Do not log-transform count data

O'Hara R.B.[1,3] & Kotze, D.J.[2]

[1] *Biodiversity and Climate Research Centre, Senckenberganlage 25, D-60325 Frankfurt am Main, Germany. Email: bohara@senckenberg.de*

[2] *Department of Biological and Environmental Sciences, PO Box 65, FI-00014, University of Helsinki, Finland. Email: johan.kotze@helsinki.fi*

[3] Corresponding auther: *Biodiversity and Climate Research Centre, Senckenberganlage 25, D-60325 Frankfurt am Main, Germany.*

Email: bohara@senckenberg.de,

tel: +49 69 798 40216,

fax: +49 69 798 40169.

Word Count: 2672

Abstract

1. Ecological count data (e.g., number of individuals or species) are often log-transformed to satisfy parametric test assumptions.
2. Apart from the fact that generalized linear models are better suited in dealing with count data, a log-transformation of counts has the additional quandary in how to deal with zero observations. With just one zero observation (if this observation represents a sampling unit), the whole dataset needs to be fudged by adding a value (usually 1) before transformation.
3. Simulating data from a negative binomial distribution, we compared the outcome of fitting models that were transformed in various ways (log, square-root) with results from fitting models using Poisson and negative binomial models to untransformed count data.
4. We found that the transformations performed poorly, except when the dispersion was small and the mean counts were large. The Poisson and negative binomial models consistently performed well, with little bias.

Keywords: transformation, Poisson, overdispersion, linear models, generalized linear models,

Introduction

Ecological data are often discrete counts - the number of individuals or species in a trap, quadrat, habitat patch, on an island, in a nature reserve, on a host plant or animal, the number of offspring, the number of colonies, or the number of segments on an insect antenna. Even though textbooks on statistical methods in ecology (e.g., Sokal & Rohlf 1995; Zar 1999; Crawley 2003; Maindonald & Braun 2007) recommend the use of the square root transformation to normalise count data, such data are often log-transformed for subsequent analysis with parametric test procedures (e.g., Gebeyehu & Samways 2002; Magura, Tóthmérész & Elek 2005; Cuesta *et al*. 2008). The reasons for this (log-transforming count data) are not clear but perhaps has to do with the common use of log transformations on all kinds of data, and the fact that textbooks usually deal with the log-transformation first, before evaluating other transformation techniques.

The main purpose of a transformation is to get the sampled data in line with the assumptions of parametric statistics (such as ANOVA, t-test, linear regression) or to deal with outliers (see Zuur, Ieno & Smith 2007; Zuur, Ieno & Elphick 2009). These assumptions include that the residuals from a model fit are normally distributed with a homogeneous variance. In addition, regression assumes that the relationship between the covariate and the expected value of the observation is linear. Classical parametric methods deal with continuous response variables (weights, lengths, concentrations, volumes, rates) with few "zero" observations. As such, a log-transformation may successfully 'normalise' such continuous data for use in parametric statistics.

Discrete response variables, such as counts data, on the other hand, often contain many "zero" observations (see Sileshi, Hailu & Nyadzi 2009) and are unlikely to have a normally distributed error structure. The question arises; can, or should, count data that include zeroes be transformed to approximate normality to be subject to parametric statistics? Maindonald & Braun (2007) argued that generalized linear models have largely removed the need for transforming count data, yet the practice is still widespread in the ecological literature (see above).

Classically, response variables are transformed to improve two aspects of the fit: linearity of the response and homogeneity of the variance ("homoscedasticity"). This can be done in an exploratory manner (e.g., Box & Cox 1964) but transformations often have sensible interpretations, e.g. the log transformation implies that the mechanisms are multiplicative on the scale of the raw data. Clearly, there is no reason to expect that a single transformation will behave optimally for both linearity and homoscedasticity, so some compromise is often needed.

More recently, generalized linear models have been developed (McCullagh

& Nelder 1989). These allow the analyst to specify the distribution that the data are assumed to have come from, which implicitly defines the relationship between the mean and variance. They can be chosen based on an understanding of the underlying process that is assumed to have generated the data, e.g. a constant rate of capture of individual members of a large population implies a Poisson distribution. If the capture rate varies randomly the data look clumped, with more zeroes but also more sites with large counts. In generalized linear modelling terminlogy this is "overdispersion", which can be handled in several ways, the most popular of which are by specifying the response as coming from a quasi-Poisson or negative binomial distribution.

Here we are interested in comparing how well the two approaches work when analysing count data. An additional wrinkle with the traditional approach of log transforming is that $\log(0) = -\infty$, so a value (usually 1) is added to the count before transformation. We are not aware of any justification for adding 1, rather than any other value, and this may bias the fit of the model. Zeroes do not present any problems in generalized linear models, as there it is the expected value that is log-transformed.

Zeroes can also be handled by using zero inflated models (e.g. Sileshi, Hailu & Nyadzi 2009; Zuur, Ieno & Elphick 2009). When modeling small counts, both zero inflated models and over-dispersed models can account for a large number of zero counts, and there may be little advantage in fitting the zero inflated model. The choice of whether to use these models will thus often depend on an understanding of the biology of the system - the assumption is that there are two types of site, where the species occurs and where it does not. The species may not be caught where it occurs, hence the zero counts can be of two classes (i.e. true absence and present but not sampled). This sort of extension of a model can be an important consideration when modelling count data (for an extreme example, the zero-truncated one-inflated negative binomial, see Kotze *et al*. 2003), but is beyond the scope of this paper.

To address this problem of data transformation we simulated data from a negative binomial distribution (since count data in ecology are often clumped, producing an expected variance that is greater than the mean (see McCullagh & Nelder 1989; White & Bennetts 1996; Dalthorp 2004)), which we then subjected to various transformations (square root, log ($y+n$)). The transformed data were analysed using parametric statistics and compared to an analysis of untransformed data in which the response variable was defined as following either a Poisson distribution with overdispersion or a negative binomial error distribution.

Methods

Data sets were simulated from a negative binomial distribution, with different values of θ (θ = 0.5, 1, 2, 5, 10, 100). Low θ (also termed k, see fig. 2 in Wright (1991)) indicates greater variance in the data, i.e. stronger clumping. For each simulation, 100 data points were simulated at each of 20 means, μ (μ=1,...,20). 500 replicate simulations were carried out for each value of θ.

The data were analysed assuming that the mean was a factor, with each mean being a different level. Models were fitted making the following assumptions about the response, y:

1. y follows a negative binomial distribution
2. y follows a Poisson distribution with overdispersion
3. sqrt(y) transformation follows a normal distribution
4. $\log_{10}(y+0.001)$ transformation follows a normal distribution
5. $\log_{10}(y+0.1)$ transformation follows a normal distribution
6. $\log_{10}(y+0.5)$ transformation follows a normal distribution
7. $\log_{10}(y+1)$ transformation follows a normal distribution

The simulations were compared by calculating the mean bias, B:

$$B=\frac{1}{S}\sum_{i=1}^{S}\hat{\mu}-\mu ,$$

and root mean squared error (RMSE):

$$RMSE=\frac{1}{S}\sum_{i=1}^{S}(\hat{\mu}-\mu)^2$$

for the simulations, where $\hat{\mu}$ is the estimated parameter, μ is the true value (known from the simulations), and S is the number of simulations.

Simulations and analyses were carried out in the R statistical programme (R Development Core Team 2009), using the MASS (Vernables & Ripley 2002) package. The code that was used is available as an online supplement.

Results

The proportion of counts that were zero are shown in Fig. 1. Naturally, the proportion decreases as the mean increases, and it also decreases as the variance (controlled by θ) decreases.

The biases for the different estimation methods are plotted in Fig. 2. The negative binomial model has negligible bias, whereas the models based on a normal distribution are all biased, particularly at low means and high variances.

The amount of bias also depends on the transformation used. With little clumping (i.e. high θ), the square root transformation has little bias, as does the log transformation when the mean is high, i.e. there are few zeroes (compare to Fig. 1).

The root mean square error shows a similar pattern, with the negative binomial distribution consistently having a low RMSE, and a high value added to the log transformation being better (Fig. 3). The behaviour of the log+1 transformation is a result of a change in sign of the bias, with the minimum at the point where the mean bias is zero (compare to Fig. 2).

The difference between the negative binomial and quasi-Poisson distribution models is insignificant. The largest absolute difference in bias was 2.4×10^{-8}, and the largest RMSE was only 1.1×10^{-8}, both of which are much smaller than the scales in Figs 2 & 3.

Discussion

When the error structure of data is simple, a transformation (usually a log or power-transformation) can be quite useful to improve the ability of a model to fit to the data by stabilising variances or by making relationships linear (Miller 1997; Piepho 2009) before applying simple linear regression. But a transformation is not guaranteed to solve these problems: there may be a trade-off between homoscedasticity and linearity, or the family of transformations used may not be able to correct one or both of these problems. Different models may therefore need to be applied, and there is now a wide variety of possibilities, of which generalized linear models and their derivatives (McCullagh & Nelder 1989) are the most popular.

For count data, our results suggest that transformations perform poorly and instead statistical procedures designed to deal with counts should be used, i.e. methods for fitting Poisson or negative binomial models to data.

The development of statistical and computational methods over the last 40 years has made it easier to fit these sorts of models, and the procedures for doing this are available in any serious statistics package.

It is perhaps not surprising that fitting the correct model to the data (i.e. the same model that was used to simulate the data) gives the best result; what is more interesting is that there is a difference in performance of the models (see also Jiao *et al*. 2004). This suggests that the choice of model does make a difference, and we would suggest that a model based on counts is more sensible, as it is easier to interpret and avoids the problems of deciding which transformation to use. The model is also more explicit, in the sense that the process that leads to a Poisson distribution of counts is clear (i.e. sampling with a uniform rate of capture), and is likely to provide a more accurate foundation for the model. The extra variability that can be added can be chosen according to the the way it affects the relationship between the mean and variance (Ver Hoef & Boveng 2007).

In our simulations, the Poisson and negative binomial models gave almost identical estimates. This suggests that the models are robust to a mis-specification of the relationship between the mean and variance. In contrast, Ver Hoef & Boveng (2007) gave an example from a real dataset where they differed in their predictions. Whilst their data set is unusual (as they acknowledge), it does serve as a warning that our result may not generalize to real data, which rarely has as balanced a design as our simulations. However, even though the choice of which type of generalized linear model to use depends on many things (O'Hara 2009; Zuur, Ieno & Elphick 2009), we do recommend that count data not be transformed to be used in parametric tests. For such data, GLMs and their derivatives are more appropriate.

Acknowledgments

The order of the authors was determined by the result of the South Africa - England cricket ODI on the 27th September 2009, which England won by 22 runs. The study was financially supported by the research funding programme “LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" of Hesse's Ministry of Higher Education, Research, and the Arts, and the Academy of Finland.

References

Box, G.E.P. & Cox D.R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society B*, **26**, 211-252.

Cuesta, D., Taboada, A., Calvo, L. & Salgado, J.M. (2008) Short- and medium-term effects of experimental nitrogen fertilization on arthropods associated with *Calluna vulgaris* heathlands in north-west Spain. *Environmental Pollution*, **152**, 394-402.

Crawley, M.J. (2003) *Statistical Computing. An Introduction to Data Analysis using S-Plus*. John Wiley & Sons Ltd., England.

Dalthorp, D. (2004) The generalized linear model for spatial data: assessing the effects of environmental covariates on population density in the field. *Entomologia Experimentalis et Applicata*, **111**, 117-131.

Gebeyehu, S. & Samways, M.J. (2002) Grasshopper assemblage response to a restored national park (Mountain Zebra National Park, South Africa). *Biodiversity and Conservation*, **11**, 283-304.

Jiao, Y., Chen, Y., Schneider, D., & Wroblewski, J. (2004) A simulation study of impacts of error structure on modeling stock-recruitment data using generalized linear models. *Canadian Journal of Fisheries and Aquatic Sciences*, **61**, 122-133.

Kotze, D.J., Niemelä, J., O'Hara R.B., Turin, H. (2003) Testing abundance-range size relationships in European carabid beetles (Coleoptera, Carabidae). *Ecography*, **26**, 553-566.

Magura, T., Tóthmérész, B. & Elek, Z. (2005) Impacts of leaf-litter addition on carabids in a conifer plantation. *Biodiversity and Conservation*, **14**, 475-491.

Maindonald, J. & Braun, J. (2007) *Data Analysis and Graphics Using R - An Example-Based Approach*, 2nd Edition. Cambridge University Press, UK.

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

Miller, R.G. Jr. (1997) *Beyond ANOVA*. Chapman & Hall/CRC Press, London.

O'Hara, R.B. (2009) How to make models add up - a primer on GLMMs. *Annales Zoologici Fennici*, **46**, 124-137.

Piepho, H-P. (2009) Data transformation in statistical analysis of field trials with changing treatment variance. *Agronomy Journal*, **101**, 865-869.

R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sileshi, G., Hailu, G. & Nyadzi, G.I. (2009) Traditional occupancy-abundance models are inadequate for zero-inflated ecological count data. *Ecological Modelling*, **220**, 1764-1775.

Sokal, R.R. & Rohlf, F.J. (1995) *Biometry*. 3rd Edition. Freeman and Company, 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.

Vernables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics with S*. 4th Edition. Springer, New York.

White, G.C. & Bennetts, R.E. (1996) Analysis of frequency count data using the negative binomial distribution. *Ecology*, **77**, 2549-2557.

Wright, D.H. (1991) Correlations between incidence and abundance are expected by chance. *Journal of Biogeography*, **18**, 463-466.

Zar, J.H. (1999) *Biostatistical Analysis*. 4th Edition. Prentice Hall, New Jersey.

Zuur, A.F., Ieno, E.N. & Elphick, C.S. (2009) A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*. DOI: 10.1111/j.2041-210X.2009.00001.x

Zuur, A.F., Ieno, E.N. & Smith, G.M. (2007) *Analysing Ecological Data*. Springer, USA.

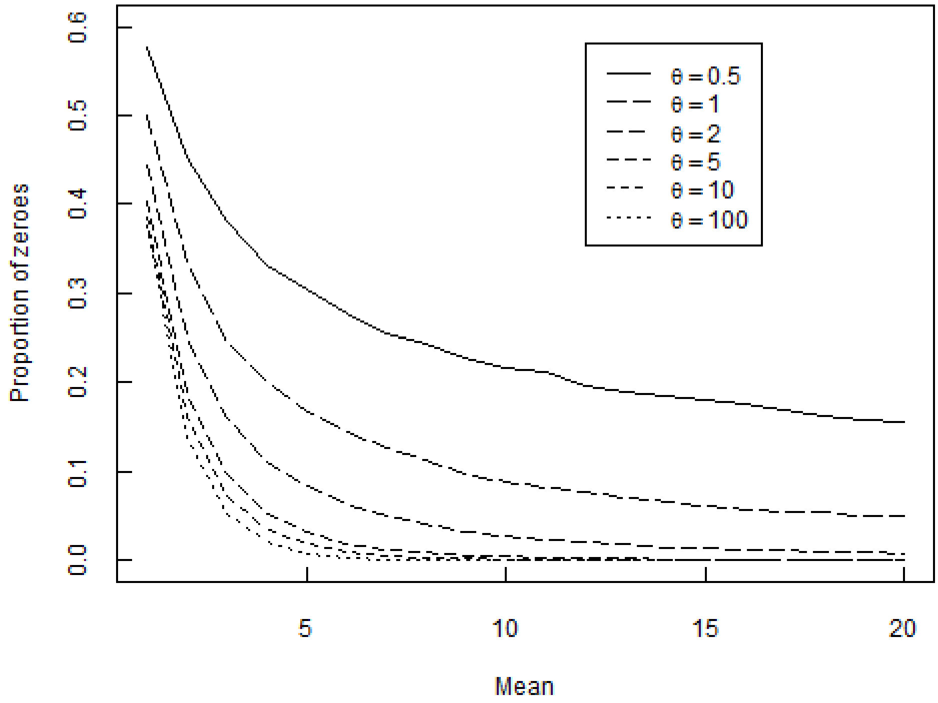

Figure 1. Proportion of values equal to zero in simulations from a negative binomial distribution. θ controls the dispersion ("clumping") in the data: a larger value of θ means lower dispersion.

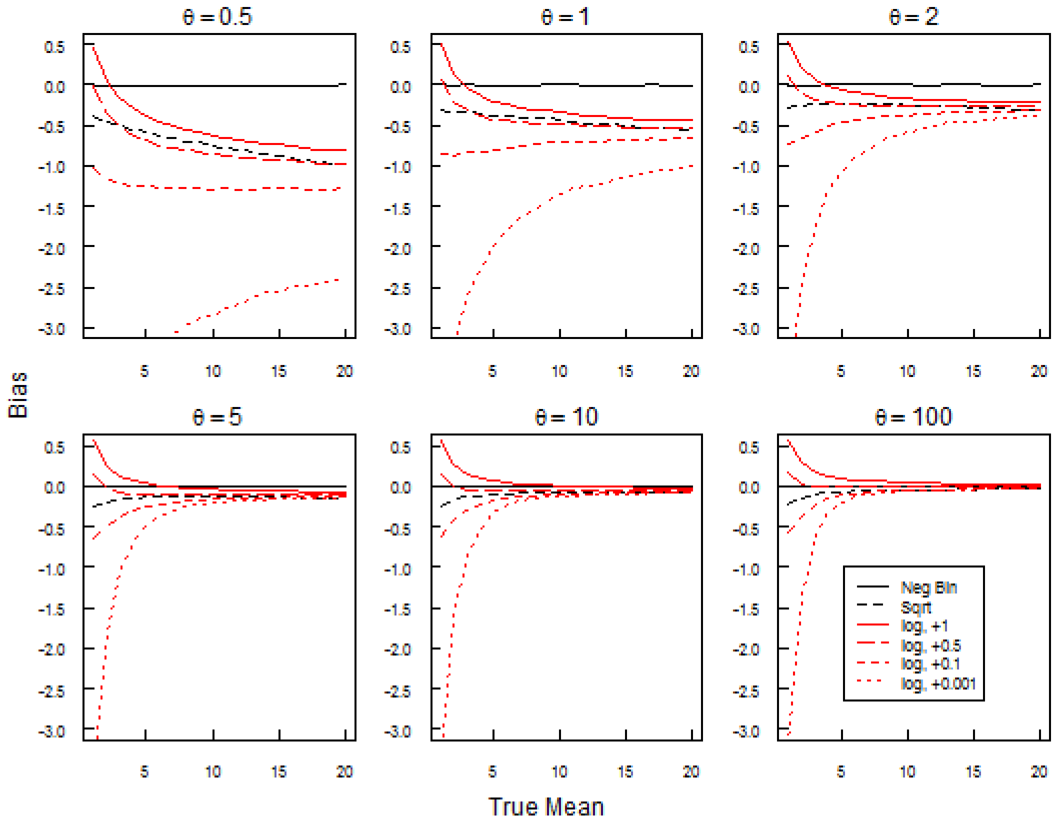

Figure 2. Estimated mean biases from 6 different models, applied to data simulated from a negative binomial distrbution. A low bias means that the method will, on average, return the "true" value.

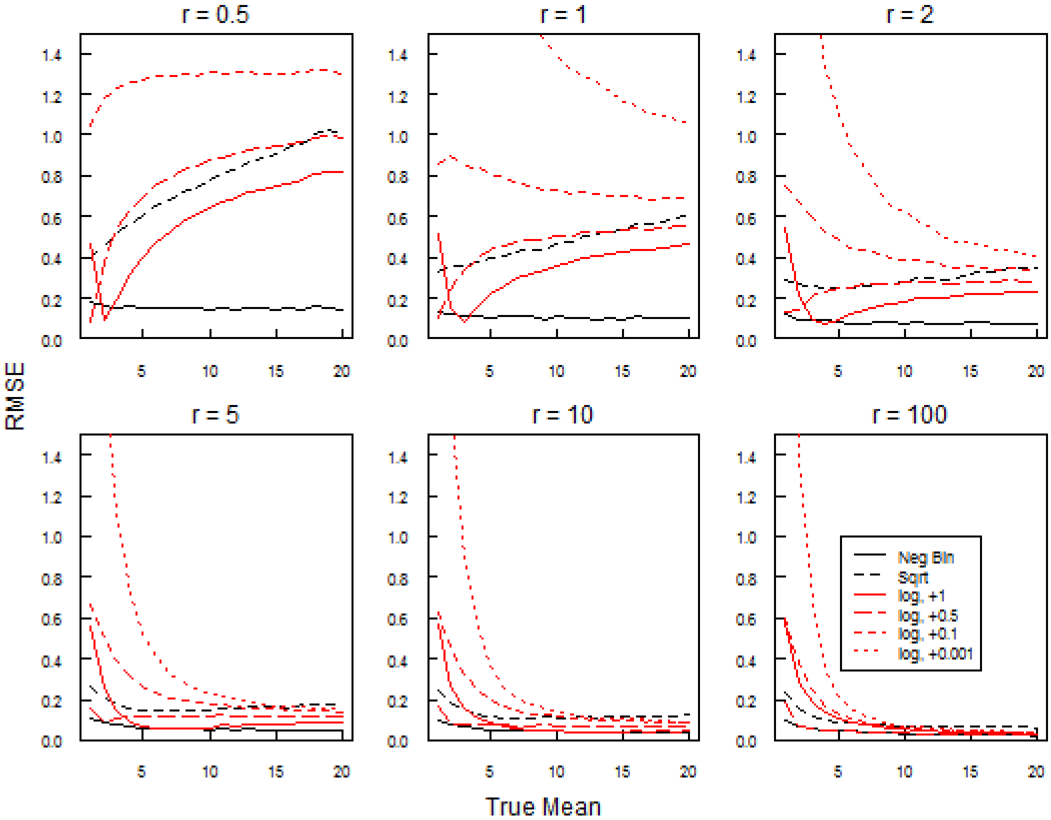

Figure 3. Estimated root mean square error from 6 different models, applied to data simulated from a negative binomial distrbution.