Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015, **6,** 828–835 doi: 10.1111/2041-210X.12386

For testing the significance of regression coefficients, go ahead and log-transform count data

Anthony R. Ives*

Department of Zoology, University of Wisconsin-Madison, Madison, WI 53706, USA

Summary

1. The rise in the use of statistical models for non-Gaussian data, such as generalized linear models (GLMs) and generalized linear mixed models (GLMMs), is pushing aside the traditional approach of transforming data and applying least-squares linear models (LMs). Nonetheless, many least-squares statistical tests depend on the variance of the sum of residuals, which by the Central Limit Theorem converge to a Gaussian distribution for large sample sizes. Therefore, least-squares LMs will likely have good performance in assessing the statistical significance of regression coefficients.

2. Using simulations of count data, I compared GLM approaches for testing whether regression coefficients differ from zero with the traditional approach of applying LMs to transformed data. Simulations assumed that variation among sample populations was either (i) negative binomial or (ii) log-normal Poisson (i.e. log-normal variation among populations that were then sampled by a Poisson distribution). I used the simulated data to conduct tests of the hypotheses that regression coefficients differed from zero; I did not investigate statistical properties of the coefficient estimators, such as bias and precision.

3. For negative binomial simulations whose assumptions closely matched the GLMs, the GLMs were nonetheless prone to type I errors (false positives) especially when there was more than one predictor (independent) variable. After correcting for type I errors, however, the GLMs provided slightly better statistical power than LMs. For log-normal-Poisson simulations, both a GLMM and the LMs performed well, but under some simulated conditions the GLMs had high type I error rates, a deadly sin for statistical tests.

4. These results show that, while GLMs have slight advantages in power when they are properly specified, they can lead to badly wrong conclusions about the significance of regression coefficients if they are mis-specified. In contrast, transforming data and applying least-squares linear analyses provide robust statistical tests for significance over a wide range of conditions. Thus, the traditional approach of transforming data and applying LMs is still useful.

Key-words: generalized linear mixed models, generalized linear models, least-squares regression, linear models, transformation, type I errors

Introduction

The last several decades have seen a huge growth in the availability and use of statistical methods for non-Gaussian data. Generalized linear models (GLMs) that can accommodate binomial, Poisson and negative binomial data are now commonplace (McCullagh & Nelder 1989). There has been a parallel, though delayed, growth in generalized linear mixed models (GLMMs) that additionally account for non-independence in the response (dependent) variable that arise from, for example, repeated measures on the same individual or at the same location (Gelman & Hill 2007; Bolker *et al.* 2009). With this rise in methods for non-Gaussian data, there are understandable calls for discarding the traditional approach of transforming data and then applying simple linear models (LMs) (O'Hara & Kotze 2010; Warton & Hui 2011; Steel *et al.* 2013).

But should we give up on transforming non-Gaussian data and applying linear models? The answer depends on the question being asked (Bolker *et al.* 2009). Probably the most common use of statistical tests in ecology and evolution is to ask whether a response variable depends on one or more predictor (independent) variables. The traditional approach to this problem would be to transform the data and fit a least-squares linear model (LM) of the form $y = b_0 + b_1x + \varepsilon$ where ε's are assumed to be independently and identically distributed. The test statistics of the regression coefficients depend on the variance of the sum of ε's, which by the Central Limit Theorem approaches a Gaussian distribution with increasing sample size regardless of the actual distribution of ε (provided the mean and variance of ε are well defined). In fact, the

*Correspondence author. E-mail: arives@wisc.edu

approach to a Gaussian distribution can be fast, with a Gaussian approximation quite good even with sample sizes <40. Furthermore, least-squares linear regression is unbiased and has the lowest variance of all non-biased estimators of regression coefficients (Judge *et al*. 1985). Therefore, although the distribution of ε may be complicated for non-Gaussian data, statistical tests based on their sum can be very good. There are still important problems that are not overcome by the Central Limit Theorem. Because non-Gaussian distributions typically have variances (and higher moments) that depend on the mean (McCullagh & Nelder 1989), non-Gaussian data are generally transformed before analysis (Sokal & Rohlf 1981). It is then necessary to perform diagnostics on the analyses to make sure the approximation to a LM is reasonably good. Furthermore, a limitation of the transformation-LM approach is that the resulting fitted model will not correctly describe the data: it is not possible to use a strictly Gaussian model to simulate non-Gaussian data, and the parameters in the LM might not have clear biological interpretations. Nonetheless, when asking simply whether y depends on x, transforming data and analysing them with a least-squares LM might be adequate; in statistical analyses, simpler is often better (Murtaugh 2007).

Statistical methods designed for non-Gaussian data have their own limitations – both technical and cultural limitations. The technical limitations stem from the mathematical and computational complexity of GLMs and GLMMs (McCullagh & Nelder 1989; Linden & Mantyniemi 2011; Okamura, Punt & Amano 2012; Bates *et al*. 2014). For example, incorporating complicated correlation structures into GLMMs, such as spatial and phylogenetic correlations, poses challenges (e.g. Ives & Helmus 2011; Ross, Hooten & Koons 2012; Ives & Garland 2014). Similarly, many of the statistical results provided by software packages for GLMs and GLMMs are themselves approximations. For example, the P-values given for coefficient estimates of GLMs from glm {stats} in R (R Core Team 2014) are based on Gaussian approximations (Wald z tests); thus, even though the model is exact, the statistical inference that researchers use is approximate. Statistical inference about parameters can also be derived from likelihood ratio tests, although these are also based on asymptotic approximations.

The cultural limitation of GLMs and GLMMs is that, because they model specific processes that a researcher might assume underlie the data, they are accepted as correct without careful attention to diagnostics (but see Gelman & Hill 2007; Bolker 2008; Bates *et al*. 2014). GLMs and GLMMs make explicit assumptions about the distribution of the data, and if these assumptions are not met by the data, the statistical results could be quite wrong (e.g. Martin *et al*. 2005; O'Hara 2005; Ver Hoef & Boveng 2007). Of course, the same is true for the transformation-LM approach. The cultural limitation is that researchers are often less cautious with GLMs and GLMMs (because these methods are 'correct') than they are with LMs applied to transformed data (which are obviously not correct).

Here, I compare statistical results from models for non-Gaussian count data with results from LMs performed on transformed data. The first set of comparisons use data simulated from a negative binomial distribution. These comparisons are motivated by the article prescriptively entitled 'Do not log-transform count data' in which O'Hara & Kotze (2010) fit negative binomial data with GLMs and with LMs following log-transformation. O'Hara and Kotze show that GLMs give better estimates of the untransformed expectations (means) than LMs; the GLM estimates have less bias and more precision. In contrast to asking about the ability of different methods to estimate expectations, I modified O'Hara and Kotze's code to produce a regression problem and asked which methods give correct type I error rates and have the greatest statistical power to identify statistically significant regression coefficients. Therefore, the statistical task I address is different from that of O'Hara & Kotze (2010), and as I show, the conclusions differ too.

The second set of comparisons addresses a situation that frequently arises in ecology: when there are multiple populations experiencing different per capita population growth rates, and each population is randomly sampled. Because population growth is multiplicative, a reasonable a priori assumption is that variation among populations is approximately log-normally distributed. Random sampling then takes place in the form of a Poisson distribution. This situation leads to a hierarchical mixture model of Gaussian 'process variation' in log abundances among populations and Poisson 'measurement variation' of those abundances. I simulated data from this hierarchical model and fit GLMs, LMs after transforming the data, and a log-normal-Poisson GLMM that matched the simulation model. The question is whether applying GLMs to ecological count data is better than transforming data and using a LM for testing significance of regression coefficients.

Materials and methods

NEGATIVE BINOMIAL SIMULATIONS

Data sets were simulated from a negative binomial distribution whose mean μ depends on a predictor variable x according to

$$\mu = \exp(b_0 + b_1 x) \qquad \text{eqn 1}$$

This contrasts with the simulation model used by O'Hara & Kotze (2010) in which there was no x variable, and the statistical task was to estimate μ. In equation 1, x is assumed to take on n evenly spaced values, scaled so that the mean is zero and the variance is one. This makes it possible to compare the effect of sample size n (n = 20 to 1000) without changing the mean and variance of x. The values of the negative binomial dispersion parameter θ were selected to generate distributions ranging from highly aggregated (θ = 0·25) with variances much higher than the mean to non-aggregated (θ = 100) when the negative binomial approaches the Poisson distribution and the variance equals the mean. Values of b_0 were investigated between log(0·2) and log(10) to vary the mean from 0·2 to 10. Finally, to test the power of the methods to reject the null hypothesis H_0:$b_1 = 0$, values of b_1 in the simulations were investigated between 0 and 0·8. The baseline parameter values that were fixed while each of the parameters was varied in turn were n = 100, θ = 1, $b_0 = 0$ and $b_1 = 0$. For each combination of parameter values, either 2000 (when b_1 was varied to test for power) or 50 000 (when varying the other parameters) simulated data sets were produced.

The models fit to the simulated data overlap with those used by O'Hara & Kotze (2010):

1 GLM in which y follows a negative binomial distribution;

2 GLM in which y follows a quasi-Poisson distribution to allow for greater-than-Poisson variances;

3 LM with a log(max(y, 0·5)) transformation, in which zeros are replaced by 0·5;

4 LM with a log(y + 1) transformation, in which one is added to all values;

5 LM with a log(y + 0·0001) transformation, in which 0·0001 is added to all values;

6 LM with a sqrt(y) transformation; and

7 LM with no transformation of y, but using exp(x) as the predictor variable.

These models were selected to have a range of characteristics. Model 1 has the same distributional assumptions as the negative binomial simulations. Model 2 matches the negative binomial simulations when θ is large and the negative binomial approaches a Poisson distribution; when θ is small (giving an aggregated distribution), the dispersion parameter of the quasi-Poisson distribution should account for the increased variance. Models 3 and 4 use standard variants of log transforms. Model 5 is a non-standard log-transformation that might be (incorrectly) assumed to perform well because it changes the data the least; in fact, it changes the data more than other log-transformations, because the value of log(0·0001) is much smaller than log(1·0001). Model 6 is a 'bad' transform that will match neither the relationship between μ and x, nor the relationship between the variance of ε and x. Finally, model 7 matches the relationship between μ and x, but the assumption about the relationship between the variance of ε and x is particularly bad.

For each simulated data set, all seven models were fit and used to test the null hypothesis H_0:$b_1 = 0$ at the $\alpha = 0{\cdot}05$ significance level using the glm {stats}, glm.nb {MASS} (Venables & Ripley 2002) or lm {stats} functions in R (R Core Team 2014). For the negative binomial model 1, significance of b_1 was tested using both a Wald test (given by glm.nb) and a likelihood ratio test (LRT). Both Wald and LRTs were tried for the quasi-Poisson model 2, although the performance of the Wald test was uniformly better and therefore only presented here. The proportion of the simulated data sets for which the null hypothesis was rejected gives the type I error rate if $b_1 = 0$ in the simulations, or the power of the tests if $b_1 \neq 0$.

Ecological and evolutionary data will often contain multiple predictor variables, raising the problem of multicollinearity in identifying which predictor variables are associated with variation in the response variable. To address this, I simulated negative binomial data with

$$\mu = \exp(b_0 + b_1 x_1 + b_2 x_2) \qquad \text{eqn 2}$$

where x_1 and x_2 are drawn from a bivariate normal distribution with correlation coefficient r ranging from 0 to 0·9. After picking values of x_1 and x_2, both variables were standardized to have mean zero and variance one. Note that in this procedure, x_1 and x_2 were randomly generated for each data set, whereas for equation 1 x was taken at even increments. For the bivariate simulations, model 7 was not fit, because x_1 and x_2 are not additive. In the simulations, $b_2 = 1$ which represents a strong effect of x_2 on y. Therefore, when the correlation r between x_1 and x_2 is high, multicollinearity could affect the test of H_0:$b_1 = 0$.

LOG-NORMAL-POISSON HIERARCHICAL SIMULATIONS

Because the total growth rate of a population is proportional to the population size, Gaussian variation in per capita population growth rates leads to approximately log-normal variation in population sizes when populations are large. To incorporate both log-normal variation in population size (process variation) and random sampling (measurement variation), I simulated data with the hierarchical model

$$\begin{aligned} \lambda &= \exp(b_0 + b_1 x + \varepsilon) \\ y &\sim \text{Poisson}(\lambda) \end{aligned} \qquad \text{eqn 3}$$

where ε is a Gaussian random variable with mean zero and standard deviation σ. Simulations were performed with σ ranging from 0 to 2 with $b_1 = 0$ to test for type I errors. Each simulated data set was fit with the models used previous, excluding model 5, plus a GLMM with the same form as the simulation model (Eqn. 3) fit using glmer (Bates *et al.* 2014).

Results

NEGATIVE BINOMIAL DATA

The GLMs (models 1 and 2) were more prone to type I errors (false positives) than LMs with transformed data (Fig. 1). Type I error is an especially unwanted property of any statistical method, because it will lead to false conclusions about patterns in the data. For the negative binomial GLM, the LRT provided better type I errors than the Wald test; this is not surprising, because the Wald test is in general a better approximation than a LRT (Engle 1984). Even using the LRT, type I errors were serious for the negative binomial GLM when sample size was small (Fig. 1a) or dispersion in the simulated data was large (small θ, Fig. 1b). This is surprising, because the data were simulated using a negative binomial distribution. The quasi-Poisson GLM often had better performance than the negative binomial GLM, although it had highly inflated type I errors in the case of multiple regression (Fig. 1d).

All of the LMs with transformed data had very acceptable type I error rates. This was true even for $b_0 = \log(0{\cdot}2) = -1{\cdot}61$ (Fig. 1c), when the mean of y is 0·2 and on average 83% of the data points in the simulations were zeros. For the case of multiple regression, even highly correlated predictor variables did not exaggerate the type I error rates for the LMs (Fig. 1d), showing that least-squares LMs with transformed data were relatively insensitive to multicollinearity for the sample size of $n = 100$. The uniformly acceptable performance of the LMs even for 'bad' transformations (models 6 and 7) that violate the true relationship between the mean of y and x, and between the variance of y and x, shows the robustness of LMs.

In the simulations to determine power, b_1 ranged from 0 to 0·8, and over this range, the proportion of data points equalling zero remained close to 0·5 for the baseline value of $\theta = 1$. Because the LRT gave better type I error rates than the Wald test for the negative binomial GLM, I only included the LRT results. Furthermore, because even the LRT had inflated type I errors (Fig. 1a), I used a likelihood ratio critical value of 3·98 rather than the critical value of 3·84 given by the standard LR χ^2 approximation; this value gave a correct type I error rate of 0·05. The negative binomial and quasi-Poisson GLM models (models 1 and 2) had the greatest power to reject the hypothesis H_0:$b_1 = 0$ at the $\alpha = 0{\cdot}05$ significance level (Fig. 2).

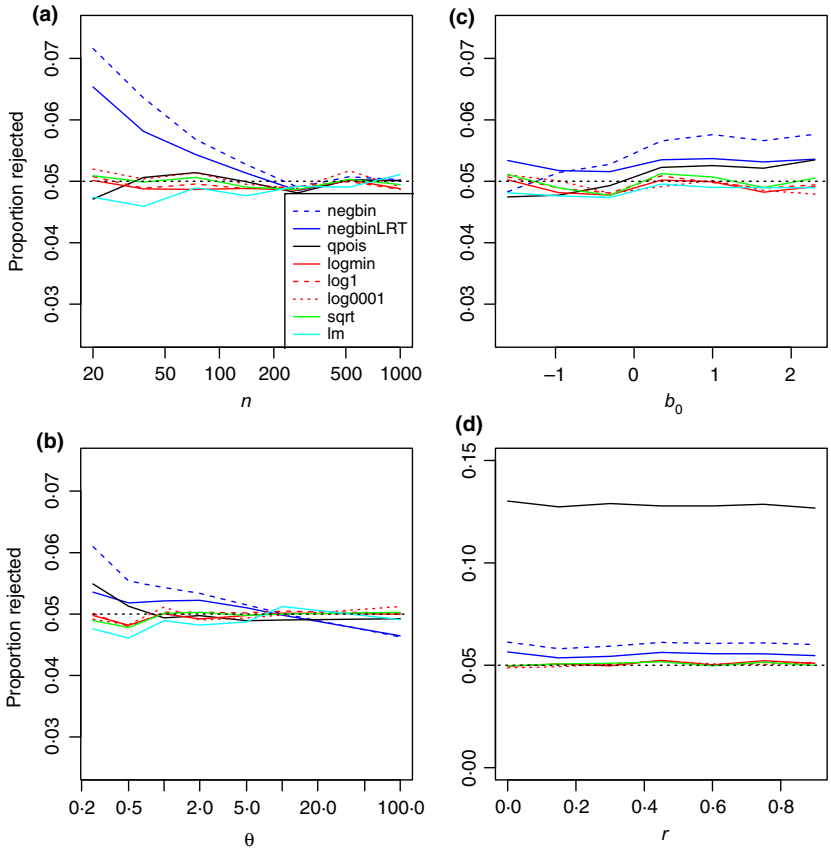

Fig. 1. Type I error rates (false positives) for the null hypothesis H_0:$b_1 = 0$ at the $\alpha = 0{\cdot}05$ significance level for seven regression models while varying (a) the sample size n, (b) the dispersion parameter of the negative binomial distribution θ, (c) the intercept b_0 and (d) the correlation between two predictor variables x_1 and x_2 for a multiple regression model. At each value of the specified variable, 50 000 data sets were simulated from a negative binomial model (a–c, Eqn. 1; d, Eqn. 2). Tests of H_0:$b_1 = 0$ were taken from the standard output from glm {stats}, glm.nb {MASS} and lm {stats} in R, and in addition, a LRT was used for glm.nb. The black dotted line gives the nominal 0·05 level for which 5% of the simulated data sets should be rejected at the $\alpha = 0{\cdot}05$ significance level. The baseline parameter values that were fixed while changing the specified parameter values in a–d were $b_1 = 0$, $n = 100$, $\theta = 1$, $b_0 = 0$ and for d $b_2 = 1$. Lines correspond to model 1 with a Wald test (negbin), model 1 with a LRT (negbinLRT), model 2 (qpois), model 3 (logmin), model 4 (log1), model 5 (log0001), model 6 (sqrt) and model 7 (lm).

Nonetheless, the power of the LMs with transformed response variables log(max(y, 0·5)) and log(y + 1) (models 3 and 4) was not much lower; for example, the chances of rejecting H_0: $b_1 = 0$ when the simulation value of $b_1 = 0{\cdot}3$ was 0·57 and 0·56 for the negative binomial and quasi-Poisson models, and 0·52 and 0·51 for the LMs with log(max(y, 0·5)) and log(y + 1) transformations. The LM with a sqrt(y) transformation and the LM with y untransformed and exp(x) (models 6 and 7) were also not bad. Only the LM with log(y + 0·0001) (model 5) showed a large loss of power, although I included this model only to illustrate that extremely bad transformations can cause loss of performance. For standard log(max(y, 0·5) and log (y + 1) transformations, however, the performance of LMs is close to that of GLMs.

The (slight) loss of power of the LMs is caused because they do not properly account for the point-by-point variance in the data. Specifically, while LS linear regression minimizes the sum of squared differences between predicted and observed values, GLMs minimize the squared deviances, where the deviances account for the non-Gaussian distributions of the data. To explain this heuristically with an extreme example, consider the case in which y is perfectly predicted by x. If y has a negative binomial distribution, then there will still be variance in y given x (i.e. non-zero deviances), and this is correctly included in the GLM. In contrast, if y has a Gaussian distribution and y is perfectly predicted by x, then there will be no residual variation in y given x. The LM will reward this case by giving strong statistical significance to b_1. However, in the case of y having a negative binomial distribution, the LM will interpret the (unavoidable) variance in y given x as unexplained variation and tax the statistical significance of b_1 accordingly. This taxation in turn leads to a reduction in the chances that the LM test will reject the null hypothesis.

LOG-NORMAL-POISSON HIERARCHICAL DATA

The log-normal-Poisson model simulates data that are common in ecological and evolutionary studies: samples are taken randomly from multiple populations that experience variation in per capita population growth rates. For the simulated data, on average 40% of the data points were zeros. All of the LMs (models 3, 4, 6 and 7) had very acceptable type I error rates. In

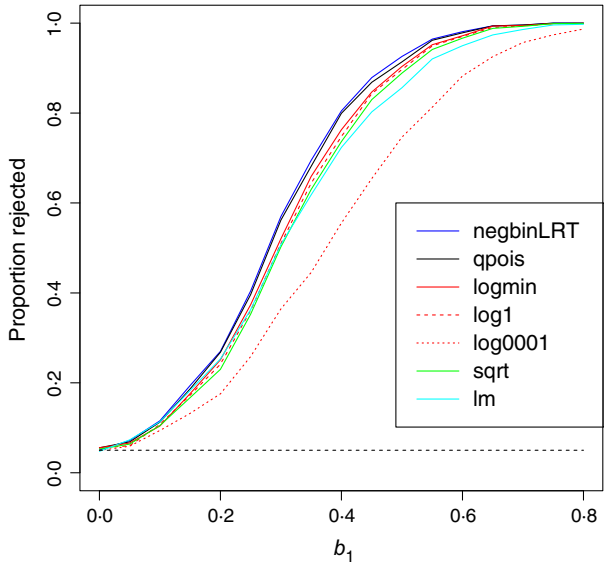

Fig. 2. Statistical power to reject the null hypothesis H_0:$b_1 = 0$ at the $\alpha = 0{\cdot}05$ significance level for seven regression models. At each value of b_1 ($b_1 = 0, 0{\cdot}05, 0{\cdot}1, \ldots, 0{\cdot}8$), 2000 data sets were simulated from a negative binomial model (Eqn. 1). Tests of H_0:$b_1 = 0$ were taken from the standard output from glm {stats} and lm {stats} in R, and for the negative binomial GLM (glm.nb), a LRT was performed with a threshold of 2*LR = 3·98 (rather than 3·84) to give correct type I errors. The black dashed line gives the nominal 0·05 level for which 5% of the simulated data sets should be rejected at the $\alpha = 0{\cdot}05$ significance level. Other parameter values were $n = 100$, $\theta = 1$ and $b_0 = 0$. Lines correspond to model 1 with a LRT (negbinLRT), model 2 (qpois), model 3 (logmin), model 4 (log1), model 5 (log0001), model 6 (sqrt) and model 7 (lm).

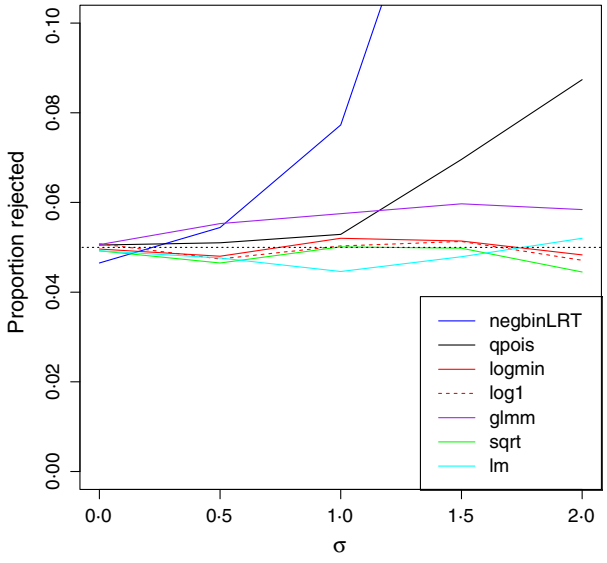

Fig. 3. Type I error rates for the null hypothesis H_0:$b_1 = 0$ at the $\alpha = 0{\cdot}05$ significance level for simulated data from a hierarchical log-normal-Poisson model (Eqn. 3). At each value of the standard deviation of the log-normal distribution σ, 10 000 data sets were simulated. Tests of H_0:$b_1 = 0$ were taken from the standard output from glm {stats}, lm {stats} and glmer {lme4} in R, and for the negative binomial GLM (glm.nb), a LRT was performed. The black dotted line gives the nominal 0·05 level for which 5% of the simulated data sets should be rejected at the $\alpha = 0{\cdot}05$ significance level. The fixed parameter values were $n = 100$, $\sigma = 1$, $b_0 = 0$ and $b_1 = 0$. Lines correspond to model 1 with a LRT (negbinLRT), model 2 (qpois), model 3 (logmin), model 4 (log1), a log-normal-Poisson GLMM (glmm), model 6 (sqrt) and model 7 (lm).

fact, the type I error rates were better than those for the log-normal-Poisson GLMM, which had slightly inflated type I error rates even though the statistical model matched the simulation model used to produce the data. Both of the GLMs had unacceptably high type I errors for larger variation among populations (σ, Fig. 3). Apparently, both GLMs are incapable of accounting for the log-normal process variation among populations. This is surprising for the negative binomial GLM, because the log-normal distribution is similar to a gamma distribution, and a hierarchical gamma-Poisson distribution is in fact a negative binomial distribution.

Because the type I errors for the negative binomial and quasi-Poisson GLMs were so poor, I did not produce power curves. If type I errors are seriously inflated, you should abandon the statistical model or use simulation techniques to correct the type I error rates. Inflated type I error rates will also falsely inflate power, so any apparent performance benefit over other statistical models is immediately suspect.

The surprising type I errors of the negative binomial GLM could have several causes: (i) incorrect specification of the relationship between the mean of y and x (linearity), (ii) incorrect specification of the relationship between the mean and variance (scedasticity), (iii) incorrect specification of the higher moments of the distribution. To investigate these, I simulated both negative binomial and log-normal-Poisson data, and fit the data with a negative binomial GLM (Fig. 4). Both simulation models have the same relationship between the mean of y and x given by equations 1 and 3, and they have the same relationship between the mean and variance (Fig. 4a) which can be proved analytically (Ben Bolker, McMaster University, *pers. comm.*). Therefore, any difference in the performance of the negative binomial GLM between the negative binomial and log-normal-Poisson simulations must be due to (iii), model mis-specification of the skew and higher moments. Indeed, the skew (Fig. 4b) and higher moments (not shown) of the log-normal-Poisson data were lower than for the negative binomial data. Thus, the negative binomial GLM mis-specifies the skew and higher moments of the log-normal-Poisson data.

The consequence of this mis-specification is that the negative binomial GLM underestimates the variance of the log-normal-Poisson distribution. To illustrate this, I simulated both negative binomial and log-normal-Poisson data sets across a range of variation ($\theta = 0{\cdot}25, 0{\cdot}5, 1, 5, 100$; $\sigma = 0, 0{\cdot}25, 0{\cdot}5, 0{\cdot}75, 1$) and for each data set computed the observed variance/mean^2 (V/M^2) as a measure of variation. For each data set, I then fit the GLM to estimate the dispersion parameter θ and compute the estimated V/M^2 under the negative binomial assumption of the GLM, namely $V/M^2 = (1 + M/\theta)/M$. Plotting the average estimated V/M^2 for each value of θ or σ against the observed V/M^2 shows that the fitted negative binomial GLM correctly describes the variance of the negative binomial data sets, but underestimates the variance of the log-normal-Poisson

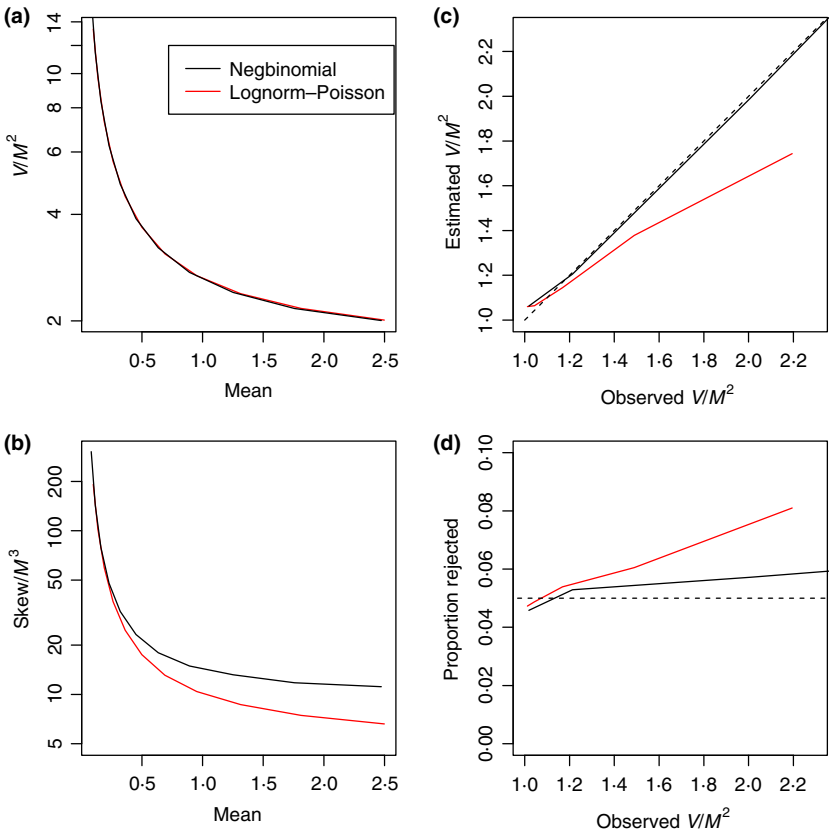

Fig. 4. Negative binomial (Eqn. 1) and log-normal-Poisson (Eqn. 3) simulation data and statistical fits of the negative binomial GLM. Data were simulated with no predictor (independent) variable (i.e. $b_1 = 0$). (a) The relationship of the variance/mean^2 ratio (V/M^2) to the mean, and (b) the relationship of the skew/mean^3 ratio (Skew/M^3) to the mean for simulated negative binomial (black) and log-normal-Poisson (red) data. For the negative binomial simulations, $\theta = 0{\cdot}62$ and b_0 ranges from log(0·1) to log(2·5); for the log-normal-Poisson simulations, $\sigma = 1$ and b_0 ranges from log (0·05) to log(1·5). (c) The mean estimated $V/M^2 = (1 + M/\theta)/M$ versus the mean observed V/M^2 from the data sets. The black dashed line gives the 1:1 line. (d) Type I errors for the negative binomial GLM applied to the simulated negative binomial (black) and log-normal-Poisson (red) data versus the observed V/M^2. Tests of $H_0{:}b_1 = 0$ were performed with a LRT using output from glm.nb {MASS} in R. The black dashed line gives the nominal 0·05 level for which 5% of the simulated data sets should be rejected at the $\alpha = 0{\cdot}05$ significance level. For c and d, 10 000 data sets were simulated at each value of the parameters governing dispersion in the negative binomial ($\theta = 0{\cdot}25$, 0·5, 1, 5, 100) and log-normal-Poisson ($\sigma = 0$, 0·25, 0·5, 0·75, 1) models. The other parameter values were $n = 100$ and $b_0 = 0$.

data sets (Fig. 4c). Presumably, this is due to the smaller skew and higher moments of the log-normal-Poisson relative to the negative binomial distribution (Fig. 4b). The type I errors are higher for the log-normal-Poisson data sets, becoming worse as the observed V/M^2 increases (Fig. 4d), which coincides with greater underestimates of the observed V/M^2 (Fig. 4c). Thus, the likely source for inflated type I error rates of the negative binomial GLM is the smaller skew and higher moments of the log-normal-Poisson data that lead to underestimates of the variance; the resulting low estimates of the variance in the GLM imply stronger information (less variability) in the data than really exists.

Discussion

For detecting a dependence of a response y variable to a predictor x variable, the strategy of transforming count data and using least-squares LMs proved to be pretty good. When applied to count data generated from a negative binomial distribution, the type I error rates were correct (Fig. 1), and there was only slight loss of statistical power compared to GLMs (Fig. 2) provided a sensible log-transformation was used (either log(max(y, 0·5)) or log(y + 1)). Similarly, when applied to hierarchical data with log-normal variation among populations that are sampled by a Poisson distribution, the type I error rates of LMs with transformed data were actually better than those of a log-normal-Poisson GLMM (Fig. 3). These results suggest that if you are interested in the significance of regression coefficients, you can go ahead and log-transform count data.

In contrast, the GLMs sometimes performed worse than the LMs, sometimes much worse. The negative binomial GLM had surprisingly inflated type I error rates for small sample sizes (Fig. 1a). Because LM tests rely on the Central Limit Theorem, LMs might be expected to perform worse than GLMs when sample sizes are small, but this is the opposite from what the simulations show. When used for multiple regression in which one predictor variable had a strong effect ($b_2 = 1$), the quasi-Poisson GLM test for significance of the other predictor variable had highly inflated type I errors

(Fig. 1d). Finally, both negative binomial and quasi-Poisson GLMs had unacceptable type I error rates when applied to the log-normal-Poisson hierarchical data (Fig. 3). These results suggest that if you are going to apply GLMs to count data, make sure you perform diagnostics and simulations to confirm your results.

This is not to say that GLMs are not useful, just that they should be used cautiously. For example, if the goal is to fit a model that can be used to simulation count data, then GLMs must be used. In contrast, the approach of log-transforming count data and applying a least-squares LM gives the expectation of the transformed variable y for a given value of x, but does not give a fully specified model of a stochastic process underlying count data. Thus, while the meaning of the regression coefficients for transformed data in a LM is clear, it is not clear how these relate to the back-transformed data. Furthermore, at least for the mean b_0, O'Hara & Kotze (2010) show that the back-transformed estimates are not as good as those from GLMs. Nonetheless, GLMs make more and somewhat more-hidden assumptions about the distribution of the data than do LMs. For example, a negative binomial GLM assumes that the variance scales according to $\mu(1 + \mu/\theta)$, while the quasi-Poisson GLM assumes the variance scales according to $\phi\mu$ where ϕ is a dispersion parameter; this contrast can give different results between these two GLMs (Ver Hoef & Boveng 2007). In the case investigated here comparing negative binomial and log-normal-Poisson data, the variance-to-mean relationship was the same for both distributions, yet the log-normal-Poisson data had lower skew and other moments. The estimate of the parameter θ from the negative binomial GLM consistently underestimated the observed variance of the log-normal-Poisson data, and this apparently inflated the type I error rates. Thus, even though the negative binomial GLM assumed the correct variance-to-mean ratio, it nonetheless did not fit the distributional characteristics (skew and higher moments) of the data well enough to give accurate significance tests for the regression coefficient.

Even if the assumptions of a GLM match the data, the tests that they provide for statistical significance of the regression coefficients are based on asymptotic assumptions guaranteed to be accurate only as sample sizes get large. Therefore, unlike the LMs applied to transformed data in which the model is approximate but the significance tests are exact, for GLMs, the model might be exact but the significance tests are still approximate. Furthermore, different approximate statistical tests have different performance. For example, the LRT for the negative binomial GLM gave more reliable type I errors than the Wald test, although the LRTs still gave worse type I errors than the log-transformation LMs.

For any regression model, LM or GLM, it is necessary to perform diagnostics to check (i) the relationship between the mean of y and x; (ii) the relationship between the mean and variance (scedasticity) and (iii) the higher moments of the distribution of residuals. For LMs linearity can be corrected using transforms, although these will also affect scedasticity. To correct for heteroscedasticity without changing linearity, weighted least-squares regression (or equivalently generalized least squares) can be used. For least-squares LMs, the only way to address higher moments is to have large sample sizes so that the sum of residuals is more Gaussian. For GLMs, linearity can be corrected without changing scedasticity by using a different link function (McCullagh & Nelder 1989; Bolker 2008). Although GLMs make rigid distributional assumptions about the relationship between the variance (and higher moments) and mean, there are several GLMs or GLMMs that can be tried; for example, the log-normal-Poisson GLMM provided much better tests than the GLMs when applied to data with less-than-negative-binomial skews (Fig. 3). Hopefully, the results presented here encourage careful selection of GLMs. This is where GLMs and GLMMs have a huge advantage over LMs. Because GLMs and GLMMs give a model fit to the data, they can be used to simulate data to explore their type I error rates, estimation bias and investigate other important statistical properties.

I have only considered count data given by Poisson-related distributions in application to regression. Nonetheless, the same conclusions about the performance of least-squares LMs applied to transformed data likely apply to other types of data and analyses. A pair of papers by Johnson, Campbell and Capuano (Campbell, Young & Capuano 1999; Young, Campbell & Capuano 1999) consider ANOVAS applied to count data and show that LMs have better type I error rates than GLMs for identifying treatment effects. Warton & Hui (2011) use simulations to evaluate statistical tests for differences in the means of two samples of binomially distributed data, either without or with additional overdispersion generated by using a logit-normal-binomial hierarchical model (their Fig. 3, Appendix S1C–F). Their logistic (binomial) GLM showed inflated type I errors when either sample sizes were small or the data were overdispersed, while their logitnormal-binomial GLMM showed inflated type I errors when there were both small sample sizes and overdispersion. In contrast, applying LMs to arcsine-transformed data, or even untransformed data, showed appropriate type I error rates, much better than either GLM or GLMM. To correct for type I errors shown by GLMs and GLMMs, Warton and Hui recommend bootstrapping. Once GLMs and GLMMs are corrected for type I errors using bootstrapping, they have only slightly greater power compared to LMs (Warton & Hui 2011; Appendix S1F). Therefore, even though Warton & Hui (2011) conclude that using the LMs is 'asinine', if the goal is to test for differences in means among samples, then their results show that LMs give a robust approach.

Although I only investigated GLMs in detail, there could potentially be similar problems with GLMMs that are hard to diagnose. GLMMs are considerably more complex computationally and statistically than their Gaussian linear mixed model (LMM) counterparts, which suggests transforming data and applying LMMs might be a good strategy when significance tests are all that is needed. Furthermore, while there are new methods for GLMMs that can incorporate complex spatial (e.g. Venables & Ripley 2002; Ross, Hooten & Koons 2012) and phylogenetic (e.g. Hadfield 2010; Ives & Helmus 2011) correlations, these can be difficult to both apply and

validate. In a simulation study of statistical methods for phylogenetically correlated binary (binomial) data, the simple approach of applying a phylogenetic LMM and ignoring the binary nature of the data gave better tests for the significance of regression coefficients than GLMMs and other phylogenetic methods designed for binary data (Ives & Garland 2014).

This work was motivated by a reviewer of a manuscript which included analyses of the distribution of species abundances among multiple sites with spatial correlation. The reviewer sympathetically acknowledged the absence of a simple way of doing the analyses with GLMMs but nonetheless asked for justification of our approach of transforming data and using a LMM. Before performing the analyses presented here, I expected that LMs with transformed data would show good type I error rates and moderate loss of power compared to GLMs and GLMMs, but I was not expecting the sometimes badly inflated type I error rates of the GLMs. These simulations show that the inappropriate application of GLMs can lead to wrong statistical conclusions. Nonetheless, this article should not be used as an argument against the appropriate use of GLMs and GLMMs. I will continue to use GLMs and GLMMs when I want a completely specified model of a statistical process, but now I will use even more caution.

Acknowledgements

I thank Ben Bolker, Philip Dixon, Bob O'Hara and David Warton for wonderfully insightful comments that helped to clarify this article. Financial support came from the US National Science Foundation, DEB-LTREB-1052160.

Data accessibility

No data were used in this article.

References

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2014) Lme4: Linear Mixed-Effects Models Using Eigen and S4. http://CRAN.R-project.org/package = lme4 [version 1.1-7].

Bolker, B.M. (2008) *Ecological Models and Data in R*. Princeton University Press, Princeton and Oxford.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.S.S. (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, **24**, 127–135.

Campbell, N.L., Young, L.J. & Capuano, G.A. (1999) Analyzing over-dispersed count data in two-way cross-classification problems using generalized linear models. *Journal of Statistical Computation and Simulation*, **63**, 263–281.

Engle, R.F. (1984) Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handbook of Econometrics*. Vol. 2 (eds Z. Griliches & M. Intriligator), pp. 775–826. Elsevier, Amsterdam.

Gelman, A. & Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY.

Hadfield, J.D. (2010) MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, **33**, 1–22.

Ives, A.R. & Garland, T. Jr (2014) Phylogenetic regression for binary dependent variables. *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology* (ed. L.Z. Garamszegi), pp. 231–261. Springer-Verlag, Berlin Heidelberg.

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

Judge, G.G., Griffiths, W.E., Hill, R.C., Lutkepohl, H. & Lee, T.-C. (1985) *The Theory and Practice of Econometrics*, 2nd edn. John Wiley & Sons, New York.

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

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**, 1235–1246.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*, 2nd edn. Chapman and Hall, London.

Murtaugh, P.A. (2007) Simplicity and complexity in ecological data analysis. *Ecology*, **88**, 56–62.

O'Hara, R.B. (2005) Species richness estimators: how many species can dance on the head of a pin? *Journal of Animal Ecology*, **74**, 375–386.

O'Hara, R.B. & Kotze, D.J. (2010) Do not log-transform count data. *Methods in Ecology and Evolution*, **1**, 118–122.

Okamura, H., Punt, A.E. & Amano, T. (2012) A generalized model for overdispersed count data. *Population Ecology*, **54**, 467–474.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ross, B.E., Hooten, M.B. & Koons, D.N. (2012) An accessible method for implementing hierarchical models with spatio-temporal abundance data. *PLoS ONE*, **7**, e49395. doi:10.1371/journal.pone.0049395.

Sokal, R.R. & Rohlf, F.J. (1981) *Biometry*, 2nd edn. W. M. Freeman & Company, New York.

Steel, E.A., Kennedy, M.C., Cunningham, P.G. & Stanovick, J.S. (2013) Applied statistics in ecology: common pitfalls and simple solutions. *Ecosphere*, **4**, 115. doi: 10.1890/ES13-00160.1.

Venables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics with S*. Springer, New York.

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.

Warton, D.I. & Hui, F.K.C. (2011) The arcsine is asinine: the analysis of proportions in ecology. *Ecology*, **92**, 3–10.

Young, L.J., Campbell, N.L. & Capuano, G.A. (1999) Analysis of overdispersed count data from single-factor experiments: a comparative study. *Journal of Agricultural Biological and Environmental Statistics*, **4**, 258–275.

Received 3 January 2015; accepted 27 March 2015
Handling Editor: Robert Freckleton

Supporting Information

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

Appendix S1. R scripts used for the simulations and analyses.