Political Science Research and Methods Vol 6, No. 2, 365–380 April 2018
© The European Political Science Association, 2017
doi:10.1017/psrm.2017.5

Enhancing Validity in Observational Settings When Replication is Not Possible*

CHRISTOPHER J. FARISS AND ZACHARY M. JONES

We argue that political sciexntists can provide additional evidence for the predictive validity of observational and quasi-experimental research designs by minimizing the expected prediction error or generalization error of their empirical models. For observational and quasi-experimental data not generated by a stochastic mechanism under the researcher's control, the reproduction of statistical analyses is possible but replication of the data-generating procedures is not. Estimating the generalization error of a model for this type of data and then adjusting the model to minimize this estimate—regularization—provides evidence for the predictive validity of the study by decreasing the risk of overfitting. Estimating generalization error also allows for model comparisons that highlight underfitting: when a model generalizes poorly due to missing systematic features of the data-generating process. Thus, minimizing generalization error provides a principled method for modeling relationships between variables that are measured but whose relationships with the outcome(s) are left unspecified by a deductively valid theory. Overall, the minimization of generalization error is important because it quantifies the expected reliability of predictions in a way that is similar to external validity, consequently increasing the validity of the study's conclusions.

Replication is explicitly focused on generating evidence in support of the external validity of an inference and involves taking a new draw from the same data-generating process used to generate the original data set by repeating the procedures specified by the research design. Unfortunately, this type of exact replication[1] is not possible in observational and quasi-experimental settings when the data-generating process is not controlled by the researcher. As we argue in this article, however, evidence regarding the reliability of predictions, generalization error, is similar to external validity, and is thus important for conclusion validity.[2] Within the framework developed by Shmueli (2010), we suggest that predictive validity should be especially important in exploratory and predictive data analyses wherein the theoretical relationships of interest are not causally identified. We additionally suggest that in explanatory analyses where a relationship is causally identified, analysis of predictive validity can contextualize effect size(s)

* Christopher J. Fariss, Assistant Professor, Department of Political Science and Faculty Associate, Center for Political Studies, Institute for Social Research, University of Michigan, Center for Political Studies (CPS) Institute for Social Research, 4200 Bay, University of Michigan, Ann Arbor, Michigan 48106-1248 USA (cjf0006@gmail.com). Zachary M. Jones, Ph.D. Candidate, Pennsylvania State University; Pond Laboratory, Pennsylvania State University, State College, PA 16801 (zmj@zmjones.com). The authors would like to thank Michael Alvarez, Neil Beck, Bernd Bischl, Charles Crabtree, Allan Dafoe, Cassy Dorff, Dan Enemark, Matt Golder, Sophia Hatz, Danny Hill, Luke Keele, Lars Kotthoff, Fridolin Linder, Mark Major, Michael Nelson, Keith Schnakenberg, and Tara Slough for many helpful comments and suggestions. This research was supported in part by The McCourtney Institute for Democracy Innovation Grant, and the College of Liberal Arts, both at Pennsylvania State University.

[1] 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.

[2] See Vapnik (1998) for a discussion of statistical learning theory, which is the theoretical investigation of the ability of algorithms which build models from data to accurately generalize to unseen data.

https://doi.org/10.1017/psrm.2017.5 Published online by Cambridge University Press

(i.e., by estimating predictive importance) and provide information about how the effect varies (Jones and Linder 2016; Athey and Imbens 2015; Wager and Athey 2015).[3]

In brief, generalization error is an unobserved measure of the accuracy of predictions from a model. Minimizing generalization error requires the estimation of this unknown quantity and adjustment of the model to minimize it. Generalization error provides information about the validity of the study in a manner similar to exact replication of a data-generating process, which provides direct evidence about the reliability of an estimate. Normal practice is to minimize prediction error on the data at hand. We instead advocate the minimization of *expected* prediction error: generalization error.[4]

Minimizing generalization error also provides a principled method for modeling complex empirical relationships because the functional form that links outcomes and explanatory variables in an empirical model is often, perhaps usually, not fully specified by the theory. That is, it is possible to increase the predictive validity (decrease the generalization error) 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. Relatedly, generalization error also provides a method of model selection. Although it will not always be the case that the relevant summary of the model is generalization error, it provides a default which at least maximizes a notion of predictive validity, which, absent a basis on which to make causal claims, may be desirable.

In the remainder of this paper, we first consider and define generalization error, external validity, replication, reproduction, and the relationship between these concepts. We then discuss generalization error, its estimation, and techniques for adjusting models to minimize it. We close with a discussion of future directions for research.

GENERALIZATION, EXTERNAL VALIDITY, REPLICATION, AND REPRODUCTION

In this article, we focus on techniques for generating evidence in support of the generalizability of an empirical model of a data-generating process, which is a function mapping explanatory variables to outcomes estimated from data which describes how the data could have arisen. We argue that generalization is similar to external validity, which can be supported by replication, but which is not possible in cases where the process generating the data is not under researcher control.

Though the terms generalizability and external validity are often used synonymously with one another, the statistical learning community defines generalizability differently than how Shadish (2010) defines external validity.[5] Generalization in the statistical learning sense refers to the transportability of a learned function (i.e., one estimated from data) to other draws from *the same* data-generating process, and is focused explicitly on prediction. A learned function that generalizes well has low prediction error on new data from the same generating process. Generalization error is *not* an estimate of the validity of cause–effect relationships (internal validity), which may or may not be plausibly causally identified in a model fit to observational or quasi-experimental data (see e.g., Dunning 2012; Keele 2015; Keele and Titiunik 2015 for more general discussions of causal identification in the social sciences).[6] Neither is

[3] In Shmueli's (2010) framework, exploratory analyses are those in which causal relationships are not identified but the relationships between the explanatory variables and the outcomes are investigated. Predictive analyses, naturally, are those in which predictions are of primary interest: for example, forecasting. Explanatory analyses are those in which causal relationships are arguably identified, and accurate estimation of these relationships is of primary interest.

[4] Although the expected loss is often used, it need not be the only choice.

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

[6] Providing evidence for the internal validity of research design is a topic that we do not consider further in this paper but again see (Dunning 2012; Keele 2015; Keele and Titiunik 2015).

https://doi.org/10.1017/psrm.2017.5 Published online by Cambridge University Press

generalizability in this sense external validity, since the latter pertains explicitly to the generalizability of cause–effect relationships learned from data.

Shadish 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" or outcome variables (2010, 4). These components make up a theoretically specified data-generating process and are linked together by a set of assumptions (premises or postulates) and a set of propositions. These assumptions and propositions link the treatment or causal variable (X) to a measurement or outcome variable (Y) within a specified empirical domain. The domain within which the theory explains the relationship between the treatment and outcome variables is bounded by scope conditions. The scope conditions of a theory are auxiliary assumptions, specifically regarding the attributes of the persons (i.e., the units) and the settings within which the persons reside (i.e., the spatial and temporal information). This auxiliary information is important because a data-generating process might change systematically for different types of persons or units (e.g., Brady 1986; Wilcox, Sigleman and Cook 1989; King et al. 2004), or across time or between places (e.g., Western 1998; Bailey 2007; Fariss 2014).

If theoretical differences between data-generating processes are not recognized and specified in the empirical model (e.g., by including important covariates capturing structural change) of said data-generating process then any predictions from the model will be biased (see e.g., Fariss 2014; Fariss Forthcoming, for a discussion of this issue as it relates to the study of human rights).[7] Thus, the scope conditions of the theory provide important information about the conditions that must be met in order for the model of the data-generating process to be a valid representation of the theory (e.g., Adcock and Collier 2001; Lake 2013; Elkins and Sides 2014; Fariss 2014). To reiterate, generalizability or generalization error is an estimate of the ability of a model to generate accurate predictions on new data from a data-generating process. In practice, what distinguishes one data-generating process from another is the scope or the domain of the theory (e.g., Lake 2013).[8] The importance of specifying the scope conditions of a theory is essential because multiple and related data-generating processes may be operating on different units or across different spatial or temporal settings. The generalization error of a model provides important information about how well the learned function generalizes to the data-generating process under study but not necessarily any other data-generating processes. Stated differently, generalization error does not necessarily provide information about the performance of a model on a sample drawn from a different population or using different treatment or outcome variables, except insofar as a different populations or measurements are similar to the process that generated the data used to fit the model (Bareinboim and Pearl 2012). In this way the definition of generalization error is analogous to exact replication because neither generalization error nor exact replication pertain to the transportability of a model of one data-generating process to another, but both pertain to the reliability of estimates: predictions and treatment effects, respectively.

Shadish (2010) makes a similar point about the external validity of causal inferences: exact replication provides evidence for the external validity of an inference only when the sample (i.e., the persons or settings) and the explanatory variables (i.e., the component parts of a

[7] Fariss (2014) demonstrates that human rights respect has actually been improving over the last 35 years. However, the standards used by monitoring groups to assess human rights practices have also become more strict over time. This contemporaneous change to the data-generating process has masked the positive trend in human rights respect.

[8] Lake (2013) emphasizes the importance of midlevel theorizing in the study world politics. In other words, Lake (2013) suggests that researchers pay close attention to the scope conditions for the theories or data-generating processes of specific research topics.

https://doi.org/10.1017/psrm.2017.5 Published online by Cambridge University Press

theoretically specified data-generating process) are fixed or at least probabilistically equivalent once accounting for measurement error. However, the external validity of a causal inference can also be enhanced by varying one or more of these components of the research design (i.e., a conceptual replication).[9] Thus, the definition of generalization error is *not* analogous to conceptual replication. With these important distinctions in mind, we now turn to a discussion of the distinctions between reproduction and the different types of replication and how these concepts are related to external validity and generalization.

As previously noted, we define replication as taking a new draw from the same data-generating process used to generate the original data. This is distinct from reproduction, which entails reproducing the same findings given the same data and statistical analysis procedure. Reproduction represents a minimal standard of transparency for scientific research and is the concept commonly referred to as the broader "replication standard" in political science (Herrnson 1995; King 1995; King 2006; Dafoe 2014).[10] 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 (i.e., an exact replication). A study based on survey data could be replicated by surveying a new set of individuals from the same target population and then conducting the same statistical analysis on the new sample (i.e., an exact replication). If the empirical relationships in these examples are generalizable to the population from which the new samples are drawn, then similar findings will be obtained, subject to the uncertainty due to sampling. Thus, exact replication, unlike reproduction, provides evidence of the external validity of a specific empirical relationship. Reproduction, though essential for the transparency of scientific research, does not provide evidence about the external validity of an empirical study.

When data are not generated by a process controlled by the researcher—replication in the sense described above is not possible (Berk 2004). If researchers view a data set as the result of a stochastic process,[11] however, it is still possible to estimate how generalizable the model's predictions are to new observations from the same data-generating process. Even quasi-experimental designs with strong evidence of internal validity are not replicable based on the definitions used above, as they take advantage of a unique exogenous shock to the social or political systems of interest (see Shadish, Cook and Campbell 2001; Dunning 2012; Keele and Titiunik 2015). Thus, the techniques we discuss next can be used to provide evidence for the generalizability of predictions drawn from both observational and quasi-experimental designs in a way that is distinct from but related to the goal of exact replication.

OVERFITTING AND UNDERFITTING

If the generalization error of a model is high, the predictions and substantive interpretation of the model are potentially unreliable. Whether or not the generalization error of a model is high,

[9] Bareinboim and Pearl (2012) develop a theory and algorithm of transportability. The algorithm is designed to identify conditions under which a causal relationship learned from an experiment can be reused in a different observational setting. This algorithmic approach is consistent with the goal of a conceptual replication in which one or more of the components of the research design is altered.

[10] 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.

[11] In the examples that follow, the processes are all strictly stationary. However, the assumptions made about the data-generating process need not necessarily be so strong. In many cases weaker assumptions are enough to obtain theoretical results (see e.g., McDonald, Shalizi and Schervish 2012).

https://doi.org/10.1017/psrm.2017.5 Published online by Cambridge University Press

however, is not made apparent by looking at the prediction error on the data used for fitting the model because of the possibility of underfitting and overfitting.

Overfitting occurs when non-systematic variation—noise—is described by an empirical model, instead of systematic variation—signal. An overfit model, by definition, has high generalization error. As is commonly recognized, it is generally the case that a model will fit the data used for estimation much more closely than data not used for estimation. A variety of procedures have been developed to prevent such overfitting, but the use of these tools is not yet common practice in political science (though see e.g., Beck, King and Zeng 2000; Ward, Greenhill and Bakke 2010; Hainmueller and Hazlett 2014; Kenkel and Signorino 2013; Hill and Jones 2014, e.g., of articles which (at least implicitly) discuss this problem).

Underfitting, a related problem, occurs when a model does not detect systematic patterns in the data, which results in higher generalization error than could have otherwise been obtained. Flexible methods—particularly non-parametric and semi-parametric methods—are attractive alternatives to restrictive parametric models because of their ability to find systematic patterns in the data that were not expected by theory (i.e., features of the data not directly encoded in the model) but which are generalizable (i.e., are systematic features of the data-generating process). The increased flexibility of such models decreases the error on the data to which the model was fit (the error on the training data), and, perhaps, across all possible data sets that could have been obtained from a specific data-generating process. In general, however, the capacity of a method to overfit the data used for estimation increases with its flexibility. What this means in practice is that finding the best method in terms of generalization error involves balancing the tradeoff between flexibility and the risk of overfitting, which, to foreshadow the next section, is a tradeoff between bias and variance. Repeated estimation of generalization error is used to make this tradeoff.

Though any estimators of generalization error could be used (e.g., adjusted R^2), off-the-shelf estimators are not necessarily the best choice. That is, reliance on one of these estimators may result in unnecessarily high generalization error for a learned function, and thus invalid model selection or misleading substantive interpretation. We suggest the use of resampling estimators, which are often better suited to this task since they rely on weaker assumptions. However, devising a reasonable estimator with complex or dependent data-generating processes may sometimes be difficult, which we believe is a important and promising area of research and an issue we discuss briefly at the close of the next section.

FINDING BALANCE BETWEEN OVERFITTING AND UNDERFITTING

In order to further elucidate the tradeoff between bias and variance—finding the right balance between the risk of overfitting and underfitting—we provide a formal exposition, which motivates a pair of Monte Carlo examples.

Prediction error is measured by a loss function l. A loss function measures the discrepancy or contrast between the observed and predicted outcomes and is a non-negative real-valued function (i.e., a function that takes as input pairs of numbers: a prediction and an observation, and returns one number that is greater than or equal to 0). For this example, our loss function is the familiar squared error loss function minimized by ordinary least squares regression. We decompose the expectation of this particular loss function, the risk, to highlight the bias–variance tradeoff, which allows us to find the model with the lowest generalization error among the class of models considered.

Here, we consider a random variables $(X, Y) \sim \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ distributed according to a joint distribution $\mathcal{P}$. $\mathcal{X}$ and $\mathcal{Y}$ represent the input spaces of the random variables X and Y,

https://doi.org/10.1017/psrm.2017.5 Published online by Cambridge University Press

respectively, with $\mathcal{P}$ a probability distribution over ordered pairs drawn from the set of possible combinations of draws from these spaces: $\mathcal{X} \times \mathcal{Y}$. A finite sample of data from $\mathcal{P}$ of length n is denoted D_n and is composed of ordered pairs (x_i, y_i), $\forall\, i = (1, \ldots, n)$ (x_i is often a vector). We denote the expected loss, that is, the average loss over the joint distribution, by R and refer to it as the risk. The empirical loss, that is, the sample average loss, is denoted by $\hat{R}_n$ and is referred to as the empirical risk.

For the aforementioned squared error loss function, the prediction function which has the minimum risk is the conditional expectation: $f^* : x \rightarrow \mathbb{E}_Y(Y \,|\, X = x)$. The risk of this optimal prediction function f^* is the variance of Y at a particular value of X (the subscript x may be dropped if Y is homoscedastic).

$$R(f^*) = \mathbb{E}_Y\left[(Y - f^*(x))^2 \,|\, X = x\right] = \sigma_x^2.$$

This is referred to as the Bayes risk: the function which makes the risk (expected loss) minimal. If this function (f^*) mapping X to Y were known, then the only error that would be made in predictions is due to irreducible variability in Y. Note that the mapping between X and Y is *not* random. When f^* is not known and $\mathcal{D}_n$ is finite (i.e., there is a finite amount of sample data), then this error rate (the Bayes error rate) cannot be achieved.

However, with a sample $\mathcal{D}_n$ drawn from $\mathcal{P}$, $\hat{f}$, an approximation to f^* can be estimated or learned. We can compute the risk of the estimated function $\hat{f}$ as well, which is necessarily larger than the risk of f^* since f^* is not known and because $\mathcal{D}_n$ is finite and thus not perfectly representative of $\mathcal{P}$. If $\mathcal{F}$ is the set of functions that can possibly be learned from $\mathcal{D}_n$ (e.g., a real two-dimensional additive function, i.e., linear regression with two explanatory variables), then the function in this class ($\mathcal{F}$) which minimizes the empirical risk, that is, the sample average loss, is frequently chosen. This does not necessarily minimize the expected loss (the risk), however.

As previously noted $\hat{f}$ is estimated from $\mathcal{D}_n$, the finite set of data used for fitting drawn from $\mathcal{P}$, the data-generating process. In the special case where $\mathcal{Y} = \mathbb{R}$ (i.e., the set of possible values for Y is the real line: regression) and the loss function is the common squared error function, the risk of the estimated function $\hat{f}$ can be written as a sum of irreducible error, the squared bias of $\hat{f}$, and the variance of $\hat{f}$.

$$\begin{aligned} R(\hat{f} \mid X = x) &= \mathbb{E}_Y\left[(\hat{f}(x) - Y)^2\right] \\ &= \underbrace{\mathbb{E}_Y\left[\left(\hat{f}(x) - \mathbb{E}_Y\left[\hat{f}(x)\right]\right)^2\right]}_{\mathrm{Var}(\hat{f}(x))} + \underbrace{\left[\mathbb{E}_Y\left[\hat{f}(x)\right] - f^*(x)\right]^2}_{\mathrm{Bias}(\hat{f}(x))^2} + \underbrace{\sigma_x^2}_{\mathrm{Var}(Y|X=x)}. \end{aligned}$$

The "excess" risk, that is, the error that is not due to irreducible randomness, is $R(\hat{f}) - R(f^*)$, the difference between the risk of the estimated function and the risk of the true function, the Bayes risk, which is the variance of Y conditional on a particular value of $X = x$. The resulting expression for the excess risk is $\mathrm{Bias}(\hat{f})^2 + \mathrm{Var}(\hat{f})$ which, again, is the prediction error *not* due to irreducible randomness in Y. Bias is the difference between the expectation of $\hat{f}$ at $X = x$ and f^* at $X = x$. Note that the bias is *not* the difference between $\hat{f}$ and Y at $X = x$, which also contains irreducible error of Y, but instead the expected difference between $\hat{f}$ and f^*. $\mathrm{Var}(\hat{f})$ measures the variability in $\hat{f}$ that comes from random variation in the training data (i.e., data from $\mathcal{P}_{\mathcal{X} \times \mathcal{Y}}$ that could have been obtained but were not).

Minimizing the risk of $\hat{f}$, $R(\hat{f})$, thus involves minimizing both bias and variance, which, as previously mentioned, involves a tradeoff. Bias can be decreased by allowing the model to more

https://doi.org/10.1017/psrm.2017.5 Published online by Cambridge University Press

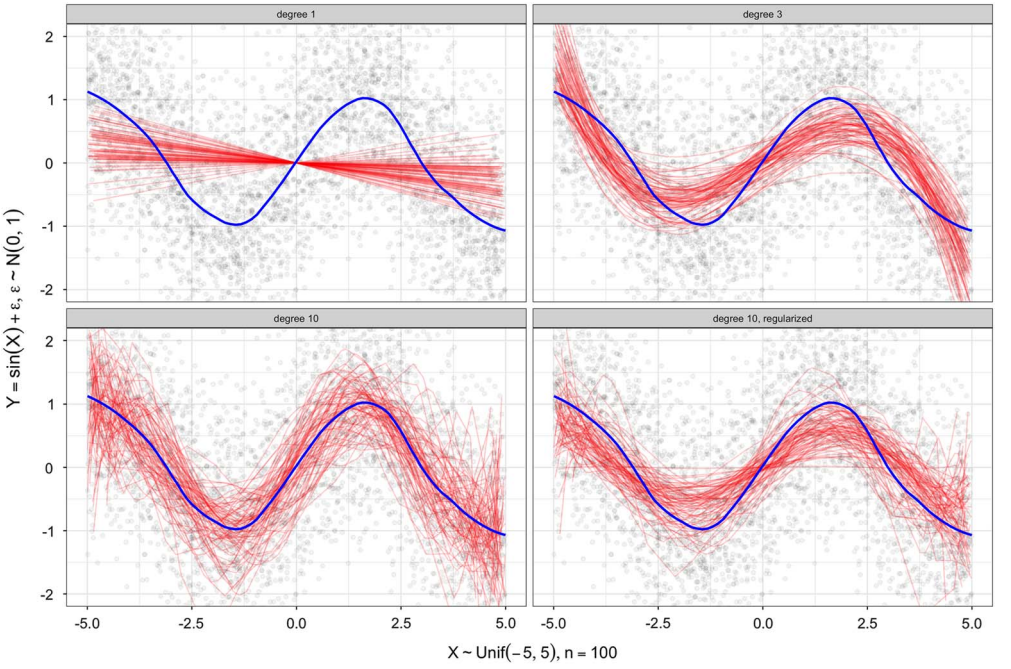


Fig. 1. Here $Y = sin(X) + \epsilon$, where $X \sim U(-5, 5)$ and $\epsilon \sim \mathcal{N}(0, 1)$
Note: The blue line shows the Monte Carlo estimate of $\mathbb{E}[Y \mid X = x]$ across ($\mathbf{x}$, $\mathbf{y}$) drawn from the data-generating process. The red lines in each panel indicate the fit of the model to a particular sample. Each sample has 100 observations and the process is repeated 1000 times (75 randomly drawn examples shown in the figure). The linear case (fit by ordinary least squares) on the top left panel clearly underfits (the bias is high), though this estimator for f^* 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. The bias is smaller 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 10-fold cross-validation (Efron et al. 2004; see also Kenkel and Signorino 2013 for a similar approach). This substantially reduces the variance of the predictions at the cost of a relatively small amount bias, producing a fit similar to that in the upper right. This fit has the smallest risk or generalization error. Table 1 gives further details.

closely fit the data, but decreasing bias by increasing a model's flexibility also increases the model's variance, as the model is more sensitive to random components of the data.[12] The tradeoff between bias and variance is not usually 1:1, however, so it often makes sense to increase one to lower the other. Finding the optimal tradeoff 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 tradeoff graphically with a simulated example (further details are shown in Table 1). Figure 2 gives another example of the bias–variance tradeoff in action with boosted regression.

[12] This assumes a fixed sample size. Having more data for a fixed level of model complexity would decrease variance.

https://doi.org/10.1017/psrm.2017.5 Published online by Cambridge University Press

TABLE 1 *Monte Carlo (1000 Samples) Estimates of the Expected Risk $R(\hat{f})$, Empirical Risk $\hat{R}_n(\hat{f})$, Excess Risk $R(\hat{f}) - R(f^*)$, and the Bayes Risk, $R(f^*)$ 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 $X \sim U(-5, 5)$*

	Expected Risk	Empirical Risk	Excess Risk	Bayes Risk
Degree 1	1.52	1.48	0.52	1.00
Degree 3	1.22	1.07	0.21	1.00
Degree 10	14.02	0.81	6.61	1.00
Degree 10 (regularized)	1.20	0.97	0.18	1.00

Note: The expected risk is minimized by the regularized linear model. Note also the divergence of the empirical risk and the expected risk as the degree of the polynomial increases. The regularized model is the most generalizable in this sense.

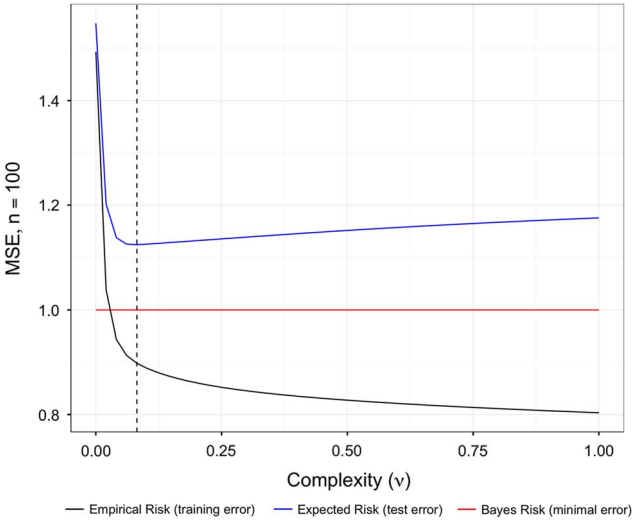


Fig. 2. A learning curve for boosted regression trees (Hothorn et al. 2010; Hothorn et al. 2014)
Note: On the x-axis the complexity parameter ν is shown, increasing from left to right (higher is more complex). ν controls the "learning rate": how quickly the model adapts to the data. On the y-axis the mean squared error of a series of fits to independent and identically distributed training data ($n = 100$). Each fit is used to predict on the training set and the test set and is averaged over 1000 Monte Carlo iterations. At low levels of complexity, variance is low and bias is high: the expected and empirical risk is similar. As the complexity of the model increases, however, the difference between the empirical and expected risk diverges, with the former decreasing below the Bayes error rate (the theoretical minimum expected risk): overfitting the data. Minimizing the expected risk prevents overfitting that occurs when the empirical risk is minimized: the complexity parameter ν which minimizes the expected risk is denoted by the dashed vertical line ($\nu = 0.08$).

Flexible methods are desirable because they minimize bias, and simple methods are desirable because they have lower variance, both of which are components of excess risk. Regularization methods (e.g., the lower-right panel of Figure 1) penalize the complexity of a model in a manner that aims to minimize generalization error by making an optimal tradeoff between bias and variance: allowing a model to adapt to the data but not so much so that it overfits. Many of these regularization methods are heuristic, that is, they are not strictly optimal but they are

https://doi.org/10.1017/psrm.2017.5 Published online by Cambridge University Press

computationally tractable. 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) (61–73; Tibshirani 1996; Hastie, Tibshirani and Friedman 2009).[13] The function minimized when using ridge regression on a continuous outcome is shown below.[14]

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

here $\mathbf{y}$ denotes a continuous real-valued outcome, p the number of predictors, β the regression coefficients, and n the number of observations which are assumed independent and centered by mean deviation (note that we omit an intercept for this reason, the empirical mean of $\mathbf{y}$ is 0). The only addition to the common least squares empirical risk function is the last term, where λ is a penalty parameter which is multiplied by the sum of the squares of each β_j. When this function is minimized at a particular value of λ, coefficients which are less useful in predicting $\mathbf{y}$ are shrunk toward 0. Thus, when this function is minimized, both the norm of the coefficient vector and the empirical risk are jointly minimized. This amounts to an application of Occam's Razor: simpler solutions (i.e., smaller coefficient vector norms) are to be preferred, all else equal. That is, the coefficients are shrunk toward 0 if they do not contribute enough to the minimization of the empirical risk. How quickly shrinkage occurs is determined by the form of the penalty (e.g., the L_1 or L_2 norms of the coefficient vector) as well as the value of λ, which is usually selected to minimize generalization error (how this selection works is discussed further below). Parameter shrinkage makes regularized estimators less sensitive to the data (decreases their variance), which, again, can prevent overfitting.

The empirical risk function minimized when using the Lasso is similar and is shown below.[15] Note that the Lasso penalty may result in some elements of β being set of to (exactly) 0, unlike the ridge penalty.

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

The selection of 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). This process which is often referred to as tuning or hyperparameter optimization. Though extensive discussion of this topic is beyond the scope of this paper, hyperparameter optimization is often much more sophisticated than an exhaustive search over a finite grid of tuning parameter values (grid search) and hence much more computationally efficient (see e.g., Bengio 2000; Bergstra and Bengio 2012). The expected risk of a model with a particular set of hyperparameters (in this case just λ) is often estimated using resampling methods. Then the value of the hyperparameter(s) which minimizes (or nearly minimizes) the resampled estimate of the generalization error is used. When data are independent and

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

[14] Ridge regression can be thought of as Bayesian or frequentist procedure (Tibshirani 1996). The Bayesian equivalent of ridge regression is a model with independent Normal priors on the regression coefficients.

[15] Like ridge regression, Lasso regression can be also thought of as Bayesian or frequentist procedures (Tibshirani 1996). Bayesian Lasso estimates are equivalent to the frequentist analogue under independent double exponential priors on the regression coefficients (Tibshirani 1996; Park and Casella 2008).

https://doi.org/10.1017/psrm.2017.5 Published online by Cambridge University Press

identically distributed we can use simple nonparametric resampling methods such as k-fold cross-validation or the bootstrap to estimate the generalization error of a model. Note also that regularization may be used with far more complex models, we discuss linear regression only because of its familiarity and simplicity.[16]

Resampling methods work by treating the data at hand $\mathcal{D}_n$ as the data-generating process $\mathcal{P}$, sampling from $\mathcal{D}_n$, and finding an estimate of f^*, $\hat{f}$, on each pseudo-sample. The bootstrap, for example, works by sampling n observations uniformly and with replacement from $\mathcal{D}_n$ (Efron 1982): analogous to simple random sampling from $\mathcal{P}$. $\hat{f}$ is estimated from this pseudo-sample and an estimate of the risk is obtained by computing the prediction error on the observations that are *not* in said psuedo-sample. k-Fold cross-validation is another common resampling estimator which divides $\mathcal{D}_n$ into k randomly selected folds (groups of $\frac{n}{k}$ observations sampled without replacement). $\hat{f}$ is learned on $k-1$ of the folds, and the risk is estimated by using $\hat{f}$ to predict on the kth held out fold. The procedure repeats so that each fold is held out from estimation of $\hat{f}$, and the risk estimates from each iteration are averaged. Many variations on these two resampling methods are available (see e.g., Efron and Tibshirani 1994; Arlot et al. 2010). Resampling methods are an active area of research, and these simple, well-known methods may not be the best choice in many situations (see e.g., Bischl et al. 2012 for more recommendations).

Methods for dependent data are available but require additional assumptions about the dependence structure of the data-generating process and can be considerably more difficult to develop and use (Lahiri 2003; Givens and Hoeting 2012). For some sorts of dependent data, such as some types of time-series data, relatively simple nonparametric resampling methods are available (e.g., the moving block bootstrap). In other instances, Bayesian hierarchical models may be most effective (e.g., Tibshirani 1996; Western 1998; Gelman 2003; Gelman 2004; Park and Casella 2008). In other cases, bounds on generalization error can be estimated by using structural risk minimization (Vapnik 1998). Discussion of structural risk minimization is beyond the scope of this paper but this appears to be a promising area of research with applications to social science data (see e.g., McDonald, Shalizi and Schervish 2012 for recent work with macroeconomic data). In general the development of methods specific to political data may be a fruitful area of methodological research.

We emphasize that better estimates of generalization error will naturally result in a more optimally tuned model which will generalize better (in terms of prediction error). Hence, estimating generalization error and then adjusting the model to minimize this quantity is a means to maximize the predictive validity of a model's predictions in situations where exact replication is not possible. We now turn to a discussion of how applied researchers might use our recommendations in future research.

EMPIRICAL VALIDATION OF UNSPECIFIED FUNCTIONAL FORMS AND MODEL SELECTION

It is often the case that the deductively valid theories used to specify models of empirical relationships in data are underdetermined.[17] What we mean by this is that the functional form

[16] Decision trees (and ensembles thereof, such as forests and boosting) can be regularized by pruning nodes, by penalizing the risk function being minimized, and by adjusting numerous other hyperparameters (see e.g., Mingers 1989; Hothorn, Hornik and Zeileis 2006). Other models such as splines can be regularized using assumptions about smoothness or roughness (see e.g., Beck and Jackman 1998; Keele 2008).

[17] In this sense many empirical models are not capable of learning (generating valid inferences) about underlying structure of the data if the theory is not specified to sufficiently constrain the parameter space. Hence,

https://doi.org/10.1017/psrm.2017.5 Published online by Cambridge University Press

that links outcomes and explanatory variables in an empirical model is usually not fully specified by the theory. We suggest that such models often do worse than they might have otherwise—in terms of predictive validity—had a more flexible functional form been selected.[18] 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 (in the sense that it increases predictive validity) way to use more flexible semiparametric and nonparametric models in observational and quasi-experimental research design settings. That is, it is possible to increase the predictive validity 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. The use of regularized nonparametric or semiparametric methods (e.g., using methods such as boosting, generalized additive models, feedforward neural networks, kernel methods, or random forests, among many others) is often a much better option than an inflexible parametric model that is not fully implied by the theory.[19] Adopting a restrictive functional form where one is not directly implied is an arbitrary data analytic choice which impedes scientific progress by obscuring unexpected features of the data, which results in lower predictive validity.

Combining both a strict functional form deduced or encoded from a theory and a more flexible functional form to capture structure in the data where either the functional form is unclear or relevant measurements have not been obtained is an active area of research. Stagewise methods such as boosting (e.g., "model-based boosting" and generalized additive models) offer well-developed implementations using well-studied statistical frameworks (Hastie and Tibshirani 1990; Friedman 2001; Hothorn et al. 2010; Schapire and Freund 2012; Wood and Wood 2015). It is also possible to combine more restrictive and more flexible functional forms by specifying latent variable models such as the latent space/factor class of models for networks (Hoff, Raftery and Handcock 2002; Hoff 2005; Handcock, Raftery and Tantrum 2007; Hoff 2009). Ensembles of models estimated using different explanatory variables and combined by using a meta/super learner is also an attractive approach commonly referred to as stacking (Breiman 1996; LeBlanc and Tibshirani 1996). Under certain conditions this approach also allows the estimation of sampling uncertainty (Sexton and Laake 2009; Mentch and Hooker 2014; Wager, Hastie and Efron 2014). Another alternative approach is to not strictly specify any parts of the empirical model (i.e., only things like continuity or the maximal depth of interaction). Under this most flexible model, the hypothesized relationships can be compared with what the model learned from the data.

As we have argued above, generalization error can be used to make the optimal tradeoff between bias and variance. In the above discussion this has been "internal" in the sense that parameters of a model are found by the iterative minimization of estimates of generalization error. Generalization error can also be used for "external" model selection (Hastie, Tibshirani and Friedman 2009; Arlot et al. 2010). This would entail, for example, the comparison of models developed by different groups of researchers or that embodied different explanations for the process that generates outcomes. Absent a compelling alternative (such as a statistic which captures a particular type of structure in the data of theoretical importance), particularly in the

(F'note continued)
regularization can be used to accomplish this important goal and help to solve such ill-posed problems. Ill-posed problems, as opposed to the well-posed problems, are those in which a unique solution is not determined by the data. In practice it is often auxiliary assumptions of convenience made about the functional form (e.g., linearity, additivity) that allow the data to determine a unique solution.

[18] Adcock and Collier (2001) label predictive validity as nomological validity.

[19] 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 2003; Gelman 2004; Gelman and Shalizi 2012).

https://doi.org/10.1017/psrm.2017.5 Published online by Cambridge University Press

cases we have focused on where the data-generating process is not under researcher control, preference for models with lower generalization error is arguably validity enhancing.[20]

We do not suggest that researchers adopt a single method, or even a particular class of methods in this paper; we simply wish to emphasize 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, though most examples are relatively new, the call for more focus on predictive checking is not new to applied political science research (see e.g., Beck, King and Zeng 2000; Ward, Greenhill and Bakke 2010; Beger, Dorff and Ward 2014; Hill and Jones 2014; Schnakenberg and Fariss 2014; Chenoweth and Ulfelder 2015; Douglass 2015; Graham, Gartzke and Fariss 2015). Given the importance of predictive checking, and the recent discussion of transparency and the replication standard—we view exact replication as specifically a form of model validation—it is an important point to re-emphasize here: regularization[21] can be used to decrease threats to predictive validity from over/underfitting an empirical model by focusing on the minimization of generalization error.[22] Moreover, "data mining" (i.e., the use of statistical/machine learning techniques), when used in the principled fashion described here, *should not* be used as a pejorative term. Instead, such tools should be adopted to help provide evidence for the predictive validity of observational and quasi-experimental designs when exact replication is not possible.

CONCLUSION

In areas of political science research where replication is not possible because the theoretically specified data-generating process is not under the direct control of the researcher (i.e., observational or quasi-experimental designs), flexible methods used with regularization can be used to decrease threats to predictive validity from over/underfitting by minimizing generalization error. This serves a similar function to exact replication in settings where the data-generating process is under the direct control of the researcher (i.e., an experimental or survey designs). We believe that this will be of use in exploratory and/or predictive data analyses where causal relationship(s) of interest are not identified, and, when they are, to contextualize effect sizes and to study the heterogeneity of the estimated effects.

To review, the estimation of generalization error allows for model comparisons that highlight underfitting: when a model generalizes poorly due to missing systematic features of the data-generating process, and overfitting: when a model generalizes poorly due to discovering non-systematic features of the data used for fitting. Relatedly, the estimation and minimization of generalization error provides a principled way to use flexible methods which are suitable for modeling relationships that are left unspecified by a deductively valid theory, which we believe is common. Lastly, model comparison based on generalization error naturally enhances predictive validity and can be a useful default when there are not valid alternatives.

[20] See Jones and Linder (2016) and Friedman (2001) for work on interpreting flexible models when the learned relationship(s) between the explanatory variables and outcomes are not directly interpretable, as is often the case.

[21] For examples of applied research in political science that use regularization see Monroe, Colaresi and Quinn (2008), and Quinn et al. (2010). Note also though that a Bayesian hierarchical model implicitly use regularization (for more details about Bayesian hierarchical models see Tibshirani 1996; Western 1998; Gelman 2003; Gelman 2004; Park and Casella 2008).

[22] 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 with inference about parameters), we consider it premature to argue that any particular phenomena is inherently unpredictable, despite there being some compelling reasons to think that this may be the case in some situations (cf. Gartzke 1999).

https://doi.org/10.1017/psrm.2017.5 Published online by Cambridge University Press

TABLE 2 *Advantages of Using Flexible, Regularized Methods Across Distinct Analytical Goals*

Type of Analysis	Advantages of Fitting a Flexible Regularized Model
Causal explanation	Contextualization of effect size, heterogeneity
Exploration	Discovery of unexpected relationships: nonlinearity, and interactions
Prediction	Decrease generalization error

While it would be desirable to provide specific recommendations, we believe that the diversity of data sources and analytic goals would make such recommendations unsatisfactory. The relative usefulness of any method depends on properties of the data (e.g., collection method, dependence structure, measurement error) and the analytic goal (e.g., causal explanation, exploration, prediction), and an application appropriate loss function applied to statistics such as prediction error, which comport with the analytic goal. However, we believe that within the framework of Shmueli (2010), there are distinct advantages to fitting flexible, regularized models, which are reiterated in Table 2.

To close, we wish to emphasize that scholars using any form of observational data or quasi-experimental data can benefit from the use of that minimize generalization error, which provides evidence for the predictive validity of empirical models. We have offered a brief introduction to the reasoning behind this approach, but much of the difficulty for applied political science research is in the development of appropriate estimators of generalization error for complex data. Again, we believe this to be a productive area for new research in political science and political methodology.

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, and Alain Celisse. 2010. 'A Survey of Cross-Validation Procedures for Model Selection'. *Statistics Surveys* 4:40–79.

Athey, Susan, and Guido Imbens. 2015. 'Machine Learning Methods for Estimating Heterogeneous Causal Effects'. *ArXiv Preprint ArXiv:1504.01132*.

Bailey, Michael A. 2007. 'Comparable Preference Estimates Across Time and Institutions for the Court, Congress, and Presidency'. *American Journal of Political Science* 51(3):433–48.

Bareinboim, Elias, and Judea Pearl. 2012. 'Transportability of Causal Effects: Completeness Results', vol. R-390. Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, Sheraton Centre Toronto, Toronto, Ontario, July 22–26, 2012.

Beck, Nathaniel, Gary King, and Langche Zeng. 2000. 'Improving Quantitative Studies of International Conflict: A Conjecture'. *American Political Science Review* 94(1):21–35.

Beck, Nathaniel, and Simon Jackman. 1998. 'Beyond Linearity by Default: Generalized Additive Models'. *American Journal of Political Science* 42(2), 596–627.

Beger, Andreas, Cassy L. Dorff, and Michael D. Ward. 2014. 'Ensemble Forecasting of Irregular Leadership Change'. *Research & Politics* 1(3): http://journals.sagepub.com/doi/abs/10.1177/2053168014557511.

Bengio, Yoshua. 2000. 'Gradient-Based Optimization of Hyperparameters'. *Neural Computation* 12(8):1889–900.

Bergstra, James, and Yoshua Bengio. 2012. 'Random Search for Hyper-Parameter Optimization'. *The Journal of Machine Learning Research* 13(1):281–305.

https://doi.org/10.1017/psrm.2017.5 Published online by Cambridge University Press

Berk, Richard A. 2004. *Regression Analysis: A Constructive Critique*, vol. 11. Thousand oaks, CA: Sage.
Bischl, Bernd, Olaf Mersmann, Heike Trautmann, and Claus Weihs. 2012. 'Resampling Methods for Meta-Model Validation With Recommendations for Evolutionary Computation'. *Evolutionary Computation* 20(2):249–75.
Brady, Henry E. 1986. 'The Perils of Survey Research: Inter-Personally Incomparable Responses'. *Political Methodology* 11:269–91.
Breiman, Leo. 1996. 'Stacked Regressions'. *Machine Learning* 24(1):49–64.
Chenoweth, Erica, and Jay Ulfelder. 2015. 'Can Structural Conditions Explain the Onset of Nonviolent Uprisings?'. *Journal of Conflict Resolution* 61(2), 2017.
Dafoe, Allan. 2014. 'Science Deserves Better: The Imperative to Share Complete Replication Files'. *PS: Political Science & Politics* 47(1):60–66.
Douglass, Rex W. 2015. 'Understanding Civil War Violence Through Military Intelligence: Mining Civilian Targeting Records from the Vietnam War'. *ArXiv Preprint arXiv:1506.05413v1*.
Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge: Cambridge University Press.
Efron, Bradley. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans*, vol. 38. Philadelphia, PA: SIAM.
Efron, Bradley, and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Boca Raton, FL: CRC press.
Efron, Bradley, Trevor Hastie, Iain Johnstone, Robert Tibshirani, and Stefan Wager. 2004. 'Least Angle Regression'. *The Annals of Statistics* 32(2):407–99.
Elkins, Zachary, and John Sides. 2014. 'The Vodka is Potent, but the Meat is Rotten1: Evaluating Measurement Equivalence Across Contexts'. Working Paper.
Fariss, Christopher J. 2014. 'Respect for Human Rights Has Improved Over Time: Modeling the Changing Standard of Accountability in Human Rights Documents'. *American Political Science Review* 108(2):297–318.
Fariss, Christopher J. Forthcoming. 'Human Rights Treaty Compliance and the Changing Standard of Accountability'. *British Journal of Political Science*. http://dx.doi.org/10.1017/S000712341500054X.
Friedman, Jerome H. 2001. 'Greedy Function Approximation: A Gradient Boosting Machine'. *Annals of Statistics* 29(5):1189–232.
Gartzke, Erik. 1999. 'War is in the Error Term'. *International Organization* 53(3):567–87.
Gelman, Andrew. 2003. 'A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-Fit Testing'. *International Statistical Review* 71(2):369–82.
Gelman, Andrew. 2004. 'Exploratory Data Analysis for Complex Models'. *Journal of Computational and Graphical Statistics* 13(4):755–779.
Gelman, Andrew, and Cosma Rohilla Shalizi. 2012. 'Philosophy and the Practice of Bayesian Statistics'. *British Journal of Mathematical and Statistical Psychology* 66(1):8–38.
Givens, Geof H., and Jennifer A. Hoeting. 2012. *Computational Statistics*, vol. 708. Hoboken, NJ: John Wiley & Sons.
Graham, Benjamin A. T., Erik A. Gartzke, and Christopher J. Fariss. 2015. 'Regime Type, Coalition Size, and Victory'. *Political Science Research and Methods*, doi:https://doi.org/10.1017/psrm.2015.52
Hainmueller, Jens, and Chad Hazlett. 2014. 'Kernel Regularized Least Squares: Reducing Misspecification Bias With a Flexible and Interpretable Machine Learning Approach'. *Political Analysis* 22:143–68.
Handcock, Mark S., Adrian E. Raftery, and Jeremy M. Tantrum. 2007. 'Model-Based Clustering for Social Networks'. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170(2):301–54.
Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition New York, NY: Springer.
Hastie, Trevor J., and Robert J. Tibshirani. 1990. *Generalized Additive Models*, vol. 43 Boca Raton, FL: CRC Press.
Herrnson, Paul S. 1995. 'Replication, Verification, Secondary Analysis, and Data Collection in Political Science'. *PS: Political Science & Politics* 28(3):452–55.

https://doi.org/10.1017/psrm.2017.5 Published online by Cambridge University Press

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

Hoff, Peter D. 2005. 'Bilinear Mixed-Effects Models for Dyadic Data'. *Journal of the American Statistical Association* 100(469):286–95.

Hoff, P. D. 2009. 'Multiplicative Latent Factor Models for Description and Prediction of Social Networks'. *Computational & Mathematical Organization Theory* 15(4):261–72.

Hoff, Peter D., Adrian E. Raftery, and Mark S. Handcock. 2002. 'Latent Space Approaches to Social Network Analysis'. *Journal of the American Statistical Association* 97(460):1090–098.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. 'Unbiased Recursive Partitioning: A Conditional Inference Framework'. *Journal of Computational and Graphical Statistics* 15(3):651–74.

Hothorn, Torsten, Peter Bühlmann, Thomas Kneib, Matthias Schmid, and Benjamin Hofner. 2010. 'Model-Based Boosting 2.0'. *The Journal of Machine Learning Research* 11:2109–113.

Hothorn, Torsten, Peter Buehlmann, Thomas Kneib, Matthias Schmid, and Benjamin Hofner. 2014. 'Model-Based Boosting'.

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

Jones, Zachary M., and Fridolin Linder. 2016. 'edarf: Exploratory Data Analysis using Random Forests'. The Journal of Open Source Software. http://dx.doi.org/10.21105/joss.00092.

Keele, Luke. 2015. 'The Statistics of Causal Inference: A View from Political Methodology'. *Political Analysis* 23:313–35.

Keele, Luke John. 2008. *Semiparametric Regression for the Social Sciences*. Hoboken, NJ: John Wiley & Sons.

Keele, Luke, and Rocío Titiunik. 2015. 'Natural Experiments Based on Geography'. *Political Science Research and Methods* 4(1):65–95.

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, Rochester, NY.

King, Gary. 1995. 'Replication, Replication'. *PS: Political Science and Politics* XXVIII:494–99.

King, Gary. 2006. 'Publication, Publication'. *PS: Political Science and Politics* XXXIX(1):119–25.

King, Gary, Christopher J. L. Murray, Joshua A. Solomon, and Ajay Tandon. 2004. 'Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research'. *American Political Science Review* 98(1):191–207.

Lahiri, Soumendra Nath. 2003. *Resampling Methods for Dependent Data*. New York, NY: Springer.

Lake, David A. 2013. 'Theory is Dead, Long Live Theory: The End of the Great Debates and the Rise of Eclecticism in International Relations'. *European Journal of International Relations* 19(3):567–87.

LeBlanc, Michael, and Robert Tibshirani. 1996. 'Combining Estimates in Regression and Classification'. *Journal of the American Statistical Association* 91(436):1641–650.

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.*

Mentch, Lucas, and Giles Hooker. 2014. 'Ensemble Trees and Clts: Statistical Inference for Supervised Learning'. *ArXiv Preprint ArXiv:1404.6473*.

Mingers, John. 1989. 'An Empirical Comparison of Pruning Methods for Decision Tree Induction'. *Machine Learning* 4(2):227–43.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. 'Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict'. *Political Analysis* 16(4):372–403.

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

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. 'How to Analyze Political Attention With Minimal Assumptions and Costs'. *American Journal of Political Science* 54(1):209–28.

https://doi.org/10.1017/psrm.2017.5 Published online by Cambridge University Press

Schapire, Robert E., and Yoav Freund. 2012. *Boosting: Foundations and Algorithms*. Cambridge, MA: MIT Press.

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

Sexton, Joseph, and Petter Laake. 2009. 'Standard Errors for Bagged and Random Forest Estimators'. *Computational Statistics & Data Analysis* 53(3):801–11.

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

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

Shmueli, Galit. 2010. 'To Explain or to Predict?'. *Statistical Science* 25(3):289–310.

Tibshirani, Robert. 1996. 'Regression Shrinkage and Selection Via the Lasso'. *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1):267–88.

Vapnik, Vladimir Naumovich. 1998. *Statistical Learning Theory* 2nd ed. New York, NY: Wiley.

Wager, Stefan, and Susan Athey. 2015. 'Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests'. *ArXiv Preprint ArXiv:1510.04342*.

Wager, Stefan, Trevor Hastie, and Bradley Efron. 2014. 'Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife'. *The Journal of Machine Learning Research* 15(1):1625–651.

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):363–75.

Western, Bruce. 1998. 'Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modeling Approach'. *American Journal of Political Science* 42(4):1233–259.

Wilcox, Clyde, Lee Sigleman, and Elizabeth Cook. 1989. 'Some Like it Hot: Individual Differences in Responses to Group Feeling Thermometers'. *Public Opinion Quarterly* 53(2):246–57.

Wood, Simon, and Maintainer Simon Wood. 2015. 'Package "Mgcv"'. R Package Version, 1–7.

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):301–20.

https://doi.org/10.1017/psrm.2017.5 Published online by Cambridge University Press