Multiple Regression Is Not Multiple Regressions: The Meaning of Multiple Regression and the Non-Problem of Collinearity

Michael B. Morrissey* and Graeme D. Ruxton†

Simple regression (regression analysis with a single explanatory variable), and multiple regression (regression models with multiple explanatory variables), typically correspond to very different biological questions. The former use regression lines to describe univariate associations. The latter describe the partial, or direct, effects of multiple variables, conditioned on one another. We suspect that the superficial similarity of simple and multiple regression leads to confusion in their interpretation. A clear understanding of these methods is essential, as they underlie a large range of procedures in common use in biology. Beyond simple and multiple regression in their most basic forms, understanding the key principles of these procedures is critical to understanding, and properly applying, many methods, such as mixed models, generalised models, and causal inference using graphs (including path analysis and its extensions). A simple, but careful, look at the distinction between these two analyses is valuable in its own right, and can also be used to clarify widely-held misconceptions about collinearity (correlations among explanatory variables). There is no general sense in which collinearity is a problem. We suspect that the perception of collinearity as a hindrance to analysis stems from misconceptions about interpretation of multiple regression models, and so we pursue discussions about these misconceptions in this light. In particular, collinearity causes multiple regression coefficients to be less precisely estimated than corresponding simple regression coefficients. This should not be interpreted as a problem, as it is perfectly natural that direct effects should be harder to characterise than univariate associations. Purported solutions to the perceived problems of collinearity are detrimental to most biological analyses.

Keywords

regression • multiple regression • collinearity • ordinary least squares • linear model • causal effect • correlation

*Dyers Brae House, School of Biology, University of St. Andrews, St. Andrews UK, KY16 9TH
michael.morrissey@st-andrews.ac.uk
†graeme.ruxton@st-andrews.ac.uk

Received 10 February 2018; Revised 5 May 2018; Accepted 7 May 2018
doi:10.3998/ptpbio.16039257.0010.003

1 Introduction

Simple and multiple regression are two of the most-used statistical procedures in biology. Statistical results from both procedures are commonly interpreted as metrics of the degree of relationship between (sometimes multiple) explanatory and response variables. This rough interpretation may generally be satisfactory for simple regressions, i.e., models involving only one explanatory variable. However, this interpretation can lead to confusion for multiple regression, where the coefficients of a multiple regression model measure something subtly but crucially different. In this paper we first discuss what multiple regression is, what coefficients arising from multiple regression analysis are, and how the purpose of multiple regression is different from that of simple regression. We then build on this to clear up some misleading advice that we consider prevalent on the consequences of (multi)collinearity, i.e. of correlations among explanatory variables.

Our primary goal is to aid empiricists in understanding what aspects of the literature on collinearity should be considered when conducting analyses of data sets wherein correlations occur among predictor variables. We argue that the biostatistical literature contains a great deal of advice that could be leading empiricists to believe that they are seriously violating assumptions of their models, when in fact they are not. We hope that by promoting more understanding of how this is the case, many analyses of biological data can be better understood. Furthermore, it may be possible that properly understanding multiple regression and collinearity will reduce the use of complex and also counterproductive perceived solutions to non-existent problems.

2 What are simple and multiple regression analyses?

Regression is a set of statistical methods that can be used to estimate functions by which a response variable is related to one or more explanatory variables. In simple regression, expected values of a response variable are described as a linear function of a single explanatory variable; in multiple regression, expected values of a response variable may be described by multiple explanatory variables.

The key statistic of simple regression is the slope of the fitted linear function. The interpretation of this slope is that groups of observations that differ by one unit in the explanatory variable differ in the means of their response variable by the value of the slope. If there is a reason to believe that no quantities that are correlated with the explanatory variable also influence the response variable, then a causal interpretation of a simple regression model's slope is possible. Such an interpretation may hold, for example, in an experiment where only the value of the explanatory variable is varied across experimental treatments. More rarely, in observational data, such a causal interpretation may be possible if we know a lot of background information, and can make use of that information to make an argument about low influence of such third variables on the response variable. Pearl et al. (2016) provide an accessible discussion of the conditions under which direct effects estimated by multiple regression may be interpreted as causal effects. The causal interpretation of simple regression is that increasing the value of an explanatory variable by one unit will increase the value of the response variable by the value of the slope. Another way to interpret simple regression is as a measure of the direct effect of a single explanatory variable on the response variable under the assumption that there are no strong indirect influences through any third variables. These interpretations require that a linear function provides a reasonable model of the relationship between the explanatory variable and the response variable.

Multiple regression simultaneously describes the direct effects of multiple explanatory vari-

ables on a response, accounting for the direct effects of each explanatory variables included in the model. These estimated direct effects are called partial regression coefficients.

The distinction between simple and partial regression coefficients is probably best demonstrated by an example. Consider two correlated explanatory variables, x_1 and x_2, one of which, x_1, influences a response variable, y, while the other, x_2, does not. Both explanatory variables will be correlated with the response variable; for both, there will be a non-zero regression slope in a simple regression. That is, the simple regression of y on x_2 will have a non-zero regression slope even though x_2 has no direct effect on y, because x_2 is correlated with x_1 and x_1 does have a direct effect on y.

The signs of the simple regression coefficients will depend on the correlation between x_1 and x_2, and the effect of x_1 on y. This situation is depicted in Figure 1, for a correlation of 0.5 between the explanatory variables, and an effect of x_1 on expected values of y of 0.5. Figure 1 parts (a) and (b) show the simple regressions of y on x_1 and x_2, respectively. Figure 1c depicts the multiple regression of y on x_1 and x_2. y is only affected by x_1, but because x_1 and x_2 are correlated, large values of x_2 are associated with large values of x_1, which in turn are linked to large values of y. Two simple regressions of y on x_1 and of y on x_2 would imply that high values of both these explanatory variables are associated with high values of y. In contrast, multiple regression simultaneously recovers the direct effect of y on x_1, and also that x_2 only has an indirect association with y through its correlation with x_1.

The example scenario in Figure 1 reflects just one of the patterns where simple and partial regression coefficients can differ. Another example may be useful before constructing a careful, but simple, statement of what partial regression coefficients are. Figure 2 shows how an explanatory variable that is uncorrelated with a response variable may nonetheless have a direct effect on the response variable. As in Figure 1, this example also has a correlation between x_1 and x_2 of 0.5. However, in this time we simulated y values with expectation $0.5x_1 - 0.25x_2$. In this scenario, the effect of x_1 on y, in combination with the positive correlation of x_1 and x_2, is a source of positive covariance between x_1 and y. However, the negative effect of x_2 on y ultimately has an equivalent negative effect on the covariance of x_2 and y, resulting in no correlation of x_2 and y. Consequently, the simple regression slope of y on x_2 is zero (Fig. 2b). However, a multiple regression of y on x_1 and x_2 (Fig. 2c) is able to recover the full picture of how both x_1 and x_2 influence y.

Partial regression coefficients thus represent very different quantities from simple regression coefficients. These quantities do not represent improved or more sophisticated inferences about slopes than are provided by simple regression. Rather, partial regression coefficients are distinct quantities that will typically correspond to different biological questions from those that can best be explored with simple regression. Partial regression coefficients represent the difference in a response variable, per unit difference in a specified explanatory variable, holding all other explanatory variables constant (despite the fact that the other explanatory variables may covary and so are not themselves constant if regressed on one another). A causal interpretation of a given partial regression coefficient is possible, under the assumption that all quantities that both (i) are correlated with the given explanatory variable and (ii) have direct effects on the response are included as explanatory variables in the model. The causal interpretation is then that a partial regression coefficient represents how much the response variable would increase, per unit increase of a given explanatory variable, if that explanatory variables value were increased while holding the values of all other explanatory variables constant. For example, imagine that individual reproductive success has been regressed on both body-length and body-mass in a multiple regression analysis. The partial regression of reproductive success on mass would represent how much reproductive success would increase if the mass of an individual were increased by one

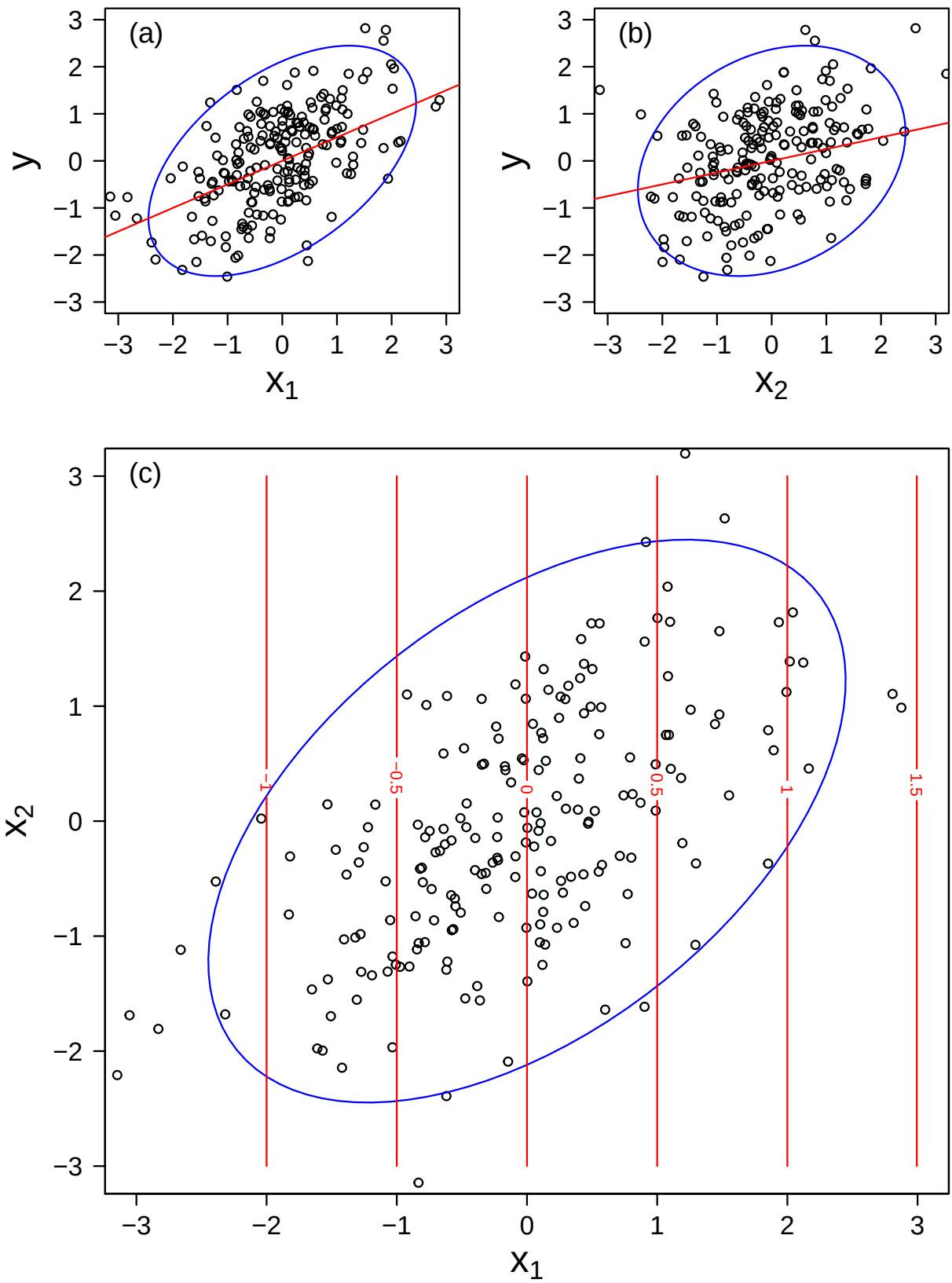

Figure 1: Multiple regression example scenario 1. When two explanatory variables (x_1 & x_2) are both correlated with a response variable (y; depicted in parts a and b), it is possible that only one directly affects the response (c). In (c), contour lines (red) indicate expected values of the response variable (i.e., y in parts a and b) corresponding to the range of joint values of the explanatory variables.

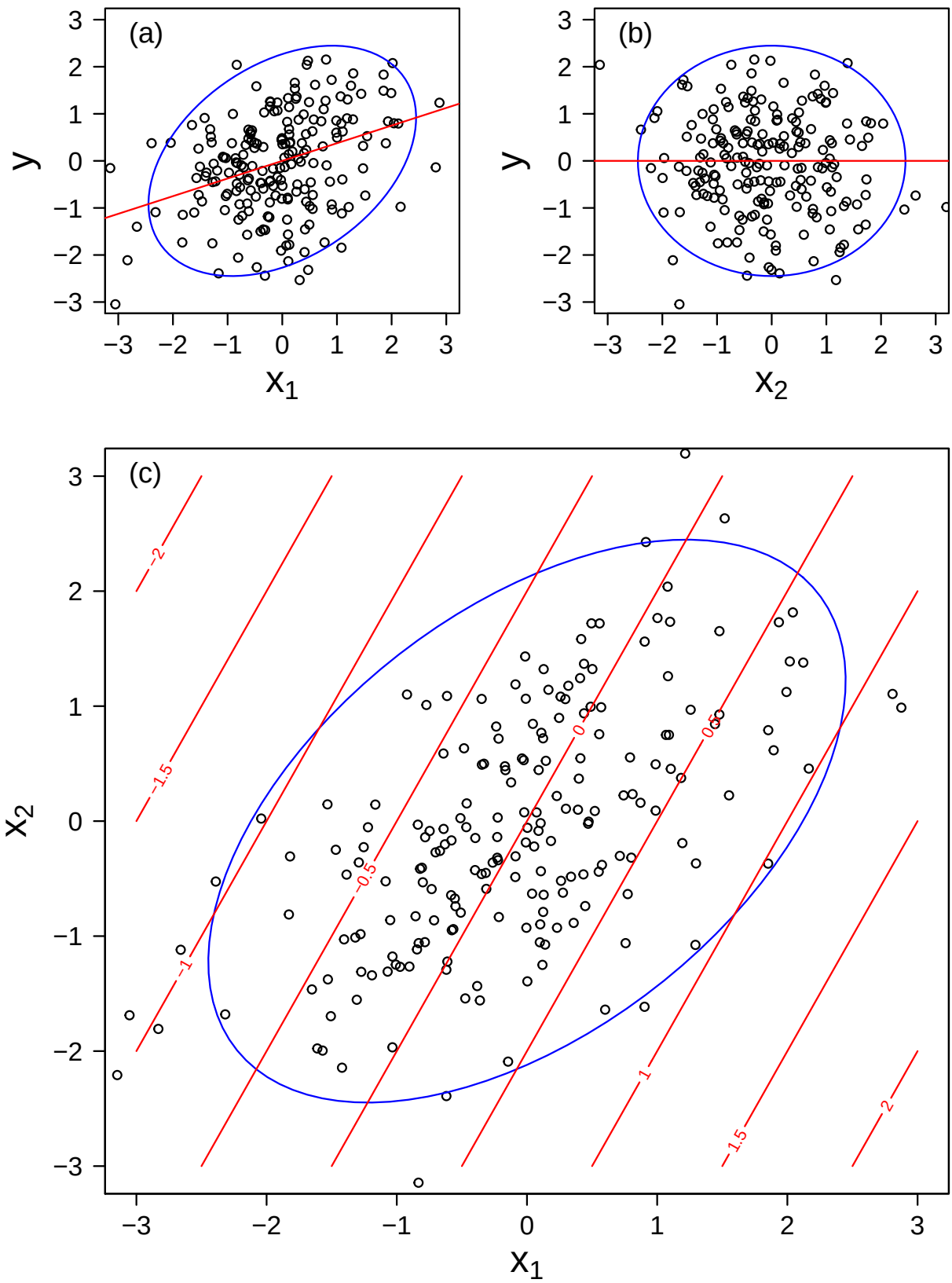

Figure 2: Multiple regression example scenario 2. An explanatory variable (x_2) that has little or no correlation with a response variable (y), e.g., (b), the explanatory variable may nonetheless have a direct effect on the response (c). As in Figure 1, contour lines in (c) indicate expected values of the response variable (i.e., y in parts a and b) corresponding to the range of joint values of the explanatory variables.

unit, while its length were held constant, regardless of whether or not a real individual that were heavier would also be longer.

3 Correlations among explanatory variables: information or problem?

Collinearity is a non-zero correlation among explanatory variables. An understanding of how simple and multiple regression are different analyses can be used to clarify some apparent misunderstandings about multi-collinearity.

To streamline our presentation, we avoid point-by-point accounting of where and how the principles we discuss conflict with existing information about (multi)collinearity in the literature. Each of the misunderstandings we discuss is explicitly expressed in multiple sources. Recent and/or influential sources with which material we present conflicts include statistics textbooks aimed at biologists (e.g., Legendre and Legendre 1998, Crawley 2007, Zur et al. 2007 and Quinn and Keough 2002), reviews (e.g., Graham 2003, Dormann et al. 2013 and Zur et al. 2010), and methodological primary research articles (e.g., Ray-Mukherjee et al. 2014 and Prunier et al. 2015). We believe that our readers will see that statistical advice contained in these and other sources is based on a range of misunderstandings, and that these can be rectified by considering the meaning of partial regression coefficients.

We focus primarily on two properties of partial effects in multiple regression models: the ability to make unbiased estimates, and the ability to make credible statements about the uncertainty in those estimates. We treat bias formally; i.e., when we discuss bias of an estimator, we are referring to the difference between the expected value of the estimator (in our case, this will be and ordinary least squares estimate of a simple or partial regression coefficient) and the true value of its estimand (the true value of the regression coefficient). Our discussion of statements about uncertainty (e.g., standard errors) focus on whether they correctly portray the (un)certainty in estimates of simple and partial regression coefficients.

Misunderstanding 1: *Collinearity is a potential problem*

If we seek tests of the overall relationship different individual explanatory variables have with the response, then indeed correlations among variables in a multiple regression model will mean that we cannot interpret partial regression coefficients, or their statistical significance, as indicators of this overall relationship. For example, in the simple example depicted in Figure 2, there is no overall relationship between x_2 and y, but multiple regression does not reflect this. However, exploring the overall relationship between one of the explanatory variables and the response variable is not the purpose of multiple regression; it is the purpose of simple regression. Multiple regression is not a system of automating multiple individual tests for relationships, nor is it a system for controlling for multiple tests for relationships. As we have seen (Figures 1 and 2) the purpose of multiple regression is to determine, among multiple correlated explanatory variables, what the direct effects of each is on the response. Correlations among predictor variables are not a problem for multiple regression; rather, they are part of the information that multiple regression uses to infer direct effects, as opposed to overall relationships (see Appendix 1 for more on this).

Our treatments of other, specific misunderstandings are all essentially elaborations on this argument. We contend that the behaviour of multiple regression analysis when predictors are correlated is perfectly natural, provided that we clearly understand that multiple regression—and the partial regression estimates that it generates—is distinct from simple regression.

Misunderstanding 2: *Collinearity leads to misleading inferences because it causes parameter estimates to change—even to change sign—when other model terms are added to, or removed from, a model*

This change in estimated coefficients is commonly seen as an indication that a model is unreliable. However, again, this concern disappears when the distinction between simple and partial regression coefficients is considered. It is perfectly reasonable that simple and partial regression coefficients for a given explanatory variable should have different signs, and indeed such a difference conveys valuable information that aids biological interpretation. Such a pattern indicates that the contributions of correlated explanatory variables to the covariance of a given focal explanatory variable with the response are greater than, and of opposite sign to, the direct effect. This same process causing differences between simple and partial regression coefficients can occur for comparisons between multiple regression models with different sets of explanatory variables. It does not mean that the method is unreliable. Rather, it means that comparison between different models provides an opportunity to understand potentially interesting reasons why explanatory variables covary with the response variable.

Misunderstanding 3: *Collinearity inflates p-values and standard errors*

"Inflating" is not a synonym for "increasing": inflation of standard errors suggests that that some methodological failing means that they are larger than they should be to correctly describe the level of uncertainty in predictions from the data. Correlations among predictor variables do not inflate either standard errors or p-values. Determining which of multiple predictor variables have direct effects on a response variable is a harder problem than determining whether each explanatory variable, individually, is related to the response variable. Consequently, we should not be surprised that, for a given data set, inference of direct effects is more challenging than inference of pairwise relationships. Consider a situation where a multiple regression of y on x_1 and x_2 suggests that neither partial regression coefficient is statistically significant, but one or both slopes of the associated simple regressions are significantly different from zero. This does not imply the collinearity has inflated type II errors in the multiple regression. Rather, it reflects the statistical and biological reality that direct effects considered in multiple regression are more difficult to characterise empirically than overall effects studied in simple regression.

We can show that statements about statistical uncertainty in estimates of partial regression coefficients remain valid under collinearity. The simplest approach is by simulation. We repeated the exercises that generated Figures 1 and 2, but with correlations among the explanatory variables x_1 and x_2 ranging between 0.99 and –0.99. The true partial regression coefficients remained the same as in Figures 1 and 2 across all values of the correlation between x_1 and x_2. We used sample sizes of 10 and 100 observations. We repeated each simulation scenario 10,000 times. For each scenario, we calculated the standard deviation of estimated simple and partial regression coefficients, and the average standard error of each regression coefficient estimator. We also recorded the distribution of p-values for the effect of x_2 on y in the simulation scenario for Figure 1 (but across a wide range of correlations between x_1 and x_2). Since the true effect of x_2 on y is zero in this scenario, a tendency of collinearity to inflate p-values would be manifested in a pattern where $p<0.05$ occurs in less than 5% of simulations.

Results of the simulations are given in Figure 3. All mean standard errors (black lines) closely match the empirical standard deviations (grey lines) of replicated simulated estimates. For extremely small sample sizes ($n = 10$ in Fig. 3a,c), standard errors from the ordinary least squares regressions are slightly smaller than the empirical standard deviation of simulated estimates. This happens regardless of the degree of collinearity (i.e., it happens even when the predictor variables are uncorrelated), and it is a result of the fact that calculations of standard

errors for ordinary least squares regression rely on large sample theory. The close match of standard deviations of estimates with mean standard errors demonstrates that statements about statistical uncertainty in multiple regression analysis are not adversely affected by correlations among predictor variables.

The p-values for the effect of x_2 on y, in the simulations where x_2 has no effect, are not inflated by sampling error (Fig. 4). Across all simulation scenarios, including those with very high correlations between predictor variables, p-values less than 0.05 occurred almost exactly 5% of the time. If collinearity were to inflate p-values, then the proportion of significant tests would be lower than α.

However, standard errors of partial regression coefficients are larger when explanatory variables are highly correlated. This quite naturally reflects that when there is less variation in any explanatory predictor variable that is independent of variation in other explanatory variables, there is less information from which to make inference about direct effects. In the limit of correlations among predictor variables approaching ± 1, both standard errors and the standard deviation of estimates become extremely large. This too is perfectly natural: if there is no information in one predictor, independent of other predictor variables, there is no information, and multiple regression analysis correctly reflects this fact with appropriately large standard errors. When correlations are perfect (or in practice, very near perfect) partial effects cannot be calculated. (The OLS mechanics fail because the matrix algebra equivalent of division by zero occurs.) This too, is perfectly natural. It is no different from the fact that division by zero would occur if one were to attempt to apply the standard formula for the mean of a variable with $n = 0$. In other words, this is another manifestation of the biological reality that inference of direct effects is harder than inference of overall effects.

Misunderstanding 4: *Collinearity leads to biased parameter estimates*

The fact that partial regression coefficients change due to inclusion or exclusion of correlated predictor variables could easily lead to a misunderstanding that the estimates from models containing correlated predictor variables must be biased. However, we can use our simulations to check whether parameter estimates are in fact biased by the presence of collinearity. Formally, bias is the difference between the expected value of an estimator and the true value of an associated estimand. In the case of multiple regression, the estimators are the values of estimated partial regression coefficients from a multiple regression analysis. The estimands are the true effects of the predictor variables on the response. Figure 5 shows the mean values of estimated partial regression coefficients for every simulated scenario represented in Figure 3. This gives the expected value of the estimators. The estimands are the simulated values. Figure 5 shows that estimator and estimand have the same values, plus or minus trivial error due to a finite number of replicate simulations. That is, there is no bias.

Misunderstanding 5: *Variance inflation factors (VIFs) can indicate when collinearity is a problem, and can inform deletion of variables from regression models to address the problem*

Armed with an understanding of the distinction between simple and partial regression coefficients, we can see clearly that collinearity is not a problem. Consequently, no statistic can tell us when this non-existent problem exists. However, measures of the amount of collinearity among a set of predictor variables are potentially very useful: Variance Inflation Factors (VIFs) are a statistic that measures the degree to which collinearity affects inferences for each variable in a multiple regression analysis. Discussions on using VIFs primarily centre on threshold values of VIFs above which collinearity is considered a problem. Curiously, these discussions rarely

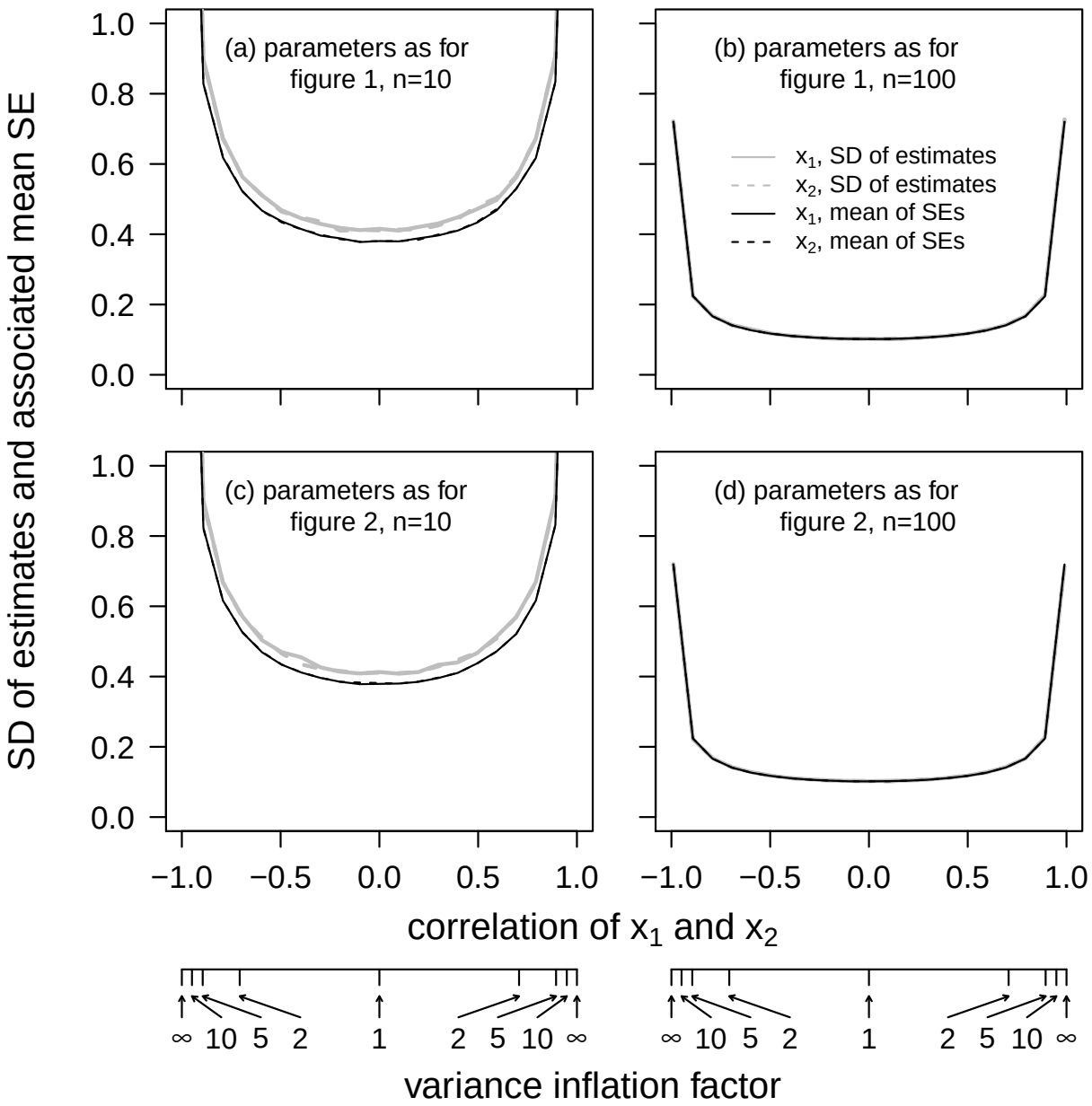

Figure 3: Simulation-based demonstration that collinearity in multiple regression analysis does not lead to unreasonable statements about precision. Solid lines show mean standard errors of estimates of the effect of x_1 on y, and dashed lines show mean standard errors estimates of the effect of x_2 on y, across a range of correlations of x_1 and x_2. Gray lines show standard deviations of estimates for each simulation scenario. If the standard errors are valid, they should match the standard deviations of estimates. (a) and (b) use the true values of the partial regression coefficients from Fig. 1, and (c) and (d) use the true parameter values from Fig. 2. Simulations for (a) and (c) use a sample size of 10, and those for (b) and (d) use a sample size of 100. The variance inflation factors for both x_1 and x_2 are identical in the bivariate analyses depicted.

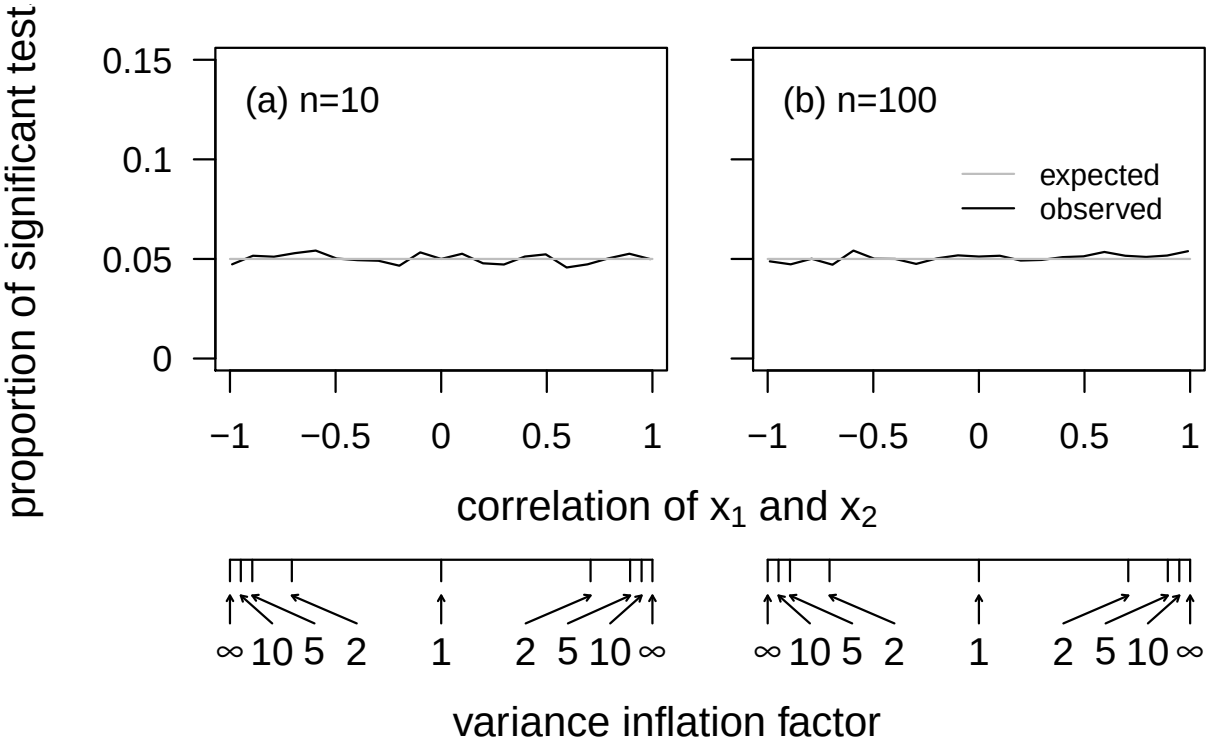

Figure 4: Simulation-based demonstration that collinearity in multiple regression analysis does not lead to inflated p-values. Black lines indicate the proportion of significant tests of the effect of x_2 on y in the simulation scenario for Fig. 1 (where the true effect is zero), across a range of correlations between x_1 and x_2, and for $n = 10$ (a) and $n = 100$ (b). The variance inflation factors for both x_1 and x_2 are identical in the bivariate analyses depicted.

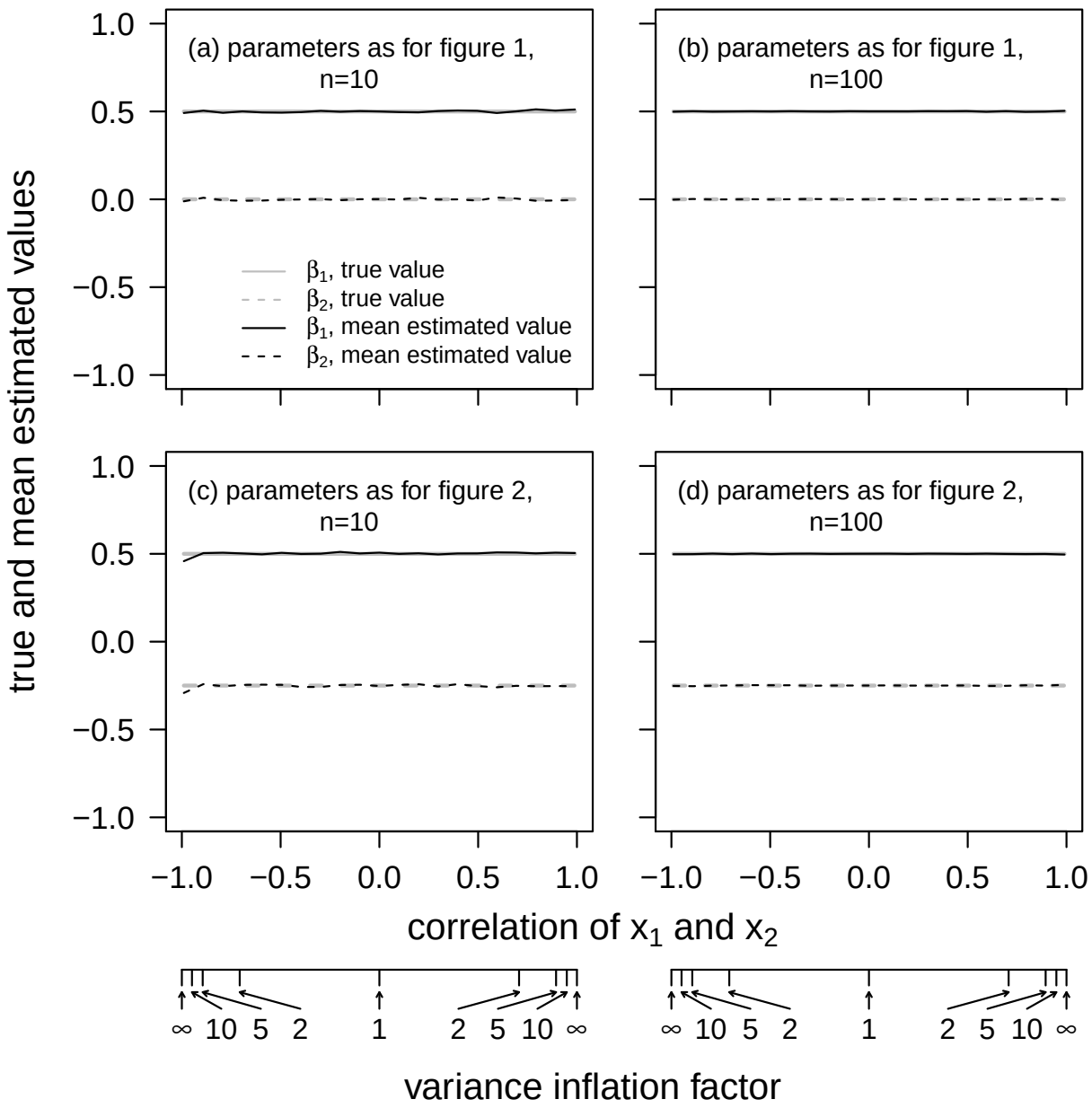

Figure 5: Simulation-based demonstration that collinearity in multiple regression analysis is not biased by any degree of colinearity among explanatory variables. Solid lines show estimates of the effect of x_1 on y, and dashed lines show estimates of the effect of x_2 on y, across a range of correlations of x_1 and x_2. Gray lines show true values of the partial regression coefficients. (a) and (b) use the true values of the partial regression coefficients from Fig. 1, and (c) and (d) use the true parameter values from Fig. 2. Simulations for (a) and (c) use a sample size of 10, and those for (b) and (d) use a sample size of 100. The variance inflation factors for both x_1 and x_2 are identical in the bivariate analyses depicted.

consider what VIFs are. The variance inflation factor associated with a given explanatory variable is the proportion by which the sampling variance (i.e., standard error squared) of its partial regression coefficient will be larger than the sampling variance of its associated simple regression coefficient, in the instructive circumstance where the residual variances of the associated simple and multiple regression models are the same. Thus, VIFs are potentially very useful, as they inform us about how much more challenging a given multiple regression analysis will be than a corresponding simple regression analysis. Researchers' 'feels' for how powerful a regression analysis might be (e.g., what sample size is typically useful in a given field) probably relate most intuitively to simple, rather than multiple, regression. VIFs can tell us how much larger our sample size must be in a given multiple regression analysis such that its precision will be equivalent to that of a simple regression analysis.

Most discussions about the magnitude of VIFs at which collinearity is understood to become an issue make no mention of sample size. Rather, many sources simply indicate values of VIF over which problems are likely to occur. However, with sufficient sample size, it is possible to resolve partial effects, even among predictors that are highly correlated. Figures 3, 4, and 5 all show, for sample sizes of $n = 10$ and $n = 100$, that multiple regression analysis is neither biased, nor generates misleading information about uncertainty, regardless of VIF (which is calculated directly from the correlation between variables). Figure 6 further illustrates the point, showing the mean standard errors from 10,000 replicate analyses for a range of sample sizes from 10 to 10^4. The analyses in these simulations attempt to resolve the direct effects of two variables that are correlated such that they have VIFs of either 1.3 or 10. (The latter is very high, according to most sources.) No true effects of either predictor exist in these simulations, both predictors have a variance of one unit, and the response has a residual variance of one unit. For both values of VIF, robust inference is achievable, given sufficient sample size. VIFs are thus probably very useful, as they may alert the analyst to variables for which partial effects may be very difficult to resolve. (This may not as immediately evident from correlations, in analyses with more than two predictors.) However, VIFs should never be treated as containing any information that can be interpreted without other quantities that determine precision. These quantities are the residual variance of the response variable, and the sample size.

Ironically, those who see collinearity as a problem often offer procedures for dealing with the perceived problem, and these procedures will generally cause bias—these are often but not exclusively based on VIFs. In the real world, the null hypothesis will rarely be true. Some quantities may have trivial direct effects on response variables of interest in any given study system, but they will not have effects of exactly zero. Imagine now that many researchers are interested in how explanatory variable x_1 influences quantity y. Suppose that x_1 has a consistent effect on y. In any given study system, there are naturally-occurring and/or experimental (e.g., time of day) variables, x_2, that might be correlated with x_1, and that might confound inferences about the effect of x_1 on y. If these x_2 variables are often removed on account of being correlated with x_1 (as recommended by for example Zur et al. 2010), then within each study, there will be a bias in the estimated effect of x_1 on y. This bias will be in the direction of the product of the correlation of x_1 and x_2 and the effect of x_2 on y. If a given x_2 variable is relevant across studies and has consistent true effects across studies, then estimates will be biased on average.

Removal of variables from models due to values of VIFs will be a detriment to biological studies in which the values of regression coefficients are of interest. This is because deletion of a variable causes the values of other partial coefficients to change. The values of the partial regression coefficients for remaining variables will then be neither simple regression coefficients nor partial effects, insofar as available data would have allowed. Removal of variables from models for any other reason will also be detrimental to biological studies. Because predictor variables

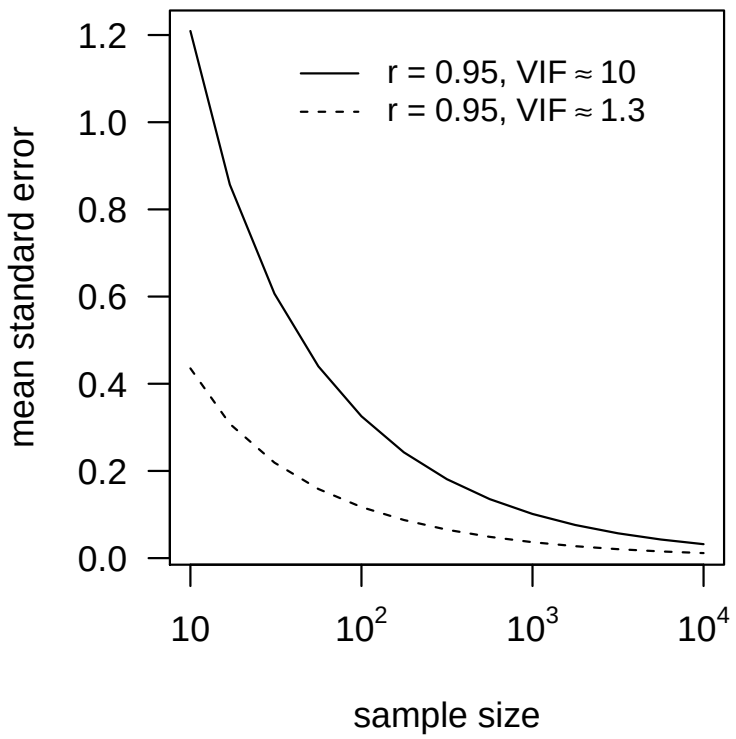

Figure 6: Simulation-based demonstration that variance inflation factors (VIFs) are meaningless unless considered together with other factors that determine precision. Lines indicate mean standard error of estimates of the partial effect of x_1 on y, as a function of sample size, in analyses where y is regressed on x_1 and x_2, where both x_1 and x_2 have variances of 1 but different correlations, where neither predictor has an effect on y, and the residual variance in y is one.

that are correlated with other explanatory variables will be associated with less precise parameter estimates, they will be disproportionately eliminated by any variable selection procedure. This includes variable selection based on p-values, and information criterion-based selection of models or sets of models.

Misunderstanding 6: *Collinearity makes predictions unreliable*

Prediction from a fitted model is the exercise of describing the expected value of the response value, given hypothetical or observed values of the explanatory variables. In prediction, values of explanatory variables at points where one wishes to obtain expectations of the response are multiplied by their associated (partial) regression coefficients. Under collinearity, a multiple regression model in which there are undoubtedly some relationships (as might be characterised with precision by simple regression) between explanatory variables and predictors might rightly have partial effects with high uncertainty (large standard errors, perhaps non-significant effects). It may initially seem that such a model will give poor predictions. If prediction occurs via the partial regression coefficients, and these coefficients have (legitimately) larger standard errors, then surely the predictions will be highly uncertain. In fact, this is not necessarily the case.

Standard errors of individual parameter estimates are not complete descriptors of uncertainty in any given fitted multiple regression model. Under collinearity, there will also be sampling error covariance among parameter estimates. Not only will there be uncertainty in the values of each parameter in isolation, but also there will be combinations of parameters that are more likely than others. Each estimate in a set of partial regression coefficients will not have been generated by the model fitting algorithm in isolation, but rather will be the most likely combination of parameter values.

An example may be most instructive in this case. Suppose that x_1 affects y, the associated regression coefficient is 0.5, and the residual variance in y is 0.75. (y thus has a total variance of 1, if the explanatory variables are standardized to have variances of 1 themselves.) Suppose that x_2 has no effect on y, and has a correlation with x_1 of 0.9. If the sample size is $n = 30$, then the simple regressions of y on x_1 and of y on x_2 would be very likely to be statistically significant, but the partial regression coefficients of the multiple regression of y on x_1 and x_2 would typically be non-significant.

In this example, the sampling covariance matrix of the partial regression coefficients is

$$\operatorname{cov}\left(\widehat{\beta}\right) = \begin{bmatrix} 0.132 & -0.118 \\ -0.118 & 0.132 \end{bmatrix},$$

in which the diagonal entries are the sampling variances of the regression slopes for the effects of x_1 and x_2 on y—for simplicity, we need not consider statistical noise in the estimate of the intercept, for the present purposes—and the diagonal value is the sampling covariance of the slopes. We give a fuller explanation of this sampling covariance matrix, and of other results in this subsection, in Appendix 2. This sampling covariance corresponds to expected standard errors of approximately 0.36 (i.e., $\sqrt{0.132}$) for both estimated partial regression coefficients. Converting the sampling covariance matrix to a correlation matrix, we get

$$\operatorname{cor}\left(\widehat{\beta}\right) = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}.$$

This means that the estimates we get are not just drawn from distributions with a standard deviation of 0.36 around the true values. But when one coefficient is overestimated, the other is almost always underestimated. The consequence of this for prediction is that the large sampling

errors that legitimately arise under collinearity cancel out in appropriate ways when predictions are made.

Suppose we wanted to know the expected value for a prediction with values of $x_1 = 1$ and $x_2 = 1$. (Since x_1 and x_2 are highly positively correlated, such a data point would be quite likely to occur in nature, and may be of interest.) Suppose also that our analysis had been lucky in that it had quite closely estimated the true values of 0.5 and 0 for the partial regression coefficients (even though they would each be non-significant), and our prediction would be $1 \cdot 0.5 + 1 \cdot 0 = 0.5$. If the standard errors (0.36 and 0.36 for the coefficients for x_1 and x_2) were to completely describe uncertainty in the model, our standard error of the prediction would be $\sqrt{0.36^2 \times 1 + 0.36^2 \times 1} = 0.51$. Our 95% prediction interval would be –0.5 to 1.5, which would probably be disappointing. Either the simple regression model of y on x_1 (the true model) or the simple regression of y on x_2 would make a better prediction. (The latter would be downwardly biased, but for many purposes the benefit of increased precision would probably outweigh cost of the bias.) However intuitive the above calculation of the standard error of the prediction may seem, it is wrong. The standard error of the prediction depends on both the sampling variance of the partial regression coefficients and their sampling correlation. (Because of the positive correlation of x_1 and x_2, there is a sampling correlation of –0.9; see above and Appendix 2.) Accounting for these, we would expect the multiple regression model to have a prediction standard error[1] of $\sqrt{0.36^2 \times 1 + 0.36^2 \times 1 + 2 \times 0.36^2 \times (-0.9)} = 0.16$. The associated 95% prediction interval (around an unbiased prediction of 0.5) would, on average, be 0.18 to 0.81. This is much more satisfying precision for a small study ($n = 30$), and could be regarded as a prediction that is significantly different from zero, even though the coefficients of the multiple regression model are both non-significant.

An issue that is commonly raised about collinearity and prediction concerns changing correlations among predictor variables from the data to which a model is fitted, to situations where a model is used for prediction. If covariances among predictors change in such scenarios, then some predictions can be highly uncertain. This is perfectly natural if it is understood as a case of out of sample prediction. Imagine a model were fitted where two predictor variables, x_1 and x_2, were strongly positively correlated. Since data with large values of x_1 and small values of x_2 will be rare or absent from the data, predictions in new settings that may contain large values of x_1 and small values of x_2 will be highly uncertain. However, it should be seen as appropriate that such scenarios generate large uncertainty: the data to which the model was originally fitted contained little information about such scenarios, and so high uncertainty should be expected. If correlations among predictors change between data sets used for model fitting and prediction, and variable selection is applied as part of the model fitting procedure, then it is possible that highly biased predictions (with inappropriately small uncertainty) will result. We do not view this as a problem arising from collinearity itself, but rather as a problem with variable selection procedures.

4 Example

Loyn (1987) investigated the effects of several properties of forest habitat patches in Australia on an index of bird abundance (average number of birds seen or heard in 20-minute surveys). Our analysis first recapitulates, and then revisits the interpretations, of previous didactic uses of Loyn's (1987) data by Quinn and Keough (2002) and Zur et al. (2009). The explanatory

[1]This is the standard formula for the variance of a product of random variables, i.e., $\mathrm{Var}(a + b) = \mathrm{Var}(a) + \mathrm{Var}(b) + 2\mathrm{Cov}(a, b)$, where a is $x_1\beta_1$ and b is $x_2\beta_2$, x_1 and x_2 are the values of the predictor variables at which we seek a prediction, and β_1 and β_2 are the partial regressions of the response on x_1 and x_2

variables for bird abundance are: (log) patch area, (log) distance of each patch to the nearest neighbour patch, (log) distance of each patch to the nearest large patch, the year in which each patch was isolated from similar habitat, patch altitude, and a 5-level factor indicating the intensity of grazing experienced in each patch. Correlations among the predictor variables are depicted in Figure 7.

There are significant associations of bird abundance with habitat patch area, year of isolation, altitude, and grazing intensity in simple regression analyses (Table 1a). In particular, larger patches, more recently isolated patches, and patches at higher altitudes have higher bird abundance. Patches with the most intense grazing have the lowest bird abundance. In a multiple regression (Table 1b) considering all variables simultaneously, the partial effects of year of isolation and altitude are much more modest and are not statistically significantly different from zero.

Should the fact that very different results, especially for year of isolation and altitude, are obtained by models containing different predictor variables be viewed as an indication that something is wrong? In particular, should we be concerned that correlations among the predictor variables could be invalidating some, or all, of our simple and multiple regression models?

There is no need for concern. It is perfectly natural that the coefficients of different models, containing different predictor variables, will have different values. It may be instructive to unpack the distinction between the simple and partial regression effects of year of isolation and altitude on bird abundance, in order to illustrate how it is natural that a simple regression effect exists in the absence of a partial effect. We can begin this unpacking by noting the correlations of grazing intensity with both year of isolation and altitude: The patches with the highest grazing intensity are those that have been isolated the longest (Fig. 7, 7a). The correlation of altitude and grazing intensity is less intense, but the main feature is similar: the patches with the highest grazing intensity are predominantly those at the lowest altitude. Figure 8 shows the correlations of year of isolation and altitude with grazing, and also shows predicted value of bird abundance from multiple regression models including year of isolation and grazing (Fig. 8a) and altitude and grazing (Fig. 8b). The coefficients of these additional illustrative models are given in the supplemental information and are qualitatively similar to effects of year of isolation and grazing intensity in the full multiple regression model (Table 1b). Both year of isolation and altitude are negatively associated with grazing, and grazing has a negative effect on bird abundance. Consequently, the grazing effect induces positive correlations of year of isolation and altitude with bird abundance, each of which is reflected in simple regressions.

Other discussions about collinearity in many sources, including influential texts, indicate that various aspects of the multiple regression model of bird abundance could be inappropriate or unreliable. However, insofar as we may have wished to generate estimates of the partial effects of different variables on bird abundance and may have wished to make valid statements about uncertainty in those partial effects, there are no aspects of these results that should be considered inappropriate as a result of the correlations that exist among the predictor variables. Importantly, the cases where significant effects exist in simple regressions (e.g., a significant simple regression of bird abundance on year of isolation) but not in the multiple regression (e.g., isolation does not have a significant partial effect on bird abundance), should not be seen as a failure of the multiple regression model. Rather, the partial effects simply represent different quantities than the simple regression effects, and as such, there should be no cause for concern that the two parameters take different values. Furthermore, the larger standard errors of the partial effects (see Table 1a,b) reflect the fact that partial regressions are legitimately more difficult to characterise than simple regressions, and are not symptoms of any inappropriate effects of collinearity either.

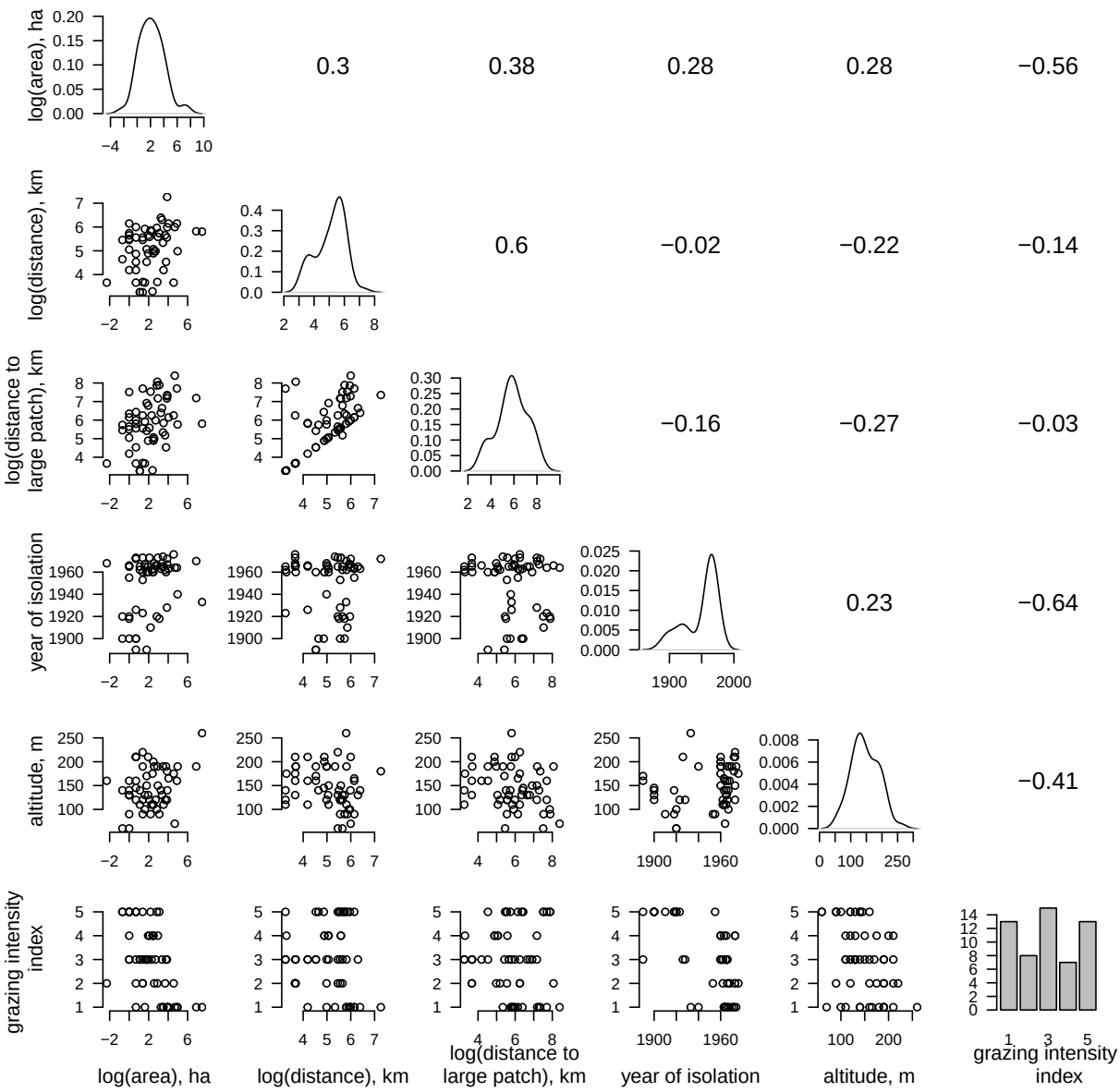

Figure 7: Correlations among predictor variables in an analysis of determinants of bird abundance. Values above the diagonal are correlation coefficients corresponding to bivariate plots below the diagonal. The grazing intensity is treated as a multi-level factor for the purpose of analyses, but treated as a continuous variable for the purpose of calculating a correlation coefficient.

Table 1: Simple (a) and partial (b) regression coefficients of the regression of Loyn's (1987) index of bird abundance on available predictor variables. Two predictor variables, year of isolation and altitude, share non-trivial univariate relationships with bird abundance (as assessed by simple regression in part a), but have little direct effect in the multiple regression model (part b). These differences arise because of correlations among predictor variables, but should not be interpreted as failures of either the simple regressions, the multiple regression model, or both. Rather, consideration of the distinction between simple and partial regression coefficients reveals how the differences between models that contain different predictor variables (in this case simple vs. multiple regression models) can be informative about how different predictor variables come to covary with response variables. Further illustration of the distinction between simple and partial effects of year of isolation and grazing are given in Figure 7. The bird abundance index is the number of birds seen or heard per 20 minute survey. We denote this quantity by A when presenting units for estimated parameters.

Regression coefficient		Units	Estimate	SE	p-value
(a) Simple regression coefficients					
(log) area		A/ln(h)	4.25	0.52	7.2×10^{-11}
(log) distance to nearest patch		A/ln(km)	1.43	1.52	0.352
(log) distance to nearest large patch		A/ln(km)	0.96	1.10	0.836
year of isolation		A/year	0.211	0.049	7.7×10^{-5}
altitude		A/m	0.095	0.031	3.3×10^{-3}
grazing intensity index	level 2	A (contrast)	−6.67	3.38	0.053
	level 3	A (contrast)	−7.34	2.84	0.013
	level 4	A (contrast)	−8.05	3.52	0.026
	level 5	A (contrast)	−22.33	2.95	6.8×10^{-10}
(b) Partial regression coefficients					
(log) area		A/ln(ha)	2.96	0.65	4.9×10^{-5}
(log) distance to nearest patch		A/ln(km)	0.14	1.16	0.904
(log) distance to nearest large patch		A/ln(km)	0.35	0.93	0.711
year of isolation		A/year	−0.013	0.058	0.827
altitude		A/m	0.011	0.024	0.656
grazing intensity index	level 2	A (contrast)	0.53	3.25	0.872
	level 3	A (contrast)	0.07	2.96	0.982
	level 4	A (contrast)	−1.25	3.20	0.698
	level 5	A (contrast)	−12.47	4.78	0.012

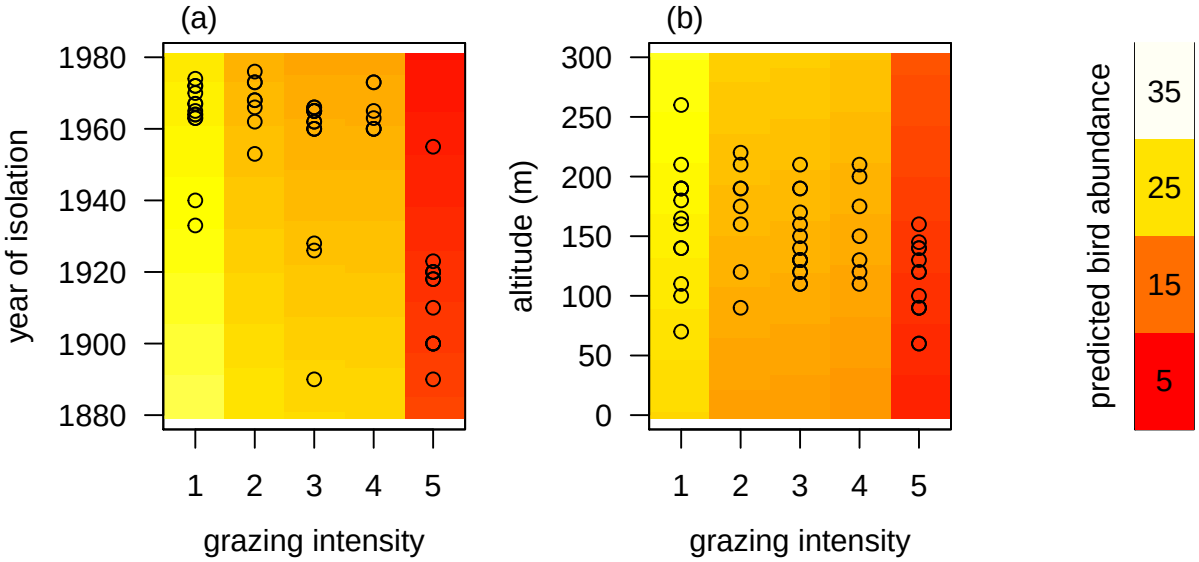

Figure 8: Illustrations of the distinction between simple and partial regression coefficients, using data on bird abundance from Loyn (1987). Grazing intensity (an index with five levels), year of isolation, and altitude are all related to the index of bird abundance in simple regressions (Table 1a). However, only grazing intensity has a direct effect in a multiple regression model containing these (and other) variables as predictors (Table 1b). (a) and (b) show the correlations of year of isolation and altitude with the grazing intensity index. In both cases, there is a negative correlation with grazing intensity, and this covariance of predictors, combined with the negative effect of high grazing on bird abundance, generates positive correlations (and corresponding positive simple regressions) of grazing intensity with year of isolation and altitude. Thus, it is not necessary to invoke any failure of the multiple regression model to explain the differences between the simple and partial regression coefficients for year of isolation and altitude; rather, the models containing different predictors help us to understand different things about how different predictor variables come to be related to bird abundance.

5 Conclusions

The views we advocate on collinearity certainly do not guarantee correct biological interpretations from multiple regression models. Partial regression coefficients and associated biological interpretations always risk being wrong: aside from statistical error due to finite sample size, unmeasured, important variables may exist in any given study. The important feature of analyses where predictor variables have been removed due to collinearity is that they unnecessarily risk being wrong, when they could have been right about a partial effect, insofar as available data would have allowed. To the extent that an analysis can be judged successful if it returns statistically significant results (or that results fulfil some other criterion that, rightly or wrongly, gets interpreted as analogous to statistical significance), excluding variables on the basis of collinearity, significance, or any other means of variable selection will generate successful analysis. To the extent that an analysis can be considered successful if it can be relied upon to give unbiased estimates of the direct effects of predictor variables on responses, so far as available data allow, and that the analysis can yield credible statements about our uncertainty in these estimates, collinearity is not a problem, and procedures to address the supposed problem lead to unsuccessful analysis.

What has led to widespread confusion about the various supposed problems that collinearity causes? This is very hard to determine. We suspect that a lack of appreciation for the distinction between simple and partial regression coefficients could be a major underlying contributor to confusion in this area. Many of the perceived problems with collinearity begin to make sense if one expects multiple regression coefficients to behave like simple regression coefficients. For example, if multiple regression were simply a mechanism for automating the application of multiple simple regressions, and a researcher were asking questions for which simple regression coefficients provided the answers, then the behaviour of multiple regression coefficients under collinearity would indeed be a cause for consternation! Given the widespread nature of misunderstandings about the behaviour of multiple regression coefficients, it seems possible that the specific questions that researchers often ask might be better served, or at least well complemented, by fitting multiple simple regression analyses (i.e., one to calculate the regression describing the total association of each predictor variable of interest) in place of or in addition to multiple regression analyses.

A plethora of methods exist that are widely understood to deal with different aspects of the problem of collinearity. In many cases, the problems of collinearity are simply misunderstandings: if we understand what partial effects are, then the fact that partial effects take different values when different covariates are included in a model will not be seen as a problem in need of a solution. Similarly, when partial effects are correctly understood to be different quantities than simple regression effects, then the fact that they have bigger standard errors and p-values should not be cause for consternation: there is no reason to expect to ask different questions of the same data set and get equally precise answers. If one question (partial regression) is harder than the other (simple regression, or regression models from which collinear terms have been removed), then we should be shocked and suspicious if we somehow obtain equally precise answers. Variants of multiple regression analysis (variable selection, different kinds of regularised regression, model averaging) generally sacrifice the desirable features of unadulterated multiple regression (in particular, unbiasedness and ability to make credible statements about uncertainty). Some of these methods may provide tangible benefits in return, in some settings. However, these alternative methods absolutely should not be adopted on the understanding that they correct or counterbalance problems with bias or uncertainty in multiple regression; such understanding is incorrect.

The purpose of this paper has been twofold. First, we have sought to provide the clearest possible illustration of the meaning of the coefficients that multiple regression provides. Second, we have used this clarification to rectify common misunderstandings pertaining to collinearity in multiple regression. We hope this focus on collinearity has been useful in its own right, but also in that it has provided an opportunity to explore, reiterate, and reinforce key concepts about multiple regression. The issues raised here about collinearity in multiple regression translate across to other statistical procedures that use multiple explanatory variables, such as generalised linear models, mixed models, generalised additive models, and causal inference using graphs (i.e., path analysis and related procedures). Collinearity is in no general way a problem for statistical analysis. Rather, collinearity is an often-interesting facet of biological systems under study; and it is a facet that familiar statistical techniques such as multiple regression are perfectly able to accommodate.

Appendix 1: Covariance among explanatory variables as a source of information in multiple regression

One way to see how correlations among explanatory variables are not assumed in multiple regression to take any particular values (e.g., small values or zero) is to examine the equations by which partial regression coefficients are estimated. These are generally presented in matrix form with vectors of responses and parameter estimates, and design matrices. Such a presentation is not conducive to general understanding. The various matrix products that occur are actually just the covariances of explanatory variables with responses, and variances and covariances of explanatory variables, scaled by the sample size. The mechanics of estimation of multiple regression coefficients actually boils down to dividing the covariances of responses with predictors by the variance-covariance matrix of the predictors. Without affecting any subsequent arguments, we can assume that explanatory variables are mean centered, and then not worry about the model intercept. The formula for estimating partial regression coefficients, $\widehat{\mathbf{b}}$, is

$$\widehat{\mathbf{b}} = \mathbf{\Sigma}^{-1}\mathbf{c},$$

where $\mathbf{\Sigma}$ is the matrix of variances and covariances among explanatory variables, and $\mathbf{c}$ is a vector of covariances of the explanatory variables with the response variable.

In the example in Figure 1, the covariance matrix of x_1 and x_2 is

$$\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix},$$

and the covariances of x_1 and x_2 with y are

$$\begin{bmatrix} 1/2 \\ 1/4 \end{bmatrix}.$$

Consequently, the partial regression coefficients are

$$\widehat{\mathbf{b}} = \mathbf{\Sigma}^{-1}\mathbf{c} = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1/2 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 4/3 & -2/3 \\ -2/3 & 4/3 \end{bmatrix} \begin{bmatrix} 1/2 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 0 \end{bmatrix},$$

which is fortunate since we simulated each value of y from a normal distribution with mean of $1/2\ x_1$, no effect of x_2, and a variance of one.

Rather than being a hindrance to multiple regression, covariance among explanatory variables is in fact the source of information that allows us to infer direct effects.

Appendix 2: A little more about sampling (co)variance and its effect on prediction

The sampling variance-covariance matrix of the parameters of a multiple regression with a given design matrix $\mathbf{X}$ is given by

$$\mathrm{Var}\left[\widehat{\mathbf{b}}\middle|\mathbf{X}\right] = \sigma_e^2(\mathbf{X}^t\mathbf{X})^{-1}.$$

The design matrix is a slightly intimidating quantity at first. But it is just a way of organizing the predictor variables. For example, in a multiple regression of some response on two predictor variables, where for the first three records the values of the first predictor variable were 1, 2, and 2, and of the second predictor variable were 2, 1 and 3, the top of the design matrix would be

$$\begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 1 & 2 & 3 \\ \vdots & \vdots & \vdots \end{bmatrix}.$$

The first column of ones indicates the model intercept, and the next two columns are the values of the two predictor variables.

Correlations among predictor variables come into the calculation of sampling variances and covariances via $(\mathbf{X}^t\mathbf{X})^{-1}$. If $\boldsymbol{\Sigma}$ is the covariance matrix of columns of $\mathbf{X}$, and $\boldsymbol{\mu}$ is the vector of means of columns of $\mathbf{X}$, then

$$(\mathbf{X}^t\mathbf{X})^{-1} = \frac{1}{n}(\boldsymbol{\Sigma}+\boldsymbol{\mu}^{\mathrm{t}}\boldsymbol{\mu})^{-1}.$$

Recall that for estimating regression coefficients, far from there being some assumption that predictors are uncorrelated, or are not correlated above some level, the correlations among predictors, whatever they may be, are actually part of the information used to estimate partial regression coefficients. Similarly, we can see here that, rather than there being some assumption about no or little correlation among predictor variables in the machinery for calculating important measures of statistical uncertainty (e.g., standard errors or p-values), whatever correlations among predictor variable occur, they are used by multiple regression analysis.

In the example in the text, the predictor variables have a correlation of 0.9, the sample size is 30 and the residual variance is 0.75. Using the expressions above (and assuming that the covariates are mean centred; this is not necessary, but it makes the maths more transparent, as we need only consider the second two columns of $\mathbf{X}$), we expect the sampling variances and covariances to be

$$\begin{aligned}\mathrm{Var}\left[\widehat{b}\middle|\mathbf{X},\sigma_e^2\right] &= \sigma_e^2(\mathbf{X}^t\mathbf{X})^{-1} = \frac{\sigma_e^2}{n}\left(\boldsymbol{\Sigma}+\mu^{\mathrm{t}}\mu\right)^{-1} = \frac{3/4}{30}\frac{1}{1-0.9^2}\begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} \\ &\approx \begin{bmatrix} 0.132 & -0.118 \\ -0.118 & 0.132 \end{bmatrix}.\end{aligned}$$

The variances of predictions (i.e., of $\widehat{\mathbf{b}}\mathbf{x}$) are then given by the standard expression for the variance of a linear transformation of a random variable,

$$\mathrm{Var}\left[\mathrm{E}[y|x,\widehat{\mathbf{b}},\mathrm{Var}\left[\widehat{b}\middle|\mathbf{X},\sigma_e^2\right]]\right] = \mathbf{x}\left(\mathrm{Var}\left[\widehat{b}\middle|\mathbf{X},\sigma_e^2\right]\right)\mathbf{x}^t.$$

So, while it may seem as though precise prediction is impossible, since the standard errors of the estimated regression coefficients are very large $\sqrt{0.132} = 0.363$, quite precise prediction

is possible, because of the sampling covariance. Using, as an example (see text) an attempt to predict y when $x_1 = 1$ and $x_2 = 1$,

$$\mathrm{Var}[E[y]] = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 0.132 & -0.118 \\ -0.118 & 0.132 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 0.0263.$$

The standard error of the prediction is $\sqrt{0.0263} = 0.162$, which is likely to represent far more useful precision than does the impression given from the standard errors of the regression coefficients alone.

Acknowledgments

We thank Maria Joao Janeiro, Andy Gardner, Luke Rendell, and Sophie Smout for discussions. MBM is supported by a University Research Fellowship from the Royal Society (London).

Literature cited

Crawley, MJ. 2007. *The R Book.* Wiley, Chichester.

Dormann, CF, Elith, J, Bacher, S, Buchmann, C, Carl, G, Carré, G, García Marquéz, JR, Gruber, B, Lafourcade, B, Leitão, PJ, Münkemüller, T, McClean, C, Osborne, PE, Reineking, B, Schr oder, B., Skidmore, AK, Zurell, D, and Lautenbach, S. 2013. "Collinearity: A Review of Methods to Deal With It and a Simulation Study of Their Performance." *Ecography* 36: 27–46.

Graham, MH. 2003. "Confronting Multicollinearity in Ecological Multiple Regression." *Ecology* 84: 2809–2815.

Legendre, P, and Legendre, L. 1998. *Numerical Ecology*. Elsevier Press, London.

Loyn, RH. 1987. "Effects of Patch Area and Habitat on Bird Abundances, Species Numbers and Tree Health in Fragmented Victorian Forests." In *Nature Conservation: The Role of Remnants of Native Vegetation*, edited by DA Saunders, GW Arnold, AA Burbidge, and AJM Hopkins, 65–77. Surrey Beatty & Sons, Chipping Norton, NSW.

Pearl, J, Glymour, M, and Jewel, NP. 2016. *Causal Inference in Statistics*. John Wiley and Sons, Chichester, UK.

Prunier, JG, Colyn, M, Legendre, X, Nimon, KF, and Flamand, MC. 2015. "Multicollinearity in Spatial Genetics: Separating the Wheat from the Chaff Using Commonality Analyses." *Molecular Ecology* 24: 263–283.

Quinn, GP, and Keough, MJ. 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge.

Ray-Mukherjee, J, Nimon, K, Mukherjee, S, Morris, DW, Slotow, R, and Hamer, M. 2014. "Using Commonality Analysis in Multiple Regressions: A Tool to Decompose Regression Effects in the Face of Multicollinearity." *Methods in Ecology and Evolution* 5: 320–328.

Zur, AF, Ieno, EN, and Elphick, CS. 2010. "A Protocol for Data Exploration to Avoid Common Statistical Problems." *Methods in Ecology and Evolution* 1: 3–14.

Zur, A, Ieno, EN, and Smith, GM. 2007. *Analysing Ecological Data.* Springer, New York.

ISSN 2475-3025