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 October 31, 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, 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 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 & 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 & 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 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" 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 & 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, 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 & 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 overlooked" (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 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 for practitioners.

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,..., 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:

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

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

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, ..., 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) 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 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 $\widehat{\lambda}_i = \exp(\widehat{\beta}_0 + \widehat{\beta}_X X)$.

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. 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 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). Let us briefly review the three alternative regression models that we will consider.

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

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

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 $Var(Y_i) = \mu_i + \mu_i^2\omega_i/(1-\omega_i)$. 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:

$$\Pr(Y = y_i|\alpha, \lambda_i) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i + 1)\Gamma(1/\alpha)} \left(\frac{1}{1 + \alpha\lambda_i}\right)^{1/\alpha} \left(\frac{\alpha\lambda_i}{1 + \alpha\lambda_i}\right)^{y_i}; \tag{4}$$

where we use a log link function for $\lambda_i = exp(\beta_0 + \beta_X X_i)$, and where $\alpha > 0$ 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 $Var(Y_i) = \lambda_i + \alpha\lambda_i^2$. 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:

$$\begin{aligned} Pr(Y_i = y_i|\alpha, \omega_i, \lambda_i) &= \omega_i + (1 - \omega_i)(1/(1 + \alpha\lambda_i))^{1/\alpha}, \quad \text{if} \quad y = 0; \\ &= (1 - \omega_i)\frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i + 1)\Gamma(1/\alpha)} \left(\frac{1}{1 + \alpha\lambda_i}\right)^{1/\alpha} \left(\frac{\alpha\lambda_i}{1 + \alpha\lambda_i}\right)^{y_i}, \quad \text{if} \quad y_i > 0; \end{aligned} \tag{5}$$

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 $\alpha > 0$ is a dispersion parameter that does not depend on covariates. 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 a 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

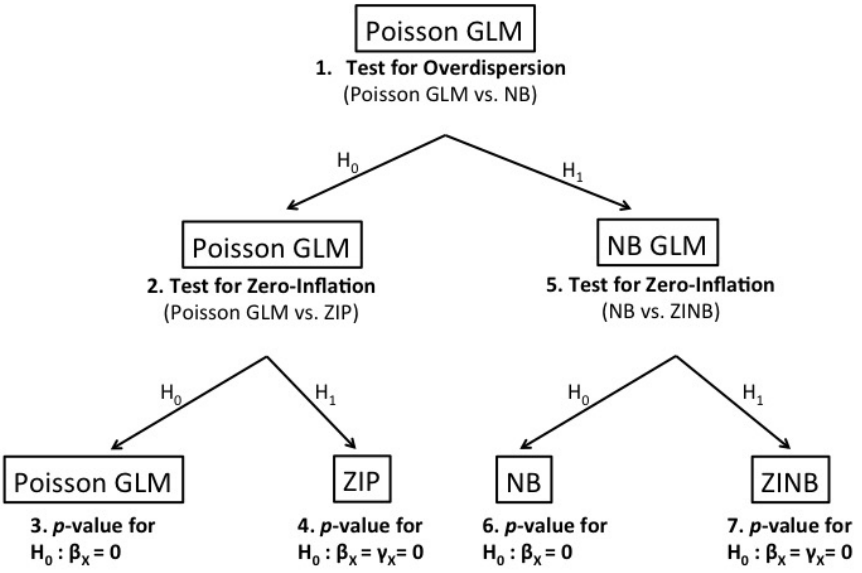

Figure 1. The multi-stage model selection procedure. The Poisson GLM is the starting point. Three score tests lead to one of four models.

instead of the Clarke test (or the Ridout score test), since the Vuong test appears to be the most widely used in practice.

We conducted a 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:

$y_i \sim NegBin(\alpha, \lambda = exp(\beta_0))$, 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:

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

For each scenario, all data are simulated under the null hypothesis with $\beta_X = 0$ (and $\gamma_X = 0$). We varied the following: $n = (30, 50, 100, 250, 500, 1000)$, $\beta_0 =$

$(0.5, 1.0, 1.5, 2.0, 2.5)$, $\alpha = (1, 1.5, 2, 3, 4)$, and $\omega = (0, 0.05, 0.1, 0.2, 0.5)$. Note that going forward we refer to:

- scenarios with $\alpha = 1$ and $\omega = 0$ as those with data simulated from the Poisson distribution;
- scenarios with $\alpha > 1$ and $\omega = 0$ as those with data simulated from the Negative Binomial distribution;
- scenarios with $\alpha = 1$ and $\omega > 0$ as those with data simulated from the Zero-inflated 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: $X_i \sim Normal(\mu = 0, \sigma^2 = 100)$, 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. We did not consider models that deal with under-dispersion, even though under-dispersed counts often arise in various ecological studies; see Lynch et al. (2014).

Note that, for the Poisson distributed data, we are simulating data with overall mean of $\lambda = exp(\beta_0) \approx (1.6, 2.7, 4.5, 7.4, 12.2, 20.1, 54.6)$. For $\lambda > 5$, any zeros in the data would be unexpected since $Pr(Y = 0|\lambda) \approx 0$. Also note that the values of α that we consider (1, 1.5, 2, 3, and 4) are perhaps low compared to what is observed empirically. Zuur et al. (2007) note: "Although it is common in ecology to have datasets with a large overdispersion, it is unclear how large a value of $[\alpha]$ is acceptable. Some authors report that $[\alpha] = 5$ or 10 is large, and other authors use overdispersion parameters of 50 or more."

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 are interested in the overall 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 initial simulations, LRTs performed rather erratically under under certain situations when the model was misspecified (e.g., when a Poisson model was fit to ZIP data, convergence issues occurred occasionally and resulted in inappropriately small LRT p-values).

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 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 the type 1 error appears to be higher than the advertised rate of 0.05 for scenarios with $n \leq 100$, $\alpha > 1$ and $\omega = 0.5$ (see Figure 2, panels 22-25).

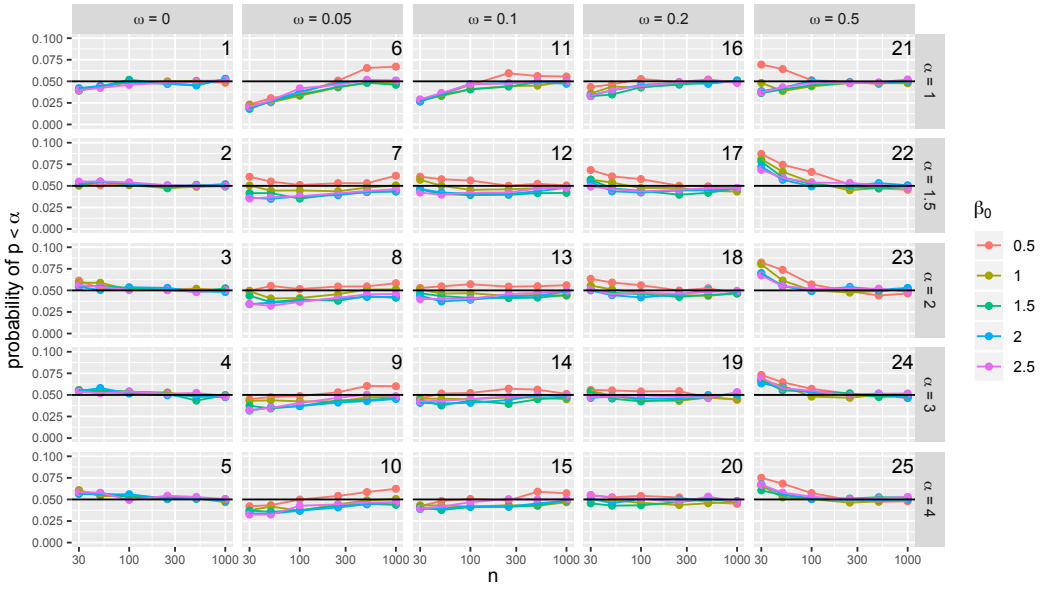

Figure 2. The level of Type 1 error under the "correct model."

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 Figure 5) and specifically scenarios with $\omega = 0$ and $\alpha > 1$, i.e., scenarios with overdispersion and no zero-inflation. With the exception of the small sample-size scenarios ($n \leq 100$) with a small amount of overdispersion ($\alpha = 1.5$), the D&L test correctly rejects the null hypothesis of no overdispersion for the vast majority of cases (Figure 5 panel 2-5). For all cases with $\alpha = 1$ and $\omega = 0$, the D&L test appears to show correct type 1 error (Figure 5 panel 1). However, somewhat unexpectedly, when $\alpha = 1$ and $\omega > 0$, the D&L test will often reject the null hypothesis of no overdispersion 5 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. See Figures 6 and 7 for the Vuong test results.

Overall, the probability that the preliminary testing scheme selects the "correct" model depends highly on β_0, ω, α, and n, see Figure 3. Note that for the ZIP data scenarios ($\omega > 0$, $\alpha = 1$; Figure 3 panels 6, 11, 16, 21), the "incorrect" ZINB model is

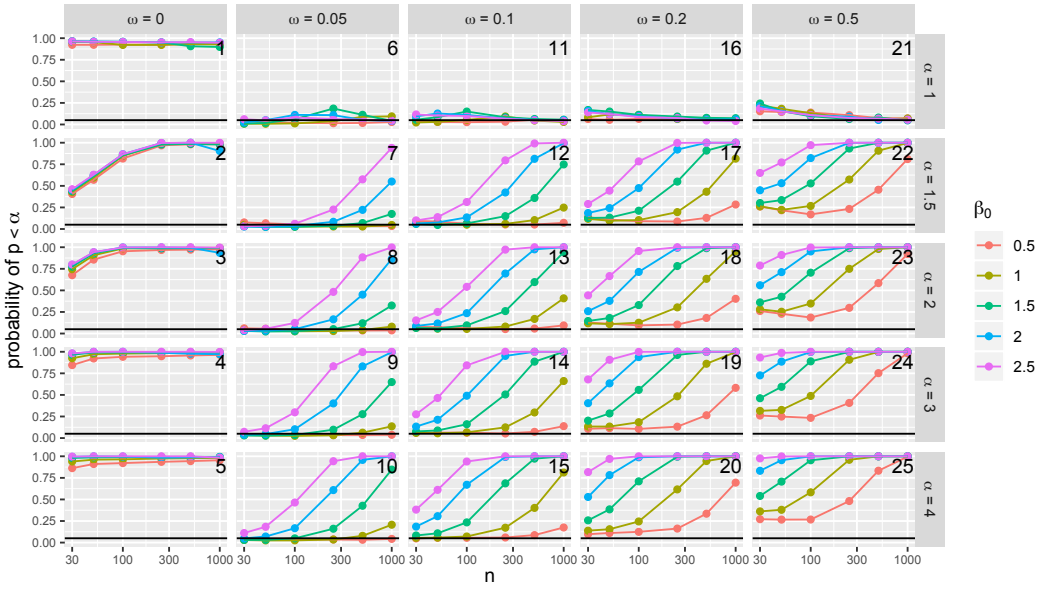

Figure 3. The probability of selecting the "correct model."

chosen in a majority of cases. This may not necessarily lead to type 1 error inflation since the "incorrect" ZINB model can be overly conservative when applied to ZIP data, see Figure 11. For ZINB data scenarios (i.e., when $\omega > 0$ and $\alpha > 1$,) in cases when the ZINB model is not selected, it is most likely that the NB model is selected instead. This also might not be overly consequential since the misspecified NB model appears to maintain correct type 1 error in many of these situations, see Figure 10.

Unconditional type 1 error - Our main question of interest is whether the null hypotheses of no association between X and Y is rejected at the 0.05 significance level when following the entire seven-step procedure outlined in Section 3. The corresponding rejection rates are plotted in Figure 4. Three observations merit comment.

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 while incorrect models are occasionally chosen (see Figure 3, panel 1), these are always more conservative models than the Poisson model. As such, model selection bias, in this case, has the innocuous effect of slightly lowering the type 1 error level.

Second, for data simulated from the ZIP distribution (i.e., when $\omega > 0$ and $\alpha = 1$), the "incorrect" ZINB model is almost always selected. However, the type 1 error under this "incorrect" ZINB model is not substantially higher than the advertised 0.05 rate (see Figure 4, panels 6,11,16,21). As such, the type 1 error level is left relatively unaffected by the model selection bias.

Third, for data simulated from the NB and ZINB distributions (i.e., when $\alpha > 1$; see Figure 4, panels 2-5, and panels 7-10, 12-15, 17-20, 22-25), model selection bias does substantially inflate the type 1 error for small sample sizes. For some scenarios, we see observed type 1 error rates at nearly double the advertised rate. For example, for the scenario with $n = 30$, $\beta_0 = 0.5$, $\alpha = 1.5$ and $\omega = 0.5$, the type 1 error is 0.095; see Figure 4, panel 22. The magnitude of this "inflation" decreases with increasing α, increasing β_0, decreasing ω, and increasing n.

Note that the type 1 error for these small n scenarios is very high even when the correct model is always selected (see Figure 11). Indeed, consider once again the example scenario with $n = 30$, $\beta_0 = 0.5$, $\alpha = 1.5$ and $\omega = 0.5$. The type 1 error obtained with the "correct" ZINB model for this example scenario is 0.087. The main reason why, following preliminary testing, this rate is increased to 0.095 is that, in approximately 6% of cases, the Poisson GLM is incorrectly selected, see Figure 14. If one were to always test this ZINB data with a Poisson GLM, the type 1 error would be more than 21%, see Figure 8, panel 22.

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. 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. In such circumstances, Tsou (2006) suggest using a "robust" Poisson

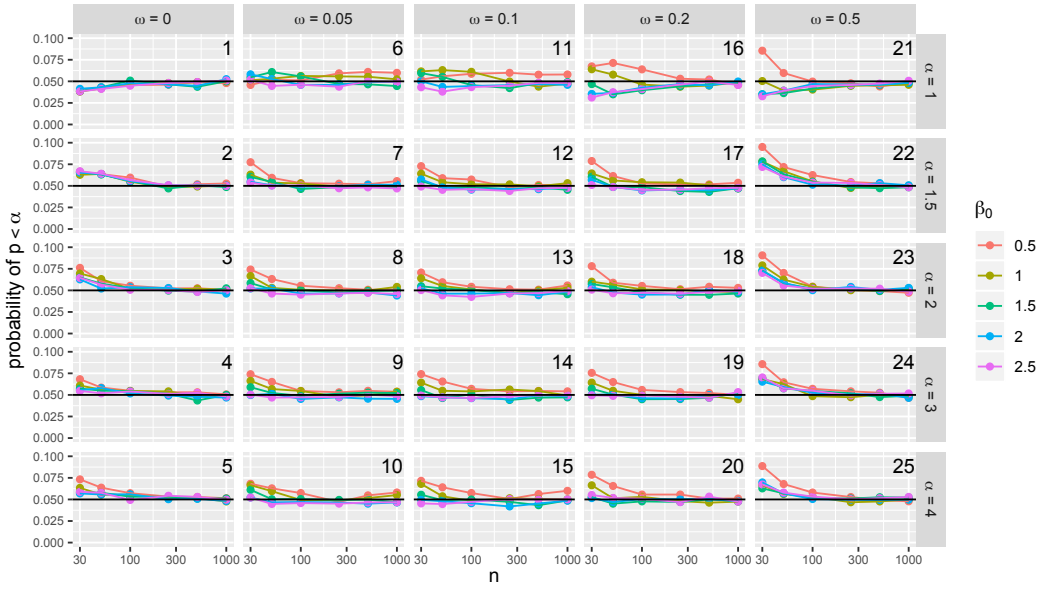

Figure 4. Type 1 error obtained following the seven step testing scheme outlined in Section 3.

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 often immense expense required to obtain data, anyone working in data-driven research will no doubt be reluctant to adopt any approach which potentially compromises power. Williamson et al. (2007) review power/sample size requirements for ZIP and ZINB models.

Researchers who do not know in advance whether or not there is overdispersion or zero-inflation, should in theory, use a "robust" model or perhaps simply use a ZINB as a "safer bet" (Perumean-Chaney et al. 2013) and pay a price in terms of efficiency. In practice, this might not be necessary. In our simulation study, we did not see a substantial degree of type 1 error inflation in the scenarios with large sample sizes ($n >$ 100). This is due to the fact that the preliminary tests we considered have relatively high power to identify situations when the distributional assumptions are questionable. And when preliminary testing does erroneously select an "incorrect" model, this is not necessarily problematic (in terms of type 1 error) since the misspecified model may not necessarily be liberal in testing. In some cases the "incorect" model may be do just the opposite and be overly conservative.

When the sample size is small, preliminary testing does have the potential to inflate the type 1 error. Indeed, our simulation study suggests that the impact of model selection bias for significance testing with small samples can be substantial. Researchers should always be cautious when interpreting results when n is small. Model selection bias is just one more reason (Button et al. 2013) to be cautiously skeptical of significance testing 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.

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.

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'.

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.

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.

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.

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.

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.

Zuur, A., Ieno, E. N. & Smith, G. M. (2007), *Analyzing ecological data*, Springer Science & Business Media.

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:

$$T_1 = \sum_{i=1}^{n} \left\{ \left(y_i - \hat{\lambda}_i \right)^2 - y_i \right\} / \left(2 \sum_{i=1}^{n} \hat{\lambda}_i^2 \right)^{1/2} \tag{6}$$

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:

$$V = \frac{\sum_{i=1}^{n}(\log dL_i)}{\sqrt{n} \cdot \sqrt{\sum_{i=1}^{n}((\log dL_i - \sum_{i=1}^{n}(\log dL_i)/n)^2/(n-1))}}, \tag{7}$$

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: $\log dL_i = \log(Pr_{ZINB}(Y_i = y_i|\widehat{\alpha}, \widehat{\omega}_i, \widehat{\lambda}_i)) - \log(Pr_{NB}(Y_i = y_i|\widehat{\alpha}, \widehat{\lambda}_i))$.

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).

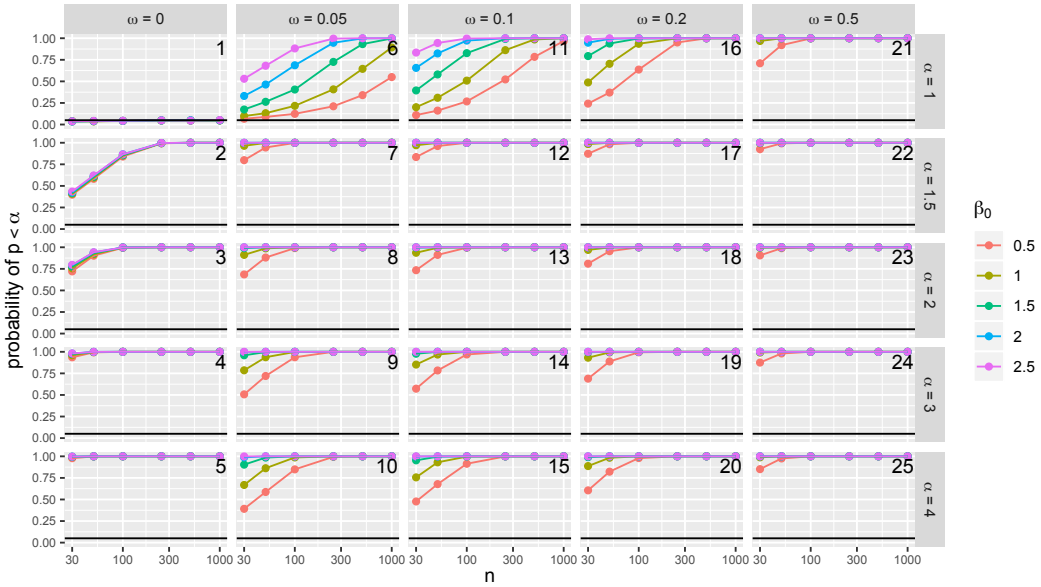

Figure 5. Probability that the D&L test rejects the null hypothesis that there is no overdispersion.

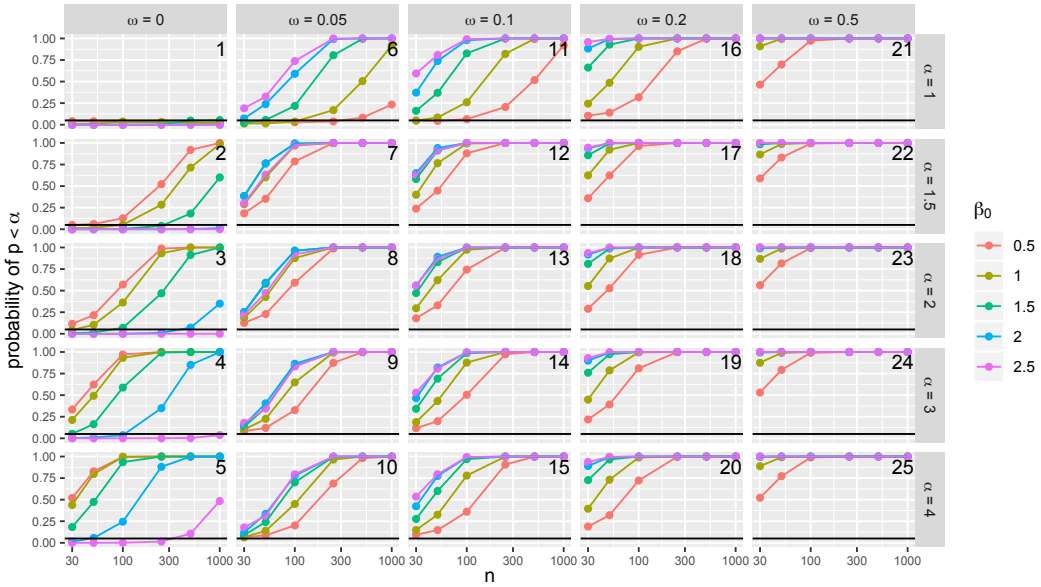

Figure 6. Probability that the Vuong test rejects the null hypothesis that there is no zero-inflation, comparing the Poisson model to the ZIP model.

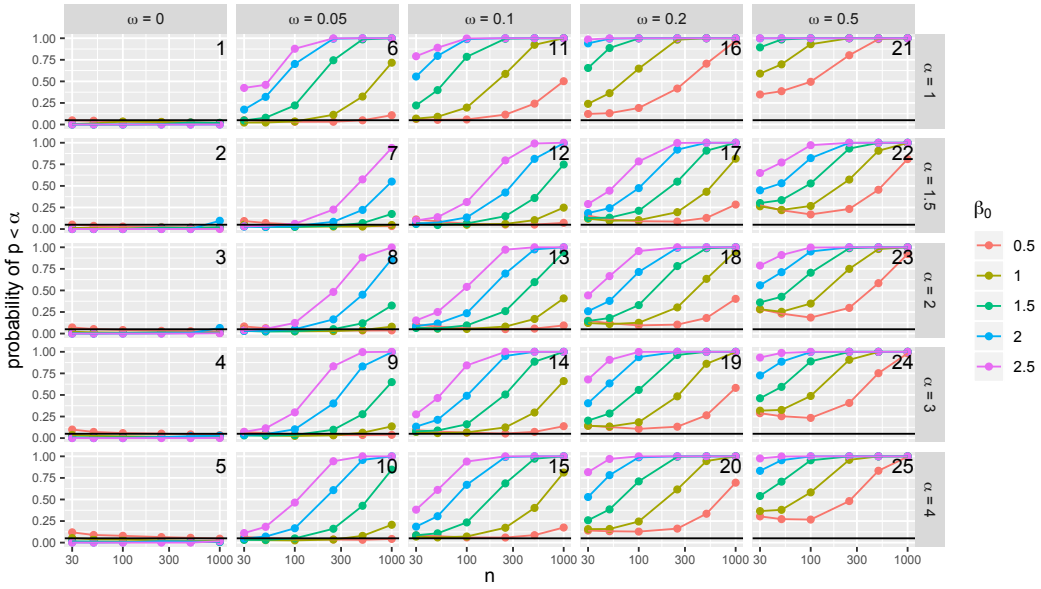

Figure 7. Probability that the Vuong test rejects the null hypothesis that there is no zero-inflation, comparing the NB model to the ZINB model.

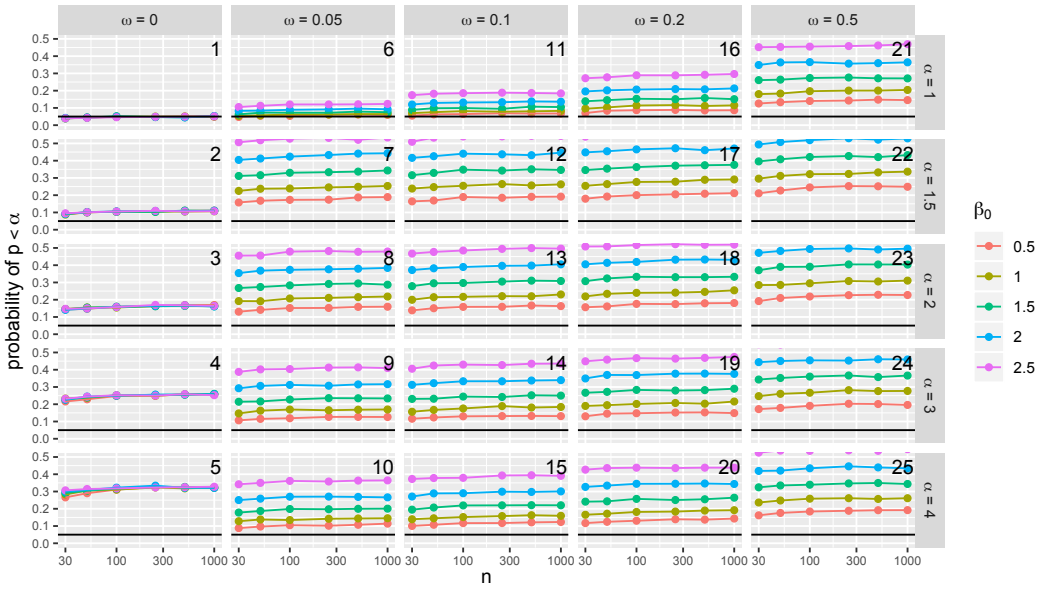

Figure 8. Probability that the Poisson model rejects the null hypothesis of $H_0 : \beta_X = 0$.

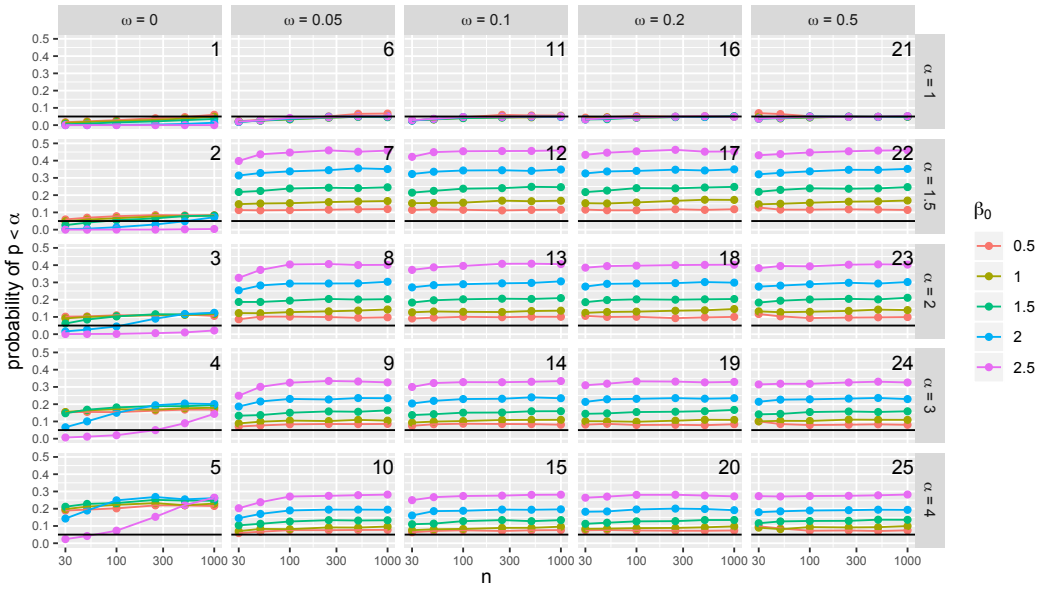

Figure 9. Probability that the ZIP model rejects the null hypothesis of $H_0 : \beta_X = 0$.

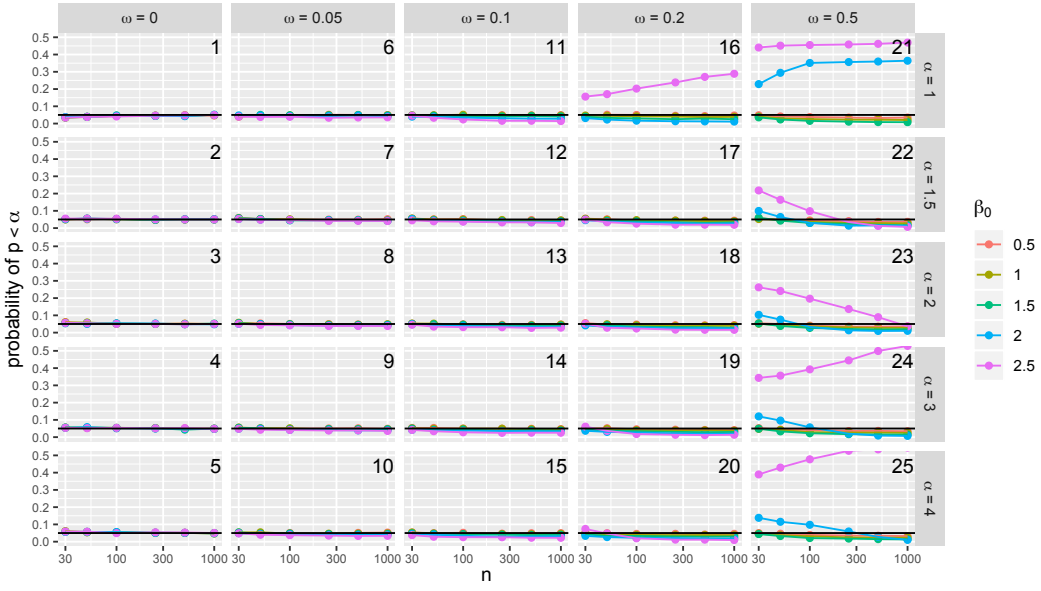

Figure 10. Probability that the NB model rejects the null hypothesis of $H_0 : \beta_X = \gamma_X = 0$.

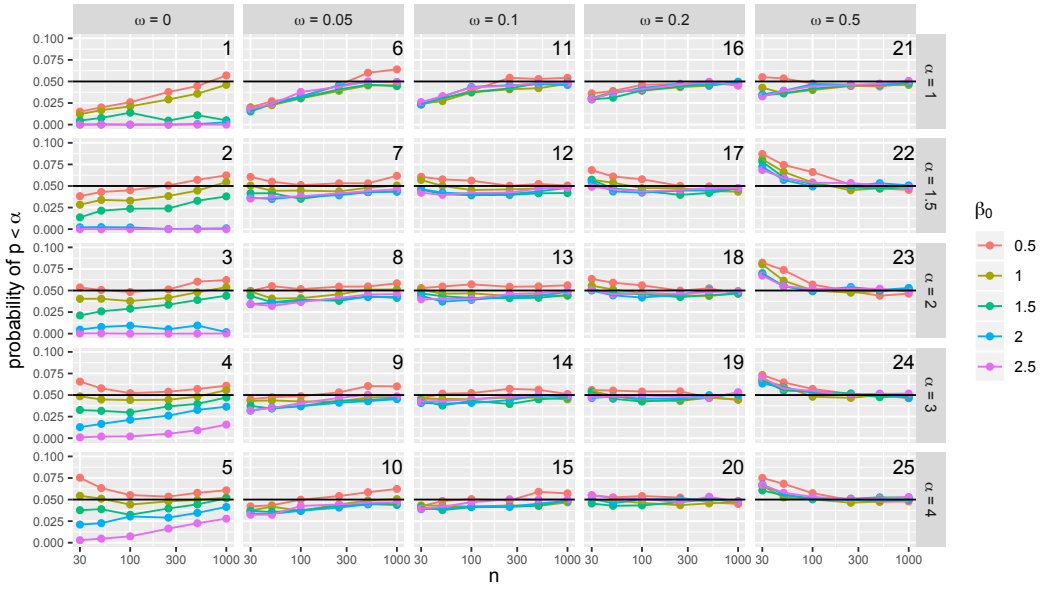

Figure 11. Probability that the ZINB model rejects the null hypothesis of $H_0 : \beta_X = \gamma_X = 0$.

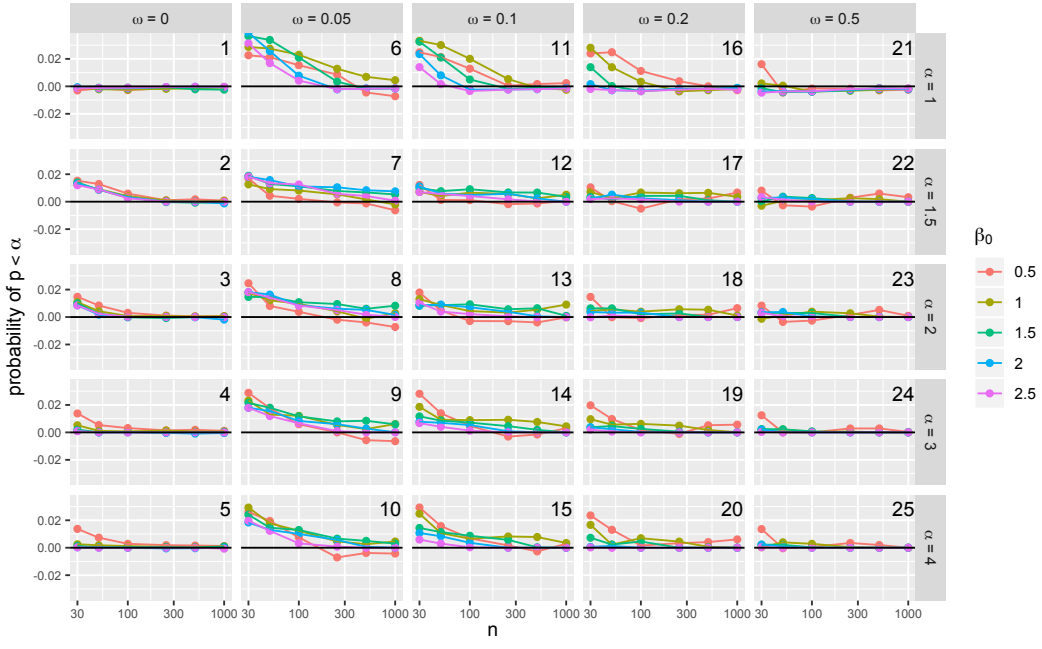

Figure 12. Difference between type 1 error under "correct" model (in Figure 2) and unconditional type 1 error (in Figure 4).

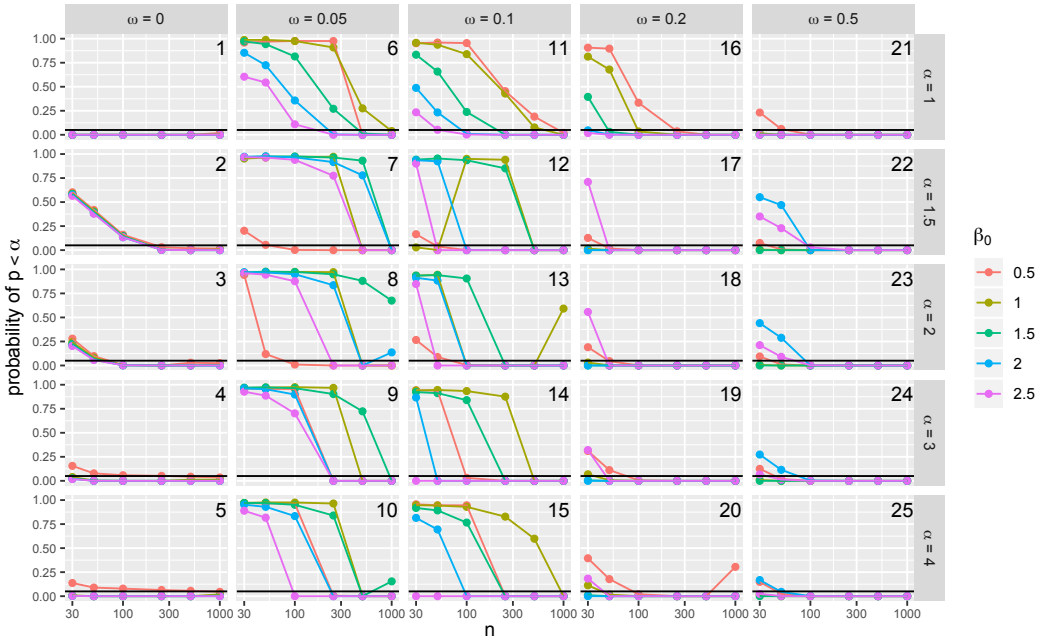

Figure 13. Probability of selecting a model that has higher type 1 error than the "correct" model (i.e., higher type 1 error than rate plotted in Figure 2.)

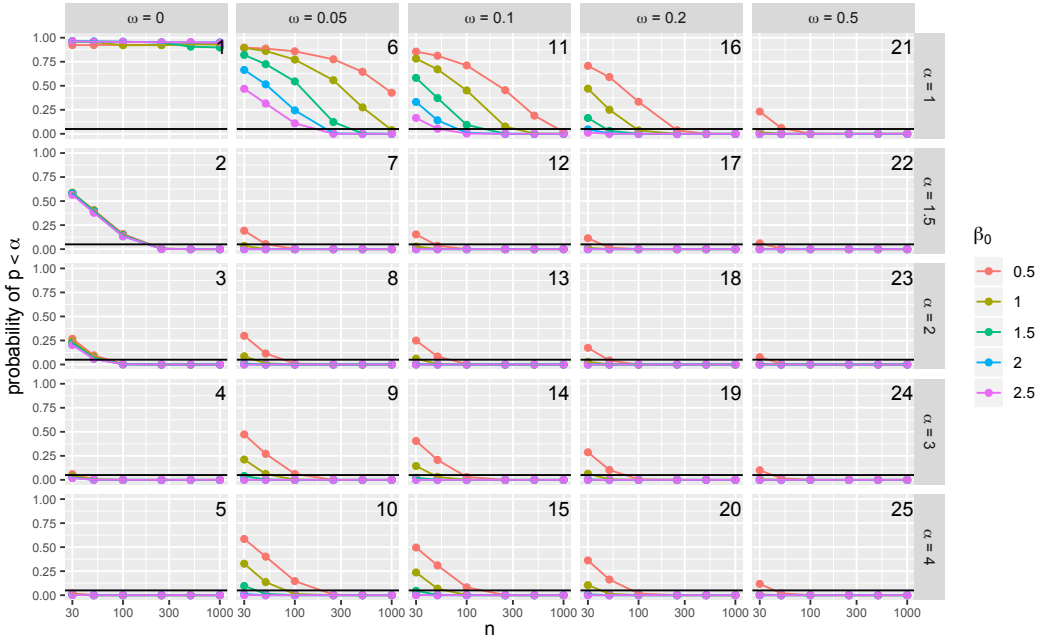

Figure 14. Proportion of datasets for which the preliminary testing scheme selects the Poisson model for analysis.

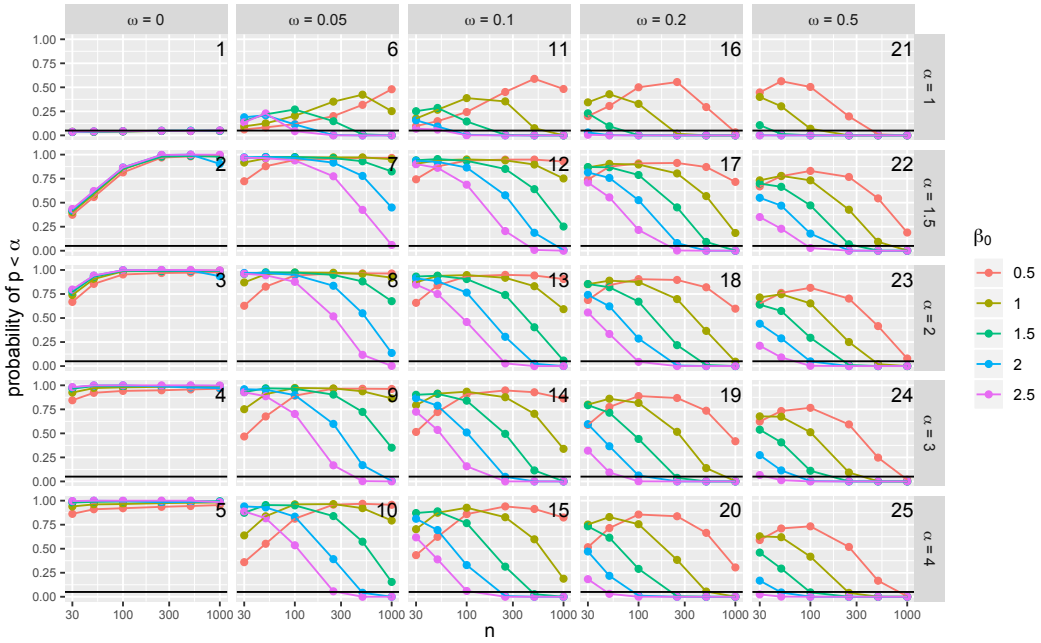

Figure 15. Proportion of datasets for which the preliminary testing scheme selects the NB model for analysis.

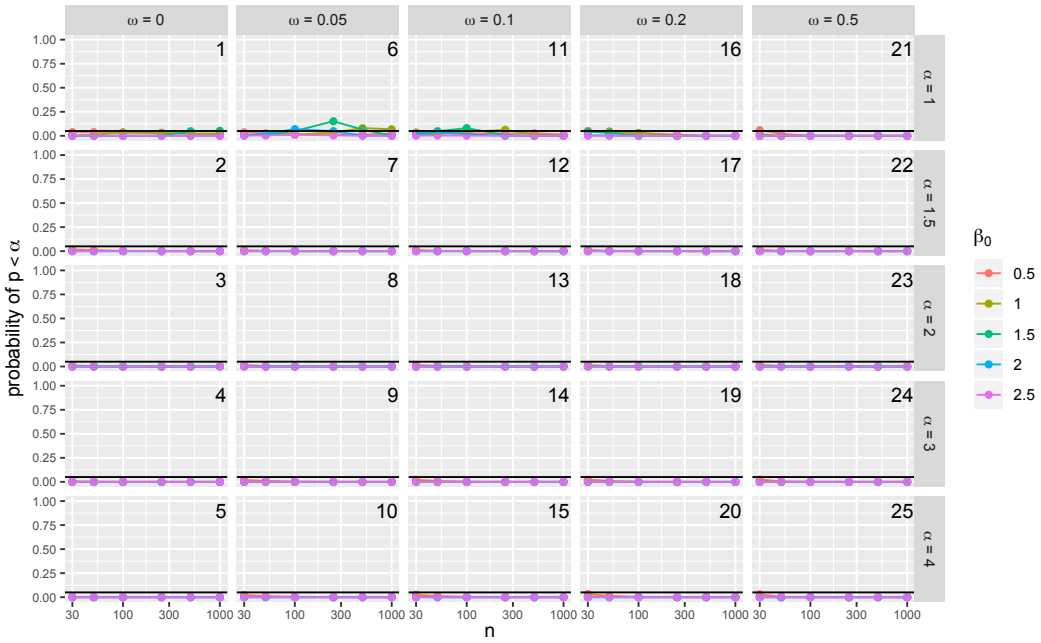

Figure 16. Proportion of datasets for which the preliminary testing scheme selects the ZIP model for analysis.

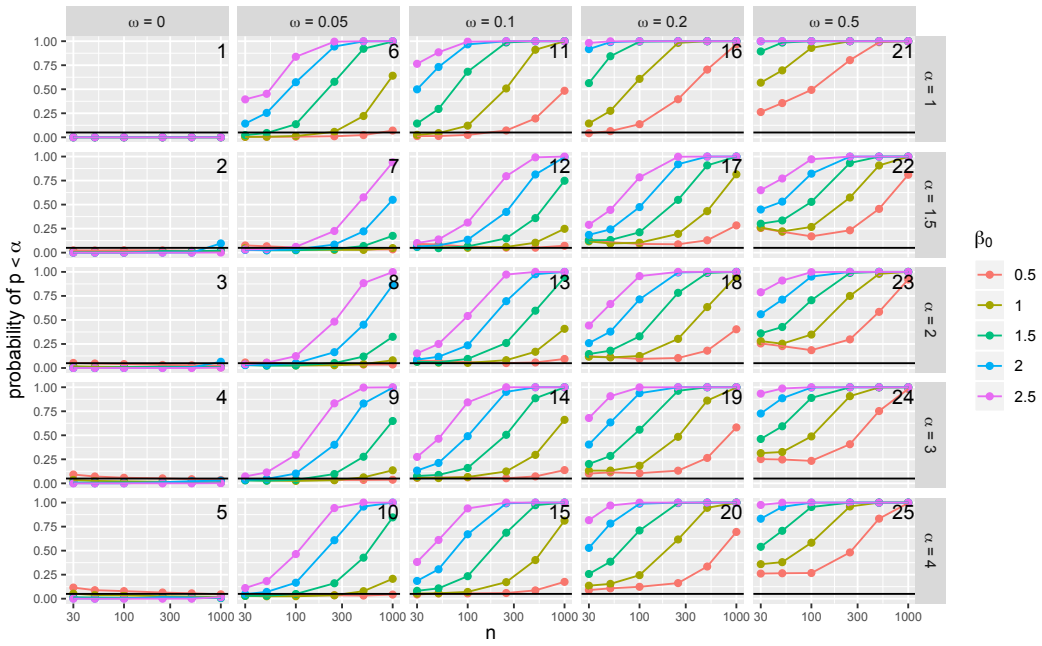

Figure 17. Proportion of datasets for which the preliminary testing scheme selects the ZINB model for analysis.