Enhancing Validity in Observational Settings When Replication is Not Possible

Christopher J. Fariss and Zachary M. Jones

Abstract

We argue that political scientists can enhance the external validity of models of observational and quasi-experimental data, when reproduction but not replication is possible, by minimizing the expected prediction error on new data: generalization error. When data are not generated by a stochastic mechanism under the researcher's control (i.e., an experimental or survey design), reproduction is possible but replication is not, since the researcher cannot take another draw from the data generating process. This situation is quite common in studies of international relations and comparative politics in which researchers are essentially interested in obtaining and analyzing a census of relevant historical observations. Estimating the generalization error of a model for these data and then adjusting the model to minimize this estimate — regularization — increases external validity by decreasing the risk of overfitting, which provides information analogous to that obtained from the replication of a data generating process. Estimating generalization error also allows for model comparison that highlights underfitting: when a model generalizes poorly due to missing systematic features of the data generating process. This is important because providing additional evidence for the external validity of an empirical model increases the validity of the conclusions of a study.

Introduction

Political scientists can enhance the external validity and the conclusion validity of applied experimental and survey research when replicating the findings from existing studies. Replication involves taking a new draw from a data generating process and then repeating the procedures specified by the research design and is explicitly focused on the external validity of an inference. This type of replication is not possible in observational and quasi-experimental settings. As we argue in this research note however, the external validity of these types of studies can still be enhanced by using methods that minimize generalization error: the expected prediction error on new observations from the data generating process.[1] Generalization error is an unobserved measure of the external validity of a model's predictions. Minimizing generalization error requires the estimation of this unknown quantity and adjustment of the model to minimize it, which provides information analogous to that obtained from the replication of a data generating process. In the remainder of this paper, we first consider external validity, replication, and reproduction. We then discuss generalization error, its estimation, and techniques for adjusting models to minimize it. We close with a discussion of future directions.

External Validity, Replication, and Reproduction

Shadish (2010) defines external validity as the "validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables"

[1] See V. N. Vapnik and Vapnik (1998) for more statistical learning theory, that is, theoretical investigation of the ability of algorithms which construct models from data to accurately generalize to unseen data.

or outcome variables (4). The terms generalizability and external validity are synonymous with one another.[2] Generalization error is an estimate of external validity based on the model's ability to generate accurate predictions on new data from the data generating process. Generalization error is not an estimate of the validity of cause-effect relationships (internal validity), which may or may not be plausibly identified in a model fit to observational or quasi-experimental data.

We define replication as taking a new draw from a data generating process, which is distinct from reproduction, which entails reproducing the same findings given the same data and research design procedures. Reproduction represents a minimal standard of transparency for scientific research and is the concept commonly referred to as the "replication standard" in political science (Dafoe 2014).[3] To replicate an experimental design, a researcher might conduct a new experiment and attempt to find the same treatment effect(s) with a new sample drawn from the same target population of interest. A study based on survey data could be replicated by surveying a new set of individuals and then conducting the same statistical analysis on the new sample.[4] If the empirical relationships in these examples are generalizable to the population from which the new samples used for replication are drawn, then similar findings will result, subject to the uncertainty due to sampling. Thus, replication, unlike reproduction, provides evidence of the external validity of an empirical relationship.

When data are not generated by a process controlled by the researcher (i.e., by experimental manipulation or sampling) — replication in the sense described above is not possible (Berk 2004). If researchers view data collected from non-controllable processes as the result of a stochastic process however, they may want to assess how generalizable the model's predictions are.[5] Even quasi-experimental designs with strong evidence of internal validity are not replicable, as they take advantage of a unique exogenous shock to the social or political systems of interest (see Dunning 2012). Thus, the techniques we discuss in this paper can be used to provide evidence for the external validity of model predictions drawn from both observational and quasi-experimental designs in a way that is district from but related to one of the goals of replication.

External Validity and Over/Under-fitting

If the generalization error of a model is high, the predictions and substantive interpretation of the model are possibly unreliable. Whether or not the generalization error of a model is high, however, is not made apparent by looking at the prediction error on the data used for fitting, due to the possibility of overfitting. Overfitting occurs when non-systematic variation — the noise — is described by an empirical model, instead of systematic variation — the signal. By definition, an overfit model has high generalization error. As is commonly recognized, it is generally the case that models fit the data used for fitting much more closely than the data not used for fitting. A variety of procedures have been developed to prevent overfitting, though the use of these tools is still not common practice

[2] For more on validity generally, see Shadish, Cook, and Campbell (2001).

[3] See King (1995) and King (2006) for earlier discussion of the replication standard in political science and see Jones (2013) for a recent perspective on reproduction. Note also that "secondary research" as defined by Herrnson (1995) may not be clearly replication or reproduction, that is, this typology is not exhaustive.

[4] We make a distinction between an exact replication and a conceptual replication. An exact replication uses the same protocol with theoretically identical and practically similar subjects, settings, treatment variables, and outcome variables. A conceptual replication might change one or more of these components of the design. Both types of replications, in addition to reproduction, are useful starting points for new research depending on the goals of the researcher.

[5] When researchers move away from situations in which it is possible to achieve probabilistic equivalence or to randomly sample from a population of interest, we have much less evidence of the internal validity of the study design and therefore must rely on an argument or analogy made about the measurements of observed variables to random variables, which are theoretical constructs (Kass 2011). Insofar as this analogy is useful, the bounds on an estimate may be reliable. However, without knowing, by design, that part of the model is correct, we require information about the fit of the model to the data, as well as prior information (whether incorporated into the model or not) to judge how reasonable the probability or confidence bounds on an estimate are.

in political science (though see e.g. Ward, Greenhill, and Bakke 2010; Beck, King, and Zeng 2000; Hill Jr. and Jones 2014; Hainmueller and Hazlett 2013; Kenkel and Signorino 2013 for examples of articles which (at least implicitly) discuss this problem).

In general, the capacity of a method to overfit increases with its flexibility. However, flexible methods are attractive because of their ability to fit interesting patterns in the data, i.e., increased flexibility decreases a method's error on the data to which the method was fit, and, perhaps, across all possible data sets that could have been obtained. Finding the best method in terms of generalization error involves balancing the tradeoff between flexibility and the risk of overfitting.

In order to further elucidate this tradeoff, we provide some a theoretical exposition and a short Monte Carlo example. Here we define error to be measured using the familiar squared-error loss function minimized by ordinary least squares (OLS). We decompose this particular loss function to highlight the bias-variance tradeoff, which allows us to find the model with the lowest generalization error amongst the class of models considered.

Define R as a risk function, which is the expected loss over the data generating process. This is the quantity we want to minimize. If we suppose that $Y = f(X) + \epsilon$, where f (we sometimes suppress that f is a function of X for notational convenience) is a function which maps $X \to Y$ then we can write R as

$$R(f) = \mathbb{E}[(f(X) - Y)]^2 = \sigma^2$$

.

This simply states that if the true function mapping X to Y were known, then the only error that would be made in predictions is due to irreducible variability in Y. When f is not known this error rate is not achieved and it is useful to decompose the risk of the estimated function $\hat{f}$. $\hat{f}$ is estimated from a training set (a set of data used for fittng) ordered pairs (x_i, y_i), $i \in (1, \ldots, n)$ drawn from (X, Y). Specifically, the risk of $\hat{f}$ can be written as a sum of irreducible error, the squared bias of $\hat{f}$, and the variance of $\hat{f}$.

$$\begin{aligned} R(\hat{f}) &= \mathbb{E}[(\hat{f}(X) - Y)^2] \\ &= \underbrace{\mathbb{E}[(\hat{f}(X) - \mathbb{E}[\hat{f}(X)])^2]}_{\text{Var}(\hat{f}(X))} + \underbrace{\mathbb{E}[(\mathbb{E}[\hat{f}(X)] - f(X))^2]}_{\text{Bias}(\hat{f}(X))^2} + \sigma^2 \end{aligned}$$

The "excess" risk is then $R(\hat{f}) - R(f) = \text{Bias}(\hat{f}(X))^2 + \text{Var}(\hat{f}(X))$, which is the prediction error not due to irreducibile randomness in Y. Bias is the root mean squared (RMS) error between the true function mapping X to Y and the estimated mapping. Note that this is *not* the difference between $\hat{f}(X)$ and Y, which also contains irreducible error of Y, but instead the RMS error of $\hat{f}(X)$ and $f(X)$. $\text{Var}(\hat{f})$ measures the variability in $\hat{f}$ that comes from random variation in Y that differs across sets of training data (i.e., data that could have been obtained but was not).

Minimizing $R(\hat{f})$ involves minimizing bias and variance, which, as previously mentioned, involves a tradeoff: decreasing bias by increasing a model's flexibility also increases the model's variance. This trade-off is not usually 1:1 however, so sometimes it makes sense to increase one to lower the other. Finding the optimal trade-off requires minimizing excess risk (generalization error minus the irreducible error in Y), $R(\hat{f}) - R(f)$. This is the same as minimizing generalization error since the irreducible randomness in Y is assumed to have expectation 0. Figure 1 shows this trade-off graphically with a simulated example (further details are shown in 1).

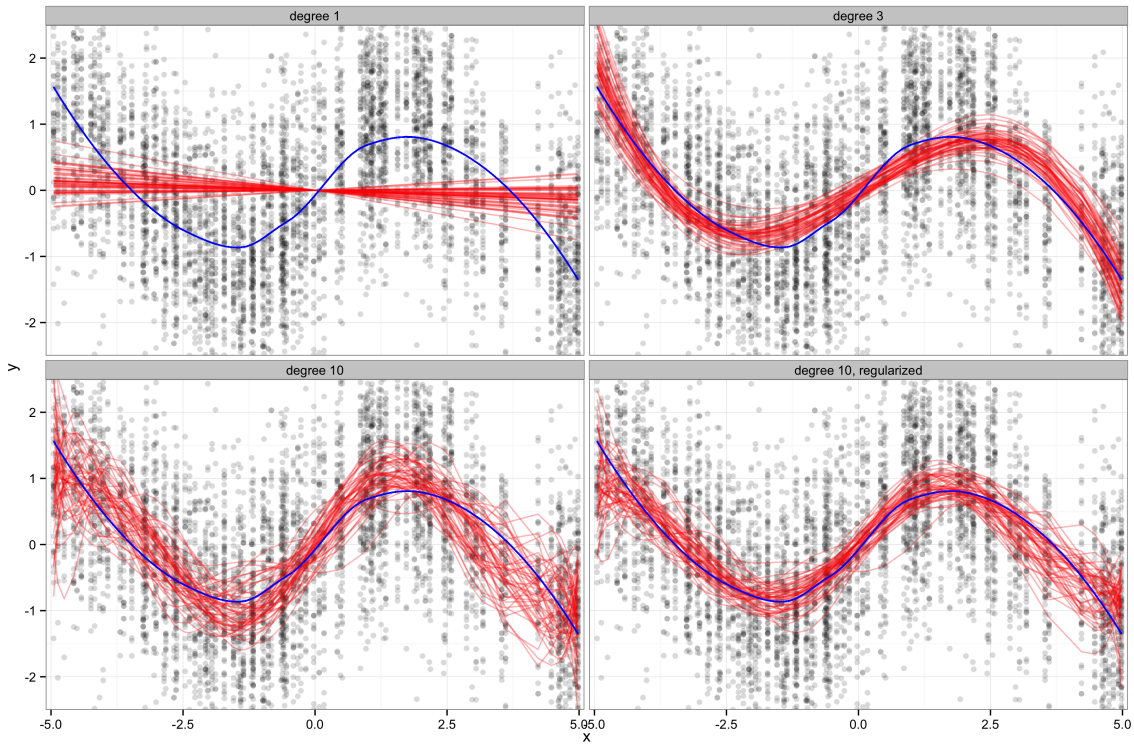

Figure 1: Here $Y = \sin(X) + \epsilon$, where $X \sim U(-5, 5)$ (fixed in repeated sampling) and $\epsilon \sim \mathcal{N}(0, 1)$. The blue line shows the Monte Carlo estimate of $\mathbb{E}[Y|X]$ across $(\mathbf{x}, \mathbf{y})$ drawn from the data generating process. The red lines in each panel indicate the fit of the model to a particular sample. Each sample has 100 observations and the process is repeated 1,000 times (50 randomly drawn examples shown in the figure). The linear case (fit by ordinary least squares) on the top left panel clearly underfits, though this estimator for $\mathbb{E}[Y|X]$ has the lowest variance. The top right panel shows a linear model with a degree 3 orthogonal polynomial expansion of $\mathbf{x}$, which has much lower bias but a higher variance. The bottom left shows a linear model with a degree 10 orthogonal polynomial. Again, the bias is low but the variance has increased relative to the top two panels due to overfitting. The model shown in the bottom right introduces a penalty term (a scalar λ) multiplied by the sum of the absolute values of the coefficients (the L_1 norm of the coefficient vector) where λ is estimated by finding the value which minimizes an estimate of the generalization error using 5-fold cross validation (Efron et al. 2004) (See Kenkel and Signorino (2013) for a similar approach). This substantially reduces variance at the cost of a relatively small amount bias, producing a fit similar to that in the upper right. Table 1 gives further details.

	bias	variance	excess risk	training risk
degree 1	0.75	0.01	0.57	1.55
degree 3	0.35	0.03	0.15	1.09
degree 10	0.02	0.10	0.10	0.90
degree 10, regularized	0.15	0.05	0.07	0.96

Table 1: Monte Carlo (1,000 samples) estimates of the bias, variance, excess risk, and training risk (MS error) of linear models with orthogonal polynomials of degree (1, 3, or 10), and a L_1 regularized linear model fit to training samples of length $n = 100$ drawn from $Y = \sin(X) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 1)$, and $X \sim U(-5, 5)$ and is fixed in repeated sampling. Despite the increase in bias, the regularized degree 10 model has lower excess risk than any other model. Additionally it is clear that the training risk is not an accurate measure of generalization error. The regularized model is the most externally valid.

Regularization (such as the shown in the lower-right panel of 1) methods penalize the complexity of a model in a manner that aims to minimize generalization error. In the case of linear models, two popular forms of regularization are ridge regression and the least absolute shrinkage and selection operator (Lasso), both of which penalize regression coefficients using the size (norm) of the coefficient vector, the sum of the absolute values of the coefficients (the L^1 norm), or the sum of the squares of the coefficients (the L^2 norm) (pp. 61-73, Hastie et al. 2009; Tibshirani 1996).[6] The risk function minimized when using ridge regression on a continuous outcome is shown below.[7]

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Here $\mathbf{y}$ denotes a continuous real-valued outcome, p the number of predictors $\boldsymbol{\beta}$ the regression coefficients, and n the number of observations which are assumed conditionally independent and centered by mean deviation. The only addition to the common least-squares risk function minimized by OLS is the last term, where λ is a penalty parameter which is multiplied by the sum of the square of each β_j. When this function is minimized at a given value of λ coefficients which are less useful in predicting $\mathbf{y}$ are shrunk towards 0.

The risk function minimized when using the Lasso is similar and is shown below.[8] Note that the Lasso penalty may result in some elements of $\boldsymbol{\beta}$ being set of to (exactly) zero, unlike the ridge penalty.

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Selecting how much to penalize the complexity of a model is an application specific problem which is usually solved by estimating generalization error at many values of the penalty parameter(s), a process which is often referred to as tuning, or hyperparameter optimization. In the above cases λ is usually selected using resampling methods such as cross-validation and the bootstrap: the value of λ that minimizes (or nearly minimizes) the resampled estimate of the generalization error is used. When data are independent and identically distributed we can use simple nonparametric methods

[6] These penalties can be combined to give the Elastic Net (Zou and Hastie 2005).

[7] Both can be thought of as Bayesian or frequentist procedures (Tibshirani 1996). The Bayesian equivalent is typically specified with independent Normal priors on the regression coefficients.

[8] Bayesian Lasso estimates are equivalent to the frequentist analogue under independent double-exponential priors on the regression coefficients (Tibshirani 1996; Park and Casella 2008).

such as k-fold cross-validation or the bootstrap to estimate the generalization error of a model.[9] When data are dependent or not assumed to have been drawn from the same distribution these methods are not generally applicable. Nonparametric resampling methods for dependent data are available but require assumptions about the dependence structure and can be considerably more difficult to develop and use (Lahiri 2003; Givens and Hoeting 2012).[10] Development in this area for problems specific to political science may be a fruitful area of methodological research. For some sorts of dependent data, such as time-series data, relatively simple nonparametric resampling methods are available (e.g. the moving block bootstrap). However, any estimate or monotone function of generalization error could be used; for example adjusted R^2, AIC, BIC, or DIC.

Better estimates of generalization error will naturally result in a more optimally tuned model which will both predict more accurately and be less likely to produce model fits that are misleading. These methods offer valuable information about the external validity of a model for data gathered from observational or quasi-experimental designs when true replication is not possible. They therefore are capable of enhancing the generalizability and conclusion validity of the reported empirical relationships.

Recommendations and Future Directions

Importantly, the use of predictive validity as a criterion for inference, another way of saying that there should be a focus on minimizing generalization error, provides a principled way to assess more flexible semiparametric and nonparametric models in observational and quasi-experimental settings.[11] It is often the case that the deductively valid theories used to specify models of empirical relationships in data are underdetermined. What we mean by this is that the functional form that links dependent and independent variables in an empirical model is usually not fully specified by the theory. We suggest that such models often do worse at predicting than they could — in terms of predictive validity — had a more flexible functional form been selected. That is, it is possible to increase the predictive power of a model by only constraining the empirical model in ways specified by the theory, and adopting a more flexible approach for other parts of the model left unspecified by the theory. The use of regularized nonparametric or semiparametric methods is often a much better option than an inflexible parametric model that is not fully implied by the theory (e.g., boosting, generalized additive models, feedforward neural networks, kernal methods, or random forests, among others). It is also possible to use an iterative process for fitting and predictive checking of a parametric model to attain a similar level of flexibility (e.g., Gelman 2004; Gelman 2003).

We do not suggest that researchers adopt a single method, or class of methods in this paper; we simply wish to note that researchers are likely selling their theories short in terms of predictive power by using overly restrictive models that are underdetermined by theory. We note that the call for more focus on predictive checking is not new to political science (e.g. Beck, King, and Zeng 2000; Hill Jr. and Jones 2014; Schnakenberg and Fariss 2014; and Ward, Greenhill, and Bakke 2010 in political science) but given its importance, and the recent focus on replication, which we view as a form of model validation, it is an important point to re-emphasize.[12]

[9] See Arlot, Celisse, and others (2010) for more on the statistical properties of cross-validation estimators of generalization error.

[10] Another approach to regularization relies on finding a theoretical upper bound on generalization error (see e.g. McDonald, Shalizi, and Schervish (2012) for recent work with dependent data).

[11] When generalization error is not estimated, it is unclear how much predictive validity (Adcock and Collier (2001) label predictive validity as nomological validity) the data at hand give to the empirical model as well as the validity of the conclusions based on the empirical model. This is so even in when the empirical model corresponds to a deductively valid theoretical model.

[12] One possible response to this is that some social phenomena may be inherently unpredictable, however, since political scientists have spent relatively little time trying to predict (compared to inference about parameters), we consider it premature to argue that any particular phenomena is inherently unpredictable, despite there being some

Conclusion

In areas of political science where replication is not possible because the data generating process is not under researcher control (i.e,. observational or quasi-experimental data), regularization can be used to decrease threats to external validity from overfitting, and a focus on the minimization of generalization error can increase the external validity of models as well. This serves a similar function to replication in settings where the data generating process is under researcher control (i.e., an experimental or survey design).

To close, we wish to emphasize that scholars of international relations and comparative politics and any other scholars using historical data should endeavor to use methods that minimize generalization error, which enhances the external validity of their models. We have offered a few suggestions towards these goals but there are likely many other methods that applied scholars might consider to strengthen the validity of their conclusions.

R Code Appendix

The R code below replicates the Monte Carlo simulation described in this paper.

```
set.seed(1987)

library(ggplot2)
library(dplyr)
library(tidyr)
library(foreach)
library(doParallel)
library(lars)
library(xtable)

cl <- makeCluster(8)
registerDoParallel(cl)

predict_lars <- function(X, y) {
    cv_res <- cv.lars(X, y, 10, type = "lar", mode = "fraction", plot.it = FALSE)
    opt_frac <- min(cv_res$cv)
    opt_frac <- cv_res$index[which(cv_res$cv == opt_frac)][[1]]
    lasso_path <- lars(X, y, type = "lar")
    lasso_fit <- predict.lars(lasso_path, type = "coefficients",
                              mode = "fraction", s = opt_frac)
    X %*% coef(lasso_fit)
}

mse <- function(x, y) mean((x - y)^2)

n <- 100
m <- 1000
sigma <- 1
x <- runif(n, -5, 5)
out <- foreach(i = 1:m, .combine = "rbind",
```

compelling reasons to think that this may be the case in some situations (c.f. Gartzke 1999).

```
                .packages = c("lars", "dplyr")) %dopar% {
    df <- data_frame("x" = x,
                     "fx" = sin(x),
                     "y" = fx + rnorm(n, 0, sigma),
                     "iteration" = rep(i, n),
                     "idx" = 1:n)
    df$"degree 1" <- lm(y ~ -1 + x, df)$fitted
    df$"degree 3" <- lm(y ~ -1 + poly(x, 3), df)$fitted
    df$"degree 10" <- lm(y ~ -1 + poly(x, 10), df)$fitted
    df$"degree 10, regularized" <- predict_lars(poly(df$x, 10), df$y)
    df
}

out <- out %>% gather(model, estimate, -one_of("iteration", "idx", "y", "x", "fx"))
out <- out %>% group_by(idx) %>% mutate(ty = mean(y), tx = mean(x))

tab <- out %>% group_by(model, idx) %>%
    summarise(bias_sq = mse(mean(estimate), fx),
              variance = mse(estimate, mean(estimate)),
              train_error = mse(estimate, y)) %>%
                  group_by(model) %>%
                      summarise("bias" = sqrt(mean(bias_sq)),
                                "variance" = mean(variance),
                                "excess risk" = bias^2 + variance,
                                "training risk" = mean(train_error))
row.names(tab) <- tab$model
tab$model <- NULL
xtable(tab, digits = 2)

pout <- out %>% filter(iteration %in% 1:50)
p <- ggplot(pout, aes(x, y, group = iteration)) +
    geom_point(alpha = .15) + facet_wrap(~ model) +
        geom_line(aes(x, estimate), alpha = .3, colour = "red") +
            stat_smooth(aes(tx, ty), colour = "blue", method = "loess", se = FALSE) +
                theme_bw() + labs(x = expression(x), y = expression(y)) +
                    coord_cartesian(c(-5.1, 5.1), c(-2.5, 2.5))
ggsave("bias_variance.png", p, width = 12, height = 8)
```

References

Adcock, Robert, and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95 (3): 529–46.

Arlot, Sylvain, Alain Celisse, and others. 2010. "A Survey of Cross-Validation Procedures for Model Selection." *Statistics Surveys* 4. The author, under a Creative Commons Attribution License: 40–79.

Beck, Nathaniel, Gary King, and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *American Political Science Review*. JSTOR, 21–35.

Berk, Richard A. 2004. *Regression Analysis: A Constructive Critique*. Vol. 11. Sage.

Dafoe, Allan. 2014. "Science Deserves Better: The Imperative to Share Complete Replication Files." *PS: Political Science & Politics* 47 (01). Cambridge Univ Press: 60–66.

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge University Press.

Efron, Bradley, Trevor Hastie, Iain Johnstone, Robert Tibshirani, and others. 2004. "Least Angle Regression." *The Annals of Statistics* 32 (2). Institute of Mathematical Statistics: 407–99.

Gartzke, Erik. 1999. "War Is in the Error Term." *International Organization* 53 (03). Cambridge Univ Press: 567–87.

Gelman, Andrew. 2003. "A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-Fit Testing." *International Statistical Review* 71 (2). Wiley Online Library: 369–82.

———. 2004. "Exploratory Data Analysis for Complex Models." *Journal of Computational and Graphical Statistics* 13 (4).

Givens, Geof H, and Jennifer A Hoeting. 2012. *Computational Statistics*. Vol. 708. John Wiley & Sons.

Hainmueller, Jens, and Chad Hazlett. 2013. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach." *Political Analysis*. SPM-PMSAPSA, mpt019.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. 2009. *The Elements of Statistical Learning*. Vol. 2. 1. Springer.

Herrnson, Paul S. 1995. "Replication, Verification, Secondary Analysis, and Data Collection in Political Science." *PS: Political Science & Politics* 28 (03). Cambridge Univ Press: 452–55.

Hill Jr., Daniel W., and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Reivew* 108 (3): 661–87.

Jones, Zachary M. 2013. "Git/GitHub, Transparency, and Legitimacy in Quantitative Research." *The Political Methodologist* 21 (1): 6–7.

Kass, Robert E. 2011. "Statistical Inference: The Big Picture." *Statistical Science* 26 (1): 1–9.

Kenkel, Brenton, and Curtis S Signorino. 2013. "Bootstrapped Basis Regression with Variable Selection: A New Method for Flexible Functional Form Estimation." *Manuscript, University of Rochester*.

King, Gary. 1995. "Replication, Replication." *PS: Political Science and Politics* XXVIII (494-499).

———. 2006. "Publication, Publication." *PS: Political Science and Politics* XXXIX (1): 119–25.

Lahiri, Soumendra Nath. 2003. *Resampling Methods for Dependent Data*. Springer.

McDonald, Daniel J, Cosma Rohilla Shalizi, and Mark Schervish. 2012. "Time Series Forecasting: Model Evaluation and Selection Using Nonparametric Risk Bounds." *ArXiv Preprint ArXiv:1212.0463*.

Park, Trevor, and George Casella. 2008. "The Bayesian Lasso." *Journal of the American Statistical Association* 103 (482). Taylor & Francis: 681–86.

Schnakenberg, Keith E., and Christopher J. Fariss. 2014. "Dynamic Patterns of Human Rights Practices." *Political Science Research and Methods* 2 (1): 1–31.

Shadish, William R. 2010. "Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings." *Psychological Methods* 12 (1): 3–17.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth Publishing.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 267–88.

Vapnik, Vladimir Naumovich, and Vlamimir Vapnik. 1998. *Statistical Learning Theory*. Vol. 2. Wiley New York.

Ward, Michael D, Brian D Greenhill, and Kristin M Bakke. 2010. "The Perils of Policy by P-Value: Predicting Civil Conflicts." *Journal of Peace Research* 47 (4). Sage Publications: 363–75.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2). Wiley Online Library: 301–20.