godwin, 2017), the one-inflated positive Poisson mixture model (Godwin, 2019), the one-inflated positive Poisson–Lindley (Tajuddin et al., 2022), and others. To our knowledge, none of these distributions have been extended to a regression framework, until now.

The development of regression models for one-inflated data entails several mathematical contributions. A novel link function for the one-inflation parameter is proposed. Marginal effects are derived. Log-likelihoods are determined. Various tests, including for no one-inflation, are presented. In addition, simulation studies verify the consistency of the proposed estimators, as well as illustrate the bias and inconsistency of the standard ZTNB estimator arising from ignoring one-inflation. A suite of tools to aid in the estimation, specification, and interpretation of the models are developed, and presented in the R package `oneinfl`. The importance of these contributions is highlighted through the finding of one-inflation in four applications where ZTNB has been the standard.

2   One-inflated distributions

This section reviews some of the one-inflated distributions that have been developed to date. We start with the Poisson distribution, truncate it, and allow for one-inflation, leading to the one-inflated positive Poisson (OIPP) distribution. Allowing for unobserved heterogeneity then leads to the one-inflated zero-truncated negative binomial (OIZTNB) distribution. In the next section, we introduce regressors and build the regression models.

2.1   Poisson distribution

The Poisson distribution is often the starting point for the analysis of count data, due to its relationship to more complicated but more suitable count distributions. The mass function for a random variable, y, that follows a Poisson distribution, is:

$$f^P = \frac{\lambda^y}{\exp(\lambda)y!} \quad ; \quad y = 0, 1, \ldots \tag{1}$$

where $\lambda > 0$ is the mean and variance of y over the period of observation. The key limiting assumption of the Poisson model is homogeneity; the mean intensity of counts over the period of observation should be the same for all individuals in the sample. In economics, violation of this assumption is usually dealt with by using the negative binomial model instead.

2.2 Departure from Poisson

Before we develop the regression model, we discuss three main issues that cause the basic Poisson distribution to require extension. (i) Zero-truncation, the impossibility of observing a 0 count due to the administrative nature of the data. (ii) Unobserved heterogeneity, where the regressors can not explain enough of the variation in counts in order to satisfy the Poisson assumption of homogeneity. (iii) One-inflation (or deflation), a phenomenon where there is an excess (or a lack of) 1 counts.

Consideration of these issues will lead to the OIPP distribution (Godwin and Böhning, 2017) and the OIZTNB distribution (Godwin, 2017). After these models have been discussed, regressors are introduced and a regression model is developed.

2.3 Truncation

Only those who "come through the door" enter the sample, and the counts begin at 1. Such is the case with many administrative data. In general, count distributions can be altered to accommodate zero-truncation simply by recognizing that the support of the distribution is changed (y cannot equal 0), and by normalizing so that the probabilities of allowable counts sum to 1:

$$f^{ZT} = \frac{f}{1 - f(0)} \quad ; \quad y = 1, 2, \dots \tag{2}$$

where f^{ZT} is the zero-truncated count distribution, f is the untruncated distribution, and $f(0)$ is the probability of a 0 under the untruncated count distribution. The zero-truncated, or positive Poisson (PP) distribution is then:

$$f^{PP} = \frac{\lambda^y}{(\exp(\lambda) - 1)y!} \quad ; \quad y = 1, 2, \dots \tag{3}$$

where λ is again the sole "rate" parameter in the distribution.

2.4 Heterogeneity and negative binomial

For the PP distribution in equation 3 to be suitable, the Poisson assumptions of independence and homogeneity are required. Homogeneity is often too restrictive. Violation of this assumption is often detected through overdispersion of the data. A potential solution is to use a regression model, since the rigid Poisson assumption of homogeneity is more easily satisfied by allowing the Poisson parameter λ to be a function of regressors. Often, however, there is still unobserved heterogeneity, and a richer model is required.

The negative binomial distribution is one such distribution capable of dealing with unobserved heterogeneity, and may be generated several ways. For example, one way is to include a gamma distributed error term in the function that links regressors to λ. The mass function for a random variable y that follows the negative binomial (NB) distribution is:

$$f^{NB} = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\Gamma(y+1)} \left(\frac{1}{1+\theta}\right)^{\alpha} \left(\frac{\theta}{1+\theta}\right)^{y} \quad ; \quad y = 0, 1, \ldots \tag{4}$$

Using equation (2), the mass function for a random variable y that follows the zero-truncated negative binomial (ZTNB) distribution is:

$$f^{ZTNB} = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\Gamma(y+1)} \left(\frac{1}{1+\theta}\right)^{\alpha} \left(\frac{\theta}{1+\theta}\right)^{y} \left(\frac{1}{1-(1+\theta)^{-\alpha}}\right) \quad ; \quad y = 1, 2, \ldots \tag{5}$$

2.5 One-inflated count distributions

In general, a count distribution may be 1-inflated by introducing an additional parameter ω, which allows for an extra (or decreased) probability of a 1-count occurring, relative to an underlying count distribution:

$$\begin{aligned} f^{OI} &= \omega + (1-\omega)f(1) \quad ; \quad y = 1 \\ &= (1-\omega)f(y) \quad ; \quad y = 2, 3, \ldots \end{aligned} \tag{6}$$

where f^{OI} is the one-inflated distribution, f is the underlying count distribution, and $f(1)$ and $f(y)$ are probabilities of specific counts occurring. The distribution in equation 6 is conceptually similar to the zero-inflated Poisson distribution (Johnson and Kotz, 1969; Lambert, 1992).

2.5.1 One-inflated positive Poisson

The one-inflated positive Poisson (OIPP) distribution (Godwin and Böhning, 2017) is designed to allow for one-inflation. Setting $f = f^{PP}$ in equation 6 gives the mass function for the OIPP distribution:

$$\begin{aligned} f^{OIPP} &= \omega + (1-\omega)\frac{\lambda}{\exp(\lambda) - 1} \quad ; \quad y = 1 \\ &= (1-\omega)\frac{\lambda^{y}}{(\exp(\lambda) - 1)y!} \quad ; \quad y = 2, 3, \ldots \end{aligned} \tag{7}$$

The OIPP distribution (7) deviates from the PP distribution via the parameter ω, and collapses to the PP distribution for $\omega = 0$. The λ parameter has the same interpretation in both PP and OIPP distributions.

The OIPP distribution is to the PP distribution as the zero-inflated Poisson (ZIP) distribution (Johnson and Kotz, 1969) is to the Poisson distribution. ω adds to the probability of a 1 count (ω may be negative), and the probability of all counts due to the Poisson process is appropriately decreased by the scaling factor $(1 - \omega)$ so that f^{OIPP} remains a proper mass function. In order for the probability of a 1 count to remain positive, $-\lambda(e^{\lambda} - \lambda - 1)^{-1} < \omega < 1$, which will be ensured in a link function for ω in a following section.

2.5.2 One-inflated zero-truncated negative binomial

As discussed in a previous section on heterogeneity, the PP and OIPP distribution may not be a good choice due to their restricitive assumption of homogeneity. If there is unobserved heterogeneity then negative binomial is more appropriate than Poisson, regardless of truncation or one-inflation. The one-inflated zero-truncated negative binomial (OIZTNB) distribution (Godwin, 2017) can be obtained by setting $f = f^{ZTNB}$ in equation 6, and has mass function:

$$
\begin{aligned}
f^{OIZTNB} &= \omega + (1 - \omega)\alpha \left(\frac{1}{1+\theta}\right)^{\alpha} \left(\frac{\theta}{1+\theta-(1+\theta)^{1-\alpha}}\right) \ ; \ y = 1 \\
&= (1 - \omega)\frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\Gamma(y+1)} \left(\frac{1}{1+\theta}\right)^{\alpha} \left(\frac{\theta}{1+\theta}\right)^{y} \left(\frac{1}{1-(1+\theta)^{-\alpha}}\right) \ ; \ y = 2,3,\ldots
\end{aligned}
\tag{8}
$$

3 One-inflated count regression models

This section presents the main contribution of the paper: the development of the one-inflated positive Poisson (OIPP) and one-inflated zero-truncated negative binomial (OIZTNB) regression models. This is accomplished by allowing the parameters of the distributions to be a function of regressors.

3.1 Canonical link for rate parameters

In order for individual characteristics or policy interventions to influence outcomes in count data models, it is customary to link the rate parameter (denoted λ) to the regressors via the canonical link:

$$
\lambda_i = \exp(\boldsymbol{X}_i \boldsymbol{\beta}),
\tag{9}
$$

and instead estimate the parameters $\boldsymbol{\beta}$ in the link function. We follow this convention. To permit individual heterogeneity, the rate parameter has been denoted λ_i in equation 9 (instead of λ), where

the subscript i denotes an individual observation, where $i = 1, \ldots, n$, and where n is the sample size. Equation 9 can be substituted directly into the mass function for the OIPP distribution (equation 7), but for the OIZTNB distribution (equation 8) an additional reparameterization is needed:

$$\theta = \frac{\lambda_i}{\alpha} \tag{10}$$

Substituting the reparameterization (equation 10) and the canonical link (equation 9) into the OIZTNB mass function (equation 8) gives rise to a NegBin-II type regression model (the more popular among the NegBin-I and NegBin-k type models).

3.2   A novel link function for inflation parameter ω_i

The inflation parameter shall now be denoted ω_i with subscript i to highlight that the model allows for individual heterogeneity, and is linked to a (possibly identical to $\boldsymbol{X}_i$) set of regressors $\boldsymbol{Z}_i$. The link function for ω_i requires care and development, and does not follow any convention that we are aware of.

It is desirable for the regression models to allow for one-*deflation*, as well as one-inflation. One-deflation (in a truncated model) is likely more empirically common than zero-deflation (in an un-truncated model) because of the nature of the phenomenon. While the two are conceptually and mathematically similar, zero-inflation typically occurs because "non-users" (always 0s) are mixed into the population of "users". 1-altered counts in administrative data, however, occur from a behavioural response due to the experience of being observed for the first time. The effect can be positive or negative; either the observational experience is a deterrent or an encouragement, and may vary by individual or context.

The zero-inflated Poisson model (Lambert, 1992) provides some guidance, where a logistic-like reparameterization is used: $\omega_i = \phi_i/(1+\phi_i)$, and regressors are linked linearly by $\phi_i = \boldsymbol{Z}_i\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ is an additional vector of parameters to be estimated. This ensures $\omega_i \in [0, 1]$ for $\boldsymbol{Z}_i\boldsymbol{\gamma} \in (-\infty, \infty)$. If the same link were used for the one-inflated models, then $f^{OI}(1) \in [f(1), 1]$[1]. This would be acceptable if 1-deflation (a lack of 1s) is ruled out a priori, but is not an appropriate restriction for the reasons outlined above. Instead, the ω_i parameter should be bound such that the probability of a 1-count lies between 0 and 1:

$$f^{OI}(1) \in [0, 1] \tag{11}$$

---

[1] f^{OI} denotes the one-inflated distribution, and f denotes the underlying un-altered count distribution, as in equation 6.

That is, the inflation parameter should be able to completely alter the probability of a 1-count to any allowable value. Equation 11 can be ensured by binding ω_i appropriately:

$$\omega_i \in \left[\frac{-f(1)}{1 - f(1)}, 1 \right] \tag{12}$$

where the domain of ω_i is found by setting $f^{OI}(1) = 0$ and $f^{OI}(1) = 1$ in equation 6. Finally, the generalized logistic function is used to link regressors to ω_i:

$$\omega_i = L_i + \frac{1 - L_i}{1 + \exp(-\mathbf{Z}_i\boldsymbol{\gamma})}, \tag{13}$$

where L_i is the lower bound from equation 12: $L_i = \frac{-f(1)}{1-f(1)}$, $\mathbf{Z}_i$ is the vector of regressors effecting one-inflation, and where $\boldsymbol{\gamma}$ is a parameter vector to be estimated. The link function in equation 13 ensures that the binds on ω_i in equation 12 are satisfied, and that $f^{OI}(y) \geq 0$ and $\sum f^{OI} = 1 \quad \forall \quad y = 1, 2, \ldots$ (provided that f is a proper mass function).

For the OIPP regression model, the lower bound L_i takes on the form:

$$L_i = \frac{-\lambda_i}{\exp(\lambda_i) - \lambda_i - 1} \tag{14}$$

and for the OIZTNB regression model the lower bound L_i takes the form:

$$L_i = -\left\{ \left(\frac{\alpha}{\alpha + \lambda_i} \right)^{-\alpha} \frac{1}{\lambda_i} \left[1 + \frac{\lambda_i}{\alpha} - \left(1 + \frac{\lambda_i}{\alpha} \right)^{1-\alpha} \right] - 1 \right\}^{-1} \tag{15}$$

Since $\lambda_i = \exp(\mathbf{X}_i\boldsymbol{\beta})$, these lower bounds L_i depend on the data and vary between individuals. This must be taken into account when determining marginal effects, and when estimating the parameters of the regression models.

In summary, the OIPP regression model is obtained by substituting the canonical link for λ_i (equation 9), the logistic link for ω_i, and the lower bound L_i (equation 14), into the OIPP mass function (equation 7). The OIZTNB regression model is similarly achieved, except the lower bound for L_i in equation 15 must instead be used.

4 Marginal effects

Of particular importance are the marginal effects that individual characteristics or policies in X and Z have on the expected number of counts $\mathbb{E}[y_i]$. Obtaining the estimated marginal effects is a standard goal of regression modelling. The following derivatives need to be determined:

$$\frac{\partial \mathbb{E}[y_i]}{\partial q_{ij}},$$

8

where q_{ij} is the jth regressor in X and/or Z, for individual i. These partial derivatives can then be evaluated at the MLEs and at values for $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$. Traditionally, in non-linear models, the MLEs of the marginal effects are evaluated at either the sample means of the data, or at all data points and then averaged. The latter is the default for the `margins` command in Stata, and we have followed this convention for the `margins` function in the `oneinfl` package, although either method can be selected.

For binary dummy variables, the "marginal effect" is instead $\mathbb{E}[y_i \mid D_{ij} = 1] - \mathbb{E}[y_i \mid D_{ij} = 0]$, for some dummy in X and/or Z. This is straightforward to estimate by evaluating the expected difference at several data points.

Model parameters from positive Poisson or zero-truncated negative binomial models are incident risk ratios when exponentiated, offering an interpretation of the impact of a regressor on the outcome variable. However, this is not possible for the OIPP and OIZTNB models, since each regressor typically enters the model through a mean link (equation 9) *and* through a one-inflation link (equation 13). Marginal effects offer a valid way to compare the impact of regressors between ZTNB and OIZTNB, and are more intuitive to economists.

4.1 Marginal effects for the OIPP model

The mean of the OIPP distribution is:

$$\mathbb{E}[y_i^{OIPP}] = \omega_i + (1 - \omega_i)\frac{\lambda_i \exp(\lambda_i)}{\exp(\lambda_i) - 1} \tag{16}$$

The first derivative of equation 16 with respect to a regressor q_{ij} in X and/or Z is:

$$\frac{\partial \mathbb{E}[y_i^{OIPP}]}{\partial q_{ij}} = \frac{\partial \omega_i}{\partial q_{ij}} \left[1 - \frac{\lambda_i \exp(\lambda_i)}{\exp(\lambda_i) - 1} \right] + \frac{\partial \lambda_i}{\partial q_{ij}} (1 - \omega_i) \frac{\exp(\lambda_i) \left[\exp(\lambda_i) - \lambda_i - 1 \right]}{\{\exp(\lambda_i) - 1\}^2}, \tag{17}$$

where

$$\frac{\partial \omega_i}{\partial q_{ij}} = \frac{\partial \lambda_i}{\partial q_{ij}} \left\{ \frac{\exp(\lambda_i) - \lambda_i \exp(\lambda_i) - 1}{[\exp(\lambda_i) - \lambda_i - 1]^2} \right\} \left[\frac{-\exp(-\boldsymbol{Z}_i \gamma)}{1 + \exp(-\boldsymbol{Z}_i \gamma)} \right]$$
$$- \frac{\partial \exp(-\boldsymbol{Z}_i \boldsymbol{\gamma})}{\partial q_{ij}} \left\{ \frac{\exp(\lambda_i) - 1}{[\exp(\lambda_i) - \lambda_i - 1][1 + \exp(-\boldsymbol{Z}_i \gamma)]^2} \right\},$$

$$\lambda_i = \exp\left(\boldsymbol{X}_i \boldsymbol{\beta}\right),$$

$$\frac{\partial \lambda_i}{\partial q_{ij}} = \begin{cases} \lambda_i \beta_j, & \text{if } q_{ij} \text{ is the } j^{th} \text{ column in } \boldsymbol{X}_i \\ 0, & \text{otherwise} \end{cases},$$

$$\frac{\partial \exp(-\boldsymbol{Z}_i \boldsymbol{\gamma})}{\partial q_{ij}} = \begin{cases} -\exp(-\boldsymbol{Z}_i \boldsymbol{\gamma}) \gamma_j, & \text{if } q_{ij} \text{ is the } j^{th} \text{ column in } \boldsymbol{Z}_i \\ 0, & \text{otherwise} \end{cases}.$$

As in most non-linear models, the marginal effects themselves are highly non-linear, and are functions of the values of the regressors themselves. That is, the marginal effects vary by individual. When the impact of a regressor is interpreted, a decision will need to be made regarding values for $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ in the evaluation of $\partial y_i / \partial q_{ij}$.

4.2   Marginal effects for the OIZTNB model

The mean of the OIZTNB distribution is:

$$\mathbb{E}[y_i^{OIZTNB}] = \omega_i + (1 - \omega_i) \left[\frac{\lambda_i}{1 - \left(1 + \frac{\lambda_i}{\alpha}\right)^{-\alpha}} \right] \tag{18}$$

The first partial derivative of equation 18 with respect to a regressor q_{ij} in X and/or Z is:

$$\frac{\partial \mathbb{E}[y_i^{OIZTNB}]}{\partial q_{ij}} = \frac{\partial \omega_i}{\partial q_{ij}} \left[1 - \frac{\lambda_i}{1 - \left(1 + \frac{\lambda_i}{\alpha}\right)^{-\alpha}} \right]$$
$$+ \frac{\partial \lambda_i}{\partial q_{ij}} \frac{(1 - \omega_i)}{1 - \left(1 + \frac{\lambda_i}{\alpha}\right)^{-\alpha}} \left[1 - \frac{\lambda_i \left(1 + \frac{\lambda_i}{\alpha}\right)^{-\alpha - 1}}{1 - \left(1 + \frac{\lambda_i}{\alpha}\right)^{-\alpha}} \right], \tag{19}$$

where

$$\frac{\partial \omega_i}{\partial q_{ij}} = \frac{\partial L}{\partial q_{ij}} \left[1 - \frac{1}{1 + \exp(-\boldsymbol{Z}_i \boldsymbol{\gamma})} \right] - \frac{\partial \exp(-\boldsymbol{Z}_i \boldsymbol{\gamma})}{\partial q_{ij}} \left\{ \frac{(1 - L)}{[1 + \exp(-\boldsymbol{Z}_i \boldsymbol{\gamma})]^2} \right\},$$

$$\frac{\partial L}{\partial q_{ij}} = -\frac{\partial f_1}{\partial q_{ij}} \left[\frac{1}{(1 - f_1)^2} \right],$$

$$\frac{\partial f_1}{\partial q_{ij}} = \frac{\partial \lambda_i}{\partial q_{ij}} \left(1 + \frac{\lambda_i}{\alpha} \right)^{-\alpha} \left[1 + \frac{\lambda_i}{\alpha} - \left(1 + \frac{\lambda_i}{\alpha} \right)^{1-\alpha} \right]^{-1}$$

$$\left\{ 1 - \frac{\alpha \lambda_i}{\alpha + \lambda_i} - \frac{1}{\alpha} \left[1 + \frac{\lambda_i}{\alpha} - \left(1 + \frac{\lambda_i}{\alpha} \right)^{1-\alpha} \right]^{-1} \left[1 - (1 - \alpha) \left(1 + \frac{\lambda_i}{\alpha} \right)^{-\alpha} \right] \right\},$$

and where λ_i and $\frac{\partial \lambda_i}{\partial q_{ij}}$ are as defined previously, and where f_1 is the probability of a 1 count according to equation 8.

5   Estimation and testing

Various issues relating to the estimation of the models are discussed in this section: maximization of the log-likelihoods; estimation of variance-covariance matrices and their related uses such as tests for no one-inflation and tests of significance; estimation of the marginal effects and their standard errors; and the visual assessment of the fitted models via barplots.

5.1   Log-likelihoods

Estimation of the one-inflated regression models are accomplished via maximum likelihood. The `optim` function in base R (R Core Team et al., 2013) is used to maximize log-likelihoods. The `oneinfl` package uses the `BFGS` algorithm as default method of optimization, but other methods may be passed to `optim`. `oneinfl` attempts to find safe starting values, having no difficulties during simulations, but nonetheless starting values may be passed to `optim` as well.

5.1.1   OIPP model

From equation 7, the log-likelihood for the one-inflated positive Poisson (OIPP) model is:

$$\ell = \sum_{i=1}^{n} \left\{ \log(1 - \omega_i) + I_1 \log \left[\frac{\omega_i}{1 - \omega_i} + \frac{\lambda_i}{\exp \lambda_i - 1} \right] \right.$$
$$\left. + (1 - I_1) \left[y_i \log \lambda_i - \log \left[\exp \lambda_i - 1 \right] - \log y_i! \right] \right\} \tag{20}$$

where $I_1 = 1$ if $y_i = 1$ and 0 otherwise, and where λ_i and ω_i have been previously defined.

5.1.2 OIZTNB model

From equation 8, a log-likelihood function for the OIZTNB model is:

$$\ell = \sum_{i=1}^{n} \left\{ \log(1 - \omega_i) + I_1 \log \left[\left(\frac{\omega_i}{1 - \omega_i} + \alpha \left(\frac{\alpha}{\alpha + \lambda_i} \right)^{\alpha} \lambda_i \right) \Big/ \left(\alpha + \lambda_i - \alpha \left(1 + \frac{\lambda_i}{\alpha} \right)^{1-\alpha} \right) \right] \right.$$

$$+ (1 - I_1) \left[\sum_{j=1}^{y_i} \log(\alpha + j - 1) - \log y_i! + \alpha \log \alpha + y_i \log \lambda_i - (\alpha + y_i) \log(\alpha + \lambda_i) \right. \qquad (21)$$

$$\left. \left. - \log \left(1 - \left(1 + \frac{\lambda_i}{\alpha} \right)^{-\alpha} \right) \right] \right\}$$

where I_1, λ_i, and ω_i are defined previously. The recursive property of the gamma function has been invoked in order to eliminate it: $\log \Gamma(\alpha + k) - \log \Gamma(\alpha) = \sum_{j=1}^{k} \log(\alpha + j - 1)$.

In both OIPP and OIZTNB models, $\log y_i!$ appears. These terms could be removed for faster optimization, but for the calculation of likelihood ratio test (LRT) statistics. R is unable to calculate $\log y_i!$ for $y_i > 170$, so in `oneinfl` we use Stirling's approximation ($\log y_i! \approx y_i \log y_i - y_i$) for values that are large.

5.2 Variance-covariance matrix

The estimated variance-covariance matrix for MLEs is the negative of the expected inverse Hessian matrix, evaluated at the MLEs:

$$\hat{\mathbb{V}}(\hat{\boldsymbol{\theta}}) = -E \left[\frac{\partial^2 \ell(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}} \partial \hat{\boldsymbol{\theta}}'} \right]^{-1} \qquad (22)$$

The estimated standard errors are the square roots of the diagonal elements of this matrix. The Hessian matrix is not derived analytically from the log-likelihoods in equations 20 or 21. Rather, the Hessian matrix is estimated numerically by `optim`. The $\hat{\mathbb{V}}(\hat{\boldsymbol{\theta}})$ matrix is used to report estimated standard errors in the `summary.oneinfl` function, for various tests, and for the estimated standard errors of marginal effects.

5.3 Tests of significance

The `summary.oneinfl` function extracts and reports the square roots of the diagonal elements from the estimated variance-covariance matrix in equation 22, and provides p-values for tests of $H_0 : \beta_j = 0$ and for $H_0 : \gamma_j = 0$. However, this does not typically amount to a standard test of "significance" in the OIPP and OIZTNB models.

The same regressors are likely to appear in both the X and Z matrices, allowing them to influence both the mean of the distribution through λ_i, and the amount of one-inflation through ω_i. Hence, tests of significance of say x_j are then joint hypotheses for when $x_j = z_m$, for example: $H_0 : \beta_j = 0$ and $\gamma_m = 0$. Such null hypotheses can be tested via a Wald statistic, and are accomplished via the `signifWald` function in the `oneinfl` package. This function uses the estimated variance-covariance matrix in equation 22, and returns the Wald statistic and associated p-value.

5.4   Tests for no one-inflation

The amount of one-inflation can be estimated by $\frac{\sum \hat{\omega}_i}{n}$ (average one-inflation) or $\frac{\sum |\hat{\omega}_i|}{n}$ (average absolute one-inflation). Both are reported in the `summary.oneinfl` function. Formal tests for the presence of one-inflation are also developed. Recalling equation 13, a test for no one-inflation amounts to $H_0 : \boldsymbol{\gamma} = \mathbf{0}$, rejection of which would indicate that the one-inflated model is more appropriate. Alternatively, the null and alternative hypotheses could be specified as, for example, $H_0 : \text{ZTNB}$ and $H_A : \text{OIZTNB}$. The null hypothesis is nested in the alternative, allowing for the asymptotically equivalent Wald, likelihood ratio, and lagrange multiplier testing. Only the Wald and likelihood ratio tests are developed here.

5.4.1   Wald test

The Wald test statistic does not require estimation of the model under the null hypothesis, and is calculated by:

$$W = \boldsymbol{\gamma}' \left[R\hat{\mathbb{V}}(\hat{\boldsymbol{\theta}})R' \right]^{-1} \boldsymbol{\gamma} \tag{23}$$

where the R matrix selects the necessary rows and columns from $\hat{\mathbb{V}}(\hat{\theta})$. This Wald statistic is $\chi^2_{(p)}$ distributed, where p are the number of regressors in Z including the intercept. Note that the null hypothesis is not on the boundary of the parameter space as one-deflation is allowed in the models through the choice of the link function in equations 11 and 12, otherwise the Wald statistic would follow a half Chi-square distribution. The Wald test statistic and associated p-value for no one-inflation are calculated by the `oneWald` function in the `oneinfl` package.

5.4.2   Likelihood ratio test

The likelihood ratio test statistic for testing the null of no one-inflation can be obtained by comparing the log-likelihood functions (evaluated at the MLEs) of OIPP and PP, or OIZTNB and ZTNB. The

statistic takes the form $LRT = -2(\hat{\ell}^{PP} - \hat{\ell}^{OIPP})$ or $LRT = -2(\hat{\ell}^{ZTNB} - \hat{\ell}^{OIZTNB})$ and is $\chi^2_{(p)}$, where p is again the rows in $\boldsymbol{\gamma}$ (also the columns of Z). The `oneLRT` function in the `oneinfl` package provides this test statistic and associated p-value by evaluating both the one-inflated and standard models.

5.5 Estimated marginal effects and their standard errors

The marginal effects in equations 17 and 19 are evaluated at the MLEs, and either averaged over all data points ("average effects"), evaluated once at the sample means of the data ("effect at means"), or at some representative case $\boldsymbol{X}_i, \boldsymbol{Z}_i$. The `margins` function in the `oneinfl` package calculates these marginal effects (defaulting to the "average effects"), and also reports the estimated standard errors for the marginal effects.

In order to determine whether the estimated marginal effects are significant, their standard errors must be estimated. There are two widely popular options available; the delta method and the bootstrap. We employ the delta method since serious computational time can be required in order to estimate some one-inflated models.

The delta method can provide the variance of a function $h(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ has a known asymptotic distribution (Normal in the case of MLEs) with variance-covariance matrix $V(\hat{\boldsymbol{\theta}})$. Justified by a first order Taylor series approximation, the asymptotic variance of $h(\hat{\boldsymbol{\theta}})$ is approximately equal to $J(\hat{\boldsymbol{\theta}})\hat{\mathbb{V}}(\hat{\theta})J(\hat{\boldsymbol{\theta}})'$. The Jacobian matrix $J(\hat{\boldsymbol{\theta}})$ is the vector of first partial derivatives of $h(\hat{\boldsymbol{\theta}})$. Standard errors for the marginal effects are estimated by setting $h(\hat{\boldsymbol{\theta}})$ equal to the marginal effects derived in equations 17 or 19, numerically approximating $J(\hat{\boldsymbol{\theta}})$ using the `numericDeriv` function in base R, and taking the square roots of the diagonal elements of $J(\hat{\boldsymbol{\theta}})\hat{\mathbb{V}}(\hat{\theta})J(\hat{\boldsymbol{\theta}})'$. These standard errors, as well as z-tests for the significance of the marginal effects, are reported in the `margins` function.

5.6 Predicted counts and visualizing the fit of the models

A model's predicted counts are obtained by evaluating its probability mass function (such as equation 8) at each $\hat{\lambda}_i$ (and $\hat{\omega}_i$ if relevant), and summing. This is facilitated in the `oneinfl` package using the `pred` function.

Visualizing the data and predicted counts is an important diagnostic tool, as often the presence of one-inflation is immediately apparent. The `oneplot` function in the `oneinfl` package generates a barplot of the actual data, as well as plotting the predicted counts for up to four models: positive

14

Poisson (PP), one-inflated positive Poisson (OIPP), zero truncated negative binomial (ZTNB), and one-inflated zero truncated negative binomial (OIZTNB). The `oneinfl` package facilitates estimation of the PP and ZTNB models partly for this purpose. Figures 1 - 4 show the result of applying `oneplot` to ZTNB and OIZTNB models.

6 Simulation studies

Simulation studies are undertaken to examine the percent bias of various maximum likelihood estimators (MLEs) under different data generating processes (DGPs) and for differently specified models. The simulations suggest massive bias of the MLEs for when one-inflation is present, but is ignored. That is, the MLEs from the standard positive Poisson (PP) or zero-truncated negative binomial (ZTNB) exhibit large bias when the true data generating process is characteristic of one-inflation. The bias of the estimators depends on the values of the parameters chosen in the DGP, but it is not difficult to find values for which there is 100% bias.

We conduct two main simulations, one each for a one-inflated positive Poisson (OIPP) DGP and for a one-inflated zero-truncated negative binomial (OIZTNB) DGP. Under each DGP, both the standard and one-inflated models are estimated. The same two regressors are used in the X and Z matrices; x_1 is generated according to a $N(10, 1)$ distribution and x_2 is a dummy variable with mean 0.5. The parameter values under OIPP are chosen to be $\boldsymbol{\beta} = (-2, 0.4, 0.2)$, $\boldsymbol{\gamma} = (-21, 2, 0.5)$. The same parameter values are chosen for OIZTNB, but with the addition of $\alpha = 10$. The simulations are run for various sample sizes, and all experiments use 10,000 replications. Random variates for OIPP and OIZTNB were generated using the `roipp` and `roiztnb` functions in `oneinfl`. The results of the experiments are reported in Table 1.

Table 1 shows very high percent biases for PP and ZTNB (between 75% and 325% bias), that do not diminish as n increases, reflecting the inconsistency of the estimators in the presence of one-inflation. In contrast, the percent biases of OIPP and OIZTNB are always less than 4%, and diminish further as n increases. Supposing that these results reflect the consistency of a correctly specified MLE, they should provide some comfort to the reader that the many mathematical expressions required to specify the log-likelihoods of the OIPP and OIZTNB regression models, and the functions coded in the `oneinfl` package, are indeed correct.

Table 1: Percent biases of the MLEs from various models, under a one-inflated DGP (either OIPP or OIZTNB).

model	n	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\alpha}$
				Poisson				
OIPP	200	0.17	0.05	-0.24	3.24	3.28	3.02	
PP		-258.08	-141.20	-75.62				
OIPP	400	-0.18	-0.12	0.00	1.29	1.29	1.52	
PP		-257.03	-140.63	-74.45				
OIPP	800	-0.01	-0.02	0.22	0.69	0.69	0.73	
PP		-256.58	-140.36	-74.43				
OIPP	1600	-0.19	-0.11	0.18	0.34	0.34	-0.08	
PP		-256.46	-140.29	-74.04				
				Negative Binomial				
OIZTNB	200	-0.10	-0.10	-0.67	3.22	3.25	2.99	27.77
ZTNB		-331.59	-194.82	-89.34				-96.02
OIZTNB	400	-0.38	-0.23	-0.08	1.27	1.28	1.50	11.26
ZTNB		-325.28	-190.76	-85.91				-96.20
OIZTNB	800	-0.08	-0.06	0.22	0.69	0.69	0.72	5.42
ZTNB		-321.50	-188.46	-85.61				-96.28
OIZTNB	1600	-0.11	-0.07	0.27	0.32	0.33	-0.14	2.44
ZTNB		-320.42	-187.74	-84.64				-96.32

7 Empirical illustrations

Four empirical illustration are examined. All data have traditionally been used to exhibit the zero-truncated negative binomial (ZTNB) model; in all cases one-inflation is found and the one-inflated zero-truncated negative binomial (OIZTNB) model is superior. The illustrations highlight how the estimated parameters and interpretations of the models change, sometimes dramatically, with the incorporation of one-inflation.

7.1 Hospital length of stay: MedPar data

A particular data set from the 1991 Arizona MedPar database has championed and exemplified the zero-truncated negative binomial (ZTNB) regression model. The data appears in articles (Hardin and Hilbe, 2007, 2015), books (Hilbe, 2011, 2014), R packages (Hilbe and Robinson, 2016; Hilbe et al., 2018), and several web vignettes for ZTNB estimation. ZTNB regression models have been estimated for subsequent waves of MedPar data in many research articles (Baser et al., 2010; Kleindorfer et al., 2009; Lasater et al., 2021), to name a few.

Figure 1: The one-inflated zero truncated negative binomial (OIZTNB) regression model, and the zero truncated negative binomial (ZTNB) regression model, fit to the MedPar data set.
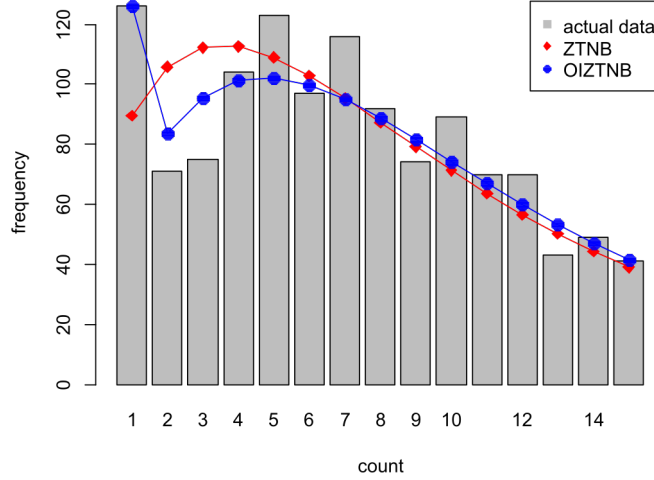

Table 2: ZTNB and OIZTNB models fit to the MedPar data.

	ZTNB	OIZTNB		ZTNB	OIZTNB
	$\hat{\beta}$	$\hat{\beta}$	$\hat{\gamma}$	marginal effects	
(Intercept)	2.333***	2.299***	-4.200***		
	(0.075)	(0.072)	(0.506)		
white	-0.132*	-0.097	0.659	-1.296*	-1.258*
	(0.075)	(0.071)	(0.481)	(0.776)	(0.734)
died	-0.251***	-0.068	2.335***	-2.238***	-2.189***
	(0.045)	(0.045)	(0.236)	(0.387)	(0.396)
type2	0.260***	0.234***	-0.541*	2.639***	2.575***
	(0.055)	(0.054)	(0.283)	(0.616)	(0.588)
type3	0.769***	0.756***	-0.751*	10.159***	10.142***
	(0.083)	(0.079)	(0.447)	(1.530)	(1.467)

Significance at the 1% (***), 5% (**), and 10% (*) levels.

The goal is to model a patient's hospital length of stay (los), which is a zero-truncated count. Some illustrations of the ZTNB model use membership to a health maintenance organization as an explanatory variable, but we instead estimate the model from "11.1: Zero-truncated count models", in Hilbe (2011). All of the explanatory variables are dummies: $white = 1$ if the patient identifies as Caucasian; $died = 1$ if the patient died during their stay; and categorical dummies for the type of admission, whether elective, urgent ($type2 = 1$) or emergency ($type3 = 1$).

One-inflation is plausible in such data, for at least two reasons. Suppose there are two types of individuals mixed into the sample; ill and healthy. Type is revealed through examination by a doctor, when $los = 1$. Those that are healthy are discharged and one-inflate the data. As a second channel for one-inflation, suppose that patients that die tend to do so on their first day of stay.

Indeed, both Wald (p-value 0.000) and LRT (p-value 0.000) tests of no one-inflation support the presence of one-inflation, which is 4.2% on average, and is 6.8% in average absolute one-inflation. Despite its association with the MedPar data set, the ZTNB regression model is likely misspecified, and the estimators are biased and inconsistent.

The estimated ZTNB and OIZTNB models, and the marginal effects from each, are displayed in Table 2. A plot of the actual data, as well as the predicted counts from the ZTNB and OIZTNB models, are displayed in Figure 1, where one-inflation is visibly apparent. All estimates and standard errors for the parameters and marginal effects change somewhat. Most notably is for that of the regressor *died*, whose association to *los* is linked to one-inflation rather than the mean of the underlying count distribution. In addition, under OIZTNB, *white* is no longer a significant variable according to the Wald test from Section 5.4.1 (p-value 0.16).

All estimation and results were generated from the `oneinfl` package. A vignette found at rtgodwin.com/oneinfl/ can reproduce all of the results in this section, and illustrates use of the package.

7.2   Hospital length of stay: Arizona Medicare data

Figure 2: OIZTNB and ZTNB regression models fit to the Arizona Medicare data set.
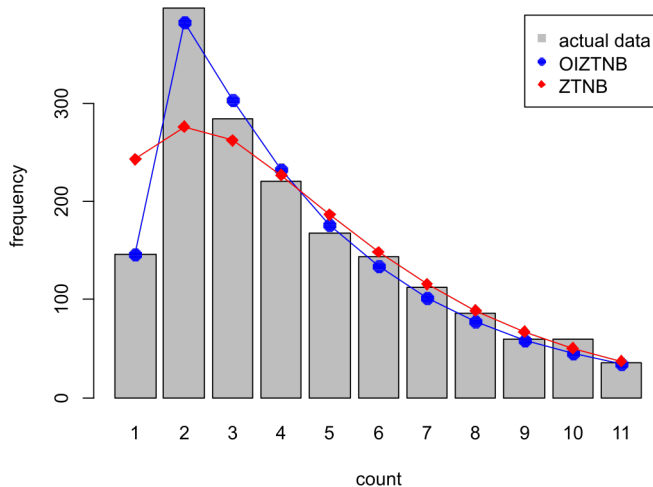
This illustration uses 1995 Arizona Medicare data for the hospital length of stay for patients undergoing a heart procedure. The data may be found in the R package `COUNT` (Hilbe and Robinson, 2016). This data has appeared as an application of the ZTNB model, and a two component truncated negative binomial mixture model, in Hilbe (2007, 2011). Note that finite mixture models are incapable of capturing one-inflation, since the mixture component would need to reach the

Table 3: ZTNB and OIZTNB models fit to the Arizona Medicare data.

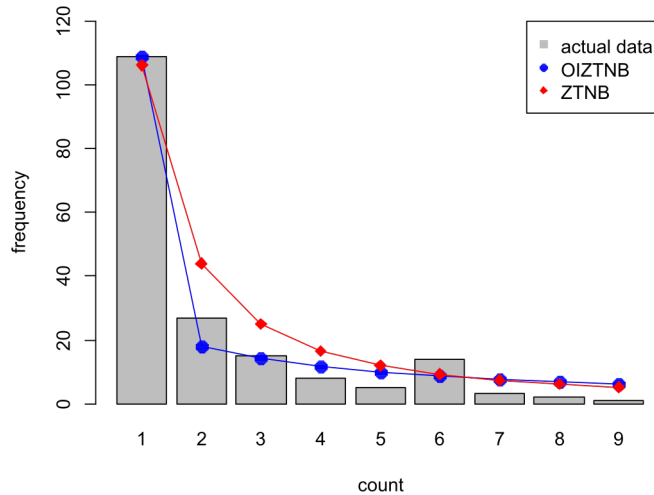|	ZTNB	OIZTNB		ZTNB	OIZTNB
	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}$	$\hat{\gamma}$	marginal effects	
(Intercept)	1.051***	0.744***	-2.131***		
	(0.043)	(0.066)	(0.199)		
gender	-0.166***	-0.178***	0.597***	-0.729***	-0.727***
	(0.035)	(0.047)	(0.205)	(0.160)	(0.165)
type1	0.737***	0.922***	-1.340***	2.699***	2.697***
	(0.040)	(0.056)	(0.183)	(0.134)	(0.136)
age75	0.131***	0.155***	-0.123	0.560***	0.552***
	(0.038)	(0.050)	(0.206)	(0.173)	(0.181)

Significance at the 1% (***), 5% (**), and 10% (*) levels.

boundary (i.e. $\lambda \to 0$). Regressors are again all dummies: $gender = 1$ for male; $type1 = 1$ for emergency/urgent admission and $= 0$ for elective admission; and $age75 = 1$ if the patient is over the age of 75.

One-deflation is present (average one-inflation is estimated to be -16.4%), and is seemingly driven by type of admission (see Table 3). A one-count is less probable under an emergency/urgent admission. Wald and LRT tests for no one-inflation again produce p-values of 0. Table 3 shows big swings in the values of the estimated $\hat{\boldsymbol{\beta}}$ between ZTNB and OIZTNB, as well as some moderate changes in the estimated marginal effects.

7.3 Recreation demand

Figure 3: OIZTNB and ZTNB regression models fit to the Loomis (2003) data set.

The data is from Loomis (2003) and is on number of trips taken to a recreation site. Estimates

Table 4: ZTNB and OIZTNB models fit to the Loomis (2003) data.

|	ZTNB	OIZTNB		ZTNB	OIZTNB
	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\gamma}}$	marginal effects	
(Intercept)	-0.326	3.415***	-3.480***		
	(1.024)	(0.430)	(0.986)		
gender	0.841***	0.469**	-0.531	14.01***	8.55***
	(0.281)	(0.196)	(0.343)	(3.55)	(1.99)
income2	1.166**	0.189	0.347	41.26*	4.154
	(0.484)	(0.285)	(0.604)	(23.85)	(7.76)
income3	-0.958**	-0.526	1.142*	-24.40**	-14.20**
	(0.416)	(0.313)	(0.588)	(9.56)	(6.08)
income4	-0.326**	-0.780*	0.003	-10.28	-17.00***
	(0.430)	(0.270)	(0.555)	(12.19)	(5.66)
travel2	-0.582	-0.499***	1.983***	-20.31	-16.78***
	(0.358)	(0.209)	(0.767)	(13.97)	(5.99)
travel3	-4.141***	-2.142***	4.828***	-54.93***	-40.75***
	(0.356)	(0.289)	(0.763)	(12.83)	(4.59)

Significance at the 1% (***), 5% (**), and 10% (*) levels.

of the effect of travel cost on the number of visits can be used to value the site. Endogenous stratification is of concern for this type of data; it is collected on-site, and risks over-sampling avid users. However, in this illustration, and in other examples for recreation demand, endogenous stratification seems to have little impact. For this data, Hilbe (2011) finds that the "zero-truncated model without adjustment for endogenous stratification is preferred." Additionally, he writes, "the extreme number of 1s in the data is the likely cause of the model not fitting." Descriptions of the variables can be found in Loomis (2003) or the R package COUNT.

ZTNB and OIZTNB regression models are estimated for the data (see Figure 3). The `oneinfl` package, which also estimates the standard PP and ZTNB models using the function `truncreg`, fails to replicate the ZTNB results for Hilbe (2011) obtained in Stata, or from the `glm.nb` function in the R package MASS. However, `oneinfl` appears superior, providing a higher estimated log-likelihood than either of these competitors.

Massive changes in the estimated $\hat{\boldsymbol{\beta}}$ and the marginal effects can be seen in Table 4, namely the effect on *income2* changes from 41.26 to 4.154. Average one-inflation is estimated to be 24% (with some observations having one-deflation), and both Wald and LRT tests (p-values 0) again support the presence of one-inflation.

Figure 4: OIZTNB and ZTNB regression models fit to the fishing trip data set.
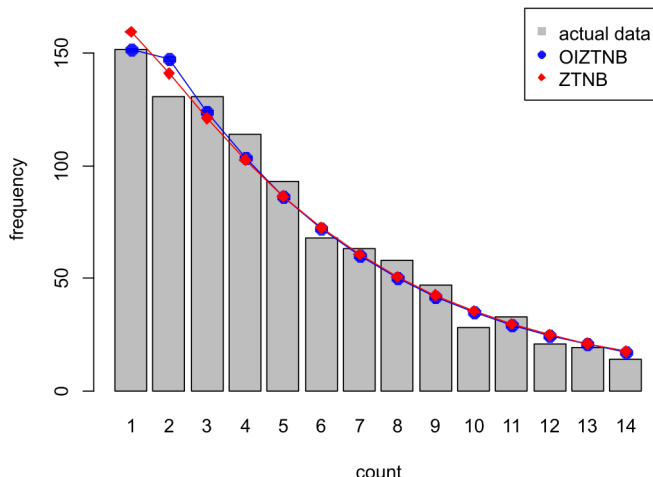
7.4 Fishing trips

In "Models for truncated counts" (Grogger and Carson, 1991), published in this journal, zero-truncated Poisson and negative binomial (PP and ZTNB) models are presented as important and viable in the fields of recreation demand, crime, transporation, and labour economics. The merits of the truncated models (in comparison to the untruncated counterparts), in particular the merits of ZTNB, are exemplified. To illustrate, the authors estimate a participation equation using data on the number of fishing trips taken by households in Alaska.

Average absolute one-inflation is estimated to be 4.4%. Most observations have one-deflation (average inflation is -2.2%). The Wald statistic (p-value 0.000) supports the presence of one-inflation, but the LRT is borderline (p-value is 0.09, unless the insignificant variables are dropped from Z, in which the p-value is 0.02). The fit of the models is illustrated in Figure 4, and estimated parameters are reported in Table 5. Most notably, a significant marginal effect (on $AVLONG$) changes sign.

Our replication efforts use data from Carson et al. (2009), with small differences in the parameter estimates from ZTNB, due to our inability to reproduce a $TRATE$ variable that was an inclusive variable from a nested logit model. Nonetheless, the highlighted discrepancies between the ZTNB and OIZTNB models should be invariant to the presence of this missing variable; when one-inflation is present (as supported by the Wald and LRT), additional variables can not explain away its presence unless they are linked to the one-inflation parameter.

Table 5: ZTNB and OIZTNB models fit to the fishing trip data set.

	ZTNB	OIZTNB		ZTNB	OIZTNB
	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}$	$\hat{\gamma}$	marginal effects	
(Intercept)	2.152***	2.188***	-2.322***		
	(0.081)	(0.103)	(0.284)		
AVLONG	-0.257***	-0.284***	0.255***	1.446***	-1.674***
	(0.022)	(0.027)	(0.046)	(0.139)	(0.166)
MISS	-0.151**	-0.147*	0.191	-0.814**	-0.795**
	(0.070)	(0.078)	(0.206)	(0.364)	(0.366)
CROWD	-0.034***	-0.026	0.075	-0.192	-0.176
	(0.030)	(0.033)	(0.091)	(0.168)	(0.186)
INC	0.000	0.001	0.002	0.002	0.004
	(0.001)	(0.001)	(0.003)	(0.005)	(0.006)
CPMILE	1.188***	1.013**	-3.028**	6.685***	6.941***
	(0.407)	(0.443)	(1.462)	(2.317)	(2.532)
FOFF1	-0.214***	-0.190***	0.416***	-1.205***	-1.233***
	(0.031)	(0.035)	(0.087)	(0.182)	(0.200)
LEISURE	0.097***	0.118***	0.026	0.544***	0.638***
	(0.030)	(0.034)	(0.092)	(0.172)	(0.192)

Significance at the 1% (***), 5% (**), and 10% (*) levels.

8   Conclusions

Several outcomes in economics are truncated counts, particularly in administrative settings. For administrative data, the individual is typically not observed unless they visit the entity that is collecting the data. For example, data on the number of health care visits typically begin at the 1-count; individuals that do not visit the doctor are not observed. A common endeavor is to determine how a policy or characteristic might influence the expected number of visits, and for data that are zero-truncated, an appropriate zero-truncated distribution must be used in order to avoid biased and inconsistent estimators.

The current standard truncated count model, the zero truncated negative binomial (ZTNB), is likely biased and inconsistent in most settings that it is employed. This is due to one-inflation. One-inflation occurs because the individual experiences being observed during their first visit. Information is revealed, and behaviour is altered, typically in a way that discourages future visits. Individuals must experience the observational setting before they can enter the data set, but the experience itself alters their behaviour. This causes altered 1 counts as the experience either encourages or dissuades future visits. While one-inflated distributions have been developed in areas outside economics, the distributions have not yet been extended to a regression framework, until

now.

This paper develops one-inflated zero truncated negative binomial (OIZTNB) and one-inflated positive Poisson (OIPP) regression models, as well as a plethora of related estimators and tools, all made available in the R package `oneinfl`. The methods developed here are applied to four data sets that have traditionally been used to champion the ZTNB regression model. In all cases, one-inflation is found, suggesting that the ZTNB model is actually misspecified. Simulation studies suggest large biases when specifying the wrong model. Accordingly, we recommend the use of the OIZTNB model in place of the ZTNB, for administrative type data.

References

Baser, O., Supina, D., Sengupta, N., Wang, L., and Kwong, L. (2010). Impact of postoperative venous thromboembolism on medicare recipients undergoing total hip replacement or total knee replacement surgery. *American Journal of Health-System Pharmacy*, 67(17):1438–1445.

Carson, R. T., Hanemann, W. M., and Wegge, T. C. (2009). A nested logit model of recreational fishing demand in alaska. *Marine Resource Economics*, 24(2):101–129.

Godwin, R. T. (2017). One-inflation and unobserved heterogeneity in population size estimation. *Biometrical Journal*, 59(1):79–93.

Godwin, R. T. (2019). The one-inflated positive poisson mixture model for use in population size estimation. *Biometrical Journal*, 61(6):1541–1556.

Godwin, R. T. and Böhning, D. (2017). Estimation of the population size by using the one-inflated positive poisson model. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 66(2):425–448.

Grogger, J. T. and Carson, R. T. (1991). Models for truncated counts. *Journal of applied econometrics*, 6(3):225–238.

Hardin, J. W. and Hilbe, J. M. (2007). *Generalized linear models and extensions*. Stata press.

Hardin, J. W. and Hilbe, J. M. (2015). Regression models for count data from truncated distributions. *The Stata Journal*, 15(1):226–246.

Hilbe, J., Robinson, A., and Robinson, M. A. (2018). Package 'msme'.

Hilbe, J. M. (2007). *Negative binomial regression*. Cambridge University Press.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press.

Hilbe, J. M. and Robinson, M. A. (2016). Package 'count'.

Johnson, N. L. and Kotz, S. (1969). *Distributions in Statistics—Discrete Distributions*. John Wiley & Sons.

Kleindorfer, D., Xu, Y., Moomaw, C. J., Khatri, P., Adeoye, O., and Hornung, R. (2009). Us geographic distribution of rt-pa utilization by hospital for acute ischemic stroke. *Stroke*, 40(11):3580–3584.

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

Lasater, K. B., Sloane, D. M., McHugh, M. D., Cimiotti, J. P., Riman, K. A., Martin, B., Alexander, M., and Aiken, L. H. (2021). Evaluation of hospital nurse-to-patient staffing ratios and sepsis bundles on patient outcomes. *American Journal of Infection Control*, 49(7):868–873.

Loomis, J. (2003). Travel cost demand model based river recreation benefit estimates with on-site and household surveys: Comparative results and a correction procedure. *Water Resources Research*, 39(4).

R Core Team, R. et al. (2013). R: A language and environment for statistical computing.

Rossmo, D. K. and Routledge, R. (1990). Estimating the size of criminal populations. *Journal of quantitative criminology*, 6:293–314.

Tajuddin, R. R. M., Ismail, N., and Ibrahim, K. (2022). Estimating population size of criminals: A new horvitz–thompson estimator under one-inflated positive poisson–lindley model. *Crime & Delinquency*, 68(6-7):1004–1034.