On Model Averaging Partial Regression Coefficients

Jeffrey A. Walker*[a]

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

November 14, 2017

Running Head: Model averaging partial regression coefficients

*walker@maine.edu

Abstract

Model averaging of partial regression coefficients has been criticized for averaging over a set of models with coefficients that have different meanings from model to model. however, partial regression coefficients from different models have the same meaning if they are used to estimate the effect parameters of 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 any X_j 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 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 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 or 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 meaningless because coefficients from different models have different units and not simply different meanings. That is, model averaging requires the belief that partial regression coefficients are "just numbers without any units attached to them." 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 has developed a rich literature in applied statistics over the last 25 years (Hoeting et al., 1999; Burnham and Anderson, 2002; Hjort and Claeskens, 2003; Hansen, 2007; Liang et al., 2011; Zhang et al., 2014) with only limited attention to the meaning of the parameter estimated

by a model averaged coefficient (D. Draper comment in Hoeting et al., 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. For similar purposes, Consonni and Veronese (2008) showed that the parameters of a single submodel have four different interpretations. In one of these, the parameter for X_j in a submodel has the same meaning as that in the full model if we model the submodel as a variant of the full model with the parameters for the excluded predictors set to zero (Equation 1) (this is a more formal version of the reply given by Hoeting et al., 1999, in response to Draper).

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad \text{full model: } \beta_2 \neq 0 \tag{1}$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad \text{sub model: } \beta_2 = 0 \tag{2}$$

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 can be meaningfully averaged if they estimate a common effect parameter (β_j), which is the direct causal or generating effect of X_j on Y. In short, 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 are explicitly (or more commonly implicitly) interested in the coefficients as estimates of the effects that *generated the data*, that is, the effect parameters. 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. Here, I show this independence using the math of causal modeling with causal (path or

structural) diagrams (Wright, 1934; Shipley, 2002; Pearl, 2009).

A motivating example: the 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 a 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 $b_{j.m}$ is an estimator of two parameters

$$\theta_{j.m} = E(Y|X_j = x_j + 1, X_m = x_m) - E(Y|X_j = x_j, X_m = x_m) \tag{3}$$

$$\beta_j = E(Y|do(X_j = x_j + 1)) - E(Y|do(X_j = x_j)) \tag{4}$$

$\theta_{j.m}$ is the partial regression parameter (the population difference in conditional means). It is a descriptive parameter. β_j is the effect parameter, the direct, generating effect of X_j on Y. The *do* operator represents what would happen in an experimental intervention that modifies X_j but leaves all other variables unchanged (Pearl, 2009). The *do* operator is what we *do* when we conduct a randomized experiment and when we simulate a data generating process (see below). The concept of effect coefficients goes back to at least the seminal work of Sewell Wright (1918, 1934) in his method of path analysis. Importantly, the definition of an effect parameter does not refer to other variables in the model. Again, β_j derives its meaning from a pre-specified causal hypothesis and this meaning is independent of which set of variables are in the full model (Pearl, 2009). $b_{j.m}$ is a biased estimate of β_j (but not $\theta_{j.m}$ – see below) if the model used to compute the partial regression coefficients excludes confounders – variables that both causally effect Y by a path independent of that through X_j and are correlated with X_j – a bias known as omitted variable bias. Omitted variable bias is an introductory textbook concept 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.

Model-averaged coefficients are meaningful estimates of effect parameters β_j

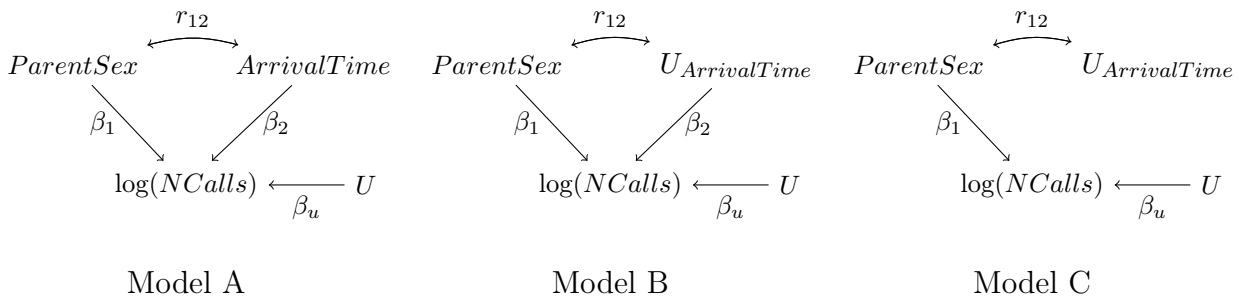

Figure 1:

To highlight the difference between the parameters, $\theta_{j.m}$ and β_j, I use the Roulin and Bersier (2007) example and show three simplified, 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 Models B and C (Figure 1), $ArrivalTime$ is unmeasured and designated $U_{ArrivalTime}$. All 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 $b_{1.2}$ and $b_{2.1}$ are unbiased estimates of the parameters $\theta_{1.2}$ and $\theta_{2.1}$. Because the model is correctly specified, 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 a unbiased estimate of θ_1. In Model B, then, the parameters θ_1 and β_1 do not coincide.

The generating Model C differs from Models A and B by the lack of a causal effect of $ArrivalTime$ on Y. The regression model in C, however, is precisely that in B. The effect parameter β_1 has the same meaning as in Models A and B, despite being a different generating model than both. And the regression coefficient b_1 is an unbiased estimate of β_1 (as in Model A but not B), because the regression model is correctly specified. The regression parameter θ_1 has the same meaning but different value as θ_1 in Model B but has the same value but different meaning as $\theta_{1.2}$ in Model A. Like model A, the parameters θ_1 and β_1 coincide because they share the same value (which will occur if the specified regression model is the generating model).

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.

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

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 specifies 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. Unfortunately, Freckleton (2010) 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 from his results which method performs better *overall*. 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

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

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

$$ArrivalTime_i \sim N(\beta_{2.1} ParentSex_i, 1) \tag{7}$$

$NCalls$ 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). $\text{Exp}(\beta_0)$ is the expected number of calls (175) during the mother's visit. $\text{Exp}(\beta_1)$ 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. $\text{Exp}(\beta_2)$ was set to 0.9. To create collinearity, $ArrivalTime$ is modeled as a function of $ParentSex$ with the magnitude $\beta_{2.1}$ set so the expected correlation between ParentSex and ArrivalTime is one of four values (0.2, 0.4, 0.6, 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 $\beta_{2.1}$ (controlling collinearity). With this generating model, $ParentSex$ has both a direct (β_1) and indirect ($\beta_{2.1}\beta_2$) effect on $NCalls$ but the focus is the direct effect. 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 $\beta_{2.1}$. The performance (the ability to estimate β_j) 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.

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 $ArrivalTime$ instead of the generating parameter $\beta_{2.1}$ (again, all of the correlations within a batch of 5000 runs were within 0.02 of the expected correlation).

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

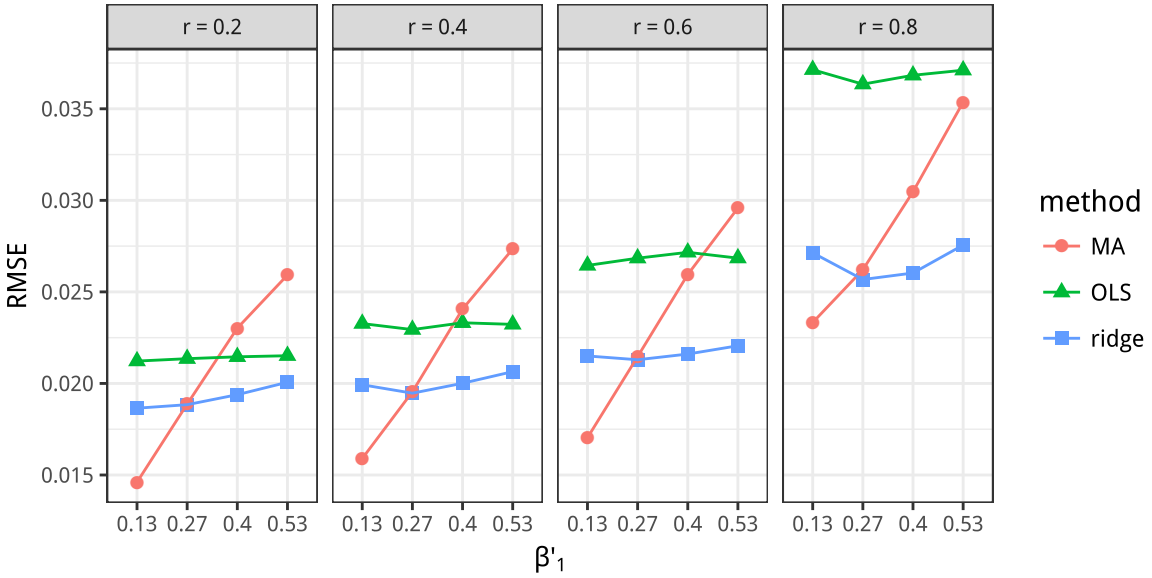

Figure 2: **Error as a function of increased collinearity**. Root mean square error in the estimate of β_1 (the effect parameter of $ParentSex$) over the 5000 runs for each combination of β_1 and the correlation r between $ParentSex$ and $ArrivalTime$. The effect parameter is standardized and presented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

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.

Partial regression coefficients from different models have the same units

If Cade (2015) were simply cautioning against model averaging when data have high multicollinearity, the simulation above would suggest otherwise, as it shows that model averaging is less affected by multicollinearity than the full model. But a warning is not Cade's point; instead, Cade argues that model averaging is meaningless if there is *any* correlation among predictors because model-averaged coefficients "have no defined units in the presence of multicollinearity." As a consequence, Cade (2015) simply rejects the only experiments where we know the truth and where we can empirically compare the

efficacy of different estimators.

Nevertheless, it is worth exploring what Cade means by "No defined units." This might mean: 1) that the actual units of $b_{j.m}$ differ among models and, consequently, the units of the averaged coefficient are undefined, or 2) that the units contain imply additional information and this additional information differs among models. Number one is simply false; a partial regression coefficient b_j has the units of the simple regression coefficient of Y on X_j regardless of the other predictors in the model (Supplement 1). Concerning number two, Cade clarifies this using the Frisch-Waugh decomposition of $\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{y})$

$$b_{j.m} = \frac{\mathrm{COV}(\tilde{X}_{j.m}, Y)}{\mathrm{VAR}(\tilde{X}_{j.m})} \tag{8}$$

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 variance of the residuals $\tilde{X}_{j.m}$ have different values among the models (because of different patterns of shared variance among the X), Cade (2015) concludes that "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." Or, elsewhere, "it is impossible to interpret the model-averaged regression coefficients 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." I think Cade is arguing that the units of $\tilde{X}_{j.m}$ differ among models because these are something like "partial" units, and because the magnitude of these partial units differs among models, the partial regression coefficients have "no defined units." This 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 among models due to differences in the magnitude of $\text{VAR}(\tilde{X}_{j.m})$. What does shrink or swell with differences in $\text{VAR}(\tilde{X}_{j.m})$ is the magnitude of the difference in $E(Y)$ given a one unit difference in X_j, but this is just re-stating the definition of a partial regression coefficient.

Conclusion

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 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 a causal hypothesis takes its meaning from a specified path and not the combination of variables in the 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. This is somewhat 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 five years should simply be rejected. Indeed, if we accept Cade's interpretation, 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.

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 not advocating model averaging as the best choice among the many for estimating effect parameters, but simply defending it as a meaningful choice. Ultimately, this may prove to be an empirical issue, which can often only be obtained via simulation, where the true effects are known.

Acknowledgments

I thank three colleagues and multiple 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.

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.

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.

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.

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.

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.

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., 1918. On the nature of size factors. *Genetics* **3**:367–374.

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

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

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.