

1 Do not log-transform count data

2
3 O'Hara R.B.^{1,3} & Kotze, D.J.²

4
5 ¹ *Biodiversity and Climate Research Centre, Senckenberganlage 25, D-*
6 *60325 Frankfurt am Main, Germany. Email: bohara@senckenberg.de*

7
8 ² *Department of Biological and Environmental Sciences, PO Box 65, FI-*
9 *00014, University of Helsinki, Finland. Email: johan.kotze@helsinki.fi*

10
11 ³ *Corresponding author: Biodiversity and Climate Research Centre,*
12 *Senckenberganlage 25, D-60325 Frankfurt am Main, Germany.*

13 *Email: bohara@senckenberg.de,*

14 *tel: +49 69 798 40216,*

15 *fax: +49 69 798 40169.*

16
17 **Word Count: 2672**

18 **Abstract**

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

35
36 Keywords: transformation, Poisson, overdispersion, linear models,
37 generalized linear models,

38 **Introduction**

39

40 Ecological data are often discrete counts - the number of individuals or
41 species in a trap, quadrat, habitat patch, on an island, in a nature reserve,
42 on a host plant or animal, the number of offspring, the number of
43 colonies, or the number of segments on an insect antenna. Even though
44 textbooks on statistical methods in ecology (e.g., Sokal & Rohlf 1995; Zar
45 1999; Crawley 2003; Maindonald & Braun 2007) recommend the use of
46 the square root transformation to normalise count data, such data are
47 often log-transformed for subsequent analysis with parametric test
48 procedures (e.g., Gebeyehu & Samways 2002; Magura, Tóthmérész &
49 Elek 2005; Cuesta *et al.* 2008). The reasons for this (log-transforming
50 count data) are not clear but perhaps has to do with the common use of
51 log transformations on all kinds of data, and the fact that textbooks
52 usually deal with the log-transformation first, before evaluating other
53 transformation techniques.

54

55 The main purpose of a transformation is to get the sampled data in line
56 with the assumptions of parametric statistics (such as ANOVA, t-test,
57 linear regression) or to deal with outliers (see Zuur, Ieno & Smith 2007;
58 Zuur, Ieno & Elphick 2009). These assumptions include that the residuals
59 from a model fit are normally distributed with a homogeneous variance. In
60 addition, regression assumes that the relationship between the covariate
61 and the expected value of the observation is linear. Classical parametric

62 methods deal with continuous response variables (weights, lengths,
63 concentrations, volumes, rates) with few "zero" observations. As such, a
64 log-transformation may successfully 'normalise' such continuous data for
65 use in parametric statistics.

66

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

75

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

84

85 More recently, generalized linear models have been developed (McCullagh

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

97

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

105

106 Zeroes can also be handled by using zero inflated models (e.g. Sileshi,
107 Hailu & Nyadzi 2009; Zuur, Ieno & Elphick 2009). When modeling small
108 counts, both zero inflated models and over-dispersed models can account
109 for a large number of zero counts, and there may be little advantage in

110 fitting the zero inflated model. The choice of whether to use these models
111 will thus often depend on an understanding of the biology of the system -
112 the assumption is that there are two types of site, where the species
113 occurs and where it does not. The species may not be caught where it
114 occurs, hence the zero counts can be of two classes (i.e. true absence and
115 present but not sampled). This sort of extension of a model can be an
116 important consideration when modelling count data (for an extreme
117 example, the zero-truncated one-inflated negative binomial, see Kotze *et*
118 *al.* 2003), but is beyond the scope of this paper.

119
120 To address this problem of data transformation we simulated data from a
121 negative binomial distribution (since count data in ecology are often
122 clumped, producing an expected variance that is greater than the mean
123 (see McCullagh & Nelder 1989; White & Bennetts 1996; Dalthorp 2004)),
124 which we then subjected to various transformations (square root, log
125 $(y+n)$). The transformed data were analysed using parametric statistics
126 and compared to an analysis of untransformed data in which the response
127 variable was defined as following either a Poisson distribution with
128 overdispersion or a negative binomial error distribution.

129

130 **Methods**

131

132 Data sets were simulated from a negative binomial distribution, with
133 different values of θ ($\theta = 0.5, 1, 2, 5, 10, 100$). Low θ (also termed k , see

134 fig. 2 in Wright (1991)) indicates greater variance in the data, i.e.
135 stronger clumping. For each simulation, 100 data points were simulated at
136 each of 20 means, μ ($\mu=1,\dots,20$). 500 replicate simulations were carried
137 out for each value of θ .

138

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

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

149

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

$$151 \quad B = \frac{1}{S} \sum_{i=1}^S \hat{\mu} - \mu ,$$

152 and root mean squared error (RMSE):

$$153 \quad RMSE = \frac{1}{S} \sum_{i=1}^S (\hat{\mu} - \mu)^2$$

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

156

157 Simulations and analyses were carried out in the R statistical programme
158 (R Development Core Team 2009), using the MASS (Vernables & Ripley
159 2002) package. The code that was used is available as an online
160 supplement.

161

162 **Results**

163

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

167

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

172

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

177

178 The root mean square error shows a similar pattern, with the negative
179 binomial distribution consistently having a low RMSE, and a high value
180 added to the log transformation being better (Fig. 3). The behaviour of the

181 log+1 transformation is a result of a change in sign of the bias, with the
182 minimum at the point where the mean bias is zero (compare to Fig. 2).

183

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

188

189 ***Discussion***

190

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

201

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

205 The development of statistical and computational methods over the last
206 40 years has made it easier to fit these sorts of models, and the
207 procedures for doing this are available in any serious statistics package.

208

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

222

223 In our simulations, the Poisson and negative binomial models gave almost
224 identical estimates. This suggests that the models are robust to a mis-
225 specification of the relationship between the mean and variance. In
226 contrast, Ver Hoef & Boveng (2007) gave an example from a real dataset
227 where they differed in their predictions. Whilst their data set is unusual
228 (as they acknowledge), it does serve as a warning that our result may not

229 generalize to real data, which rarely has as balanced a design as our
230 simulations. However, even though the choice of which type of
231 generalized linear model to use depends on many things (O'Hara 2009;
232 Zuur, Ieno & Elphick 2009), we do recommend that count data not be
233 transformed to be used in parametric tests. For such data, GLMs and their
234 derivatives are more appropriate.

235

236 **Acknowledgments**

237

238 The order of the authors was determined by the result of the South Africa
239 - England cricket ODI on the 27th September 2009, which England won by
240 22 runs. The study was financially supported by the research funding
241 programme "LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-
242 ökonomischer Exzellenz" of Hesse's Ministry of Higher Education,
243 Research, and the Arts, and the Academy of Finland.

244

245 **References**

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

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

- 252 Crawley, M.J. (2003) *Statistical Computing. An Introduction to Data*
253 *Analysis using S-Plus*. John Wiley & Sons Ltd., England.
- 254 Dalthorp, D. (2004) The generalized linear model for spatial data:
255 assessing the effects of environmental covariates on population
256 density in the field. *Entomologia Experimentalis et Applicata*, **111**,
257 117-131.
- 258 Gebeyehu, S. & Samways, M.J. (2002) Grasshopper assemblage response
259 to a restored national park (Mountain Zebra National Park, South
260 Africa). *Biodiversity and Conservation*, **11**, 283-304.
- 261 Jiao, Y., Chen, Y., Schneider, D., & Wroblewski, J. (2004) A simulation
262 study of impacts of error structure on modeling stock-recruitment data
263 using generalized linear models. *Canadian Journal of Fisheries and*
264 *Aquatic Sciences*, **61**, 122-133.
- 265 Kotze, D.J., Niemelä, J., O'Hara R.B., Turin, H. (2003) Testing abundance-
266 range size relationships in European carabid beetles (Coleoptera,
267 Carabidae). *Ecography*, **26**, 553-566.
- 268 Magura, T., Tóthmérész, B. & Elek, Z. (2005) Impacts of leaf-litter
269 addition on carabids in a conifer plantation. *Biodiversity and*
270 *Conservation*, **14**, 475-491.
- 271 Maindonald, J. & Braun, J. (2007) *Data Analysis and Graphics Using R -*
272 *An Example-Based Approach*, 2nd Edition. Cambridge University Press,
273 UK.
- 274 McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*, 2nd
275 Edition. Chapman & Hall, London.

- 276 Miller, R.G. Jr. (1997) *Beyond ANOVA*. Chapman & Hall/CRC Press,
277 London.
- 278 O'Hara, R.B. (2009) How to make models add up - a primer on GLMMs.
279 *Annales Zoologici Fennici*, **46**, 124-137.
- 280 Piepho, H-P. (2009) Data transformation in statistical analysis of field
281 trials with changing treatment variance. *Agronomy Journal*, **101**, 865-
282 869.
- 283 R Development Core Team (2009) R: A language and environment for
284 statistical computing. R Foundation for Statistical Computing, Vienna,
285 Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- 286 Sileshi, G., Hailu, G. & Nyadzi, G.I. (2009) Traditional occupancy-
287 abundance models are inadequate for zero-inflated ecological count
288 data. *Ecological Modelling*, **220**, 1764-1775.
- 289 Sokal, R.R. & Rohlf, F.J. (1995) *Biometry*. 3rd Edition. Freeman and
290 Company, New York.
- 291 Ver Hoef, J.M, & Boveng, P.L. (2007) Quasi-Poisson vs. negative binomial
292 regression: how should we model overdispersed count data? *Ecology*,
293 **88**, 2766-2772.
- 294 Vernables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics with S*.
295 4th Edition. Springer, New York.
- 296 White, G.C. & Bennetts, R.E. (1996) Analysis of frequency count data
297 using the negative binomial distribution. *Ecology*, **77**, 2549-2557.
- 298 Wright, D.H. (1991) Correlations between incidence and abundance are
299 expected by chance. *Journal of Biogeography*, **18**, 463-466.

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

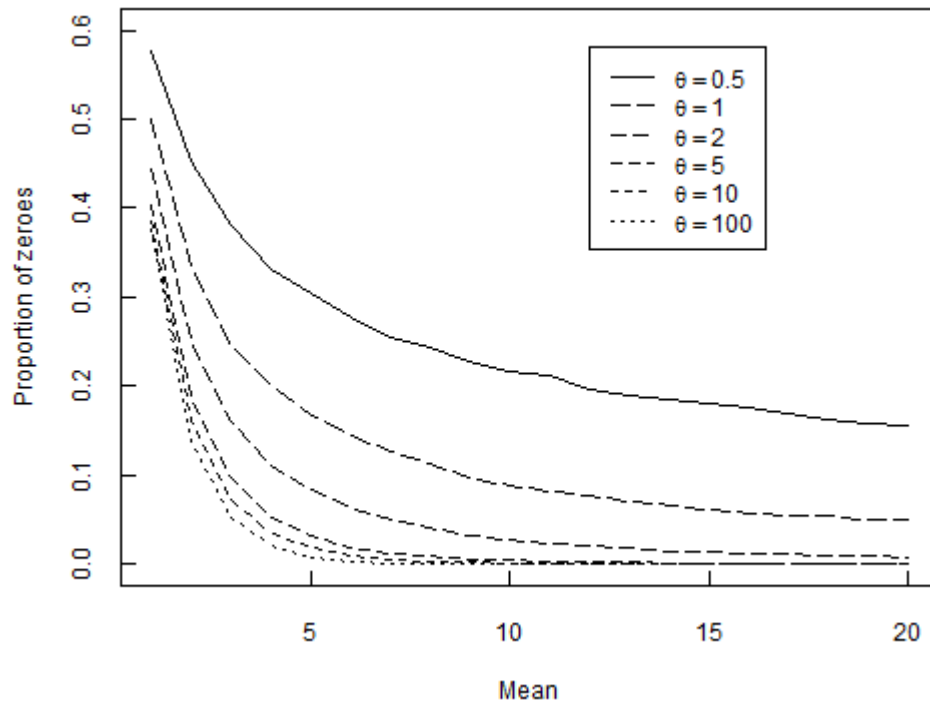
302 Zuur, A.F., Ieno, E.N. & Elphick, C.S. (2009) A protocol for data
303 exploration to avoid common statistical problems. *Methods in Ecology*
304 *and Evolution*. DOI: 10.1111/j.2041-210X.2009.00001.x

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

307

308

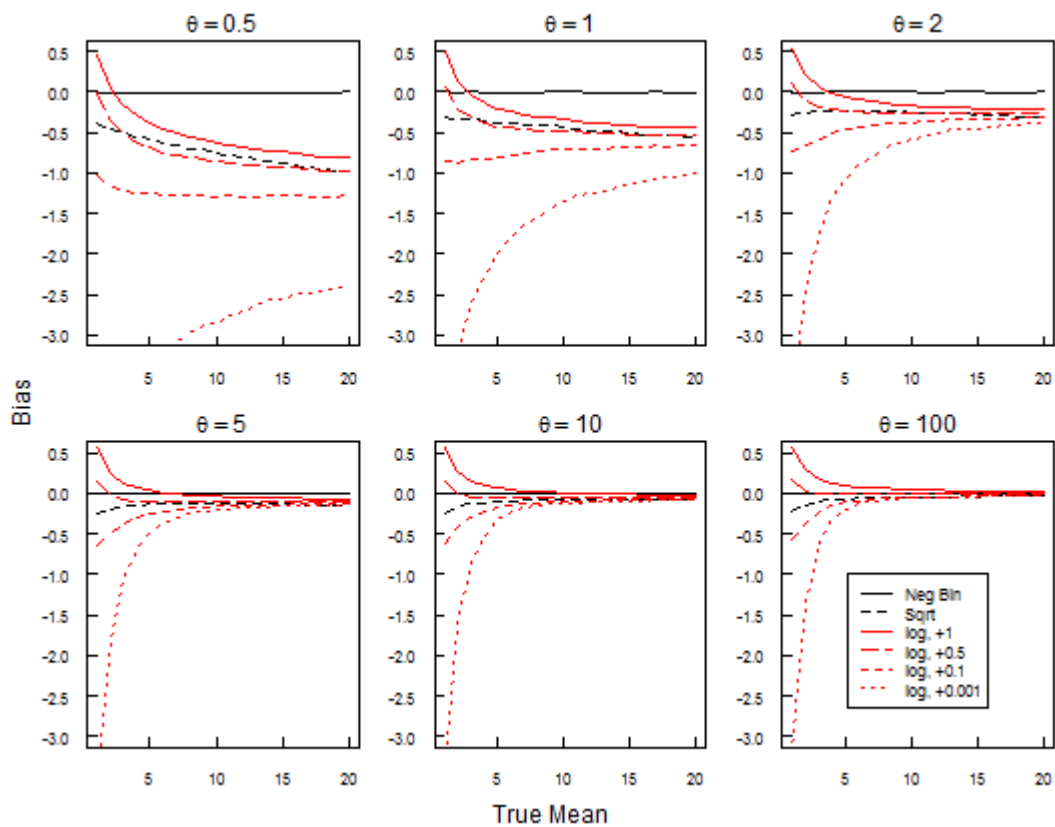
309



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

314

315

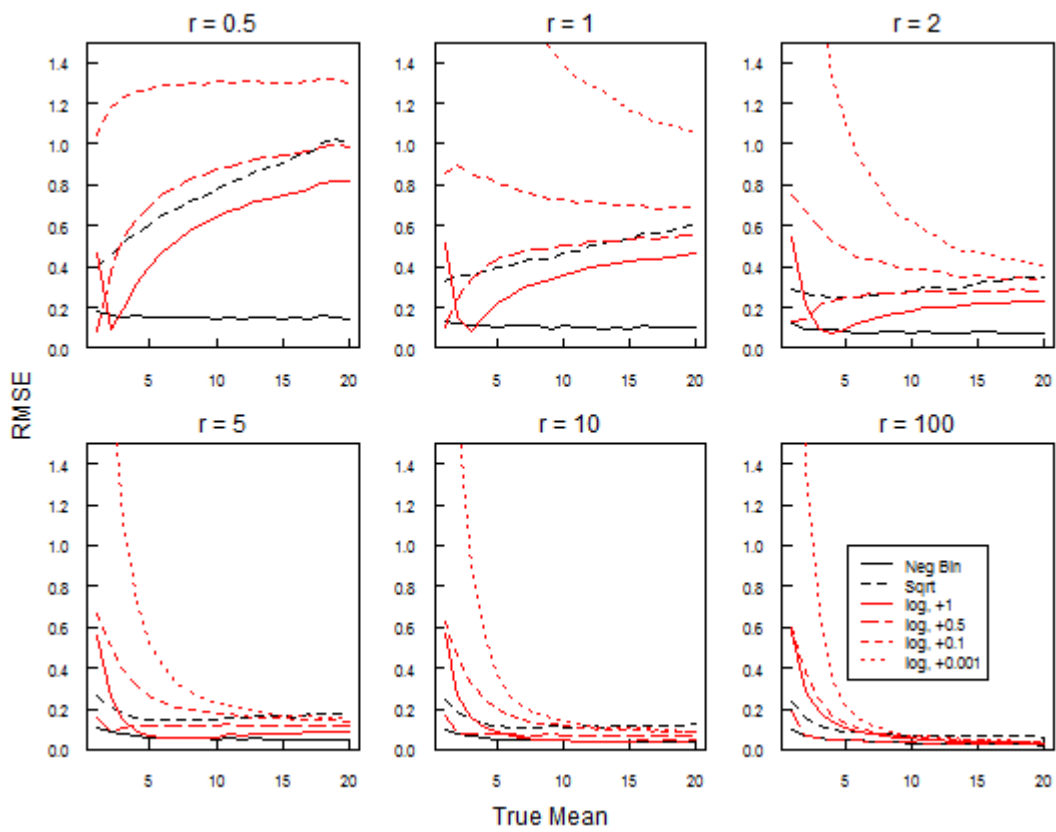


316 **Figure 2.** Estimated mean biases from 6 different models, applied to data
317 simulated from a negative binomial distribution. A low bias means that the
318 method will, on average, return the "true" value.

319

320

321



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