Perils and pitfalls of mixed-effects regression models in biology

Matthew Silk [Corresp., 1, 2], **Xavier Harrison** [1], **David J Hodgson** [Corresp. 1]

[1] Centre for Ecology and Conservation, University of Exeter, Penryn, Cornwall, United Kingdom

[2] Environment and Sustainability Institute, University of Exeter, Penryn, Cornwall, United Kingdom

Corresponding Authors: Matthew Silk, David J Hodgson
Email address: matthewsilk@outlook.com, d.j.hodgson@exeter.ac.uk

Biological systems, at all scales of organisation from nucleic acids to ecosystems, are inherently complex and variable. Biologists therefore use statistical analyses to detect signal among this systemic noise. Statistical models infer trends, find functional relationships and detect differences that exist among groups or are caused by experimental manipulations. They also use statistical relationships to help predict uncertain futures. All branches of the biological sciences now embrace the possibilities of mixed-effects modelling and its flexible toolkit for partitioning noise and signal. The mixed-effects model is not, however, a panacea for poor experimental design, and should be used with caution when inferring or deducing the importance of both fixed and random effects. Here we describe a selection of the perils and pitfalls that are widespread in the biological literature, but can be avoided by careful reflection, modelling and model-checking. We focus on situations where incautious modelling risks exposure to these pitfalls and the drawing of incorrect conclusions. Our stance is that statements of significance, information content or credibility all have their place in biological research, as long as these statements are cautious and well-informed by checks on the validity of assumptions. Our intention is to reveal potential perils and pitfalls in mixed model estimation so that researchers can use these powerful approaches with greater awareness and confidence. Our examples are ecological, but translate easily to all branches of biology.

1 **Perils and pitfalls of mixed-effects regression models in biology**

2 Matthew J. Silk[1,2], Xavier A. Harrison[1] and David J. Hodgson[1]*

4 [1]Centre for Ecology and Conservation, University of Exeter Penryn Campus, Penryn, Cornwall.

5 TR10 9FE.

6 [2]Environment and Sustainability Institute, University of Exeter Penryn Campus, Penryn,

7 Cornwall. TR10 9FE.

8 *corresponding author: D.J.Hodgson@exeter.ac.uk

Abstract

Biological systems, at all scales of organisation from nucleic acids to ecosystems, are inherently complex and variable. Biologists therefore use statistical analyses to detect signal among this systemic noise. Statistical models infer trends, find functional relationships and detect differences that exist among groups or are caused by experimental manipulations. They also use statistical relationships to help predict uncertain futures. All branches of the biological sciences now embrace the possibilities of mixed-effects modelling and its flexible toolkit for partitioning noise and signal. The mixed-effects model is not, however, a panacea for poor experimental design, and should be used with caution when inferring or deducing the importance of both fixed and random effects. Here we describe a selection of the perils and pitfalls that are widespread in the biological literature, but can be avoided by careful reflection, modelling and model-checking. We focus on situations where incautious modelling risks exposure to these pitfalls and the drawing of incorrect conclusions. Our stance is that statements of significance, information content or credibility all have their place in biological research, as long as these statements are cautious and well-informed by checks on the validity of assumptions. Our intention is to reveal potential perils and pitfalls in mixed model estimation so that researchers can use these powerful approaches with greater awareness and confidence. Our examples are ecological, but translate easily to all branches of biology.

Introduction

33   Linear mixed-effects models (LMMs) and generalised linear mixed effects models

34   (GLMMs) have, in recent decades, gained prevalence as statistical tools in biological research.

35   They have replaced blocked, split-plot and hierarchical analysis of variance as the tool of choice

36   for experimental biologists, and have helped countless surveyors of natural systems to measure

37   hierarchical partitions of variation, and to deal with non-independence among their subjects.

38   (G)LMMs are flexible, powerful, well-served by popular statistical software, and relatively easy

39   to build and interpret. However, inevitably, they are also easy to get wrong. As peer reviewers,

40   too often we see (G)LMMs being used naively and without validation.

41   (G)LMMs differ from simple linear models (LMs) and generalised linear models (GLMs)

42   by incorporating random effects among the explanatory variables alongside fixed effects

43   (Gelman & Hill, 2006; Bolker et al., 2009; Zuur, Hilbe & Ieno, 2013; Harrison et al., 2018).

44   Fixed effects represent variables with intercepts, means or slopes to be estimated. Random

45   effects describe experimental or survey units as members of groups, with these groups drawn

46   from a larger population of other, unmeasured groups. For example, we might measure features

47   of 10 populations of seabirds, but many hundreds of other populations of seabirds will not be

48   measured in our study. Fitting 'population ID' acknowledges that each population represents a

49   different group, that observations within a group are likely non-independent (i.e. seabirds in the

50   same population will likely be more similar to each other than to seabirds in other populations),

51   *and* that we have only measured a subsample of possible groups. The random effect components

52   of mixed effects models infer the variance associated with group membership (Gelman & Hill,

53   2006; Schielzeth & Nakagawa, 2013). Consequently, the use of random effects is effective when

54   measurements of a response variable within "groups" or random effect levels tend to be more

55 similar to each other than to those in the wider population. By inferring variance among groups

56 rather than group-specific parameters, random effects absorb fewer degrees of freedom than

57 fixed effects and this is exploited in many analyses to absorb real but 'unwanted' variation or

58 covariance in the data (Bolker et al., 2009; Harrison et al., 2018). In other analyses, such as those

59 used in quantitative genetics, the relative magnitudes of the variance components are of primary

60 interest, because they provide estimates for the variation in traits explained at different levels of a

61 model's hierarchy.

62   The inclusion of random effects in biological regression models is often seen as a

63 panacea for dealing with difficult or structured datasets. One will often read statements in papers

64 that a grouping variable was 'fitted as a random effect to control for non-independence among

65 groups'. However, if not used appropriately, (G)LMMs can fail to achieve their purpose, or can

66 yield false positive tests of importance of fixed effects (Schielzeth & Forstmeier, 2008;

67 Forstmeier, Wagenmakers & Parker, 2017; Arnqvist, 2020). The correct specification of random

68 effects is essential to successful analysis, but guidance is often couched in specialist statistical

69 language, code or algebra and can be hard to find. Here we use non-specialist language to

70 highlight the risks of coupling fixed and random effects naively in models. We explain seven

71 common perils and pitfalls (Table 1), use simple data simulations to demonstrate the risk of

72 inappropriate mixed-effect modelling, illustrate the problems caused by these pitfalls and show

73 the reader to diagnose them in their own data and models.

74

75 **Mixed model basics**

76   The fundamental assumption shared by simple linear models and mixed effects models is

77 that residual error for each measurement of the response variable is an independent draw from a

78   Normal distribution with mean zero and fixed variance. (G)LMMs are mixed because they

79   include both fixed and random effects, and linear because they describe the relationship between

80   the observed data and the explanatory variables as the additive contribution of intercepts and

81   linear slopes (of the fixed effects), and deviations from these expectations caused by random

82   effects and residual noise. The adjective Generalised in GLMMs refers to the use of canonical

83   link functions that help deal with non-Normal response variables. For both LMMs and GLMMs

84   random effect group means are typically assumed to be Normally distributed on the link scale,

85   although other distributions can be considered (described in more detail below). In general, the

86   choice of whether a variable should be included as a random effect or a fixed effect is best

87   decided by the researcher working on a problem based on a combination of question-driven and

88   practical considerations (Schielzeth & Nakagawa, 2013; Harrison et al., 2018). We expand on

89   other practical considerations in subsequent sections.

90   As a verbal model, any LMM can be described as a response variable (y) being an

91   additive function of a global intercept, change caused by fixed effects, deviation caused by

92   membership of a random effect category, and any residual variation not explained by the fixed or

93   random effects. This is most elegantly described using statistical algebra (Fig. 1; this will be the

94   only mathematical equation we will present). A simple cartoon of LMM inference is provided in

95   Figure 2. While it is possible to fit mixed effect models in a variety of software (e.g. SAS [SAS

96   Institute Inc.]; SPSS [IBM Corp.]; STATA [StataCorp. 2017]), we focus on the fitting of LMMs

97   and GLMMs in the R statistical environment (R Core Team, 2019) using the packages lme4

98   (Bates et al., 2015) and nlme (Pinheiro et al., 2014). We, along with many of our biologist

99   colleagues, enjoy the open source philosophy and community spirit associated with R. It is also

100 free and extremely powerful.

101

102

Perils and Pitfalls

104   In this guide, we identify seven major risks of misuse of (G)LMMs. We explain each of

105   them in simple terms, discuss when they are most likely to cause problems and provide solutions

106   in the form of better-formulated models. Some of the problems stem from an assumption that

107   simple (G)LMMs are a panacea for non-independent data and can be used universally to model

108   difficult datasets. However, complicated hierarchical designs must be treated with appropriate

109   caution, as (G)LMMs are not infallible if specified incorrectly (Arnqvist, 2020). Further issues

110   can arise because these hierarchical models are amenable to the standard toolkit of significance

111   tests, model simplification or multi-model inference taught in undergraduate statistics modules.

112   Analysts, taught by frequentists to infer significance from p-values, commonly struggle to

113   understand how to infer the importance of fixed and random effects reported by (G)LMMs.

114   There is no simple answer here, but simple mistakes can be avoided. The perils and pitfalls we

115   describe are focused on attempts to infer the importance of fixed-effect explanatory variables,

116   but most remain relevant to attempts to understand the importance of hierarchical variance

117   components.

118

Peril #1: (G)LMM p-values can be anti-conservative

120   A common worry for the frequentist statistician is the risk that the p-value of a test (i.e. the

121   probability of finding signal as extreme, or even more extreme, than the observed signal,

122   supposing that the null hypothesis is true) might exceed 5% even when the null hypothesis is

123   true. This situation, also known as an inflated type I error rate or being anti-conservative, risks

124   the false conclusion of the importance of a fixed effect explanatory variable (false positive) and

125   has contributed to a replication crisis in some research fields (Forstmeier, Wagenmakers &

126   Parker, 2017). Anti-conservative inference from (G)LMMs can arise when sample sizes are

127   small because maximum likelihood estimates of variance components become biased. This

128   problem is fixed by Restricted Maximum Likelihood (REML) algorithms, which are the default

129   settings for (G)LMMs in most software. However, REML models are not amenable to standard

130   model simplification or tests of significance. It is common practice to convert REML models to

131   Maximum Likelihood and simplify these using likelihood ratio tests. These tests follow chi-

132   square distributions only approximately. As sample size increases, the approximation improves,

133   but for small sample sizes the outcome tends to be anti-conservative. Neither can analysts safely

134   fall back on information theoretic approaches such as the well-known Akaike Information

135   Criterion to compare REML models. AIC is a function of the model's likelihood and the number

136   of parameters estimated. However, REML likelihoods contain a nuisance factor that is dealt with

137   differently by different software, meanwhile the number of parameters estimated is linked to the

138   number of degrees of freedom associated with the random effects, and that number is hard to

139   define (Bolker et al., 2009). Adjustments for AIC (such as AICc) that cope with small sample

140   sizes are approximations and do not completely deal with this problem.

141   Guidance on what qualifies as a small sample size is lacking (Gelman & Hill, 2006), but it

142   is safe to assume that some datasets in biology will suffer from these issues. Small datasets are

143   also at risk from over-fitting of models, which can also lead to high rates of false positives

144   (Forstmeier, Wagenmakers & Parker, 2017), and with more complex model structures it can be

145   more challenging to identify situations where overfitting is an issue. These potential issues can

146   arise regardless of the approach taken to statistical inference, although they can be exacerbated

147 by model simplification (Forstmeier & Schielzeth, 2011; Forstmeier, Wagenmakers & Parker,

148 2017)

149   A simple simulation illustrates how this issue arises in small datasets. Imagine a simple

150 treatment-control experiment that applies a treatment to half of the members of each of six

151 families. Here, family ID is the random intercept term, and the mean value for the response

152 variable for each family is drawn from a Normal distribution with mean = 0 and standard

153 deviation = 1. We then simulate values for the response variable so that the experimental

154 treatment itself has no effect on values of the response variable (mean = family mean and

155 standard deviation = 1). Here, family ID is the dominant source of variation in the response

156 variable, and the null hypothesis is true and should be accepted. We conducted a poorly

157 replicated version of this experiment, in which only one individual from each family is measured

158 in each experimental group (treatment and control). We then tested for statistical significance of

159 the treatment effect using mixed-effects models and: a) by determining if the 95% confidence

160 interval of the treatment effect in the full model overlapped zero; b) a likelihood ratio test that

161 compared Maximum Likelihood models with and without the fixed effect of treatment; and c) a

162 comparison of the AIC of the models with and without the treatment effect. We used a sigma

163 threshold of 0.05 for approach b) and considered the effect of treatment to be important for

164 approach c) when the AIC of the model including the fixed effect of treatment was more than 2

165 AIC units less than the null model. All three approaches produced high rates of false positives: in

166 122 out of 1000 (12.2%) simulations, 105 out of 1000 (10.5%) and 100 out of 1000 (10%)

167 respectively (with identical results from nlme and lme4). This is around double the 5% threshold

168 desired by frequentists and indicates a high false positive rate. The analysis is anti-conservative.

169 The problem of anti-conservatism fades with increasing sample sizes: when each of six families

170   contribute ten control and ten treated individuals, the error rates recorded were 5.9%, 5.7% and

171   3.3% (5.1% for nlme) respectively. Consequently, while this is a pitfall to be aware of when

172   using (G)LMMs in small datasets, it is likely to be less important with adequate replication.

173

174   **Peril #2: (G)LMMs do not cure all types of pseudoreplication**

175   Pseudoreplicated experimental designs are those in which fixed effects vary at a

176   hierarchical scale higher than the measured subject. If, for example, whole families are exposed

177   to experimental treatment, and family members resemble each other beyond the impact of

178   treatment, then individual members of families do not represent independent measures of the

179   effect of treatment and should not be considered true biological replicates. In an extreme

180   situation where there are many non-independent observations within each of a small number of

181   random effect levels this could cause the number of residual degrees of freedom (those used to

182   infer residual variance) to greatly exceed the number of experimental units to which treatments

183   were applied (number of families). While, traditionally, F-ratio tests would reveal this problem,

184   likelihood ratio tests hide it from the analyst because they use only the degrees of freedom

185   associated with the explanatory variables (Arnqvist, 2020). AIC values also hide the problem by

186   providing no information on degrees of freedom at all. When (G)LMMs are fitted in R this

187   problem is managed by fitting fixed effects at the correct level of hierarchy. This deals with the

188   pseudoreplication pitfall but can result in poor replication of the fixed effects if the number of

189   random effect groups is low (Fig. 3). The model may struggle to partition variance explained by

190   the random effects from unexplained (residual) variance. This results in a marked decline in the

191   accuracy of model inference and increased risk of encountering anti-conservative p-values from

192   likelihood ratio tests (see Peril #1) with no way for the inexperienced analyst to check.

193   To demonstrate, we repeated our simulation from the previous section, but this time

194   applied experimental treatments to whole families (so that every individual in each family was in

195   the same experimental group). We simulated 600 observations in total, then in separate analyses

196   we divided observations among 6, 8, 10, 12, 15, 20 or 24 families (with concomitant reduction in

197   sample size per family, ranging from 100 to 25). Each family was allocated either to the

198   treatment or the control group. In this design, the individuals in each family do not provide

199   independent measures of the effect of treatment. As before, we simulated treatment to have zero

200   effect so the null hypothesis is true and should be accepted. When the number of families was

201   small we observed a high false positive rate, even though the per-family sample size was large:

202   the effect of treatment was statistically significant in >10% of simulations, except when a t-test

203   using the mean and standard error of the treatment effect from the nlme model was used (Fig. 3).

204   The false positive rate then declined to approximately 5% in cases where there more families

205   were sampled (e.g. our simulation runs with 20 or 24 families). Our simulations therefore

206   demonstrate that caution is required in interpreting the statistical significance of group-level

207   explanatory variables when the number of random effect levels is small.

208   There are additional circumstances where using random effects to attempt to control for

209   non-independence may not be straightforward. Imagine an experiment where we grow two

210   genotypes of wheat in replicate planters containing varying, finite amounts of phosphate to test a

211   hypothesis about the performance of each genotype in different conditions. You would be correct

212   in assuming that measurements from within the same feeding trial are not independent. One

213   potential option here would be to fit planter ID as a random effect to control for this. However,

214   higher uptake by one genotype within an experimental block means less phosphate available for

215   the other, causing the performance of each genotype within a planter to be negatively correlated.

216 A standard "blocked" (G)LMM with planter ID as a random effect would not control for this

217 issue.

218   The risk of anti-conservatism does not mean that (G)LMMs have no value in

219 pseudoreplicated designs. However, it does mean that the analyst must have a clear

220 understanding of the hierarchical design of the model, in particular whether fixed effects are

221 observation-level (apply uniquely to each observation in the dataset) or group-level (apply

222 uniquely to each level of the random effect) (Schielzeth & Nakagawa, 2013). It is also important

223 to be aware of the number of independent experimental units, the extent and nature of

224 pseudoreplication and the ability of the statistical software to test the importance of signal at the

225 correct level of the design hierarchy. Experimental units, i.e. those which can be claimed to be

226 independent replicates of the experiment or survey design, must be replicated sufficiently to

227 provide confidence in their inference *and* to avoid Type I errors (false positives).

228

229 **Peril #3: Too few random effect categories**

230   The standard definition of a random effect is that it should describe membership of a

231 group that is part of a random sample of a larger population of groups. Random effect

232 hyperparameters infer the variance of means and slopes among groups. Inference of any variance

233 parameter requires several independent replicates and improves with increasing sample size of

234 the number of groups. Hence, accurate estimation of random effects can only occur when several

235 groups are represented in a dataset (Harrison, 2014, 2015; Harrison et al., 2018). In practice,

236 however, mixed effects models are commonly used to capture and absorb any non-independence

237 due to group membership. When only few groups have been measured, inference of random

238 effect hyperparameters will be poor; models might be degenerate; and there will be

239 consequences for the correct inference of fixed effect parameters and tests of their importance.

240 For example, fitting sex as a random effect in a model implies the choice of calculating a

241 variance using only two values: we wouldn't infer a simple standard deviation from a sample of

242 two individuals, so we shouldn't use the (G)LMM algorithm to make the same crude inference.

243 In general, it is suggested to only fit a categorical variable as a random effect if it has 5-6 levels

244 or more (Bolker et al., 2009; Harrison, 2015).

245   We illustrate this problem in Figure 4 using simple data simulations. A sample of 200

246 individuals is subdivided into between 2 and 40 groups. A single covariate has a positive effect

247 on our response variable. Full code for the example is provided in the supplementary material.

248 Estimates of random effect variance are biased small when there are fewer than six or seven

249 groups (Fig. 4A), with the greatest problems caused by having only two or three groups. The

250 inaccuracy of estimates of the random effect variance shows an exponential decline. This means

251 that models tend to underestimate the standard error of the fixed-effect covariate's slope when

252 there are too few random effect levels (Fig. 4B), indicating increased risk of Type I (false

253 positive) errors when there are fewer random effect levels.

254

255 **Peril #4: Effect of predictor varies among random effect categories**

256   Many studies that use (G)LMMs use a random intercepts model, in which the random

257 effect measures or controls for variation among group means. However, there is no guarantee

258 that this is a sufficient control for variation among groups, especially when modelling a covariate

259 whose effect may itself vary among groups (Schielzeth & Forstmeier, 2008; Barr et al., 2013;

260 Bell, Fairbrother & Jones, 2018). The use of random slopes models is advisable in these contexts

261 (Barr et al., 2013). Random-intercept models risk anti-conservative tests of importance, and

262 typically predict responses of unsampled groups worse than models that include random slopes

263   (Bell, Fairbrother & Jones, 2018). To provide an ecological example, we simulated a relationship

264   between oxidative stress and parasite burden in 20 populations of wild amphibian (n= 20 frogs

265   measured per population), where parasite burden is a non-integer measure of genomic

266   equivalents of a parasite. We allowed both mean parasite burden (random intercepts) and the

267   strength of relationship between parasites and oxidative stress (random slopes) to vary by

268   population. We also randomly assigned populations to a 2-level factor as either 'treatment' or

269   'control', so having no effect on parasite burden. We fitted a 'maximal' model with Gaussian

270   error structure allowing for an interaction between oxidative stress and treatment, and containing

271   either only random intercept for population ID, or both a random intercept and random slope for

272   oxidative stress given population. After 10,000 iterations, when fitting only a random intercept,

273   the model erroneously estimated the treatment*oxidative stress interaction to be significant (95%

274   confidence intervals not crossing zero) in 30% of cases. Fitting a random slope model reduced

275   the incidence of Type I (false positive) errors to 6.7%, far closer to the desired 5% rate. For

276   further details on this phenomenon, and a more detailed set of simulations, see (Schielzeth &

277   Forstmeier, 2008). Clearly, not including a random slope results in elevated type I (false

278   positive) error rates.

279   An important risk of random-intercept models is that unusual groups can cause false

280   inference of whole-population effects. Imagine a test for a linear relationship between body size

281   and the size of a status signal in a hypothetical bird species consisting of 10 different subspecies.

282   There is a positive body size-signal size relationship for only one of these subspecies (Fig. 5a). A

283   random intercepts model may conclude that there is a positive effect of body size on status signal

284   size across *all* subspecies if the relationship in that single subspecies is sufficiently strong (Fig.

285   5b). In contrast a random slopes model that controls for differences in this relationship among

286   subspecies is likely to detect no overall effect, but also show that the slope for one subspecies

287   was unusual. The difference between the random slopes and random intercept models does not

288   occur because of differences in their estimates of the effect of body size (Fig. 5c) but instead

289   because the random intercept model consistently underestimates the standard error around this

290   estimate (Fig. 5d). Note that often random slope models can return low estimates of variance for

291   the among-group slopes, but this is to be expected if there is only one group with a strong

292   relationship with the response. This is especially the case if the random effects sample size is

293   large. Measuring the 'importance' of random effects based on these variance components is a

294   contentious issue (Harrison et al., 2018), but random effects shouldn't *automatically* be

295   immediately discounted/removed just because a variance component appears negligible.

296

297   **Peril #5: Correlations between fixed and random effects, and informative**

298   **group sizes**

299   Mixed models assume that values of fixed effect variables are independent of the

300   groups/clusters used as levels of a random effect. When this assumption is violated, the models

301   suffer 'confounding by cluster' (Seaman, Pavlou & Copas, 2014a). Confounding by cluster will

302   therefore be a potential problem when there is both within- and between-group variation in the

303   values of an explanatory variable. It can lead to misleading estimates for fixed effect parameters

304   and biased estimations of variance components (Neuhaus & McCulloch, 2006). A number of

305   potential solutions have been suggested (Seaman, Pavlou & Copas, 2014a), although one of the

306   most accessible and intuitive is to decompose the effect of a fixed effect variable into between-

307   group (and therefore group-level) and within-group (observation-level) covariates (Neuhaus &

308   McCulloch, 2006). For example, a study determining the relationship between fitness and body

309   size in a population made up of multiple social groups could contain a group-level effect of mean

310   body size, and an individual-level effect that is group-mean centred, so that it describes the size

311   of that individual relative to other individuals in the group.

312   A similar problem can arise if the cluster/grouping size (number of samples for each level

313   of the random effect) is correlated with one or more fixed effect variables (Seaman, Pavlou &

314   Copas, 2014b). For example, if modelling the association between maternal stress (response

315   variable) and offspring weight at fledging (explanatory variable) in a passerine bird, clutch ID

316   would typically be included to control for variation in offspring weight among different breeding

317   attempts (as might be caused by genetic differences, parental effects or the shared environment,

318   for example). However, if females that are more stressed tend to lay smaller clutches then the

319   size of the cluster is informative. Informative cluster sizes can result in biased inference,

320   especially when fixed and random effects are correlated (Neuhaus & McCulloch, 2011). This is a

321   particular problem for models that only include a random intercept, where failing to account for

322   informative cluster sizes can lead to biased estimates for the intercept. However, for most

323   covariates model estimates are unaffected by the presence of informative cluster sizes (Neuhaus

324   & McCulloch, 2011). One potential solution for dealing with informative cluster sizes is by

325   including cluster size as a covariate, however this is not a sensible choice when cluster size lies

326   on the causal pathway linking a particular explanatory variable to a response variable (Seaman,

327   Pavlou & Copas, 2014a): inclusion of the cluster size as a fixed covariate risks masking the real

328   influence of the main explanatory variable. Another more complex approach is to jointly model

329   the response variable of interest and cluster size (Dunson, Chen & Harry, 2003; Chen, Zhang &

330   Albert, 2011; Seaman, Pavlou & Copas, 2014a).

331

Peril #6: Random effect category means are not Normally distributed

332

333   In many situations random effect means or slopes deviate from a Normal distribution.

334   They might be clustered along the y-axis, or might be skewed or heavy-tailed. Skew and kurtosis

335   are rarely considered by biologists, but are important when comparing variances among samples

336   (Hosken, Buss & Hodgson, 2018). Typically, the random effect hyperparameter used in

337   (G)LMMs is the standard deviation of a Normal distribution. It is possible to specify alternative

338   hyperparameters and allow random effects to have non-Normal distributions (Zhang et al., 2008;

339   Molenberghs et al., 2010, 2012; Fabio, Paula & de Castro, 2012) but this requires non-standard

340   modelling algorithms. In general, both LMMs and GLMMs have been found to be impressively

341   robust against misspecification of the random effects distribution (McCulloch & Neuhaus, 2011;

342   Neuhaus, McCulloch & Boylan, 2013; Schielzeth et al., 2020). The estimates of fixed effect

343   parameters for individual-level variables are particularly robust (McCulloch & Neuhaus, 2011),

344   while estimation of random effects and variances is more susceptible to misspecification

345   (McCulloch & Neuhaus, 2011). However, when fixed effects are correlated with random effects,

346   and if other assumptions of mixed models are not met, then misspecification of the random

347   effects distribution can influence parameter estimates (Neuhaus & McCulloch, 2011). Diagnostic

348   tools such as those found in the R package 'sjPlot' (Ludecke, 2019) allow users to diagnose the

349   validity of their assumption of normality of random effects, and should be used where

350   appropriate. While LMMs and GLMMs are typically robust to misspecification, other

351   approaches are available when these assumptions are violated, for example the application of

352   mixture models (Hamel, Yoccoz & Gaillard, 2017). Direct comparison of choices of random

353   effect distributions can be made using Bayesian mixed effects models coded in standard software

354   such as JAGS (Plummer, 2003) or Stan (Carpenter et al., 2017).

355

Peril #7: Random effect categories used as a proxy for covariance structure

357 It is not unusual for non-independence among residuals, caused by spatial, temporal or

358 phylogenetic autocorrelation, to be modelled using categorical random effects (e.g. postcode,

359 year or taxonomy). Despite increased prevalence of regressions that explicitly model covariance

360 structures (Housworth, Martins & Lynch, 2004; Dormann et al., 2007; Miller, Franklin &

361 Aspinall, 2007; Hadfield & Nakagawa, 2010; Ives & Helmus, 2011; Rousset & Ferdy, 2014),

362 many researchers elect to control for these dependencies with categorical variables as an

363 alternative. While some coarse control for spatial (or temporal) structure is better than none at

364 all, the success of using a categorical random effect can vary depending on the nature of the true

365 covariance structure in the data. First, random effects do not control for all types of non-

366 independence and are useful only when there is positive autocorrelation (i.e. measurements

367 closer together in time or space tend to be more similar). Second, the success of using a

368 categorical proxy for a covariance structure will depend on how the granularity of the effect used

369 (e.g. site ID versus region ID) compares with the scale of any autocorrelation.

370 We demonstrate here that the success of categorical proxies for continuous covariation

371 depends on how the granularity of the random effect compares with the scale of any

372 autocorrelation. We simulated two Gaussian fields that influence the residual error in the

373 relationship between a single explanatory variable (Temperature) and single response variable

374 (Height). The first (Fig. 6a) varied at a much finer scale than the second (Fig. 6d). We varied the

375 slope of the relationship, so that we could examine the impact on model performance at a various

376 effect sizes. There were 320 measurements in total, occurring in eight regions that each

377 contained four sites, with 10 individuals measured per site. We fitted three models: one directly

378 modelled spatial covariance in the residuals (using Generalised Least Squares); one using site as

379   a categorical random effect; and one using region as a categorical random effect. For the model

380   in which the spatial field varied at a finer scale, model estimates from the models that used site

381   as a random effect were in much closer agreement with the explicitly spatial model (Fig. 6b) and

382   much more precise (Fig. 6c) than those that used region as a random effect. In contrast, when the

383   scale of the spatial field was broader, the accuracy and precision of model estimates was much

384   more similar (Fig. 6e-f). Biases in the accuracy and precision of model predictions will affect

385   Type I (false positive) and Type II (false negative) error rates. This helps illustrate the benefit of

386   selecting a categorical effect to control for spatial or temporal structure in the data that is as well

387   matched to the autocorrelation in the data as possible.

388

389   **Detecting and resolving problems with mixed model estimation**

390   The issues identified here can all be identified with adequate data checks and checks of

391   model fit (Zuur & Ieno, 2016; Harrison et al., 2018).

392   First, we recommend having a clear idea of the structure and hierarchy of the model.

393   Identifying categorical variables with too few levels for random effect estimation can help avoid

394   Peril 3, while identifying the level of the hierarchy that fixed effect variables apply at

395   (observation-level versus group-level) can help identify situations where there is a risk of Peril 2

396   having an impact. Similarly, a clear idea of the hypothesis and model structure can help identify

397   scenarios where a random slopes model might be more appropriate than a random intercepts

398   model (Peril 4).

399   Second, we emphasise the importance of data exploration (Zuur, Ieno & Elphick, 2010).

400   Plots of the raw data can help identify unusual random effect categories (helping to avoid Peril

401   6), small sample sizes (Peril 1) or confounding by cluster and informative cluster sizes (both part

402   of Peril 5). Similarly, raw data plots might help provide an initial idea of the scale of any spatial

403 or temporal structure to the data and so be informative in deciding an appropriate random effect

404 (Peril 7).

405   Third, we highlight the value of conducting full model checks (Zuur & Ieno, 2016;

406 Harrison et al., 2018). Plots of residuals for each level of the random effect are important in

407 identifying heteroscedasticity in the residuals between levels of the random effect, or

408 relationships between residuals and fitted values that necessitate random slopes models (Peril 4).

409 Semi-variograms check the need for explicit modelling of covariance structures to avoid Peril 7

410 (Fletcher & Fortin, 2018). The simplest method to assess the extent to which the random effects

411 distribution differs from normal (to avoid Peril 6) are quantile-quantile plots. However, various

412 other diagnostic tools are available (Drikvandi, Verbeke & Molenberghs, 2017; Efendi et al.,

413 2017). We demonstrate a comprehensive but simple approach to model checking in the

414 supplementary material. We provide code to allow users to be confident in generating these

415 checks themselves as well as using functions provided by existing R packages. New R packages

416 such as DHARMa (Hartig, 2020) are making these model checks more available and accessible

417 even for more complex GLMMs.

418   There are also now R packages that make more sophisticated modelling approaches

419 available to a wider audience. For example, brms (Bürkner, 2017) provides an interface to the

420 Bayesian statistical model fitting software Stan (Carpenter et al., 2017). Users can exploit coding

421 syntax similar to lme4, and an array of model specifications are available that deal with various

422 pitfalls described here. Similarly, glmmTMB (Brooks et al., 2017) provides software to facilitate

423 the fitting of more complex error distributions and correlation structures using similar syntax to

424 lme4.

425   (G)LMMs are reassuringly robust to some of the individual pitfalls we have described

426   here (Schielzeth et al., 2020). However, there is very little understanding of the combined impact

427   of multiple violations of the core assumptions: there is plenty of scope for simulation studies of

428   the power and anticonservatism of mixed effects models. The pitfalls described here exist and

429   are prevalent in the biological literature, with little opportunity for readers to check the validity

430   of published models and results. Hence, raising awareness of these potential pitfalls and

431   considering them together can help researchers use mixed models more assuredly and avoid

432   pitfalls that impact on their statistical inference.

433

434   **Conclusions**

435   Mixed modelling approaches are firmly embedded as state-of-the-art for the analysis of

436   biological data. As these methods become available to a growing user base it is necessary to

437   reveal the perils and pitfalls associated with them (Arnqvist, 2020). Often with biological

438   datasets it is not possible to meet the statistical assumptions of these models perfectly and it may

439   be necessary to make compromises in model design. We have provided guidance on a set of

440   typical perils and pitfalls by providing insights from the statistical literature and illustrating key

441   points using simulated case studies and examples. While (G)LMMs can be robust to some of the

442   perils we describe (Schielzeth et al., 2020), there remains little understanding of their combined

443   impacts. Readers can only assess the quality of inference from mixed effects models if their

444   hierarchical structures, levels of true replication, and checks of validity, are described clearly and

445   honestly. This review will help readers be more aware of some of the key perils in mixed model

446   design and so improve statistical inference when testing biological hypotheses. We hope that by

447   providing an overview of these perils, we can help researchers feel better informed and more

448   confident that they are using mixed modelling approaches to draw correct conclusions from their

449   data.

450

Acknowledgements

452   The authors thank Tom Houslay and Julian Evans for useful discussions while developing the

453   paper.

References

Arnqvist G. 2020. Mixed Models Offer No Freedom from Degrees of Freedom. *Trends in Ecology & Evolution* 35:329–335.

Barr DJ, Levy R, Scheepers C, Tily HJ. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68:255–278.

Bates D, Maechler M, Bolker B, Walker S. 2015. Fitting Linear Mixed-Effects Models using lme4. *Journal of statistical Software* 67:1–48.

Bell A, Fairbrother M, Jones K. 2018. Fixed and random effects models: making an informed choice. *Quality & Quantity*:1–24.

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White J-SS. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* 24:127–135. DOI: 10.1016/j.tree.2008.10.008.

Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, Skaug HJ, Machler M, Bolker BM. 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R journal* 9:378–400.

Bürkner P-C. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80:1–28.

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A. 2017. Stan: A probabilistic programming language. *Journal of statistical software* 76.

Chen Z, Zhang B, Albert PS. 2011. A joint modeling approach to data with informative cluster size: robustness to the cluster size model. *Statistics in medicine* 30:1825–1836.

Dormann CF, M McPherson J, B Araújo M, Bivand R, Bolliger J, Carl G, G Davies R, Hirzel A, Jetz W, Daniel Kissling W. 2007. Methods to account for spatial autocorrelation in the

479   analysis of species distributional data: a review. *Ecography* 30:609–628.

480   Drikvandi R, Verbeke G, Molenberghs G. 2017. Diagnosing misspecification of the

481   random-effects distribution in mixed models. *Biometrics* 73:63–71.

482   Dunson DB, Chen Z, Harry J. 2003. A Bayesian approach for joint modeling of cluster size and

483   subunit-specific outcomes. *Biometrics* 59:521–530.

484   Efendi A, Drikvandi R, Verbeke G, Molenberghs G. 2017. A goodness-of-fit test for the random-

485   effects distribution in mixed models. *Statistical methods in medical research* 26:970–983.

486   Fabio LC, Paula GA, de Castro M. 2012. A Poisson mixed model with nonnormal random effect

487   distribution. *Computational Statistics & Data Analysis* 56:1499–1510.

488   Fletcher R, Fortin M-J. 2018. Spatial Dependence and Autocorrelation. In: Fletcher R, Fortin M-

489   J eds. *Spatial Ecology and Conservation Modeling*. New York: Springer, 133–168.

490   Forstmeier W, Schielzeth H. 2011. Cryptic multiple hypotheses testing in linear models:

491   overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology*

492   65:47–55.

493   Forstmeier W, Wagenmakers E, Parker TH. 2017. Detecting and avoiding likely false-positive

494   findings–a practical guide. *Biological Reviews* 92:1941–1968.

495   Gelman A, Hill J. 2006. *Data analysis using regression and multilevel/hierarchical models*.

496   Cambridge university press.

497   Hadfield JD, Nakagawa S. 2010. General quantitative genetic methods for comparative biology:

498   phylogenies, taxonomies and multi-trait models for continuous and categorical characters.

499   *Journal of evolutionary biology* 23:494–508.

500   Hamel S, Yoccoz NG, Gaillard J. 2017. Assessing variation in life-history tactics within a

501   population using mixture regression models: a practical guide for evolutionary ecologists.

502   *Biological Reviews* 92:754–775.

503   Harrison XA. 2014. Using observation-level random effects to model overdispersion in count

504   data in ecology and evolution. *PeerJ* 2:e616. DOI: 10.7717/peerj.616.

505   Harrison XA. 2015. A comparison of observation-level random effect and Beta-Binomial models

506   for modelling overdispersion in Binomial data in ecology & evolution. *PeerJ* 3:e1114. DOI:

507   10.7717/peerj.1114.

508   Harrison XA, Donaldson L, Correa-Cano ME, Evans J, Fisher DN, Goodwin CED, Robinson

509   BS, Hodgson DJ, Inger R. 2018. A brief introduction to mixed effects modelling and multi-

510   model inference in ecology. *PeerJ* 6:e4794. DOI: 10.7717/peerj.4794.

511   Hartig F. 2020. DHARMa: Residual Diagnostics for Hierarchical (Multi-level/Mixed)

512   Regression Models. *R package version 0.3.1*:http://florianhartig.github.io/DHARMa/.

513   Hosken DJ, Buss DL, Hodgson DJ. 2018. Beware the F test (or, how to compare variances).

514   *Animal behaviour* 136:119–126.

515   Housworth EA, Martins EP, Lynch M. 2004. The phylogenetic mixed model. *The American*

516   *Naturalist* 163:84–96.

517   Ives AR, Helmus MR. 2011. Generalized linear mixed models for phylogenetic analyses of

518   community structure. *Ecological Monographs* 81:511–525.

519   Ludecke D. 2019. siPlot: Data visualization for Statistics in Social Science: R package version

520   2.8.1. *R package version 2.8.3*:https://CRAN.R-project.org/package=sjPlot.

521   McCulloch CE, Neuhaus JM. 2011. Misspecifying the shape of a random effects distribution:

522   why getting it wrong may not matter. *Statistical science*:388–402.

523   Miller J, Franklin J, Aspinall R. 2007. Incorporating spatial dependence in predictive vegetation

524   models. *ecological modelling* 202:225–242.

525   Molenberghs G, Verbeke G, Demétrio CGB, Vieira AMC. 2010. A family of generalized linear

526   models for repeated measures with normal and conjugate random effects. *Statistical*

527   *science*:325–347.

528   Molenberghs G, Verbeke G, Iddi S, Demétrio CGB. 2012. A combined beta and normal random-

529   effects model for repeated, overdispersed binary and binomial data. *Journal of Multivariate*

530   *Analysis* 111:94–109.

531   Neuhaus JM, McCulloch CE. 2006. Separating between-and within-cluster covariate effects by

532   using conditional and partitioning methods. *Journal of the Royal Statistical Society: Series*

533   *B (Statistical Methodology)* 68:859–872.

534   Neuhaus JM, McCulloch CE. 2011. Estimation of covariate effects in generalized linear mixed

535   models with informative cluster sizes. *Biometrika* 98:147–162.

536   Neuhaus JM, McCulloch CE, Boylan R. 2013. Estimation of covariate effects in generalized

537   linear mixed models with a misspecified distribution of random intercepts and slopes.

538   *Statistics in medicine* 32:2419–2429.

539   Pinheiro J, Bates D, DebRoy S, Sarkar D. 2014. R Core Team (2014) nlme: linear and nonlinear

540   mixed effects models. R package version 3.1-117. *Available at h ttp://CRAN. R-project.*

541   *org/package= nlme*.

542   Plummer M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs

543   sampling. In: *Proceedings of the 3rd international workshop on distributed statistical*

544   *computing*. Vienna, Austria., 10.

545   R Core Team. 2019. R: A Language and Environment for Statistical Computing. *R Foundation*

546   *for Statistical Computing*. DOI: 10.1007/978-3-540-74686-7.

547   Rousset F, Ferdy J. 2014. Testing environmental and genetic effects in the presence of spatial

548   autocorrelation. *Ecography* 37:781–790.

549   Schielzeth H, Dingemanse NJ, Nakagawa S, Westneat DF, Allegue H, Teplitsky C, Reale D,

550   Dochtermann NA, Garamszegi LZ, Araya-Ajoy YG. 2020. Robustness of linear mixed-

551   effects models to violations of distributional assumptions. *Methods in Ecology and*

552   *Evolution*.

553   Schielzeth H, Forstmeier W. 2008. Conclusions beyond support: overconfident estimates in

554   mixed models. *Behavioral Ecology* 20:416–420.

555   Schielzeth H, Nakagawa S. 2013. Nested by design: model fitting and interpretation in a mixed

556   model era. *Methods in Ecology and Evolution* 4:14–24.

557   Seaman S, Pavlou M, Copas A. 2014a. Review of methods for handling confounding by cluster

558   and informative cluster size in clustered data. *Statistics in medicine* 33:5371–5387.

559   Seaman SR, Pavlou M, Copas AJ. 2014b. Methods for observed-cluster inference when cluster

560   size is informative: A review and clarifications. *Biometrics* 70:449–456.

561   Zhang P, Song PX, Qu A, Greene T. 2008. Efficient Estimation for Patient-Specific Rates of

562   Disease Progression Using Nonnormal Linear Mixed Models. *Biometrics* 64:29–38.

563   Zuur AF, Hilbe JM, Ieno EN. 2013. *A Beginner's Guide to GLM and GLMM with R: A*

564   *Frequentist and Bayesian Perspective for Ecologists*. Highland Statistics Limited.

565   Zuur AF, Ieno EN. 2016. A protocol for conducting and presenting results of regression-type

566   analyses. *Methods in Ecology and Evolution* 7:636–645.

567   Zuur AF, Ieno EN, Elphick CS. 2010. A protocol for data exploration to avoid common

568   statistical problems. *Methods in ecology and evolution* 1:3–14.

569

570

Table 1. The perils of mixed modelling highlighted in this paper together with their potential consequences and solutions to avoid them.

Peril	Example	Consequences	Potential solutions
#1. Anticonservative significance tests at low sample size	Comparing crop yields in split-plot experiments with few replicates	P-value of Wald-like Chi-square test of significance is too low, causing high rates of Type I error	If better replication is not possible, use corrections for small sample size and accept that answers are approximations; alternatively, move to statements of credibility based on Bayesian analyses.
#2. Pseudoreplicated with group-level predictors	Infer the effect of maternal traits on the performance of several offspring per mother.	Risks inflation of confidence/significance/information/credibility if pseudoreplication is not recognised; even if recognised, risks Type I errors if true replication is small.	Have a firm grasp of the design level at which true replication occurs, and of the correct mixed-model specification; if better replication of experimental units is not possible, use corrections for small sample size; alternatively, move to statements of credibility based on Bayesian analyses with correct specification of design hierarchy
#3. Too few levels of a random effect	Fitting sex as a random effect	Model degeneracy Biased estimation of random effect variance Inaccurate estimation of random effect variance Major issues for questions related to random effects – errors for questions related to fixed effects less substantial	Fit the variable as a fixed effect rather than random effect Use a Bayesian model with strong priors for the size of the random effect variance component. Such an approach requires caution and an ability to justify the inclusion of prior wisdom.
#4. Random intercepts when groups vary in their response to a treatment	Testing a relationship between body size and competitive advantage in multiple populations of the same species	Increased Type I error rate when there is variation in slopes between different random effect levels. This is regardless of any correlation between the fixed and random effects.	Use of random slopes model instead of random intercepts only

#5a. Confounding by cluster	Multiple observation of foraging behaviour are made at a succession of sites to test the hypothesis that foraging rates are associated with disturbance levels. However, sites have different mean levels of disturbance. Both disturbance (fixed effect) and site (random effect) are used as explanatory variables	Biased estimates of fixed and random effect parameters	Use within-group mean centring of variables alongside a group-level covariate
#5b. Informative cluster sizes	A model with offspring weight as a response and maternal pathogen load as an explanatory variable with maternal ID as a random effect, if high pathogen loads also cause reduced litter sizes	Possibility of biased estimates for fixed effect parameters if they are correlated with the random effect especially if the model only includes a random intercept	Fit cluster size as a covariate where appropriate (see main text)\n\nJoint modelling of the response variable of interest and cluster size in a multivariate model
#6. Group means are not normally distributed	An unmeasured variable causes differences between sub-populations\n\nSkewed differences in a trait between sub-populations that are unexplained by fixed effects	Fixed effect estimates robust unless a) mis-specification of random effects is extreme or b) fixed effects are correlated with random effects\n\nRandom effect estimates can become less accurate and systematically biased	Use non-Gaussian random effect distributions (challenging, but made available by use of Bayesian models)\n\nFit the variable as a fixed effect instead.\n\nTake extra care to check for other violations of mixed model assumptions

#7. Use of categorical random effects for autocorrelated data	Region is used as a random effect to control for spatial autocorrelation when modelling abundance of a species in response to a range of habitat variables	Poorly fitting model (with inaccurate predictions) and increased Type I error rate(?) unless scale of random effect is correct	Use correlograms to check for covariance in the residuals of the model to ensure that categorical random effect is effective

Figure 1

A mathematical and verbal representation of a simple mixed effects model

__y__ is the response variable, __β₀__ is the global intercept (the expectation of y when all fixed effects are zero, and for members of an average group in the random effect), __xᵢ__ is the measured value of the iᵗʰ fixed explanatory variable, __βᵢ__ is the additive expected change caused by the value of each of the fixed explanatory variables, __γ__ is a draw from the distribution of category means for a Normally distributed random effect (with mean of zero and variance equal to the random effect variance), and __ε__ is a draw from the Normal distribution of residuals (with mean of zero and variance equal to the residual variance).

$$\mathbf{y} = \beta_0 + \sum \beta_i \mathbf{x}_i + \gamma + \varepsilon$$

| Response variable | Global intercept | Fixed effect parameters | Fixed effect variables | Random effect variance | Residual variance |

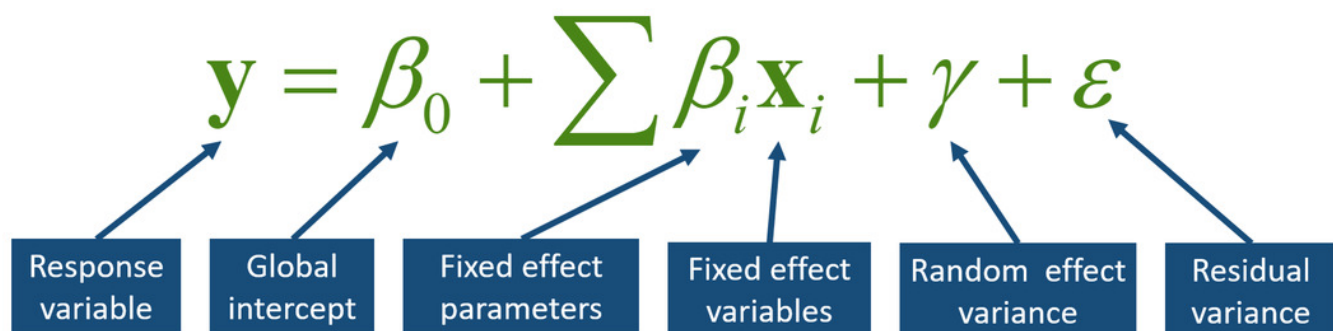



Figure 2

Figure 2. Simple examples of linear mixed models.

In (a & b), offspring from six families (different colours) are allocated to one of two treatments (a categorical fixed effect). In (a), a **random intercepts model**, the treatment (fixed) effect is the slope of the line connecting each family cluster. The differences in the elevation of the line for each family shows the random intercept, with each family's intercept being drawn from a Normal distribution. The random effect absorbs variation in intercepts among the families, helping to reveal the independent, global influence of treatment (black dashed line). In (b), a **random slopes model** reveals a global slope having absorbed variation among families in both intercept and effect of treatment. In (c & d), the response variable is instead regressed against a covariate (continuous fixed effect) measured, for each offspring, between zero and one. The random intercepts model (c) reveals the global slope (black dashed line) of the relationship between response and explanatory variables, having absorbed variation in the intercept among families. The random slopes model in (d) reveals a global slope having absorbed variation among families in both intercept and slope. The random-intercepts model infers parallel changes caused by the fixed effect among groups, while random-slopes allows the model to infer variation in slopes as well as intercepts.

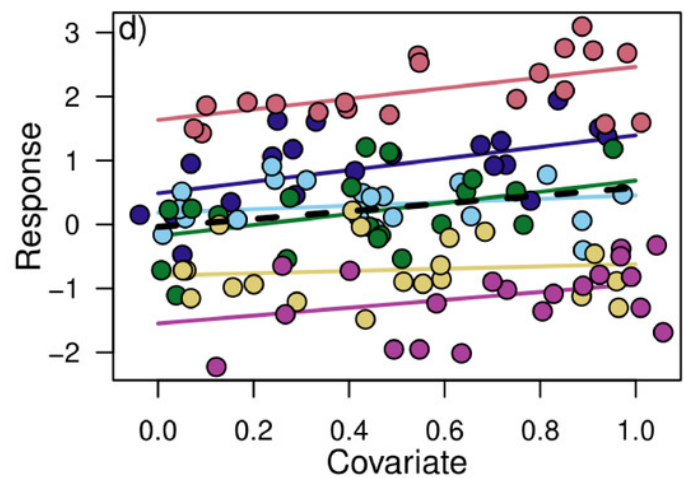
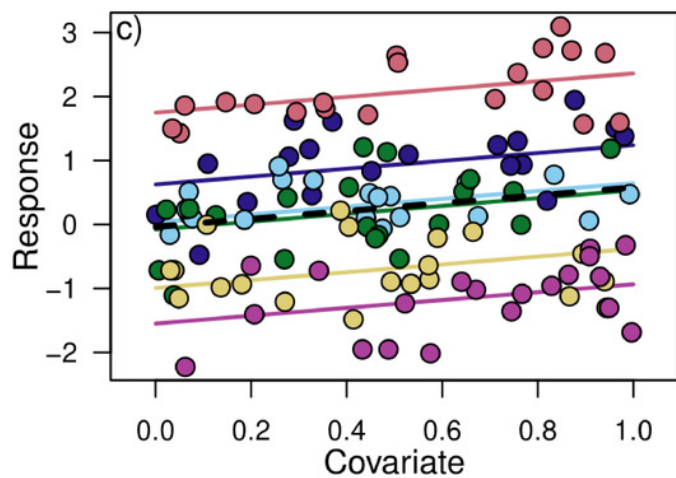
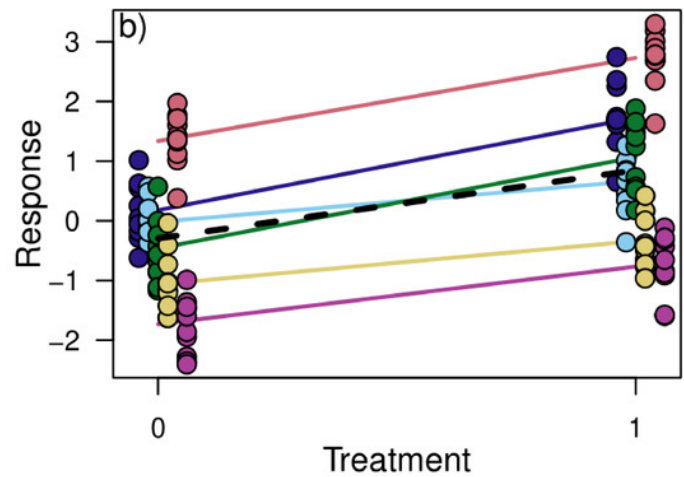
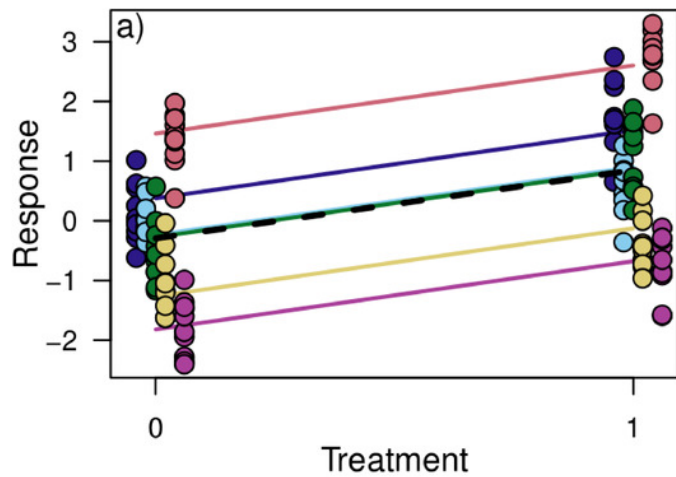



Figure 3

Figure 3. The false positive rate for group-level variables in linear mixed-effect models with small number of random effects using common forms of statistical inference.

For each number of random effect levels (x axis) we simulated 1000 datasets containing no effect of an experimental treatment applied at the group-level and recorded the number of times that a statistically significant effect of treatment was recorded using a) model estimates and 95% confidence intervals from the full model; b) a Likelihood ratio test between the model with an effect of treatment and a null model; and c) ΔAIC > 2 between the two models. All code for this example is provided in the Supplementary Material.

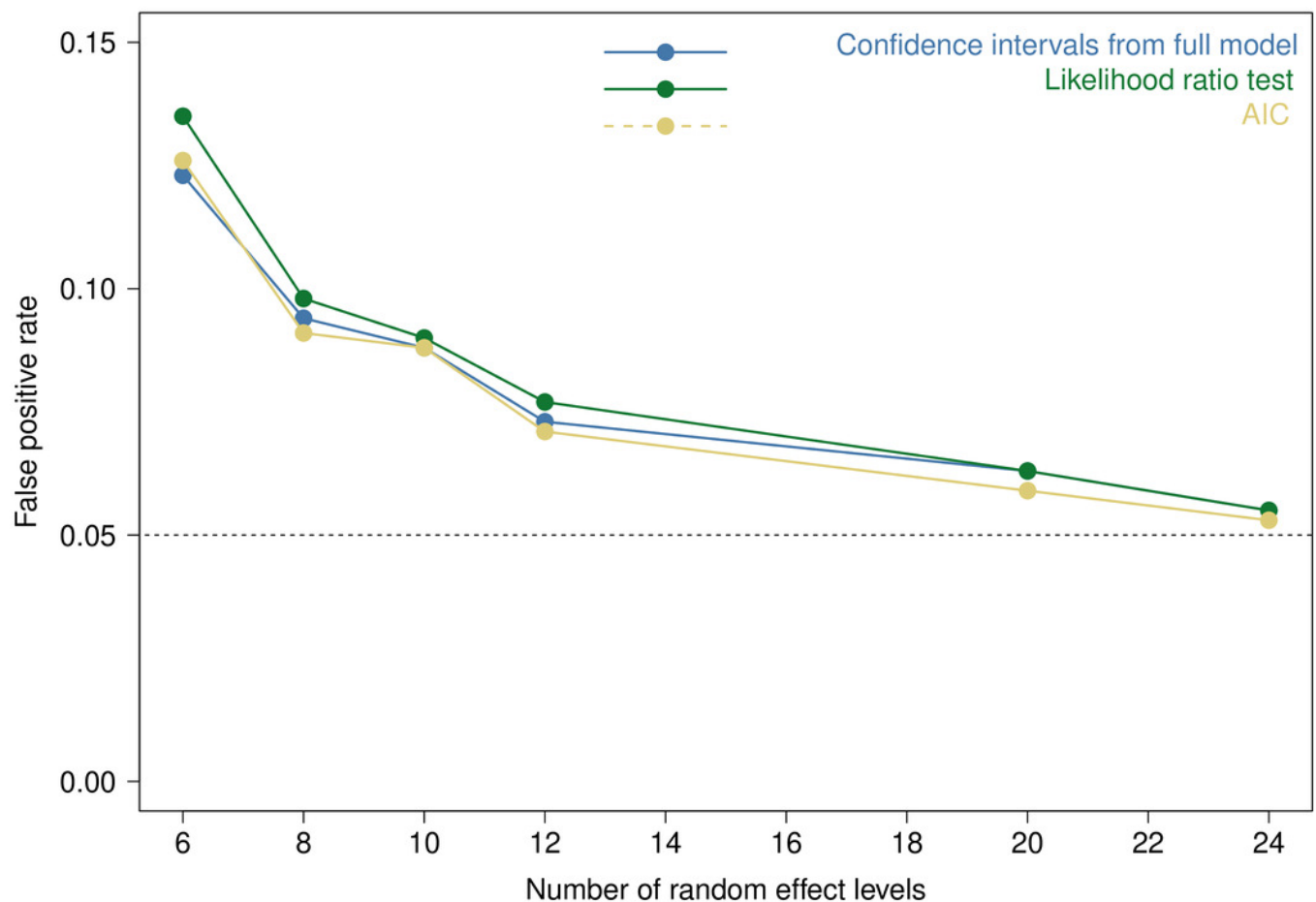



Figure 4

Figure 4.[i] The effect of the number of levels of a random effect on a) the random effect variance (displayed here as a standard deviation) in a random intercepts model, and b) the standard error around the effect of a fixed effect covariate in the

The red line in a) indicates the true value of the between-group standard deviation. The true effect size for the covariate effect is 1 and the residual deviance had a standard deviation of 3 (see Supplementary Material).

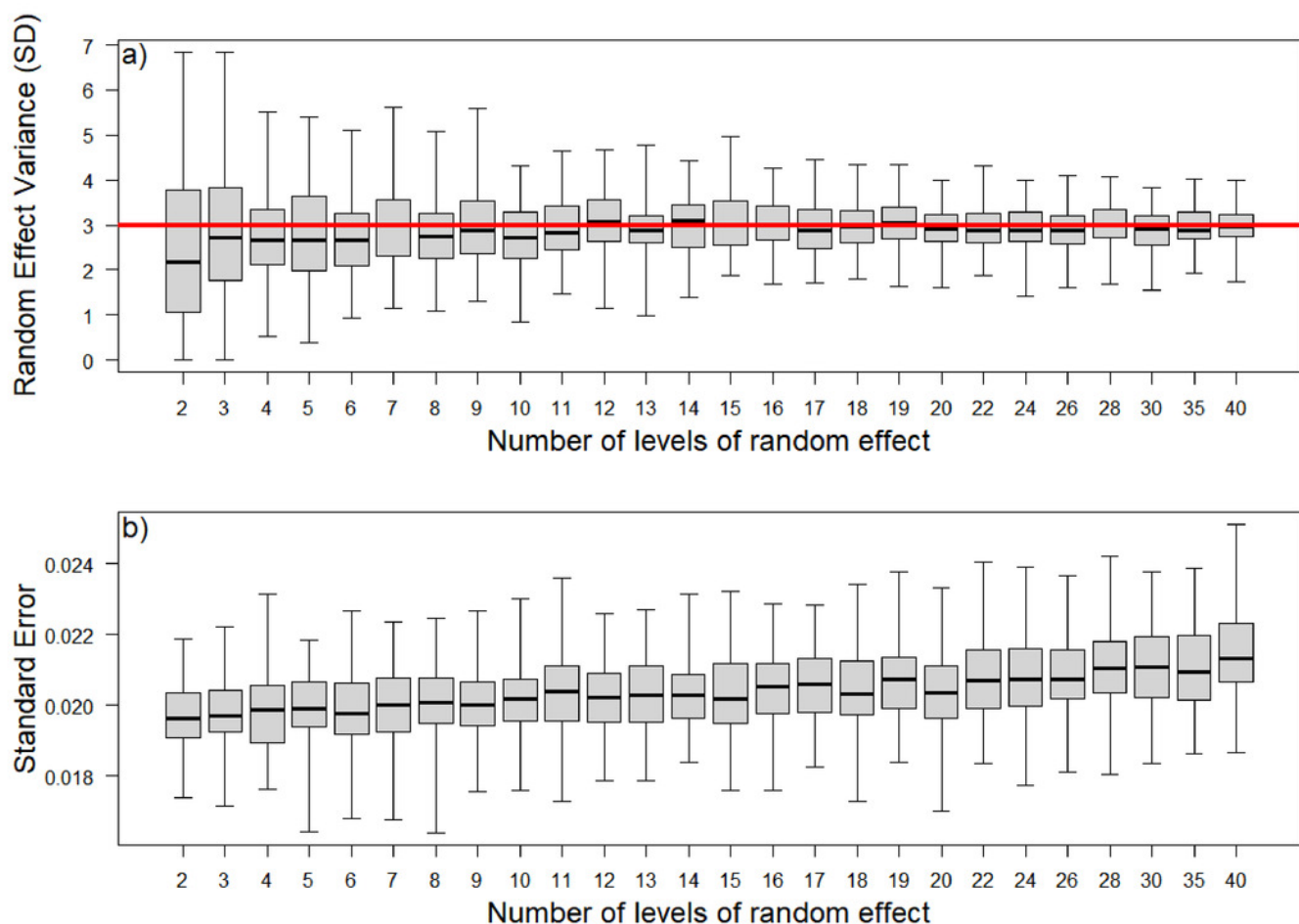
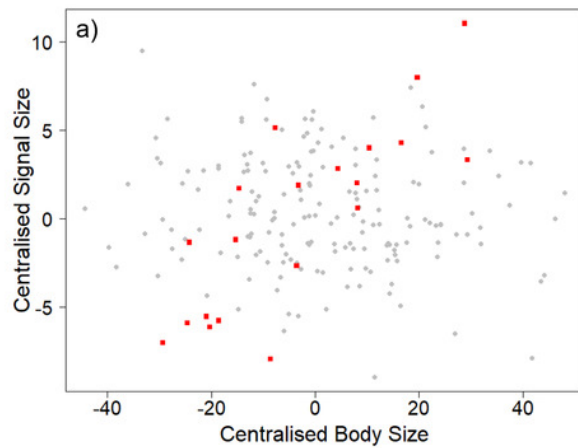



Figure 5

Figure 5. Random slopes models avoid the exaggerated influence of anomalous groups.

Simulated datasets contain 10 groups, only one of which harboured a relationship between body size and status signal size [a)]. Grey points represent data from the 9 typical groups, and red points data from the anomalous group. For a variety of anomalous-group slopes (0, 0.1, 0.2) we measured the proportion of models, out of 100 simulations, that inferred a significant global relationship between body size and signal size [b)]. The difference in type I error rate was not caused by differences in estimate for the overall relationship between body size and signal size [c)], but instead an underestimate of the error around this estimate in the random intercept only model [d)]. In these plots the polygons are density plots of inferences of global slope from the random intercepts models (blue) and the random slopes models (yellow). All R code for this example is provided in the Supplementary material.

b)

Body size effect in exception group	Type of Model	Type I error rate
0	Random intercept only	1%
0.1	Random intercept only	17%
0.2	Random intercept only	32%
0	Random slopes	1%
0.1	Random slopes	6%
0.2	Random slopes	3%

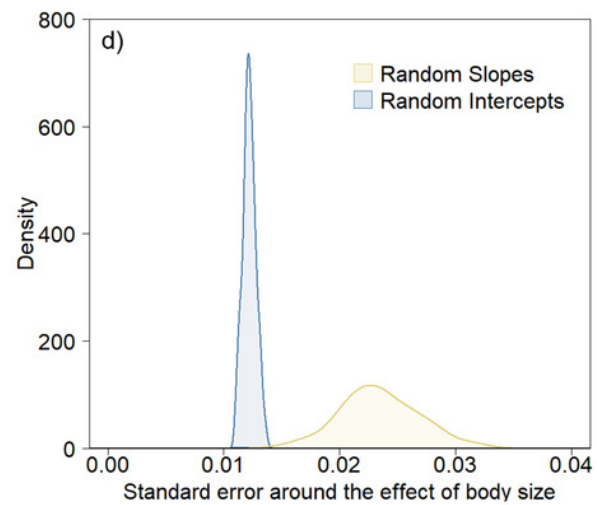
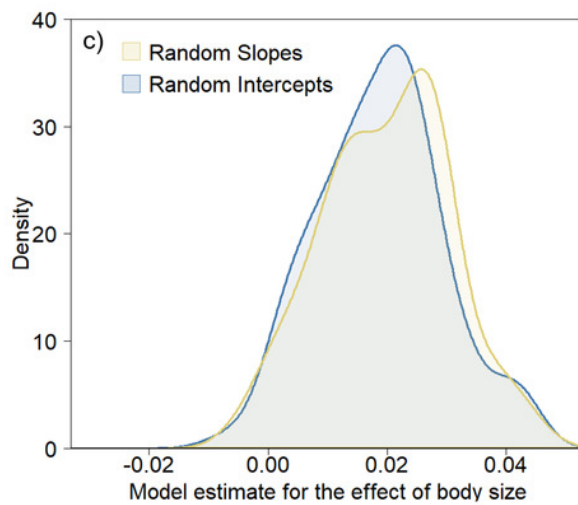



Figure 6

Figure 6. An illustration of the appropriateness of categorical effects to control for autocorrelated data.

In this example dataset, a response variable measuring height varies according to temperature and there is spatial autocorrelation in the residual error at either a fine spatial scale [a)] or a broader spatial scale [d)]. Sampled individuals (small, semi-transparent black points) are associated with particular sites (white diamonds), with 10 individuals per site. Each site is associated with a particular region (centre marked with light grey square), with four sites sampled in each region [all depicted in a) and d)]. We fitted three models, each with height as the response variable and temperature as the explanatory variable: a generalised least squares model with a spatial correlation structure (for comparison), an LMM with site as a random effect (comparison in blue) and an LMM with region as a random effect (comparison in yellow). b) and e) show the difference between model estimates for the effect of temperature between the LMMs and the model with spatial autocorrelation structure for the case with fine scale spatial autocorrelation and broader scale autocorrelation respectively. c) and f) show the same comparison for standard error around those model estimates. All R code for this example is provided in the Supplementary Material.