Accepted Manuscript

Model-averaged confidence intervals for factorial experiments

David Fletcher, Peter Dillingham

PII: S0167-9473(11)00180-0
DOI: 10.1016/j.csda.2011.05.014
Reference: COMSTA 5011

To appear in: *Computational Statistics and Data Analysis*

Received date: 30 November 2010
Revised date: 17 May 2011
Accepted date: 18 May 2011

Please cite this article as: Fletcher, D., Dillingham, P., Model-averaged confidence intervals for factorial experiments. *Computational Statistics and Data Analysis* (2011), doi:10.1016/j.csda.2011.05.014

Model-Averaged Confidence Intervals for Factorial Experiments

David Fletcher [a,1] and Peter Dillingham [a,2]

[a] Department of Mathematics and Statistics, University of Otago, PO Box 56, Dunedin, New Zealand

[1] Corresponding author: Email dfletcher@maths.otago.ac.nz
Fax +64 3 479 8427
Phone +64 3 479 7762

[2] Current address: George Perkins Marsh Institute, Clark University, 950 Main Street, Worcester, MA 01610-1477, U.S.A.

Abstract

We consider the coverage rate of model-averaged confidence intervals for the treatment means in a factorial experiment, when we use a normal linear model in the analysis. Model-averaging provides a useful compromise between using the full model (containing all main effects and interactions) and a "best model" obtained by some model-selection process. Use of the full model guarantees perfect coverage, whereas use of a best model is known to lead to narrow intervals with poor coverage. Model-averaging allows us to achieve good coverage using intervals that are also narrower than those from the full model. We compare four information criteria that might be used for model-averaging in this setting: AIC, AIC_c, AIC_c^* and BIC. In this setting, if the full model is "truth", all the criteria will have perfect coverage rates asymptotically. We use simulation to assess the coverage rates and interval widths likely to be achieved by a confidence interval with a nominal coverage of 95%. Our results suggest that AIC performs best in terms of coverage rate; across a wide range of scenarios and replication levels, it consistently provides coverage rates within 1.5 percentage points of the nominal level, while also leading to reductions in interval-width of up to 30%, compared to the full model. AIC_c performed worst overall, with a coverage rate that was up to 5.2 percentage points too low. We recommend that model-averaging become standard practice when summarising the results of a factorial experiment in terms of the treatment means, and that AIC be used to perform the model-averaging.

Keywords: coverage rate, information criterion, model uncertainty, model weight.

1. Introduction

In many application areas, it is increasingly common to allow for model uncertainty when providing parameter estimates and confidence intervals. This extra level of uncertainty arises because we cannot be sure that a model-selection process will always lead to the same best model (Chatfield 1995). Traditionally parameter estimation has been carried out using a single model, often after model selection, and is therefore conditional upon choice of that model. Model-averaging has been proposed as a means of allowing for some of the model uncertainty, in that it is conditional upon a set of models rather than a single best model (Buckland et al. 1997; Burnham and Anderson 2002; Claeskens and Hjort 2008). A model-averaged estimate of a parameter is a weighted mean of a set of single-model estimates for the parameter, where the weights are typically chosen using Akaike's information criterion (*AIC*), Bayes' information criterion (*BIC*), or using bootstrap methods (Buckland et al. 1997).

Methods for calculating a confidence interval around a model-averaged estimate have been considered by Buckland et al. (1997), Burnham and Anderson (2002), Claeskens and Hjort (2008) and Hjort and Claeskens (2003). Hjort and Claeskens (2003) assessed the asymptotic properties of some of these methods, but did not consider the coverage rates that might be achieved with real data. Lukacs et al. (2010) used simulation in the context of linear regression to show that the coverage rate for a model-averaged confidence interval for each parameter in the generating model was close to the nominal level, and was a substantial improvement over stepwise regression. Wheeler and Bailer (2009) used simulation to assess the coverage rates of model-averaged confidence intervals obtained using *AIC* and *BIC* in the context of dose-response relationships, and

found that *AIC* performed marginally better. Chen, Giannakouros and Yang (2007) considered the use of model-averaging in the context of factorial ANOVA, but did not focus on coverage rates of confidence intervals.

The purpose of this paper is twofold. First, to propose that model-averaging become the default method for estimating treatment means in a factorial experiment. We illustrate the benefits of its use in this context via a simulation study. Second, in the simulation study we also compare four information criteria that might be used to perform model-averaging: three variations of *AIC*, and *BIC*. In Section 2 we define our notation and several methods for model-averaging. In Section 3 we illustrate the use of model-averaging by analysing the results from a 2^3 factorial experiment. In Sections 4 and 5 we describe and present the results from a simulation study of the coverage properties of confidence intervals for treatment means. We conclude with recommendations and ideas for further research in Section 6.

2. Notation and Methods

Suppose we use a normal linear model to analyse data from a factorial experiment and we wish to estimate θ, the expected value of the response variable *Y* for a particular treatment combination. An analysis based on a single model involves use of the following formula to obtain a 95% confidence interval for θ

$$\hat{\theta} \pm t\sqrt{\hat{V}\left(\hat{\theta}\right)} \tag{1}$$

where $\hat{\theta}$ is the estimate of θ obtained from the model and t is the 97.5th percentile of the t-distribution with degrees of freedom equal to the error degrees of freedom for that model. If we fit a set of *R* models, a model-averaged estimate of θ is

$$\bar{\theta} = \sum_i w_i \hat{\theta}_i \tag{2}$$

where w_i is the weight attached to model *i* and $\hat{\theta}_i$ is the estimate obtained from that model $(i = 1,...,R)$. A 95% confidence interval for θ can be calculated as

$$\bar{\theta} \pm 1.96 \sqrt{\sum_{i=1}^{R} w_i \left\{ \left(\frac{t_i}{1.96} \right)^2 \hat{V}\left(\hat{\theta}_i\right) + \left(\hat{\theta}_i - \bar{\theta}\right)^2 \right\}} \tag{3}$$

where $\hat{V}\left(\hat{\theta}_i\right)$ is an estimate of the conditional variance of $\hat{\theta}_i$ given model *i* and t_i is the 97.5th percentile of the t-distribution with degrees of freedom equal to the error degrees of freedom for model *i*. Equation 3 uses the estimate of the unconditional variance of $\bar{\theta}$ given in Equation 6.12 of Burnham and Anderson (2002), with a modification recommended by them to allow for the uncertainty in $\hat{V}\left(\hat{\theta}_i\right)$ (Burnham and Anderson 2002, p.164). In order to check the robustness of our conclusions, we also considered the following formula, based on an earlier estimate of the unconditional variance of $\bar{\theta}$ proposed by Buckland et al (1997)

$$\bar{\theta} \pm 1.96 \sum_{i=1}^{R} w_i \sqrt{\left(\frac{t_i}{1.96} \right)^2 \hat{V}\left(\hat{\theta}_i\right) + \left(\hat{\theta}_i - \bar{\theta}\right)^2} \tag{4}$$

(see Burnham and Anderson 2002, p164). The Cauchy-Schwarz inequality (Solomentsev 2001) implies that Equation 4 will lead to slightly narrower intervals than Equation 3.

We consider four information criteria that have been proposed for calculating the model weights: three variations of *AIC*, and *BIC*. We do not consider the bootstrap, as we wish to focus on methods that are not computer-intensive. All four criteria specify the weight for model *i* as

$$w_i = \frac{\exp\left(-\Delta IC_i/2\right)}{\sum_{i=1}^{R}\exp\left(-\Delta IC_i/2\right)} \tag{5}$$

where $\Delta IC_i = IC_i - IC_{\min}$, IC_i is the value of the criterion for model *i* and $IC_{\min}$ is its minimum value across all *R* models. For the normal linear model, the criteria are given by (we omit the subscript *i* for simplicity of notation):

$$AIC = n\log\hat{\sigma}^2 + 2p \tag{6}$$

$$AIC_c = n\log\hat{\sigma}^2 + 2np/\left(n-p-1\right) \tag{7}$$

$$AIC_c^* = n\log\tilde{\sigma}^2 + p \tag{8}$$

$$BIC = n\log\hat{\sigma}^2 + p\log n \tag{9}$$

where $\hat{\sigma}^2 = E/n$, $\tilde{\sigma}^2 = E/\left(n-p-1\right)$, E is the error sum of squares, *n* is the number of observations and p is the number of parameters, including the intercept and σ^2. Claeskens and Hjort (2008) is a useful reference for details of these four criteria.

3. Example

We illustrate the use of model-averaging when analysing the results from a factorial experiment with data from a completely randomised design involving three factors, each at two levels, taken from Mead (1990, p.39). Eight frogs and eight toads were kept in either moist or dry conditions and half were then injected with a water-balance hormone. The response variable was the percent increase in weight after immersion in water for two hours, with the factors being *species* (frog or toad), *conditions* (moist or dry) and *hormone* (yes or no). The factors and their levels are shown in Table 1, together with the data. The analysis of variance table from fitting the full model is shown in Table 2. This suggests that a potential best model is one involving just the main effects (and possibly the *species-hormone* interaction).

Suppose we consider selection of a best model and use of model-averaging over the set of all possible models (subject to the usual hierarchical restrictions inherent in factorial models). This set is shown in Table 3, together with the model weights according to the four information criteria. AIC gives the highest weights to models 13, 15, 17, 18 (all of which contain all main effects plus at least one two-way interaction) and model 19 (the full model). AIC_c prefers smaller models, as expected, giving highest weights to 5, 8 and 13 (all of which contain the main effects of *species* and *conditions*). AIC_c^* and BIC give almost identical weights, with highest weights to 13, 15, 17 and 18. In terms of a best model, use of AIC suggests the full model, while the other criteria all suggest model 13 (which contains all main effects and the *species-hormone* interaction).

Figure 1 shows the estimates and 95% confidence intervals for each combination of factor levels, using model 19 (the full model and the best according to AIC), model 13

(the best according to all criteria except AIC) and, for illustration, model-averaging with AIC. As predicted, use of model 13 leads to the narrowest intervals, and the width of each model-averaging interval lies between the width for models 19 and 13. Interestingly, despite the differences in the model weights in Table 3, the four criteria give similar results when using model-averaging.

4. Simulations

We simulated data from the following model for a 2^3 study involving r replicates:

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \varepsilon_{ijkl} \quad (10)$$

where μ is the overall effect, $\{\alpha_i, \beta_j, \gamma_k\}$ are the main effects, $\{\alpha\beta_{ij}, \alpha\gamma_{ik}, \beta\gamma_{jk}\}$ are the two-way interactions, $\alpha\beta\gamma_{ijk}$ is the three-way interaction, and ε_{ijkl} is the error term, with $\mathrm{V}(\varepsilon_{ijkl}) = \sigma^2$ $(i = 1, 2;\ j = 1, 2;\ k = 1, 2;\ l = 1, ..., r)$. The value of μ can have no impact on the coverage rates and widths of confidence intervals, so we arbitrarily set $\mu = 0$. In addition, for simplicity we set $\sigma^2 = 1$ as we expected the results to be influenced by the "signal-to-noise" ratio, i.e. by the effect sizes *relative* to σ^2.

We used the following random-effects generating model in the simulations. First, each of the three sets of parameter values (main effects, two-way interactions, three-way interaction) was assigned a "magnitude" of 2 (*High*), 1 (*Medium*) or 0.1 (*Low*). As $\sigma^2 = 1$, these can be thought of as signal-to-noise ratios. Second, for each run, the value for each parameter in a set was generated randomly from a normal distribution with mean

zero and standard deviation equal to the magnitude for that set. Thus for a model in which the main effects are *High*, the two-way interactions are *Medium* and the three-way interaction is *Low* (denoted as model-scenario *HML*), the parameters were generated as follows

$$\begin{aligned} &\alpha_i \sim N(0,2^2) \qquad \beta_j \sim N(0,2^2) \qquad \gamma_k \sim N(0,2^2) \\ &\alpha\beta_{ij} \sim N(0,1^2) \qquad \alpha\gamma_{ik} \sim N(0,1^2) \qquad \beta\gamma_{jk} \sim N(0,1^2) \\ &\alpha\beta\gamma_{ijk} \sim N(0,0.1^2) \end{aligned} \tag{11}$$

We considered a total of 10 model-scenarios: *LLL*, *MLL*, *HLL*, *MML*, *HML*, *MMM*, *HMM*, *HHL*, *HHM*, and *HHH*. These were chosen in order to focus on those we would expect to encounter in practice (Burnham and Anderson 2002, p.21; Mead 1988, p.368). Thus the main effect magnitude was always greater than or equal to the two-way interaction magnitude, which in turn was greater than or equal to the three-way interaction magnitude.

The choice of 0.1 for the *Low* magnitude meant that model-scenarios including an *L* would provide essentially the same results as those we would obtain by assuming a magnitude of zero for those terms, i.e. they would be equivalent to assuming that the true model was a reduced model. For example, the model-scenario *HML* has three-way interactions that are so small that the true model is very close to one without a three-way interaction. We did not consider model-scenarios with magnitudes of zero because we do not believe they reflect reality (Buckland et al. 1997; Burnham and Anderson 2004): there will always be some three-way interaction in such an experiment, even if it is very small.

Each simulation run involved generating the parameter values and then generating the data using the model in Equation 10. We chose to use this "random-effects generating model" in order to consider the coverage properties across a range of models for each model-scenario. This allowed us to obtain broad conclusions about the performance of the methods, compared to the use of fixed parameter values in all simulation runs.

We considered eight methods for calculating a 95% confidence interval for θ. These involved use of either a best model (and Equation 1) or model-averaging (and Equation 3 or 4), for each of the four information criteria. For each simulation run, and for each method, we calculated a confidence interval for the expected value of each of the eight combinations of factor levels. We calculated the coverage rate, over all simulations, for each combination, and summarised these by calculating the mean rate over the eight combinations. Likewise, we calculated the percent reduction in the interval-width, relative to that for the full model, and summarised these by calculating their mean over the eight combinations. Note that we did not estimate coverage rates from using the full model, as these must all be 0.95 (see the Introduction).

In order to assess the effect of the number of replicates, we considered $r = 2, 4,$ and 8. We also considered the extreme case $r = 10^4$, primarily in order to check the large-sample behaviour of some of the methods. Likewise, in order to assess the impact of the choice of model-set, we repeated the simulations (with $r = 2, 4,$ and 8) using a natural smaller set of models: the null model, the model with all main effects, the model with all main effects and two-way interactions, and the full model (models 1,8,18 and 19). In all simulations, we used 25,000 simulation runs, which ensured that the

coverage rate for each combination was estimated with a binomial standard error of approximately 0.001. All calculations were performed in R Version 2.10.1 (2009).

5. Results

Use of Equations 3 and 4 led to very similar coverage rates, the differences being small relative to those between both the criteria and the scenarios. Overall, for all three variations of *AIC*, the two equations performed equally well. For *BIC*, use of Equation 3 always provided slightly better coverage than Equation 4. For simplicity of presentation, we therefore focus attention on the results for Equation 3.

Table 4 shows the error in the coverage rate (difference between the mean coverage rate and 0.95) for each method, with a negative value indicating a coverage rate that is too low. The model-scenarios have been ordered according to the coverage rates achieved using the best model according to AIC, as this will roughly correspond to increasing complexity in the best model.

As expected, model-averaging was always better than use of a best model, giving a coverage error of at most 5.2% compared to over 20% with a best model, and we do not consider coverage for the best-model approach any further. Overall, model-averaging using AIC consistently performed better than the other criteria, its error being at most 1.5% and remaining stable across model-scenarios and levels of replication. AIC_c^* was almost as good, with an error of at most 2%. The largest errors occurred using AIC_c for $r = 2$. The only scenarios for which AIC_c had a smaller error than AIC were those in which all interactions were of *Low* magnitude, for $r = 4$ and 8. Unlike the other criteria, the performance of BIC did not seem to improve as r increased, its largest error being

for model-scenario *LLL* with $r = 8$. The simulations that we carried out with $r = 10^4$, showed that for this model-scenario *BIC* still had a coverage error of 2.8% compared to 0.1% for the other criteria. For all the criteria, the error was generally greater for model-scenarios *MML*, *HML*, *MMM* and *HMM*, especially for AIC_c and $r = 2$. This is to be expected, as these four scenarios are likely to have the greatest uncertainty in model selection.

Table 5 shows the mean percent reduction in interval-width, relative to the full model, for each method. Again as expected, use of a best model provided much narrower intervals than model-averaging, which led to the poor coverage discussed above. For model-averaging, use of AIC and AIC_c^* led to slightly wider intervals than the other two methods, as would be expected from the coverage rates. Overall, the greatest reductions in interval-width occurred for model-scenarios in which the interactions were of low magnitude (*LLL*, *MLL* and *HLL*), with interval-widths increasing as the complexity of the best model increased. Interestingly, for model-scenarios *HHM* and *HHH* and $r = 2$, AIC_c produced intervals wider than those from the full model.

The results obtained from the simulations using the smaller model set, were similar to those for the orignal model set, so the details are not given here. For AIC and AIC_c^*, the results were virtually unchanged, with the coverage rate being in error by at most 1.6% and 2.2% respectively. For AIC_c and BIC the errors in coverage rate were generally even worse than for the original model set.

6. Discussion

Analysis of a factorial experiment is a natural setting in which to use model-averaging, when we are interested in estimating the treatment means. Note that it is not appropriate to use model-averaging for estimating main effects and interactions, as interpretation of these changes with the model fitted (Davison 2003, p.470). Our results suggest that model-averaged confidence intervals can have coverage rates close to the nominal level *and* be narrower than those from the full model. As expected, use of a best model led to coverage rates that were too low, in line with previous work on model selection (e.g. Hurvich and Tsai 1990; Lukacs et al 2010); our sole purpose in including this approach was to emphasise again its shortcomings, particularly compared with model-averaging, and because it is still common practice for the results from a factorial experiment to be summarised using a best model.

An important conceptual issue in model selection and model-averaging is whether or not it is reasonable to assume that the "true model", i.e. the process that generated the data, is in the set of models being considered in the analysis, a related issue being whether or not it is reasonable to assume that the true model is simple or complex (Burnham and Anderson 2004; Link and Barker 2006). A factorial experiment is one of the few types of study in which these issues are clear-cut: if the factors have been chosen carefully and randomisation has been suitably employed, the full model, containing all main effects and interactions, can be regarded as the true model. If the data are normally distributed, intervals obtained from this model will have perfect coverage, at the expense of being wider than those obtained using model-averaging. Model-averaging can be thought of as a good compromise between use of the full model and a best model. Where the model weights indicate that smaller models are preferable, model-averaging will

provide narrower intervals than the full model; unlike use of a best model, model-averaging allows for model uncertainty and should therefore lead to coverage rates that are closer to the required level.

When just one of the candidate models is the true model, as here, both *AIC* and *BIC* are consistent, i.e. the weight they give to this model will tend to one as the sample size increases (Claeskens and Hjort 2008, Sections 4.1-4.3). For a factorial experiment, use of model-averaging based on either of these criteria is therefore asymptotically equivalent to use of the full model, which has perfect coverage. Although this result is useful, it does not provide any guidance as to the coverage rates model-averaging will lead to with real data, for which we needed simulation.

For model-averaging, AIC and AIC_c^* consistently performed best, BIC was slightly worse, and AIC_c was the worst overall. AIC_c provided a slight improvement on AIC for the scenarios where $r = 4$ or 8 and there were strongly-tapering effects, i.e. when giving higher weight to smaller models might be beneficial. Both AIC_c and AIC_c^* were designed as bias-corrected improvements on AIC when we have a normal linear model, and AIC_c has been adopted by many researchers as the method of choice, especially when the sample size is small compared to the number of parameters (Burnham and Anderson 2002; Claeskens and Hjort 2008; Hurvich and Tsai 1989; Sugiura 1978). However, the theory underlying the development of AIC_c and AIC_c^* did not focus on coverage rates of the resulting confidence intervals. Our results suggest that AIC_c can perform worse than AIC in terms of coverage rate, even for small sample

sizes. Although AIC_c^* performed almost as well as AIC it cannot be easily generalised to other settings (Claeskens and Hjort 2008).

As noted earlier, the results for the model-scenarios that include an *L* should be close to those we would obtain if we were to assume that the true model was a reduced model. Our study has therefore included scenarios in which we might expect AIC_c to perform better than AIC. For example, the model-scenario *HML* will provide results very close to those we would obtain if we assumed that the true model is a model with no three-way interaction. For such a true model, we might expect AIC to perform worse than AIC_c, putting more weight on the full model, which would be larger than necessary. Having included such model-scenarios, our conclusions about the relative performance of AIC and AIC_c should therefore be robust. As usual, there is always a balance to be struck between the achieving the nominal coverage rate and keeping the interval width small: we believe it is important that a method provides a reliable coverage rate before considering the width of the interval.

Previous comparisons of these criteria have tended to focus largely on selection of a best model, either in terms of the probability that it is the true model (consistency) or in terms of the predictive mean squared error (*PMSE*) associated with using the best model (e.g. Burnham and Anderson 2002; Burnham and Anderson 2004; Claeskens and Hjort 2008). Our results suggest that in this setting AIC_c and AIC_c^* perform worse than AIC in terms of coverage rates.

Model-averaging is well-established within the Bayesian approach to inference (Raftery et al. 1997; Volinsky et al. 1997) and we suggest there is also a need for research

on the coverage properties of credible intervals using Bayesian model-averaging (Little 2006).

When using model-averaging, there is a need to consider the choice of model set, as inference is conditional upon this set. In a factorial experiment, a natural choice is the set of all possible models. If the number of factors is large, or we have prior information, we may wish to use a smaller model set (Burnham and Anderson 2002). We therefore considered use of a smaller model set in our simulation study, and found that the differences between methods were not greatly affected by the choice of model set.

We have considered a range of model-scenarios and used a random-effects generating model in order to strengthen the robustness of our conclusions. Alternative values could be considered for the magnitudes of the main effects and interactions. Likewise, in non-experimental settings it might be worth considering the possibility that the true model is more complex that any of the candidate models, and a simulation study could reflect this. It is an interesting open question as to whether *AIC* performs best in such settings.

Acknowledgements

We are grateful to David Anderson and Ken Burnham for commenting on a draft of the paper, as well as to the reviewers for their helpful comments.

References

Buckland, S.T., Burnham, K.P., Augustin, N.H., 1997. Model selection: an integral part of inference. Biometrics 53, 603-618.

Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference: a practical information-theoretic approach, second ed. Springer, New York.

Burnham, K.P., Anderson, D.R., 2004. Multimodel inference: understanding AIC and BIC in model selection, Sociological Methods and Research, 33, 261-304.

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society, Series A 158, 419-466

Chen, L., Giannakouros, P. Yang, Y., 2007. Model combining in factorial data analysis. Journal of Statistical Planning and Inference 137, 2920–2934

Claeskens, G., Hjort, N.L., 2008. Model selection and model averaging. Cambridge University Press, Cambridge.

Davison, A.C., 2003. Statistical models. Cambridge University Press, Cambridge.

Hjort, N.L., Claeskens, G., 2003. Frequentist model average estimators. Journal of the American Statistical Association 98, 879-899.

Hurvich, C.M. Tsai, C.-L., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307.

Hurvich, C.M. Tsai, C.-L., 1990. The impact of model selection on inference in linear regression. The American Statistician 44, 214-217.

Link, W.A. Barker, R.J., 2006. Model weights and the foundations of multimodel inference. Ecology 87, 2626–2635.

Little, R.J., 2006. Calibrated Bayes: A Bayes/frequentist roadmap. American Statistician 60, 213–223.

Lukacs, P.M., Burnham, K.P. Anderson, D.R., 2010. Model selection bias and Freedman's paradox. Annals of the Institute of Statistical Mathematics 62, 117–125.

Mead, R., 1988. The design of experiments: statistical principles for practical applications. Cambridge University Press, Cambridge.

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

Raftery, A.E., Madigan, D., Hoeting, J.A., 1997. Bayesian model averaging for linear regression models. Journal of the American Statistical Association 92, 179-191.

Solomentsev, E.D., 2001. Cauchy inequality, in: Hazewinkel, B. (Ed.), Encyclopaedia of Mathematics, Springer, New York.

Sugiura, N., 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. Communications in Statistics, Theory and Methods A7, 13–26.

Volinsky, C.T., Madigan, D., Raftery, A.E., Kronmal, R.A., 1997. Bayesian model averaging in proportional hazard models: assessing the risk of a stroke. Journal of the Royal Statistical Society, Series C 46, 433-448.

Wheeler, M.W. Bailer, A.J., 2009. Comparing model averaging with other model selection strategies for benchmark dose estimation. Environmental and Ecological Statistics 16, 37–51.

Table 1. Factors, combination labels, and percentage change in weight for each individual, from the water uptake experiment.

Species (S)	Condition (C)	Hormone (H)	Combination	Percent Change in Weight	
				Individual 1	Individual 2
Frog	Dry	No	FDN	2.5	17.7
		Yes	FDY	13.7	7.4
	Wet	No	FWN	0.9	2.9
		Yes	FWY	3.8	2.9
Toad	Dry	No	TDN	17.7	25.2
		Yes	TDY	28.4	27.9
	Wet	No	TWN	2.3	–1.6
		Yes	TWY	28.4	14.2

Table 2. Analysis of variance obtained from fitting the full model to the data from the water uptake experiment involving three factors: Species, Condition and Hormone. In this table S, C and H denote the main effects; SC, SH and CH denote the two-way interactions; SCH denotes the three-way interaction.

	df	SS	MS	F	p
S	1	514.2	514.2	15.0	0.005
C	1	469.8	469.8	13.7	0.006
H	1	218.3	218.3	6.4	0.036
SC	1	39.4	39.4	1.2	0.315
SH	1	165.8	165.8	4.8	0.059
CH	1	58.1	58.1	1.7	0.229
SCH	1	43.9	43.9	1.3	0.291
Error	8	274.4	34.3		
Total	15	1783.9			

Table 3. Full set of models for the water-uptake example, with model weights using four different methods. S, C and H denote the main effects of *species*, *conditions* and *hormone* respectively; SC, SH and CH denote the corresponding two-way interactions, and SCH denotes the three-way interaction.

		Model weight			
Number	Model	*AIC*	AIC_c	AIC_c^*	*BIC*
1	Null	0.000	0.009	0.000	0.001
2	S	0.000	0.028	0.002	0.002
3	C	0.000	0.021	0.001	0.002
4	H	0.000	0.005	0.000	0.000
5	S+C	0.006	0.185	0.019	0.021
6	S+H	0.001	0.021	0.002	0.002
7	C+H	0.001	0.015	0.002	0.002
8	S+C+H	0.030	0.267	0.070	0.066
9	S+C+SC	0.003	0.031	0.008	0.008
10	S+H+SH	0.001	0.009	0.002	0.002
11	C+H+CH	0.000	0.003	0.001	0.001
12	S+C+H+SC	0.019	0.033	0.032	0.029
13	S+C+H+SH	0.161	0.272	0.269	0.241
14	S+C+H+CH	0.025	0.043	0.043	0.038
15	S+C+H+SC+SH	0.131	0.021	0.141	0.134
16	S+C+H+SC+CH	0.018	0.003	0.019	0.018
17	S+C+H+SH+CH	0.197	0.032	0.212	0.201
18	S+C+H+SC+SH+CH	0.184	0.001	0.112	0.128
19	S+C+H+SC+SH+CH+SCH	0.222	0.000	0.065	0.105

Table 4. Difference between mean coverage rate (percent) and nominal level for different methods of calculating a 95% confidence interval, for 10 model-scenarios and 3 levels of replication in a 2^3 factorial study, the mean being across the eight combinations of the factor levels. The model scenario denotes the magnitude of the main effects, two-way interactions and three-way interaction respectively, where L=*Low*, M=*Medium* and H=*High*. Confidence intervals using model-averaging were obtained using Equation 3. Results are from 25,000 simulations, and have a standard error of approximately 0.001.

		Mean for Best model				Mean for Model-averaging			
Replication	Model-scenario	*AIC*	AIC_c	AIC_c^*	*BIC*	*AIC*	AIC_c	AIC_c^*	*BIC*
2	LLL	-12.6	-12.7	-13.5	-13.6	-0.1	0.2	0.1	0.1
	MLL	-9.2	-13.0	-10.7	-11.6	-1.0	-2.3	-1.3	-1.5
	HLL	-7.0	-8.1	-7.7	-8.0	-0.9	-1.6	-1.1	-1.1
	MML	-5.6	-14.8	-7.4	-7.8	-1.4	-4.3	-2.0	-2.0
	HML	-5.2	-11.8	-6.7	-6.8	-1.5	-3.6	-2.0	-1.9
	MMM	-4.0	-16.1	-6.4	-6.3	-1.2	-5.2	-1.9	-1.8
	HMM	-3.8	-13.0	-5.7	-5.6	-1.3	-4.2	-1.9	-1.8
	HHL	-3.0	-7.2	-3.8	-3.7	-0.9	-2.3	-1.2	-1.2
	HHM	-2.4	-8.1	-3.4	-3.2	-0.9	-2.6	-1.3	-1.1
	HHH	-1.5	-8.5	-2.4	-2.2	-0.6	-2.7	-0.9	-0.8
4	LLL	-13.9	-15.2	-14.4	-15.9	1.3	0.8	1.2	0.0
	MLL	-8.2	-9.6	-8.7	-11.3	0.1	-0.6	-0.1	-1.2
	HLL	-5.9	-6.4	-6.1	-7.1	0.3	-0.1	0.3	-0.4
	MML	-4.4	-6.8	-4.9	-7.9	-0.7	-1.6	-0.8	-1.9
	HML	-4.2	-6.2	-4.6	-6.9	-0.7	-1.5	-0.8	-1.6
	MMM	-3.3	-6.5	-4.0	-7.0	-0.9	-2.1	-1.2	-2.2
	HMM	-3.1	-6.0	-3.7	-6.2	-0.8	-2.0	-1.0	-2.0
	HHL	-2.3	-3.2	-2.5	-3.3	-0.2	-0.7	-0.3	-0.7
	HHM	-2.1	-3.8	-2.4	-3.8	-0.5	-1.2	-0.7	-1.2
	HHH	-1.3	-2.5	-1.5	-2.4	-0.4	-0.9	-0.5	-0.8
8	LLL	-16.4	-17.7	-16.8	-22.8	0.6	0.1	0.4	-3.8
	MLL	-7.3	-8.0	-7.5	-11.3	0.4	0.0	0.3	-1.9
	HLL	-5.7	-6.0	-5.8	-7.4	0.5	0.3	0.4	-0.9
	MML	-3.4	-4.1	-3.6	-7.0	-0.4	-0.7	-0.5	-2.0
	HML	-3.3	-3.9	-3.4	-6.3	-0.3	-0.6	-0.4	-1.7
	MMM	-2.3	-3.2	-2.5	-6.2	-0.6	-1.0	-0.7	-2.3
	HMM	-2.2	-3.1	-2.4	-5.8	-0.6	-0.9	-0.6	-2.1
	HHL	-1.9	-2.2	-2.0	-3.2	-0.1	-0.3	-0.1	-0.7
	HHM	-1.7	-2.3	-1.9	-3.9	-0.5	-0.7	-0.6	-1.4
	HHH	-0.7	-1.1	-0.8	-2.0	-0.2	-0.3	-0.2	-0.7

Table 5. Mean percent reduction in interval-width, relative to the full model, for different methods of calculating a 95% confidence interval, for 10 model-scenarios and 3 levels of replication in a 2^3 factorial study, the mean being across the eight combinations of the factor levels. Confidence intervals using model-averaging were obtained using Equation 3. Results are from 25,000 simulations.

		Mean for Best model				Mean for Model-averaging			
Replication	Model-scenario	*AIC*	AIC_c	AIC_c^*	*BIC*	*AIC*	AIC_c	AIC_c^*	*BIC*
2	LLL	48	60	54	56	30	44	35	37
	MLL	35	45	40	40	22	30	25	25
	HLL	31	38	35	35	20	27	23	23
	MML	20	28	24	24	11	10	13	12
	HML	19	25	22	22	11	10	12	12
	MMM	13	21	17	17	6	0	7	7
	HMM	13	18	16	16	6	0	7	7
	HHL	14	16	16	15	8	4	9	9
	HHM	9	10	11	11	5	-5	5	5
	HHH	5	3	7	7	2	-15	2	2
4	LLL	48	54	50	58	27	34	29	39
	MLL	31	36	33	38	17	21	18	22
	HLL	28	32	29	33	16	20	17	20
	MML	15	18	16	19	6	7	7	7
	HML	14	17	15	18	6	7	7	7
	MMM	8	12	9	12	2	2	2	2
	HMM	8	11	9	11	2	2	2	2
	HHL	10	12	11	12	4	6	5	5
	HHM	6	8	6	8	1	1	1	1
	HHH	3	4	3	4	0	0	0	0
8	LLL	46	49	47	58	24	28	25	40
	MLL	29	31	29	36	15	17	15	22
	HLL	26	28	27	32	14	16	15	20
	MML	12	13	12	15	5	5	5	6
	HML	11	13	12	15	5	5	5	6
	MMM	5	6	5	9	1	1	1	1
	HMM	5	6	5	8	1	1	1	1
	HHL	8	9	9	10	4	4	4	5
	HHM	4	4	4	6	1	1	1	1
	HHH	2	2	2	3	0	0	0	0

Figure Captions

Figure 1. Estimates and 95% confidence intervals for each combination of factor levels in the water uptake experiment, using model 19 (black), model 13 (red) and model-averaging using *AIC* (blue). Model 13 contains all main effects plus the SH interaction; Model 19 contains all main effects and interactions. The treatment combination denotes the level of each factor, where F=Frog, T=Toad, D=Dry, W=Wet, N=No and Y=Yes (Table 1).

Figure 1.