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

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

Abstract

1. 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 model assumptions are satisfied. If the residuals appear to be overdispersed or if there is 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.

2. Here 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. Specifically, we investigate model selection procedures recently recommended by Blasco-Moreno et al. (2019) using either a series of score tests or the AIC statistic to select the best model.

3. We find that, when sample sizes are small, model selection based on preliminary score tests (or the AIC) can lead to potentially substantial type 1 error inflation. When sample sizes are large, model selection based on preliminary score tests is less problematic.

4. Ignoring the possibility of overdispersion and zero inflation during data analyses can lead to invalid inference. However, if one does not have sufficient power to test for overdispersion and zero inflation, *post hoc* model selection may also lead to substantial bias. This "catch-22" suggests that, if sample sizes are small, a healthy skepticism is warranted whenever one rejects the null hypothesis.

KEYWORDS

model selection bias, overdispersion, zero inflation, zero-inflated models

1. Introduction

Despite the ongoing debate surrounding the use (and misuse) of significance testing in ecology (Murtaugh 2014, Dushoff et al. 2019) (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 many other fields, e.g., finance, psychology, neuroscience, and microbiome research, (Bening & Korolev 2012, Loeys et al. 2012, Zoltowski & Pillow 2018, Xu et al. 2015)) is the Poisson generalized linear model (GLM) (Nelder & 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, conditional on the covariates, 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 be 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 refers to an excess of variability, while zero-inflation refers to an excess of zeros (Yang et al. 2010). If model residuals are overdispersed or have 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 assumptions required of a Poisson model hold 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 & Mäntyniemi 2011). Note that when selecting between the QP and NB models, the best choice is not always straightforward; see Ver Hoef & 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 (based on the residuals from a Poisson GLM), 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 selected, 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 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 & 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" (Hayes 2020). 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. Unless the preliminary tests and the main hypothesis tests are entirely independent, this can result in model selection bias.

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 & 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 & 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."

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 & Loken 2013, Fraser et al. 2018). In some fields, such as psychology, the issue

is finally being recognized. Williams & Albers (2019) conclude that "it is currently unclear how [psychology] researchers should deal with distributional assumptions" since "diagnosing and responding to distributional assumption problems" may result in "error rates [that] vary considerably from the nominal error rates."

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 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. It is difficult to anticipate what these consequences might be. Often, while model selection bias is problematic from a theoretical perspective, it does not lead to substantial problems in practice. We restrict our attention to two model selection procedures, one based on conducting score tests, and another based on calculating AIC statistics.

In Section 2, we review commonly used models and 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.

2. Models for the analysis of count data

Let us consider the simplest version of the Poisson GLM. Let Y_i, for i in $1, \ldots, n$, be the outcome of interest observed for n independent samples. Let X_i, for i in $1, \ldots, n$, represent a single covariate of interest. If the covariate of interest is categorical with k different categories (e.g., k different 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:

$$Y_i \sim Poisson(\lambda_i = exp(\beta_0 + \beta_X X_i)), \text{or equivalently:} \tag{1}$$

$$Pr(Y_i = y_i|\beta_0, \beta_X) = \frac{(exp(\beta_0 + \beta_X X_i))^{y_i} exp(-exp(\beta_0 + \beta_X X_i))}{y_i!}, \text{ for } i \text{ in } 1, ..., n; \tag{2}$$

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 $1, \ldots, n$.

Parameter estimates, $\widehat{\beta_0}$, and $\widehat{\beta_X}$, can be obtained by maximum likelihood estimation via 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 & Moolgavkar (1988) or more recently Uusipaikka (2008). Maximum likelihood estimation via iterative Fischer scoring is implemented as a default for the `glm()` function in `R`; see Dunn & Smyth (2018). Profile likelihood confidence intervals are provided by default when using the `confint()` function for GLMs; see Ripley et al. (2013).

To test 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 likelihood ratio test (LRT), or Wald test will provide a p-value to evaluate this hypothesis; see Zeileis et al. (2008). The LRT and Wald test are asymptotically equivalent. For the likelihood ratio test, the Z statistic is obtained by calculating the null and residual deviance as $Z_{LRT} = D_1 - D_0$, where :

$D_0 = 2\sum_{i=1}^{n} \left\{ Y_i \log\left(Y_i / exp(\widehat{\beta}_0)\right) - \left(Y_i - exp(\widehat{\beta}_0)\right) \right\}$, and:

$D_1 = 2\sum_{i=1}^{n} \left\{ Y_i \log\left(Y_i / \widehat{\lambda}_i\right) - \left(Y_i - \widehat{\lambda}_i\right) \right\}$, where $\widehat{\lambda}_i = \exp(\widehat{\beta}_0 + \widehat{\beta}_X X_i)$.

If the distributional assumptions of the Poisson GLM are met and the null hypothesis holds, the Z statistic will follow (asymptotically) a χ^2 distribution with $df = k-1$ degrees of freedom, and the p-value is calculated as: p-value $= P_{\chi^2}(Z, df = k-1)$. (For the Wald test, with $k = 2$, the Z-statistic is defined as $Z_{Wald} = \left(\widehat{\beta_X}/\text{se}(\widehat{\beta_X})\right)^2$, where $\text{se}(\widehat{\beta_X})$ is the standard error of the maximum likelihood estimate (MLE); see Hilbe & Greene (2007) for details when $k > 2$).

If the distributional assumptions do not hold, the Z statistic will be compared with the wrong reference distribution invalidating any significance test (and associated confidence intervals). Therefore, in order to conduct valid inference, researchers will typically carry out an extensive model selection procedure. Note that model selection must always be based on model residuals and not on the distribution of the response variable (which is erroneously done on occasion). 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 & Valero 2006));
- conducting score tests (e.g., the $D\&L$ score test for Poisson vs. NB regression (Dean & 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, AICc and BIC).

In this paper, for simplicity, we will only consider three alternative models in addition to the Poisson model described above: the (type 2) NB, the ZIP, and the (type 2) ZINB regression models as described in Blasco-Moreno et al. (2019). Let us briefly review these three alternative regression models.

(1) The ZIP regression model - We will consider the following zero-inflated Poisson model where the probability of a structural zero, ω_i, is a function of the covariate X_i. Specifically,

$$
\begin{aligned}
Pr(Y_i = y_i|\omega_i, \lambda_i) &= \omega_i + (1-\omega_i)exp(-\lambda_i), \quad \text{if} \quad y = 0; \qquad (3)\\
&= (1-\omega_i)exp(-\lambda_i)\lambda_i^{y_i}/y_i!, \quad \text{if} \quad y_i > 0;
\end{aligned}
$$

where we have a log link function for λ_i and a logit link function for ω_i (for i in $1, \ldots, n$) such that:

$$
\begin{aligned}
\lambda_i &= exp(\beta_0 + \beta_X X_i), \quad \text{and} \qquad (4)\\
\omega_i &= \left(\frac{exp(\gamma_0 + \gamma_X X_i)}{1 + exp(\gamma_0 + \gamma_X X_i)}\right). \qquad (5)
\end{aligned}
$$

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: $\mathrm{E}(Y_i) = \lambda_i(1-\omega_i) = \mu_i$ and $\mathrm{Var}(Y_i) = \mu_i + \mu_i^2\omega_i/(1-\omega_i)$. The dispersion index is therefore equal to $d = \mathrm{Var}(Y_i)/\mathrm{E}(Y_i) = 1 + \lambda_i\omega_i$. As $\omega_i \to 0$, we have that Y_i reverts to follow a Poisson distribution with mean λ_i. A null hypothesis of no association between X and Y is specified as: $H_0 : \beta_X = \gamma_X = 0$.

(2) The (type 2) NB regression model - We will consider the following NB regression model:

$$
\Pr(Y = y_i|\nu, \lambda_i) = \frac{\Gamma(y_i + \nu)}{\Gamma(y_i + 1)\Gamma(\nu)}\left(\frac{1}{1 + \lambda_i/\nu}\right)^{\nu}\left(\frac{\lambda_i/\nu}{1 + \lambda_i/\nu}\right)^{y_i}; \qquad (6)
$$

where we use a log link function for $\lambda_i = exp(\beta_0 + \beta_X X_i)$, and where $\nu > 0$ 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: $\mathrm{E}(Y_i) = \lambda_i$, and $\mathrm{Var}(Y_i) = \lambda_i + \lambda_i^2/\nu$. The dispersion index is therefore equal to $d = Var(Y_i)/E(Y_i) = 1 + \lambda_i/\nu$. As $\nu \to \infty$, we have that Y_i reverts to follow a Poisson distribution with mean λ_i. A null hypothesis of no association between X and Y is specified as: $H_0 : \beta_X = 0$.

(3) The (type 2) ZINB regression model - We will consider the following ZINB

regression model:

$$
\begin{aligned}
Pr(Y_i = y_i|\nu, \omega_i, \lambda_i) &= \omega_i + (1-\omega_i)(1/(1+\lambda_i/\nu))^{\nu}, \quad \text{if} \quad y = 0; \\
&= (1-\omega_i)\frac{\Gamma(y_i+\nu)}{\Gamma(y_i+1)\Gamma(\nu)}\left(\frac{1}{1+\lambda_i/\nu}\right)^{\nu}\left(\frac{\lambda_i/\nu}{1+\lambda_i/\nu}\right)^{y_i}, \quad \text{if} \quad y_i > 0;
\end{aligned} \tag{7}
$$

where we use a log link function for λ_i and a logit link function for ω_i as described in equations (4) and (5); and where $\nu > 0$ is a dispersion parameter that does not depend on covariates. The type 2 ZINB model implies the following about the mean and variance of the data: $\mathrm{E}(Y_i) = \lambda_i(1-\omega_i)$, and $\mathrm{Var}(Y_i) = (1-\omega_i)(\lambda_i + \lambda_i^2(\omega_i + 1/\nu))$. The dispersion index is therefore equal to $d = Var(Y_i)/E(Y_i) = 1 + \lambda_i(\omega_i + 1/\nu)$. A null hypothesis of no association between X and Y is specified as: $H_0 : \beta_X = \gamma_X = 0$.

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 that the distributional assumptions of the Poisson GLM may be violated. We will therefore consider the following multi-stage testing procedure in our investigation. This follows the recommendations of Blasco-Moreno et al. (2019) yet represents a simplification of the typical process followed by researchers. Walters (2007) also recommends a similar multi-step model selection procedure.

For the illustrative purposes of this paper, we consider the Dean & Lawless (1989) score test (D&L test) for oversdispersion and the Vuong (1989) test for zero-inflation (see Appendix for details) in the following seven step procedure:

- **Step 1.** Conduct the $D\&L$ score test for overdispersion (H_0: Poisson vs. H_1: NB).
- - **Step 2.** If the $D\&L$ score test fails to reject the null, conduct a Vuong test for zero-inflation (H_0: Poisson vs. H_1: ZIP). Otherwise, proceed to Step 5.
  - - **Step 3.** If the Vuong test for zero-inflation fails to reject the null, fit the Poisson GLM and calculate the p-value ($H_0 : \beta_X = 0$ vs. $H_1 : \beta_X \neq 0$). Otherwise, proceed to Step 4.
    - **Step 4.** If the Vuong test for zero-inflation rejects the null, fit the ZIP model and calculate the p-value ($H_0 : \beta_X = \gamma_X = 0$).
  - **Step 5.** If the $D\&L$ score test rejects the null, conduct the Vuong test for zero-inflation (H_0: NB vs. H_1: ZINB).
  - - **Step 6.** If the the Vuong test for zero-inflation fails to reject the null, fit the NB model and calculate the p-value ($H_0 : \beta_X = 0$). Otherwise, proceed to Step 7.
    - **Step 7.** If the Vuong test for zero-inflation rejects the null, fit the ZINB regression model and calculate the p-value ($H_0 : \beta_X = \gamma_X = 0$).

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. First, based on the $D\&L$ score test, "the data is clearly overdispersed and a NB model was preferred to a Poisson." The authors also conduct Vuong and Clarke tests: "The Vuong and Clarke tests rejected the Poisson and NB models in favour of their zeroinflated versions[...]." We decided to consider the Vuong test in our simulations instead of the Clarke test (or the Ridout score test), since the Vuong test appears to be the most widely used in practice. We also investigate another, simpler, model selection

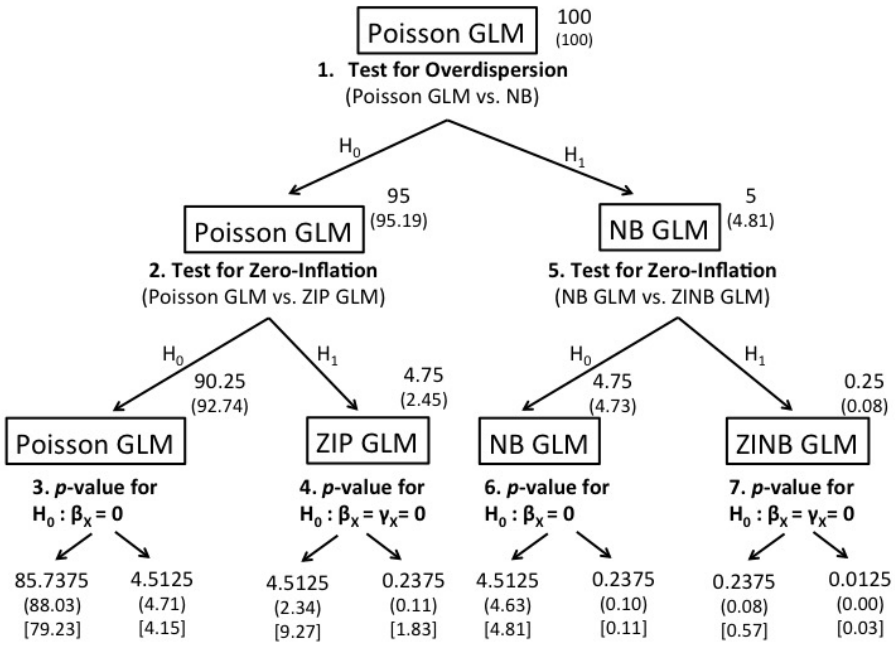

Figure 1. The multi-stage model selection procedure. The Poisson GLM is the starting point. Three score tests lead to one of four models. Numbers in the top right-hand corner of each node indicate the expected number of datasets (out of a total of 100) to reach each outcome if the data was Poisson (with $\beta_X = 0$), and each of the tests were truly independent (with a $\alpha = 0.05$ type 1 error rate). Numbers in parentheses correspond to results from the simulation study (Scenario "3", with $\phi = \infty$, $\omega = 0$, $\beta_0 = 0.5$ and $n = 250$). Numbers in (curved) parentheses are those obtained in following the seven-step procedure; numbers in [square] parentheses are those obtained via AIC. The unconditional type 1 error rate obtained in the simulation study following the seven-step procedure is 4.92% (= 4.71 + 0.11 + 0.10 + 0.00). The unconditional type 1 error rate obtained the simulation study when selecting the best model via AIC is 6.12% (= 4.15 + 1.83 + 0.11 + 0.03).

strategy: among the four models considered, the model with lowest AIC is chosen and the corresponding p-value for the association between X and Y is calculated (Brooks et al. 2019).

We conducted a large-scale simulation study 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:

$y_i \sim NegBin(\nu, \lambda = exp(\beta_0))$, for $i = 1, ..., n$;

(3) the Zero-Inflated Poisson distribution:

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

(4) the Zero-Inflated Negative Binomial distribution:

$y_i \sim ZINegBin(\nu, \omega, \lambda = exp(\beta_0))$, for $i = 1, ..., n$.

For each scenario, all data are simulated under the null hypothesis (i.e., with $\beta_X = 0$ and $\gamma_X = 0$). We varied the following: the sample size, $n = (50, 100, 250, 500, 1000, 2000)$, the intercept, $\beta_0 = (0.5, 1.0, 1.5, 2.0, 2.5)$, and the probability of a structural zero, $\omega = (0, 0.05, 0.1, 0.2, 0.5)$. We also varied the degree of overdispersion by setting $\phi = \nu/\lambda = (\infty, 2, 1, 1/2, 1/3)$ (so that data simulated from the Negative Binomial distribution has a dispersion index of $d = 1 + \lambda/\nu = 1 + 1/\phi = (1.0, 1.5, 2.0, 3.0, 4.0)$).

- scenarios with $\phi = \infty$ and $\omega = 0$ as those with data simulated from the Poisson distribution;
- scenarios with $\phi < \infty$ and $\omega = 0$ as those with data simulated from the Negative Binomial distribution;
- scenarios with $\phi = \infty$ and $\omega > 0$ as those with data simulated from the Zero-inflated Poisson distribution; and
- scenarios with $\phi < \infty$ 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, with mean of zero and variance of 100: $X_i \sim Normal(0, 100)$, 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 & 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 of random covariates is typically more appropriate for observational studies. Also note that the simulation study can only test for rates of false-positives (since $\beta_X = 0$ and $\gamma_X = 0$ for all scenarios). We are not testing for excessive false-negatives (and overly wide confidence intervals) which are also undesirable (Brooks et al. 2019).

Note that, for the Poisson distributed data, we are simulating data with overall

mean of $\lambda = exp(\beta_0) \approx (1.6, 2.7, 4.5, 7.4, 12.2)$. For $\lambda > 5$, zeros in the data are quite rare since $Pr(Y = 0|\lambda) \approx 0$. We did not consider models that deal with under-dispersion, even though under-dispersed counts may arise in various ecological studies; see Lynch et al. (2014).

In total, we considered 750 distinct scenarios and for each simulated 15,000 unique datasets. For each dataset, we conducted the seven step procedure and recorded all p-values and whether or not the null hypotheses is rejected at the 0.05 significance level under the entire procedure. We also recorded all AIC statistics. We are interested in the unconditional type one error. We specifically chose to conduct 15,000 simulation runs so as to keep computing time within a reasonable limit while also reducing the amount of Monte Carlo standard error to a negligible amount (for looking at type 1 error with $\alpha = 0.05$, Monte Carlo SE will be approximately $0.0018 \approx \sqrt{0.05(1 - 0.05)/15,000}$; see Morris et al. (2019)).

To test the association between X and Y with each of the regression models, we conducted a Wald test to obtain the necessary p-value since in `R`, the p-values in the default summary.glm output are from Wald tests. Moreover, in initial simulations, LRTs performed rather erratically in rare situations when the model was misspecified (e.g., when a Poisson model was fit to ZIP data, convergence issues occasionally occurred).

4. Simulation study results

Analysis under the "correct model" - We first wish to confirm that the models under investigation deliver correct type 1 error when used as intended. In other words, suppose the "correct model" is known a-*priori* and is used regardless of any preliminary testing, would we obtain the desired 0.05 level of type 1 error? See Figure 2 which plots the rejection rates corresponding to this question.

In summary, we see that for data simulated from the Poisson distribution (Figure 2, panel 1), empirical type 1 error is slightly smaller than 0.05 for small sample-size scenarios ($n \leq 100$) and approximately 0.05 otherwise. We also note that for data

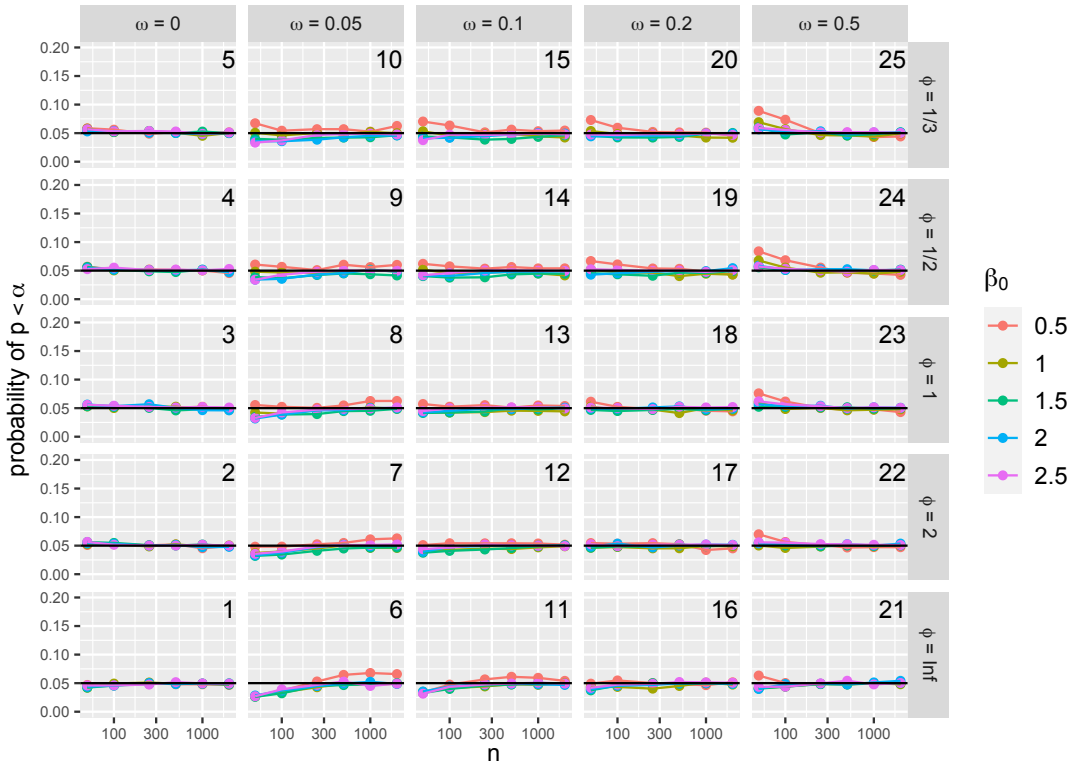

Figure 2. The empirical level of Type 1 error obtained under the "correct" model. For panel 1, the "correct" model is the Poisson GLM; for panels 2-5 the "correct" model is the NB GLM; for panels 6, 11, 16, 21, the "correct" model is the ZIP GLM; and for other panels, the "correct" model is the ZINB GLM.

simulated from the NB distribution (Figure 2, panels 2-5), empirical type 1 error is approximately 0.05 for all $n \geq 100$ and for all β_0. For data simulated from the ZIP distribution (see Figure 2, panels 6, 11, 16, 21), empirical type 1 error can be substantially conservative (i.e., less than 0.05) for small values of n and small values of ω. Finally, for ZINB data, we note that, when n is small, the type 1 error appears to be higher than the advertised rate of 0.05 for some scenarios and less than 0.05 for others. For example, with $n = 100$, $\beta_0 = 0.5$, $\phi = 1/3$, and $\omega = 0.5$, the type 1 error is 0.074, whereas, when $n = 100$, $\beta_0 = 2.5$, $\phi = 2$, and $\omega = 0.05$, the type 1 error is 0.040 (see Figure 2, panels 7 and 25).

We also note that none of the models appear to be "robust" to model misspecification. The Poisson model applied to non-Poisson data leads to very high rejection rates (so high they are often off-the chart in Appendix - Figure 10). The ZIP model also performs poorly when applied to non-ZIP data (see Appendix - Figure 11), as does the NB model when applied to non-NB data (see Appendix - Figure 12), and the ZINB model (see Appendix - Figure 13) when applied to non-ZINB data (specifically when applied to Poisson data and NB data).

As such, it seems inadvisable to recommend simply fitting a ZIP or ZINB to Poisson data if one is uncertain about the possibility of zero-inflation, or overdispersion. As the sample size, n, increases (and as β_0 decreases), the type 1 error rates obtained when the ZIP and ZINB models are fit to Poisson data increase well beyond 0.05 (see Figures 11 and 13, panel 1). This unexpected result may be due to the fact that these models are testing a null hypothesis that lies on the boundary of the parameter space (i.e., $\omega = 0$).

Preliminary testing - The next question is "how often do the preliminary tests reject their null hypotheses?" We also wish to determine how often the preliminary testing scheme successfully identifies the "correct" model.

Let us first consider the D&L score test (see Appendix - Figure 7) and specifically scenarios with $\omega = 0$ and $\phi < \infty$, i.e., scenarios with overdispersion but no structural zero-inflation. With the exception of the small sample-size scenarios ($n \leq 100$) with

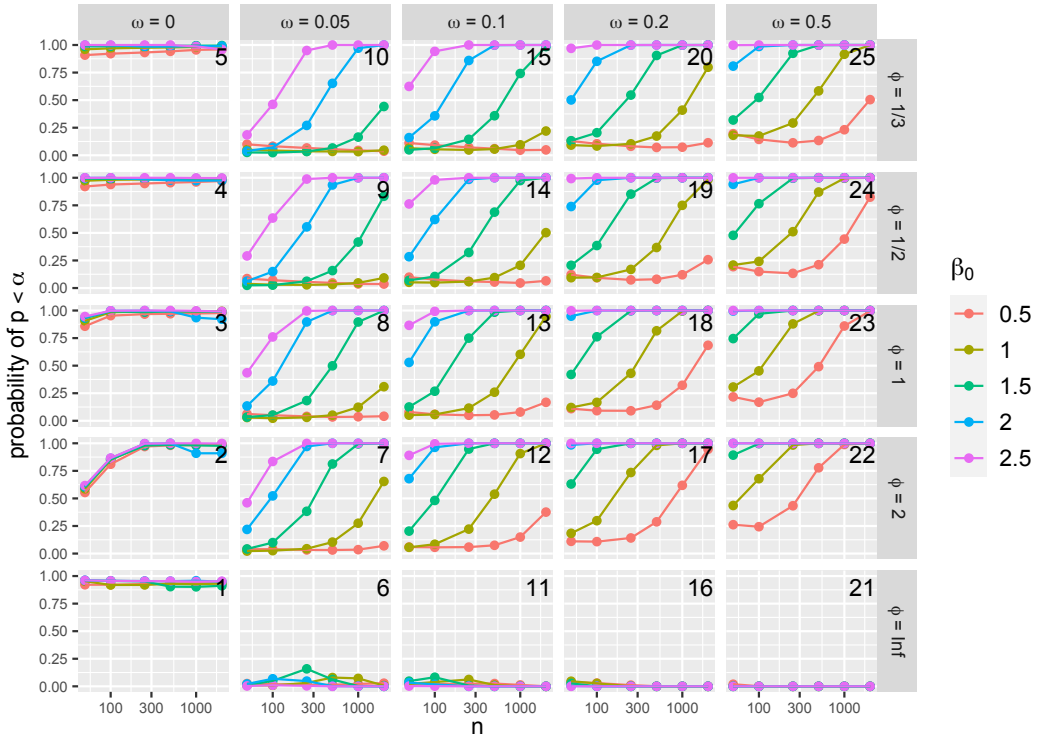

Figure 3. The probability of selecting the "correct model" after following the seven step testing scheme outlined in Section 3.

a small amount of overdispersion, the D&L test correctly rejects the null hypothesis of no overdispersion for the vast majority of cases (Appendix - Figure 7, panels 2-5). For all cases with $\phi = \infty$ and $\omega = 0$, the D&L test appears to show approximately correct type 1 error, with rejection rates ranging from 0.0368 to 0.0514 (see Figure 7, panel 1). However, when $\phi = \infty$ and $\omega > 0$, the D&L test will often reject the null hypothesis of no overdispersion; see Figure 7, panels 6, 11, 16, 21. The rate of rejection increases with increasing sample size, with increasing ω, and with increasing β_0. Strictly speaking, rejection in these cases is correct since an excess of zeros ($\omega > 0$) does contribute to overdispersion. However, it must be noted that using the NB model for overdispersion when the underlying issue is zero-inflation is not appropriate and can yield biased parameter estimates; see Harrison (2014). When the NB model is fit to ZIP data, we record type 1 error rates either much too low or much too high, depending on ω, β_0, and n; see Figure 12, panels 6, 11, 16, and 21.

Now let us discuss the Vuong test for zero-inflation. See Appendix - Figures 8 and 9 for the Vuong test results. Note that the "Poisson vs. ZIP" Vuong test will often

reject the null of no zero-inflation for NB data (when $\phi < \infty$ and $\omega = 0$; Appendix - Figure 8, panels 2-5). In contrast, the "NB vs. ZINB" Vuong test will rarely reject the null of no zero-inflation for NB data (Appendix - Figure 9, panels 2-5). In this way, the Vuong test acts as a second-line defense against erroneously selecting the Poisson model. If the D&L score test fails to select the NB model in Step 1, the 'Poisson vs. ZIP" Vuong test will be used in Step 3 and in many cases (particularly when n and β_0 are large) will reject the Poisson model in favour of the ZIP. The ZIP model, when used for NB, is not ideal, but is definitely preferable to the Poisson model; compare Appendix - Figures 10 and 11, panels 2-5. For example, when there is only a modest amount of overdispersion, with $\phi = 1$ and $\omega = 0$, the type 1 error rate obtained with the Poisson GLM is above 0.15 for all n and β_0 values (see Figure 10, panel 3). In contrast, for the same data, the type 1 error rate obtained with the ZIP GLM is well bellow 0.15 for all n and β_0 values (see Figure 11, panel 3).

Overall, the probability that the preliminary seven-step testing scheme selects the "correct" model depends highly on β_0, ω, ϕ, and n, see Figure 3. With Poisson data, if each of the diagnostic tests were truly independent (and each had a $\alpha = 0.05$ type 1 error rate), then the probability of selecting the "correct" model should be 90.25% ($= 0.95 \times 95\%$); see Figure 1. The numbers we obtain from the simulation study range from 90.21% to 96.19%.

For the ZIP data scenarios ($\omega > 0$, $\phi = \infty$; Figure 3, panels 6, 11, 16, 21), the "incorrect" ZINB model is chosen in a majority of cases. This may not necessarily lead to type 1 error inflation since the "incorrect" ZINB model is often conservative when applied to ZIP data; see Appendix - Figure 13. For ZINB data scenarios (i.e., when $\omega > 0$ and $\phi < \infty$), in cases when the ZINB model is not selected, it is most likely that the NB model is selected instead. This might not necessarily lead to type 1 error inflation since misspecified NB model appears to maintain a type 1 error rate at or bellow the advertised rate in many of these situations (specifically when $\phi < 2$ and $\omega < 0.2$); see Appendix - Figure 12.

Post-testing unconditional type 1 error - Our main question of interest is whether or not the null hypotheses of no association between X and Y is rejected at the desired 0.05 significance level when following the entire seven-step procedure outlined in Section 3. The corresponding rejection rates are plotted in Figure 4. Table 1 list rejection rates and model selection rates for a select number of scenarios. Let us consider the results for each distribution.

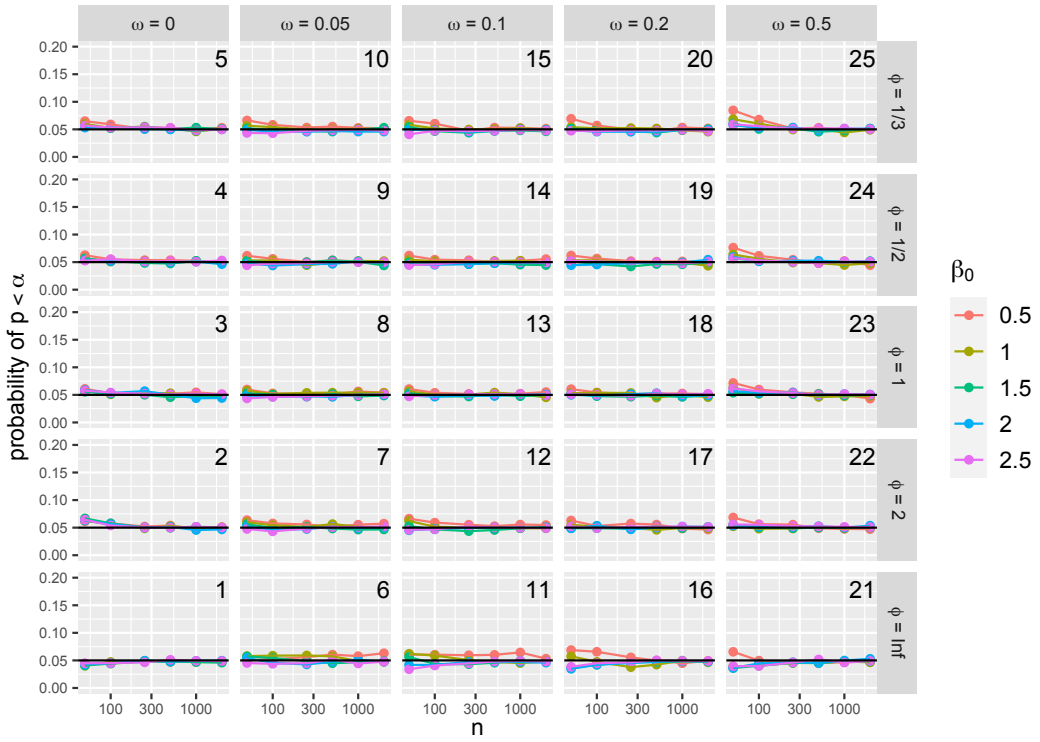

Figure 4. Type 1 error obtained following the seven step testing scheme outlined in Section 3.

First, for data simulated from the Poisson distribution (Figure 4, panel 1), empirical type 1 error appears to be unaffected by model selection bias. This is due to the fact that incorrect models are rarely selected, even when sample sizes are small (see Figure 3, panel 1). Consider two specific scenarios:

- Scenario "3" ($n = 250$, $\beta_0 = 0.5$, $\phi = \infty$, and $\omega = 0$) - When $\beta_0 = 0.5$ and $n = 250$, the Poisson model is correctly selected in approximately 93% of cases while the NB and ZIP models are selected in about 5% and 2% of cases, respectively. Numbers in the top right-hand corner of each node in Figure 1 indicate the expected number of datasets (out of a total of 100) to reach each outcome if the

	Poisson GLM	ZIP GLM	NB GLM	ZINB GLM
Scenario "3"				
($n = 250$, $\beta_0 = 0.5$, $\phi = \infty$, $\omega = 0$; Poisson)				
Pr(reject H_0)	0.05	0.04	0.05	0.04
Pr(reject $H_0 \mid M$ selected by tests)	0.05	0.04	0.02	0.00
Pr(M selected by tests)	0.93	0.02	0.05	0.00
Pr(reject $H_0 \mid M$ has lowest AIC)	0.05	0.16	0.02	0.04
Pr(M has lowest AIC)	0.83	0.11	0.05	0.01
Scenario "6"				
($n = 2,000$, $\beta_0 = 0.5$, $\phi = \infty$, $\omega = 0$; Poisson)				
Pr(reject H_0)	0.05	0.07	0.05	0.06
Pr(reject $H_0 \mid M$ selected by tests)	0.05	0.09	0.04	0.30
Pr(M selected by tests)	0.93	0.02	0.05	0.00
Pr(reject $H_0 \mid M$ has lowest AIC)	0.05	0.30	0.04	0.13
Pr(M has lowest AIC)	0.83	0.10	0.06	0.01
Scenario "36"				
($n = 2,000$, $\beta_0 = 0.5$, $\phi = 2$, $\omega = 0$; NB)				
Pr(reject H_0)	0.11	0.08	0.05	0.07
Pr(reject $H_0 \mid M$ selected by tests)	–	–	0.05	0.12
Pr(M selected by tests)	0.00	0.00	0.98	0.02
Pr(reject $H_0 \mid M$ has lowest AIC)	–	–	0.05	0.24
Pr(M has lowest AIC)	0.00	0.00	0.87	0.13
Scenario "43"				
($n = 50$, $\beta_0 = 1.5$, $\phi = 2$, $\omega = 0$; NB)				
Pr(reject H_0)	0.10	0.04	0.06	0.02
Pr(reject $H_0 \mid M$ selected by tests)	0.11	0.00	0.04	0.18
Pr(M selected by tests)	0.40	0.00	0.60	0.00
Pr(reject $H_0 \mid M$ has lowest AIC)	0.10	0.09	0.04	0.03
Pr(M has lowest AIC)	0.33	0.08	0.56	0.03
Scenario "182"				
($n = 100$, $\beta_0 = 0.5$, $\phi = 2$, $\omega = 0.05$; ZINB)				
Pr(reject H_0)	0.12	0.08	0.05	0.05
Pr(reject $H_0 \mid M$ selected by tests)	0.12	0.29	0.05	0.14
Pr(M selected by tests)	0.09	0.00	0.87	0.04
Pr(reject $H_0 \mid M$ has lowest AIC)	0.12	0.13	0.05	0.14
Pr(M has lowest AIC)	0.06	0.19	0.67	0.08
Scenario "302"				
($n = 100$, $\beta_0 = 0.5$, $\phi = \infty$, $\omega = 0.1$; ZIP)				
Pr(reject H_0)	0.06	0.05	0.05	0.04
Pr(reject $H_0 \mid M$ selected by tests)	0.07	0.17	0.03	0.15
Pr(M selected by tests)	0.71	0.02	0.25	0.02
Pr(reject $H_0 \mid M$ has lowest AIC)	0.06	0.10	0.03	0.07
Pr(M has lowest AIC)	0.51	0.31	0.16	0.02

Table 1. Rejection rates and model selection rates for a number of selected scenarios from the simulation study. These numbers can be used to calculate the overall unconditional type 1 error rates. For example, for Scenario "302", the type 1 error obtained after model selection via AIC is 0.071 ($=0.06 \times 0.51 + 0.10 \times 0.31 + 0.03 \times 0.16 + 0.07 \times 0.02$); and the type 1 error obtained after model selection via sequential score tests is 0.060 ($=0.07 \times 0.71 + 0.17 \times 0.02 + 0.03 \times 0.25 + 0.15 \times 0.02$).

data was Poisson (with $\beta_X = 0$), and each of the tests were truly independent (with a $\alpha = 0.05$ type 1 error rate). The numbers in parentheses correspond to results from the simulation study for this scenario.

For those datasets directed to the NB and ZIP models, the null hypothesis of no association between X and Y is rejected with probability of 0.021 ($=0.10/(4.63+0.10)$) and 0.044 ($=0.11/(2.34+0.11)$), respectively. As such, model selection bias, in this case, has the innocuous effect of ever so slightly low-

ering the type 1 error level: the Poisson GLM fit to this data provides in a type 1 error rate of 0.050, whereas the unconditional type 1 error rate obtained after following the seven-step procedure is 0.049 (=0.0471+0.0011+0.0010+0.0000).

- Scenario "6" ($n = 2,000$, $\beta_0 = 0.5$, $\phi = \infty$, and $\omega = 0$) - When $\beta_0 = 0.5$ and $n = 2,000$, the Poisson model is correctly selected in approximately 93% of cases while the NB and ZIP models are selected in about 5% and 2% of cases, respectively. While the NB model is conservative for this data (Pr(reject H_0|NB model selected by tests) = 0.037), the ZIP model is not (Pr(reject H_0|ZIP model selected by tests) = 0.089). However, the impact is indiscernible: the unconditional type 1 error rate obtained after following the seven-step procedure is 0.047.

Second, for data simulated from the ZIP distribution (i.e., when $\omega > 0$ and $\phi = \infty$), the "incorrect" ZINB model is almost always selected due to the fact that the model selection procedure tests for zero-inflation only after first testing for overdispersion. However, the type 1 error under this "incorrect" ZINB model is, for most scenarios, not substantially higher than the advertised 0.05 rate, (see Appendix - Figure 13, panels 6, 11, 16, 21). There are, however, exceptions where model selection bias is apparent. Consider, for example, scenario "302":

- Scenario "302" ($n = 100$, $\beta_0 = 0.5$, $\phi = \infty$ and $\omega = 0.1$) - The unconditional type 1 error obtained after following the seven step procedure is 0.060 (see Figure 4, panel 11). Note that, when applied to the data ignoring the results of the diagnostic tests, both the ZIP and the ZINB models demonstrate conservative rejection rates (of 0.047 and 0.040, respectively). However, amongst the simulated datasets for which the ZIP model is selected (by the D&L and Vuong diangostic tests), the ZIP model has a rejection rate of 0.168. Amongst the simulated datasets for which the ZINB model is selected, the ZINB model has a rejection rate of 0.154. This clearly shows that the diagnostic tests (the D&L and Vuong tests) and the subsequent hypothesis tests ($H_0 : \beta_X = \gamma_X = 0$) are not independent of one another. In this instance, the D&L test will not only screen for overdispersion, but will also direct the data towards a model that is

more likely to reject $H_0 : \beta_X = \gamma_X = 0$, thereby inflating the type 1 error.

With data simulated from the NB distribution (i.e., when $\phi < \infty$ and $\omega = 0$; see Figure 4, panels 2-5), we see that model selection bias can lead to modest type 1 error inflation when n is small. When sample sizes are sufficiently large, there is little evidence of any substantial type 1 error inflation caused by model selection bias. Consider for example "Scenario 43":

- Scenario "43" ($n = 50$, $\beta_0 = 1.5$, $\phi = 2$ and $\omega = 0$) - The unconditional type 1 error obtained after following the seven step procedure is 0.067 (see Figure 4, panel 2), whereas the type 1 error obtained with the "correct" NB model is 0.056. This inflation is due to the fact that, for this data, there is a 40% probability of selecting the Poisson model following the seven-step procedure and that Pr(reject H_0|Poisson model is selected by tests) = 0.11; see Table 1.

Finally, consider data simulated from the ZINB distribution (i.e., when $\phi < \infty$ and $\omega > 0$; see Figure 4, panels 7-10, 12-15, 17-20, 22-25). We see type 1 error rates much higher than 0.05 for some scenarios (e.g., when n is small and ω is large). For example, consider scenario "182":

- Scenario "182" ($n = 100$, $\beta_0 = 0.5$, $\phi = 2$ and $\omega = 0.05$) - The unconditional type 1 error obtained after following the seven step procedure is 0.058 (see Figure 4, panel 7), whereas the type 1 error obtained with the "correct" ZINB model is 0.048. Note that, when applied to the data ignoring the results of the diagnostic tests, both the NB and the ZINB models demonstrate appropriate rejection rates (of 0.051 and 0.048, respectively; see Table 1). However, amongst the simulated datasets for which the ZINB model is selected (by the D&L and Vuong diagnostic tests) the ZINB model has a rejection rate of 0.142. This clearly shows that, to the detriment of the type 1 error rate, the diagnostic tests and the subsequent hypothesis test for $H_0 : \beta_X = \gamma_X = 0$ are not independent.

AIC model selection - We also investigated model selection using the AIC. We were curious as to how often the "correct" model is the model with the lowest AIC.

Figure 5 plots the results. We see that the probability that the AIC statistic selects the "correct" model depends highly on β_0, ω, ϕ, and n. Overall, across all scenarios we considered, the AIC selected the correct model for 77% of datasets whereas the seven-step model selection based on score tests selects the correct model for 58% of datasets.

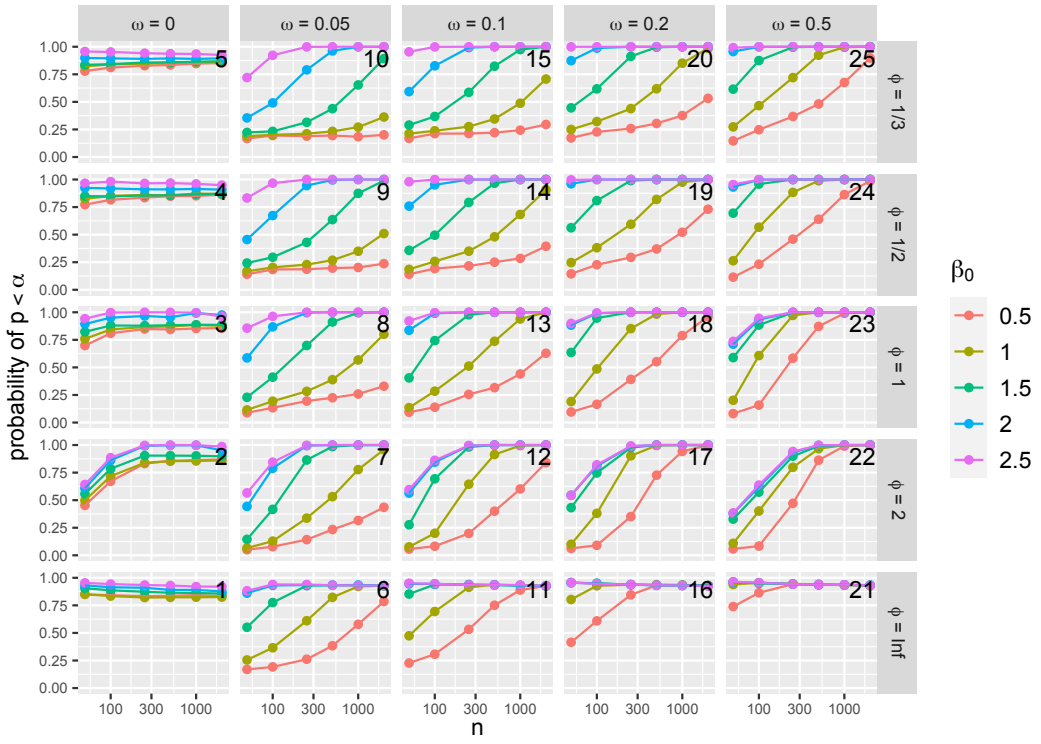

Figure 5. The probability that the "correct" model is the one with the lowest AIC.

More specifically, for the NB data scenarios ($\omega = 0$, $\phi < \infty$; Figure 5, panels 2-5), the "correct" NB model is chosen using the AIC in a large majority of cases for most scenarios. In contrast, for ZIP data (i.e., when $\omega > 0$ and $\phi = \infty$), the AIC is less capable of determining the "correct model" when β_0, n, and ω are small. For ZINB data scenarios (i.e., when $\omega > 0$ and $\phi < \infty$), the probability of selecting the correct model using the AIC ranges substantially and increases (somewhat predictably) with increasing n and increasing β_0.

We also wish to determine whether or not the null hypotheses of no association between X and Y is rejected at the 0.05 significance level when following model selection via AIC. Figure 6 shows that, when β_0 is small, there are several scenarios

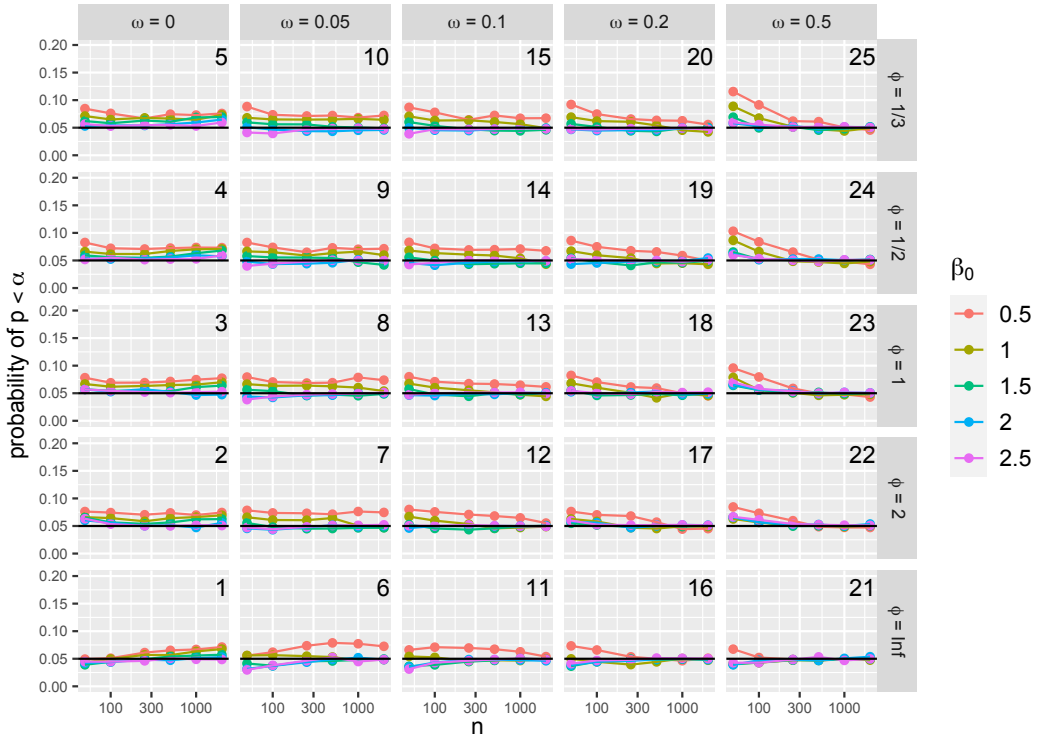

Figure 6. Type 1 error obtained from model with the lowest AIC.

in which the unconditional type 1 error is much too high. Perhaps most surprisingly, with Poisson data (i.e., scenarios with $\omega = 0$ and $\phi = \infty$), when $\beta_0 = 0.5$, the unconditional type 1 error increases with increasing n, from 0.049 to 0.071 (see Figure 6, panel 1). Consider again Scenario "3" ($n = 250$, $\beta_0 = 0.5$, $\phi = \infty$, and $\omega = 0$) and Scenario "6" ($n = 2,000$, $\beta_0 = 0.5$, $\phi = \infty$, and $\omega = 0$); see Table 1. When $n = 250$ (Scenario "3"), the probability that the AIC incorrectly selects the ZIP GLM is 0.11, and Pr(M has lowest AIC) = 0.16. The unconditional type 1 error rate = 0.06 (=$0.05 \times 0.83 + 0.16 \times 0.11 + 0.02 \times 0.05 + 0.04 \times 0.01$). For Scenario "6", the probability that the AIC incorrectly selects the ZIP GLM is 0.10, and Pr(M has lowest AIC) = 0.30. The unconditional type 1 error rate = 0.07 ($= 0.05 \times 0.83 + 0.30 \times 0.10 + 0.04 \times 0.06 + 0.13 \times 0.01$).

With NB data (i.e., scenarios with $\omega = 0$ and $\phi < \infty$), the unconditional type 1 error is also much higher than 0.05, even when n and β_0 are large. This is due to the fact that the ZINB model, when erroneously selected in a minority of cases, reject the null at rates much much higher than 0.05. This particularly true when n is large.

Consider for example, Scenario "36" ($n = 2,000$, $\beta_0 = 0.5$, $\phi = 2$, and $\omega = 0$); see Table 1. Amongst the 87% of datasets for which the AIC correctly selects the NB model, the null hypothesis of no association between X and Y is rejected with probability of exactly 0.050. However, amongst the remaining 13% of datasets for which the ZINB model is erroneously selected, the probability of rejecting the null hypothesis of no association between X and Y is 0.240. As a result the unconditional type 1 error rate is 0.074 ($= 0.240 \times 0.13 + 0.87 \times 0.05$).

In summary, while the AIC is able to select the "correct" model more often than the sequential score testing scheme, there appears to be more potential for type 1 error inflation. How can this be? Consider once again Scenario "302" (with $n = 100$, $\beta_0 = 0.5$, $\phi = \infty$ and $\omega = 0.1$). Following the sequential score tests, the "correct" ZIP model was only selected with a 2% probability. With model selection via AIC, the "correct" ZIP model was selected with a 31% probability. However, in the presence of model selection bias, selecting the "correct" model more often is, somewhat paradoxically, not always preferable. Indeed, for Scenario "302", the unconditional type 1 error obtained after following the seven step procedure is 0.060, whereas the unconditional type 1 error obtained after model selection via AIC is 0.071; see Table 1. We see a similar phenomenon with Scenario "182" (with $n = 100$, $\beta_0 = 0.5$, $\phi = 2$ and $\omega = 0.05$). The "correct" ZINB model is chosen more often with the AIC than with the score tests (8% vs. 4%). However, the type 1 error obtained after model selection via AIC is 0.074, vs. 0.058 after model selection by sequential score tests.

5. Conclusion

If the population distribution is known in advance, model selection bias will not be a problem. If the assumptions required of the Poisson distribution are known to be wrong, alternative models that do not depend on these assumptions can be used and ideally a valid model can be pre-specified prior to obtaining/observing any data. However, outside of a highly controlled laboratory experiment, this may not be realistic. The potentially problematic (and most likely scenario) is when one cannot, with a high

degree of confidence, determine the distributional nature of the data before observing the data. What should be done in these circumstances? Tsou (2006) suggest using a "robust" Poisson regression model "so that one need not worry about the correctness of the Poisson assumption." However, when the distributional assumptions of the Poisson GLM do hold, Tsou (2006) acknowledge that the "robust approach might not be as efficient." Given the potentially immense expense required to obtain data, anyone working in data-driven research will no doubt be reluctant to adopt any approach which compromises statistical power.

Researchers who do not know in advance whether or not there is overdispersion or zero-inflation, might decide to simply use a ZINB as a "safer bet" (Perumean-Chaney et al. 2013) and pay a price in terms of efficiency (Williamson et al. 2007). However, this is problematic. We observed that the ZIP and ZINB models, when fit to ordinary Poisson data, can lead to type 1 error well above the advertised rate when sample sizes are large. (Future work should consider whether hurdle models (Rose et al. 2006) are similarly problematic.) Instead, if there is sufficient data, researchers should proceed with a model selection procedure, ideally one based on efficient score tests.

Our simulation study suggests that, if sample sizes are sufficiently large, there should be no need to worry about model selection bias from a series of sequential score tests. However, when sample sizes are small, our simulation study demonstrated that model selection bias can lead to potentially substantial type 1 error inflation. Model selection based on the AIC cannot be recommended. We observed that, when the true underlying distribution of the data is Poisson, using the AIC to select the "best" model can often lead to substantial type 1 error inflation. Future work should investigate the suitability of other information criteria (e.g., BIC, AICc).

Ignoring the possibility of overdispersion and zero inflation during data analyses can lead to invalid inference. However, if one does not have sufficient power to confidently test for overdispersion and zero inflation, it may be best to simply use a model that can accommodate for these possibilities (e.g., use a robust model) instead of going through a model selection procedure that might inflate the type 1 error. In summary, if one does not have the power to test for distributional assumptions, testing

for distributional assumptions may not be wise. And if one does have a sufficiently large sample size to test for distributional assumptions, testing for distributional assumptions may be very beneficial. Note that our simulation study did not include any covariates and in studies where there are several covariates, it will no doubt be difficult to determine what constitutes a "sufficiently large" sample size. To conclude, be reminded that researchers should always be cautious when interpreting results when n is small (Button et al. 2013). Model selection bias is just one more reason to have a healthy skepticism of NHST based on small sample sizes.

References

Amrhein, V., Greenland, S. & McShane, B. (2019), 'Scientists rise up against statistical significance', *Nature* **567**, 305–307.

Anderson, D. R. (2007), *Model based inference in the life sciences: a primer on evidence*, Springer Science & Business Media.

Bening, V. E. & Korolev, V. Y. (2012), *Generalized Poissonmodels and their applications in insurance and finance*, Walter de Gruyter.

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.

Breiman, L. (1992), 'The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error', *Journal of the American Statistical Association* **87**(419), 738–754.

Brooks, M. E., Kristensen, K., Darrigo, M. R., Rubim, P., Uriarte, M., Bruna, E. & Bolker, B. M. (2019), 'Statistical modeling of patterns in annual reproductive rates', *Ecology* **100**(7), e02706.

Buckland, S. T., Burnham, K. P. & Augustin, N. H. (1997), 'Model selection: an integral part of inference', *Biometrics* pp. 603–618.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S.

& Munafò, M. R. (2013), 'Power failure: why small sample size undermines the reliability of neuroscience', *Nature Reviews Neuroscience* **14**(5), 365–376.

Campbell, H. & Dean, C. (2014), 'The consequences of proportional hazards based model selection', *Statistics in Medicine* **33**(6), 1042–1056.

Chen, Q. & Giles, D. E. (2011), 'Finite-sample properties of the maximum likelihood estimator for the Poissonregression model with random covariates', *Communications in Statistics-Theory and Methods* **40**(6), 1000–1014.

Cox, D. R. (1983), 'Some remarks on overdispersion', *Biometrika* **70**(1), 269–274.

Dean, C. & Lawless, J. F. (1989), 'Tests for detecting overdispersion in Poissonregression models', *Journal of the American Statistical Association* **84**(406), 467–472.

Desmarais, B. A. & Harden, J. J. (2013), 'Testing for zero inflation in count models: Bias correction for the vuong test', *The Stata Journal* **13**(4), 810–835.

Dunn, P. K. & Smyth, G. K. (2018), Generalized linear models: Estimation, *in* 'Generalized Linear Models With Examples in R', Springer, pp. 243–263.

Dushoff, J., Kain, M. P. & Bolker, B. M. (2019), 'I can see clearly now: reinterpreting statistical significance', *Methods in Ecology and Evolution* **10**(6), 756–759.

Fisher, R. A. (1950), 'The significance of deviations from expectation in a Poisson-series', *Biometrics* **6**(1), 17–24.

Fraser, H., Parker, T., Nakagawa, S., Barnett, A. & Fidler, F. (2018), 'Questionable research practices in ecology and evolution', *PloS one* **13**(7), e0200303.

Freckleton, R. (2009), 'The seven deadly sins of comparative analysis', *Journal of Evolutionary Biology* **22**(7), 1367–1375.

Gelman, A. & Loken, E. (2013), 'The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time', *Department of Statistics, Columbia University* .

Greene, W. H. (1994), 'Accounting for excess zeros and sample selection in Poissonand negative binomial regression models'.

Harrison, X. A. (2014), 'Using observation-level random effects to model overdispersion in count data in ecology and evolution', *PeerJ* **2**, e616.

Hayes, A. (2020), 'Using the data twice', https://www.alexpghayes.com/blog/using-the-data-twice/.

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

Hurvich, C. M. & Tsai, C. (1990), 'The impact of model selection on inference in linear regression', *The American Statistician* **44**(3), 214–217.

Kahan, B. C. (2013), 'Bias in randomised factorial trials', *Statistics in Medicine* **32**(26), 4540–4549.

Kelly, C. (2019), 'Rate and success of study replication in ecology and evolution', *PeerJ* **7**(e7654).

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

Lindén, A. & Mäntyniemi, S. (2011), 'Using the negative binomial distribution to model overdispersion in ecological count data', *Ecology* **92**(7), 1414–1421.

Loeys, T., Moerkerke, B., De Smet, O. & Buysse, A. (2012), 'The analysis of zero-inflated count data: Beyond zero-inflated Poissonregression.', *British Journal of Mathematical and Statistical Psychology* **65**(1), 163–180.

Lynch, H. J., Thorson, J. T. & Shelton, A. O. (2014), 'Dealing with under-and over-dispersed count data in life history, spatial, and community ecology', *Ecology* **95**(11), 3173–3180.

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.

Morris, T. P., White, I. R. & Crowther, M. J. (2019), 'Using simulation studies to evaluate statistical methods', *Statistics in Medicine* **38**(11), 2074–2102.

Murtaugh, P. A. (2014), 'In defense of p values', *Ecology* **95**(3), 611–617.

Nelder, J. A. & Wedderburn, R. W. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society: Series A (General)* **135**(3), 370–384.

Nosek, B. A., Spies, J. R. & Motyl, M. (2012), 'Scientific utopia II. restructuring incentives and practices to promote truth over publishability', *Perspectives on Psychological Science* **7**(6), 615–631.

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.

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

Richards, S. A. (2008), 'Dealing with overdispersed count data in applied ecology', *Journal of Applied Ecology* **45**(1), 218–227.

Ridout, M., Hinde, J. & Demétrio, C. G. (2001), 'A score test for testing a zero-inflated Poissonregression model against zero-inflated negative binomial alternatives', *Biometrics* **57**(1), 219–223.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D. & Ripley, M. B. (2013), 'Package mass', *Cran R* **538**.

Rochon, J., Gondan, M. & Kieser, M. (2012), 'To test or not to test: Preliminary assessment of normality when comparing two independent samples', *BMC Medical Research Methodology* **12**(1), 81.

Rose, C. E., Martin, S. W., Wannemuehler, K. A. & Plikaytis, B. D. (2006), 'On the

use of zero-inflated and hurdle models for modeling vaccine adverse event count data', *Journal of Biopharmaceutical Statistics* **16**(4), 463–481.

Shuster, J. J. (2005), 'Diagnostics for assumptions in moderate to large simple clinical trials: do they really help?', *Statistics in Medicine* **24**(16), 2431–2438.

Stephens, P. A., Buskirk, S. W., Hayward, G. D. & Martinez Del Rio, C. (2005), 'Information theory and hypothesis testing: a call for pluralism', *Journal of Applied Ecology* **42**(1), 4–12.

Tsou, T.-S. (2006), 'Robust Poissonregression', *Journal of Statistical Planning and Inference* **136**(9), 3173–3186.

Uusipaikka, E. (2008), *Confidence intervals in generalized regression models*, Chapman and Hall/CRC.

Van den Broek, J. (1995), 'A score test for zero inflation in a Poissondistribution', *Biometrics* pp. 738–743.

Venzon, D. & Moolgavkar, S. (1988), 'A method for computing profile-likelihood-based confidence intervals', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **37**(1), 87–94.

Ver Hoef, J. M. & Boveng, P. L. (2007), 'Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data?', *Ecology* **88**(11), 2766–2772.

Vuong, Q. H. (1989), 'Likelihood ratio tests for model selection and non-nested hypotheses', *Econometrica: Journal of the Econometric Society* pp. 307–333.

Walters, G. D. (2007), 'Using Poissonclass regression to analyze count data in correctional and forensic psychology: A relatively old solution to a relatively new problem', *Criminal Justice and Behavior* **34**(12), 1659–1674.

Wedderburn, R. W. (1974), 'Quasi-likelihood functions, generalized linear models, and the gaussnewton method', *Biometrika* **61**(3), 439–447.

Wells, C. S. & Hintze, J. M. (2007), 'Dealing with assumptions underlying statistical tests', *Psychology in the Schools* **44**(5), 495–502.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B. & Freckleton, R. P. (2006), 'Why do we still use stepwise modelling in ecology and behaviour?', *Journal of Animal Ecology* **75**(5), 1182–1189.

Whittingham, M. J., Swetnam, R. D., Wilson, J. D., Chamberlain, D. E. & Freckleton, R. P. (2005), 'Habitat selection by yellowhammers emberiza citrinella on lowland farmland at two spatial scales: implications for conservation management', *Journal of applied ecology* **42**(2), 270–280.

Williams, M. N. & Albers, C. (2019), 'Dealing with distributional assumptions in preregistered research', *Meta-Psychology* **3**.

Williamson, J. M., Lin, H., Lyles, R. H. & Hightower, A. W. (2007), 'Power calculations for zip and zinb models', *Journal of Data Science* **5**(4), 519–534.

Wilson, P. (2015), 'The misuse of the vuong test for non-nested models to test for zero-inflation', *Economics Letters* **127**, 51–53.

Xu, L., Paterson, A. D., Turpin, W. & Xu, W. (2015), 'Assessment and selection of competing models for zero-inflated microbiome data', *PloS one* **10**(7), e0129606.

Yang, Z., Hardin, J. W. & Addy, C. L. (2010), 'Score tests for zero-inflation in overdispersed count data', *Communications in StatisticsTheory and Methods* **39**(11), 2008–2030.

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

Zoltowski, D. & Pillow, J. W. (2018), Scaling the Poissonglm to massive neural datasets through polynomial approximations, *in* 'Advances in Neural Information Processing Systems', pp. 3517–3527.

Zorn, C. J. (1998), 'An analytic and empirical examination of zero-inflated and hurdle Poissonspecifications', *Sociological Methods & Research* **26**(3), 368–400.

Zuur, A. F., Ieno, E. N. & Elphick, C. S. (2010), 'A protocol for data exploration to avoid common statistical problems', *Methods in Ecology and Evolution* **1**(1), 3–14.

6. Appendix

Let us briefly review the Dean & Lawless (1989) score test for overdispersion and the Vuong test for zero-inflation.

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

$$T_1 = \sum_{i=1}^{n} \left\{ \left(y_i - \hat{\lambda}_i \right)^2 - y_i \right\} / \left(2 \sum_{i=1}^{n} \hat{\lambda}_i^2 \right)^{1/2} \tag{8}$$

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 Vuong test for zero-inflation - The Vuong test statistic is calculated as follows:

$$V = \frac{\sum_{i=1}^{n} (\log dL_i)}{\sqrt{n} \cdot \sqrt{\sum_{i=1}^{n} ((\log dL_i - \sum_{i=1}^{n} (\log dL_i)/n)^2/(n-1))}}, \tag{9}$$

where, if the Poisson model is compared to it's zero-inflated counterpart, the ZIP model, we define: $\log dL_i = \log(Pr_{ZIP}(Y_i = y_i | \widehat{\omega}_i, \widehat{\lambda}_i)) - \log(Pr_{Pois}(Y_i = y_i | \widehat{\lambda}_i))$. If the NB model is compared to the ZINB model, we define: $\log dL_i = \log(Pr_{ZINB}(Y_i = y_i | \widehat{\nu}, \widehat{\omega}_i, \widehat{\lambda}_i)) - \log(Pr_{NB}(Y_i = y_i | \widehat{\nu}, \widehat{\lambda}_i))$.

The V statistic, under the null, will follow the Normal distribution and a p-value is calculated as: p-value $= 1 - P_N(|V|)$. Note that Desmarais & Harden (2013) have suggested an adjustment to the Vuong test which, for larger samples, may have greater efficiency. Also, note that the Vuong test for zero-inflation, while widely used in practice, is somewhat controversial, see Wilson (2015).

7. R-Code: Models and score tests

```
library("pscl", "lmtest")
######################
### MODELS:

######################
### Poisson model ###
PoissonGLM <- glm(y ~ newX, family = poisson)
Poisson_pval <- waldtest(PoissonGLM)[ "Pr(>F)"][2,]
Poisson_AIC <- AIC(PoissonGLM)

######################
### NB model ###
NB_AIC<-Inf
NB_pval <- 0.99
tryCatch({
NB_mod  <- glm.nb(y ~ newX)
NB_pval <- waldtest(NB_mod)[ "Pr(>F)"][2,]
NB_AIC  <- AIC(NB_mod)
},
 error=function(e){})

######################
### ZIP model ###
ZIP_AIC <- Inf
zip_mod <- 99
ZIP_pval <- 0.99
zip_modNA <- TRUE

tryCatch({
zip_mod   <- zeroinfl(y ~ newX|newX, dist = "poiss")
zip_modNA <- sum(is.na(unlist((summary(zip_mod))$coefficients)))>0
},
error=function(e){})

if(is.double(zip_mod)| zip_modNA){
tryCatch({
zip_mod <- zeroinfl(y ~ newX|newX , dist = "poiss", EM=TRUE)},
error=function(e){})}
```

```
tryCatch({
ZIP_pval <- waldtest(zip_mod)[ "Pr(>Chisq)"][2,]
ZIP_AIC <- AIC(zip_mod)
 },
 error=function(e){})

######################
### ZINB model ###
ZINP_AIC  <- Inf
zinb_mod<-99
ZINB_pval <- 0.99
zinb_modNA <- TRUE

tryCatch({
zinb_mod <- zeroinfl(y ~ newX|newX , dist = "negbin")
zinb_modNA <- sum(is.na(unlist((summary(zinb_mod))$coefficients)))>0
},
error=function(e){})

if(is.double(zinb_mod) | zinb_modNA){
tryCatch({
zinb_mod <- zeroinfl(y ~ newX|newX  , dist = "negbin", EM=TRUE); summary(zinb_mod)
},
 error=function(e){})}


tryCatch({
ZINB_pval<-waldtest(zinb_mod)[ "Pr(>Chisq)"][2,]
dimK<-dim(summary(zinb_mod)$coefficients$count)[1]
ZINP_AIC <-AIC(zinb_mod)
},
error=function(e){})

######################
### SCORE TESTS:

#######
## vuong_f test :  compare model1 to model2 (modified from "pscl" package)
#######
```

```
vuong_f<-function (m1, m2, digits = getOption("digits"))
{
    m1y <- m1$y
    m2y <- m2$y
    m1n <- length(m1y)
    m2n <- length(m2y)
    if (m1n == 0 | m2n == 0)
        stop("Could not extract dependent variables from models.")
    if (m1n != m2n)
        stop(paste("Models appear to have different numbers of observations.\n",
            "Model 1 has ", m1n, " observations.\n", "Model 2 has ",
            m2n, " observations.\n", sep = ""))
    if (any(m1y != m2y)) {
        stop(paste("Models appear to have different values on dependent variables.\n"))
    }
    p1 <- predprob(m1)
    p2 <- predprob(m2)
    if (!all(colnames(p1) == colnames(p2))) {
        stop("Models appear to have different values on dependent variables.\n")
    }
    whichCol <- match(m1y, colnames(p1))
    whichCol2 <- match(m2y, colnames(p2))
    if (!all(whichCol == whichCol2)) {
        stop("Models appear to have different values on dependent variables.\n")
    }
    m1p <- rep(NA, m1n)
    m2p <- rep(NA, m2n)
    for (i in 1:m1n) {
        m1p[i] <- p1[i, whichCol[i]]
        m2p[i] <- p2[i, whichCol[i]]
    }
    k1 <- length(coef(m1))
    k2 <- length(coef(m2))
    lm1p <- log(m1p)
    lm2p <- log(m2p)
    m <- lm1p - lm2p
    bad1 <- is.na(lm1p) | is.nan(lm1p) | is.infinite(lm1p)
    bad2 <- is.na(lm2p) | is.nan(lm2p) | is.infinite(lm2p)
    bad3 <- is.na(m) | is.nan(m) | is.infinite(m)
```

```
    bad <- bad1 | bad2 | bad3
    neff <- sum(!bad)
    if (any(bad)) {
        cat("NA or numerical zeros or ones encountered in fitted probabilities\n")
        cat(paste("dropping these", sum(bad), "cases, but proceed with caution\n"))
    }
    aic.factor <- (k1 - k2)/neff
    bic.factor <- (k1 - k2)/(2 * neff) * log(neff)
    v <- rep(NA, 3)
    arg1 <- matrix(m[!bad], nrow = neff, ncol = 3, byrow = FALSE)
    arg2 <- matrix(c(0, aic.factor, bic.factor), nrow = neff,
        ncol = 3, byrow = TRUE)
    num <- arg1 - arg2
    s <- apply(num, 2, sd)
    numsum <- apply(num, 2, sum)
    v <- numsum/(s * sqrt(neff))
    names(v) <- c("Raw", "AIC-corrected", "BIC-corrected")
    pval <- rep(NA, 3)
    msg <- rep("", 3)
    for (j in 1:3) {
        if (v[j] > 0) {
            pval[j] <- 1 - pnorm(v[j])
            msg[j] <- "model1 > model2"
        }
        else {
            pval[j] <- pnorm(v[j])
            msg[j] <- "model2 > model1"
        }
    }
    out <- data.frame(v, msg,(pval))
    names(out) <- c("Vuong z-statistic", "H_A", "p-value")

    return(out)
}
####################

############
### the D&L score test for overdispersion:
lambdahat <- yhat <- predict(PoissonGLM, type="response")
T_1 <- sum((y-lambdahat)^2 - y)/ sqrt(2*sum(lambdahat^2))
```

```
LRT_pval <- DLtest_pval <-  pnorm(T_1, lower.tail=FALSE)

###########
### the Vuong test for zero-inflation:
vuong_P_ZIP_pval <- vuong_NB_ZINB_pval <- 1

if(exists("zip_mod") & sum(y==0)>1 ){
tryCatch({
vv_P_ZIP<-(vuong_f(PoissonGLM, zip_mod))
vuong_P_ZIP_pval <-as.numeric(as.character(vv_P_ZIP[1,3]))
},
 error=function(e){})}

if(exists("zinb_mod")  & sum(y==0)>1 ){
tryCatch({
vv_NB_ZINB<-(vuong_f(NB_mod, zinb_mod))
vuong_NB_ZINB_pval <-as.numeric(as.character(vv_NB_ZINB[1,3]))
},
 error=function(e){})}
#########################################################
```

8. Appendix Figures

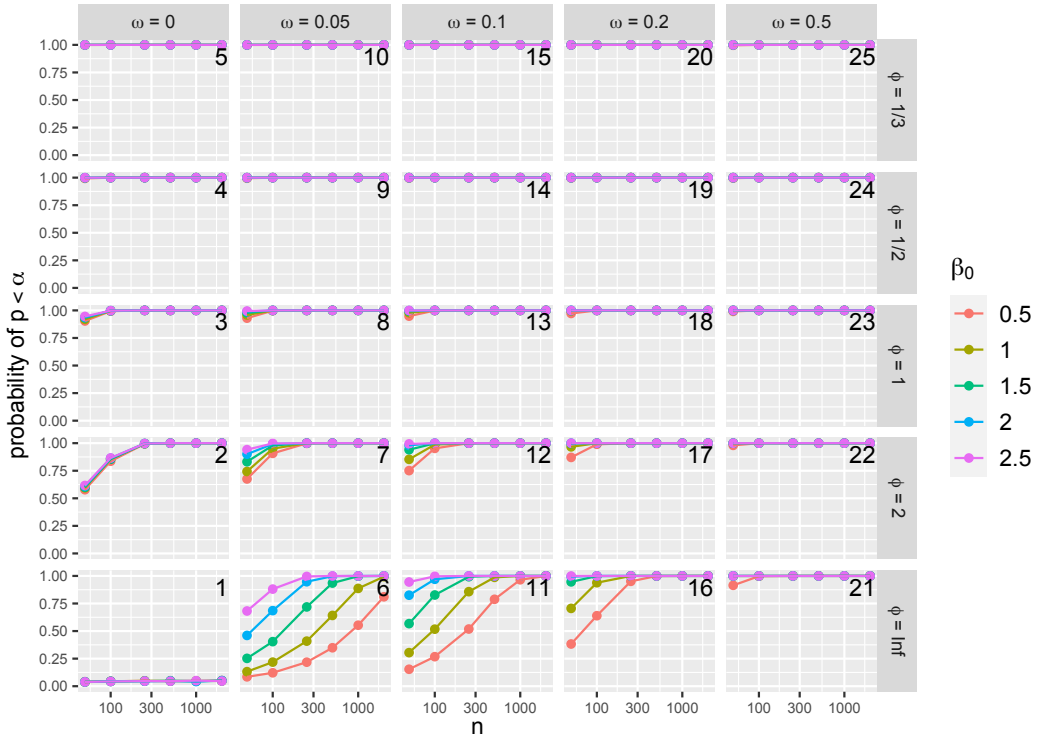

Figure 7. Probability that the D&L test rejects the null hypothesis that there is no overdispersion.

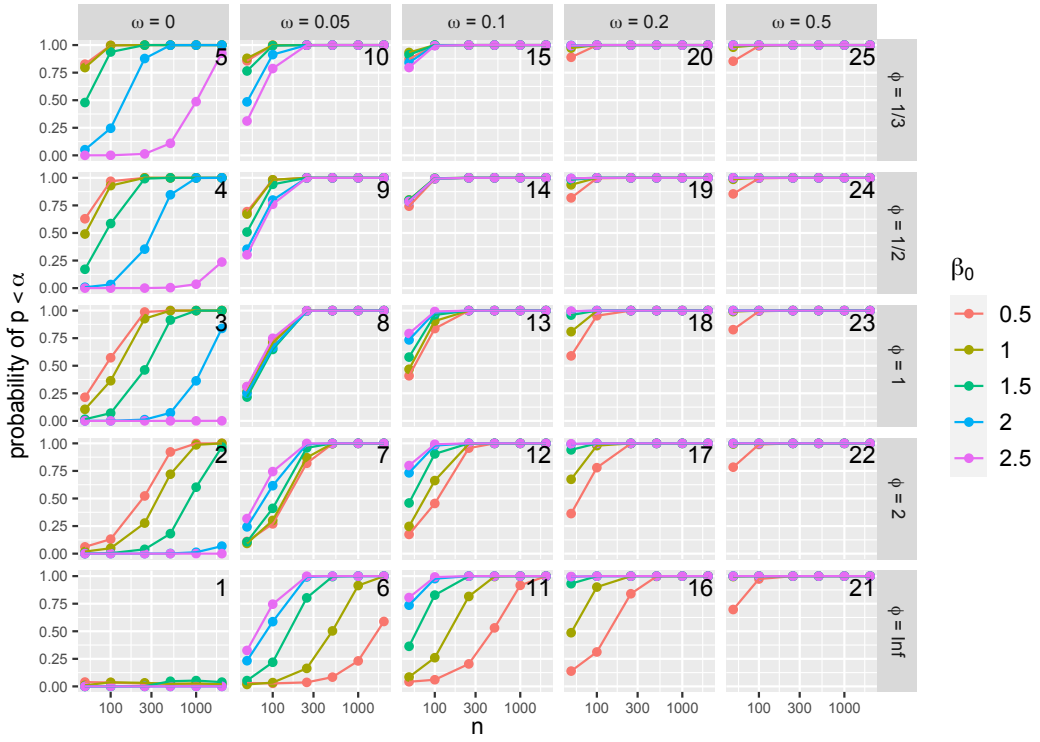

Figure 8. Probability that the Vuong test rejects the null hypothesis that there is no zero-inflation, comparing the Poisson model to the ZIP model.

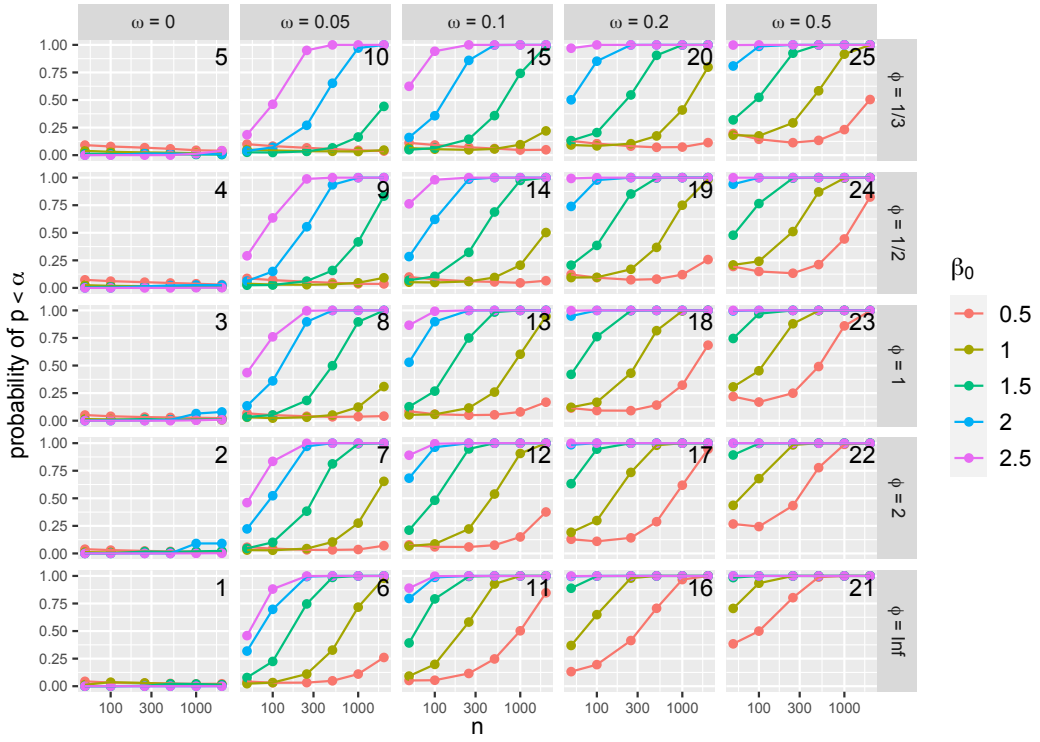

Figure 9. Probability that the Vuong test rejects the null hypothesis that there is no zero-inflation, comparing the NB model to the ZINB model.

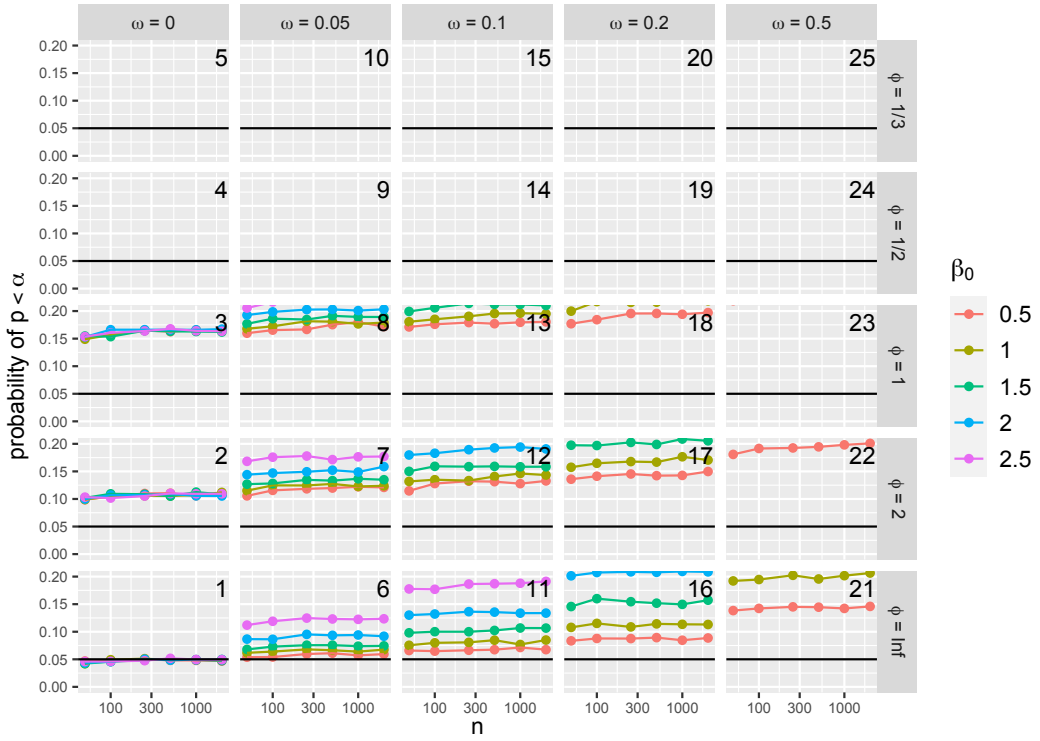

Figure 10. Probability that the Poisson model rejects the null hypothesis of $H_0 : \beta_X = 0$.

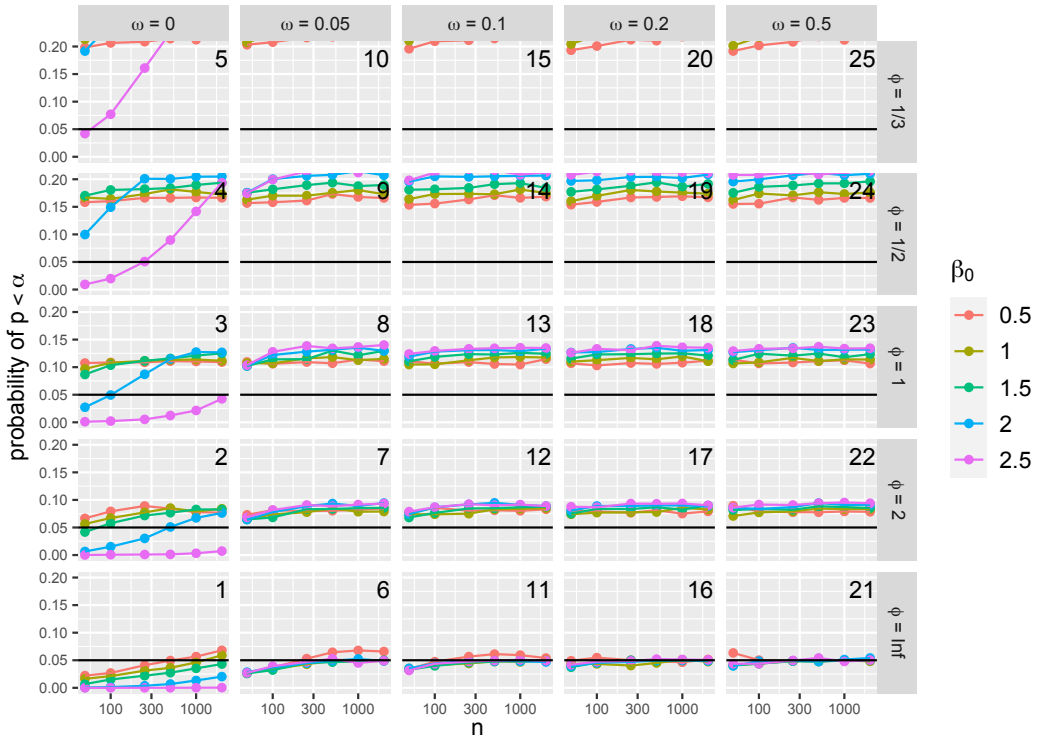

Figure 11. Probability that the ZIP model rejects the null hypothesis of $H_0 : \beta_X = 0$.

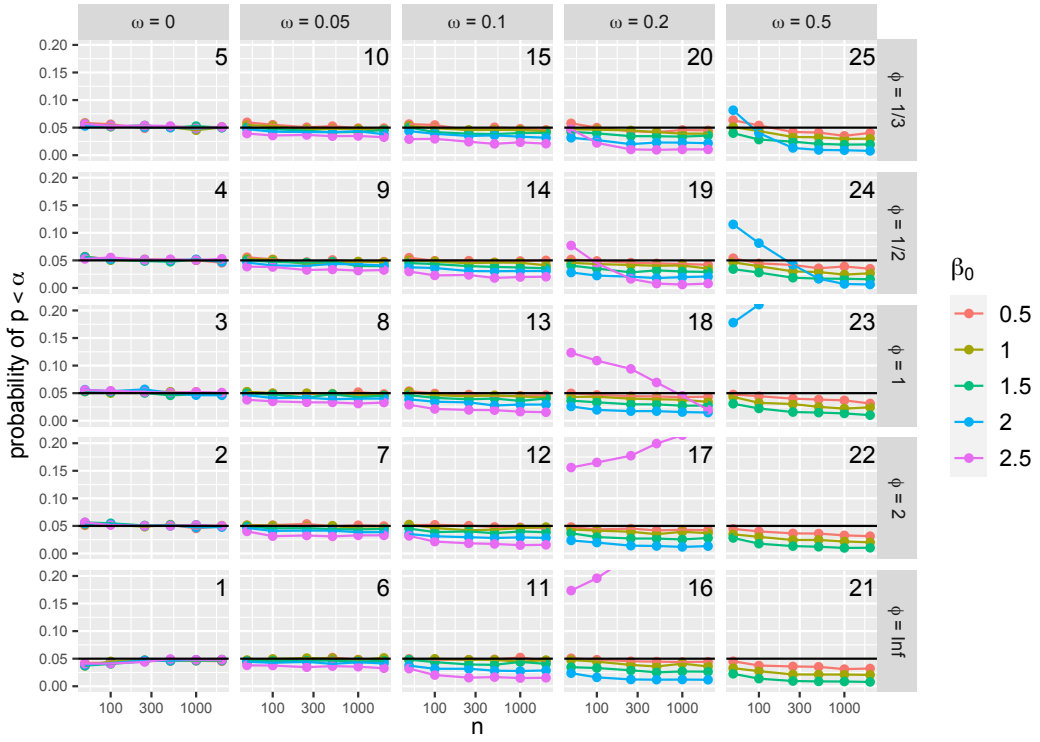

Figure 12. Probability that the NB model rejects the null hypothesis of $H_0 : \beta_X = \gamma_X = 0$.

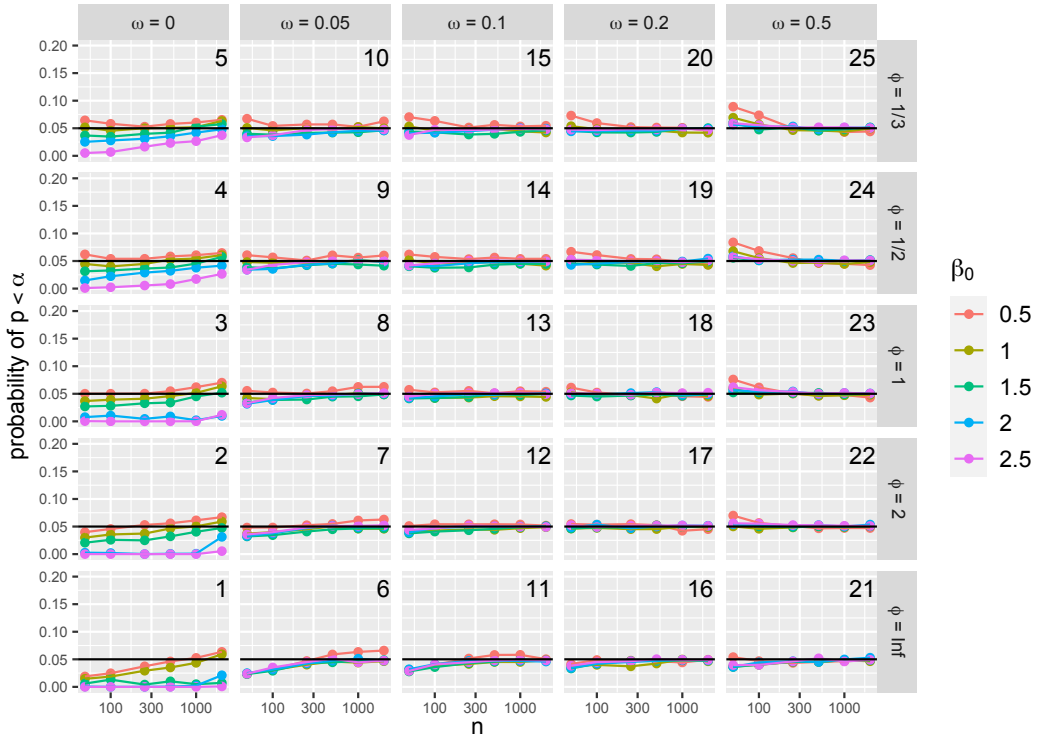

Figure 13. Probability that the ZINB model rejects the null hypothesis of $H_0 : \beta_X = \gamma_X = 0$.

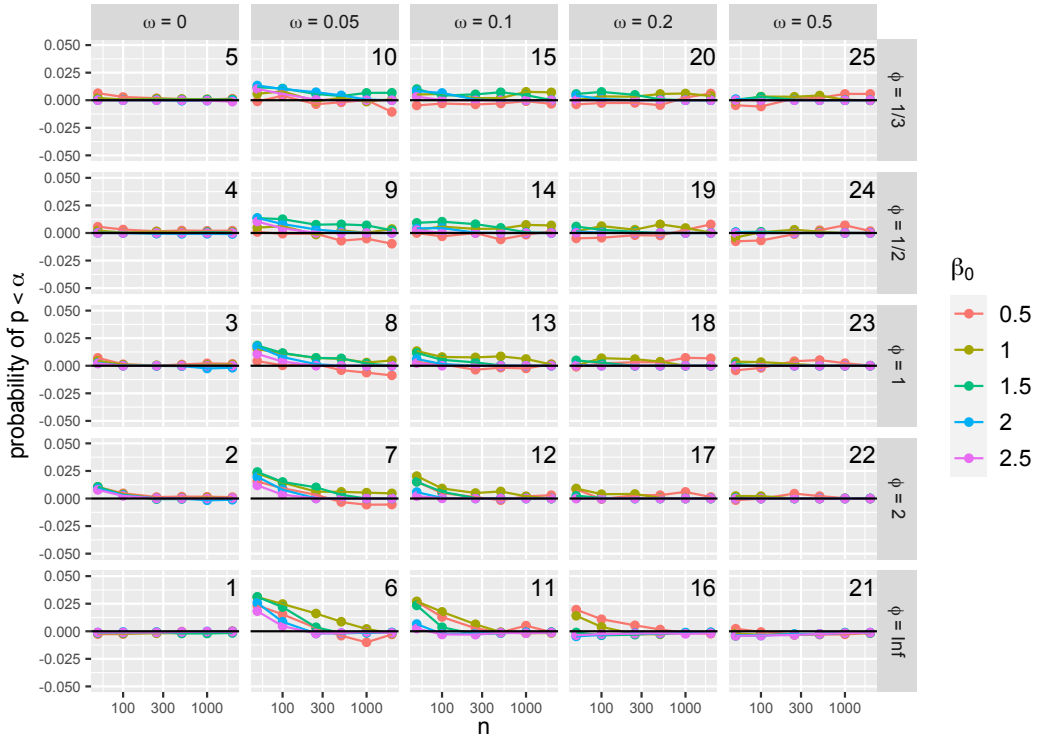

Figure 14. Difference between type 1 error under "correct" model (in Figure 2) and unconditional type 1 error (in Figure 4).

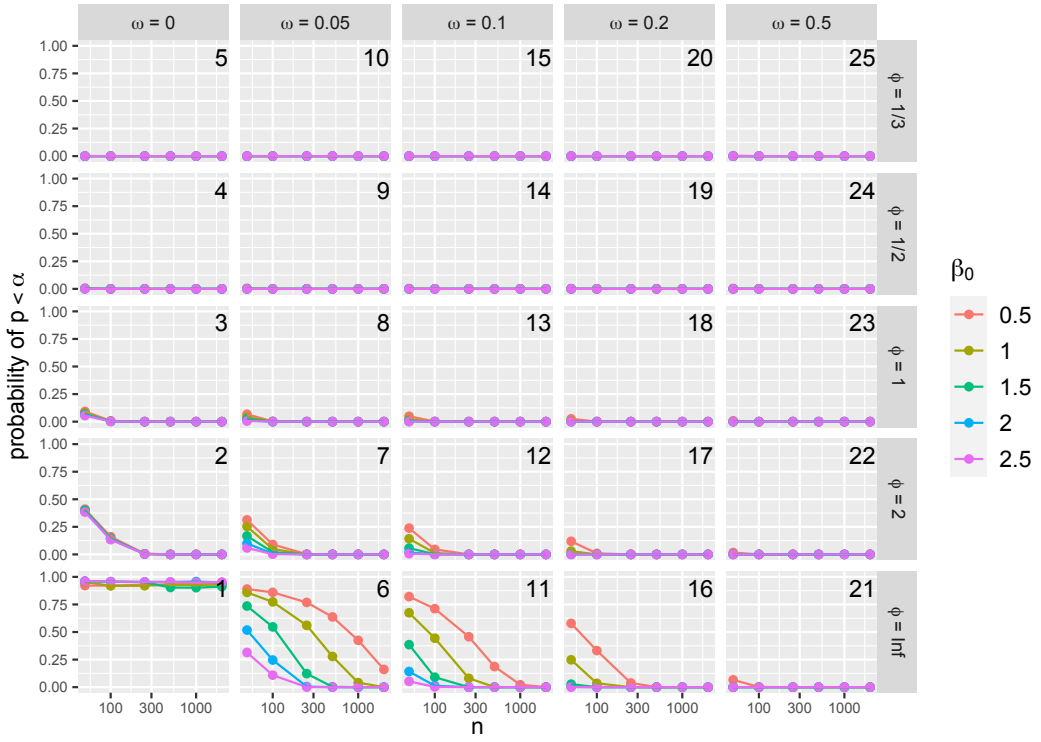

Figure 15. Proportion of datasets for which the preliminary testing scheme selects the Poisson model for analysis.

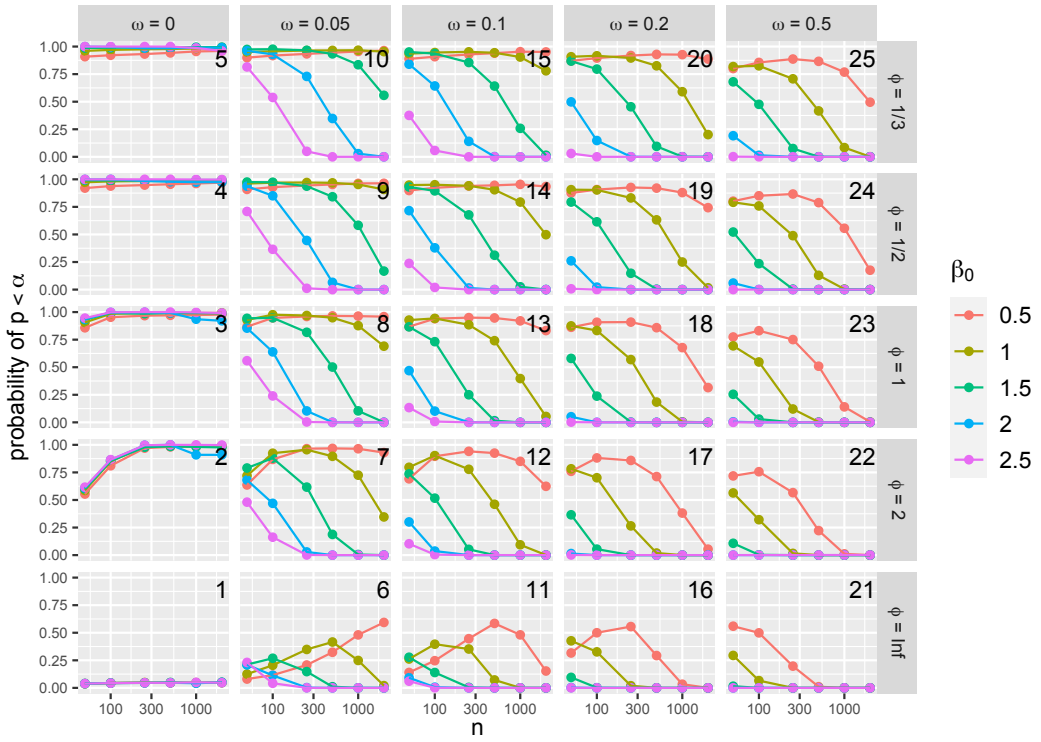

Figure 16. Proportion of datasets for which the preliminary testing scheme selects the NB model for analysis.

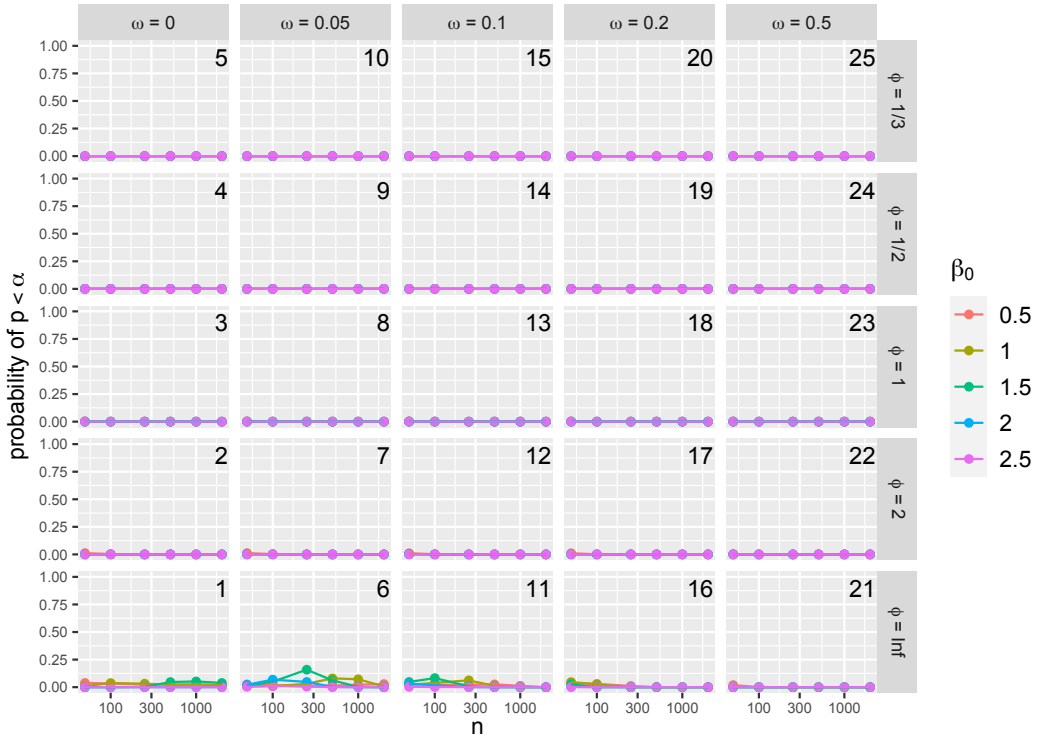

Figure 17. Proportion of datasets for which the preliminary testing scheme selects the ZIP model for analysis.

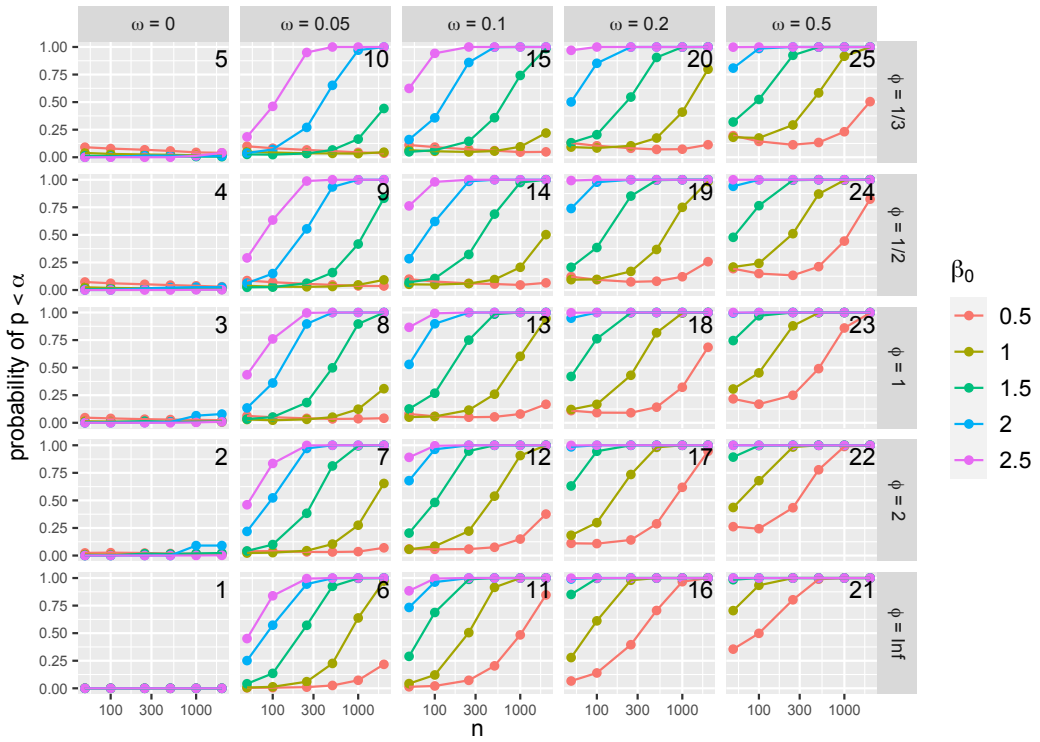

Figure 18. Proportion of datasets for which the preliminary testing scheme selects the ZINB model for analysis.