

# On Model Averaging Partial Regression Coefficients

Jeffrey A. Walker<sup>\*a</sup>

<sup>a</sup>Department of Biological Sciences, University of Southern Maine,  
Portland, ME 04103, USA

December 4, 2017

1 Running Head: Model averaging

---

\*walker@maine.edu

## 2 Abstract

3 Model averaging partial regression coefficients has been criticized because coefficients  
4 conditioned on different covariates estimate regression parameters with different inter-  
5 pretations from model to model. This criticism ignores (or rejects) the long tradition of  
6 using a partial regression coefficient to estimate an effect parameter (or Average Causal  
7 Effect), which gives the direct generating or causal effect of an independent variable on  
8 the response variable. The regression parameter is a descriptor and its meaning is con-  
9 ditional on the covariates in the model. It makes no claims about causal or generating  
10 effects. By contrast, an effect parameter derives its meaning from a causal model and  
11 not from a set of covariates. A multiple regression model implicitly specifies a causal  
12 model with direct, causal paths from each predictor to the response. Consequently,  
13 the partial regression coefficient for any predictor has the same meaning across all sub-  
14 models if the goal is estimation of the causal effects that generated the response. In a  
15 recent article, Cade (2015) went beyond this “different parameter” criticism and sug-  
16 gested that, in the presence of *any* multicollinearity, averaging partial regression coef-  
17 ficients is invalid because they have no defined units. I argue that Cade’s interpreta-  
18 tion of the math is incorrect. While partial regression coefficients may be meaningfully  
19 averaged, model averaging may not be especially useful. To clarify this, I compare ef-  
20 fect estimates using a small Monte-Carlo simulation. The simulation results show that  
21 model-averaged (and ridge) estimates have increasingly better performance, relative  
22 to full model estimates, as multicollinearity increases, despite the full regression model  
23 correctly specifying the causal effect structure (that is, even when we know the truth, a  
24 method that averages over incorrectly specified models outperforms the correctly speci-  
25 fied model).

26 **Keywords** multiple regression, average causal effect, causal model, conditional effect,

27 monte carlo simulation, ridge regression, model selection, multicollinearity.

## 28 Introduction

29 Model averaging is an alternative to model selection for either effect estimation or pre-  
30 diction (Draper, 1995; Hoeting et al., 1999; Burnham and Anderson, 2002). For effect  
31 estimation, model averaging is attractive because it recognizes that for data typical in  
32 biology, *all* measured predictors will have some non-zero association with the response  
33 variable independent of that shared with other predictors. Consequently, model av-  
34 eraging encourages the worthy goal of emphasizing effect estimation and not simply  
35 the identification of some “best” subset of predictors. Model-selection often follows an  
36 all-subsets regression, a practice that is criticized for mindless model building. Never-  
37 theless, averaging across all or a best subset of models shrinks regression coefficients  
38 toward zero, which has the effect of contracting error variance. Consequently, model-  
39 averaging can outperform model selection and even the full model under some condi-  
40 tions, where performance is measured by a summary of the long run frequency of error  
41 (Raftery et al., 1997; Hjort and Claeskens, 2003).

42 Despite these features of model-averaging, model averaged partial regression coef-  
43 ficients have been criticized in the recent ecology literature because their computation  
44 requires averaging over a set of coefficients that are conditional on a specific set of co-  
45 variates (Cade, 2015; Banner and Higgs, 2017). That is, coefficients from different mod-  
46 els have different interpretations and cannot be meaningfully averaged. Cade (2015)  
47 took this criticism further, and specifically argued that in the presence of any multi-  
48 collinearity, averaging coefficients is invalid because an averaged coefficient has “no de-  
49 fined units.” Cade’s criticism is not the typical caution against the estimation of partial  
50 regression coefficients in the presence of high multicollinearity because of a high vari-

51 ance inflation factor but an argument that model-averaged coefficients in the presence  
52 of *any* correlation among the predictors are, quite literally, meaningless. Cade’s critique  
53 is receiving much attention, as evidenced by the 108 Google Scholar citations in about  
54 two years.

55 These critiques are noteworthy given that model averaging regression coefficients  
56 has developed a rich literature in applied statistics over the last 20 years (Hoeting et al.,  
57 1999; Burnham and Anderson, 2002; Hjort and Claeskens, 2003; Hansen, 2007; Liang  
58 et al., 2011; Zhang et al., 2014; Zigler and Dominici, 2014) with only limited attention  
59 to the meaning of the parameter estimated by a model averaged coefficient (Draper,  
60 1999; Candolo et al., 2003; Raftery and Zheng, 2003). Berger et al. (2001) noted the  
61 issue not in the context of a meaningless average but in the context of modeling the  
62 prior distribution. Consonni and Veronese (2008) also considered the meaning of the  
63 parameters in a submodel and showed four different interpretations. In two of these  
64 (their interpretations  $M_A^*$  and  $M_B^*$ ), the parameter for a regression coefficient in a sub-  
65 model has the same meaning as that in the full model. Specifically, consider the full  
66 model  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$  and the submodel  $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$ .  $\beta_1$  is  
67 the same parameter in both models if we consider the submodel to be the full model  
68 with  $\beta_2 = 0$ . This “zero effect” interpretation is effectively that given by Hoeting et al.  
69 (1999) in their response to Draper (1999).

70 Here, I offer a defense of model averaging partial regression coefficients that is re-  
71 lated to this zero effect interpretation of the parameters: the coefficients  $b_{j,m}$  from dif-  
72 ferent submodels  $m$  can be meaningfully averaged because they estimate an *effect pa-*  
73 *rameter* ( $\beta_j$ ) common to all models, where an effect parameter (or Average Causal Ef-  
74 fect) is the direct causal or generating effect of  $X_j$  on  $Y$  (Angrist and Pischke, 2008;  
75 Pearl, 2009). In short, while a partial regression coefficient is a conditional statistic,

76 it can be used to estimate two different parameters, a regression parameter (a condi-  
77 tional “effect” that functions as a descriptor) and an effect parameter (a causal effect  
78 that states how something was generated). If a researcher wishes to *describe* a sta-  
79 tistical relationship conditional on a specific set of covariates, then model averaging  
80 would indeed be averaging different things and an averaged value would have an awk-  
81 ward (or not especially useful) interpretation. Often, however, researchers use multiple  
82 regression to explicitly (or more commonly implicitly) estimate the causal effects that  
83 *generated the data*. Importantly, an effect parameter derives its meaning from a pre-  
84 specified causal hypothesis and this meaning is independent of the set of variables in  
85 the full model (Pearl, 2009). Consequently, averaging estimates of these parameters is  
86 perfectly meaningful.

87 I begin my paper with a motivating example. I then extend Grace and Bollen (2005)  
88 by employing well-known, formal definitions of two different (effect and regression) pa-  
89 rameters in order to address the criticism that model averaging regression coefficients  
90 averages over different things. I use path models to clarify these concepts. I then ad-  
91 dress Cade’s specific criticism that averaging partial regression coefficients is invalid be-  
92 cause these coefficients have different units. Finally, after arguing that averaging partial  
93 regression is meaningful, I address the question, “is it useful?”, with a simulation. The  
94 goal of the simulation is not meant to be an exhaustive exploration of model averaging  
95 but simply to show that under conditions of low to moderate power, model averaged  
96 regression coefficients outperform estimates from the full model even when the full re-  
97 gression model correctly identifies the causal structure.

## 98 **A motivating example: the causal effect of parental sex** 99 **on offspring calling in owls**

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

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

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

120 The second is the effect parameter, or Average Causal Effect, which is the direct, gener-  
121 ating or causal effect of  $X_j$  on  $Y$ , and variously defined as

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

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

122 Equation 2 is the counterfactual definition of a causal effect while equation 3 is an in-  
123 terventional definition of a causal effect. The counterfactual definition is what would  
124 happen if we could measure individuals under two conditions but only  $X_j$  has changed.  
125 The *do* operator represents what would happen in a hypothetical intervention that  
126 modifies  $X_j$  but leaves all other variables unchanged (Pearl, 2009). In both definitions,  
127 the meaning of  $\beta_j$  is not conditional on other  $X$  (Definition 2 and Equation 5 in Pearl,  
128 1995). In the formal language of graphical causal models, an effect coefficient's mean-  
129 ing is derived from a pre-specified causal hypothesis of a *potential effect* in the form of  
130 a directed path from  $X_j$  to  $Y$ . The absence of an arrow is a hypothesis of no causal ef-  
131 fect. By contrast, the presence of an arrow allows for empirical estimates that are close  
132 to, or effectively, zero, and in this way an effect parameter is similar to the “null effect”  
133 interpretation of the parameters of the full model described above (Hoeting et al., 1999;  
134 Consonni and Veronese, 2008).

135 The concept of effect coefficients goes back to beginning of multiple regression, by  
136 George Yule, who developed least squares multiple regression in order to estimate the  
137 causal effects of the changing demographics of pauperism of 19th century Britain (Yule,  
138 1899)(partial regression coefficients were first published by Yule's mentor and colleague  
139 Karl Pearson three years earlier). Importantly, Yule's conception of cause was effec-  
140 tively that encoded by the *do*-operator (many others at the time essentially equated  
141 causation with correlation, see for example Niles, 1922). The concept of graphical causal

142 models was first developed by the seminal work of Sewell Wright (1921, 1934) in his  
143 method of path analysis. Wright did not develop path analysis to discover causal re-  
144 lationships but to quantify causal effects from a pre-specified causal hypothesis in the  
145 form of paths (arrows) connecting causes (causal variables) to effects (response vari-  
146 ables) (see below). Wright used partial regression coefficients as the effect (path) co-  
147 efficients. By contrast to these causal uses of regression coefficients, the “difference in  
148 conditional means” concept of a regression coefficient began to emerge only following  
149 Fisher (1922).

150 A reasonable concern is, how does an effect parameter, which represents the ef-  
151 fect of hypothetical differences in  $X_j$  with all other  $X$  unchanged, apply to observa-  
152 tional data, where a change in the value of  $X_j$  is always associated with changes in  
153 other predictor variables? The answer is, the formal definitions clarify the assump-  
154 tions needed to use a partial regression coefficient as an estimate of an effect parame-  
155 ter. More specifically, a partial regression coefficient  $b_{j,m}$  is a consistent estimate of the  
156 effect parameter  $\beta_j$  if the regression model correctly identifies the causal structure and  
157 does not exclude confounding variables. A confounder of the effect of  $X_j$  on  $Y$  is any  
158 variable that both causally effect  $Y$  by a path independent of that through  $X_j$  and is  
159 correlated with  $X_j$ . A partial regression coefficient is a biased estimate of  $\beta_j$  if the re-  
160 gression model excludes confounders for  $X_j$  – a bias known as omitted variable bias.  
161 Yule (1899) explicitly recognized and discussed the consequence of omitted confounders  
162 in the first multiple regression analysis.

163 In contrast to effect coefficients, a partial regression coefficient  $b_{j,m}$  is not a biased  
164 estimate of  $\theta_{j,m}$  if other  $X$  that both contribute to  $Y$  and are correlated to  $X_j$  are omit-  
165 ted from the regression model, because here  $b_{j,m}$  is estimating a parameter conditional  
166 on the same  $m$ . Consequently, omitted variable bias and confounding are irrelevant or

167 meaningless in the context of regression as mere description. Not surprisingly then,  
 168 omitted variable bias and confounding are introductory textbook concepts in disciplines  
 169 that commonly use regression for causal modeling, including econometrics and epidemi-  
 170 ology, but not disciplines where explicit causal modeling is uncommon, including biol-  
 171 ogy generally, and ecology and evolution, specifically (but see Shipley, 2002; Pugsek  
 172 et al., 2003).

173 **Path models clarify the difference between  $\beta_j$  and  $\theta_{j.m}$**

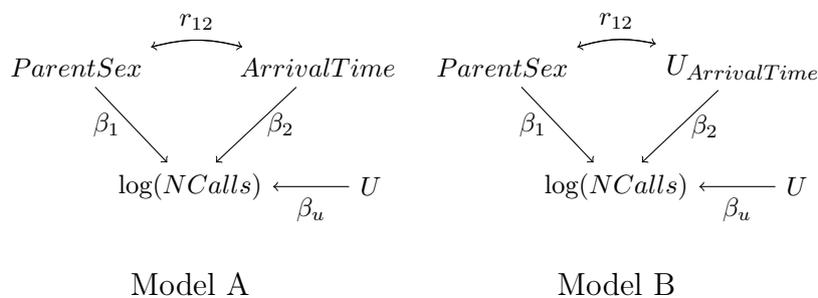


Figure 1:

174 To highlight the difference between the parameters,  $\theta_{j.m}$  and  $\beta_j$ , I use the Roulin  
 175 and Bersier (2007) example and show two generative models of  $\log(\text{NCalls})$  (Figure 1).  
 176 The effect coefficients  $\beta_1$  and  $\beta_2$  are the direct, causal effect of *ParentSex* and *ArrivalTime*  
 177 on  $Y$ . In Model A, *ArrivalTime* is measured and included in the regression model. In  
 178 Model B, *ArrivalTime* is unmeasured and designated  $U_{\text{ArrivalTime}}$ . Both models also  
 179 include  $U$ , which represents all un-modeled factors, other than *ArrivalTime*, that con-  
 180 tribute to the variance in  $Y$ . The generating models also include a correlation  $r_{12}$  be-  
 181 tween *ParentSex* and *ArrivalTime*. An important assumption of these models is that  
 182 the unmeasured factor  $U$  is not correlated with either of the  $X_j$  as indicated by a lack  
 183 of a double-headed arrow (otherwise it would be a confounding variable – see below).

184 In Model A, the partial regression coefficients  $b_{1.2}$  (the coefficient of  $X_1$  conditioned

185 on  $X_2$ ) and  $b_{2,1}$  are unbiased estimates of the parameters  $\theta_{1,2}$  and  $\theta_{2,1}$ . Because the re-  
186 gression model includes all confounders, the regression coefficients are also unbiased es-  
187 timates of the generating parameters  $\beta_1$  and  $\beta_2$ . In Model A, then, the parameters  $\theta_{j,m}$   
188 and  $\beta_j$  coincide, in the sense that both have the same value. This coincidence occurs  
189 only under very limited conditions.

190 Model B is the same generating model as Model A, but the regression model omits  
191 *ArrivalTime*. The effect parameter  $\beta_1$  has precisely the same meaning in Model B as  
192 it did in Model A, but the regression coefficient  $b_1$  is a biased estimate of  $\beta_1$  because  
193 of the omitted variable. This bias is  $r_{12}\beta_2$ . By contrast, the regression parameter  $\theta_1$  in  
194 Model B has a different meaning than  $\theta_{1,2}$  in Model A, and differs from the latter by  
195  $r_{12}\beta_2$ , but the regression coefficient  $b_1$  is an unbiased estimate of  $\theta_1$ . In Model B, then,  
196 the parameters  $\theta_1$  and  $\beta_1$  do not coincide.

197 Note that the missing confounder *ArrivalTime* in Model B does not bias the esti-  
198 mate of  $\theta_1$  but does bias the estimate of  $\beta_1$  (again, by  $r_{12}\beta_2$ ). Consequently, if the goal  
199 is mere description, omitted confounders are not a concern. But if the goal is causal  
200 modeling, missing confounders are (or should be) a major concern. Omitted confounders  
201 result in standard errors of effect coefficients that are (often far) too small, which re-  
202 sults in inflated confidence in effect magnitudes and even signs (Walker, 2014).

203 Importantly, while the meaning and/or value of the regression parameter  $\theta_{j,m}$  dif-  
204 fers among the models, the meaning and value of the effect parameter  $\beta_j$  is constant  
205 among the models. Because the meaning of the  $\beta_j$  is invariable across generating mod-  
206 els, estimates of  $\beta_j$  have the same meaning – they are estimates of  $\beta_j$  – across regression  
207 models, regardless of the set of covariates specified in the model. And because partial  
208 regression coefficients *as estimates of  $\beta_j$*  have the same meaning across regression mod-  
209 els, partial regression coefficients from different models can be meaningfully averaged.

## 210 Partial regression coefficients from different models have 211 the same units

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

$$b_{j,m} = \frac{\text{COV}(\tilde{X}_{j,m}, Y)}{\text{VAR}(\tilde{X}_{j,m})} \quad (4)$$

221 where  $\tilde{X}_{j,m}$  is the component of the variation of  $X_j$  not shared with the other  $X$ , which  
222 is simply the vector of residuals of the regression of  $X_j$  on the set of covariates  $X_m$ .  
223 Because the unshared variance ( $\tilde{X}_{j,m}$ ) shrinks and swells from model to model, I inter-  
224 pret Cade as stating (Table 1) that the units of  $X_j$  itself shrinks and swells from model  
225 to model. And, consequently, a unit difference in  $X_j$  shrinks and swells from model to  
226 model. This conclusion is a misunderstanding of the math; the magnitude of a unit or  
227 unit difference is defined by the actual units and not the variance of  $\tilde{X}_{j,m}$ . The units  
228 of the residuals of weight on height are full kilograms, not partial kilograms. In the  
229 owl example, if *ArrivalTime* is measured in hours, a one unit difference is one hour  
230 regardless if we are referring to the raw measures or the residuals of *ArrivalTime* on  
231 *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"	the units of $b_{j,m}$ differ among models
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"	
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"	a unit difference in $X_j$ differs among models
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 ( $\Delta X_i$ ) is not the same across all models."	

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

## 233 **Model-averaged coefficients outperform full-model co-** 234 **efficients when power is low to moderate**

235 Even though partial regression coefficients as estimates of effect coefficients can be  
236 meaningfully averaged, the averaged coefficients may not be very useful, compared to  
237 the coefficients computed from the full model. Here, I show that, despite averaging over  
238 incorrectly specified models, model-averaged coefficients can outperform the full-model  
239 coefficients for estimating effect parameters even when the full regression model cor-  
240 rectly identifies the generating model. I use a Monte Carlo simulation experiment to  
241 measure the total (root mean square) error in the estimates relative to the known, gen-  
242 erating parameters. Freckleton (2010) used a similar simulation to show how full model  
243 and model-averaged estimates perform with increased multicollinearity and showed that

244 the error variance of the full model estimates increases more rapidly than that of the  
245 model-averaged estimates but that the model-averaged estimates were increasingly bi-  
246 ased with increased multicollinearity. Here I extend these results by combining both  
247 forms of error into one measure. Because I am specifically comparing the relative per-  
248 formance of the estimators at different levels of multicollinearity, I also compare ridge  
249 regression estimates. The simulation is not meant to be a comprehensive comparison of  
250 model averaging estimators but simply a pedagogical case study of why model average  
251 estimators should be considered as reasonable alternatives to the full model.

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

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

$$\log(\mu_i) = \beta_0 + \beta_1 \text{ParentSex}_i + \beta_2 \text{ArrivalTime}_i \quad (6)$$

$$\text{ParentSex}_i = Z_i \quad (7)$$

$$\text{ArrivalTime}_i \sim N(\beta_z Z_i, 1) \quad (8)$$

$$Z_i = 0 \text{ or } 1 \quad (9)$$

253 *NCalls* are sampled for  $n = 27$  nests, once for each parent. There is no nest effect in  
254 the generation of the data (nor is one modeled in the regression).  $\text{Exp}(\beta_0)$  is the ex-  
255 pected number of calls (175) during the mother's visit.  $\text{Exp}(\beta_1)$  was set to one of four  
256 values 0.99, 0.98, 0.97, and 0.96, which is equivalent to standardized effects (Cohen's  $d$ )  
257 of -0.13, -0.27, -0.40, and -0.53. The expected reductions in calls during the father's vis-  
258 its holding *ArrivalTime* constant are 1.75, 3.50, 5.25, and 7.00.  $\text{Exp}(\beta_2)$  was set to 0.9.  
259  $Z$  is the common cause of *ParentSex* and *ArrivalTime*, which creates a correlation  
260 (collinearity) between the two effects.  $Z$  and *ParentalSex* are equal numerically (fe-  
261 male=0, male= 1) but are not conceptually equivalent.  $Z$  is the sex determining factor

262 while *ParentalSex* is the phenotypic feature that allows a chick to identify the parent  
263 as dad or mom. The expected correlation between *ParentSex* and *ArrivalTime* is set  
264 (using the parameter  $\beta_z$ ) to one of four values: 0.2, 0.4, 0.6, and 0.8. Each iteration,  
265 the empirical correlation is checked and only used if it is within 0.02 of the expected  
266 correlation. 5000 iterations were run for each combination of  $\beta_1$  (controlling effect size  
267 and power) and  $\beta_z$  (controlling collinearity). The power to reject a null direct effect  
268 ( $\beta_1 = 0$ ) at a type I error rate of 5% was computed using the 5000 runs for each com-  
269 bination of the causal parameters  $\beta_1$  and  $\beta_z$ . The performance (the ability to estimate  
270  $\beta_1$ ) of model averaged, full model, and ridge estimates were quantified using the long-  
271 run error  $RMSE = \sqrt{\frac{\sum(b_1 - \beta_1)^2}{5000}}$ , which accounts for both error variance and bias. For  
272 the model-averaged estimates,  $b_1$  is the model-averaged coefficient.

273 The entire simulation was implemented in R (R Core Team, 2015) and the script is  
274 available in Supplement 1. In each run, the generating coefficients were estimated us-  
275 ing the full model, model averaging, and ridge regression. Model-averaged coefficients  
276 were computed using AICc weights and over all models using the dredge and model.avg  
277 function in the MuMIn package. Coefficients of predictors excluded from a model were  
278 assigned a value of zero (using the row of the coefficient table with the label "full").  
279 For the ridge regression, I used the cv.glmnet function (setting alpha=0) in the glmnet  
280 package (Friedman et al., 2010) and used the default 10-fold cross-validation to com-  
281 pute the optimal tuning parameter.

282 The performance of the three methods are shown in (Figure 2, where the  $X$ -axis  
283 is the effect parameter of *ParentSex* standardized by the average expected variance  
284 to generalize the results. More specifically,  $\beta'_1 = \beta_1/\sigma$ , where  $\sigma$  is the square root of  
285  $(\text{Exp}(\beta_0)/2 + \text{Exp}(\beta_0 + \beta_1)/2)$ . Again, these standardized effects are -0.13, -0.27, -0.40,  
286 and -0.53. While -0.13 is a “small” standardized effect, this is a common value in ecol-

287 ogy, where an estimated average standardized effect size is 0.18-0.19 (Møller and Jen-  
288 nions, 2002). The label for each panel shows the mean correlation (over the 5000 runs)  
289 between *ParentSex* and *ArrivalTime* (again, all of the correlations within a batch of  
290 5000 runs were within 0.02 of the expected correlation). Power for the four standard-  
291 ized effects ranged from 0.06 - 0.08, 0.09 - 0.16, 0.13 - 0.30, and 0.20 - 0.47 (within each  
292 effect size, power decreased with increased  $r_{12}$ ).

293 The simulation shows increased RMSE for all estimators as collinearity increases  
294 and the qualitative pattern of relative performance among models remains about the  
295 same as collinearity increases (Figure 2). Quantitatively, the full model RMSE increases  
296 71% (averaged over the four levels of  $\beta_1$  as the correlation increases from 0.2 to 0.8. By  
297 contrast, the model averaged and ridge RMSE increase 40% and 37%. Model averag-  
298 ing outperforms the full model when power is relatively low despite the full model cor-  
299 rectly specifying the generating model. When collinearity is high, power is relatively  
300 low at all tested effect sizes of  $\beta'_1$  and, consequently, model averaging outperforms the  
301 full model at all levels of  $\beta'_1$ . Ridge regression outperforms model averaging over much  
302 of the space except at the lowest power.

## 303 Conclusion

304 Banner and Higgs (2017) prefer the descriptive use and language of multiple regres-  
305 sion, especially in observational studies that do not give “careful attention to principles  
306 and methods of causal modeling,” an opinion of which I’m sympathetic (Walker, 2014).  
307 In their abstract, Banner and Higgs (2017) state that the “use of model averaging im-  
308 plicitly assumes the same parameter exists across models so that averaging is sensi-  
309 ble. While this assumption may initially seem tenable, regression coefficients associated  
310 with particular explanatory variables may not hold equivalent interpretations across

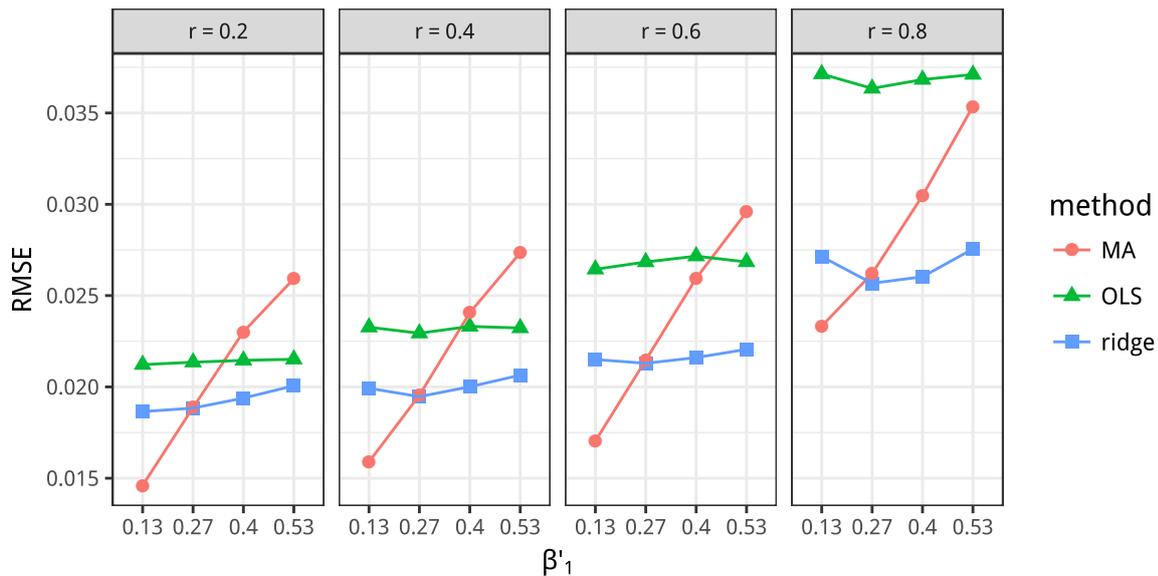


Figure 2: **Error as a function of increased collinearity.** Root mean square error in the estimate of  $\beta_1$  (the effect parameter of *ParentSex*) over the 5000 runs for each combination of  $\beta_1$  and the correlation  $r$  between *ParentSex* and *ArrivalTime*. The effect parameter is standardized and presented as Cohen’s  $d$ . Key to methods: MA = model average, OLS = full model, Ridge = ridge regression.

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

319 Cade (2015) explicitly rejects simulation experiments similar to that above, arguing  
 320 that “the statistical performance suggested by distributions of their simulated model-  
 321 averaged estimates is of questionable merit” because model averaging is “nonsensical”  
 322 because the averaged coefficients have no defined units. Even if we ignore what Cade  
 323 means by “units” and agree that model averaging averages over coefficients with differ-  
 324 ent meanings, I find the conclusion surprising; if a method has pretty good empirical

325 results relative to other estimators, I'd be inclined to use it, even if it's not entirely in-  
326 tellectually satisfying (Breiman, 2001). But averaging partial regression coefficients is  
327 intellectually satisfying if causal modeling because both a unit difference in  $X_j$  and the  
328 parameter estimated by  $b_{j,m}$  is the same among models.

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

344 Given the long and rich history of model averaging within several applied fields, in-  
345 cluding economics and epidemiology, its application to a diverse array of problems, and  
346 its relationship to other well known methods, model averaging as a method probably  
347 does not need defending. Here, I am advocating neither naive multiple regression for  
348 causal modeling nor model averaging as the best choice among many for estimating ef-  
349 fect parameters, but simply defending model-averaged regression coefficients as a mean-

350 ingful choice.

## 351 Acknowledgments

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

## 353 References

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

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

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

361 Breiman, L., 2001. Statistical Modeling: The Two Cultures. *Statistical Science* **16**:199–  
362 231.

363 Burnham, K. P. and D. R. Anderson, 2002. Model selection and multimodel inference,  
364 2nd Edition. Springer-Verlag, New York, NY.

365 Cade, B. S., 2015. Model averaging and muddled multimodel inferences. *Ecology*  
366 **96**:2370–2382.

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

- 370 Consonni, G. and P. Veronese, 2008. Compatibility of Prior Specifications Across Lin-  
371 ear Models. *Statistical Science* **23**:332–353.
- 372 Draper, D., 1995. Assessment and Propagation of Model Uncertainty. *Assessment*  
373 **57**:45–97.
- 374 Draper, D., 1999. Comment—Bayesian model averaging: a tutorial. *Statistical Science*  
375 **14**:405–409.
- 376 Fisher, R. A., 1922. The Goodness of Fit of Regression Formulae, and the Distribution  
377 of Regression Coefficients. *Journal of the Royal Statistical Society* **85**:597–612.
- 378 Freckleton, R. P., 2010. Dealing with collinearity in behavioural and ecological data:  
379 model averaging and the problems of measurement error. *Behavioral Ecology and*  
380 *Sociobiology* **65**:91–101.
- 381 Friedman, J., T. Hastie, and R. Tibshirani, 2010. Regularization paths for generalized  
382 linear models via coordinate descent. *Journal of Statistical Software* **33**:1–22.
- 383 Grace, J. B. and K. A. Bollen, 2005. Interpreting the Results from Multiple Regres-  
384 sion and Structural Equation Models. *Bulletin of the Ecological Society of America*  
385 **86**:283–295.
- 386 Hansen, B. E., 2007. Least squares model averaging. *Econometrica* **75**:1175–1189.
- 387 Hjort, N. L. and G. Claeskens, 2003. Frequentist model average estimators. *Journal of*  
388 *the American Statistical Association* **98**:879–899.
- 389 Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999. Bayesian model  
390 averaging: a tutorial. *Statistical Science* **14**:382–417.

- 391 Liang, H., G. Zou, A. T. K. Wan, and X. Zhang, 2011. Optimal weight choice for fre-  
392 quentist model average estimators. *Journal of the American Statistical Association*  
393 **106**:1053–1066.
- 394 Møller, A. P. and M. D. Jennions, 2002. How much variance can be explained by eco-  
395 logists and evolutionary biologists? *Oecologia* **132**:492–500.
- 396 Niles, H. E., 1922. Correlation, causation and Wright's theory of " path coefficients".  
397 *Genetics* **7**:258.
- 398 Pearl, J., 1995. Causal diagrams for empirical research. *Biometrika* **82**:669–688.
- 399 Pearl, J., 2009. Causal inference in statistics: An overview. *Statistics Surveys* **3**:96–146.
- 400 Pugesek, B. H., A. Tomer, and A. Von Eye, 2003. Structural equation modeling: appli-  
401 cations in ecological and evolutionary biology. Cambridge University Press.
- 402 R Core Team, 2015. R: A Language and Environment for Statistical Computing. R  
403 Foundation for Statistical Computing, Vienna, Austria.
- 404 Raftery, A. E., D. Madigan, and J. A. Hoeting, 1997. Bayesian model averaging for  
405 linear regression models. *Journal of the American Statistical Association* **92**:179–191.
- 406 Raftery, A. E. and Y. Zheng, 2003. Discussion: performance of bayesian model averag-  
407 ing. *Journal of the American Statistical Association* **98**:931–938.
- 408 Roulin, A. and L.-F. Bersier, 2007. Nestling barn owls beg more intensely in the pres-  
409 ence of their mother than in the presence of their father. *Animal Behaviour* **74**:1099–  
410 1106.
- 411 Rubin, D. B., 1974. Estimating causal effects of treatments in randomized and nonran-  
412 domized studies. *Journal of educational Psychology* **66**:688.

- 413 Shipley, B., 2002. Cause and correlation in biology: a user's guide to path analysis,  
414 structural equations and causal inference. Cambridge: Cambridge University Press.
- 415 Walker, J. A., 2014. The effect of unmeasured confounders on the ability to estimate  
416 a true performance or selection gradient (and other partial regression coefficients).  
417 *Evolution* **68**:2128–2136.
- 418 Wright, S., 1921. Correlation and causation. *Journal of agricultural research* **20**:557–  
419 585.
- 420 Wright, S., 1934. The method of path coefficients. *The Annals of Mathematical Statis-*  
421 *tics* **5**:161–215.
- 422 Yule, G. U., 1899. An Investigation into the Causes of Changes in Pauperism in Eng-  
423 land, Chiefly During the Last Two Intercensal Decades (Part I.). *Journal of the*  
424 *Royal Statistical Society* **62**:249.
- 425 Zhang, X., G. Zou, and H. Liang, 2014. Model averaging and weight choice in linear  
426 mixed-effects models. *Biometrika* **101**:205–218.
- 427 Zigler, C. M. and F. Dominici, 2014. Uncertainty in Propensity Score Estimation:  
428 Bayesian Methods for Variable Selection and Model-Averaged Causal Effects. *Journal*  
429 *of the American Statistical Association* **109**:95–107.
- 430 Zuur, A. F. and E. N. Ieno, 2016. A protocol for conducting and presenting results of  
431 regression-type analyses. *Methods in Ecology and Evolution* **7**:636–645.