ORIGINAL RESEARCH PAPER

The consequences of checking for zero-inflation and overdispersion in the analysis of count data

Harlan Campbell, harlan.campbell@stat.ubc.ca

ARTICLE HISTORY

Compiled September 30, 2019

ABSTRACT

Count data are ubiquitous in ecology and the Poisson generalized linear model (GLM) is commonly used to model the association between counts and explanatory variables of interest. When fitting this model to the data, one typically proceeds by first confirming that the data is not overdispersed and that there is no excess of zeros. If the data appear to be overdispersed or if there is any zero-inflation, key assumptions of the Poison GLM may be violated and researchers will then typically consider alternatives to the Poison GLM. An important question is whether the potential model selection bias introduced by this data-driven multi-stage procedure merits concern. In this paper, we conduct a large-scale simulation study to investigate the potential consequences of model selection bias that can arise in the simple scenario of analyzing a sample of potentially overdispersed, potentially zero-heavy, count data.

KEYWORDS

poisson regression, overdispersion, negative-binomial, model-selection bias

1. Introduction

Despite the ongoing debate surrounding the use (and misuse) of significance testing in ecology (Murtaugh, 2014) (and in other fields (Amrhein et al., 2019)), hypothesis testing remains prevalent. Indeed, many research fields have been criticized for publishing studies with serious errors of testing and interpretation, and ecologists have been accused of being "confused" about when and how to conduct appropriate hypothesis tests (Stephens et al., 2005). One issue that receives a substantial amount of attention is that of failing to check for possible violations of distributional assumptions. According to Freckleton (2009), using statistical tests that assume a given distribution on the data while failing to test for the assumptions required of said distribution is one of "seven deadly sins."

One of the most popular statistical models in ecology (and in other fields, e.g., psychology, neuroscience and microbiome data, (Loeys et al., 2012; Zoltowski and Pillow, 2018; Xu et al., 2015)) is the Poisson generalized linear model (GLM) (Nelder and Wedderburn, 1972). With count outcome data, a Poisson GLM is the most common starting point for testing an association between a given outcome, Y, and a given covariate of interest, X. The Poisson GLM assumes the outcome data are the result of independent sampling from a Poisson distribution where, importantly, the mean and variance are equal. However, in practice, count data will often show more variation than is implied by the Poisson distribution and the use of Poisson models is not always appropriate (Cox, 1983).

Count data frequently exhibit two (related) characteristics: (1) overdispersion and (2) zero-inflation. (Overdispersion and zero-inflation are related to each other since an excess of zeros also contributes to overdispersion.) If the data are indeed overdispersed or if there is indeed an excess of zeros, assumptions underlying the Poisson GLM will not hold and ignoring this will lead to serious errors (e.g., biased parameter estimates and invalid standard errors). It is therefore routine practice for researchers to check if the data fulfill the assumptions required of the Poisson model and adopt an alternative

statistical model in the event that they do not; see Zuur et al. (2010).

In the case of overdispersion, popular alternatives to the Poison GLM include the Quasi-Poisson (QP) model (Wedderburn, 1974) and the Negative Binomial (NB) model (Richards, 2008; Lindén and Mäntyniemi, 2011). Note that when selecting between the QP and NB models, the best choice is not always straightforward, is often data-driven, and is often based on rather subjective criteria; see Ver Hoef and Boveng (2007). In the case of zero-inflation, popular alternatives to the Poison GLM include the Zero-Inflated Poisson model (ZIP) (Martin et al., 2005; Lambert, 1992) and the Zero-Inflated Negative-Binomial model (ZINB) (Greene, 1994).

A multi-stage procedure will typically have researchers testing for overdispersion and zero-inflation in a preliminary stage, before testing the main hypothesis of interest (i.e., the association between Y and X) in a second stage; see Blasco-Moreno et al. (2019). If the first stage tests are not significant, the Poisson GLM is fit, regression coefficients are estimated along with their standard errors, and p-values are calculated allowing one to test for the association between Y and X. On the other hand, if the first stage test for overdispersion is significant, a QP or a NB model will be fit to the data. Or, alternatively, if the first stage test for zero-inflation is significant, a ZIP model may be used. In cases when there is evidence of both overdispersion and zero-inflation, more complex models such as the ZINB model or hurdle models will often be considered; see Zorn (1998).

Such a multi-stage, multi-test procedure may appear rather reasonable, and goodness-of-fit tests (e.g., likelihood ratio tests; the Vuong and Clarke tests (Clarke, 2007; Wilson, 2015)) are frequently reported to confirm that the model-selection is appropriate. However, recently, some researchers have warned against preliminary testing for distributional assumptions; e.g., Shuster (2005) and Wells and Hintze (2007). Their warnings are based on the following concern. Since the preliminary tests are applied to the same data as the main hypothesis tests, this multi-stage procedure amounts to "using the data twice" and may result in model selection bias. In other words, a hypothesis test using a model selected based on preliminary testing fails to take into account one's uncertainty with regards to the distributional properties of the data.

The model selection bias at issue here is not the better known model selection bias associated with deciding *post*-hoc which variables to include in the model, e.g., the model selection bias associated with stepwise regression (Hurvich and Tsai, 1990; Whittingham et al., 2005). Instead, here we are concerned with the potential bias introduced when deciding *post*-hoc which distributional assumptions should be accepted. The implications of considering *post*-hoc alternatives (or adjustments) to accommodate for distributional assumptions have been previously considered in other contexts. Three examples come to mind.

First, in the context of time-to-event data, the consequences of checking and adjusting for potential violations of the proportional hazards (PH) assumption required of a Cox PH model are considered by Campbell and Dean (2014). The authors find that the "common two-stage approach" (in which one selects a model based on a preliminary test for PH) can lead to a substantial inflation of the type 1 error, even in scenarios where there is no violation of the PH assumption.

Second, in the simple context of testing the means of two independent samples, Rochon et al. (2012) investigate the consequences of conducting a preliminary test for normality (e.g., the Shapiro-Wilk test). The authors conclude that while "[f]rom a formal perspective, preliminary testing for normality is incorrect and should therefore be avoided," in practice, "preliminary testing does not seem to cause much harm, at least for the cases we have investigated."

Finally, in the context of clinical trials, factorial trials are an efficient method of estimating multiple treatments in a single trial. However, factorial trials rely on the strict assumption of no interaction between the different treatments. Kahan (2013) investigates the consequences of conducting a preliminary test for the interaction between treatment arms (as is often recommended). By means of a simulation study, Kahan (2013) shows that the estimated treatment effect from a factorial trial under the "two-stage analysis" can be severely biased, even in the absence of a true interaction.

Model selection bias is considered a "quiet scandal in the statistical community" (Breiman, 1992) and is now all the more important to understand given recent concerns

with research reproducibility and researcher incentives (Kelly, 2019; Nosek et al., 2012; Gelman and Loken, 2013; Fraser et al., 2018). In ecology, some have warned about model selection bias (e.g., Buckland et al. (1997)), but the problem "remains widely over-looked" (Whittingham et al., 2006). Indeed, ecologists will readily admit that "this problem is commonly not appreciated in modelling applications" (Whittingham et al., 2005). Anderson (2007) notes that: "Model selection bias is subtle but its effects are widespread and little understood by many people working in the life sciences."

In this short paper, we conduct a large-scale simulation study to investigate the potential consequences of model selection bias that can arise in the simple scenario of analyzing a sample of potentially overdispersed, potentially zero-inflated, count data. In Section 2, we review commonly used "preliminary tests" for the distributional assumptions required of the Poisson GLM. In Section 3, we outline the framework of a simulation study to investigate the consequences of checking for zero-inflation and overdispersion. In Section 4, we discuss the results of this simulation study and we conclude in Section 5 with a summary of findings and general recommendations for practitioners.

2. Testing for distributional assumptions

We will consider the simplest version of the Poisson GLM. Let Y_i, for i in 1,..., n, be the outcome of interest observed for n independent samples. Let X_i, for i in 1,...,n, represent a single covariate of interest. If the covariate of interest is categorical with k different categories (e.g., k species of fish), X_i will be a vector with length equal to $k-1$; otherwise it will be a single scalar and $k=2$. The simplest Poisson regression model, with a standard *log* link, will have that:

$$Y_i \sim Poisson(\lambda_i = exp(\beta_0 + \beta_X X_i)), \text{ for } i \text{ in } 1, ..., n; \quad (1)$$

where β_0 is the intercept, and β_X is the coefficient (or coefficient-vector of length

$k-1$) representing the association between X and Y. Note that this model implies the following equality: $E[Y_i] = Var[Y_i] = \lambda_i$, for i in 1, ..., n. Parameter estimates, $\widehat{\beta_0}$, and $\widehat{\beta_X}$, can be obtained by iterative Fischer scoring. A confidence interval for β_X is typically calculated by the standard profile likelihood approach where one inverts a likelihood-ratio test; see Venzon and Moolgavkar (1988) and Uusipaikka (2008).

For testing, in order to determine whether there is an association between Y and X, we define the following hypothesis test: $H_0 : \beta_X = 0$ vs. $H_1 : \beta_X \neq 0$. A simple chi-squared statistic, Z, can be obtained to evaluate this hypothesis by calculating the null and residual deviance as follows:

$$D_0 = 2\sum_{i=1}^{n}\left\{Y_i \log\left(Y_i/\widehat{\beta_0}\right) - \left(Y_i - \widehat{\beta_0}\right)\right\},$$
$$D_1 = 2\sum_{i=1}^{n}\left\{Y_i \log\left(Y_i/\widehat{\lambda_i}\right) - \left(Y_i - \widehat{\lambda_i}\right)\right\}, \text{ where } \widehat{\lambda_i} = \exp(\widehat{\beta_0} + \widehat{\beta_X} X).$$

If the distributional assumptions of the Poisson GLM are met and the null hypothesis holds, the test statistic, $Z = D_1 - D_0$, will follow (asymptotically) a χ^2 distribution with $df = k-1$ degrees of freedom, and the p-value is calculated as: p-value $= P_{\chi^2}(Z, df = k-1)$. However, if the distributional assumptions do not hold, Z will be compared with the wrong reference distribution invalidating any significance test (and any associated confidence intervals). Therefore, in order to conduct valid inference, researchers will typically carry out an extensive model selection procedure. Blasco-Moreno et al. (2019) outline and illustrate a proposed protocol. Such a procedure is typically based on:

- measuring indices (e.g., the dispersion index (Fisher, 1950); the zero-inflation index (Puig and Valero, 2006));
- conducting score tests (e.g., the $D\&L$ score test for Poisson vs. NB regression (Dean and Lawless, 1989); the vdB score test for Poisson vs. ZIP regression (Van den Broek, 1995); the score test for ZIP vs ZINB regression (Ridout et al., 2001));
- and evaluating candidate models with goodness-of-fit tests (e.g., likelihood ratio tests; the Vuong and Clarke tests) and model selection criteria (e.g., AIC and BIC).

In this paper, for simplicity, we will only consider three alternative models: the (type 2) NB, the ZIP, and the (type 2) ZINB regression models as described in Blasco-Moreno et al. (2019). Furthermore, we will restrict our attention to using three score tests: (1) the $D\&L$ score test for testing the Poisson model against the NB (Dean and Lawless, 1989); (2) the vdB score test for testing the Poisson model against the ZIP (Van den Broek, 1995); and (3) the Ridout score test for testing the ZIP against the ZINB (Ridout et al., 2001). Let us briefly review the three alternative regression models that we will consider.

(1) The ZIP regression model - We will consider the the following zero-inflated poisson model where the weights, ω_i, are a function of the covariate X_i. Specifically,

$$\begin{aligned} Pr(Y_i = y_i|\omega_i, \lambda_i) &= \omega_i + (1-\omega_i)exp(-\lambda_i), \quad \text{if} \quad y = 0; \\ &= (1-\omega_i)exp(-\lambda_i)\lambda_i^{y_i}/y_i!, \quad \text{if} \quad y_i > 0; \end{aligned} \tag{2}$$

where we use a log link function for $\lambda_i = exp(\beta_0 + \beta_X X_i)$; and a logit link function for $\omega_i = \left(\frac{exp(\gamma_0+\gamma_X X_i)}{1+exp(\gamma_0+\gamma_X X_i)}\right)$. The ZIP model has that $0 \leq \omega_i \leq 1$ and $\lambda_i > 0$, and implies the following about the mean and variance of the data: $E(Y_i) = \lambda_i(1-\omega_i) = \mu_i$ and $Var(Y_i) = \mu_i + \mu_i^2\omega_i/(1-\omega_i)$. We will conduct a likelihood ratio test for the null hypothesis of no association between X and Y ($H_0 : \beta_X = \gamma_X = 0$).

(2) The (type 2) NB regression model - Consider the following model:

$$\Pr(Y = y_i|\alpha, \lambda_i) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha\lambda_i}\right)^{1/\alpha} \left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right)^{y_i}; \tag{3}$$

where we use a log link function for $\lambda_i = exp(\beta_0 + \beta_X X_i)$, and where $\alpha \geq 0$ is a dispersion parameter that does not depend on covariates. The type 2 NB model implies the following about the mean and variance of the data: $E[Y_i] = \lambda_i$, and $Var(Y_i) = \lambda_i + \alpha\lambda_i^2$. We will conduct a likelihood ratio test for the null hypothesis of no association

between X and Y ($H_0 : \beta_X = 0$).

(3) The (type 2) ZINB regression model - Consider the following:

$$\begin{aligned} Pr(Y_i = y_i|\omega_i, \lambda_i) &= \omega_i + (1-\omega_i)(1/(1+\alpha\lambda_i))^{1/\alpha}, \quad \text{if} \quad y = 0; \\ &= (1-\omega_i)\frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)}\left(\frac{1}{1+\alpha\lambda_i}\right)^{1/\alpha}\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right)^{y_i}, \quad \text{if} \quad y_i > 0; \end{aligned} \tag{4}$$

where we use a log link function for $\lambda_i = exp(\beta_0 + \beta_X X_i)$; a logit link function for $\omega_i = \left(\frac{exp(\gamma_0+\gamma_X X_i)}{1+exp(\gamma_0+\gamma_X X_i)}\right)$; and where $\alpha \geq 0$ is a dispersion parameter that does not depend on covariates. We will conduct a likelihood ratio test for the null hypothesis of no association between X and Y ($H_0 : \beta_X = \gamma_X = 0$).

Let us know briefly review the three score tests that we will consider.

(1) The $D\&L$ score test - Dean and Lawless (1989) proposed calculating the following score statistic for testing overdispersion:

$$T_1 = \sum_{i=1}^{n}\left\{\left(y_i - \hat{\lambda}_i\right)^2 - y_i\right\} / \left(2\sum_{i=1}^{n}\hat{\lambda}_i^2\right)^{1/2} \tag{5}$$

Under the null hypothesis of no overdispersion, the T_1 statistic converges to a standard Normal distribution and the p-value is calculated as: p-value $= P_{\mathcal{N}}(T_1)$.

(2) The vdB score test - Van den Broek (1995) proposed calculating the following score test statistic for checking for zero-inflation (i.e., testing the Poisson against a ZIP):

$$S = \frac{\left\{\sum_{i=1}^{n}\left(\frac{I(\{y_i=0\})}{exp(-\hat{\lambda}_i)} - 1\right)\right\}^2}{\left\{\sum_{i=1}^{n}\left(\frac{1}{exp(-\hat{\lambda}_i)} - 1\right)\right\} - \hat{\lambda}^\top \mathbf{X}\left[\mathbf{X}^\top \text{diag}(\hat{\lambda})\mathbf{X}\right]^{-1}\mathbf{X}^\top\hat{\lambda}}. \tag{6}$$

Under the null of no zero-inflation, this S statistic has an asymptotic χ^2_1 distribution and the p-value can be calculated as: p-value $= 1 - P_{\chi^2}(S, df = 1)$.

(3) The Ridout score test - Ridout et al. (2001) proposed calculating the following score test statistic for testing a zero-inflated Poisson regression model against a zero-inflated negative binomial alternative :

$$T = \sqrt{\hat{J}^{\alpha\alpha}} \frac{1}{2} \sum_{i=1}^{n} \left([(y_i - \hat{\lambda}_i)^2 - y_i] - I(\{y_i = 0\}) \hat{\lambda}_i^2 \frac{\hat{\omega}_i}{Pr(Y_i = 0)} \right), \tag{7}$$

where $\hat{\lambda}_i$ and $\hat{\omega}_i$ are the point estimates for parameters λ_i and ω_i obtained from the ZIP; and where $\sqrt{\hat{J}^{\alpha\alpha}}$ is the estimated standard error of the β_0 intercept parameter in the ZIP model. Asymptotically, if the data do indeed follow a Zero-Inflated Poisson distribution, the T statistic has a **chi-squared** distribution with df=1.

3. Methods

As discussed in the previous section, prevailing practice for the analysis of count data is first to try to fit one's data with a Poisson GLM and only consider alternatives in the event that a preliminary test indicates a that the distributional assumptions may be violated. We will therefore consider the following multi-stage testing procedure in our investigation. This procedure follows the recommendations of Blasco-Moreno et al. (2019) yet represents a simplification of the typical process followed by researchers.

- **Step 1.** Conduct the *vdB* score test for zero-inflation (H_0: Poisson vs. H_1: ZIP).
- - **Step 2.** If the *vdB* score test fails to reject the null, conduct the $D\&L$ score test for overdispersion (H_0: Poisson vs. H_1: NB). Otherwise, proceed to Step 5.

- **Step 3.** If the $D\&L$ score test fails to reject the null, fit the Poisson GLM and calculate the p-value. Otherwise, proceed to Step 4.
- **Step 4.** If the $D\&L$ score test rejects the null, fit the NB regression model and calculate the p-value.

- **Step 5.** If the vdB score test rejects the null, conduct the Ridout score test (H_0: Poisson vs. H_1: NB).

- **Step 6.** If the Ridout score test fails to reject the null, fit the ZIP and calculate the p-value. Otherwise, proceed to Step 7.
- **Step 7.** If the Ridout score test rejects the null, fit the ZINB regression model and calculate the p-value.

Figure 1 illustrates the multi-stage model selection procedure with the Poisson GLM as the starting point. Note that, in their example analysis of plant-herbivore interaction data, Blasco-Moreno et al. (2019) conduct a version of the above procedure with a slightly different order. First, based on the $D\&L$ score test, "the data is clearly overdispersed and a NB model was preferred to a Poisson." After this first preliminary test, Blasco-Moreno et al. (2019) conduct a Ridout score test and as a consequence "the ZIP model was rejected in favour of the ZINB model." The authors therefore accept the ZINB model as the chosen model and state that a "comparison between Poisson and ZIP regression (Van der Broek score test) was not needed because both models were already rejected."

We conducted two large-scale simulation studies in which samples of data were drawn from four different distributions:

(1) the Poisson distribution:
$y_i \sim Poisson(\lambda = exp(\beta_0))$, for i in 1,...n;

(2) the (type 2) Negative Binomial distribution:
indent $y_i \sim NegBin(\alpha, \lambda = exp(\beta_0))$, for $i = 1, ..., n$;

(3) the Zero-Inflated Poisson distribution:

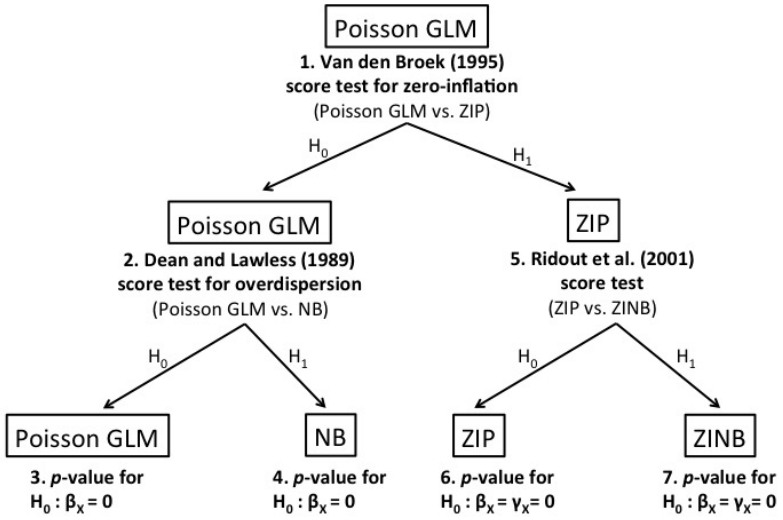

Figure 1. The multi-stage model selection procedure. The Poisson GLM is the starting point. Three score tests lead to one of four models.

$y_i \sim ZIPoisson(\omega, \lambda = exp(\beta_0))$, for $i = 1, ..., n$; and

(4) the Zero-Inflated Negative Binomial distribution:
$y_i \sim ZINegBin(\alpha, \omega, \lambda = exp(\beta_0))$, for $i = 1, ..., n$.

We varied the following: $n = (60, 160, 260, 500, 800, 1200, 2500)$, $\beta_0 = (0.5, 1, 1.5, 2, 2.5, 3, 4)$, $\alpha = (1, 1.5, 2, 3, 4)$, and $\omega = (0, 0.025, 0.1, 0.2, 0.5)$. Note that going forward we reffer to scenarios with $\alpha > 1$ and $\omega = 0$ as those with data simulated from the Negative Binomial distribution; scenarios with $\omega > 0$ and $\alpha = 1$ as those with data simulated from the Zero-inflated Poisson distribution; scenarios with $\alpha = 1$ and $\omega = 0$ as those with data simulated from the Poisson distribution; and scenarios with $\alpha > 1$ and $\omega > 0$ as those with data simulated from the Zero-Inflated Negative Binomial distribution. We considered X_i as a univariate continuous covariate from a Normal distribution: $X_i \sim Normal(\mu = 0, \sigma = 10)$, for i in 1,..., n (as such, $k = 2$). Note that the covariate matrix X is simulated anew for each individual simulation run. Therefore, we are considering the case of *random* regressors. Chen and Giles (2011) discuss the difference between fixed and random covariates. The assumption of fixed covariates is generally considered only in experimental settings whereas an assumption