A Defense of Model Averaging

Jeffrey A. Walker*[a]

[a]Department of Biological Sciences, University of Southern Maine, Portland, ME 04105, USA

May 3, 2017

Abstract

Model averaging of partial regression coefficients has been criticized for averaging over a set of models with coefficients that are either incommensurable or describe fundamentally different things if there is any correlation (multicollinearity) among the predictors. It is easy to show that partial regression coefficients conditional on different covariates are commensurable. And, partial regression coefficients from different models can have the same meaning if they are used to estimate the effects in a causal model, which derive their meaning from the specified paths and not from the set of covariates. A multiple regression model implicitly specifies a causal model with direct, causal paths from each predictor to the response. Consequently, the partial regression coefficient for a predictor has the same meaning across all sub-models if the goal is estimation of the causal effects that generated the response. In order to clarify the effects of multicollinearity on model-averaged estimates, I compare effect estimates using a small Monte-Carlo simulation. The simulation results show that model-averaged and ridge estimates have increasingly better performance, relative to model selection and

*walker@maine.edu

full model estimates, as multicollinearity increases.

Keywords multiple regression, causal model, conditional effect, monte carlo simulation, ridge regression, model selection, multicollinearity

Introduction

Model averaging is an alternative to model selection for either effect estimation or prediction. Model averaging is attractive because it recognizes that all measured predictors will have some non-zero association with the response independent of that shared with other predictors. Consequently, model averaging encourages the worthy goal of emphasizing effect estimation and the consequences of effect magnitude and not simply the identification of some "best" subset of predictors. Model averaging works well in practice because it improves the long run performance of either effect estimation or prediction compared to model selection (Raftery et al., 1997; Hjort and Claeskens, 2003). In ecology, model averaging is probably most frequently used to estimate the coefficients of a linear model using multiple regression but model averaging can also be used to estimate the structural (path or effect) coefficients of a causal graph (Madigan and Raftery, 1994).

Recently, model averaging of partial regression coefficients has been criticized for averaging over a set of coefficients that are conditional on a specific set of covariates and, consequently, have different "interpretations" and cannot be meaningfully averaged (Cade, 2015; Banner and Higgs, 2017). Cade (2015) specifically argued that in the presence of any multicollinearity, coefficients from different models are incommensurable. Cade's criticism is not the typical caution against the estimation of partial regression coefficients in the presence of high multicollinearity because of increased error but an argument that model-averaged coefficients in the presence of *any* correlation

among the predictors are meaningless. This effectively includes any analysis of observational data. Cade's critique is receiving much attention, as evidenced by the 74 Google Scholar citations in less than two years.

These critiques are noteworthy given that model averaging has developed a rich literature in theoretical and applied statistics over the last 25 years (Hoeting et al., 1999; Burnham and Anderson, 2002; Claeskens et al., 2008) with only passing comments questioning the meaning of model-averaged coefficients. Using simple algebra, it is easy to show that partial regression coefficients from different models are commensurable and, consequently, the meaningfulness of model averaging rests on how one uses the partial regression coefficients. In short, if we wish to *describe* an effect conditional on a specific set of covariates, then model averaging would indeed be averaging different things. But if one is using partial regression coefficients to estimate (implicitly or explicitly) a causal effect, then model averaging averages over estimates of the same thing.

Both Banner and Higgs (2017) and Cade (2015) discuss issues related to model averaging in inferring variable importance, which is not the focus here because other than for large screening designs, I believe estimates of the direction and magnitude of an effect are much more valuable for understanding a system than classifying predictors as important or not (Greenland et al., 1991). Cade (2015) suggests that model averaging effects standardized by partial variances is meaningful. While these standardized effects may give us a sense of the rank order of (standardized) variable importance, they cannot be used to quantitatively model the consequences of interventions in the system.

Model averaging

Throughout this paper, I will be referring to a regression model such as $Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_p X_p$ that has 2^p submodels (including the full model) that each contain a unique combination of the predictors. The focus here is on the estimation of the effect (the partial regression coefficient) of the jth predictor

$$\hat{b}_j = \sum w_m b_{j,m} \tag{1}$$

where m indexes the model and w_m is a weight measuring model evidence, which is the posterior probability in Bayesian model averaging. The optimal weight for frequentist model averaging is an area of active research (Hjort and Claeskens, 2003; Hansen, 2007; Liang et al., 2012). Within the frequentist model averaging literature, there is some question over the inclusion of models in which the predictor is excluded (which implicitly assigns a value of zero to its coefficient). Inclusion of the models where the assigned effect is zero has the effect of shrinking the estimate $\hat{b}_j$ toward zero, which makes model averaging an implicit shrinkage estimator and gives model averaging its desirable properties.

Partial regression coefficients from different models are commensurable

A small example suffices to show that partial regression coefficients from different models are commensurable. Let Y be measured in mm, X_1 in °C, and X_2 in μatm. The units of the simple regression of Y on X_1 are mm °C^{-1}. The units of Y on X_1 adjusted for X_2 are less obvious, but are easily computed using matrix algebra and focussing

on the units. The two partial regression coefficients are contained in the vector $\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. The upper left cell of $\mathbf{X}^\top\mathbf{X}$ is the sum of the squared elements of X_1 and has units °C^2. Likewise, the lower right cell of $\mathbf{X}^\top\mathbf{X}$ is the sum of the squared elements of X_2 and has units μatm^2. The lower left and upper right cells of $\mathbf{X}^\top\mathbf{X}$ is the sum of the cross products of the elements of X_1 and X_2 and has units °C μatm. The units of the inverse of this matrix are just the inverse units. The top (first) cell of $\mathbf{X}^\top\mathbf{y}$ is the sum of the cross products of the elements of X_1 and Y and has units °C mm. The bottom (second) cell of $\mathbf{X}^\top\mathbf{y}$ is the sum of the cross products of the elements of X_2 and Y and has units μatm mm. The partial regression coefficient of Y on X_1 is sum of the top left element of $(\mathbf{X}^\top\mathbf{X})^{-1}$ times the top element of $\mathbf{X}^\top\mathbf{y}$ and the top right element of $(\mathbf{X}^\top\mathbf{X})^{-1}$ times the bottom element of $\mathbf{X}^\top\mathbf{y}$. The first of these products has units °C$^{-2}\times$°C mm = mm °C^{-1}, which is the same for the simple regression coefficient. There is nothing about the units of a partial regression coefficient that contains information that the value is conditional on some set of covariates.

Nevertheless, Cade (2015) believes partial regression coefficients from different models are incommensurable because they "have no defined units in the presence of multicollinearity" and further criticizes those who justify model averaging for "ignoring that the individual regression coefficients have units and interpretations that are not identical across the R candidate models". By "no defined units," Cade does not mean units in the strict sense, otherwise the statement is easily shown to be incorrect (as above). Instead, Cade argues that because of the different combinations of covariates in the matrix $(\mathbf{X}^\top\mathbf{X})^{-1}$ that scales the sum of the cross products $(\mathbf{X}^\top\mathbf{y})$ to compute the partial regression coefficients, "a unit change in the predictor variable does not necessarily mean the same thing across all candidate models." This is a mistaken interpretation of the algebra. Premultiplying $\mathbf{X}^\top\mathbf{y}$ by $(\mathbf{X}^\top\mathbf{X})^{-1}$ does not determine what a unit change

of X_j is; a unit change in X_j is a unit change in X_j regardless of the covariates in the model, or the magnitude of the elements in $(\mathbf{X}^\top\mathbf{X})^{-1}$, or the magnitudes of the partial variances. If I simply replicated a study, would it be non-sensical to average X_1 over the two studies because the magnitude of the values of $(\mathbf{X}^\top\mathbf{X})^{-1}$, or the sampled partial variances of X_1, differed between studies? What does differ among the X_j from different models is not the magnitude of a "unit change" but the partial regression coefficient as an integrated whole, due to differences in both $(\mathbf{X}^\top\mathbf{X})^{-1}$ *and* $\mathbf{X}^\top\mathbf{y}$. Thus, the sense, or non-sense, of averaging across models rests on how one is using the partial regression coefficient.

The meaningfulness of model averaging depends on the goal of the analysis

Banner and Higgs (2017) caution against model averaging partial regression coefficients by arguing that the "meaning of the relationship associated with a partial regression coefficient is not exchangeable across models" because the coefficients from different models are conditioned on different sets of covariates, an argument that echoes Draper's comment on Hoeting et al. (1999). Effectively, this argument limits the usage of a partial regression coefficient to mere description. Often, however, we are explicitly (or more commonly implicitly) interested in the coefficients as estimates of the effects that generated the data, that is, as measures of causal effect. Here, I expand this concept using two different strategies, the first is more in the tradition of textbook explanations of multiple regression while the second is squarely within the tradition of causal modeling with causal (path or structural) diagrams using directed acyclic graphs.

A partial regression coefficient is an estimate of a slope

A partial regression coefficient (b_j) is an estimate of the slope (β_j) of the line formed by the intersection of the true regression (hyper)plane and the X_jY plane. The coefficient b_j from all sub-models of the full model estimate β_j, including the models where this coefficient is forced to zero because the jth predictor is dropped from the model. The larger models have potentially more information, but they also have potentially more noise and, as a consequence, the estimates from the smaller models may be better. Regardless, all coefficient b_j from all of the models estimate the same "thing".

To show this graphically, here are two simple generating models of a response Y to two variables X_1 and X_2

$$Y^{(a)} = 0.5X_1^{(a)} + 0.5X_2^{(a)} + \varepsilon^{(a)} \tag{2}$$

$$Y^{(b)} = 0.5X_1^{(b)} + 0.5X_2^{(b)} + \varepsilon^{(b)} \tag{3}$$

where $\varepsilon = N(0, \sigma)$. Both functions model precisely the same plane representing the response $E(Y|X_1, X_2)$. The difference in the generating models lies in their correlation structure of the independent variables. In model a (Equation 2), the expected correlation between X_1 and X_2 is 0.0 (this is the correlation used to generate the data). In model b (Equation 3), the expected correlation between X_1 and X_2 is 0.8. The online figure (https://plot.ly/ middle.professor/0/a-vs-b/) shows an actualization of the models using an interactive 3D scatterplot and I encourage readers to explore this plot. All of the points effectively lie on the true plane because I set σ in both models to 0.0001. The interactive 3D plot quite elegantly shows differences in the distributions of the points on the regression plane between the two models and it is this difference in the distributions on the regression plane that cause the different expected values of the

simple regression coefficient of Y on X_1 and the partial regression coefficient of Y on X_1 in model b but not a. The true (generating) coefficient of Y on X_1 can be interpreted as the slope of the line formed by the intersection of the regression plane with the X_1Y plane. This slope is easily seen by plotting Y vs. X_1 but limiting the data to those with values of X_2 between -0.05 and +0.05 (Figure 1A). The goal of the analysis is to estimate the slope of this intersection with limited information, by which I mean missing (and possibly confounding) predictors.

When the full dataset is plotted (Figure 1B), the projected scatter of points completely masks the fact that all of these points, from both data sets, effectively lie on a single plane. This projection of all of the points onto the X_1Y plane represents the 2D data fit by the simple regression of Y on X_1. This simple regression coefficient estimates the slope of the intersection of the regression plane but with limited information, and, because of the correlation structure in data b, the estimated coefficient is biased upwards. The upward bias is easily explained by the different distribution of points in the X_1Y space. Data set b is rich with points that have either large X_1 and very large Y or small X and very small Y. Because of the high correlation between X_1 and X_2 in data set b, there are many cases with these combined values and because both generating effects are positive, the expected response will be very large (or small in the case of small X_1). Data set a has few points in these regions of the space because the combination of unusually large (or small) X_1 and X_2 is uncommon. The opposite pattern occurs in the large X_1-small Y or small X_1-large Y parts of the space. A small Y in combination with a large X_1 would be extremely uncommon in data set b because of the large positive correlation between the X, which nearly guarantees a large Y.

If we interpret the jth partial regression coefficient not as a *specific* conditional effect but simply as an estimate of the slope of the line formed by the intersection of the

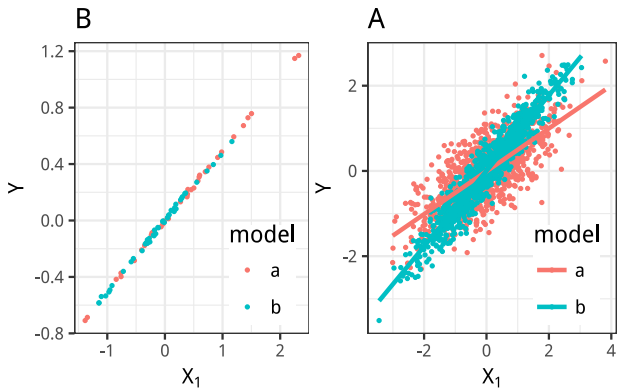


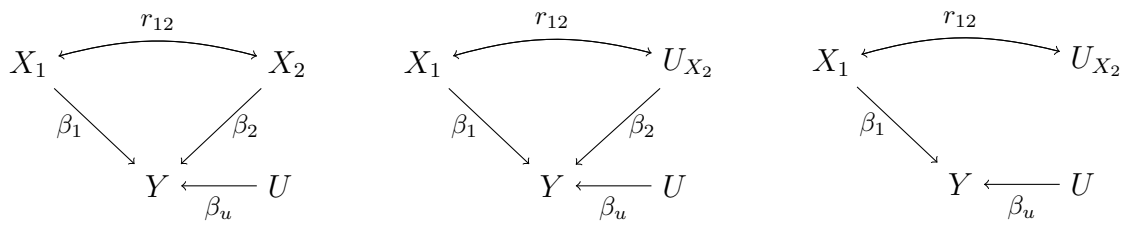
Figure 1: **Simulated data from two models that describe the same regression plane**. A) The data are limited to the cases with $-0.05 < X_2 < 0.05$, which shows the slope of the line formed by the intersection of the regression plane and the X_1Y plane. B) The full data set, which masks the fact that all of these points lie on the same plane and shows how points from model a and model b have different distributions on this plane. Finally, this view shows the scatter relevant to the simple regression of Y on X_1.

true regression (hyper)plane and the X_jY plane (Figure 1A), then b_j from all models estimate this slope. In the particular example given above, it is clear that the coefficient conditioned on X_2 is a better (much!) estimate but this is only because the noise is extremely small and we know the truth. With real, and much noisier, data, we don't know which of the models results in better estimates, this is, after all, the whole point of model selection and model averaging.

A partial regression coefficient is an estimate of a path coefficient

Banner and Higgs (2017) prefer a descriptive usage of partial regression coefficients because they believe that causal usage is largely "inappropriate for observational studies without paying careful attention to principles and methods of causal modeling." I agree that explicitly modeling causal relationships between variables using the principles and methods of causal graphs (Wright, 1934; Shipley, 2002; Pearl, 2009) adds rigor and clarity to a causal analysis. Nevertheless, many researchers use partial regression coefficients as estimates of causal effect, either explicitly using causal graphs or implic-

itly using something like multiple regression. Any multiple regression equation can be transformed into a simple, causal (path or structural) model where the partial regression coefficient for X_j estimates the path (or "effect") coefficient β_j. For example, the linear regression model $Y = b_1X_1 + b_2X_2 + \varepsilon$ implies the left path model in Equation 4.



(4)

where the path coefficient X_j is the direct, causal effect of X_j on Y (Wright, 1934) and U contains all unmodeled factors that contribute to the error ε (in this left model, and the other two, all variables and coefficients are variance standardized to simplify the math). An important assumption of this path model is that this unmeasured factor U is not correlated with either of the X as indicated by a lack of a double-headed arrow (otherwise it would be a confounder – see below). Given this model, the partial regression coefficients b_1 and b_2 are unbiased estimates of the path coefficients.

Importantly, a path coefficient is defined by the graph and not the data (Pearl, 2009). Formally, $\beta_1 = E(Y|do(X_1 = x_1 + 1)) - E(Y|do(X_1 = x_1))$ where the *do* operator represents what would happen in an experimental intervention that manipulates X_1 but leaves any other variables unchanged (Pearl, 2009). I think of β_1 as X_1's direct contribution to Y in a simulation of the graph in the left panel of Equation 4,

```
y <- beta_1*x1 + beta_2*x2 + rnorm(n)*sigma
```

Using either the formal definition or my informal interpretation, the meaning of β_1 is independent of what other covariates are in the model. By contrast, a partial regression

coefficient is defined by the data: $b_1 = E(Y|X_1 = x_1+1, X_k = X_k) - E(Y|X_1 = x_1, X_k = X_k)$. The partial regression coefficient X_1 is the expected difference in Y if we *observe* a one unit difference in X_1 within any subset of data with equal X_2. A path coefficient is the expected *change* in Y if we intervene and assign X_j a one unit increase but leave X_2 at it's original value. For the left side model, these two concepts coincide, in the sense that both concepts produce the same value, but this coincidence occurs only under very limited conditions.

To see this, consider the path model in the middle (Equation 4), which is the same as the one on the left except that X_2 is relabeled U_{X_2} to indicate that it is unmeasured. The path coefficient β_1 has precisely the same meaning here as it did in the left graph, but this is not true of the regression coefficient b_1. In a regression analysis that implies the middle graph, b_1 is not conditioned on any other variable, while in a regression that implies the left graph, b_1 is conditioned on X_2. If we use the regression implied by the middle graph to estimate β_1, the resulting coefficient, *as an estimate of* β_1, will be biased by the amount $r_{12}\beta_2$. Consequently, while the meaning of β_1 is the same in both the left side and middle graphs of Equation 4, the estimated values of β_1 will differ if this is all the information we have and we use a partial regression to estimate its value. In this model, the partial regression coefficient and the path coefficient do not coincide.

Now consider the path model to the right (Equation 4), which is precisely like the middle model except that there is no effect of X_2 on Y, which is indicated by the missing arrow between the two variables. And again, X_2 is unmeasured. The path coefficient β_1 has, of course, the same meaning in this graph as in the others. The meaning of b_1 as the simple regression coefficient not conditioned on any other X is the same as in the middle graph but not as in the left graph, where the coefficient is conditional on X_2. But the expected value of b_1 is the same as that in the left but not middle graph.

That is, the meanings of the partial regression coefficients differ among the graphs, the expected values of the partial regression coefficients differ among the graphs, and how these differ itself differs. β_1, by contrast, has the same meaning if I'm simulating the left, the middle, or the right graph. This is because a path coefficient takes it's meaning from the arrow connecting the two variables, and not the set of all variables in the graph or how these variables are connected.

The missing confounder illustrated by U_{X_2} in the middle (but not right side) graph (equation 4) highlights the major concern of causal modeling with observational designs using simple multiple regression or causal graphs. In observational designs, there are likely to be many unmeasured U that both directly affect Y and are correlated with the X. Regardless of weather we prefer model selection or model averaging over a set of models, all of these coefficients are biased by some unknown and unknowable amount. This unknown error results in standard errors that are (often far) too small and in inflated confidence in our inference (Walker, 2014). Certainly this omitted variable bias is a far bigger concern of observational designs than averaging coefficients from different models.

Banner and Higgs (2017) prefer the descriptive usage of partial regression coefficients because "the word 'effect' implies a causal relationship between the input variable and the response, which is often inappropriate for observational studies without paying careful attention to principles and methods of causal modeling." I am attracted to graphical causal models for observational designs because these models, compared to multiple regression type models, allow a much more flexible relationship among variables and, generally, the assumptions necessary for causal interpretation of the coefficients are much more transparent. Nevertheless, as shown above, any multiple regression model can be directly expressed as a causal graphical model and partial regression

coefficients are a classical method for estimating path coefficients (Wright, 1934). Coefficients in a causal graphical model can also be averaged across models (Madigan and Raftery, 1994). Importantly, a causal interpretation of coefficients in a causal graphical model and a causal interpretation of partial regression coefficients in a multiple regression type model require the exact same set of strong assumptions.

How might model averaging mislead us?

Neither Cade (2015) nor Banner and Higgs (2017) criticize model averaging based on Monte Carlo simulation results. Cade simply rejects Monte Carlo results, arguing that "the statistical performance suggested by distributions of their simulated model-averaged estimates is of questionable merit." To Cade, model-averaged coefficients have meaningless magnitudes. Throughout his paper, Cade argues that model-averaged estimates are meaningless given *any* correlation between the predictors. For meaningful model-averaged estimates, the covariance matrix of the predictors must be orthogonal as in an experimental design with no interaction effects.

In his discussion of the GPA data, however, Cade (2015) argues that "averaging across estimates β_{1j} for X_1 is likely to be much more misleading because of changes in the scaling of units than is averaging estimates for the other predictors." Here, misleadingness has a scale. Regardless, in what sense is the average of the coefficients of X_1 across models misleading? I would be mislead if the model-averaged estimate were far from the truth but this doesn't seem to be what Cade means by "misleading." Instead, Cade seems to use misleading to mean "nonsensical." But we don't need the example to demonstrate this. If it is nonsensical to average coefficients across models it is nonsensical to average coefficients across models. That is, there cannot be a scale of "misleadingness" among values that are all meaningless. If by "misleading" Cade means some-

thing like mean square error, then this is true of any multiple regression problem; some effects are estimated with smaller error than others because of the covariance structure. But if I'm deciding on a method, I don't care about relative error among effects within a method, but absolute error for the same effect across methods.

Model averaging in the presence of multicollinearity

Given the usual cautions about multicollinearity in the ecology literature, one might mistakenly believe that Cade's critique of model averaging only applies to data with high multicollinearity. Regardless, the critique does raise the question of how model averaging compares to other estimators in the presence of high multicollinearity. Freckleton (2010) used a small simulation to show how full model and model-averaged estimates perform with increased multicollinearity and showed that the error variance of the full model estimates increases more rapidly than that of the model-averaged estimates but that the model-averaged estimates were increasingly biased with increased multicollinearity. Unfortunately, Freckleton did not combine the two sources of error into a single measure (such as mean square error) for comparison and, as a consequence, I cannot determine which method performs better in the sense of long run error.

Here I use a small simulation to explore the effect of increasing multicollinearity on full model and model-averaged estimates. I also report the results for a model-selection strategy and the results using a ridge estimate. The simulation is not meant to be comprehensive but simply instructive.

Simulated data were generated using the generating function

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \sigma_y N(0,1) \tag{5}$$

where

$$\sigma_y^2 = 1 - (\beta_1^2 + \beta_2^2 + \beta_3^2 + 2\beta_1\beta_2\rho_{12} + 2\beta_1\beta_3\rho_{13} + 2\beta_2\beta_3\rho_{23}) \tag{6}$$

where ρ_{ij} is the expected correlation between x_i and x_j (see below). This error is scaled so that the effect coefficients β are on a standardized scale (the coefficients are not actually standardized and they have units). The predictors are random variables with expected mean zero and variance one but are correlated due to the shared, common factors z_1 and z_2 and generated using the functions

$$x_1 = \alpha_1 z_1 + \alpha_2 z_2 + \sigma_1 N(0,1) \tag{7}$$

$$x_2 = \alpha_1 z_1 + \alpha_2 z_2 + \sigma_2 N(0,1) \tag{8}$$

$$x_3 = \alpha_2 z_2 + \sigma_3 N(0,1) \tag{9}$$

where the z are independent, random normal variables drawn from $N(0,1)$. The predictors have expected variance of one by setting $\sigma_1^2 = (1 - \alpha_1^2 - \alpha_2^2)$, $\sigma_2^2 = (1 - \alpha_1^2 - \alpha_2^2)$, and $\sigma_3^2 = (1 - \alpha_2^2)$. The expected correlations are $\rho_{12} = \alpha_1^2 + \alpha_2^2$, $\rho_{13} = \alpha_2^2$, and $\rho_{23} = \alpha_2^2$. These equations for σ and ρ are easily derived using Sewell Wright's rules for path coefficients (Wright, 1934; Sokal and Rohlf, 2012).

The focus of the simulation is the estimate of β_1 given different levels of correlation between x_1 and x_2. I allowed β_1 to vary among three values: 0.15, 0.3, or 0.6. While 0.15 is a small effect, this is a common value in multiple regressions in ecology, where the estimated average standardized effect size is 0.18-0.19 (Møller and Jennions, 2002). The other two effects were set to $\beta_2 = 0.1$ and $\beta_3 = 0.2$. The level of multicollinearity was controlled by varying α_1 among the values 0.0, 0.5, 0.7, and 0.85. α_2 was set to 0.4. The expected correlation between x_1 and x_2 for the four values of α_1 are 0.16,

0.41, 0.65, and 0.88. Two sample sizes were simulated, $n = 20$ and $n = 100$. 2000 simulated data sets were generated for each combination of α_1 (controlling the level of collinearity), β_1 (controlling the level of the focal effect) and n.

The entire simulation was implemented in R (R Core Team, 2015) and the script is available in a supplement. In each run, the generating coefficients were estimated using the full model, model selection, model averaging, and ridge regression. For the model selection estimates, I used the dredge function in the MuMIn package (Bartoń, 2015) to fit all submodels and chose the best model as that with the smallest AICc. Model-averaged coefficients were computed over all fits returned by dredge using the model.avg function in the MuMIn package. Coefficients of predictors excluded from a model were assigned a value of zero (using the row of the coefficient table with the label "full"). For the ridge regression, I used the lm.ridge function in the MASS package (Venables and Ripley, 2002) and used cross-validation to compute the optimal tuning parameter. The performance of model averaging relative to the full model estimates, model-selection, and ridge regression is quantified using the long-run error $RMSE = \frac{\sum(b_1-\beta_1)^2}{2000}$, which is a function of both error variance and bias. b_1 is the estimate of β_1 from the regression model.

At the effect size common in ecology and evolution (0.15), model averaging outperforms other estimators when the sample size is small ($n = 20$) while ridge regression outperforms other estimators when the sample size is big ($n = 100$) (Figure 2. At intermediate effect sizes, ridge regression performs better than other estimators at both sample sizes, with the exception of model averaging when the correlation is very high ($r = 0.88$) and sample size is small. For the largest effect size, model averaging performs better than other estimators when sample size is large. When sample size is small, the full-model performs best when the correlation is small but model averaging

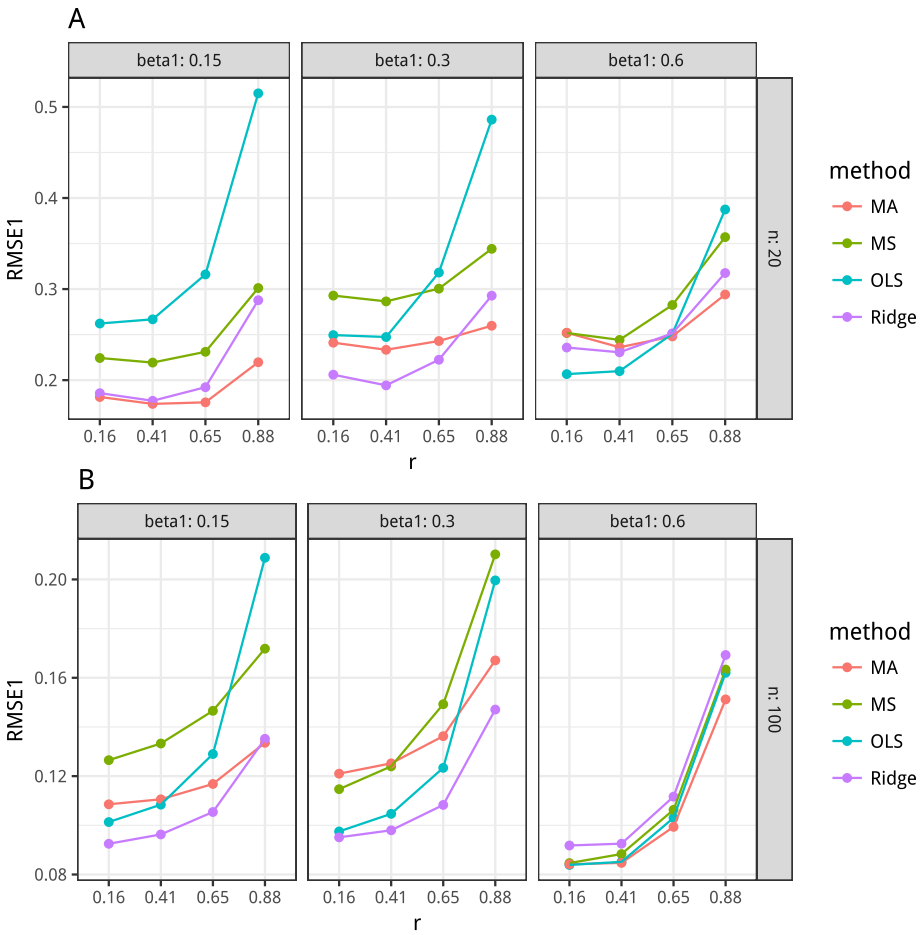


Figure 2: **Error as a function of increased collinearity**. Root mean square error in the estimate of β_1 over the 2000 runs for each combination of β_1 (the focal effect), α_1 (the generating coefficient that controls the level of collinearity, r, between x_1 and x_2), and n (sample size). Both x_1 and x_2 have a small correlation with a third predictor (x_3). A) n=20. B) n=100.

performs best when the correlation is large.

The simulation shows increased RMSE for model-averaged estimates as collinearity (the correlation among predictors) increases. In this sense, a model-averaged estimate is more misleading as collinearity increases (which results in "larger changes in the scaling of units" among models). However the effect of the correlation on the RMSE is stronger for the full model; it's estimates are even more misleading with increased collinearity.

Conclusion

Given the long and rich history of model averaging within statistics, its application to a diverse array of problems, and its relationship to other well known methods, model averaging as a method probably does not need defending. Additionally, I am not advocating model averaging as the best choice among the many for analyzing relationships between variables, but simply defending it as a meaningful choice. Ultimately, this would seem to be an empirical issue, which can often only be obtained via simulation, where the true effects are known. Many ecologists are careful to use the language of "prediction" when describing multiple regression results of observational studies but slip into causal language when discussing the importance of the results. This is okay. We should embrace causal interpretation if our goal is understanding the consequences of intervention in a system. But what if we truly are interested only in prediction? Do coefficients conditioned on different co-variates have the same meaning in a predictive model? I would argue yes, as these coefficients are simply weights with no other implied meaning.

Acknowledgments

I thank three colleagues and two anonymous reviewers for improving the clarity of the manuscript.

References

Banner, K. M. and M. D. Higgs, 2017. Considerations for assessing model averaging of regression coefficients. *Ecological Applications* **27**:78–93.

Bartoń, K., 2015. MuMIn: Multi-Model Inference. R package version 1.15.1.

Burnham, K. P. and D. R. Anderson, 2002. Model selection and multimodel inference, 2nd Edition. Springer-Verlag, New York, NY.

Cade, B. S., 2015. Model averaging and muddled multimodel inferences. *Ecology* **96**:2370–2382.

Claeskens, G., N. L. Hjort, and others, 2008. Model selection and model averaging, volume 330. Cambridge University Press Cambridge.

Freckleton, R. P., 2010. Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. *Behavioral Ecology and Sociobiology* **65**:91–101.

Greenland, S., M. Maclure, J. J. Schlesselman, C. Poole, and H. Morgenstern, 1991. Standardized regression coefficients: a further critique and review of some alternatives. *Epidemiology* **2**:387–392.

Hansen, B. E., 2007. Least squares model averaging. *Econometrica* **75**:1175–1189.

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

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

Liang, H., G. Zou, A. T. Wan, and X. Zhang, 2012. Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association* **106**:1053–1066.

Madigan, D. and A. E. Raftery, 1994. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association* **89**:1535–1546.

Møller, A. P. and M. D. Jennions, 2002. How much variance can be explained by ecologists and evolutionary biologists? *Oecologia* **132**:492–500.

Pearl, J., 2009. Causal inference in statistics: An overview. *Statistics Surveys* **3**:96–146.

R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

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

Shipley, B., 2002. Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference. Cambridge: Cambridge University Press.

Sokal, R. R. and F. J. Rohlf, 2012. Biometry. W H Freeman & Company.

Venables, W. N. and B. D. Ripley, 2002. Modern Applied Statistics with S. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Walker, J. A., 2014. The effect of unmeasured confounders on the ability to estimate a true performance or selection gradient (and other partial regression coefficients). *Evolution* **68**:2128–2136.

Wright, S., 1934. The method of path coefficients. *The Annals of Mathematical Statistics* **5**:161–215.