Behav Ecol Sociobiol (2011) 65:91–101
DOI 10.1007/s00265-010-1045-6

ORIGINAL PAPER

Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error

Robert P. Freckleton

Received: 1 January 2010 / Revised: 10 August 2010 / Accepted: 10 August 2010 / Published online: 14 September 2010

Abstract There has been a great deal of recent discussion of the practice of regression analysis (or more generally, linear modelling) in behaviour and ecology. In this paper, I wish to highlight two factors that have been under-considered, collinearity and measurement error in predictors, as well as to consider what happens when both exist at the same time. I examine what the consequences are for conventional regression analysis (ordinary least squares, OLS) as well as model averaging methods, typified by information theoretic approaches based around Akaike's information criterion. Collinearity causes variance inflation of estimated slopes in OLS analysis, as is well known. In the presence of collinearity, model averaging reduces this variance for predictors with weak effects, but also can lead to parameter bias. When collinearity is strong or when all predictors have strong effects, model averaging relies heavily on the full model including all predictors and hence the results from this and OLS are essentially the same. I highlight that it is not safe to simply eliminate collinear variables without due consideration of their likely independent effects as this can lead to biases. Measurement error is also considered and I show that when collinearity exists, this can lead to extreme biases when predictors are collinear, have strong effects but differ in their degree of measurement error. I highlight techniques for dealing with and diagnosing these problems. These results reinforce that automated model selection techniques should not be relied on in the analysis of complex multivariable datasets.

Keywords Regression · Model selection · Information theory

Introduction

In many respects, the 'gold standard' in hypothesis testing in behaviour, ecology and evolutionary biology is the randomised experiment, in which factors of interest are manipulated over a range of values. When examining the effects of different factors simultaneously, randomised experiments allow the effects of each of the variables examined to be isolated and measured individually through fully factorial designs (e.g. Grafen and Hails 2002; Ruxton and Colgrave 2002). The framework for statistical testing of data from designed experiments is extremely comprehensive and sophisticated (Sokal and Rohlf 1995).

In many situations, however, experimental approaches cannot be used and alternative methods are required. For instance, long-term monitoring (e.g. Leigh and Johnston 1994) and comparative analyses of data across groups of species (e.g. Harvey and Pagel 1991) are examples of commonly employed approaches to data gathering that do not usually use experimental methods. In general, observational approaches use data that are gathered passively without manipulation, and rely on natural variation in the variables of interest. If the natural variation in the system is large enough, then statistical analyses can be used to examine the effects of factors of interest. Statistical analyses

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).

R. P. Freckleton (✉)
Department of Animal and Plant Sciences,
University of Sheffield,
Sheffield S10 2TN, UK
e-mail: r.freckleton@sheffield.ac.uk

(usually regression analysis or linear modelling) are performed as if this variation had been created through experimental manipulation with the aim of determining underlying causal relationships.

The downsides of observational approaches are twofold. First, confounding variables may be responsible for generating observed patterns, which may lead to incorrect conclusions. For example, spatial or temporal autocorrelation (Haining 1990; Chatfield 1996), or phylogenetic non-independence (Felsenstein 1988; Harvey and Pagel 1991) are well-known confounding factors in the analysis of behavioural and ecological data.

The second problem is that in complex datasets with a range of predictors, there is frequent correlation between the predictors. For instance, in climate data from temperate regions it is often found that hot summer weather is accompanied by dry conditions, and hence rainfall is low when temperatures are high. Consequently, it is difficult to disentangle the effects of temperature and rainfall using data that are gathered under normal conditions. In regression analysis, collinearity of this sort among predictors can generate problems of analysis and interpretation. Thus, a variable of interest may correlate strongly with several predictors; however, if these predictors are correlated, the independent effects of each may be hard to disentangle (e.g. see Freckleton et al. 1998 for an example in a regression context). This problem is one that is difficult to address and to effectively deal with, and that I discuss in this paper.

One of the most straightforward ways to deal with collinear variables is to use a data reduction method such as principal components or factor analysis (Draper and Smith 1998; Quinn and Keough 2002). For instance, in the hypothetical example, above summer temperature and rainfall is the product of prevailing weather conditions and thus a single summary variable (e.g. the first principal component) may accurately represent the data. However, often the correlated variables may be expected to have independent effects: as a hypothetical example, plant growth increases with temperature, but decreases with decreasing rainfall. However, if temperature and rainfall are negatively correlated as outlined above, a single axis would not allow the countervailing effects of these two variables to be disentangled. An alternative approach uses diagnostics and adjustments based on propensity scores (Rosenbaum and Rubin 1983) to account for imbalances in observational studies when the assignment of observations amongst groups is non-random, potentially yielding biases akin to those resulting from collinearity in regression. For fitting regression models, other approaches exist such as ridge regression and various shrinkage techniques (Draper and Smith 1998; see below for model averaging based on Akaike's information criterion (AIC)-information theoretic (IT) which is a shrinkage method). These improve parameter estimates, or estimates of variance to account for collinearity when fitting single models. However, it is in fact the case (although not always widely appreciated) that least squares estimates of statistical model parameters are robust to moderate and even high levels of collinearity (Draper and Smith 1998). Estimates of parameter variance may, however, be very sensitive affecting hypothesis tests.

The problem of collinearity is particularly an issue in model selection (e.g. Grafen and Hails 2002). In model selection, the aim is to find a model with the best fit to data with not too many parameters. However, if predictors are correlated, models with different predictors may have similar fits to data. This is a problem that is particularly important when using automated techniques, causing them to identify suboptimal models as the 'best'.

For problems in model analysis in behaviour and ecology, interest has focussed on model averaging (Burnham and Anderson 1998, 2002; Rushton et al. 2004; Link and Barker 2006; Johnson and Omland 2004). Model averaging recognises that there are two forms of uncertainty in modelling. The first is the uncertainty in parameter estimates, for example measured by standard errors and confidence intervals for parameters. The second source of uncertainty is in the model: usually the 'true' model is unknown, and there is a probability that each candidate model is the 'true' model. This probability can be measured and incorporated as a source of uncertainty. There are many ways to perform model averaging; a recent comprehensive review is that of Claeskens and Hjort (2008). In ecology, the methods of Burnham and Anderson (1998, 2002) based on AIC have become widely used. Model averaging uses information on the fit of all models to data, not just the best-fitting model. The contribution of each model to the final analysis depends on its relative fit, with better fitting models playing a greater role than poor-fitting ones. However, the consequences of collinearity for model averaging methods are not clear.

Least squares methods should yield unbiased parameter estimates even in the presence of moderate amounts of collinearity. This is because least squares estimates can be shown to be the best linear unbiased estimators (BLUE) for a given model. It is possible to show with simulations (e.g. Freckleton 2002, below) that this means that ordinary multiple regression is relatively insensitive to collinearity, probably more than most researchers imagine. However, this result is dependent on the predictors being measured without error. If measurement error in predictors exists, biases result in parameter estimates with the bias increasing with the amount of measurement error (Carroll et al. 2006). If different levels of measurement error exist in collinear variables, this will affect the outcome. One might expect differences in measurement errors to be common: for instance, temperature will be proportionately much more accurately measured using a thermometer than rainfall is using a rain gauge.

In this paper, I wish to highlight two issues. First, that ordinary least squares (OLS) and model-averaging methods *can* perform well in the presence of even quite high levels of collinearity, with model-averaging methods outperforming OLS approaches under certain conditions. However, high or different levels of collinearity between predictors, or different measurement errors between these alter this conclusion, and yield problems for both methods: this is the real problem of collinearity in ecological data analysis. If measurement errors are appreciable, then these should be quantified and the possible effects can be modelled.

Methods

Linear model

Here, I consider a simple model for collinear data. The data ($\mathbf{y}$) consist of n observations of y_i and are modelled as a function of predictors $\mathbf{x}_1$ and $\mathbf{x}_2$. $\mathbf{x}_1$ and $\mathbf{x}_2$ are assumed to have been standardised. The *data model* is then:

$$\mathbf{y} = a + b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \mathbf{e}_y \quad (1)$$

The effects of the two predictors are assumed to be linear and are modelled by the parameters b_1 and b_2. Without losing generality I set a=0, but included this as an estimated parameter in the statistical models below (because it would be unknown in reality). $\mathbf{e}_y$ is an error term, where each observation has an associated error e_i. This is assumed to be normally distributed with zero mean and variance $(1-b_1)^2+(1-b_2)^2$. This is a standard linear model, and the mathematical theory for such models is well developed. In the simulations described below, I define β=b_2/b_1, i.e. β is the ratio of the effect of x_2 relative to that of x_1.

Collinearity

In order to model collinearity I assumed that the two predictors are correlated, according to a simple linear model:

$$\mathbf{x}^*{}_2 = r\mathbf{x}_1 + \mathbf{e}_x \quad (2)$$

The asterisk denotes that the value of this predictor is unstandardised. $\mathbf{x}_2$ was standardised before entering into Eq. 1. r is the correlation between the two predictors, and $\mathbf{e}_x$ is normally distributed with zero mean and variance $(1–r)^2$.

Observation model

Linear models assume that the predictors are measured with no error. In reality, this assumption is often not correct. In order to model parameter error, I denote $\mathbf{x}_1^{obs}$ and $\mathbf{x}_2^{obs}$ as the *measured* values of x_1 and x_2, respectively. These are then related to the true values by:

$$\begin{aligned}\mathbf{x}_1^{obs} &= \mathbf{x}_1 + \mathbf{e}_1\\ \mathbf{x}_2^{obs} &= \mathbf{x}_2 + \mathbf{e}_2\end{aligned} \quad (3)$$

where, $\mathbf{e}_1$ and $\mathbf{e}_2$ are normally distributed with zero mean and variances σ_1^2 and σ_2^2, respectively.

Model fitting

Two approaches to fitting models were contrasted. First a linear model including both predictors was fitted using OLS and estimated parameters were recorded from each simulation.

Second, an information–theoretic approach based on Akaike's information criterion was used (Burnham and Anderson 1998, 2002; Burnham et al. 2010). I used the methods described by Burnham and Anderson (2002). The approach compares the fits of a suite of candidate models using AIC. The absolute size of the AIC is unimportant; instead, the difference in AIC values between models indicates the relative support for the different models. In order to compare models, I calculated an "Akaike weight", w_i for each model. For a set of R models, the w_i sum to 1 and have a probabilistic interpretation: of these models, w_i is the probability that model i would be selected as the best fitting model if the data were collected again under identical circumstances.

Because the w_i are probabilities, it is possible to sum these for models containing given variables (Burnham and Anderson 2002). For instance, if one considers some variable k, one can calculate the sum of the Akaike weights of all the models including k, and this is the probability that of the variables considered, variable k would be in the best approximating model were the data collected again under identical circumstances.

Model averaging uses the average of parameter estimates or model predictions from each candidate model, weighted by its Akaike weight. For parameter b_j, the model averaged estimate was calculated as:

$$\overline{b}_j = \sum_{i=1}^{R} w_i \widehat{b}_{j,i}^{+} \quad (4)$$

In which, w_i is the Akaike weight of model i, and $\widehat{b}_{j,i}^{+}$ is the estimate of b_j if predictor j is included in model i, or is zero otherwise. In the simulation analysis, there are four models that are included in this model averaging procedure:

Model 1: $y=a$
Model 2: $y = a + b_1x_1$
Model 3: $y = a + b_2x_2$
Model 4: $y = a + b_1x_1 + b_2x_2$

In this example, the set of models was minimised; however in practice, the set may be expanded to consider interactions between variables, or more complex models containing additional predictors. A specific problem with doing this is that the problems of collinearity will be magnified by adding models including interactions between collinear variables, as the interaction term would necessarily also be highly collinear.

Simulations

I conducted a series of simulations to demonstrate the effects of collinearity and measurement error in the predictors on parameter estimates and their sampling variances. To examine the effect of collinearity, I set b_1 at a value of 0.5, a moderate effect. I set n=100 as this is a value typical of that used in many comparative analyses. The value of β was set at zero (no effect of predictor 2) or 1.5 (a stronger effect of predictor 2 than predictor 1). The correlation between the predictors, r, was then varied between 0 and 1. At each parameter combination, I conducted 10,000 simulations and in each case recorded the fitted parameter estimates, generated using the two methods described above.

To illustrate the impacts of measurement error, I repeated the above simulations but adding measurement error to predictor 1. I simulated error in this way as the aim was to demonstrate how the effect of unquantified error in one predictor could lead to mistaken inferences about the effects of other predictors. The measurement error standard deviation of $\mathbf{e}_1$ was set at 0.5.

Example data

In order to illustrate how model averaging and OLS methods might perform in a dataset containing collinearity, I performed an analysis of the "foxes" dataset from Grafen and Hails (2002). This is apparently a dataset on factors that influence overwinter survival in foxes, which is a function of average individual weight. Thirty groups were studied and data recorded on the size of the group, the availability of food, the area of each territory, as well as the average weight of foxes in each group. Two of the potential predictors, group size and food availability are strongly correlated with each other (r=0.88, P<0.0001), so that collinearity is an issue in this dataset. Models were fit and model-averaged parameter estimates generated as described above. Model-averaged estimates of parameter variances were generated using the formula in Burnham and Anderson (2002) and Anderson (2008).

Results

Collinearity and model averaging versus OLS

The broad difference in the performance of OLS and model averaging methods is illustrated with an example in Fig. 1. Figure 1 shows how increasing the degree of collinearity between the predictors affects sampling distributions for one set of parameter values in which the effect of one parameter is nil and the other has a stronger effect. If OLS is used to estimate model parameters, then the estimated mean value is, on average, unbiased, irrespective of whether the predictors are uncorrelated (Fig. 1a), moderately correlated (Fig. 1b) or strongly correlated (Fig. 1c). However, as is well known, the sampling variance is affected by collinearity, becoming large as the degree of collinearity increases.

Figure 1d–f shows what happens to the parameter estimates using the AIC-IT approach in this example. For zero and moderate levels of correlation, the parameter estimates are unbiased and the sampling distributions can be narrower than for those obtained using the OLS approach (Fig. 1d, e). However, when the correlation between the predictors is strong, the parameter estimates become biased, with the effect of the weak predictor being over-estimated and that of the strong predictor under-estimated (Fig. 1f)

These results are generalised in Figs 2 and 3. Figure 2 shows two cases, one where one predictor has zero effect and the other a stronger effect, as in Fig 1 (Fig. 2a, b). In the other case, both predictors have strong effects (Fig. 2c, d). As described above, parameter estimates are essentially unbiased using the OLS approach (Fig. 2a, c). When one of the predictors is weak, the parameter estimates become slightly biased for high collinearity when estimated using the AIC-IT approach (Fig. 2b). However, when the effects of both parameters are strong, the estimates obtained from the AIC-IT approach are unbiased and the general pattern is the same as that obtained using the OLS approach. The relationship between parameter estimates and sampling variance is summarised in Fig. 3. When the correlation between predictors is low to moderate, the sampling variance of estimates from the AIC-IT method is lower for the zero predictor than for the others (Fig. 3a). On the other hand, when both predictors have strong effects, the sampling variance is similar at low to moderate correlations for all methods.

The differences between the performance of the methods is relatively straightforward to understand. When the effect of x_1

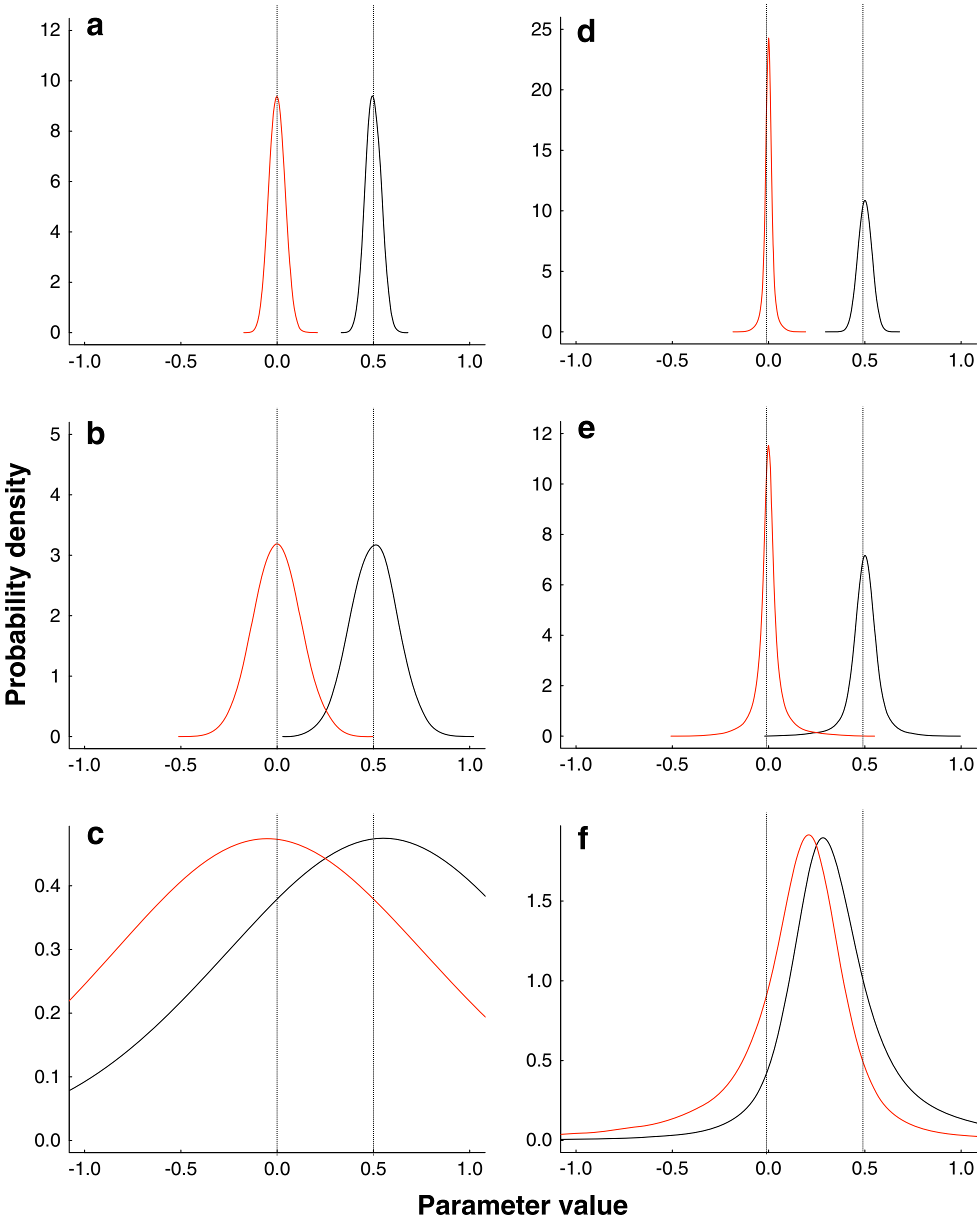


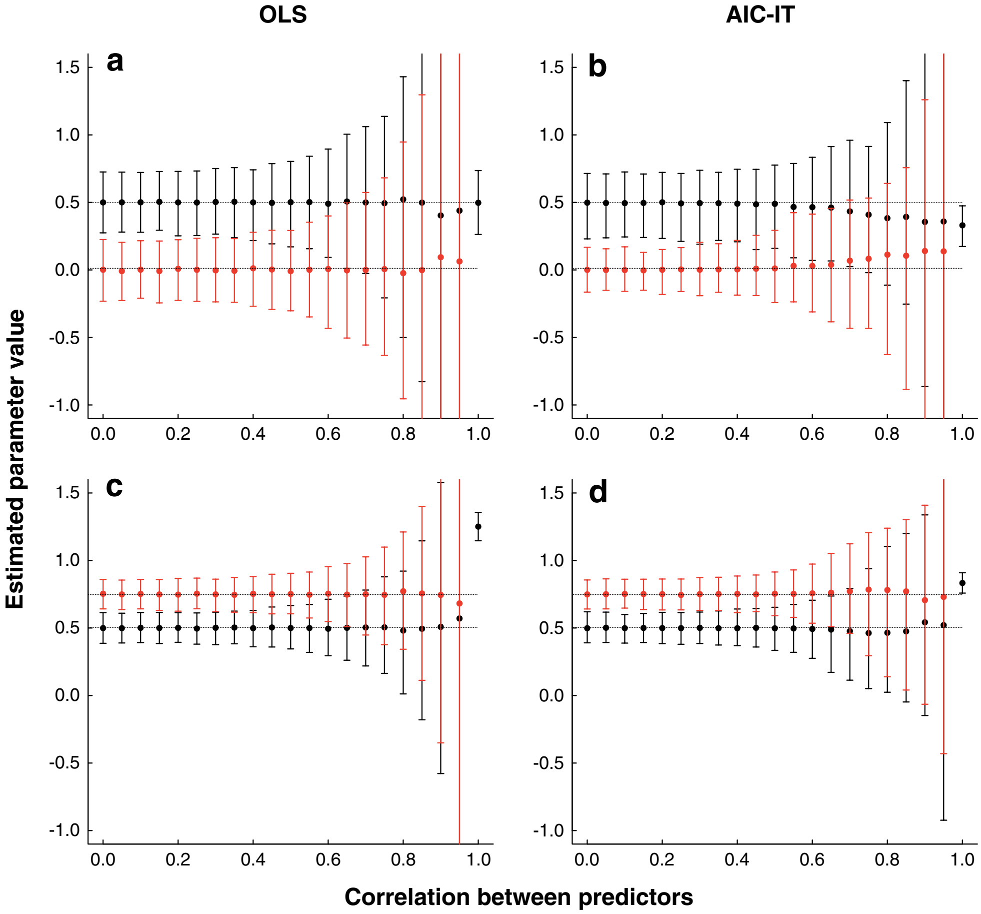
Fig. 1 Examples of the sampling distributions of parameters estimated using OLS (**a**–**c**) and AIC–IT (**d**–**f**). Data were generated according to a linear model with two covariates, x_1 and x_2, with each dataset containing 10 000 observations. The model is described in detail in the text. The values of the slopes were set as 0 for x_1 (*red*) and 0.5 for x_2 (*black*). The sampling distributions were generated from 10 000 replicates at each parameter combination. The simulations were repeated at different levels of *r*, the correlation between the predictions: *r* was set at 0 (**a**, **d**), 0.5 (**b**, **e**) and 0.95

is strong and that of x_2 is weak, model averaging tends to give higher weight to model 2 and low weights to models 1, 3 and 4. The estimates of b_2 in models 3 and 4 are downweighted (using the weighting scheme in Eq. 4) and consequently the estimates for this parameter are "shrunk" (see below for a discussion of shrinkage estimators) towards zero. The sampling variance for b_2 is thus lowered relative to the other models, particularly relative to model 4 which is the model fitted by OLS. As the correlation between the predictors increases, a problem arises because it becomes increasingly difficult to distinguish between models 2 and 3. This results in bias in the two parameter estimates (Figs 1d and 2b) as weight is given to the incorrect model (model 3).

When the effects of both predictors are strong, a high weight is given to model 4, and the others given low weights. This means that the model used for estimation by the AIC-IT method is basically the same model that is fitted by OLS.

A final small, but potentially important point, to emerge from Fig. 2 concerns the apparently aberrant points in Fig. 2c, d. The point in question is the estimate of the effect of x_1 when the two predictors are perfectly correlated (r=1). At this point because the two predictors are indistinguishable, only one parameter can be estimated. The estimate from OLS is ~1.25 (the sum of b_1 and b_2), for the AIC-IT method it is ~0.8. In the case of OLS, it is easy to see what is happening: because x_1 and x_2 are the same, the effect of x_1 is estimated to be the sum of the effects of the two variables. In general, it is easy to show that the slope estimated for a single predictor x_1

Fig. 2 Simulated parameter estimates and 95% percentiles for different fitting methods and predictor values. Data were simulated as described in the text, and parameters estimated using OLS (**a**, **c**) ort AIC-IT (**b**, **d**). The slope parameters were set as as 0 for x_1 (*red*) and 0.5 for x_2 (*black*) (**a**, **b**) or 0.75 for x_1 (*red*) and 0.5 for x_2 (*black*) (**c**, **d**)



used singly in preference to co-fitting with a second correlated predictor x_2 is b_1+rb_2. The bad consequences of this for practical regression analysis are discussed below.

Example data: model-averaging versus OLS

The analysis of the example dataset is summarised in Table 1. The full model for the data indicates that food (b=3.20, se=1.35, P=0.026) and group size (b=−0.69, se=0.15, P<0.0001) are both significant predictors of weight, whilst area is not important (b=0.35, se=0.22, P=0.12). Comparing the magnitude of the coefficients with their standard errors to estimate effect size, the effect of group size was much greater (4.6) than that of food (1.6). The overall model fit is moderate (R^2=0.41). As noted above, the collinearity between the predictors is strong (r=0.87). The model-averaged parameter estimate for food is a bit higher than that obtained using OLS (3.46 compared with 3.20 from the OLS model), whereas for group size the estimate is similar (−0.65 compared with −0.69). The model-averaged estimate for area is rather lower than that from the OLS model (0.22 compared with 0.35). The parameter variances are generally similar (0.23 vs. 0.22 for area; 1.38 vs. 1.35 for food; and 0.16 vs. 0.15 for group size). Although, based on theory, we would expect the variances to be lower for the model-averaged parameters, in practice this may be offset by some additional uncertainty resulting from the model selection process.

In this example, the OLS and model-averaged parameters are overall very similar indeed, despite the high degree of collinearity. In large part, this results from the overall moderate effects of the predictors on the response variable. Given that we would expect the model-averaged parameters to behave in a more stable manner under such a high level of collinearity, the similarity of the results should lead us to conclude (1) that the results of the OLS model are not hugely biased by the underlying collinearity in the data; (2) model selection should be a useful way to proceed in this dataset as the analysis indicates that the two best models contain the two collinear variables, but that the model containing the third (area) barely improves the relative fit to the data.

Measurement error in the face of collinearity

The worst effects of measurement error on estimation of regression parameters by OLS (essentially the same result would also be obtained from AIC-IT) are illustrated in Fig. 4. In Fig. 4 the two predictors have reasonably strong effects (b_1=0.75, b_2=0.5). There is measurement error in x_2 which has the consequence that, even in the absence of collinearity between the predictors, the effect of x_2 is under-estimated. This bias resulting from error in predictors is well known, and termed ‘attenuation’ (e.g. Carroll et al. 2006), however the effect does not seem widely appreciated.

The effects of this bias become extremely important when collinearity between the variables exists. As shown in Fig. 4b, c, collinearity results in the effect of x_2 being progressively underestimated, and the effect of x_1 over-estimated. What is happening is that the measurement error in x_2 results in under-estimation in the effect of this variable and, as the collinearity between the predictor increases, the effect of x_2 is mis-attributed to x_1.

The effect is generalised in Fig. 5. As shown in Fig. 5 there is no bias, unsurprisingly, when the effect of the predictor with error is zero (Fig. 5a, b), irrespective of the estimation method. However, problems arise when both variables have non-zero effects and the degree of collinearity increases. Irrespective of the estimation method employed, there is bias in the estimates of both parameters with the consequence that at moderate levels of collinearity (r>0.6) estimates of both parameters are very different from their true values. Indeed even at low values of r there is some bias because of statistcal attenuation, i.e., the measurement error reduces the estimate of the predictor with error in it.

How does this relate to the example data and analysis summarised in Table 1? Unfortunately, measurement error is not estimated for these data. However, intuitively one would expect that in a detailed behavioural study, group

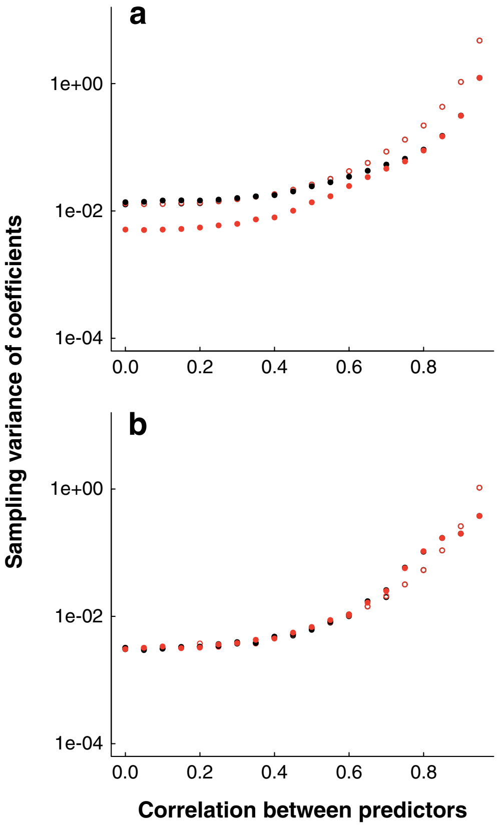


Fig. 3 The variances recorded for the simulations in Fig. 2. **a** The variances for the OLS (*open symbols*) and AIC-IT (*closed*) estimates of parameters when the slope parameters were set at 0 for x_1 (*red*) and 0.5 for x_2 (*black*) (**b**) as (**a**), but slopes were set at 0.75 for x_1 (*red*) and 0.5 for x_2 (*black*). Note that the variances are the same for most parameter values

Table 1 Analysis of a dataset containing multicollinearity using AIC-IT methods

Model	Intercept	se	Area	se	Food	se	Group Size	se	cAIC	w_i
0	4.59	0.12	–	–	–	–	–	–	64.86	0.00
1	4.41	0.49	–	–	0.25	0.68	–	–	67.19	0.00
2	5.05	0.37	–	–	–	–	–0.12	–	65.46	0.00
3	4.31	0.42	0.10	0.14	–	–	–	–	66.82	0.00
4	3.98	0.39	–	–	4.51	1.11	–0.66	0.15	53.77	0.46
5	4.44	0.49	0.22	0.29	–0.70	1.42	–	–	69.22	<0.001
6	4.49	0.35	0.66	0.19	–	–	–0.47	0.13	56.78	0.10
7	4.00	0.38	0.35	0.22	3.20	1.35	–0.69	0.15	53.86	0.44
Parameter estimate	4.04	0.41	0.22	0.23	3.46	1.38	–0.65	0.16		
Probability	>0.99		0.54		0.90		1.00			

The full details of the dataset are given in the text. Eight models were considered containing all combinations of the three predictors. The parameters were estimated for each model separately, shown together with standard errors (*se*) for each parameter in each model. Bias-corrected AIC values (cAIC) and model weights (w_i) were calculated and used to generate model averaged parameter estimates and inclusion probabilities for parameters (bottom two lines)

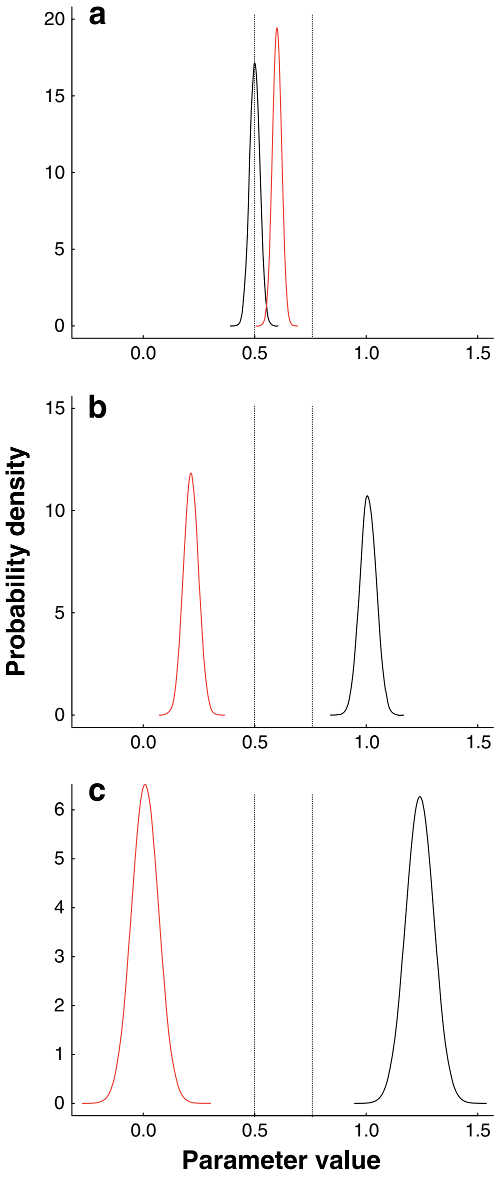


Fig. 4 Examples of sampling distributions of parameter estimates in data containing measurement error in one variable. Data were generated according to a linear model with two covariates, x_1 and x_2, with each dataset containing 10,000 observations. The model is described in detail in the text. The values of the slopes were set as 0.75 for x_1 (*red*) and 0.5 for x_2 (*black*). The sampling distributions of the OLS estimates were generated from 10,000 replicates at each parameter combination. The simulations were repeated at different levels of r, the correlation between the predictions: r was set at 0 (**a**), 0.5 (**b**) and 0.95 (**c**). x_1 was assumed to contain error with the standard deviation of the error set at 0.5. The true parameter values are indicated by the *vertical lines*

size would be more accurately measured than food availability as the former is based on direct counts of animals over a long period. Thus, without further information, one hypothesis is that the difference in magnitude of the effects of the predictors (group size is estimated to have a stronger effect than food availability) could be a consequence of different levels of measurement errors in the two predictors.

Discussion

Linear modelling with multiple predictors will always be an important and commonly used technique in behavioural and ecological research. There has been a great deal of debate about how such analyses are best conducted (e.g. Burnham and Anderson 1998, 2002; Garcia-Berthou 2001; Freckleton 2002; Link and Barker 2006) with major shifts in what is considered best practice (Rushton et al. 2004). The main assumptions of such analyses are well known, however the consequences of collinearity and measurement error for different techniques are only rarely examined. The key points I wish to make in this paper are that: (1) when using methods such as model averaging, there is a possibility that bias reduction by parameter shrinkage may yield incorrect results; (2) when measurement error and collinearity occur simultaneously, there are potentially severe problems for both estimation and hypothesis testing.

An important point to realise is that when all predictors have strong effects *OLS and AIC-IT will be largely dependent on the same model*. The methods differ most when predictors with weak effects are included: *AIC-IT is the more efficient method for reducing the parameter estimates for predictors with weak or no effects*, and this conclusion holds even in the face of weak to moderate collinearity. However, as noted in the previous paragraph, *AIC-IT methods will yield biased parameter estimates when predictors are highly correlated, particularly when their effects are rather different.*

Bias and variance

It is well known that there is a statistical trade-off between bias and variance in parameter estimates. This is exemplified in Fig. 1 in which the OLS parameter estimates are unbiased regardless of levels of collinearity; those from the model averaging method have lower variance, but begin to exhibit bias under high levels of collinearity. Conventional diagnostics such as variance inflation factors allow the extent of this effect to be estimated. The AIC-IT parameter estimates have a lower sampling variance for weak predictors because estimation is more heavily weighted to models that do not incorporate the weak predictors. The model averaged parameter estimates Eq. 4 are known as shrinkage estimators, because these have sampling distributions ‘shrunk’ back towards zero. This property of model-averaged parameters is discussed in detail by Burnham and Anderson (2002). The important point here is that under low to moderate levels of collinearity the

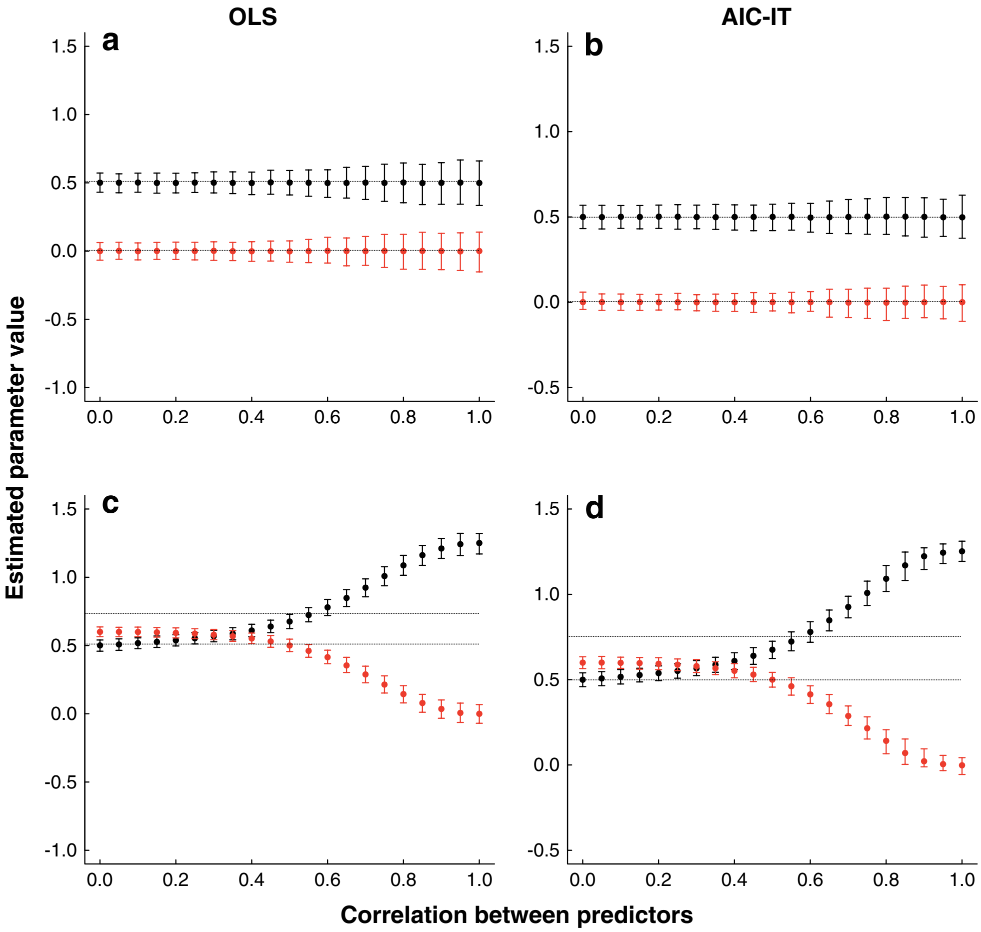


Fig. 5 Effects of correlation between predictors and measurement error in predictors on parameters estimated using OLS (**a**, **c**) and AIC-IT (**b**, **d**). Data were generated according to a linear model with two covariates x_1 and x_2, with x_2 being measured with error. Each simulated dataset containing 10,000 observations. The model is described in detail in the text. The values of the slopes were set as 0 for x_1 (*red*) and 0.5 for x_2 (*black*) in (**a**) and (**c**) and 0.75 for x_1 (*red*) and 0.5 for x_2 (*black*) in (**b**) and (**d**)

AIC-IT estimates have lower variance than OLS and thus may be preferable.

What to do with collinear variables?

The usual advice with a pair of collinear variables is to combine them in some way or to eliminate one or the other (see introduction). The results above reveal that this may not always be the best course of action. In the analysis, it was noted when x_2 is absent from the model, the slope for the effect of x_1 would be b_1+rb_2, where r is the correlation between x_2 and x_1. Thus, the effect of x_1 would be systematically over or under-estimated, depending on the sign of the correlation between the predictors and the signs of the slopes. The consequence of this is that the way to deal with collinear variables will depend on their nature and interpretation. In the hypothetical example in the introduction, rainfall and temperature would be expected to be negatively correlated with each other because hot conditions are usually associated with low rainfall. Low rainfall depresses growth of plants, whereas high temperatures promote growth. In this instance, it would not make sense to combine the two variables, or to omit one of them, even if they are highly correlated. Doing so would risk underestimating the effects of the included variable and of mismodelling the underlying determinants of growth. On the other hand, if the dataset contained both rainfall and soil moisture as predictors, these two variables are simply different ways of measuring the same quantity, i.e. the amount of water available. As a consequence, it would seem sensible to combine these, or just include one or the other.

If collinearity exists in a dataset, then there is no justification at all for using automated regression selection, such as stepwise regression. These techniques are widely criticised in the statistical literature and beyond because they result in biased parameter estimates with degenerate sampling distributions and a high probability of Type I errors (Chatfield 1996; Burnham and Anderson 1998, 2002; Whittingham et al. 2006). Collinearity reflects important structure in the data, which needs to be understood and dealt with explicitly. Although some authors have suggested ways in which Stepwise methods may be amended to deal with such issues (e.g. Hegyi and Garamszegi 2010) such suggestions are computational kludges, and ignore the wider problems (e.g. see Burnham et al. 2010). In most cases that researchers believe they should be conducting selection (e.g. using stepwise methods as envisaged by Hegyi and Garamszegi (2010), they would probably be better advised to use a full model (Whittingham et al. 2006; Forstmeier and Schielzeth 2010) and the results here largely endorse that conclusion.

AIC-IT methods are generally robust to collinearity. However, some problems can arise when predictors are highly correlated, particularly when their effects are rather different. The most extreme problem arises when one predictor has a weak effect, but is strongly correlated with another which has a strong effect. This situation can easily arise in real data. In a behavioural ecological context, for example, Székely et al. (2004) found a strong correlation between sexual dimorphism in body size and dimorphism in bill size in shorebirds. However, when compared with other variables in the dataset, the underlying correlates of these two measures of size were rather different. If not recognised, this type of variation can lead to substantial bias in both parameters (e.g. see Fig. 1f). These methods should be used with caution in such cases. In the example discussed above, one pragmatic fix would have been to omit models 2 and 3 from consideration when x_1 and x_2 are highly correlated. The justification for doing this would be that the models are essentially indistinguishable. The relative fits of model 1 and model 4 would allow the effects of the two predictors to be measured together and contrasted with a model including neither and should yield unbiased parameter estimates. In this example, this would essentially be the same as conducting an OLS analysis.

Other techniques exist for analysing data which contain collinearity. Ridge regression is one such approach (e.g. Draper and Smith 1998). This is a method that was developed with a view to allowing parameter estimation in cases if the collinearity between predictors is so extreme that the normal equations used to solve OLS problems contain a singular cross-product of the predictors. The method works by adding a parameter that modifies the normal equations to reduce the inflation of variance that results from collinearity. The resultant parameter estimates will be biased; however, this bias is traded off against reduced variance in parameter estimates. The downsides of the method are that the choice of parameter is arbitrary (although there are diagnostics that can be employed). The technique is most useful in generating predictions from a given model, and is less useful for comparing a suite of models.

Measurement error in predictors

Measurement error in predictors is almost never quantified or dealt with. This is despite the known issues with measurement error in regression models (Carroll et al. 2006) and the consequences for inference and estimation in practical analyses. In one of the few analyses to have addressed this, Linden and Knape (2009) showed that measurement error in predictors can result in the underestimation of environmental impacts on population dynamics, so the consequences for not estimating this error are demonstrably important.

In principle, if the measurement error in predictors has been quantified this can be dealt with. Several techniques exist for doing this, including simulation extrapolation (SIMEX; Cook and Stefanski 1994; Stefanski and Cook 1995), Bayesian methods (Fox and Glas 2003), multi-level models (Goldstein 1995), expectation maximization (Schafer 1987) or likelihood methods (Carroll et al. 1984). The technology exists with which to deal with measurement error; the limitation is that measurement errors in data are rarely quantified. The need to quantify measurement error has been emphasised in the ecological literature in the context of population modelling (Shenk et al. 1998; Ellner et al. 2002; Dennis et al. 2006; Freckleton et al. 2006; Linden and Knape 2009) and it is increasingly appreciated that this is an important component of variability in data that needs to be accounted for.

The simulations reported in Figs 4 and 5 were designed to illustrate that if measurement error differs between collinear predictors, then the consequences for estimation and inference can be particularly severe. If predictors are correlated with each other, but have different effects on the response variable, then differences between in the degree of measurement error can lead to extremely biased results. This is by no means a contrived situation: for instance, the example of plant growth given above is one in which two strongly negatively correlated drivers (temperature and rainfall) can have contrasting effects for the same underlying process (hot weather leads to a positive effect of temperature positive and a negative effect of rainfall). The simulation results indicate that if this is the case, then the resultant models can yield incorrect parameter estimates, the inference based on those parameters is wrong, and parameters have low sampling variance. Because of measurement error, the underlying correlation between the predictors would not be identified and the likely problem not diagnosed.

Dealing with this issue is difficult, and obviously impossible if measurement error has not been quantified. The main recommendation is that, assuming error has been estimated, one should proceed with caution if the relative level of error differs greatly between predictors, even if the level of collinearity is low. This is because measurement error can mask correlations between the variables and because differences in the level of error will be manifest in the relative values of slopes for the predictors.

Conclusions

Linear models and multiple regressions are extremely powerful tools, especially when combined with large datasets. The downside is that frequently data are generated non-experimentally and we are reliant on natural variation in observational data. The price to pay is that frequently we do not understand the structure of the data, and that correlations between variables and their error structure can complicate analyses. The statistical tools exist to deal with such complexity. However, we need to be aware of the possible pitfalls and of how they can be diagnosed.

Acknowledgements The author is funded by a Royal Society University Research Fellowship. Thanks to Tom Webb, the referees and László Garamszegi for comments on an earlier version of the MS.

References

Anderson DR (2008) Model-based inference in the life sciences. Springer, New York

Burnham KP, Anderson DR (1998) Model selection and multimodel inference. Springer, Berlin

Burnham KP, Anderson DR (2002) Model selection and multimodel inference. Springer, Berlin

Burnham KP, Anderson D, Huyvaert K (2010) AICc model selection in ecological and behavioural science: some background, observations and comparisons. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1029-6

Carroll RJ, Spiegelman CH, Gordon Lan KK, Bailey KT, Abbott RD (1984) On errors-in-variables for binary regression models. Biometrika 71:19–25

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu C (2006) Measurement error in nonlinear models: a modern perspective. Chapman & Hall, London

Chatfield C (1996) The analysis of time series. Chapman & Hall, London

Claeskens G, Hjort NL (2008) Model selection and model averaging. Cambridge University Press, Cambridge

Cook JR, Stefanski LA (1994) Simulation-extrapolation estimation in parametric error models. J Am Stat Soc 89:1314–1328

Dennis B, Ponciano JM, Lele SR, Taper ML, Staples DF (2006) Estimating density dependence, process noise and observation error. Ecol Monogr 76:323–341

Draper NR, Smith H (1998) Applied regression analysis. Blackwell Scientific, Oxford

Ellner SP, Seifu Y, Smith RH (2002) Fitiing population dynamic models to time-series data by gradient matching. Ecology 83:2256–2270

Felsenstein J (1988) Phylogenies and quantitative characters. Ann Rev Ecolog Syst 19:445–471

Forstmeier W, Schielzeth H (2010) Cryptic multiple hypothesis testing in linear models: overestimated effect sizes and the winner's curse. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1038-5

Fox J-P, Glas C (2003) Bayesian modelling of measurement error in predictor variables using item response theory. Psychometrika 68:169–191

Freckleton RP (2002) On the misuse of residuals in ecology: regression of residuals versus multiple regression. J Anim Ecol 71:542–545

Freckleton RP, Watkinson AR, Thomas TH, Webb DJ (1998) Yield of sugar beet in relation to weather and nutrients. Agric For Meteorol 93:39–51

Freckleton RP, Watkinson AR, Green RE, Sutherland WJ (2006) Census error and the detection of density dependence. J Anim Ecol 75:837–851

Garamszegi LZ (2010) Information-theoretic approaches in statistical analysis in behavioural ecology: an introduction. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1028-7

Garcia-Berthou E (2001) On the misuse of residuals in ecology: testing regression residuals vs. the analysis of covariance. J Anim Ecol 70:708–711

Goldstein H (1995) Multilevel statistical models. Eward Arnold, London

Grafen A, Hails R (2002) Modern statistics for the life sciences. Oxford University Press, Oxford

Haining R (1990) Spatial data analysis in the social and environmental sciences. Cambridge University Press, Cambridge

Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford

Hegyi G, Garamszegi LZ (2010) Using information theory as a substitute for stepwise regression in ecology and behavious. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1036-7

Johnson JB, Omland KS (2004) Model selection in ecology and evolution. Trends Ecol Evol 19:101–108

Leigh RA, Johnston AE (1994) Long-term experiments in agricultural and ecological science. In CAB International, Wallingford

Linden A, Knape J (2009) Estimating environmental effects on population dynamics: consequences of observation error. Oikos 118:675–680

Link WA, Barker RJ (2006) Model wieghts and the foundations of multimodel inference. Ecology 87:2626–2635

Quinn G, Keough M (2002) Experimental design and data analysis for biologists. Cambridge University Press, Cambridge

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55

Rushton SP, Ormerod SJ, Kerby G (2004) New paradigms for modelling species distributions. J Appl Ecol 41:193–200

Ruxton GD, Colgrave N (2002) Experimental design for the life sciences. Oxford University Press, Oxford

Schafer DW (1987) Covariate measurement error in generalized linear models. Biometrika 74:385–391

Shenk TM, White GC, Burnham KP (1998) Sampling variance effects on detecting density dependence from temporal trends in natural populations. Ecol Monogr 68:445–463

Sokal RR, Rohlf FJ (1995) Biometry. W.H. Freeman & Co., New York

Stefanski LA, Cook JR (1995) Simulation extrapolation: the measurement error jackknife. J Am Stat Assoc 90:1247–1256

Székely T, Freckleton RP, Reynolds JD (2004) Sexual selection explains Rensch's rule of size dimorphism in shorebirds. Proc Natl Acad Sci 101:12224–12227

Whittingham MJ, Stephens PA, Bradbury R, Freckleton RP (2006) Why do I still use stepwise regression? J Anim Ecol 42:270–280