Methods in Ecology and Evolution

Revisiting advice on the analysis of count data

Michael B. Morrissey and Graeme D. Ruxton

30 January, 2020

Dyers Brae House

School of Biology

University of St Andrews

St Andrews, Scotland, KY16 9TH

`michael.morrissey@st-andrews.ac.uk`

`graeme.ruxton@st-andrews.ac.uk`

Author contributions: MBM and GDR both contributed to identifying the key issue discussed in this paper, to constructing instructive scenarios to illustrate the issue, to developing our re-interpretation of the earlier results, and to writing the manuscript.

Abstract

(1) O'Hara and Kotze (2010; Methods in Ecology and Evolution 1: 118-122) present simulation results that appear to show very poor behaviour (as judged by bias and overall accuracy) of linear models (LMs) applied to count data, especially in relation to generalised linear model (GLM) analysis.

(2) We considered O'Hara and Kotze's (2010) comparisons, and determined that the finding occurred primarily because the quantity that they estimated in their simulations of the LM analysis (the mean of a transformation of the count data) was not the same quantity that was simulated and to which the results were compared (the logarithm of the mean of the count data). We correct this discrepancy, re-run O'Hara and Kotze's simulations, and add additional simple analyses.

(3) We found that the apparent superiority of the GLMs over LMs in O'Hara and Kotze's (2010) simulations was primarily an artefact of divergence in the meanings of results from the two analyses. After converting results from LM analyses of transformed data to estimators of the same quantity as provided by the GLM, results from both analyses rarely differed substantially. Furthermore, under the circumstances considered by O'Hara and Kotze, we find that an even simpler implementation of LM analysis, inference of the mean of the raw data, performs even better, and gives identical results to the GLM.

(4) While the analysis of count data with generalised linear models can certainly provide many benefits, we strongly caution against interpreting O'Hara and Kotze's (2010) results as evidence that simpler approaches are severely flawed.

Introduction

Many variables of interest in statistical analyses of biological data come from non-normal distributions. These variables may be most appropriate to analyse with generalised linear models (GLMs; Nelder and Wedderburn 1972, McCullagh and Nelder 1989). It has become increasingly common in the last two decades for biologists to employ GLMs, and in fact strong opinions have developed that earlier approaches to dealing with non-normal variable types are likely to be highly inappropriate. A key example is the analysis of count variables, i.e., of quantities that take non-negative integer values, such as counts of offspring or counts of behaviours. Models with count variables as responses might previously have used linear models (LMs; or methods subsumed by linear models) fitted using ordinary least squares (OLS) methods, either of untransformed counts, or after transformation using one of several methods. Transforming counts by logging (generally after adding a value of one, to avoid taking the log of any zero counts) was very common (Sokal and Rohlf 1995). In recent years, it has been more common to use GLMs that model errors in models of count variables using the Poisson distribution, or to use use other, even more flexible, models for the error structure, for example, GLMs employing the negative binomial distribution. The general expectation of clear superiority of GLMs is encapsulated in the title of a much-cited paper by O'Hara and Kotze (2010): "Do not log-transform count data". These authors' definitive advice is based in very large part on a simulation study comparing the two approaches, and appears to reveal catastrophic performance of LM analysis and excellent behaviour of GLM analysis.

O'Hara and Kotze (2010) compared different approaches for estimating the mean of a distribution, on the log scale, from count data. Their principal comparison was between (i) the location parameter in a negative binomial GLM (which is the log of the mean of the counts), and (ii) the mean of a logged distribution (to which a constant has been added to avoid the log(0) problem). O'Hara and Kotze compare these two coefficients directly; however, we feel such a comparison is problematic for two reasons.

First, the analysis of the $log(y + 1)$ data is compared to the log of the mean of count data, y, without the added 1 (or any other constant). It seems unlikely that a thoughtful researcher would take an estimate of the mean in such an analysis as representative of the (log) mean. One would not expect, in general, the mean of a random variable y (transformed or otherwise), and the mean of a random variable $y + a$ (similarly transformed), to be equivalent.

Second, putting the "+1" issue aside, the mean of a transformation of a random variable is not generally equal to (i.e., cannot be compared in a simulation study) the transformation applied to the mean. Consider the log transformation applied to variable x that follows a log-normal distribution. Such a variable, once log transformed, will have a mean of μ and a standard deviation of σ. These coefficients, μ and σ, are traditionally used as the parameters of a log-normal distribution. However, the mean of the original distribution is not e^{μ}. Rather, $E[y] = e^{\mu + \frac{\sigma^2}{2}}$. Thus, $log(E[y]) \neq \mu$. The general statement of this inequality is that for an arbitrary non-linear transformation $f()$ of a random variable x, $E[f(x)] \neq f(E[x])$. Particularly when applied to convex functions (in which case $f(E[x]) < E[f(x)]$), this principle is known as Jensen's inequality (Jensen 1906). In this article, we will be primarily concerned with the bahaviour of random variables under logarithmic transformation; since this is concave function, $log(E[x]) > E[log(x)]$.

The coefficient estimated by O'Hara and Kotze (2010) in the negative binomial log-link GLM analysis is the logarithm of the mean of the response, $log(E[y])$, and their calculations of bias and accuracy (RMSE) relate negative binomial GLM-based estimates of $log(E[y])$ to the true values of $log(E[y])$; this is a logical comparison. However, the analysis in which they fitted an identity-link linear model to log transformed data did not estimate $log(E[y])$. Rather, it estimated $E[log(y)]$; note that we are setting aside the +1 issue, where in fact, the LM analysis estimated $E[log(y + 1)]$. However, this estimator was nonetheless compared to $log(E[y])$ in calculations of bias and accuracy, and this is clearly not a similarly logical comparison.

We believe that these issues are avoidable, and that re-evaluating the evidence presented by O'Hara and Kotze (2010) in the light of such logical corrections should be illuminating. Accordingly, we performed similar analyses to those presented by O'Hara and Kotze (2010), but we transformed outputs of both the negative binomial GLM analysis and the linear model applied to logged data such that they are comparable. We considered both the log scale and the original data scale. We considered performance through different approaches (bias, and overall accuracy or RMSE, as considered by O'Hara and Kotze 2010), of all models applied in the original paper, and also of a linear model applied to the untransformed data.

Simulations

Our simulation scheme followed O'Hara and Kotze's (2010) simulations directly in almost all respects. For each simulation we generated a random sample y from a negative binomial distribution with a mean we shall denote $E[y]$ and with an overdispersion parameter θ. This parameterisation of the negative binomial distribution, common in ecology but not necessarily elsewhere, is explained in somewhat more detail in the appendix. Briefly, the negative binomial distribution converges on the Poisson distribution with $VAR[y] = E[y]$ for large values of θ, with $VAR[y] = E[y] + \frac{E[y]^2}{\theta}$. The properties of the negative binomial distribution with this parameterisation are elaborated in the supplemental materials. Each sample had $n = 100$. We investigated values of $E[y]$ in $[1, 2, 3, ..., 20]$, and values of the dispersion parameter θ in $[0.5, 1, 2, 5, 10]$. Each simulation scenario was replicated 10^4 times. Our first set of simulations exactly followed O'Hara and Kotze's (2010) procedure and simulated datasets that contained $n = 100$ values for each of the twenty values of $E[y]$, for a total of $n_{total} = 2000$ samples in each replicate analysis. Each of the 10^4 replicate simulations for each combination of parameters ($E[y]$ and θ) thus generated and estimated an intercept for each of the twenty groups with different means, and a common overdispersion parameter or residual variance. We also condicted analyses where each of the 20 groups with different means was analysed individually, generating separate estimates of the mean and disperters for each group. Finally, we also conducted all simulations with a smaller sample size of $n_{=}20$ for each factor level (i.e., each group with a true mean between 1 and 20 counts) within each replicate analysis.

Models

We employed three different models that estimate $E[y]$, $log(E[y])$, or the mean of the transformation $E[log(y + 1)]$. First, we applied a negative binomial GLM with a log link function to estimate $log(E[y])$,

$$y_i \sim NB\left(e^{\alpha_{NB}}, \theta\right), \tag{1}$$

where i indexes observations of the count variable, $NB\left(\right)$ denotes a negative binomial distribution parameterised via its expectation and a dispersion parameter θ; we note however, that the GLM doesn not assume that the data follow a negative binomial distribution (although our simulated data do), but rather that the variance of residuals is related to the mean in the same way as it is in the negative binomial distribution (McCullagh and Nelder 1989; see the supplementary materials for more on this relationship). We denote the key parameter directly estimated by each model as α with a distinguishing subscript. In the negative binomial model, α_{NB} directly estimates $log(E[y])$.

We fitted the negative binomial GLM (equation 1) using a modification of the glm.nb() function from the package MASS (Venables and Ripley 2002). We modified the function to default to fitting a Poisson GLM with a log link when the algorithm to determine the value of the θ reached very large values but did not converge (such that the negative binomial distribution converges on a Poisson distribution; see further explanation in the appendix). Otherwise, the algorithm behaves well, but generates warning messages that must be suppressed. The modified algorithm may not necessarily be suitable for analyses beyond the simulations conducted here; the modified source is available with all other code used in the present study.

Next we fitted an (identity link) linear model with $log(y + 1)$ as a response variable,

$$log(y_i + 1) = \alpha_{logLM} + e_i, \tag{2}$$

where α_{logLM} is a direct estimator of $E[log(y + 1)]$, and e_i are residuals, with estimated variance σ^2_{logLM}. This model assumes that residuals, e_i, of the $log(y + 1)$ transformed data, are independent and have constant variance.

We fitted the linear model of the transformed data (equation 2, and of untransformed data, equation 3, see below) using the lm() function in the base R package version 3.4.1 (R Core Team 2017).

Finally, we fitted an (identity link) linear model to the untransformed data,

$$y_i = \alpha_{LM} + e_i, \tag{3}$$

where e_i are residuals on the untransformed scale (and as such are distinct from those in the second model), and α_{LM} is an estimator of $E[y]$. We denote the estimated variance of residual in this model by σ^2_{LM}. This model assumes that residuals, e_i, of the untransformed cound data, y, are independent and have constant variance.

Obtaining parameters of interest

There are two principal quantities that could potentially be of interest for a count variable: its mean ($E[y]$), and the log of its mean ($log(E[y])$); the mean of the transformation (i.e., $E[log(y)]$ or $E[log(y + 1)]$) is potentially also of interest, but as $log(E[y])$ was the focal estimand in O'Hara and Kotze (2010), we focus on it. We devised estimators of each of $E[y]$ and $log(E[y])$, and associated standard errors, from each of the three analytical models (described in equations 1, 2, and 3) that we fitted to the simulated datasets. Expressions for these estimators are given in table 1. Explanations of how these estimators are derived are given in the supplemental materials, as are expressions that may be useful if standard errors of derived quantities given in table 1 are used in practice.

Evaluation of model performance

First, we evaluated the performance of the model at estimating the mean of the negative binomial variables on the $log(y + 1)$ scale. For this, we calculated the mean of α_{logLM} across simulations, for each combination of μ and θ. We compared this to the true mean of each transformed negative binomial distribution, which we calculated according to

$$E[log(y+1)] = \Sigma_{y=0}^{\infty} log(y+1) p_{negbin}(y, E[y], \theta),$$

where $p_{negbin}(y, \mu, \theta)$ is the density of a negative binomial distribution with mean μ and dispersion parameter θ, evaluated at y. In practice we did the summation over y up to $y = 1000$. We summed the estimate of the mean of the $log(y+1)$ transformed data across all 1000 replicate simulations, and plotted these against the expected value, for all values of $E[y]$ and all values of θ. Deviation from the 1:1 line would indicate that there is some inherent bias in linear models as a mechanism for estimating location parameters for this type of data.

Next, we evaluated the performance of each estimator of the log of the mean of the count variable, and of the mean of the count variable, according to the two criteria used by O'Hara and Kotze (2010): bias and overall accuracy. We also evaluated the performance of the standard errors of each estimator (i.e., square roots of estimation variances).

We estimated the bias of each estimator using the standard formula

$$bias = E[\hat{\phi}] - \phi \ ,$$

where ϕ is the true value of some quantity, i.e., ϕ is the estimand (in our case, the true simulated values of either $log(E[y])$ of $E[y]$), and $\hat{\phi}$ is an estimator of ϕ (i.e., quantities directly estimated from the models described in section *Models*, or derived in section *Transformations*). We estimate $E[\hat{\phi}]$ for each estimate of the mean (or logarithm of the mean) of our simulated count variables as the mean of the estimate across the 10^4 replicate simulations for each combination of parameters.

We estimated the overall accuracy of each analysis using the standard metric root mean squared error (RMSE). This is defined as

$$RMSE = \sqrt{E[(\hat{\phi} - \phi)^2]} \ .$$

Similarly to our calculations of bias, we estimate $E[(\hat{\phi} - \phi)^2]$ as the average taken over all replicate simulations for any given combination of parameters. Our main results consider bias and RMSE, since these are the aspects of model performance considered by O'Hara and Koze (2010). However, a range of further analyses of these simulation results is clearly of potential interest. In the supplemental materials, we provide results about bias and precision on different scales (Figures S.2 through S.5), and for smaller sample sizes ($n = 20$ per group; figures S.6 and S.7). We provide a brief investigation of the performance of standard errors in the supplemental material (figures S.8 and S.9).

Results

OLS estimates of the mean of the $log(y+1)$ transformed data closely matched the true means of the $log(y+1)$ transformation for all values true of $E[y]$ and θ (figure 1). This indicates that there is no inherent bias in the linear model analysis of the transformed data itself; estimates of the mean of the $log(y+1)$ are unbiased. This follows from least squares theory: regardless of the distribution of the $log(y+1)$ transformed data the OLS estimate of their mean is unbiased (Rao 1973; Judge et al. 1980). Therefore, any problems with estimates of quantities such as $E[y]$ or $log(E[y])$ will reflect deficiencies in the transformations that we apply.

For all parameter values, the estimates of $log(E[y])$ obtained with the negative binomial GLM and the linear model applied to the raw count data are unbiased (figure 2a-e). Both of these analyses yielded essentially identical overall accuracy, as measured by RMSE, which was better than the accuracy of the other approaches that we considered. The GLM analysis, which matches the data-generating model exactly, provided valid standard errors (figures S.8 and S.9) across all parameter values. Standard errors from the linear model were valid when the mean of each group was estimated separately (figures S.8 and S.9, parts f-j), but were generally poor, expecially in relative terms (figures S.9a-e) when a single resiudal variance was estimated for across all groups with true mean counts from 1 to 20, which spanned very large ranges of true residual variation.

Measures of the performance of the mean of the $log(y+1)$ data, treated as an estimator of $log(E[y])$, as investigated by O'Hara and Kotze's (2010), are presented in figure 2. In our results, these behave identically to the results given in O'Hara and Kotze's (2010). This quantity is, on average, larger than $log(E[y])$ for small true values of $E[y]$, and is smaller than $log(E[y])$ for large true mean values of the count variable, particularly when overdispersion is high (figure 2a-e).

When we applied the approximate estimators of $log(E[y])$ from the LM analysis of the $log(y+1)$ data, the performance of these estimators was far better than the impression given if $E[\widehat{log(y+1)}]$ is taken to be an estimator of $log(E[y])$. The approximate estimators provided reasonably unbiased inferences of $log(E[y])$ for most parameter values, certainly far better than if the mean of the $log(y+1)$ data is taken as an estimator of $log(E[y])$, except for the highest levels of overdispersion ($\theta = 0.5$; figure 2a-e). These estimators were far more accurate for estimation of $log(E[y])$ than $E[\widehat{log(y+1)}]$, as judged by RMSE (figure 2f-j). The first-order approximations to their standard errors performed reasonably, except for at very high levels of overdispersion, and for the lowest means (Figures S.8 and S.9). Some modest differences occur between the two approximations of $log(E[y])$, based on the LM analysis of the $E[log(y+1)]$, and the associated approximations of their standard errors. At the highest levels of overdispersion, the approximation based on the 2$^{\text{nd}}$ order Taylor series (eq. 9 in table 1) had better RMSE than the log-normal approximation (eq. 7 in table 1; figure 2f). The log-normal approximation for standard errors performed better for low means of the count variable (figure 2f-j), but the first order approximation for standard errors better reflected the true SD of the estimator for larger means. All the results we have considered so far (figure 2) come from scenarios where a single model is fitted to analyse the means of the twenty groups with different means. These analyses all assume a single residual variance, which is used in the approximations for $log(E[y])$. If each group mean is estimated is a separate model, with a separate residual variance, the performance of the estimators, with respect to both bias and RMSE is even better (figure 3).

For comparability with O'Hara and Kotze's (2010) results, we present our main results for inference for the logarithm of the mean of the count variable y. Equivalent plots to figures 2 and 3 are provided for all results on the scale of the observed count variable, both in absolute terms (i.e., where units are counts; figures S.1 & S.2), and in relative terms (where bias, RMSE, and standard errors are presented in units of the true mean; figures S.3 & S.4). These results agree closely with those for the log scale for all key interpretations given in this section.

Discussion

Figures 2 and 3 of O'Hara and Kotze (2010) present the results of their analyses. Their conclusion was that no matter whether bias or RMSE is considered as a measure of estimation reliability, the GLM method often substantially outperformed the log-transformation method, and there were no circumstances where the reverse was true. Our figures have a very different interpretation. Specifically, whether considering bias or RMSE, (*i*) most of the discrepancy in the original analyses was due to the fact that the LM analysis of transformed data estimates a different quantity than the GLM analysis (figures 2 & 3), (*ii*) once suitably transformed, estimates from the GLM and the linear model applied to transformed data are very similar across most of the range of scenarios examined (figures 2 & 3), and (*iii*) the performance of the GLM and the linear model applied directly to the raw count data scale are practically indistinguishable across the range of scenarios examined. Importantly, the analyses of transformed data are not nearly as severely biased as O'Hara and Kotze's (2010) results indicated; their very negative results are primarily a consequence of comparing two different quantities. The biases in our simulations involving back-transformed parameters should not be seen as arising from errors in the OLS estimation applied to the transformed data; these analyses yield unbiased estimates of the mean of the distribution of the transformed data (figure 1). Rather, the biases that persist after back-transformation (figures 2 and 3) will be a result of the standard types of approximations used in the derivations of the back-transformations (specifically, using the delta method, Dorfman 1938, Ver Hoef 2012, and approximations based on properties of the log-normal distribution, Aitchison & Brown 1957, see the supplemental materials for details). It may be possible to use newer methods to derive even better back-transformations (Khuri et al., 2015)

It is possible to explain why O'Hara and Kotze saw the patterns that they did. When the true mean of the response variable is low then the failure to account for the +1 correction is the main source of bias in their comparison (but this is absent from our comparison). This is the positive bias for the transformation-methods that can be seen in their Figure 2 for low values of the true mean. However for the samples in their (and our) simulation study variance increases with increasing mean value, so for high mean values their comparison (but not ours) predicts a negative bias for the transformation methods because the mean on the log scale is less than the log of the mean on the count data scale. For completeness we note that for both bias and RMSE both the "normal residuals" and "second order" approximations perform relatively well except when the data are strongly overdispersed (in the present context, have error variance greater than that expected for the Poisson distribution). In situations where these two methods perform less well, neither is universally better than the other. We note also that all these general trends related to how effectively the models estimate the mean also extend to the empirical standard deviation and the estimated standard error associated with the estimated mean value.

Our results provide a comparison between what would be recovered by a negative binomial GLM and a linear model using standard ordinary least squares (OLS) formulations. We find that the linear model estimates the mean as well as the negative binomial GLM. We should keep in mind that the negative binomial GLM had an advantage over all the other models considered in our comparison: the negative binomial model that we selected for the GLM was an exact match to the function used to generate the samples. In practice we will rarely, if ever, be in a situation where we know with certainty exactly the data structure to select for our GLM to provide a perfect match to the underlying system than is being sampled. So the fact that this advantage did not lead to substantially better performance than the simple linear model is particularly noteworthy. It will also be surprising to many at first, as it is widely believed that the linear model is based on the assumption that the residuals are normally distributed, and (especially for small θ), the residuals in our simulations will have been far from normal. In fact, OLS mechanics (and thus linear models) do not assume normal residuals (Rao 1973). This assumption only comes into play when generating p values (and then is probably most important at small sample sizes). However, it should be noted that standard mechanics for generating p values in GLMs are asymptotic, and thus approximate for finite sample sizes. Furthermore, GLMs themselves rely on specifying particular ling functions and mean-variance relationships. While the GLMs that we fitted in this study exactly match the link functions and distributional assumptions of the data simulation scheme, in practice, these model features will never perfectly match real biological data. It is thus possible for broken assumptions of normal residuals (insofar as such an assumption is actually made) in LMs to be less consequential than the various problems that can arise in the applications of GLMs, even for generating p values (Ives 2015).

We do not intend to deny that generalised linear model analysis will often provide great benefits for the analysis of biological data, nor that generalised models will often be the most appropriate methods for many types of analysis that arise in ecology and evolution. However, our revisions of O'Hara and Kotze's (2010) findings may nonetheless warrant some general changes to available advice on how LM-based analysis of data from arbitrary distributions should be perceived. Though one may themselves prefer other methods, results by those who opt for simpler methods should not be judged harshly or dismissed, simply because their distributional assumptions are not perfectly met – this alone does not necessarily lead to catastrophic failure of a statistical model. Similarly, results in the literature based on older methods may still in many instances be regarded as reliable. Approximations given here for converting results from linear models of $log(y + 1)$, potentially with standard errors, may facilite the use of such older results in new meta-analyses. Furthermore, when analyses of a single dataset using LMs and GLMs appear to give different answers, it is quite possible that the apparent discrepancy arises from mis-specification or mis-interpretation of the GLM results, as was the case for some key aspects of O'Hara and Kotze's (2010). In our experience, analysts typically attribute such discrepancies to the inadequacy of a LM, often invoking assumptions of OLS analysis that do not exist. In such cases, we have often found that results from LMs and GLMs are highly congruent, once errors in the implementation - or more often interpretation – of GLMs are corrected. The tendency to mis-attribute divergence between LM and GLM results to poor performance of linear models is further evidenced by the >500 citations that have been made to O'Hara and Kotze's (2010) paper, apparently without any close look at the mechanics of its LM and GLM analyses revealing that the key comparisons therein were not based on comparable quantities.

Acknowledgements

MBM is supported by a University Research Fellowship from the Royal Society (London).

Data accessibility

Code to conduct all simulations, genreate all figures, and compile the manuscript is provided in a repository at https://github.com/mbmorrissey/count_data_analysis and DOI: 10.5281/zenodo.3631965

References

J. Aitchison and J.A.C. Brown. 1957. The Lognormal Distribution, with special reference to its uses in economics. Cambridge University Press, Cambridge UK.

R. Dorfman. 1938. A note on the delta-method for finding variance formulae. The Biometric Bulletin 1: 129-137.

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

J.L.W.V. Jensen. 1906. Sur les fonctions convexeses et les inégalietés entre les valeurs moyennes. Acta Mathematica 30: 175-193.

G.G. Judge, W.E. Griffiths, R.C. Hill, H. Lutkepohl, and T.-C. Lee. 1980. The theory and practice of econometrics, 2^{nd} *ed.* Wiley, New York.

P. McCullagh asn J.A. Nelder. 1989. Generalized Linear Models. Chapman and Hall, New York.

A. Khuri, S. Mukhopadhyay, and M. Khuri. 2015. Approximating moments of continuous functions of random variables using Bernstein polynomials. Statistical Methodology 24: 37-51.

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

C.R. Rao. 1973. Linear statistical inference and its applications, 2^{nd} *ed.* Wiley, New York.

R Core Team. 2017. R: A language and environment for statistical computing. Vienna `https://www.R-project.org/`

R.R. Sokal and F.J. Rohlf. 1995. Biometry, 3rd ed. W. H. Freeman and Company, New York.

Venables, W.N., and B.D. Ripley. 2002. Modern applied statistics with S, 4^{th} ed. Springer, New York.

J.M. Ver Hoef. 2012. Who invented the delta method? The American Statistician 66: 124-127.

Table 1: Estimators of the mean of a count variable, $\widehat{E[y]}$, and the log of the mean of a count variable, $\widehat{log(E[y])}$, obtained from the parameters of three different statistical models.

model	equation with relevant terms	$\widehat{E[y]}$	$\widehat{log(E[y])}$	supplementary equation for estimation variance
glm analysis of y	eq. 1	(4) $\widehat{E[y]} = e^{\alpha_{NB}}$	(5) $\widehat{log(E[y])} = \alpha_{NB}$	eq. S.1
lm analysis of $log(y + 1)$, log-normal transformation	eq. 2	(6) $\widehat{E[y]} = e^{\alpha_{logLM}+\frac{\sigma^2_{logLM}}{2}} - 1$	(7) $\widehat{log(E[y])} = log(e^{\alpha_{logLM}+\frac{\sigma^2_{logLM}}{2}} - 1)$	eqs. S.3 & S.5
lm analysis of $log(y + 1)$, 2$^{\text{nd}}$-order approximation	eq. 2	(8) $\widehat{E[y]} = e^{\alpha_{logLM}}(1 + \frac{\sigma^2_{logLM}}{2}) - 1$	(9) $\widehat{log(E[y])} = log(e^{\alpha_{logLM}}(1 + \frac{\sigma^2_{logLM}}{2}) - 1)$	eqs. S.4 & S.6
lm analysis of y	eq. 3	(10) $\widehat{E[y]} = \alpha_{LM}$	(11) $\widehat{log(E[y])} = log(\alpha_{LM})$	eq. S.7

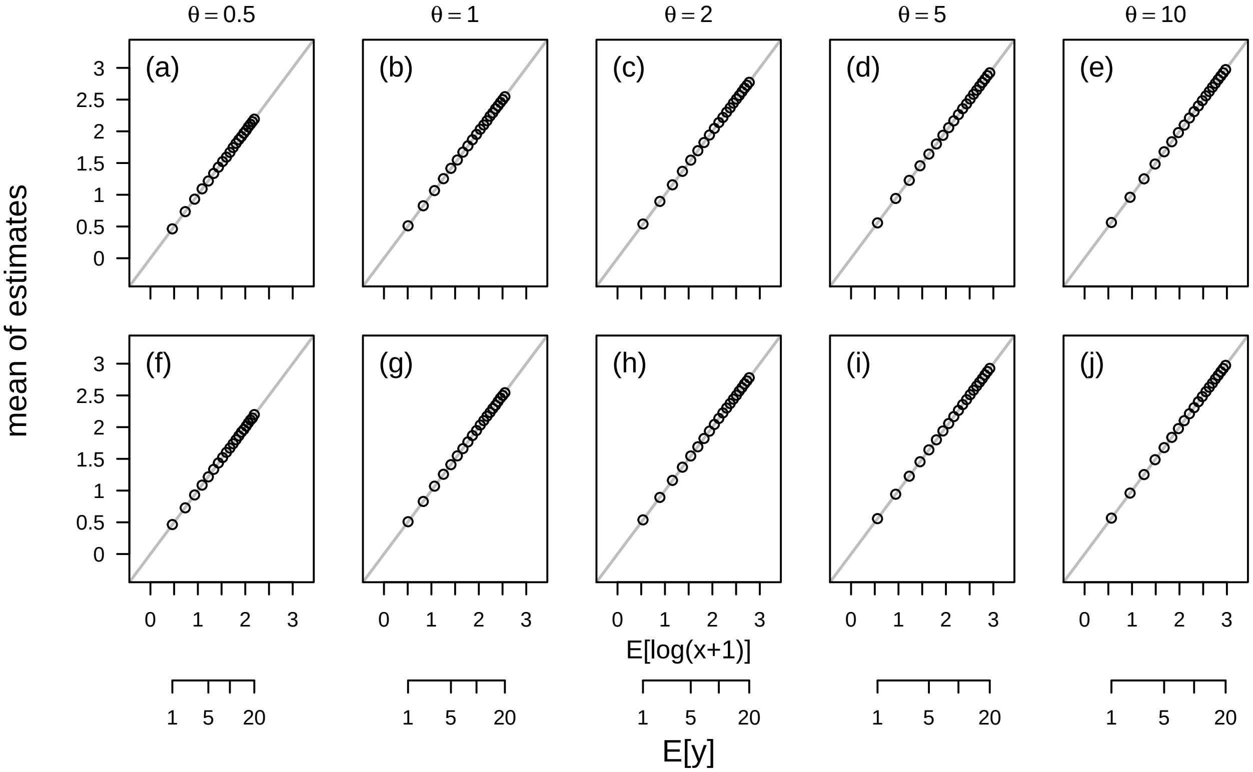

Figure 1: Bias in estimation of the mean of negative binomial variables, transformed according to $log(y + 1)$. True simulated mean values are plotted on the x-axis, and the means of simulation results are plotted on the y-axis. As such, points falling on the one-to-one line (grey) indicate simulation scenarios in which the analysis of $log(y + 1)$ transformed data is unbiased at recovering the mean on the $log(y + 1)$ scale. Plots a-e (top row) are generated from simulations where a single model estimates means of groups with true values from 1 to 20, with a common dispersion parameter or residual variance. Plots f-j (bottom row) are generated from simulations where a separate model estimates the mean and disperson parameter or residual variance for each group with a different true (simulated) mean value.

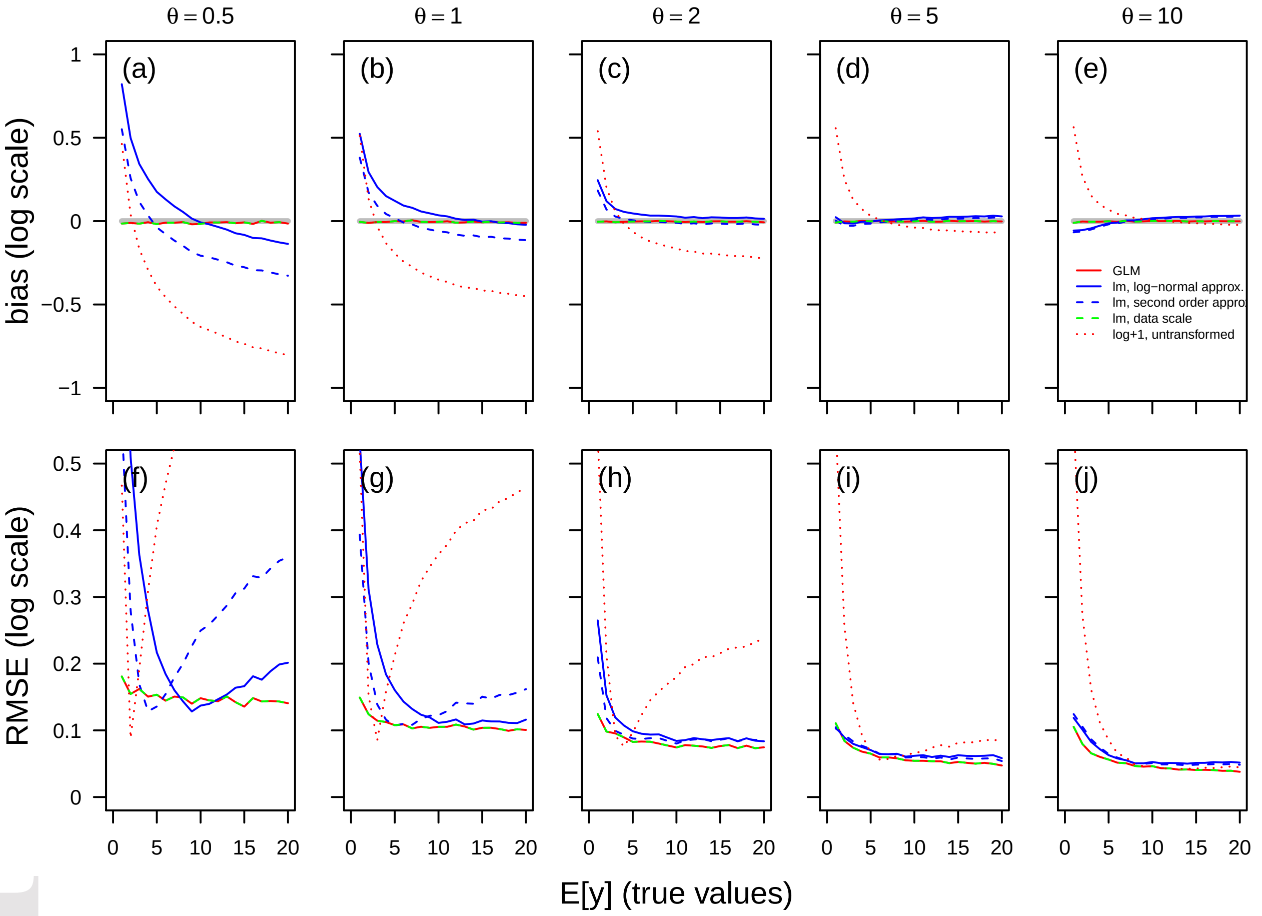

Figure 2: Bias (a-e) and overall accuracy (f-j) of inferences of the logarithm of the mean of a count variable. Data ($n = 100$) for a count variable x were simulated from a negative binomial distribution with mean $E[y]$ and size parameter θ. Expressions for the two transformations of the analysis of $log(y + 1)$ data are given in equations 7 and 9 of table 1. 10000 replicate simulations of each simulation were conducted and estimators of $log(E[y])$ were constructed from a suite of GLM and LM analyses.

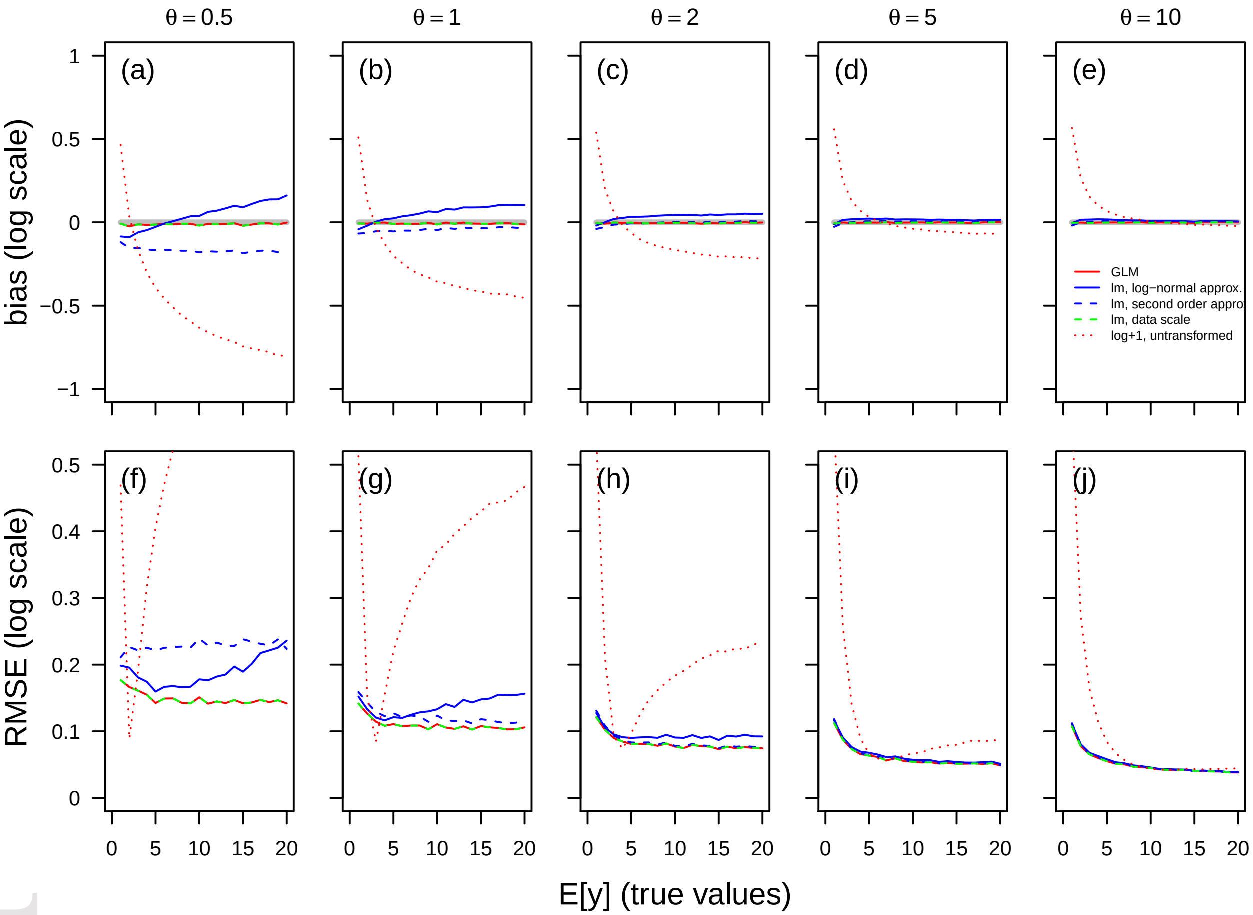

Figure 3: Bias (a-e) and overall accuracy (f-j) of inferences of the logarithm of the mean of a count variable. Simulations are as for figure 2, except that each simulation involves fitting separate models for each level of the predictor variable. Data ($n = 100$) for a count variable x were simulated from a negative binomial distribution with mean $E[y]$ and size parameter θ. Expressions for the two transformations of the analysis of $log(y + 1)$ data are given in equations 7 and 9 of table 1. 10000 replicate simulations of each simulation were conducted and estimators of $log(E[y])$ were constructed from a suite of GLM and LM analyses.