Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2010, **1**, 118–122 doi: 10.1111/j.2041-210X.2010.00021.x

Do not log-transform count data

Robert B. O'Hara[1]* and D. Johan Kotze[2]

[1]*Biodiversity and Climate Research Centre, Senckenberganlage 25, D-60325 Frankfurt am Main, Germany and*
[2]*Department of Environmental Sciences, PO Box 65, University of Helsinki, Helsinki FI-00014, Finland*

Summary

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 data set 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 quasi-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 quasi-Poisson and negative binomial models consistently performed well, with little bias.
5. We recommend that count data should not be analysed by log-transforming it, but instead models based on Poisson and negative binomial distributions should be used.

Key-words: generalized linear models, linear models, overdispersion, Poisson, transformation

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. Densities of individuals are often counts too: a count in an area of unit size (in the analysis of these data, the actual area of a count can be included as an offset; see below). 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 normalize 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 and linear regression) or to deal with outliers (see Zuur, Ieno, & Smith 2010; Zuur, Ieno, & Elphick 2009a). 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 and rates) with few 'zero' observations. As such, a log-transformation may successfully 'normalize' such continuous data for use in parametric statistics.

Discrete response variables, such as count 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 (GLMs) 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

*Correspondence author. E-mail: bohara@senckenberg.de
Correspondence site: http://www.respond2articles.com/MEE/

© 2010 The Authors. Journal compilation © 2010 British Ecological Society

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, GLMs 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 terminology this is 'overdispersion', which can be handled in several ways, the most popular of which is 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 GLMs, as there it is the expected value that is log-transformed.

Zeroes can also be handled by using zero inflated models (e.g. Sileshi *et al.* 2009; Zuur *et al.* 2009, chapter 11, Zuur, Ieno, & Elphick 2010). When modelling counts, both zero inflated models and overdispersed models can account for a large number of zero counts, and there may be little advantage in fitting the zero inflated model.

To address this problem of data transformation, we simulated data from a negative binomial distribution (as 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 and $\log(y + n)$]. The transformed data were analysed using parametric methods and compared with an analysis of untransformed data in which the response variable was defined as following either a Poisson distribution with overdispersion (i.e. a quasi-Poisson distribution) or a negative binomial error distribution.

Generalized linear models

A GLM is an extension of the well-known linear models, like regression and ANOVA (O'Hara 2009). The key idea is that, like linear models, the expected value of a data point (i.e. its mean, which we can call μ) is modelled as the sum – called a linear predictor – of different terms. A linear model assumes that the data point comes from a normal distribution, with this sum as the mean. A GLM extends this by, firstly, allowing more distributions than a normal to be used. For count data, the Poisson distribution is used as a good model of the data. Then, a function of the linear predictor is used as the mean of the distribution. So, for count data, y_i, for the ith observation we have

$$y_i \sim \text{Poisson}(\lambda_i) \qquad \text{eqn 1}$$

$$\log \lambda_i = \mu_i. \qquad \text{eqn 2}$$

Here, log( ) is the function that links the linear predictor to the expected value of the data point: it is called a *link function*. If we had a single continuous covariate x (for example), μ_i might be

$$\mu_i = \alpha + \beta x_i, \qquad \text{eqn 3}$$

exactly as in a simple regression. This is equivalent to a multiplicative model for λ_i, i.e.

$$(\lambda_i) = e^{\alpha+\beta x_i} = e^{\alpha}(e^{x_i})^{\beta}. \qquad \text{eqn 4}$$

If we were interested in estimating the density (δ) of individuals in a plot of area a, the expected (mean) number in the plot would be $a\delta$. Then, comparing with Eqn (4), we see that the density is e^{α}, and the area is $e^{x\beta}$. Hence, we can estimate the density by 'regressing' against $\log(a)$ using eqn (3), fixing $\beta = 1$: this is called using $\log(a)$ as an offset.

One further point needs to be clarified. The Poisson distribution assumes that the mean and variance are equal. Real data do not follow this, and the variance (v) is often much larger than the mean (λ). This biological reality – called overdispersion – can be incorporated into a model in several ways. These all estimate the amount of extra variation but make different assumptions about how this extra variation scales with the mean. Here, we use a quasi-Poisson distribution, which assumes $v = \sigma\lambda$, and the negative binomial distribution, which assumes $v = \lambda + \lambda^2/\theta$ (σ and θ are both overdispersion parameters). Ver Hoef & Boveng (2007) provide a more detailed discussion and comparison of these assumptions.

Materials and methods

Data sets were simulated from a negative binomial distribution, with different values of θ ($\theta = 0{\cdot}5, 1, 2, 5, 10, 100$). 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 mean values, λ ($\lambda = 1,\ldots, 20$). Five hundred 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 $\log_{10}(y + 0{\cdot}001)$ transformation follows a normal distribution
5 $\log_{10}(y + 0{\cdot}1)$ transformation follows a normal distribution
6 $\log_{10}(y + 0{\cdot}5)$ transformation follows a normal distribution
7 $\log_{10}(y + 1)$ transformation follows a normal distribution

© 2010 The Authors. Journal compilation © 2010 British Ecological Society,

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):

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

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. We calculated these on the log scale, i.e. $\mu = \log(\lambda)$. This is the scale on which the parameters are estimated in all of the models except the square-root transformation; so, for the latter model we transformed the parameters onto the log scale.

Simulations and analyses were carried out in the R statistical program (R Development Core Team 2009), using the MASS (Vernables & Ripley 2002) package. The code that was used is available as an online supplement (Appendix S1 in Supporting Information).

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 quasi-Poisson and negative binomial models behave similarly; so, only the latter is presented; see below). The negative binomial model has negligible bias, whereas the models based on a normal distribution are all biased, particularly at low mean values and high variances. The square-root transformation has a lower bias than any of the log-transformations, unless the mean is low.

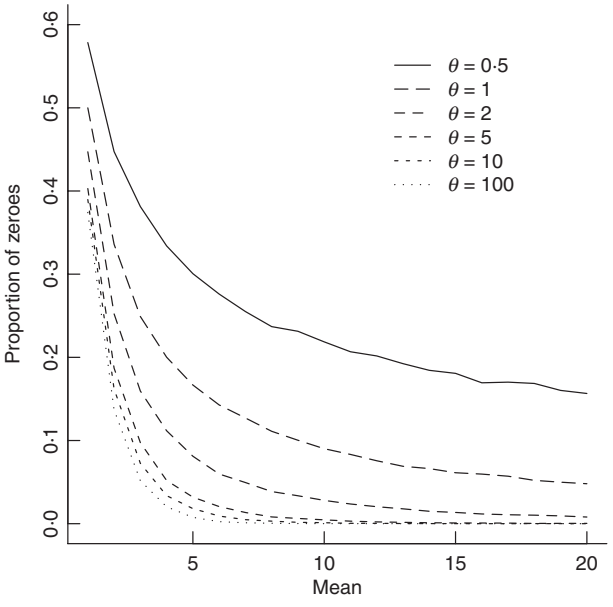

Fig. 1. Proportion of values equal to zero in simulations from a negative binomial distribution. θ controls the dispersion (clumping) in the data: a larger value of θ means lower dispersion.

The amount of bias also depends on the transformation used. When there is little variation (i.e. high θ, when the negative binomial distribution approaches the Poisson), the square-root transformation has little bias, as does the log-transformation when the mean is high, i.e. there are few zeroes (compare with Fig. 1).

The root mean-squared 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 with Fig. 2).

The difference between the negative binomial and quasi-Poisson distribution models is insignificant. The largest absolute difference in bias was $2{\cdot}4 \times 10^{-8}$, and the largest RMSE was only $1{\cdot}1 \times 10^{-8}$, both of which are much smaller than the scales in Figs 2 and 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 stabilizing variances or by making relationships linear (Miller 1997; Piepho 2009) before applying simple linear regression. However, 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 GLMs and their derivatives (McCullagh & Nelder 1989) are the most popular.

For count data, our results suggest that transformations perform poorly. An additional problem with regression of transformed variables is that it can lead to impossible predictions, such as negative numbers of individuals. 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

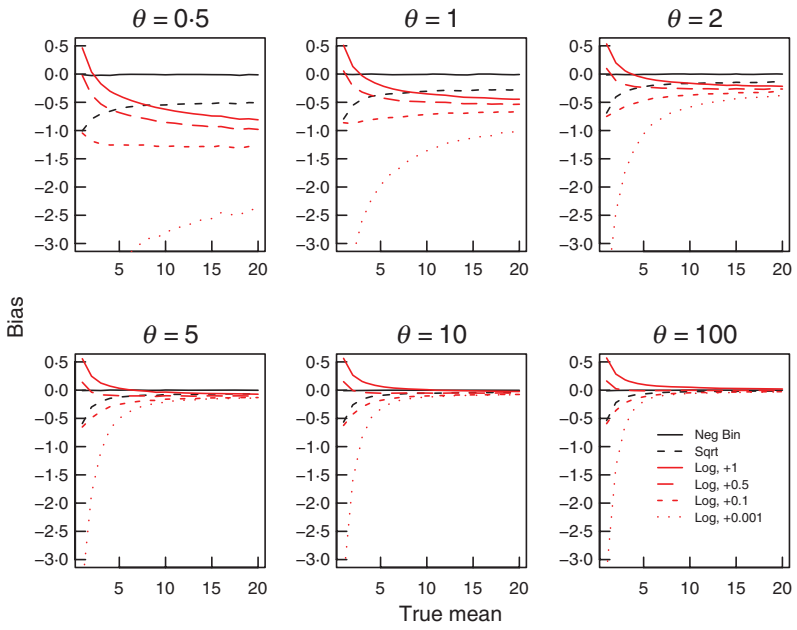

Fig. 2. Estimated mean biases from six different models, applied to data simulated from a negative binomial distribution. A low bias means that the method will, on average, return the 'true' value. Note that the curves for a quasi-Poisson model would be indistinguishable from a negative binomial curve.

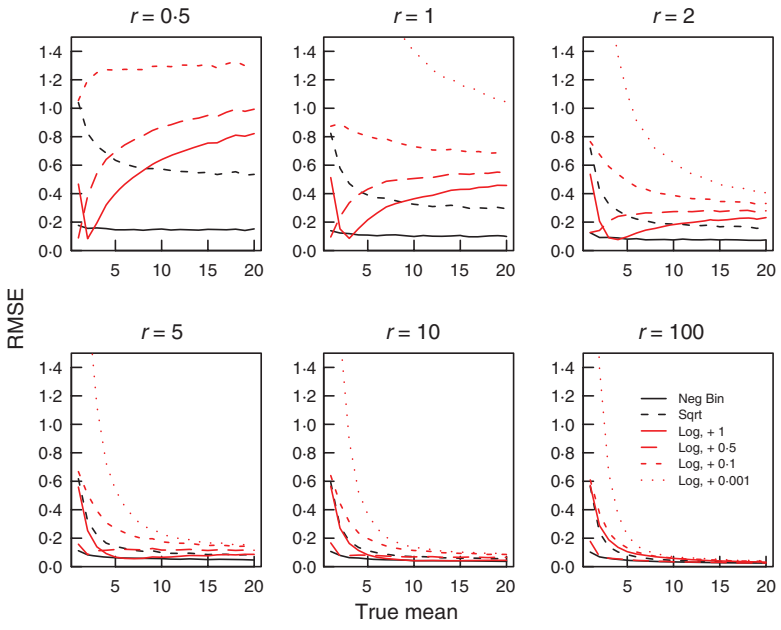

Fig. 3. Estimated root mean-squared error from six different models, applied to data simulated from a negative binomial distribution. Note that the curves for a quasi-Poisson model would be indistinguishable from a negative binomial curve.

foundation for the model. The extra variability that can be added can be chosen according to 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 data set 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. The two models differ in their relationships between the mean and variance; so, if distinguishing them becomes important, this can be done by plotting $(y_i - \lambda_i)^2$ against λ_i: it will be linear for a quasi-Poisson model

but quadratic for a negative binomial model. A clear curve in the plot would therefore suggest that a negative binomial model will provide a better fit. In practice, it is probably advisable to bin the data, i.e. calculate the average mean values and variances for data points with similar mean values, as this will make the plots less messy (Ver Hoef & Boveng 2007).

Even though the choice of the type of GLM depends on many things (O'Hara 2009; Zuur, Ieno & Elphick 2010), 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 27 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. We thank Alain Zuur and an anonymous referee for their helpful comments on an earlier version of this manuscript.

References

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

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

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.

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.

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 edn. Cambridge University Press, Cambridge.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*, 2nd edn. 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 edn. Freeman and Company, New York, New York, USA.

Ver Hoef, J.M. & Boveng, P.L. (2007) Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, **88**, 2766–2772.

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

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 edn. Prentice Hall, Englewood Cliffs, New Jersey, USA.

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

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A. & Smith, G.M. (2009) *Mixed Effects Models and Extensions in Ecology with R*. Springer, New York, NY, USA.

Zuur, A.F., Ieno, E.N. & Elphick, C.S. (2010) A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, **1**, 3–14.

Received 13 December 2009; accepted 19 January 2010
Handling Editor: Robert P. Freckleton

Supporting Information

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

Appendix S1. Simulation code for R.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.