Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7,** 882–890 doi: 10.1111/2041-210X.12552

FORUM

Three points to consider when choosing a LM or GLM test for count data

David I. Warton[1]*, Mitchell Lyons[2], Jakub Stoklosa[1] and Anthony R. Ives[3]

[1]*School of Mathematics and Statistics and Evolution & Ecology Research Centre, University of New South Wales, NSW 2052, Australia;* [2]*School of Biological, Earth and Environmental Sciences, University of New South Wales, NSW 2052, Australia; and* [3]*Department of Zoology, University of Wisconsin-Madison, Madison, WI 53706, USA*

Summary

1. The two most common approaches for analysing count data are to use a generalized linear model (GLM), or transform data, and use a linear model (LM). The latter has recently been advocated to more reliably maintain control of type I error rates in tests for no association, while seemingly losing little in power. We make three points on this issue.

2. *Point 1* – Choice of statistical model should primarily be made on the grounds of data properties. Choice of testing procedure should be considered and addressed as a separate issue, after model choice. If models with the appropriate data properties nonetheless have statistical problems such as type I error control (*i.e.* type I error rate greatly exceeds the intended significance level), the best solution is to keep the model but fix the problems.

3. *Point 2* – When a test has problems with type I error control, it can usually be corrected, but this may require departure from software default approaches. In particular, resampling is a good solution for small samples that can be easy to implement.

4. *Point 3* –Tests based on models that better fit the data (*e.g.* a negative binomial for overdispersed count data) tend to have better power properties and in some instances have considerably higher power.

5. We illustrate these issues for a 2 × 2 experiment with a count response. This seemingly simple problem becomes hard when the experimental design is unbalanced, and software default procedures using LMs or GLMs can have difficulties, although in both cases the issues can be fixed.

6. We conclude that, when GLMs are thought to fit count data well, and when any necessary steps are taken to correct type I error rates, they should be used rather than LMs. Nonetheless, standard LM tests are often robust and can have good type I error control, so there is an argument for their use for counts when diagnostics are difficult and statistical models are complex, although at some risk of loss of power and interpretability.

Key-words: data transformation, generalized linear models, multivariate analysis, power analysis, type I error

Introduction

Generalized linear models (GLM) are conventionally taught as the primary method for analysis of count data, key components of their specification being a statement of how the mean response relates to a set of predictors and how the variance is assumed to vary as the mean varies (McCullagh & Nelder 1989; Wood 2006). In ecology, there has been some recent discussion in this journal of their relative merit (O'Hara & Kotze 2010; Ives 2015) as compared to transforming data and using ordinary least squares regression (hereafter a 'linear model', LM). O'Hara & Kotze (2010) advocated the use of GLMs, arguing they have greater efficiency when estimating parameters, and avoid issues with transformation bias and interpretability when analysing transformed data. Ives (2015) responded with simulations demonstrating that in the special case of testing for significance of regression coefficients, LMs have more reliable type I error control (*i.e.* the type I error produced by the statistical test is closer to the intended significance level) and lose little power; thus, LMs provide a robust alternative. So what should we do in practice?

In this article, we make three points on the issue. First, we believe that model choice and type I error control are best considered as two separate objectives. The goal should be to choose a model that captures key data properties (using *a priori* knowledge and diagnostic tools) and then consider how to use it to make valid inferences. Second, when choosing a testing procedure, good type I error control can usually be achieved, although this may require departure from software defaults. For example, standard GLM tests can have incorrect (inflated) type I error in small samples, but this can usually be fixed using resampling. Third, differences in power can be quite

*Correspondence author. E-mail: david.warton@unsw.edu.au

[Correction note: an additional supporting information file was added on 5 October 2016 after first online publication.]

substantial, with significant gains from using a better model for the data. In unbalanced designs in particular, GLMs can have considerably higher power than LMs for count data. We close with some concrete recommendations for how one should analyse counts in practice, and make some recommendations for other related contexts, such as multivariate analysis.

Simulations – unbalanced 2 × 2 design

The ideas in this study will be illustrated using simulations of overdispersed counts collected in a 2 × 2 unbalanced sampling design, with sample sizes across the four possible treatment groups in the ratio 9:3:3:1. We considered type I error rates and power of tests for the main effect of the second factor ('Factor B') given an effect of the first factor (which was fixed at a slope of one). Two elements of the design were varied – whether or not the two factors were correlated (see Table 1), and the distribution that overdispersed counts were simulated from. Data were either negative binomial or Poisson-lognormal, these two distributions having similar properties, including the same mean–variance relationship. This sampling design, while relatively simple, presents a challenge – with unequal sample sizes, unequal variances and simulating from two distributions that are hard to distinguish from each other in practice.

Point 1 – Choose your model based on data properties

A statistical model is most likely to achieve its goals when it as closely as possible reflects the true underlying data-generating mechanism – both correctly identifying the main signals in the data and correctly modelling and accounting for the main sources of variation (*e.g.* Venables & Ripley 2013). So when specifying a statistical model, the primary goal should be getting the model right for the data at hand. Towards this goal, the best tools are *a priori* knowledge and model diagnostic tools like residual plots (see Warton *et al.* 2015, for example).

Two key data properties are as follows: the mean model, that is how the mean of the response varies as predictors vary, and the variance model, that is how the variance changes as the mean changes or as the predictors change (McCullagh & Nelder 1989). Counts are known to have a strongly increasing mean–variance relationship (*e.g.* Warton 2005; Ver Hoef & Boveng 2007), meaning that if there are differences between means, one should also expect differences in variances.

Table 1. Sampling ratios in each treatment combination for the unbalanced 2 × 2 design used in simulations

(a) Uncorrelated				(b) Correlated			
		Factor B				Factor B	
		Low	high			Low	high
Factor A	Low	1	3	Factor A	Low	1	9
	High	3	9		High	3	3

In all cases, sample sizes were a multiple of the values specified above, leading to a design with (a) uncorrelated factors and (b) correlated factors.

GLMs include several natural models for counts (Lawless 1987; Cameron & Trivedi 2013), with two key components that match the two key data properties above – a model for the mean as a function of predictors and a distributional assumption on data that implies a particular mean–variance relationship (*e.g.* if counts are Poisson, the variance is equal to the mean). A potential issue, however, is that there are many possible distributions that counts could come from and hence many possible mean-variance relationships. For example, counts may or may not be overdispersed; there are a few ways overdispersion could be introduced (Hilbe 2007; Lindén & Mäntyniemi 2011); the data could be zero-inflated (Cameron & Trivedi 2013; Yee 2015). Each of these issues implies a different type of model, so it is important to check assumptions and in particular to check for overdispersion and account for it if present. These issues are illustrated using diagnostic plots for Dunn–Smyth residuals (Dunn & Smyth 1996) from Poisson and negative binomial GLMs (Fig. 1a and b, respectively), fitted to negative binomial data simulated using the correlated design of Table 1. (Negative binomial models can be fitted on R using the `glm.nb` function in the MASS package, Venables & Ripley 2013; or using the `manyglm` function in the MVABUND package, Wang *et al.* 2012; which also produces Dunn–Smyth residuals.) The Poisson model did not account for overdispersion, and hence, it underestimated the variance, especially in groups with larger counts. This was evident as a systematic fan-shaped pattern in the residual vs. fits plot (Fig. 1a, left), and with residuals often taking more extreme values than expected on the normal quantile plot (Fig. 1a, right), in contrast to residual plots for the correct negative binomial model (Fig. 1b).

Ordinary linear models (*e.g.* ANOVA) have a structural mismatch between the model and data properties; specifically, they assume equal variance and so cannot account for the mean-variance relationship typically seen in counts. A standard approach is to transform data such that the variance is approximately stabilized, so that it does not vary as the mean count varies. Transformation comes at some cost to interpretability (O'Hara & Kotze 2010; Ives 2015). While, in a GLM, we specify a model for the mean of the response (y); hence, parameters can be interpreted in terms of effects on mean response, in a LM on transformed data, we model the mean of transformed data, which is something entirely different and occasionally nonsensical (*e.g.* the fitted values can be negative, as occurred in our simulations). Further, if there are too many small counts, transformation cannot satisfy the equal variance assumption (DI Warton, M Lyons, J Stoklosa, unpublished data). While these issues can be seen for any transformation, we have illustrated throughout using the $\log(y + 1)$ transformation, as one of the best-performing options (Ives 2015). In our simulated example, a LM on $\log(y + 1)$-transformed data did not account for the heteroscedasticity generated by the negative binomial count process, leading to a fan-shaped pattern in the residual vs. fits plot not unlike in Poisson regression (Fig. 1c, left). But in contrast to the Poisson case,

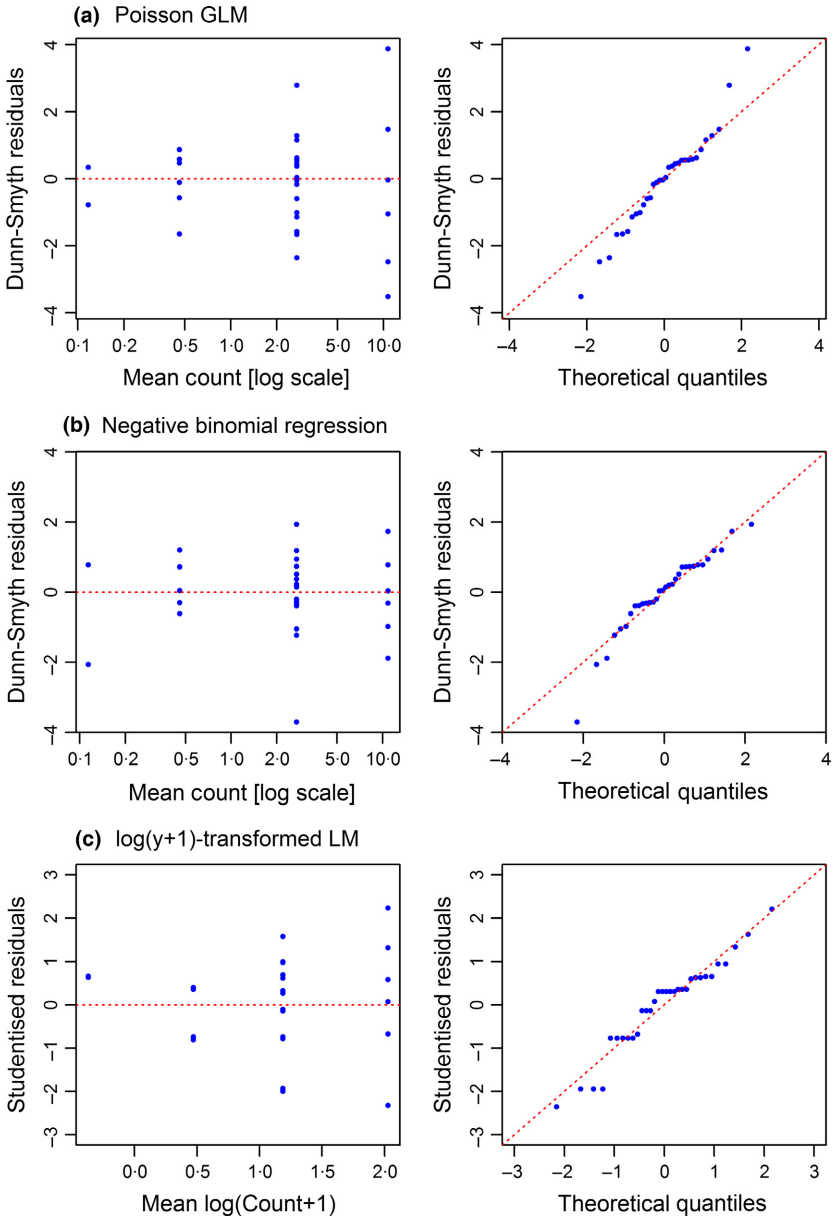

Fig. 1. Diagnostic plots of candidate models for counts simulated from a negative binomial distribution in a 2×2 sampling design. Residual vs. fits plots (left column) and normal quantile plots (right column) are used to check model fit of: (a) a Poisson GLM; (b) a negative binomial regression; (c) a linear model on log $(y + 1)$-transformed counts. Dunn–Smyth residuals (Dunn & Smyth 1996) are used for GLMs, which will be standard normal if model assumptions are correct. Note in (a) and (c) there is a fan shape in residual plots (left column), suggestive of misspecification of the model, especially of the mean–variance assumption.

the variance is not tied to the mean, so this model can account for overdispersion, as is evident in the better fit of residuals on the normal quantile plot (Fig. 1c, right). Notice also that even though the transformed counts cannot be less than zero, fitted values for one group were negative (Fig. 1c, left)! One approach to dealing with heteroscedasticity in the LM is to use weighted or generalized least squares, where weightings are used to capture changes in variance (Carroll & Ruppert 1988). In an ANOVA-type design, this can be implemented by weighting observations in different treatment groups (as defined under the *null* hypothesis) according to the inverse of their sample variances. It is less obvious how the method could be implemented in other settings, for example regression with quantitative predictors. This is related to GLMs – a GLM can be understood as a type of weighted least squares, but where the weights are constrained to satisfy an assumed mean-variance relationship. The lack of constraint here makes this particular weighted least squares method more robust (see below).

A challenge with model choice is that diagnostic tools are not always able to detect problems with the model, especially in small samples. With the sample size of 32 used in Fig. 1, under repeated simulation, diagnostic plots identified the Poisson GLM as misspecified about 80% of the time, but with a sample size of 16, model misspecification went unnoticed about 50% of the time. This emphasizes the importance of *a priori* knowledge in model choice. We often know for example to expect overdispersion, and in such cases, we can plan to use a negative binomial model in preference to a Poisson model. This also raises the issue of robustness: if diagnostics are unable to distinguish between competing models for small samples, perhaps we should use methods of fitting models that are robust to (at least moderate) violations of assumptions.

For example, a negative binomial GLM with standard inference methods lacks robustness to failure of distributional assumptions (Lawless 1987; Ives 2015), but alternatives will be discussed below.

Point 2 – Poor type I error control can be fixed

Ives (2015) studied type I error control and found that standard tests based on GLMs can have inflated type I error in small samples, even when the GLM is the correct model for the data at hand. Clearly, it is important when testing a hypothesis to use a procedure that controls type I error, in the sense that the rate of falsely declaring significance is kept close to a prespecified significance level. If the true type I error rate is unknown, and could be much larger than intended, it becomes difficult to interpret any statistical result. Sometimes, error rates are clearly an issue – in other work, we have seen type I error rates over 50%, when the intended significance level was 5%, and Karp *et al.* (2015) make a similar claim. But, for example, is 7% an acceptable error rate when intending a significance level of 5%? The authors have differing opinions about this. On the one hand, no test is exact, and 7% seems close to 5% (and sometimes it is very hard to get any closer than this in small samples). On the other hand, if the true error rate was 7%, then a result with, say, $P = 0{\cdot}036$ might not actually be statistically significant at the true 5% level. Although there may be some disagreement concerning exactly how tightly type I error needs to be controlled, everybody would agree that good type I error control is always desirable.

LMs undoubtedly have good type I error control when testing hypotheses of no association, that is when the null hypothesis specifies that all means are equal. A large part of the robustness of LMs arises somewhat incidentally because the 'all means equal' null hypothesis implies all variances are equal, irrespective of the assumed mean–variance relationship or data transformation.

If a GLM has been found that gives a good match to data properties, but it is known to have poor type I error control, there is usually a solution to the problem that does not involve changing the underlying model. For example, likelihood ratio tests (and also score tests) for GLMs tend to have better type I error properties than Wald statistics (Ives 2015; Fig. 1), as well as better power properties (Væth 1985). Further, the usual χ^2 distribution is not always a good approximation, and resampling is often a good idea when sample size is small. (In the Supplementary Material, we also considered using an F distribution for GLMs as well as for LMs, with some success.) These strategies involve departing from the usual software defaults, but are not difficult to implement.

Resampling can substantially improve type I error control, and depending on the design, might even enable an exact test. For a hypothesis of no association, permutation tests enumerating all possible permutations are known to provide an exact test (Manly 2007), and this remains true for either LM or GLM – even if the model has been misspecified. For a hypothesis of no effect within each of several groups, exact tests can be devised using what is often referred to as restricted permutation (Brown & Maritz 1982). For other situations, either a parametric bootstrap (Davison & Hinkley 1997) or a special form of residual resampling (*unpublished data*, DI Warton & YA Wang) enables tests that are close to exact, when the model being simulated from is correct. But note that how accurate such tests are may depend on how plausible the underlying model is for the data at hand. The parametric bootstrap or residual resampling usually solves the problem of obtaining accurate P-values for a correctly specified model, but might not if the model is wrong.

All the above resampling options are straightforward to implement for LMs and GLMs using existing software – for example, using the R package MVABUND (Wang *et al.* 2012), you just call the `manylm` or `manyglm` function instead of `lm` or `glm`, and the resampling method is controlled by specifying the `resamp` argument in the ANOVA call as desired. The BOOT package (Canty & Ripley 2014), included in default installations of R, is only incrementally more difficult to use and can be combined with any model or test statistic. The PERMUTE package (Simpson 2015) enables construction of permutation testing schemes for a range of different designs.

As an illustration, consider a type I error simulation under the designs of Table 1, with both negative binomial and Poisson-lognormal data, testing for an effect of Factor B when there is in fact no effect. Counts were small, with grand means of 1·4 and 1·0, respectively, for negative binomial and Poisson-lognormal data. We generated 5010 simulated data sets and estimated type I error rate as the proportion of rejections at the 0·05 significance level (which should remain close to 0·05 for a valid test). We analysed the data using Poisson and negative binomial GLMs, and a LM on log $(y + 1)$-transformed counts, obtaining P-values using standard methods (χ^2 for GLMs and F for LMs). We also included residual resampling, in which appropriately defined residuals were randomly permuted among observations (*unpublished data*, DI Warton & YA Wang), and we included restricted resampling, where counts were permuted within levels of Factor A. Finally, for the LMs, we also tried weighted least squares with an F distribution, obtaining weights empirically for each of the two groups of counts defined under the null hypothesis.

Standard negative binomial tests using the χ^2 distribution had inflated type I error for small sample sizes (Fig. 2), which was for the most part controlled when using resampling, as seen elsewhere (Szöcs & Schafer 2015). Even with residual resampling, type I error was still about 6% in small samples ($n = 16$). When data were Poisson-lognormal, χ^2 tests using negative binomial regression had slightly inflated type I error even at large sample sizes (Fig. 2). Residual resampling seemed to address this problem. Another solution is to use a different estimator of overdispersion, a method-of-moments estimator, which is robust to changes of distribution (Lawless 1987); however, we found this gives quite inflated error rates in small samples (Supplementary Material, Figure S1). The best solution we found for this particular design was restricted permutation (Brown & Maritz 1982), where observations were permuted within each level of Factor A. This is in fact an exact test for

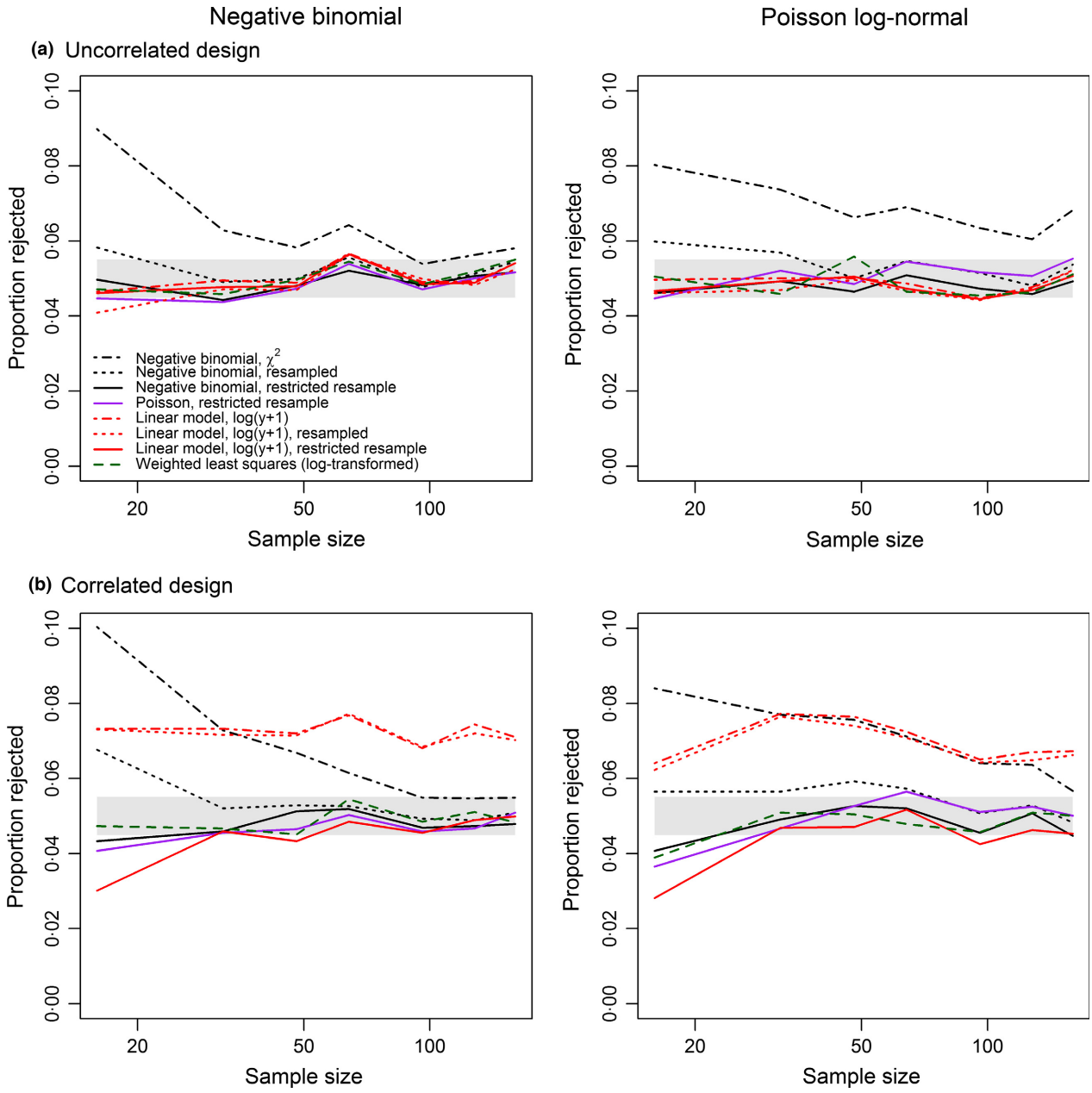

Fig. 2. Type I error simulation results for an unbalanced 2 × 2 design. Results are presented for overdispersed counts sampled from a negative binomial (left column) or Poisson-lognormal (right column) distribution. Simulations used (a) uncorrelated factors or (b) correlated factors, as in Table 1. Negative binomial regression tests include computing *P*-values using a χ^2 distribution, residual resampling or restricted resampling. The Poisson regression uses restricted resampling; Poisson χ^2 tests had very high type I error and are not presented because they are off the scale (around 0·2). For linear models, counts were log(y + 1)-transformed for a standard *F*-test, residual resampling, restricted resampling or a weighted linear model. The shaded region is a 95% confidence band around the significance level of 0·05; an exact test will usually lie in this band. Note that linear models (with or without residual resampling) had poor type I error control for correlated designs and that negative binomial χ^2 tests had inflated type I error when data were Poisson-lognormal, although residual resampling went some way towards correcting this. Restricted resampling or using a weighted linear model provides good control in both cases.

this design. All GLM tests returned similar results across correlated and uncorrelated designs.

Linear model tests had good type I error in the uncorrelated design (Fig. 2a), but poor type I error control in the correlated design, even for large sample sizes (Fig. 2b). Residual resampling did not fix the problem, because this does not address the heteroscedasticity. Restricted resampling, however, kept type I error near 0·05. Another good solution for LMs was weighted least squares obtaining weights under the two-group null hypothesis (Fig. 2). In further simulations where weights were obtained under the four-group alternative hypothesis, this method had inflated type I error (Supplementary Material), emphasizing the need for the weights to be obtained under the null. The likely reason for this is that the weights need to be independent of the effects being tested, as the weights are treated as *a priori* in analysis.

Poisson regression, which fails to account for overdispersion, led to χ^2 tests with quite inflated type I error at all sample sizes – always in the 0·17–0·22 range and off the top of the scale on Fig. 2. However, if using restricted resampling, an exact test was possible, as for negative binomial GLM and the LM. This reinforces the conclusion that it is sometimes possible to get a valid test even from a bad model, emphasizing the separation between model choices and testing procedures that we advocate.

Point 3 – A better model can have much better power

We used a power simulation to compare among models and inference methods, with the same experimental designs as previously, but with a large effect size (the coefficient for Factor B set to two) and a moderate sample size (32). We looked at the effect of varying the mean count by varying the intercept between −3 and 1, roughly corresponding to a range for mean count between 0·5 and 30. We compared the power of the four tests that had reliable type I error control – negative binomial GLM, Poisson GLM and LM under restricted resampling and the weighted least squares LM (Fig. 3).

Our main result (Fig. 3) is that power differences can be quite substantial in unbalanced designs. Negative binomial tests had consistently high power, sometimes having almost twice the power of LMs (*e.g.* 39% vs. 21%) when the effect was hard to detect, or having half the type II error rate (100−85 = 15% vs. 100−70 = 30%) when the effect was easier to see. This is not unexpected since the GLM used the correct model for negative binomial data (Fig. 3, left column) and assumed the correct mean and variance models for Poisson-lognormal data (Fig. 3, right column).

A secondary result was that power of LM tests was especially low when counts were small. One reason for this is that as the mean gets smaller, the relative difference in variances across samples increases, so heteroscedasticity becomes more of a problem when counts are small. The weighted LM had similar power properties, because the weights were computed under the null hypothesis, so heteroscedasticity across the groups being tested was not accounted for. If weights were computed under the alternative hypothesis, then different (and probably better) power properties would be seen, but with type I error issues that would need to be addressed.

Discussion

RECOMMENDATIONS

Our results show that potential problems with GLMs involving type I error control can be overcome. When used carefully, they have a natural advantage over LMs in application to count data. GLMs were developed specifically with the analysis of counts in mind, and they usually offer a more realistic model for counts with advantages in interpretability. However, as with any model, assumptions need to be carefully checked to ensure the chosen GLM has a good match to key data properties. Even then, standard GLM tests may have inflated type I error control if the sample size is small (<30, say), *for example* 8% rather than 5%. However, improved results can be obtained by using resampling to assess significance. Using a method-of-moments estimator in combination with resampling (available in the MVABUND package, `via theta.-method= Chi2`) could be expected to further improve robustness to violation of distributional assumptions.

A reasonable alternative to a negative binomial GLM is a Poisson-lognormal GLMM, especially for situations with random factors, which would require a GLMM anyway. However, this is more computationally intensive and hence more difficult to combine with resampling-based testing. We also did not test the robustness of the negative binomial GLM against zero-inflated count data, and we recommend that zero-inflated distributions be considered in situations where structural zeros might be expected – although sometimes the zeros can be adequately explained by covariates (Warton 2005).

LMs can be useful as an alternative to GLMs, especially for large counts, if it is possible to transform to approximately satisfy linear modelling assumptions. This is difficult when counts are small. LMs will be robust to overdispersion, unlike Poisson GLMs, but not to heteroscedasticity (especially in correlated experimental designs, Fig. 2b), which could be approximately addressed in a LM framework using weighted least squares. While this can improve type I error control, LMs can have substantial loss of power compared to GLMs. Nonetheless, LMs may be useful for fitting complex models (for example, a model with a spatial random effect to a large data set), if the increase in computational burden is non-trivial when moving away from the assumption that data are normally distributed. Another important disadvantage is loss of interpretability (O'Hara & Kotze 2010; Ives 2015). For example, the slope estimated by a $\log(y + 1)$-transformed LM is not equal to the slope parameter in a negative binomial simulation model – instead, the slope gives the change in expectation of $\log(y + 1)$ per unit change in x, which is unlikely to be of specific interest to a researcher.

If predictor variables are quantitative, rather than the two-level factors of our simulation experiment, this raises additional challenges. The methods that most effectively controlled type I error in our simulations were restricted resampling and weighted least squares (weighting by the inverse of sample variance for each group defined under the null hypothesis), but these methods are only available when data can be split into groups of identically distributed observations, not generally possible when predictors are numerical. These methods could be approximated by binning data into groups of approximately identically distributed observations, but exactly how to bin the data, and the consequences of this approximation for test properties, is currently not well understood. Of the methods we considered that can be used directly with quantitative x variables, the most reliable method we found was residual resampling around a well-chosen model, but success relies on a

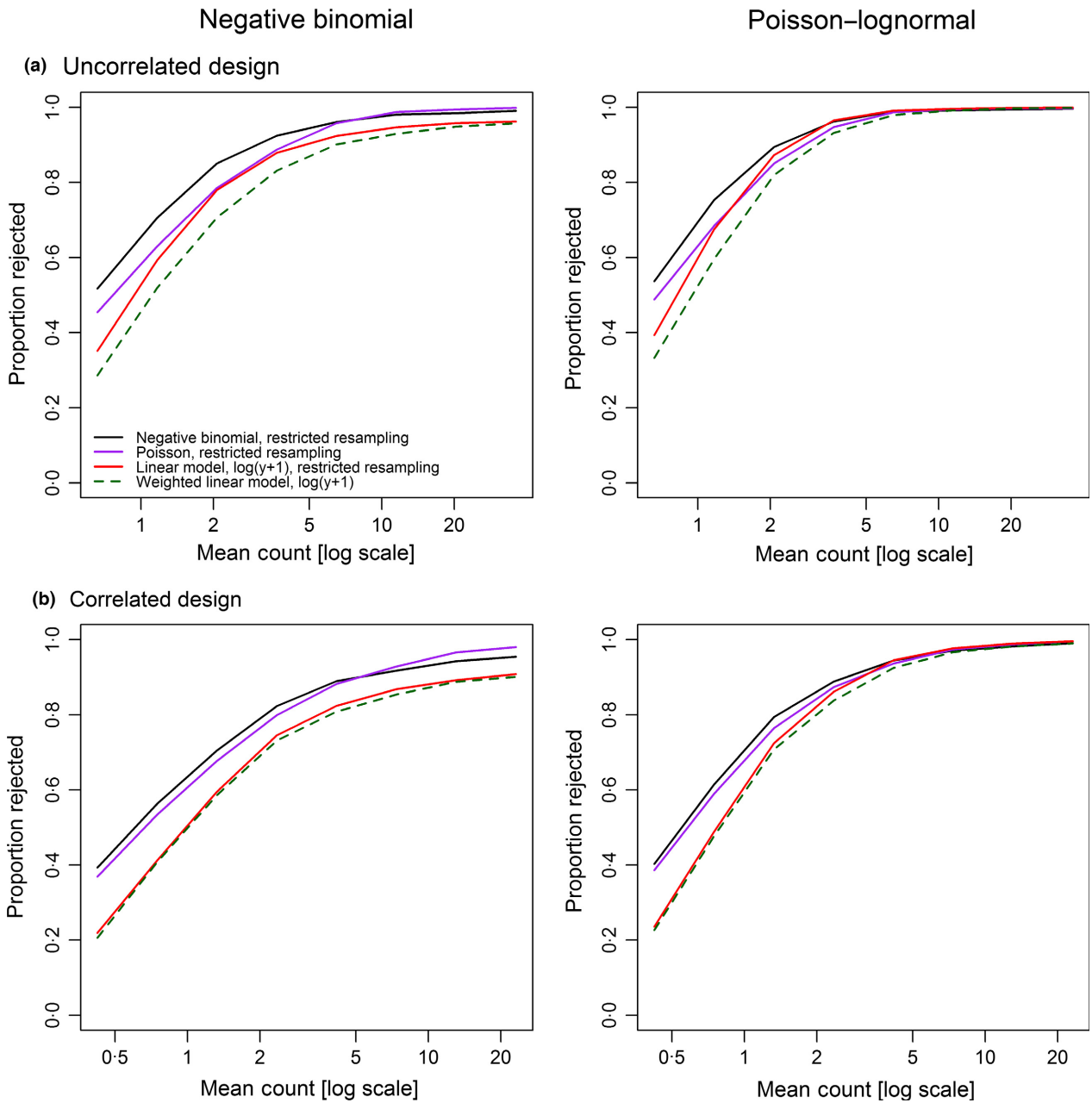

Fig. 3. Power simulation results for negative binomial counts, as a function of the mean, with a sample size of 32, for overdispersed counts sampled from a negative binomial (left column) or Poisson-lognormal (right column) distribution. Simulations used (a) uncorrelated factors or (b) correlated factors, as in Table 1. Four tests were compared: negative binomial regression, Poisson regression and log-transformed linear models all with restricted resampling and a log-transformed weighted linear model.

good choice of model for key data properties (Fig. 2), and tests are not exact.

Model choice is not easy for small samples, which may not have the resolution to diagnose model misspecification. It is made more difficult with quantitative predictors because the mean model becomes an important factor to consider in model choice – the assumed relationship between x and the mean of y can change as the model or data transformation is changed, which did not happen for example in our Fig. 2 simulations. Hence, in small samples, it is advisable to err on the side of using methods known to have some robustness to misspecification – for GLMs of counts, this usually means assuming overdispersion, accounting for it using a method-of-moments estimator (Lawless 1987), and using resampling for inference. Standard LMs, at least in uncorrelated designs, seem quite robust, and this is their main advantage.

BEYOND OUR SIMULATIONS

We selected the 2 × 2 unbalanced design because of the statistical challenges it introduced, specifically, dealing with the combined effect of heteroscedasticity and unbalanced samples. How likely is one to encounter such a situation in practice?

Heteroscedasticity will arise in count data any time the mean varies, as this implies that the variance changes too. Thus, it is out of the researcher's control, and one could reasonably expect it quite often, although data transformation can usually lessen its effect.

The extent of balance in sampling can often be controlled by the researcher through study design, especially in designed experiments, which can limit the severity of the problems LMs encountered here (Figs 2 and 3). However, sometimes, it cannot be closely controlled, particularly in field surveys with multiple predictors, or when site selection is heavily constrained (*e.g.* by availability).

On the basis of our simulation results (Fig. 3), it may be tempting to expect GLMs to always have higher power than LMs when analysing count data. In Supplementary Material, we report on additional simulations that extend Ives (2015) results, and lead to a similar conclusion. However, even for negative binomial data, the negative binomial test of Fig. 3 is not uniformly most powerful – meaning that there will be some simulation scenarios where a competing statistic, possibly even one based on LMs, has higher power. Further, sometimes power differences are quite small, especially in quite balanced designs, as in Ives (2015) or Figure S4a of Supplementary Material. But on balance a model that better suits the data will tend to have better power and is preferable for other reasons, such as interpretability.

IMPLICATIONS FOR OTHER INFERENCE TECHNIQUES

These results have important implications for analysis of multivariate abundances in ecology, where it is typically the case that many taxa, and often biologically important taxa, have small counts and many zeros in some groups of samples. As such, there is little justification for transforming data and using *ad hoc* standardizations, in place of methods that explicitly model changes in variance. Somewhat concerningly, this area is currently the one where the data transformation approach seems to be most common (*e.g.* Anderson *et al.* 2011; Legendre & De Cáceres 2013). We hope this changes in the near future, given the rapidly growing list of alternative multivariate methods based on GLMs and their extensions (Yee 2006; Dunstan, Foster & Darnell 2011; Ives & Helmus 2011; Wang *et al.* 2012; Harris 2015; Hui *et al.* 2015; Warton *et al.* 2015).

The focus in our simulations was on hypothesis testing, but often a different inferential tool is needed. For example, confidence interval estimation is more suited to problems where there is some key parameter to be estimated, or information criteria might be better for finding which of a set of competing models has strongest support. So what implications do the results of this study have for these contexts? Broadly speaking, the issues highlighted in this study arise in any context in which inference is desired. In particular, there is a one-to-one correspondence between confidence intervals and hypothesis tests (e.g. Royle & Dorazio 2008), so in situations where inaccurate type I error rates are likely, one should also expect poor coverage probability of confidence intervals. The central points from this study apply irrespective of the method of inference, specifically: model choice should be guided by data properties not the properties of ensuing inference tools; standard inference tools can be quite approximate in small samples, which can usually be addressed by changing the method of inference (*e.g.* using resampling, cross-validation,...) without changing the model; and an inference procedure based on a model that more closely reflects data properties is likely to have better properties (*e.g.* efficiency, power). This last point implies that there is no unequivocally best solution for all situations, but that a good strategy is to seek a plausible model for the data at hand.

Acknowledgements

DIW, JS and ML are supported by the Australian Research Council projects FT120100501, DP150100823 and LP150100972, respectively, and ARI is supported by the US National Science Foundation, DEB-LTREB-1052160. Thanks to editors, Philip Dixon and an anonymous reviewer for their comments.

Data accessibility

All data used in this manuscript are simulated according to code available online as Supporting Material.

References

Anderson, M.J., Crist, T.O., Chase, J.M., Vellend, M., Inouye, B.D., Freestone, A.L. *et al.* (2011) Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecology Letters*, **14**, 19–28.

Brown, B.M. & Maritz, J.S. (1982) Distribution-free methods in regression. *Australian Journal of Statistics*, **24**, 318–331.

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

Canty, A. & Ripley, B. (2014). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-13.

Carroll, R.J. & Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman & Hall, Ltd., London, UK.

Davison, A.C. & Hinkley, D.V. (1997) *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

Dunn, P. & Smyth, G. (1996) Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244.

Dunstan, P.K., Foster, S.D. & Darnell, R. (2011) Model based grouping of species across environmental gradients. *Ecological Modelling*, **222**, 955–963.

Harris, D.J. (2015) Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, **6**, 465–473.

Hilbe, J.M. (2007) *Negative Binomial Regression*. Cambridge University Press, Cambridge.

Hui, F.K., Taskinen, S., Pledger, S., Foster, S.D. & Warton, D.I. (2015) Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, **6**, 399–411.

Ives, A.R. (2015) For testing the significance of regression coefficients, go ahead and log-transform count data. *Methods in Ecology and Evolution*, **6**, 828–835.

Ives, A.R. & Helmus, M.R. (2011) Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monographs*, **81**, 511–525.

Karp, D.S., Gennet, S., Kilonzo, C., Partyka, M., Chaumont, N., Atwill, E.R. & Kremen, C. (2015) Comanaging fresh produce for nature conservation and food safety. *Proceedings of the National Academy of Sciences*, **112**, 11126–11131.

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

Legendre, P. & De Cáceres, M. (2013) Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. *Ecology Letters*, **16**, 951–963.

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

Manly, B.F.J. (2007) *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall, London, UK.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*. Chapman & Hall, London, UK.

O'Hara, R.B. & Kotze, D.J. (2010) Do not log-transform count data. *Methods in Ecology & Evolution*, **1**, 118–122.
Royle, J.A. & Dorazio, R.M. (2008) *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. Academic Press, London, UK.
Simpson, G.L. (2015). *permute: Functions for Generating Restricted Permutations of Data*. R package version 0.8-4.
Szöcs, E. & Schafer, R.B. (2015) Ecotoxicology is not normal. *Environmental Science and Pollution Research*, **22**, 13990–13999.
Væth, M. (1985) On the use of Wald's test in exponential families. *International Statistical Review*, **53**, 199–214.
Venables, W.N. & Ripley, B.D. (2013). *Modern Applied Statistics with S-PLUS*. Springer, New York, USA.
Ver Hoef, J.M. & Boveng, P.L. (2007) Quasi-Poisson vs negative binomial regression: how should we model overdispersed count data? *Ecology*, **88**, 2766–2772.
Wang, Y., Naumann, U., Wright, S.T. & Warton, D.I. (2012) `mvabund` – an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, **3**, 471–474.
Warton, D.I. (2005) Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics*, **16**, 275–289.
Warton, D.I., Foster, S.D., De'ath, G., Stoklosa, J. & Dunstan, P.K. (2015). Model-based thinking for community ecology. *Plant Ecology*, **216**, 669–682.
Wood, S. (2006) *Generalized Additive Models: An Introduction with R*. CRC press, Boca Raton, Florida, USA.
Yee, T. (2006) Constrained additive ordination. *Ecology*, **87**, 203–213.
Yee, T.W. (2015) *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer-Verlag, New York, NY.

Received 8 November 2015; accepted 26 January 2016
Handling Editor: Holger Schielzeth

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Fig. S1. Type I error simulation results for negative binomial regression tests using a method-of-moments estimator of overdispersion, and for weighted LM, for an unbalanced 2×2 design.

Fig. S2. Type I error simulation results for negative binomial data (for 2000 simulated datasets), as a function of sample size for three designs: (a) Using a regularly spaced single predictor as in Figure 1a of Ives (2015); (b) A right-skewed design where the regularly spaced predictor was log-transformed; (c) Using an unbalanced two-sample design with a 3:1 sampling ratio across the two factor levels.

Fig. S3. Type I error simulation results for Poisson-lognormal data (5000 simulated datasets), as a function of extent of overdispersion measured in two ways: (a) as the standard deviation of the lognormal variable, as in Fig. 3 of Ives (2015); (b) Back- transformed to the scale of a negative binomial overdispersion parameter using $e^{\sigma^2} - 1$.

Fig. S4. Power simulation results for negative binomial counts (2000 simulated datasets), as a function of the mean, where the predictor variable x is: (a) regularly spaced; (b) skewed (taking the logarithm of the regularly spaced values); (c) an unbalanced two-sample design, with 25% of values in one group and the remainder in the other.