On Model Averaging Regression Coefficients

Jeffrey A. Walker*[a]

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

October 8, 2018

Running Head: Model averaging

*walker@maine.edu

Abstract

Model-averaged partial regression coefficients have been criticized for averaging over a set of models with coefficients that have different meanings from model to model. This criticism arises because statisticians since Fisher believe that the meaning of a coefficient in a regression model arises from probabilistic conditioning ($P(Y|X)$) and that coefficients are mere descriptors of conditional association (or "differences in conditional means"). Because this association parameter is conditional on a specific set of covariates, the parameter for a predictor varies from model to model. The coefficients in many applied regression models, however, takes their meaning from causal conditioning ($P(Y|do(X))$) and these coefficients estimate causal effect parameters (or simply, causal effects or Average Treatment Effects). Causal effect parameters are also differences in conditional expectations, but the event conditioned on is not the set of covariates in a regression model but a hypothetical intervention. Because an effect parameter for a predictor takes its meaning from causal and not probabilistic conditioning, it is the same from model to model, and an averaged coefficient has a straightforward interpretation as an estimate of a causal effect. But, because the effect parameter is the same from model to model, the estimates of the parameter will be biased in all but at most one model. By contrast, with probabilistic conditioning, the coefficients are consistent estimates of their parameter in every model. Confounding and omitted variable bias, which are central to explanatory modeling, are meaningless in regression modeling as mere description.

Keywords multiple regression, causal effect, conditional effect, model selection.

Introduction

Statistical models have multiple purposes, including description, prediction, and causal explanation (Mac Nally, 2000; Shmueli, 2010). This variation in purpose is a source of much confusion on the interpretation of regression coefficients both in the primary literature and textbooks. In much of the ecology literature, the purpose of the modeling is not explicitly stated and is difficult to infer because the language used is ambiguous. Ecologists often avoid using causal language even if the questions pursued are, ultimately, causal. Larry Wassermen's comment that "there is always a causal interpretation lurking behind an association" (Wasserman, 1999) is often true in ecology. Here, I focus on statistical modeling for estimating parameters with causal meaning and the "interpretation" of these parameters. Interpretability of model coefficients is not a concern in purely predictive models *used for prediction only* and, consequently, researchers can exploit non-parametric and non-linear models for improved prediction (Breiman, 2001). By contrast, if we are modeling a system that we hope to manage with manipulations of key variables, or if key variables are changing due to anthropogenic effects (warming, acidification, urbanization), or if we simply want to understand how a system works, then we want our predictive models to estimate parameters with causal meaning.

Model simplification (or variable selection) using automated algorithms (stepwise or all-subsets regression) is a common modeling strategy in ecology. An alternative to variable selection is model averaging (Draper, 1995; Hoeting et al., 1999; Burnham and Anderson, 2002), in which the estimated parameter for X_j is

$$\hat{\theta} = \sum_{m=1}^{M} w_m \begin{pmatrix} \theta^{(m)} \\ \\ 0 \end{pmatrix} \tag{1}$$

where $\theta^{(m)}$ is some parameter estimate from model m and w_m is a weight based on some measure of model fit, generally penalized by the number of parameters in the model. The parameter averaged could be the predicted value for individual i or the coefficient (b_j) of predictor j (X_j). If the parameter being averaged is a coefficient and the predictor is not in model m, then $\theta^{(m)}$ is assigned the value 0. Consequently, a model averaged coefficient is, under some conditions, a shrinkage estimate and can have lower variance, at the cost of some bias, than an estimate from the full model (Hoeting et al., 1999; Hansen, 2007) (in the ecology literature, there has been some tendency to only average over models that include the parameter, in which case, the averaged coefficient is not a shrinkage estimate. In addition, a prediction computed from these "conditional" averages would not equal a model-averaged prediction).

Two extended criticisms of model-averaged coefficients have arisen in the recent ecology literature. Banner and Higgs (2017) echo the occasional comment from the statistics literature (Draper, 1999; Candolo et al., 2003; Raftery and Zheng, 2003; McElreath, 2018) that the coefficient $b_j^{(m)}$ from different models estimate different parameters with different meanings and, consequently, any interpretation of an averaged coefficient is awkward at best. Cade (2015) introduced a novel criticism. Specifically, he argued that in the presence of any correlation (his "multicollinearity") among the predictors, averaging coefficients is invalid because the units for a coefficient differ among models and, consequently, an averaged coefficient has "no defined units." Cade's criticism is neither the different-parameter criticism of model-averaged coefficients nor the typical caution against the estimation of regression coefficients in the presence of high collinearity (which some in the literature have mistakenly believed). Even if we agree that a regression coefficient from different models can estimate the same parameter, Cade's criticism is still potentially valid.

These criticisms of model-averaged coefficients have not been adequately addressed although there is some discussion in the statistics literature on the interpretation of the coefficients from a set of nested models. Berger et al. (2001) noted different potential interpretations but in the context of how to model the prior distribution and not in the context of a meaningless average. Consonni and Veronese (2008) also considered the meaning of the parameters in a sub-model and showed four different interpretations. In two of these (their interpretations M_A^* and M_B^*), the parameter for a regression coefficient in a sub-model 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 sub-model to be the full model with $\beta_2 = 0$. This "zero effect" interpretation is effectively the interpretation given by Hoeting et al. (1999) in their response to Draper (1999).

Here, I argue that model-averaging averages biased estimates of the same parameter and not consistent estimates of different parameters. Because the coefficients from different models estimate the same parameter, the interpretation (or meaning) is straightforward. While I am defending model-averaged coefficients as valid and meaningful constructs, I am specifically not defending model-averaging as the preferred estimator of model parameters.

Probabilistic vs. causal conditioning

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, and a food manipulation treatment). 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 probability density of the response variable in descriptive and explanatory models are conditioned on different events and consequently, the coefficients of descriptive and explanatory models estimate different parameters (Robins et al., 2000; Pearl, 2009a; Shalizi, 2017; Hitchcock, 2018). In descriptive modeling, the probability density of the response is

$$\mathrm{P}(Y = y|X = x) \tag{2}$$

where X is a set of variables that has been observed and included as covariates the statistical model. Shalizi (2017) refers to Equation 2 as probabilistic conditioning. By contrast, in explanatory modeling, the probability density of the response is

$$\mathrm{P}(Y = y|do(X = x)) \tag{3}$$

where the *do* operator specifies a real or hypothetical *intervention* that moves a specific set of X variable to the value x but leaves other X unchanged (Pearl, 2009b; Shalizi, 2017; Hitchcock, 2018). "Other X" is undefined; these could be other measured variables or unmeasured variables. It doesn't matter, because this distribution is not conditional on these other variables (Definition 2 and Equation 5 in Pearl, 1995). Shalizi (2017) refers to Equation 3 as causal conditioning.

Because of this difference in conditioning, the regression coefficients in descriptive and explanatory models estimate different parameters. A regression coefficient in descriptive modeling estimates the parameter

$$\alpha_{j.k} = E(Y|X_j = x_j + 1, X_k = x_k) - E(Y|X_j = x_j, X_k = x_k) \tag{4}$$

while a regression coefficient in explanatory modeling estimates the parameter

$$\beta_j = E(Y|do(X_j = x + 1)) - E(Y|do(X_j = x)) \quad \text{(Pearl, 1995, 2009a)} \tag{5}$$

α is a conditional association while β is a causal effect (Pearl (2009b) uses "effect coefficient"). $\alpha_{j.k}$ is often defined as a "difference in conditional means" but β_j is also a difference in conditional means. Their meaning differs because the event conditioned on differs. This event is the observation of specific X variables for $\alpha_{j.k}$ and a hypothetical intervention for β_j. Or, as Judea Pearl puts it, α is conditioned on "seeing" while β is conditioned on "doing" (Pearl and Mackenzie, 2018). Importantly, the definition (or interpretation) of β_j is not conditional on other X, even if these are measured and included in a regression model.

Both α and β are estimated with the coefficients of linear models (or models that are linear in their parameters), including classical linear models, generalized linear models, and hierarchical models. A difference between descriptive and explanatory modeling though is how we think about the model that generated the data (the "generating" or "true" model) (Table 1). In explanatory modeling, the true model is a model of the processes that actually generated the data. In a descriptive modeling, the true model is a model of the processes that *could* generate the data if the modeled variables were the only variables generating the data.

This difference in the generating model has consequences on inference (Hernán, 2018). When the parameters take their meaning from probabilistic conditioning (i.e. descriptive modeling), the generating model differs among regression models. By contrast, when the parameters take their meaning from causal conditioning, the generating model is the same among regression models (because it is a model of the actual generating processes) but any parameters that are missing from the regression model are swept into the error term of the generating model (since these aren't estimated by the regression model). The result is an error term that is correlated with the X variables in the generating model. But a regression model assumes uncorrelated errors and, consequently, in estimation, the error variance due to the missing parameter is passed to the other coefficients. The consequence on estimation is an incorrect residual (it is not correlated with the other X) and biased estimates of the causal parameters. This bias is due to the omitted X variables and is often called omitted variable bias. The missing X variables are missing confounders (more strictly, a variable is only a confounder of X_j if it is both correlated with X_j and has a causal effect on the response through some path that is independent of the X_j). In explanatory models, a modeled coefficient is a consistent estimate of the causal parameter β_j only if the statistical model correctly identifies the causal structure and does not exclude confounding variables, or if the statistical model uses well-known methods to block confounders (Pearl, 2009a).

Table 1: Generating models for a set of nested regression models differ between probabilistic and causal conditioning. Superscripts indicate different parameters among models

Nested Models	Generating Model
(1) $Y = b_0 + b_1X_1 + b_2X_2 + e$	Prob: $Y = \alpha_0^{(1)} + \alpha_1^{(1)}X_1 + \alpha_2^{(1)}X_2 + \varepsilon^{(1)}$ Causal: $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \varepsilon^{(1)}$
(2) $Y = b_0 + b_1X_1 + e$	Prob: $Y = \alpha_0^{(2)} + \alpha_1^{(2)}X_1 + \varepsilon^{(2)}$ Causal: $Y = \beta_0 + \beta_1X_1 + \varepsilon^{(3)}$, $\varepsilon^{(3)} = \beta_2X_2 + \varepsilon^{(1)}$

This occurs only when the parameters of the causal generating model coincide (have the same value) with the parameters of the probabilistic generating model (that is, $\beta_1 = \alpha_1$ and $\beta_2 = \alpha_2$, etc.). By contrast, a modeled coefficient in descriptive models is not a biased estimate of α_j if X with direct causal effects on Y are omitted from the statistical model. In fact, omitted variable bias and confounding are irrelevant or meaningless in the context of descriptive modeling, an important point that doesn't seem to be greatly appreciated (but see, for example, Gelman and Hill, 2007; Hernán, 2018). The recognition of this distinction in ecology is unknown, but I have had reviewers of two different manuscripts (this manuscript and (Walker, 2014)) explain to me that omitted variable bias is not really bias because OLS is an unbiased estimator. More generally, few textbooks recognize the existence of both probabilistic and causal conditioning, and most confuse them, using something like Equation 4 to define a coefficient but then using causal vocabulary ("how each X influences Y", "the change in Y given a one unit change in X", "how the X generated Y") to describe the meaning of the coefficients. Morgan and Winship (2007) give a short summary of this confusion. See also Gelman and Hill (2007). Robins et al. (2000), Shalizi (2017), and Hitchcock (2018) are very accessible accounts of the difference between probabilistic and causal conditioning and the association vs. causal parameters.

While the formal recognition of probabilistic vs. causal conditioning is fairly recent, the general idea goes back to the beginning of multiple regression, by George Yule, who used least squares multiple regression to estimate the causal effects of the changing demographics of pauperism of 19th century Britain (Yule, 1899; Freedman, 1999; Hepple, 2001) (partial regression coefficients were first developed as part of a "system of correlations" 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 statisticians at the time essentially equated causation with correlation, see for example Niles, 1922). The concept of causal parameters was formalized in 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. In contrast to these causal interpretations of a regression coefficient, the "difference in conditional means" concept of a regression coefficient began to emerge only following Fisher (1922), who seems to have been the first to think of regression as a function for a *probabilistic* conditional expectation (Aldrich, 2005). Pearl and Mackenzie (2018) present a very readable history of the conflict between probabilistic and causal conditioning.

The argument that model-averaged coefficients are meaningless because the coefficients from different models are conditioned on different covariates and, therefore, estimates different parameters, is true *only if* one is descriptive modeling. But, if one is model-averaging the coefficients, then the goal of the modeling is probably causal, even if the authors have been inculcated by the norms of their research community that their coefficients have an association interpretation only (Wasserman, 1999; Hernán, 2018). And if the goal is causal (explanatory inference), then coefficients from different models are estimating the same parameter – the casual parameter – and averaged coefficients are meaningful.

Regression coefficients from different models have the same units

In his influential 2015 criticism of model-averaged regression coefficients, Cade (2015) developed a unique argument against model averaging. Specifically, Cade argued 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." This was not a caution of estimation in the presence of high collinearity (the focus of other collinearity papers) but an argument that all model-averaged coefficients from any observational study are uninterpretable regardless of the magnitude of correlation among the predictors. It is therefore imperative that we understand what Cade means by "no defined units." This might mean that 1) the units of $b_{j.k}$ differ among models, or 2) a unit difference in X_j differs among models (Table 2). The first interpretation is simply false; a regression coefficient $b_{j.k}$ has the units of the simple regression coefficient of Y on X_j regardless of the other predictors in the model and there is nothing in the *units* of a regression coefficient that contains information that the value of the coefficient is conditional on some set of covariates (Supplement 1). Cade clarifies the second interpretation using the Frisch-Waugh decomposition of $\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{y})$

$$b_{j.k} = \frac{\mathrm{COV}(\tilde{X}_{j.k}, Y)}{\mathrm{VAR}(\tilde{X}_{j.k})} \tag{6}$$

where $\tilde{X}_{j.k}$ 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 ($\tilde{X}_{j.k}$) shrinks and swells from model to model, I interpret Cade as stating (Table 2) 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

Table 2: 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 $b_{j.k}$ 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

model. If it is invalid to average coefficients computed from Equation 6 because of variation in the denominator, then we should not average coefficients from the same model applied to different samples, as these denominators will also differ from sample to sample. But this is moot, because the magnitude of a unit or unit difference is defined by the actual units and not the variance of $\tilde{X}_{j.k}$. 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 $\mathrm{VAR}(\tilde{X}_{j.k})$.

Conclusion

I have argued that, if one is explanatory modeling, including predictive modeling where we hope to use the model to make informative interventions in the system, a model-averaged coefficient from nested regression models has a straightforward interpretation as an estimate of a causal effect parameter (equation 5). While this estimate is conditional on the other covariates in the model, its interpretation or "meaning" is not. Its interpretation takes its meaning from "causal conditioning" and not "probabilistic conditioning."

Averaging coefficients that potentially estimate different parameters is not unique to model-averaging of coefficients from nested regression models. Meta-analysis averages effects from different studies and these effects are commonly estimated with different sets of covariates. This averaging is justified because the coefficients are estimating causal parameters that derive their meaning from causal and not probabilistic conditioning. More problematically, meta-analysis averages effects across samples with different distributions of measured confounders. Because of this, Pearl Pearl (2012); Pearl and Bareinboim (2014) criticizes meta-analysis for averaging apples and oranges. But Pearl is not arguing that averaged effects from a meta-analysis fail to have a straightforward interpretation, only that we need more better tools for combining estimates. Regardless, this more insidious issue of meta-analysis is not relevant to model-averaged coefficients of nested regression models because the same sample is used for all estimates.

Some reviewers of earlier versions of the manuscript have been very negative not because they offered a critique of my argument but because of 1) issues related to how to model the prior if Bayesian model averaging or how to compute a weight if frequentist model averaging, 2) issues related to reliable estimates of standard errors, 3) is-

sues with generalized linear models in which a prediction computed as the average of predictions on the response scale is not equal to a back-transformed prediction computed from averaged coefficients on the link scale. These are interesting critiques but not relevant to my argument, only to model-averaging more generally. Two other very visceral criticisms on the manuscript are 1) model-averaging encourages mindless modeling, and 2) why average over models we know to be wrong (that is, we know they have missing confounders). Mindless modeling is not a property of an estimator but of a researcher, and no strategy is immune to mindless-modeling. Model-averaging is a tool, like OLS regression, or penalized regression (including ridge regression), hierarchical regression, or Bayesian regression. Using extensive prior knowledge to filter out thousands of small and trivial effects from a system, a careful researcher will construct a causal model with one or a few, focal predictor variables and all confounding variables that are believed to be important in the estimation of the focal effects. The goal is, given this causal model, what can we infer about the effects? The researcher then chooses a tool to estimate these effects. But why choose model-averaging, since we know this averages over models that we know to be wrong? One answer is that model-averaged estimates can have lower MSE than full model estimates as a result of smaller error variance and despite biased estimates (Freckleton, 2010). Regardless, I am not advocating that model-averaging is the optimal tool for estimating causal effects, only that it is a tool that generates coefficients with straightforward interpretations.

Acknowledgments

I thank Dr. Brian Inouye, Dr. Brian Cade, and multiple anonymous reviewers for improving the clarity of the manuscript.

References

Aldrich, J., 2005. Fisher and Regression. *Statistical Science* **20**:401–417.

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.

Freedman, D., 1999. From association to causation: Some remarks on the history of statistics. *Statistical Science* .

Gelman, A. and J. Hill, 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

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

Hepple, L. W., 2001. Multiple regression and spatial policy analysis: George Udny Yule and the origins of statistical social science. *Environment and Planning D: Society and Space* **19**:385–407.

Hernán, M. A., 2018. The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. *American Journal of Public Health* **108**:616–619.

Hitchcock, C., 2018. Causal Models. https://plato.stanford.edu/entries/causal-models/.

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

Mac Nally, R., 2000. Regression and model-building in conservation biology, biogeography and ecology: The distinction between–and reconciliation of–'predictive'and 'explanatory'models. *Biodiversity & Conservation* **9**:655–671.

McElreath, R., 2018. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. CRC Press.

Morgan, S. L. and C. Winship, 2007. Counterfactuals and Causal Inference. Methods and Principles for Social Research. Cambridge University Press.

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

Pearl, J., 2009b. Causality. Cambridge University Press.

Pearl, J., 2012. The Do-Calculus Revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 4–11, Catalina, CA. AUAI Press.

Pearl, J. and E. Bareinboim, 2014. External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science* **29**:579–595.

Pearl, J. and D. Mackenzie, 2018. The Book of Why: The New Science of Cause and Effect. Basic Books.

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

Robins, J. M., M. A. Hernán, and B. Brumback, 2000. Marginal Structural Models and Causal Inference in Epidemiology:. *Epidemiology* **11**:550–560.

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.

Shalizi, C. R., 2017. Advanced Data Analysis from an Elementary Point of View. \url{http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/}.

Shmueli, G., 2010. To Explain or to Predict? *Statistical Science* **25**:289–310.

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.

Wasserman, L., 1999. Comment. *Journal of the American Statistical Association* **94**:704–706.

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.

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.