Ann Inst Stat Math (2010) 62:117–125
DOI 10.1007/s10463-009-0234-4

Model selection bias and Freedman's paradox

Paul M. Lukacs · Kenneth P. Burnham · David R. Anderson

Received: 16 October 2008 / Revised: 10 February 2009 / Published online: 26 May 2009

Abstract In situations where limited knowledge of a system exists and the ratio of data points to variables is small, variable selection methods can often be misleading. Freedman (Am Stat 37:152–155, 1983) demonstrated how common it is to select completely unrelated variables as highly "significant" when the number of data points is similar in magnitude to the number of variables. A new type of model averaging estimator based on model selection with Akaike's AIC is used with linear regression to investigate the problems of likely inclusion of spurious effects and model selection bias, the bias introduced while using the data to select a single seemingly "best" model from a (often large) set of models employing many predictor variables. The new model averaging estimator helps reduce these problems and provides confidence interval coverage at the nominal level while traditional stepwise selection has poor inferential properties.

Keywords Akaike's information criterion · Confidence interval coverage · Freedman's paradox · Model averaging · Model selection bias · Model selection uncertainty · Multimodel inference · Stepwise selection

P. M. Lukacs (✉)
Colorado Division of Wildlife, 317 W. Prospect Road,
Fort Collins, CO 80526, USA
e-mail: Paul.Lukacs@state.co.us

K. P. Burnham · D. R. Anderson
U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit,
Colorado State University, 1484 Campus Delivery, Fort Collins, CO 80523, USA

1 Introduction

It is well known that statistical testing for variable selection results in models with poor inferential properties (Rencher and Pun 1980; Freedman 1983; Miller 2002, p. 85; McQuarrie and Tsai 1998, p. 427). Yet, it is common in scientific literature to see exploratory analyses using multiple linear or logistic regression on many variables and variable selection based on hypothesis testing. In these situations the number of models is often similar in magnitude to the size of the sample. The analysis is exploratory, yet the results are often presented as if they are confirmatory because some of the explanatory variables have large t-values (e.g. >2) giving the impression of importance.

Freedman (1983) demonstrated that variable selection methods based on testing for significant F statistics commonly include explanatory variables with no relation to the response and spuriously inflate R^2 when such irrelevant variables are present. This problem is often referred to as "Freedman's paradox." Freedman's paradox results in spurious effects appearing important. Others have demonstrated the same effect from other perspectives (Rencher and Pun 1980; Hurvich and Tsai 1990; George and McCulloch 1993). All of these investigations affirm that model selection is important, but selecting a single best model does not solve the problem regardless of what method is used to select the model. New approaches to model selection appear in recent books by Massart (2007) and Claeskens and Hjort (2008), but these methods do not address model selection bias.

Freedman's paradox is an extreme case of model selection bias, the bias introduced while using the data to select a single seemingly "best" model from a large set of models employing many predictor variables. Yet, model selection bias also occurs when an explanatory variable has a weak relationship with the response variable. The relationship is real, but small. Therefore, it is rarely selected as significant. When the variable is selected it is usually because the effect was overestimated with that dataset. These weak relationships do not necessarily result in a model with too many parameters (overfit); rather, parameters are poorly estimated. Miller (2002, p. 165) suggests that there is little that can be done to avoid selection bias. We propose an approach to substantially lessen the problem.

Model averaging is one of several methods for making formal inference from multiple models (Burnham and Anderson 2002). This approach is quite different from standard variable selection methods where inference is made only from the selected model. Model averaging admits from the beginning of the analysis that there is substantial uncertainty as to what model is best and what combination of variables is important. On the contrary, selection methods such as stepwise selection pick a single best model. Inference is then conditional on this model and variables not in the model are, therefore, deemed unimportant. These are two very different approaches.

Here we contrast model averaging and stepwise selection for their performance in identifying variables unrelated to the response. Often variables included in exploratory analyses have little or no influence on the response variable and there is therefore the potential for many spurious conclusions. Our objective is to illustrate model averaging versus stepwise selection in protecting against including spu-

rious variables, i.e. to guard against Freedman's paradox. In addition, we contrast model averaging and stepwise selection for their performance with weakly related variables.

2 Methods

The methods used here follow Freedman (1983) where applicable. One thousand matrices of 40 rows and 21 columns were generated using normally distributed random numbers with mean zero and variance one. The 40 rows represented data where $n = 40$. Column one was the response variable and columns 2–21 were explanatory variables. Freedman (1983) used 100 rows and 51 columns. We were unable to use such a large number of variables because the computing load was prohibitive.

All possible first-order linear regression models are fit to these simulated data ($2^{20} = 1{,}048{,}576$ models). All models include an intercept. Information-theoretic methods are used to rank and weight each model (Burnham and Anderson 2002). Models are ranked with AIC_c where

$$\mathrm{AIC}_c = -2 \log L\left(\hat{\beta}, \hat{\sigma}^2 | g_j, \text{data}\right) + 2K + \frac{2K\,(K+1)}{n-K-1} \tag{1}$$

where g_j is the jth model in the set of candidate models and K is the number of estimable parameters in model g_j. Also, $\hat{\beta}$ and $\hat{\sigma}^2$ are the maximum likelihood estimates of the K parameters. We chose to use this slight modification of Akaike's AIC (Akaike 1973) for two reasons: (1) because AIC_c was explicitly developed for the Guassian linear models (Sugiura 1978) and (2) because it performs a little better than AIC in situations where sample size is small relative to the number of parameters (Hurvich and Tsai 1989). Of course as sample size n becomes large relative to K, AIC_c converges to AIC.

Quantitative measures of the strength of evidence begin with an estimate of Kullback–Leibler information loss,

$$\Delta_j = \mathrm{AIC}_{cj} - \min \mathrm{AIC}_c \tag{2}$$

(Burnham and Anderson 2002, p. 74). Then, $\exp\left(-\frac{1}{2}\Delta_j\right)$ is the likelihood of model j given the data (Akaike 1979) and

$$\Pr\left[g_j \,|\, x\right] = w_j = \frac{\exp\left(-\frac{1}{2}\Delta_j\right)}{\sum_{i=1}^{R} \exp\left(-\frac{1}{2}\Delta_i\right)} \tag{3}$$

is the probability that model j is the best of all R models given the data (Akaike 1978, 1979). These model probabilities have been termed "Akaike weights, w_i" as they serve as weights in model averaging and related applications. Given a set of R models representing R science hypotheses, one of these model is, in fact, the best model in a Kullback–Leibler sense. Model selection under an information-

theoretic approach attempts to estimate the probability of model j being this K-L best model.

These Akaike weights or model probabilities are formal probabilities, but differ from Bayesian posterior probabilities, even if uninformative prior model probabilities are assumed. While the information-theoretic and Bayesian model probabilities have different meanings, we have seen several cases where numerical values for real data are quite similar.

Inference about model parameters, β_i, is based on a new type of model averaging estimator which averages $\hat{\beta}_i$ across all models, those including and excluding $\hat{\beta}_i$. The estimator is

$$\tilde{\bar{\beta}}_i = \sum_{j=1}^{R} w_j \hat{\beta}_{ij} \tag{4}$$

where $\hat{\beta}_{ij} \equiv 0$ if variable i is not included in model j (Burnham and Anderson 2002, p. 152). The $\tilde{\bar{\beta}}_i$ estimator is different than other model averaging estimators which only average over models including $\hat{\beta}_i$, often denoted $\hat{\bar{\beta}}_i$ (Buckland et al. 1997). The two estimators converge for dominant variables, but for variables with a weak relationship to the response $|\tilde{\bar{\beta}}_i| < |\hat{\bar{\beta}}_i|$ because $\tilde{\bar{\beta}}_i$ is a type of shrinkage estimator. The unconditional variance of $\tilde{\bar{\beta}}_i$ is estimated as

$$\hat{\text{var}}[\tilde{\bar{\beta}}_i] = \sum_{j=1}^{R} w_j \left[\hat{\text{var}}\left(\hat{\beta}_{ij} \mid g_j\right) + \left(\hat{\beta}_{ij} - \tilde{\bar{\beta}}_i\right)^2 \right] \tag{5}$$

where $\hat{\text{var}}(\hat{\beta}_{ij}|g_j) \equiv 0$ if β_i is not included in model j. The variance is given in Burnham and Anderson (2002, p. 345). Confidence intervals were computed here as $\tilde{\bar{\beta}}_i \pm 2.03\sqrt{\hat{\text{var}}[\tilde{\bar{\beta}}_i]}$. The value 2.03 is used here because it is the critical value of a t-distribution with 35 degrees of freedom which is approximately the model averaged degrees of freedom for the models in the model set. The exact critical value used here is of little importance because the degrees of freedom are large enough to produce a critical value near two regardless of the exact number of degrees of freedom used. The random number generation and analysis process was repeated 1,000 times. Achieved confidence interval coverage for each β_i was computed from the simulation replicates. Analysis was performed in SAS Proc IML (SAS Institute Inc. 2001). SAS code is available from the authors.

The same procedure of generating random numbers and performing first order linear regression on the data was also done using stepwise selection. Variables were included at $p \leq 0.15$ (Rawlings 1988). Variables were removed at $p > 0.15$. These α levels for inclusion and removal are the SAS default settings for stepwise selection. Due to the much more rapid computing of stepwise regression, 10,000 simulation replicates were performed. Confidence intervals are computed for β_i each time $\hat{\beta}_i$ is in a model selected as best. The analysis was performed in SAS Proc REG (SAS Institute Inc. 2001).

To further illustrate the issue of model selection bias, we generated 1,000 data sets with explanatory variables. Variable 1 was weakly correlated with the response,

$$Y = 0.05X_1 + \varepsilon, \varepsilon \sim N(0, 1) \tag{6}$$

The remaining seven variables were uncorrelated with the response. Each data set contained 40 observations. All first order linear regression models were fit to the data. The parameter estimates were model averaged, as above. In addition, stepwise selection was used to select a best model for comparison.

3 Results

While $\hat{\beta}_i$ is an unbiased estimate of β_i under both model averaging and stepwise selection in the situation where $\beta_i = 0$, the distributions of the $\hat{\beta}_i$ are quite different for the two methods (Fig. 1). The model averaged estimates are asymptotically normally distributed. The $\hat{\beta}_i$ from the single best stepwise model estimates have a bimodal distribution which excludes the true value of zero and this procedure frequently selects spurious variables (Freedman's paradox).

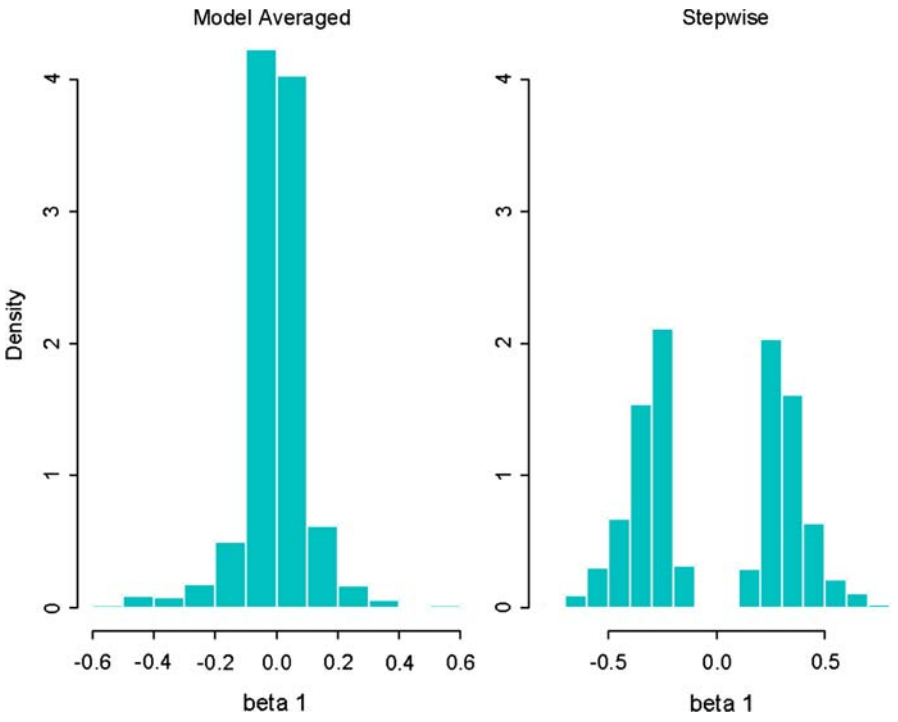

Fig. 1 Histograms of $\hat{\beta}_1$ scaled to densities for model averaging and stepwise selection for linear regression where true $\beta_1 = 0$

Table 1 Confidence interval coverage for model averaged parameter estimates from 1,000 Monte Carlo simulation replicates and stepwise selection from 10,000 simulation replicates

Parameter	Model averaging	Stepwise selection
β_1	0.963	0.444
β_2	0.957	0.403
β_3	0.948	0.416
β_4	0.961	0.406
β_5	0.958	0.419
β_6	0.944	0.407
β_7	0.967	0.431
β_8	0.965	0.413
β_9	0.956	0.425
β_{10}	0.959	0.421
β_{11}	0.954	0.431
β_{12}	0.960	0.435
β_{13}	0.953	0.405
β_{14}	0.949	0.419
β_{15}	0.972	0.404
β_{16}	0.963	0.425
β_{17}	0.955	0.409
β_{18}	0.963	0.417
β_{19}	0.958	0.405
β_{20}	0.960	0.447

For these simulation $n = 40$ and all $\beta_i \equiv 0$

Confidence interval coverage for the model averaged parameters is very near the nominal 95% level (Table 1). Confidence interval coverage for parameters selected by stepwise selection is poor, averaging only 41.9%.

For the linear regression analysis with a weak correlation, model selection bias is clearly evident for stepwise selection, $\hat{\mathbf{E}}[\bar{\hat{\beta}}] = 0.168$. When model averaging is used the average estimated parameter value is much closer to the true value of 0.05, $\hat{\mathbf{E}}[\bar{\tilde{\beta}}] = 0.031$. Bias divided by standard error is much larger for stepwise selection, $|(0.168 - 0.05)|/0.183 = 0.645$, than it is for model averaging, $|(0.031 - 0.05)|/0.159 = 0.119$.

4 Discussion

Model selection bias is often severe when making inference from an estimated best model, particularly when using stepwise selection. The bimodal distribution of parameter estimates of uncorrelated variables suggests standard confidence intervals are of limited use for inference about the parameter. The problem becomes worse for weakly correlated variables because a large bias begins to appear. Model averaging provides

much reduced bias in estimates of the parameters unrelated to the response variable and the parameter estimates have an approximately normal distribution.

Inference from parameter estimates and variances based on model averaged results provide a more realistic portrayal of the uncertainty in the estimates. If little is known about the system generating the data and little can be done to reduce the set of candidate models, then model selection uncertainty will be high, particularly if sample size is small. Given an analyst has a set of exploratory data with a sample size of the same order as the number of variables, the analyst is likely to make relatively few spurious conclusions if the $\tilde{\tilde{\beta}}_i$ model averaging estimator is used.

On the contrary, stepwise selection gives variances which are too small. Even more striking with stepwise selection is the arbitrary nature of the inclusion α level. We chose $\alpha = 0.15$ because it is the default in many statistical software packages. If we chose $\alpha = 0.05$ for inclusion and exclusion of a variable, 95% confidence interval coverage is 0% for the case of $\beta \equiv 0$.

Stepwise selection fares poorly when examining weakly correlated variables. In our example, the average estimated β is more than three times larger than the true value. More importantly, the bias in $\hat{\beta}$ relative to its standard error is large. In this situation, stepwise selection is not selecting the weakly correlated variable too frequently. Rather, it only selects the variable when it is greatly overestimated. Model averaging, while not perfect, provides better estimates and a more realistic variance. Interestingly, the model averaged variance is smaller than the stepwise selected variance, yet produces a confidence interval with better coverage because of the reduced bias. Therefore, model averaging was better here than stepwise selection in terms of both bias and coverage in the case of weakly correlated variables.

We compared model averaging to stepwise selection, but we wish to emphasize that it is not stepwise selection specifically that is central to the problem. The real issue is selecting a single best model and making inference only from that model when the size of the sample is small. Most methods will produce results similar to stepwise selection, we just chose stepwise as an example.

The use of AIC_c as a model selection criterion in this context is based on the philosophy that there are important variables to be found. Our use of no important variables in the example is intended to be an extreme case of Freedman's paradox and used to demonstrate the effectiveness of model averaging. When undertaking a study at least a few of the explanatory variables are expected to be related to the response variable. Model averaging based on AIC_c has been shown to be effective for variables that have a real effect on the response (Burnham and Anderson 2002). Anderson (2008, pp. 129–132) provides a detailed example.

Methods exist to perform model averaging in a Bayesian framework (Hoeting et al. 1999). It has been shown AIC and AIC_c can be derived as a Bayesian result and the Akaike weights are posterior model probabilities (Burnham and Anderson 2002, p. 303). This derivation is based on a specific prior model probability. With this relationship between the information-theoretic and Bayesian derivations of the model weights, both methods are likely to produce similar results. In this paper we wish to emphasize the importance of model averaging to reduce Freedman's paradox rather than to debate the relative merits of different model averaging methods.

An important issue exists in the definition of a particular β when other variables are included or excluded from a model. The parameter β_i is defined as the conditional relationship of X_i on Y given the other variables in the model are held constant. The estimated effect of β_i may change given the estimated values of other β_j in the model. One may therefore wonder about the legitimacy of averaging β_i cross models with different sets of parameters. Here we are addressing the case of a variable that is unrelated to the dependent variable. Therefore, it is guaranteed that the independent variable is pairwise orthogonal with all of the other independent variables. Given the variables are pairwise orthogonal, the estimated slopes will be the same regardless of what variables are in the model.

Our success in reducing the number of spurious effects and model selection bias is due to two reasons. First, the model averaged estimates provide estimates with less bias, while selection when using hypothesis testing procedures results in $\hat{\beta}_i$ being biased away from zero; on average, $\hat{\beta}_i$ is unbiased, but in any given selected model $\hat{\beta}_i$ is far from 0. With model averaging, the distribution of $\bar{\hat{\beta}}_i$ has a mode at zero. Second, the unconditional variances provide a better estimate of the uncertainty and therefore provide better confidence interval coverage. Part of the effectiveness of model averaging in this example is due to the use of AIC_c rather than AIC. AIC tends to select overfit models when sample size is small relative to the number of parameters in the model. Simply ranking models with AIC_c and not model averaging, or using a stepwise algorithm based on AIC, will produce similar results as the stepwise selection presented here. The critical step to reducing the number of spurious effects is to model average. Others are also finding that model averaging is advantageous (e.g. Burnham and Anderson 2004; Yang 2007; Wheeler and Bailer 2007; Wheeler 2009).

We do not advocate using all possible models when attempting to answer scientific questions. In research situations, some information is likely to be known from past research or theory to help develop a scientifically reasonable set of a priori candidate models. Then the data could be analyzed under this reduced set of models and model averaged inferences can be made. The a priori model set allows only reasonable models into the candidate set and it alleviates the computing burden in model averaging over all first order models.

Our analysis took a large amount of computing time because of the analysis of many simulation replicates. An analysis of a single data set can be done in a reasonable amount of time ($\sim$10 min). Model averaging is not out of reach of data analysts who need rapid results. Computing speeds continue to increase and this method easily lends itself to parallel processing. Therefore, a problem which is currently out of reach may be reasonably accomplished in the near future.

Our emphasis in this research is to demonstrate how information-theoretic methods and model averaging can help guard against spurious results when analyzing exploratory data and data sets with small sample sizes. Scientists often worry about failing to find an effect, but they should be equally concerned about how easy it is to, unknowingly, obtain a spurious effect. Freedman (1983, p. 152) contains a very telling statement: "To sum up, in a world with a large number of unrelated variables and no clear a priori specifications, uncritical use of standard methods will lead to models that appear to have a lot of explanatory power. That is the main—and negative—message

of the present note." We agree and add that model averaging substantially reduces this problem.

Acknowledgments U.S. Geological Survey/Biological Resources Division and the Colorado Division of Wildlife provided support for this work.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov, F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.

Akaike, H. (1978). On the likelihood of a time series model. *The Statistician, 27*, 217–235.

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika, 66*, 237–242.

Anderson, D. R. (2008). *Model based inference in the life sciences: primer on evidence*. New York: Springer.

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* (2nd ed.). New York: Springer.

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

Claeskens, G., Hjort, N. L. (2008). *Model selection and model averaging*. New York: Cambridge University Press.

Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician, 37*, 152–155.

George, E. I., McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association, 88*, 881–889.

Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with discussion). *Statistical Science, 14*, 382–417.

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.

Massart, P. (2007). *Concentration inequalities and model selection*. Berlin: Springer.

McQuarrie, A. D. R., Tsai, C.-L. (1998). *Regression and time series model selection*. Singapore: World Scientific Publishing Co.

Miller, A. J. (2002). *Subset selection in regression* (2nd ed.). New York: Chapman and Hall.

Rawlings, J. O. (1988). *Applied regression analysis: a research tool*. Belmont: Wadsworth, Inc.

Rencher, A. C., Pun, F. C. (1980). Inflation of R^2 in best subset regression. *Technometrics, 22*, 49–53.

SAS Institute, Inc. (2001). SAS version 8.02, Cary, NC.

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.

Wheeler, M. W. (2009). Comparing model averaging with other model selection strategies for benchmark dose estimation. *Environmetrics and Ecological Statistics, 16*, 37–51.

Wheeler, M. W., Bailer, A. J. (2007). Properties of model-averaged BMDLs: a study of model averaging in dichotomous response risk estimation. *Risk Analysis, 27*, 659–670.

Yang, Y. (2007). Prediction/estimation with simple linear models: is it really simple? *Econometric Theory, 23*, 1–36.