On Model Averaging Partial Regression Coefficients

Jeffrey A. Walker*[a]

[a]Department of Biological Sciences, University of Southern Maine, Portland, ME 04103, USA

December 4, 2017

Running Head: Model averaging

*walker@maine.edu

Abstract

Model averaging partial regression coefficients has been criticized because coefficients conditioned on different covariates estimate regression parameters with different interpretations from model to model. This criticism ignores (or rejects) the long tradition of using a partial regression coefficient to estimate an effect parameter (or Average Causal Effect), which gives the direct generating or causal effect of an independent variable on the response variable. The regression parameter is a descriptor and its meaning is conditional on the covariates in the model. It makes no claims about causal or generating effects. By contrast, an effect parameter derives its meaning from a causal model and not from a 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 any predictor has the same meaning across all submodels if the goal is estimation of the causal effects that generated the response. In a recent article, Cade (2015) went beyond this "different parameter" criticism and suggested that, in the presence of *any* multicollinearity, averaging partial regression coefficients is invalid because they have no defined units. I argue that Cade's interpretation of the math is incorrect. While partial regression coefficients may be meaningfully averaged, model averaging may not be especially useful. To clarify this, 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 full model estimates, as multicollinearity increases, despite the full regression model correctly specifying the causal effect structure (that is, even when we know the truth, a method that averages over incorrectly specified models outperforms the correctly specified model).

Keywords multiple regression, average causal effect, 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 (Draper, 1995; Hoeting et al., 1999; Burnham and Anderson, 2002). For effect estimation, model averaging is attractive because it recognizes that for data typical in biology, *all* measured predictors will have some non-zero association with the response variable independent of that shared with other predictors. Consequently, model averaging encourages the worthy goal of emphasizing effect estimation and not simply the identification of some "best" subset of predictors. Model-selection often follows an all-subsets regression, a practice that is criticized for mindless model building. Nevertheless, averaging across all or a best subset of models shrinks regression coefficients toward zero, which has the effect of contracting error variance. Consequently, model-averaging can outperform model selection and even the full model under some conditions, where performance is measured by a summary of the long run frequency of error (Raftery et al., 1997; Hjort and Claeskens, 2003).

Despite these features of model-averaging, model averaged partial regression coefficients have been criticized in the recent ecology literature because their computation requires averaging over a set of coefficients that are conditional on a specific set of covariates (Cade, 2015; Banner and Higgs, 2017). That is, coefficients from different models have different interpretations and cannot be meaningfully averaged. Cade (2015) took this criticism further, and specifically argued that in the presence of any multicollinearity, averaging coefficients is invalid because an averaged coefficient has "no defined units." Cade's criticism is not the typical caution against the estimation of partial regression coefficients in the presence of high multicollinearity because of a high variance inflation factor but an argument that model-averaged coefficients in the presence of *any* correlation among the predictors are, quite literally, meaningless. Cade's critique is receiving much attention, as evidenced by the 108 Google Scholar citations in about two years.

These critiques are noteworthy given that model averaging regression coefficidents has developed a rich literature in applied statistics over the last 20 years (Hoeting et al., 1999; Burnham and Anderson, 2002; Hjort and Claeskens, 2003; Hansen, 2007; Liang et al., 2011; Zhang et al., 2014; Zigler and Dominici, 2014) with only limited attention to the meaning of the parameter estimated by a model averaged coefficient (Draper, 1999; Candolo et al., 2003; Raftery and Zheng, 2003). Berger et al. (2001) noted the issue not in the context of a meaningless average but in the context of modeling the prior distribution. Consonni and Veronese (2008) also considered the meaning of the parameters in a submodel and showed four different interpretations. In two of these (their interpretations M_A^* and M_B^*), the parameter for a regression coefficient in a submodel has the same meaning as that in the full model. Specifically, consider the full model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ and the submodel $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$. β_1 is the same parameter in both models if we consider the submodel to be the full model with $\beta_2 = 0$. This "zero effect" interpretation is effectively that given by Hoeting et al. (1999) in their response to Draper (1999).

Here, I offer a defense of model averaging partial regression coefficients that is related to this zero effect interpretation of the parameters: the coefficients $b_{j.m}$ from different submodels m can be meaningfully averaged because they estimate an *effect parameter* (β_j) common to all models, where an effect parameter (or Average Causal Effect) is the direct causal or generating effect of X_j on Y (Angrist and Pischke, 2008; Pearl, 2009). In short, while a partial regression coefficient is a conditional statistic, it can be used to estimate two different parameters, a regression parameter (a conditional "effect" that functions as a descriptor) and an effect parameter (a causal effect that states how something was generated). If a researcher wishes to *describe* a statistical relationship conditional on a specific set of covariates, then model averaging would indeed be averaging different things and an averaged value would have an awkward (or not especially useful) interpretation. Often, however, researchers use multiple regression to explicitly (or more commonly implicitly) estimate the causal effects that *generated the data*. Importantly, an effect parameter derives its meaning from a pre-specified causal hypothesis and this meaning is independent of the set of variables in the full model (Pearl, 2009). Consequently, averaging estimates of these parameters is perfectly meaningful.

I begin my paper with a motivating example. I then extend Grace and Bollen (2005) by employing well-known, formal definitions of two different (effect and regression) parameters in order to address the criticism that model averaging regression coefficients averages over different things. I use path models to clarify these concepts. I then address Cade's specific criticism that averaging partial regression coefficients is invalid because these coefficients have different units. Finally, after arguing that averaging partial regression is meaningful, I address the question, "is it useful?", with a simulation. The goal of the simulation is not meant to be an exhaustive exploration of model averaging but simply to show that under conditions of low to moderate power, model averaged regression coefficients outperform estimates from the full model even when the full regression model correctly identifies the causal structure.

A motivating example: the causal effect of parental sex on offspring calling in owls

A recent article on best practices involving regression-like models (Zuur and Ieno, 2016) used as an example the data of Roulin and Bersier (2007), who showed that – and entitled their paper – "nestling barn owls beg more intensely in the presence of their mother than in the presence of their father." This title might simply be a description of the major result, that is, a difference in conditional means (on the set of covariates in the study, including time of arrival, time spent in nestbox, a food manipulation treatment, and all interactions with parental sex). In the discussion, however, Roulin and Bersier (2007) state that "differential begging to mother and father implies that offspring should be able to recognize the identity of each parent." That is, the chick behavior is in direct response to the sex of the parent, or, put differently, the sex of the parent bringing food causally modifies the intensity of chick calling.

This example serves to introduce the major argument of the present note: the parameters estimated by the coefficients of a linear model used for *causal* inference are fundamentally different from the parameters estimated by the coefficients of a linear model used for *describing* differences in conditional means. A partial regression coefficient $b_{j.m}$ is a difference in conditional means – it is the difference in the mean response between two groups that vary in x_j by one unit but have the same values for all other covariates (X_m). The partial regression coefficient $b_{j.m}$ estimates two parameters. The first is the familiar regression (conditional effect) parameter, a descriptive parameter describing the population difference in conditional means

$$\theta_{j.m} = E(Y|X_j = x_j + 1, X_m = x_m) - E(Y|X_j = x_j, X_m = x_m) \quad (1)$$

The second is the effect parameter, or Average Causal Effect, which is the direct, generating or causal effect of X_j on Y, and variously defined as

$$\beta_j = E(Y_{x_j=x+1} - Y_{x_j=x}) \quad \text{(Rubin, 1974)} \tag{2}$$

$$\beta_j = E(Y|do(X_j = x+1)) - E(Y|do(X_j = x)) \quad \text{(Pearl, 1995, 2009)} \tag{3}$$

Equation 2 is the counterfactual definition of a causal effect while equation 3 is an interventional definition of a causal effect. The counterfactual definition is what would happen if we could measure individuals under two conditions but only X_j has changed. The *do* operator represents what would happen in a hypothetical intervention that modifies X_j but leaves all other variables unchanged (Pearl, 2009). In both definitions, the meaning of β_j is not conditional on other X (Definition 2 and Equation 5 in Pearl, 1995). In the formal language of graphical causal models, an effect coefficient's meaning is derived from a pre-specified causal hypothesis of a *potential effect* in the form of a directed path from X_j to Y. The absence of an arrow is a hypothesis of no causal effect. By contrast, the presence of an arrow allows for empirical estimates that are close to, or effectively, zero, and in this way an effect parameter is similar to the "null effect" interpretation of the parameters of the full model described above (Hoeting et al., 1999; Consonni and Veronese, 2008).

The concept of effect coefficients goes back to beginning of multiple regression, by George Yule, who developed least squares multiple regression in order to estimate the causal effects of the changing demographics of pauperism of 19th century Britain (Yule, 1899)(partial regression coefficients were first published by Yule's mentor and colleague Karl Pearson three years earlier). Importantly, Yule's conception of cause was effectively that encoded by the *do*-operator (many others at the time essentially equated causation with correlation, see for example Niles, 1922). The concept of graphical causal models was first developed by the seminal work of Sewell Wright (1921, 1934) in his method of path analysis. Wright did not develop path analysis to discover causal relationships but to quantify causal effects from a pre-specified causal hypothesis in the form of paths (arrows) connecting causes (causal variables) to effects (response variables) (see below). Wright used partial regression coefficients as the effect (path) coefficients. By contrast to these causal uses of regression coefficients, the "difference in conditional means" concept of a regression coefficient began to emerge only following Fisher (1922).

A reasonable concern is, how does an effect parameter, which represents the effect of hypothetical differences in X_j with all other X unchanged, apply to observational data, where a change in the value of X_j is always associated with changes in other predictor variables? The answer is, the formal definitions clarify the assumptions needed to use a partial regression coefficient as an estimate of an effect parameter. More specifically, a partial regression coefficient $b_{j.m}$ is a consistent estimate of the effect parameter β_j if the regression model correctly identifies the causal structure and does not exclude confounding variables. A confounder of the effect of X_j on Y is any variable that both causally effect Y by a path independent of that through X_j and is correlated with X_j. A partial regression coefficient is a biased estimate of β_j if the regression model excludes confounders for X_j – a bias known as omitted variable bias. Yule (1899) explicitly recognized and discussed the consequence of omitted confounders in the first multiple regression analysis.

In contrast to effect coefficients, a partial regression coefficient $b_{j.m}$ is not a biased estimate of $\theta_{j.m}$ if other X that both contribute to Y and are correlated to X_j are omitted from the regression model, because here $b_{j.m}$ is estimating a parameter conditional on the same m. Consequently, omitted variable bias and confounding are irrelevant or meaningless in the context of regression as mere description. Not surprisingly then, omitted variable bias and confounding are introductory textbook concepts in disciplines that commonly use regression for causal modeling, including econometrics and epidemiology, but not disciplines where explicit causal modeling is uncommon, including biology generally, and ecology and evolution, specifically (but see Shipley, 2002; Pugesek et al., 2003).

Path models clarify the difference between β_j and $\theta_{j.m}$

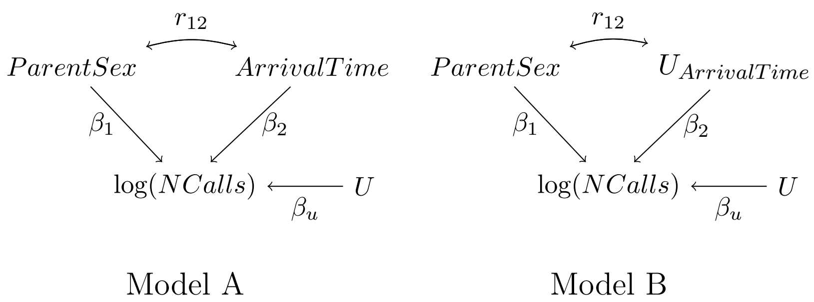

Figure 1:

To highlight the difference between the parameters, $\theta_{j.m}$ and β_j, I use the Roulin and Bersier (2007) example and show two generative models of $\log(NCalls)$ (Figure 1). The effect coefficients β_1 and β_2 are the direct, causal effect of $ParentSex$ and $ArrivalTime$ on Y. In Model A, $ArrivalTime$ is measured and included in the regression model. In Model B, $ArrivalTime$ is unmeasured and designated $U_{ArrivalTime}$. Both models also include U, which represents all un-modeled factors, other than $ArrivalTime$, that contribute to the variance in Y. The generating models also include a correlation r_{12} between $ParentSex$ and $ArrivalTime$. An important assumption of these models is that the unmeasured factor U is not correlated with either of the X_j as indicated by a lack of a double-headed arrow (otherwise it would be a confounding variable – see below).

In Model A, the partial regression coefficients $b_{1.2}$ (the coefficient of X_1 conditioned on X_2) and $b_{2.1}$ are unbiased estimates of the parameters $\theta_{1.2}$ and $\theta_{2.1}$. Because the regression model includes all confounders, the regression coefficients are also unbiased estimates of the generating parameters β_1 and β_2. In Model A, then, the parameters $\theta_{j.m}$ and β_j coincide, in the sense that both have the same value. This coincidence occurs only under very limited conditions.

Model B is the same generating model as Model A, but the regression model omits $ArrivalTime$. The effect parameter β_1 has precisely the same meaning in Model B as it did in Model A, but the regression coefficient b_1 is a biased estimate of β_1 because of the omitted variable. This bias is $r_{12}\beta_2$. By contrast, the regression parameter θ_1 in Model B has a different meaning than $\theta_{1.2}$ in Model A, and differs from the latter by $r_{12}\beta_2$, but the regression coefficient b_1 is an unbiased estimate of θ_1. In Model B, then, the parameters θ_1 and β_1 do not coincide.

Note that the missing confounder $ArrivalTime$ in Model B does not bias the estimate of θ_1 but does bias the estimate of β_1 (again, by $r_{12}\beta_2$). Consequently, if the goal is mere description, omitted confounders are not a concern. But if the goal is causal modeling, missing confounders are (or should be) a major concern. Omitted confounders result in standard errors of effect coefficients that are (often far) too small, which results in inflated confidence in effect magnitudes and even signs (Walker, 2014).

Importantly, while the meaning and/or value of the regression parameter $\theta_{j.m}$ differs among the models, the meaning and value of the effect parameter β_j is constant among the models. Because the meaning of the β_j is invariable across generating models, estimates of β_j have the same meaning – they are estimates of β_j– across regression models, regardless of the set of covariates specified in the model. And because partial regression coefficients *as estimates of* β_j have the same meaning across regression models, partial regression coefficients from different models can be meaningfully averaged.

Partial regression coefficients from different models have the same units

In addition to the "different meanings" criticism, Cade (2015) argues that model averaging is invalid if there is *any* correlation among predictors because model-averaged coefficients "have no defined units in the presence of multicollinearity." It is therefore imperative that we explore what Cade means by "No defined units." This might mean that 1) the units of $b_{j.m}$ differ among submodels, or 2) a unit difference in X_j differs among submodels (Table 1). The first interpretation is simply false; a partial regression coefficient $b_{j.m}$ has the units of the simple regression coefficient of Y on X_j regardless of the other predictors in the model (Supplement 1). Cade clarifies the second interpretation using the Frisch-Waugh decomposition of $\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{y})$

$$b_{j.m} = \frac{\mathrm{COV}(\tilde{X}_{j.m}, Y)}{\mathrm{VAR}(\tilde{X}_{j.m})} \tag{4}$$

where $\tilde{X}_{j.m}$ is the component of the variation of X_j not shared with the other X, which is simply the vector of residuals of the regression of X_j on the set of covariates X_m. Because the unshared variance ($\tilde{X}_{j.m}$) shrinks and swells from model to model, I interpret Cade as stating (Table 1) that the units of X_j itself shrinks and swells from model to model. And, consequently, a unit difference in X_j shrinks and swells from model to model. This conclusion is a misunderstanding of the math; the magnitude of a unit or unit difference is defined by the actual units and not the variance of $\tilde{X}_{j.m}$. The units of the residuals of weight on height are full kilograms, not partial kilograms. In the owl example, if $ArrivalTime$ is measured in hours, a one unit difference is one hour regardless if we are referring to the raw measures or the residuals of $ArrivalTime$ on $ParentSex$. One hour (or one kilogram or one degree Celsius) does not shrink or swell

Table 1: Two interpretation of statements in Cade (2015)

Statement	Interpretation
1. "Their AIC model averaging of regression coefficients acts as if they are just numbers without any units attached to them" 2. "It is impossible to interpret the model-averaged regression coefficients (Tables 2–5) in terms of a $\Delta y/\Delta X_i$ because we do not know what units should apply to the denominator because it no longer refers to any specific covariance structure among the predictor variables"	the units of $b_{j.m}$ differ among models
3. "Multicollinearity implies that the scaling of units in the denominators of the regression coefficients may change across models such that neither the parameters nor their estimates have common scales" 4. "the model averaging of regression coefficients ends up being done across estimates ($\beta_i = \Delta y/\Delta X_i$) without common denominators and is nonsensical because a unit change in the predictor variable (ΔX_i) is not the same across all models."	a unit difference in X_j differs among models

among models due to differences in the magnitude of VAR($\tilde{X}_{j.m}$).

Model-averaged coefficients outperform full-model coefficients when power is low to moderate

Even though partial regression coefficients as estimates of effect coefficients can be meaningfully averaged, the averaged coefficients may not be very useful, compared to the coefficients computed from the full model. Here, I show that, despite averaging over incorrectly specified models, model-averaged coefficients can outperform the full-model coefficients for estimating effect parameters even when the full regression model correctly identifies the generating model. I use a Monte Carlo simulation experiment to measure the total (root mean square) error in the estimates relative to the known, generating parameters. Freckleton (2010) used a similar 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. Here I extend these results by combining both forms of error into one measure. Because I am specifically comparing the relative performance of the estimators at different levels of multicollinearity, I also compare ridge regression estimates. The simulation is not meant to be a comprehensive comparison of model averaging estimators but simply a pedagogical case study of why model average estimators should be considered as reasonable alternatives to the full model.

Data simulating the owl call data (Roulin and Bersier, 2007) were generated with

$$NCalls \sim \text{Poisson}(\mu_i) \tag{5}$$

$$\log(\mu_i) = \beta_0 + \beta_1 ParentSex_i + \beta_2 ArrivalTime_i \tag{6}$$

$$ParentSex_i = Z_i \tag{7}$$

$$ArrivalTime_i \sim N(\beta_z Z_i, 1) \tag{8}$$

$$Z_i = 0 \text{ or } 1 \tag{9}$$

$NCalls$ are sampled for $n = 27$ nests, once for each parent. There is no nest effect in the generation of the data (nor is one modeled in the regression). Exp(β_0) is the expected number of calls (175) during the mother's visit. Exp(β_1) was set to one of four values 0.99, 0.98, 0.97, and 0.96, which is equivalent to standardized effects (Cohen's d) of -0.13, -0.27, -0.40, and -0.53. The expected reductions in calls during the father's visits holding $ArrivalTime$ constant are 1.75, 3.50, 5.25, and 7.00. Exp(β_2) was set to 0.9. Z is the common cause of $ParentSex$ and $ArrivalTime$, which creates a correlation (collinearity) between the two effects. Z and $ParentalSex$ are equal numerically (female=0, male= 1) but are not conceptually equivalent. Z is the sex determining factor while $ParentalSex$ is the phenotypic feature that allows a chick to identify the parent as dad or mom. The expected correlation between $ParentSex$ and $ArrivalTime$ is set (using the parameter β_z) to one of four values: 0.2, 0.4, 0.6, and 0.8. Each iteration, the empirical correlation is checked and only used if it is within 0.02 of the expected correlation. 5000 iterations were run for each combination of β_1 (controlling effect size and power) and β_z (controlling collinearity). The power to reject a null direct effect ($\beta_1 = 0$) at a type I error rate of 5% was computed using the 5000 runs for each combination of the causal parameters β_1 and β_z. The performance (the ability to estimate β_1) of model averaged, full model, and ridge estimates were quantified using the long-run error $RMSE = \sqrt{\frac{\sum(b_1 - \beta_1)^2}{5000}}$, which accounts for both error variance and bias. For the model-averaged estimates, b_1 is the model-averaged coefficient.

The entire simulation was implemented in R (R Core Team, 2015) and the script is available in Supplement 1. In each run, the generating coefficients were estimated using the full model, model averaging, and ridge regression. Model-averaged coefficients were computed using AICc weights and over all models using the dredge and 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 cv.glmnet function (setting alpha=0) in the glmnet package (Friedman et al., 2010) and used the default 10-fold cross-validation to compute the optimal tuning parameter.

The performance of the three methods are shown in (Figure 2, where the X-axis is the effect parameter of $ParentSex$ standardized by the average expected variance to generalize the results. More specifically, $\beta_1' = \beta_1/\sigma$, where σ is the square root of $(\text{Exp}(\beta_0)/2 + \text{Exp}(\beta_0 + \beta_1)/2)$. Again, these standardized effects are -0.13, -0.27, -0.40, and -0.53. While -0.13 is a “small” standardized effect, this is a common value in ecology, where an estimated average standardized effect size is 0.18-0.19 (Møller and Jennions, 2002). The label for each panel shows the mean correlation (over the 5000 runs) between $ParentSex$ and $ArrivalTime$ (again, all of the correlations within a batch of 5000 runs were within 0.02 of the expected correlation). Power for the four standardized effects ranged from 0.06 - 0.08, 0.09 - 0.16, 0.13 - 0.30, and 0.20 - 0.47 (within each effect size, power decreased with increased r_{12}).

The simulation shows increased RMSE for all estimators as collinearity increases and the qualitative pattern of relative performance among models remains about the same as collinearity increases (Figure 2). Quantitatively, the full model RMSE increases 71% (averaged over the four levels of β_1 as the correlation increases from 0.2 to 0.8. By contrast, the model averaged and ridge RMSE increase 40% and 37%. Model averaging outperforms the full model when power is relatively low despite the full model correctly specifying the generating model. When collinearity is high, power is relatively low at all tested effect sizes of β_1' and, consequently, model averaging outperforms the full model at all levels of β_1'. Ridge regression outperforms model averaging over much of the space except at the lowest power.

Conclusion

Banner and Higgs (2017) prefer the descriptive use and language of multiple regression, especially in observational studies that do not give "careful attention to principles and methods of causal modeling," an opinion of which I'm sympathetic (Walker, 2014). In their abstract, Banner and Higgs (2017) state that the "use of model averaging implicitly assumes the same parameter exists across models so that averaging is sensible. While this assumption may initially seem tenable, regression coefficients associated with particular explanatory variables may not hold equivalent interpretations across

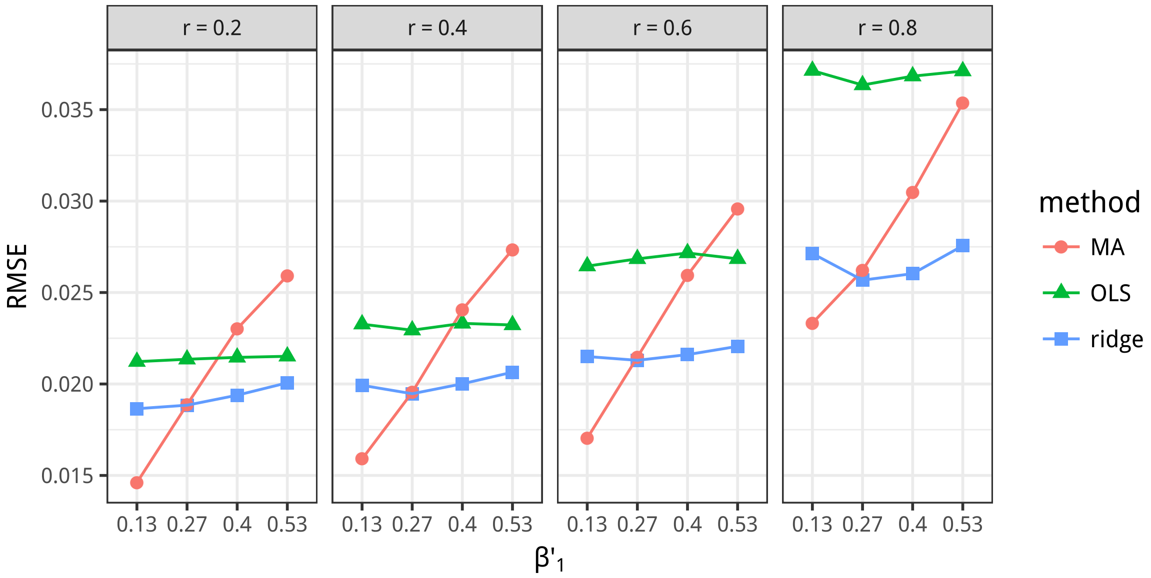

Figure 2: **Error as a function of increased collinearity**. Root mean square error in the estimate of β_1 (the effect parameter of $ParentSex$) over the 5000 runs for each combination of β_1 and the correlation r between $ParentSex$ and $ArrivalTime$. The effect parameter is standardized and presented as Cohen's d. Key to methods: MA = model average, OLS = full model, Ridge = ridge regression.

all of the models in which they appear, making explanatory inference about covariates challenging." This statement fails to recognize that an effect parameter β_j is distinct from a regression (difference in conditional means) parameter $\theta_{j.m}$ and the former but not the latter has the same meaning across models, because an effect parameter takes its meaning from a specified causal hypothesis and not the combination of variables in a regression model (Pearl, 2009) (Figure 1). As emphasized by Pearl (2009), the two parameters only coincide if the regression model correctly specifies the generating model.

Cade (2015) explicitly rejects simulation experiments similar to that above, arguing that "the statistical performance suggested by distributions of their simulated model-averaged estimates is of questionable merit" because model averaging is "nonsensical" because the averaged coefficients have no defined units. Even if we ignore what Cade means by "units" and agree that model averaging averages over coefficients with different meanings, I find the conclusion surprising; if a method has pretty good empirical results relative to other estimators, I'd be inclined to use it, even if it's not entirely intellectually satisfying (Breiman, 2001). But averaging partial regression coefficients is intellectually satisfying if causal modeling because both a unit difference in X_j and the parameter estimated by $b_{j.m}$ is the same among models.

Both Cade (2015) and Banner and Higgs (2017) consider model averaging the predicted outcome to be sensible, because the estimated parameter is the same among models. This computation is also uncontroversial in the applied statistics literature. But if Cade's interpretation of partial regression coefficients is correct (that these contain partial units of X_j), how do we generate a prediction with sensible units, as this computation requires postmultiplication of the model matrix, which includes full units of X, by the vector of coefficients, which have partial units of X? Cade's interpretation implies that all model-averaging is meaningless and the rich literature developed over the last twenty years should simply be rejected. More generally, if we accept the "averaging over coefficients with different meanings" criticism, then we must throw out additional tools in our kit. For example, meta-analysis requires averaging effects over multiple studies, many of which have been conditioned on different sets of covariates. And the measured effects from randomized experiments that are conditioned on covariates would no longer be estimates of average causal effects but could only be interpreted as conditional treatment effects that are irrelevant to the larger population.

Given the long and rich history of model averaging within several applied fields, including economics and epidemiology, 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. Here, I am advocating neither naive multiple regression for causal modeling nor model averaging as the best choice among many for estimating effect parameters, but simply defending model-averaged regression coefficients as a meaningful choice.

Acknowledgments

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

References

Angrist, J. D. and J.-S. Pischke, 2008. Mostly harmless econometrics. An Empiricist's Companion. Princeton University Press.

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

Berger, J. O., L. R. Pericchi, J. K. Ghosh, T. Samanta, F. De Santis, J. O. Berger, and L. R. Pericchi, 2001. Objective Bayesian methods for model selection: introduction and comparison. *Lecture Notes-Monograph Series* pages 135–207.

Breiman, L., 2001. Statistical Modeling: The Two Cultures. *Statistical Science* **16**:199–231.

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.

Candolo, C., A. C. Davison, and C. G. B. Demétrio, 2003. A note on model uncertainty in linear regression. *Journal of the Royal Statistical Society: Series D (The Statistician)* **52**:165–177.

Consonni, G. and P. Veronese, 2008. Compatibility of Prior Specifications Across Linear Models. *Statistical Science* **23**:332–353.

Draper, D., 1995. Assessment and Propagation of Model Uncertainty. *Assessment* **57**:45–97.

Draper, D., 1999. Comment—Bayesian model averaging: a tutorial. *Statistical Science* **14**:405–409.

Fisher, R. A., 1922. The Goodness of Fit of Regression Formulae, and the Distribution of Regression Coefficients. *Journal of the Royal Statistical Society* **85**:597–612.

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.

Friedman, J., T. Hastie, and R. Tibshirani, 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**:1–22.

Grace, J. B. and K. A. Bollen, 2005. Interpreting the Results from Multiple Regression and Structural Equation Models. *Bulletin of the Ecological Society of America* **86**:283–295.

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. K. Wan, and X. Zhang, 2011. Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association* **106**:1053–1066.

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

Niles, H. E., 1922. Correlation, causation and Wright's theory of" path coefficients". *Genetics* **7**:258.

Pearl, J., 1995. Causal diagrams for empirical research. *Biometrika* **82**:669–688.

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

Pugesek, B. H., A. Tomer, and A. Von Eye, 2003. Structural equation modeling: applications in ecological and evolutionary biology. Cambridge University Press.

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.

Raftery, A. E. and Y. Zheng, 2003. Discussion: performance of bayesian model averaging. *Journal of the American Statistical Association* **98**:931–938.

Roulin, A. and L.-F. Bersier, 2007. Nestling barn owls beg more intensely in the presence of their mother than in the presence of their father. *Animal Behaviour* **74**:1099–1106.

Rubin, D. B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **66**:688.

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

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., 1921. Correlation and causation. *Journal of agricultural research* **20**:557–585.

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

Yule, G. U., 1899. An Investigation into the Causes of Changes in Pauperism in England, Chiefly During the Last Two Intercensal Decades (Part I.). *Journal of the Royal Statistical Society* **62**:249.

Zhang, X., G. Zou, and H. Liang, 2014. Model averaging and weight choice in linear mixed-effects models. *Biometrika* **101**:205–218.

Zigler, C. M. and F. Dominici, 2014. Uncertainty in Propensity Score Estimation: Bayesian Methods for Variable Selection and Model-Averaged Causal Effects. *Journal of the American Statistical Association* **109**:95–107.

Zuur, A. F. and E. N. Ieno, 2016. A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution* **7**:636–645.