

Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse

Wolfgang Forstmeier · Holger Schielzeth

Received: 12 May 2010 / Revised: 23 July 2010 / Accepted: 29 July 2010 / Published online: 19 August 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract Fitting generalised linear models (GLMs) with more than one predictor has become the standard method of analysis in evolutionary and behavioural research. Often, GLMs are used for exploratory data analysis, where one starts with a complex full model including interaction terms and then simplifies by removing non-significant terms. While this approach can be useful, it is problematic if significant effects are interpreted as if they arose from a single a priori hypothesis test. This is because model selection involves cryptic multiple hypothesis testing, a fact that has only rarely been acknowledged or quantified. We show that the probability of finding at least one ‘significant’ effect is high, even if all null hypotheses are true (e.g. 40% when starting with four predictors and their two-way interactions). This probability is close to theoretical expectations when the sample size (N) is large relative to the number of predictors including interactions (k). In contrast, type I error rates strongly exceed even those expectations

when model simplification is applied to models that are over-fitted before simplification (low N/k ratio). The increase in false-positive results arises primarily from an overestimation of effect sizes among significant predictors, leading to upward-biased effect sizes that often cannot be reproduced in follow-up studies (‘the winner's curse’). Despite having their own problems, full model tests and P value adjustments can be used as a guide to how frequently type I errors arise by sampling variation alone. We favour the presentation of full models, since they best reflect the range of predictors investigated and ensure a balanced representation also of non-significant results.

Keywords Bonferroni correction · Effect size estimation · Generalised linear models · Model selection · Multiple regression · Multiple testing · Parameter estimation · Publication bias

Communicated by L. Garamszegi

This contribution is part of the Special Issue “Model selection, multimodel inference and information-theoretic approaches in behavioural ecology” (see Garamszegi 2010).

W. Forstmeier (✉) · H. Schielzeth
Max Planck Institute for Ornithology,
Eberhard-Gwinner-Str.,
82319 Seewiesen, Germany
e-mail: forstmeier@orn.mpg.de

Present Address:
H. Schielzeth
Department of Evolutionary Biology,
Evolutionary Biology Centre, Uppsala University,
Norbyvägen 18D,
SE-752 36 Uppsala, Sweden
e-mail: holger.schielzeth@ebc.uu.se

Introduction

Generalised linear models (GLMs) are widely used for exploratory data analysis. In observational studies, model selection procedures are commonly applied in order to find a parsimonious combination of predictors to explain some phenomenon of interest (Miller 1984; Quinn and Keough 2002). Such exploratory searches for the best-fitting model are also used in experimental studies, because an experimental treatment might show unforeseen interactions with other factors or covariates. Automated procedures of model simplification, however, often make us forget that this constitutes a case of multiple hypotheses testing that will lead to high rates of type I errors (Zhang 1992; Whittingham et al. 2006; Mundry and Nunn 2009) as well as biased effect size estimates (Burnham and Anderson 2002; Lukacs et al. 2010).

The issue of multiple hypotheses testing is most problematic in cases where data on a large number of explanatory variables are collected to explain some phenomenon of interest, but a priori information on which predictors (and essentially, if any) influence the response is not available. For instance, a bird's attractiveness in a choice test may be unrelated to each of 20 song characteristics we measured simply because variation in song might have evolved to signal individual identity but not individual quality (Tibbetts and Dale 2007). In this case, the commonly used significance threshold of $\alpha=0.05$ would imply that, irrespective of sample size, we would in the majority of cases end up with a reduced model that includes 'significant' effects. Note that the obviously high number of 20 predictors is quickly reached when interactions are included in the full model from which one starts with simplification. Similar, though less extreme, situations might arise even in experimental studies, when the experimental treatment effect is tested in combination with a number of different covariates and/or in interaction with them. So whenever we recognise that there is a problem of multiple hypotheses testing, we need a reference against which to compare our findings. We need to know how often complex models will lead to 'significant' minimal models by chance alone, i.e. when all null hypotheses are actually true. In the present paper, we aim to establish how this baseline rate of type I errors depends on the number of predictors including interactions (k) in the initial full model and on sample size (N). As we show in a literature survey below, researchers tend to focus their attention to the outcome of the selection process, namely the minimal model, while the details of full model fitting, like the number of interactions examined, often do not even get reported in publications. Hence, there clearly seems to be a lack of awareness of this problem in the empirical research literature.

The widespread use of model simplification and the presentation of minimal rather than full models (see literature survey below) also imply that researchers tend to selectively focus their attention on significant effects, while non-significant effects are often discarded, i.e. fixed to zero during model simplification. As a consequence of this selection process, the obtained parameter estimates for the 'significant' effects will tend to be biased upwards (away from the null hypothesis). This is true for predictors that are truly of zero effect, but it also applies to predictors of small effect (Lukacs et al. 2010). The subsequent difficulty to reproduce initially significant findings that arose from multiple testing has been termed the 'winner's curse' for whole large-scale association studies, where large numbers of tests are conducted (Zöllner and Pritchard 2007; Ioannidis et al. 2009). A similar trend that affect sizes tend to decay in replication studies has also been observed in

studies in the field of behavioural ecology (Jennions and Møller 2002).

The point of overestimated effect sizes among 'significant' parameter estimates has been made by Burnham and Anderson (2002) as an argument to favour information-theoretic (IT) approaches over null-hypothesis testing. IT approaches allow a ranking of models according to their support by the data. Even though this permits multi-model inference, confidence sets of IT-selected models are also defined based on thresholds. We argue that any threshold criterion leading to the selective interpretation of some relationships from a pool of investigated ones will face the same problem (see also Ioannidis 2008). Burnham and Anderson (2002) also emphasise that standard errors in linear models are conditional on the model structure and criticise stepwise selection procedures for their failure to incorporate model structure uncertainty into estimates of precision, i.e. the standard errors (see also Chatfield 1995). So one can argue that the very use of model simplification illustrates that there is uncertainty about which set of predictors will be influential, and this uncertainty is not reflected any more by the standard errors of the minimal model.

Throughout the paper, we define type I errors as finding a significant predictor, when actually all null hypotheses of no effect are true. Although we focus on type I errors, we do not want to promote binary thinking in a significant/non-significant dichotomy (Stephens et al. 2007; Garamszegi et al. 2009). However, since the distinction in significant and non-significant effects is still common, we want to emphasise that threshold-based interpretation has consequences for the distribution of effect size estimates. While, on average, individual effect size estimates are correct, the selective interpretation of only some (i.e. the most significant findings) will lead to an upward bias in effect size estimates. We argue that if full models were reported more frequently, this would also reduce publication bias, since non-significant parameter estimates would get presented as well (Anderson et al. 2000).

Literature survey

Fitting GLMs with more than one predictor is common practice in the study of ecology, evolution and behaviour. A survey of the September 2007 issues of *The American Naturalist*, *Animal Behaviour*, *Ecology*, *Evolution* and *Proceedings of the Royal Society Series B* showed that 28 out of 50 empirical studies (56%) use GLMs with two or more explanatory variables. These models were often used to test multiple hypotheses simultaneously, even though it seemed impossible to draw a clear line between 'hypothesis testing' and 'controlling for confounding variables'. Six out

of 28 studies (21%) fitted quite large models with six to 17 explanatory variables, and three out of these six studies presented models where the sample size was less than three times the number of explanatory variables ($N/k < 3$). Additional to these main effects, interactions were fitted in 80% of the studies using GLMs. Multiple hypotheses testing was not limited to observational studies; experimental studies often considered various treatment-by-covariate interactions besides the treatment main effect (six out of 11 experimental studies in our survey).

A common approach was to use model simplification based on the significance of individual predictors (backward elimination starting with the least significant predictor; Derksen and Keselman 1992; Whittingham et al. 2006). This was done in all models ($N=16$) with more than five predictors (including interactions) in our survey, except for one study that used forward selection of predictors. Problematically, model simplification tends to disguise the multiple-testing problem. It was often hard to reconstruct the initial size of the full model before simplification (e.g. whether all or only some two- or three-way interactions were included) and hence the number of parameters that were actually estimated.

Simulation

To examine how the frequency of type I errors increases with the number of predictors, we generated random data where a single dependent variable was to be explained by several two-level factors (and their interactions). Data for the dependent variable was drawn from a standard normal distribution ($\mu=0$ and $\sigma=1$). The explanatory variables were simulated such that they were balanced and uncorrelated with each other to the extent maximally allowed by a given sample size. Predictor levels were coded -0.5 and 0.5 , respectively, to ensure that main effects were interpretable even in the presence of interactions (Aiken and West 1991; Schielzeth 2010). We generated the design matrix first and then sampled response data completely independently of the design matrix. Independent sampling ensured that there was no correlation between predictors and response in the population, i.e. all null hypotheses of no effect were true and all effects were zero by definition. We then fitted linear models to the data and screened the resulting analysis of variance (ANOVA) tables for significant predictors at $\alpha=0.05$.

We generated datasets with two different sample sizes ($N=50$ and $N=200$) and one to six variables, and ran the simulations once with only main effects and once including all two-way interactions besides the main effects. The sample size of 200 was chosen so that there were sufficient data for even the largest model to be fitted (with up to $k=21$

predictors, namely six factors and their 15 two-way interactions). In contrast, with $N=50$ we wanted to examine how the type I error rate changes when the initial full model is over-fitted. Different statistics text books vary widely in their advice regarding desirable N/k ratios (e.g. Crawley 2007 recommends $N > 3k$, while Field 2005 suggests $N \geq (50+8k)$), and partly for this lack of consensus, researchers may be unaware of the dangers involved in over-fitting the initial full model (see also Chatfield 1995). Hence, with our examples, we do not want to imply that fitting large models to low sample sizes is appropriate, but rather we wanted to explore the effect of over-fitting on type I error rates.

Backward model simplification was automated using the convenient and widely used step function in R 2.8.1 (Venables and Ripley 2002; Crawley 2007). This function removes predictors one at a time and chooses models based on their Akaike information criterion (AIC) values, but retains main effects when they are involved in interactions. In our simulation, where every predictor takes 1df, this AIC-based simplification is equivalent to always removing the least significant term if removal does not impair model fit. We are not primarily interested in comparing different model selection algorithms, which has been done in previous publications (e.g. Mundry and Nunn 2009). Using other model selection procedures might slightly alter the precise values, but would not change the general picture as long as some fixed criterion is used to decide whether or not to include a predictor. We implemented our simulation in R 2.8.0 (R Development Core Team 2008) and ran 5,000 simulations for all factor combinations.

In a first step, we applied no model simplification, but rather searched the full model t -table for whether it contained a significant effect ($P < 0.05$). Not very surprisingly, the probability of finding at least one significant predictor in a GLM (i.e. the rate of type I errors) depends primarily on the number of parameters estimated in the model (Fig. 1a). With larger sample sizes ($N=200$), the realised type I error rate is very close to what would be expected from the number of tests applied $\alpha' = 1 - (1 - \alpha)^k$ (grey lines in Fig. 1). Hence, there is hardly any difference compared to conducting k independent univariate test. With smaller sample sizes ($N=50$) the increase in type I errors is slightly less pronounced than expected from the above formula. This is because there are fewer degrees of freedom for the error when testing individual predictors in a multiple-predictor model rather than individually. Quite intuitively, the loss of degrees of freedom for the error becomes relatively more important the lower the sample sizes.

In a second step, we applied model simplification and screened the t -table of the minimal model (sometimes called the ‘best model’) for whether it contained a significant effect ($p < 0.05$). In this case, the increase of

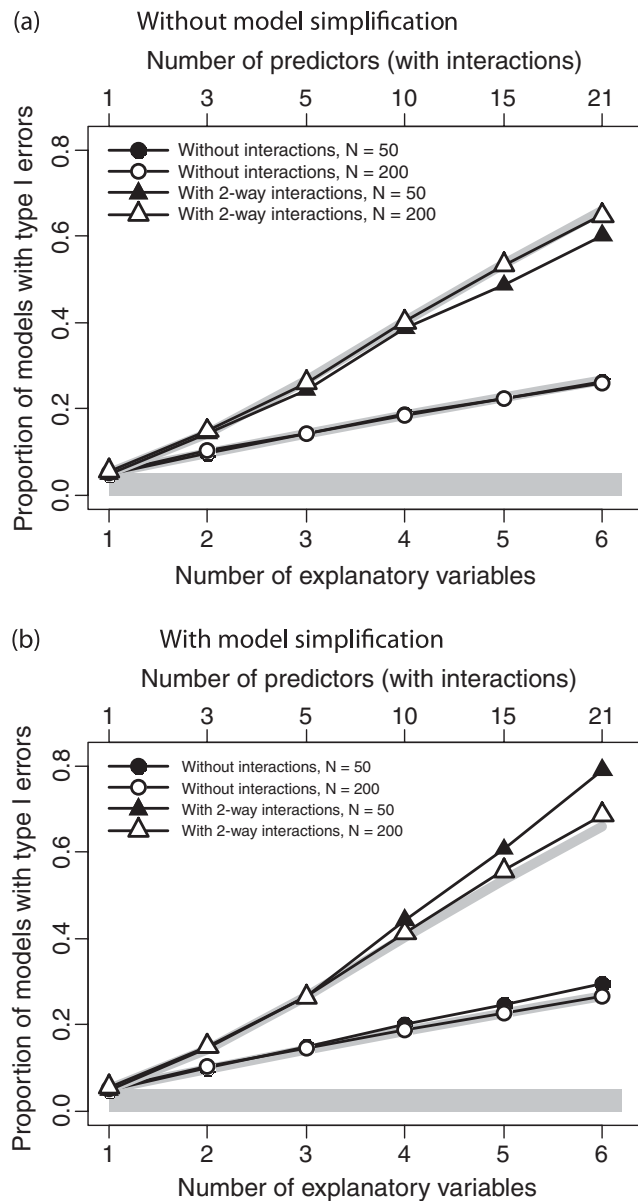


Fig. 1 Simulated proportion of models containing at least one significant predictor, despite all null hypotheses being true (type I error rate). **a** Results without model simplification, **b** results with model simplification. The grey zone shows the zone of $\alpha=0.05$. Grey lines show the expected number of false-positives from a set of independent tests: $\alpha'=1-(1-\alpha)^k$, where k is the number of tests applied

type I errors with the number of predictors turns out to be slightly steeper than expected from $\alpha'=1-(1-\alpha)^k$ (Fig. 1b). Thus, model simplification slightly increases the type I error rate, but the difference is small for large sample sizes ($N=200$ in our simulation). Hence, for large datasets, type I error rates will not depend much on whether model simplification is carried out or whether full model tables are screened for the most significant effects. This also implies that if one applies model simplification to increase

the confidence in ‘important’ estimates, there actually is not much to be gained from model simplification for large datasets. Although simplified models tend to have smaller standard errors, the effect is marginal for large N/k ratios. With small sample sizes, in contrast, the type I error rate after model simplification is noticeably higher than expected. In Fig. 1b this effect becomes apparent with four predictors (N/k ratio of 12.5), but it is most pronounced for very low N/k .

In order to illustrate the source of type I error inflation, we sampled point estimates and standard errors from the full model for significant and non-significant predictors. The set of ‘significant’ effects is what is likely to be presented and discussed in a publication. Standard errors are only marginally smaller for significant predictors compared to those for non-significant predictors (Fig. 2a), while point estimates were substantially larger (on average more than three times as large) for significant predictors (Fig. 2b). This shows the cause of type I error inflation: It is almost entirely due to an overestimation of effect sizes among significant effects rather than differences in the size of the standard errors.

Type I, type II errors and effect sizes

High rates of type I errors are typically countered by making the threshold criteria more stringent (e.g. Bonferroni correction, false-discovery rate) or by applying full model tests. However, such corrections have repeatedly been criticised for promoting type II errors, for being overly conservative or for representing a biologically insensible global null hypothesis (e.g. Perneger 1998; Anderson et al. 2000; Nakagawa 2004). Hence, whether or not such corrections are sensible to apply clearly depends on the type of research question. Sometimes we are confronted with situations where the dependent variable will most likely be affected by each of the predictors we measured: For example, singing activity in a bird will likely depend on sex, age, date, time, temperature, wind, rain and food availability. In such situations, very stringent model simplification criteria would only promote conducting type II errors (Perneger 1998; Anderson et al. 2000; Nakagawa 2004). In contrast, there will be other situations, like the example on song traits versus attractiveness introduced above, where it actually seems quite possible that the dependent variable is not affected by any of the predictors measured (see also Mundry 2010). So how can we safely reject the global null hypothesis in those latter cases, where all null hypotheses may be true even for a single reason (e.g. song does not signal individual quality)? Below, we want to give a guideline to judge, whether an effect of a predictor is actually stronger than expected by chance given the number of predictors that were tested.

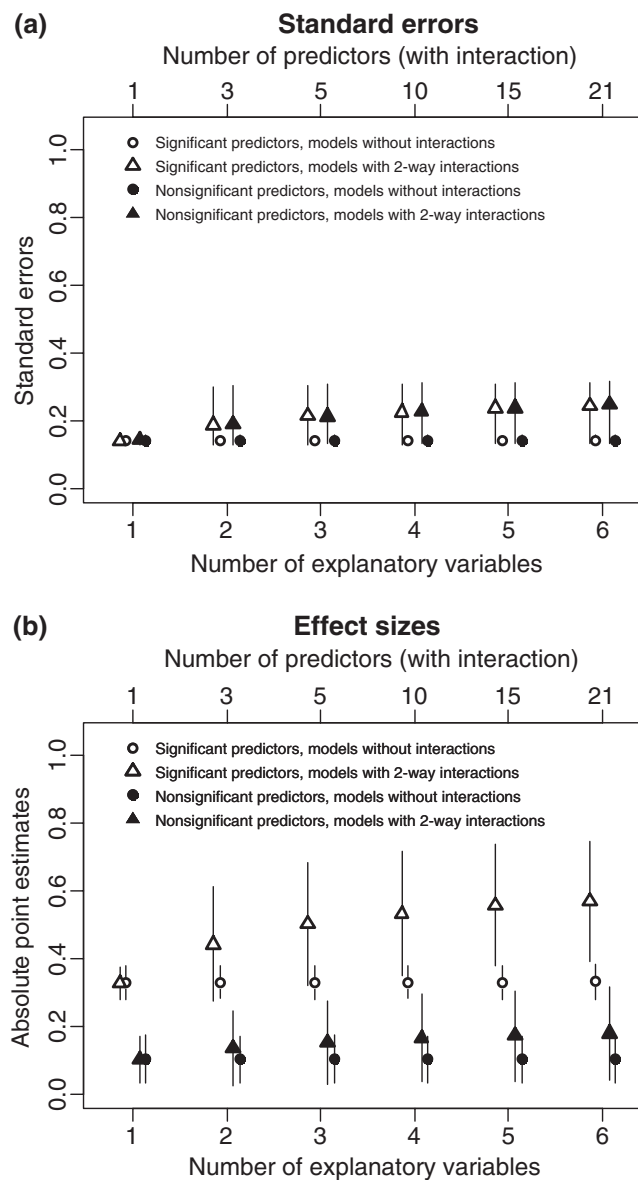


Fig. 2 (a) Standard errors and (b) point estimates for significant (white symbols) and non-significant parameters (black symbols). One significant (when available) and one non-significant predictor were sampled from each full model. Means and 95% confidence intervals are shown from 5,000 simulation runs. Sample sizes was $N=200$ for each dataset

Note that independent of any type I versus type II error discussion, there is a remaining issue with multiple testing and effect size estimation. Even if effects are indeed true and hence findings do not constitute type I errors, effect sizes are often overestimated if their true effects are small, but the sample size is too low to reliably detect small effects with high confidence (Zöllner and Pritchard 2007; Ioannidis 2008). This is particularly problematic for typical studies in behavioural ecology, where effect sizes are often small and sample sizes low (Jennions and Møller 2003). This is another aspect of the winner's curse that even if the

claimed effects are true, effect size estimates based on selection from a larger set of candidate effects might give misleadingly high effect sizes (Göring et al. 2001; Ioannidis 2008). Replication with independent datasets would evidently be a desirable strategy (Dochtermann and Jenkins 2010), but, unfortunately, independent replication is relatively rare in behavioural ecology (Kelly 2006). Intriguingly, if meta-analyses are conducted on related studies they show exactly the expected trend of diminishing effect sizes (Jennions and Møller 2002).

Examination of the full model

To safeguard against type I errors, the full model can be tested against a null model (with an intercept only) using a likelihood ratio test. This approach is particularly advisable if one considers the whole model as one complex biological hypothesis (e.g. song characteristics affect a bird's attractiveness in a complex way that is not known before data inspection). This procedure effectively keeps the type I error rate at the nominal level (Fig. 3a). Applying a full model test before looking at individual predictors is somewhat analogous to an ANOVA test for multi-level factors before applying post hoc tests to pinpoint the differences. However, this approach clearly introduces a dichotomous decision, since it is based on a significance threshold. Furthermore, full model tests do not clarify, which effects are 'true' and which are type I errors, since the full model will not only become significant with one strong, but also with several weak effects. Therefore, the full model test can only give an indication for whether or not the observed sum of effects is likely to arise by sampling variation alone.

Once the significance of the full model has been established, one may want to proceed with model simplification to narrow down the significant predictors. Many statisticians advice using model simplification, because parameter estimates from the full model may be unreliable (large standard errors) due to the many parameters that are estimated simultaneously. However, one should keep in mind that the standard errors (as well as point estimates) are conditional on the model structure (Burnham and Anderson 2002). Hence, standard errors of estimates will become smaller after model reduction (although with large N/k ratios the difference is only marginal), but these may actually be too small since they do not yet account for the uncertainty about the model structure. It is for this reason that we advocate presenting the estimates and standard errors from the full model rather than from the reduced model. In any case, testing the reduced model against a null model (as opposed to testing the full model against a null model) will be misleading, since this applies the test after

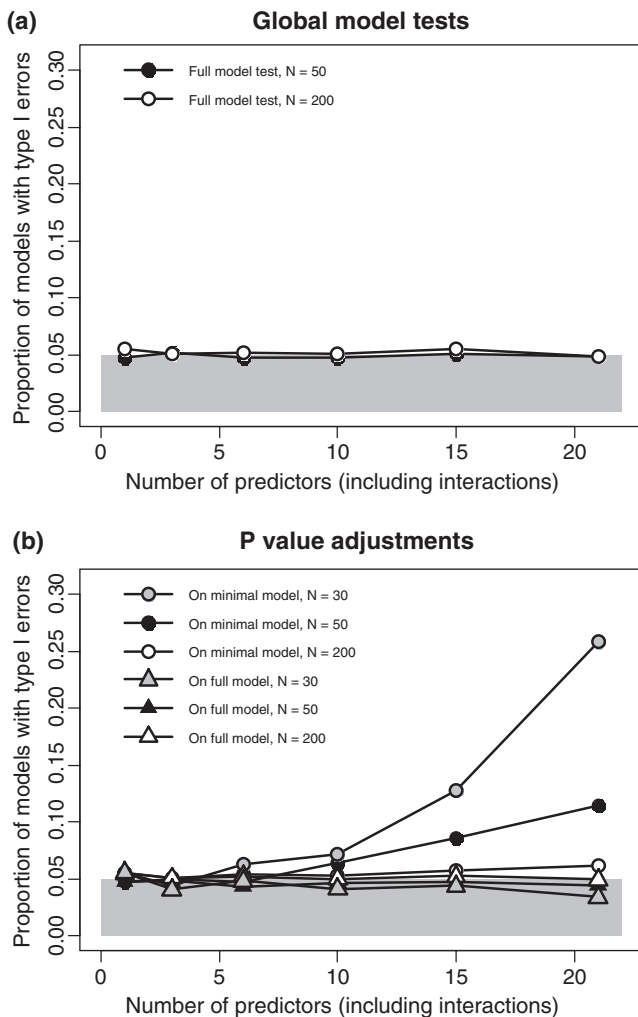


Fig. 3 Simulated proportion of models containing at least one significant predictor, despite all null hypotheses being true (type I error rate) after (a) global model tests, (b) *P* value adjustments. The x-axis shows the number of predictors in the full model (including two-way interactions), hence representing one to six explanatory variables and their two-way interactions. The grey zone shows the zone of $\alpha=0.05$

the significant predictors have been singled out. Even though this is probably the most widespread way of presenting results, it is uninformative as to whether the global null hypothesis can be rejected.

In some cases it is clear a priori that one of the factors is highly influential (e.g. age might be known to influence attractiveness, while the effect of song traits is questionable). We might want to control for this factor rather than testing it. In this case, full model tests will always become significant when tested against a null model. Therefore, the full model should be tested against a model containing only the predictor with known influence. When a significant result shows that there is more than the known effect, further examination of the other predictors can be applied (Blanchet et al. 2008).

Our literature survey shows that only two out of 28 studies presented full model tests, and in many of the other studies it was not even clear whether the full model included all or only some of the interactions. Hence, we would like to promote paying greater attention to full models. Presenting a full model has several advantages: (1) The full model test shows whether the global null hypothesis can be safely rejected. (2) The full model often is the best representation of the range of hypotheses initially considered by the investigator. Only rarely one measures predictors under the assumption that it will not affect the dependent variable. (3) Full model tables present all parameter estimates and hence many tests of hypotheses, which reduces publication bias and helps the meta-analytic study of small effects.

P value adjustment

The second approach is to control table-wide type I error rates by using sequential Bonferroni correction (Holm 1979; Hochberg 1988; Rice 1989; Wright 1992) or false-discovery rate (FDR) control (Benjamini and Hochberg 1995; Storey and Tibshirani 2003). The latter tends to be less conservative and therefore gives a better balance between type I and type II errors (Verhoeven et al. 2005). However, the use of Bonferroni versus FDR does not make much of a difference for our simulation, since we were interested in the proportion of models that contained at least one type I error and in both cases, the strictest test is against α/k . Therefore, and because of its widespread use, we will focus on Bonferroni adjustments in our analyses.

In a model with initially six factors and their 15 two-way interactions (hence 21 predictors) the smallest *P* value in the table should be below $0.05/21=0.0024$. When this is applied to the full model (before model reduction) the table-wide type I error rate is effectively 5% (Fig. 3b). In sharp contrast, after model reduction, the same strict α -level of 0.0024, in our example, will still produce approximately 12% type I errors if sample sizes and N/k ratios are low (here $N=50$, $N/k=2.4$; Fig. 3b). To better illustrate the shape of this relationship between N and k and the type I error rate, we include an extra simulation for $N=30$ in Fig. 3b. The problem accelerates rapidly with the degree of over-fitting: minimal models derived from six explanatory variables and their 15 interactions fitted to $N=30$ data points will yield as extreme *P* values as would only be found when conducting 286 independent tests ($P < 0.05/286=0.00017$ in 5% of the cases). Although this example is certainly extreme ($N/k=1.43$), it illustrates that model simplification on complex models with insufficient sample size is highly misleading. Our simulations show that the disproportionate increase in type I errors is not a

simple function of the N/k ratio, but, as a rule of thumb, the effect becomes strong if $N < 3k$.

The advantage of Bonferroni-type corrections is that it can be applied by the reader if the number of predictors in the full model is known and if exact P values are given. However, in our literature review it was often difficult to reconstruct the size of the full model that is required for an independent evaluation (see also Mundry 2010). Moreover, eight studies (29%) did not report exact P values in their minimal models that would have allowed for later P value adjustment. Finally, with model simplification and insufficient sample sizes ($N < 3k$), type I errors are still inflated even after Bonferroni correction (Fig. 3b).

Conservatism and publication bias

As we show above, type I errors can be reduced by either full-model tests or by Bonferroni correction, but this does not solve the problem of biased effect sizes due to selective interpretation. Effect sizes of significant effects will remain biased, and they will become even more biased when we focus all our attention on those estimates that ‘survive’ even a Bonferroni correction. So will the present paper that essentially calls for more statistical conservatism lead to even greater publication bias? For the following reason, we do not think so. There are three types of studies competing for publication space: (1) studies that claim positive evidence and that withstand statistical rigour, (2) studies that claim positive evidence but not quite convincingly so, and (3) comprehensive descriptive studies that make no claims of positive evidence. Publication bias arises when editors and referees prefer the first two kinds of studies over the third. Now, with increasing statistical conservatism, studies of the second type might be less favoured, giving more publication space to comprehensive exploratory and descriptive studies. Moreover, with our call for presenting full model tables, non-significant findings might also get published more often alongside with significant ones.

Correlated predictors

In this paper, we so far have assumed that the different explanatory variables are not correlated with each other. In real datasets, however, predictors are often correlated and this will lead to unstable parameter estimates. Since parameter estimates are conditional on the model (Burnham and Anderson 2002; Lukacs et al. 2010), estimates for a particular predictor might change signs depending on whether or not a correlated predictor is included. Some of those situations are avoidable (Schielzeth 2010), but if they are not, these predictors need special consideration. We

recommend to remove highly correlated predictors wherever possible (and to include only one of them or a major axis combination of both). If this is not possible, it is worth assessing the change in estimates when either one of them or both of them are included in a model. Such exploration can be reported in a publication and will clarify the dependence of the estimates on a specific model structure. However, even if predictors are correlated, full model estimates are often appropriate (Freckleton 2010).

Information-theoretic approaches and model averaging

We have outlined so far, why we think the evaluation of full models is often appropriate. This is especially true for the alternative of drawing inferences from the full model versus a ‘best model’ after model simplification. Similar criticism against best model inferences has been brought forward by proponents of the IT approach, which focuses on multi-model inferences instead of relying on a single best model (Burnham and Anderson 2002; Lukacs et al. 2010). IT approaches allow the comparison of non-nested models, which is, of course, a big advantage. Most model selection procedures in the field of behavioural ecology, however, focus on models that are nested within a global model and this is also the focus of our paper. Using information criteria such as AIC, it is possible to combine estimates from different nested models to one estimate for each predictor (Burnham and Anderson 2002). IT-based model averaging allows the estimation of confidence intervals that incorporate model estimates of the sampling variance for each model and of the model selection uncertainty. Lukacs et al. (2010) have demonstrated that model averaging yields unbiased point estimate, but unfortunately they did not consider estimates from the full model. It would be fruitful to study whether the confidence intervals estimates are closer to the nominal level (i.e. 95% confidence intervals should include the true value in 95% of all cases) when based on model averaging rather than on full models.

Three questions to consider before analysis

Research questions vary in so many aspects that it seems impossible to give advice that would equally apply to all situations. Nevertheless, there are three questions that one might want to critically examine in order to arrive at the best solution for a given situation:

1. Is there a need to test the global null hypothesis? Is it possible that the dependent variable could be unaffected by all the predictors studied? Sometimes, the global null hypothesis is biologically irrelevant (Perneger

1998), and the aim of analysis might only be to rank the various predictors in terms of the variance they explain. Hence, considering that there is also a type II error, the greatest possible statistical conservatism might not always be the best choice (Nakagawa 2004). However, note that the global null hypothesis should not be rejected prematurely with reference to the literature, as the literature necessarily contains a very large number of type I errors as well (Ioannidis 2005).

2. How much prior information is there on the given study system? So is the analysis still largely exploratory, or is there a clear focus on one or a few predictors? If your original interest is in only one specific predictor, say an experimental treatment, do you want to also consider treatment by covariate interactions? If so, a targeted Bonferroni correction might be sensible. Also, one should keep in mind that when the focus is shifted from a main effect to an interaction, it is often also required to alter the model structure with regard to random effects (see Schielzeth and Forstmeier 2009), because misspecified random effects pose an additional risk of inflated type I errors.
3. How large is the sample size? It is very important to stay with the number of predictors in the full model (including interactions) at least below $N/3$ (preferentially much lower). Otherwise, one should exclude a priori the least interesting or biologically least likely interactions. Note that such decisions cannot be made after examining the data.

In case of explorative data analysis in a study system without clear a priori predictions, we suggest using full model tests or P value adjustments as guidelines to evaluate if a significant effect might arise by sampling variation in a given set of predictors tested. Being realistic about the possibility to find overestimated effect sizes among a number of predictors tested might help planning promising follow-up studies (Ioannidis 2008). In any case, we suggest reporting all effects along with their standard errors, since this has the greatest value for the scientific community. In our view, effects that are estimated because they are connected to hypotheses of interest should not be removed because of being non-significant. While non-significance usually suggests a small effect size, even the sign of a non-significant effect might be of interest and the width of the confidence interval indicates an upper limit of the effect.

Conclusions

With the present paper, we hope to increase the awareness that elegant looking minimal models may be just as loaded with problems of multiple hypotheses testing as the long

tables that normally evoke a call for Bonferroni correction. Furthermore, we think that, in many cases, full models provide a better substrate for an independent evaluation by the reader. It is important to know from which full model the minimal model was derived, how many factors and interactions it contained, and one has to make sure that the initial full model was not over-fitted. As a consequence we recommend that publications should always provide (1) information on the structure of the full model and (2) either the test statistic of the full model and/or preferentially all parameter estimates (with their standard errors) of the full model.

Acknowledgements We thank Niels Dingemans, Alain Jacot, Laszlo Garamszegi, Bart Kempenaers and three anonymous Referees for critical comments on earlier versions of the manuscript. The study was funded by the German Science Foundation (DFG), Emmy Noether Fellowship (FO 340/1-3 to WF).

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Aiken LS, West SG (1991) Multiple regression: testing and interpreting interactions. Sage Publications, Newbury Park
- Anderson DR, Burnham KP, Thompson WL (2000) Null hypothesis testing: problems, prevalence, and an alternative. *J Wildl Manage* 64:912–923
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Methodol* 57:289–300
- Blanchet FG, Legendre P, Borcard D (2008) Forward selection of explanatory variables. *Ecology* 89:2623–2632
- Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Springer, Berlin
- Chatfield C (1995) Model uncertainty, data mining and statistical inference. *J R Stat Soc, A* 158:419–466
- Crawley MJ (2007) The R book. Wiley, Chichester
- Derksen S, Keselman HJ (1992) Backward, forward and stepwise automated subset-selection algorithms: frequency of obtaining authentic and noise variables. *Br J Math Stat Psychol* 45:265–282
- Dochtermann NA, Jenkins SH (2010) Developing multiple hypotheses in behavioural ecology. *Behavioral Ecology and Sociobiology*. doi:10.1007/s00265-010-1039-4
- Field A (2005) Discovering statistics using SPSS. Sage, London
- Freckleton RP (2010) Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. *Behavioral Ecology and Sociobiology*. doi:10.1007/s00265-010-1045-6
- Garamszegi LZ (2010) Information-theoretic approaches in statistical analysis in behavioural ecology: an introduction. *Behavioral Ecology and Sociobiology*. doi:10.1007/s00265-010-1028-7
- Garamszegi LZ, Calhim S, Dochtermann N, Hegyi G, Hurd PL, Jørgensen C, Kutsukake N, Lajeunesse MJ, Pollard KA, Schielzeth H, Symonds MRE, Nakagawa S (2009) Changing

- philosophies and tools for statistical inferences in behavioral ecology. *Behav Ecol* 20:1376–1381
- Göring HHH, Terwilliger JD, Blangero J (2001) Large upward bias in estimation of locus-specific effects from genomewide scans. *Am J Hum Genet* 69:1357–1369
- Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75:800–802
- Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat* 6:65–70
- Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* 2:696–701
- Ioannidis JPA (2008) Why most discovered true associations are inflated. *Epidemiology* 19:640–648
- Ioannidis JPA, Thomas G, Daly MJ (2009) Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet* 10:318–329
- Jennions MD, Møller AP (2002) Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. *Proc R Soc Lond B Biol Sci* 269:43–48
- Jennions MD, Møller AP (2003) A survey of the statistical power of research in behavioral ecology and animal behavior. *Behav Ecol* 14:438–445
- Kelly CD (2006) Replicating empirical research in behavioral ecology: how and why it should be done but rarely ever is. *Q Rev Biol* 81:221–236
- Lukacs PM, Burnham KP, Anderson DR (2010) Model selection bias and Freedman's paradox. *Ann Inst Stat Math* 62:117–125
- Miller AJ (1984) Selection of subsets of regression variables. *J R Stat Soc, A* 147:389–425
- Mundry R (2010) Issues in information theory based statistical inference: a commentary from a frequentist's perspective. *Behavioral Ecology and Sociobiology*. doi:10.1007/s00265-010-1040-y
- Mundry R, Nunn CL (2009) Stepwise model fitting and statistical inference: turning noise into signal pollution. *Am Nat* 173:119–123
- Nakagawa S (2004) A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behav Ecol* 15:1044–1045
- Perneger TV (1998) What's wrong with Bonferroni adjustments? *Br Med J* 316:1236–1238
- Quinn GP, Keough MJ (2002) *Experimental design and data analysis for biologists*. Cambridge University Press, Cambridge
- R Development Core Team (2008) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna
- Rice WR (1989) Analyzing tables of statistical tests. *Evolution* 43:223–225
- Schielzeth H (2010) Simple means to improve the interpretability of regression coefficients. *Meth Ecol Evol* 1:103–113
- Schielzeth H, Forstmeier W (2009) Conclusions beyond support: overconfident estimates in mixed models. *Behav Ecol* 20:416–420
- Stephens PA, Buskirk SW, del Rio CM (2007) Inference in ecology and evolution. *Trends Ecol Evol* 22:192–197
- Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. *Proc Natl Acad Sci USA* 100:9440–9445
- Tibbetts EA, Dale J (2007) Individual recognition: it is good to be different. *Trends Ecol Evol* 22:529–537
- Venables WN, Ripley BD (2002) *Modern applied statistics with S*, 4th edn. Springer, New York
- Verhoeven KJF, Simonsen KL, McIntyre LM (2005) Implementing false discovery rate control: increasing your power. *Oikos* 108:643–647
- Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP (2006) Why do we still use stepwise modelling in ecology and behaviour? *J Anim Ecol* 75:1182–1189
- Wright SP (1992) Adjusted *P*-values for simultaneous inference. *Biometrics* 48:1005–1013
- Zhang P (1992) Inference after variable selection in linear regression models. *Biometrika* 79:741–746
- Zöllner S, Pritchard JK (2007) Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *Am J Hum Genet* 80:605–615