Oxford Handbooks Online

Statistics and International Security

Adam M. Lauretig and Bear F. Braumoeller

The Oxford Handbook of International Security
Edited by Alexandra Gheciu and William C. Wohlforth

Print Publication Date: Mar 2018
Subject: Political Science, International Relations, Political Methodology
Online Publication Date: Apr 2018 DOI: 10.1093/oxfordhb/9780198777854.013.10

Abstract and Keywords

There is a rich legacy of quantitative work in Security Studies, with scholars using regression to make a variety of discoveries about questions of interest. Unfortunately, much of this work pays scant attention to the differences among description, causation, and prediction. This chapter draws on existing work in political science, economics, and statistics to illustrate the distinctions among these approaches and the models and assumptions appropriate for each. The chapter closes with the hope that better quantitative research will lead to improvements in the field of international security and bring everyday methods more in line with the traditions of strong theorizing and effective data-gathering. It also provides resources for the reader to further explore the ideas presented in the chapter.

Keywords: prediction, causality, causal inference, machine learning, potential outcomes, security studies

10.1 Introduction

STUDENTS of international security have been gathering data about conflict for a very long time (see, e.g., Wright 1942). Over time, the sophistication of quantitative analysis has grown as new generations of scholars seek out newer and better methods to capture important features of international relations data (Beck and Zeng 1999; Braumoeller 2003; Cranmer and Menninga 2012). There have been useful correctives along the way—see especially Achen (2002), Clarke and Primo (2007), and Schrodt (2014)—and in all candor, those correctives have largely fallen on willfully deaf ears. Nevertheless, the

overall trend in the statistical study of international security has been in the direction of progress.

There's only one problem: In a very fundamental sense, we don't know what we're doing.[1] Quantitative International Relations (IR) scholars overwhelmingly tend to use statistical models for purposes for which they were not intended and should not be utilized, treating description, causal inference, and prediction as interchangeable. This is a very real problem, and it is time that we face up to it.

The origins of the problem are not too difficult to describe. Most quantitative IR scholars can remember having received the warning that statistical models like regression do nothing more than describe correlations among variables, and that correlation does not mean causation. We all nodded and made a note of that point. Nevertheless, when we read quantitative work in our IR seminars, we looked for stars on the variables of interest and assumed that the statistically significant variable "caused" the effect of interest.

When it came time to do our own research, most of us simply threw our list of variables into a generalized linear model, turned the crank,[2] looked for $p < .05$, and then wrote up the results. We did not actually use the word "cause" in our write-ups, but (p. 134) overwhelmingly the analyses were presented as causal—increasing the variable of interest by 1 would cause a β increase in y. Sometimes, we would even say our variables "predict" a β-unit change in y, further muddying the conceptual waters.

IR scholars are far from unique in this regard. Morgan and Winship (2014) note that, starting in the 1960s and 1970s, methodologists in the social sciences were eager to claim that the methods they described were amenable to a causal interpretation. In the "age of regression" that followed, practitioners proved all too eager to throw caution to the wind when it came to interpreting the partial correlations represented by regression coefficients.

Our point in this chapter is simple: it is time to come back down to earth and actually do the sort of work that we have been claiming to do for decades. In this chapter, we demonstrate that doing so is far from impossible, though it does demand that observational data be treated with the respect and caution that they deserve.

Regression models, logits, probits, and other generalized linear models (GLMs) are fundamentally *descriptive* models: they describe the correlation between a dependent variable and some vector of independent variables. If correlation is all that is needed—if, for example, we just want to know whether democratic dyads are less likely to go to war than other pairs of states—description suffices. If we want to know *why* democratic dyads are less likely to go to war, however, we are attempting to make a causal claim. If we seek to assess the future conflict propensities of pairs of states based on their level of democracy, we are attempting to *predict* their future behavior.

Causal models are carefully designed to achieve *identification*: due to the design of the study or the statistics used to carry it out, we believe that we are identifying a causal effect, where manipulating X*changes* Y, if units under study are otherwise identical, rather than simply a correlation. Identification cannot simply be assumed. If the research design warrants a causal interpretation (say, the data come from an experiment), GLMs, crosstabs, and simple difference-of-means tests can bear a causal interpretation. If the design does not warrant a causal interpretation, as it typically does not in observational settings, some other research design or modeling procedure must typically be brought to bear before results can credibly be said to have a causal interpretation.

While causal modeling is built around the idea of approximating a laboratory experiment using statistical methods, the goal in *prediction* is to maximize one's ability to predict an outcome Y given a set of variables X, with the model evaluated according to some measure of how well it can predict observations that were not used to generate its parameters (Ward et al. 2010) rather than by the statistical significance of individual coefficients.

In the remainder of this chapter, we will explore methods such as matching and instrumental variables for causal inference, and the elastic net and random forests for prediction, all of which are available in the free statistical software R. We encourage the interested reader to explore these packages and their worked examples for themselves in order to get a feel for how these procedures work.

(p. 135) 10.2 Causal Inference

Experiments are held as the gold standard for causal evidence: the analyst can manipulate everything of interest and randomization ensures that the subjects vary systematically only in terms of the treatment (Hyde 2015). However, experimental results in International Relations often encounter problems of external validity. For example, people experiencing war may respond differently than undergraduates competing for money in a laboratory (Driscoll and Maliniak 2016). Simply put, there is no substitute for real-world observational data for conflict scholars. Unfortunately, this is where causal inference becomes difficult.

In order to gain some traction on the problem, we must first specify what we mean by "causation."

10.2.1 Manipulation and Potential Outcomes

The manipulation account of causation, originally developed in the philosophy of science literature (Sekhon 2004 provides an accessible review), is based on the idea that, if a

proposed cause actually does produce a given effect, the manipulation of the cause will result in the manipulation of the effect (but not vice versa) (Holland 1986).

The manipulation account of causation was introduced to the statistics literature as the *potential outcomes framework* by Rubin (1974) and has become the most popular framework for causal inference in current political science research methodology (Keele 2015). The potential outcomes framework was designed to approximate an experiment when the analyst only has observational data to estimate causal effects. Mathematically, the potential outcomes framework states that in a population with an outcome Y and a treatment D (using notation from Morgan and Winship 2014), assuming binary treatments and outcomes, every individual has a potential treatment $d_i = 1$ or $d_i = 0$, and a potential outcome y_i^1 when $d_i = 1$ or y_i^0 when $d_i = 0$. However, the scholar runs into what Holland (1986: 947) named the "fundamental problem of causal inference": we only observe one treatment and outcome for each individual.

$$Y = Y^1 \text{ if } D = 1$$

$$Y = Y^0 \text{ if } D = 0$$

or, more simply:

$$Y = DY^1 + (1 - D)Y^0$$

(p. 136)

If we *could* re-run history and observe treated and untreated outcomes for the same individuals, causal inference would be simple: we could just compare the two and see how big a difference the treatment makes.

In practice, we have to condition for the effects of confounders. Randomizing the treatment is the simplest way to do so. The effects of the cause can then be observed in the difference in means of the treated and control group. The "broadest possible average effect" (Morgan and Winship 2014: 46), which can be calculated using this set-up is the average treatment effect, which, using the expectation operator $E[.]$ from probability theory is:

$$E[Y^1] - E[Y^0]$$

By subtracting the average of the treated outcomes from the control outcomes, we can observe the effect of treatment in the population. If we satisfy two simple (but strong) assumptions, we can claim that this subtraction represents a *causal* effect, and that this effect is identified. In Figure 10.1, this can be seen by subtracting the means of the two distributions, which results in the average treatment effect.

Two assumptions need to be satisfied to make a causal claim: the stable unit treatment value assumption (SUTVA) and the independence assumption. SUTVA "is simply the a priori assumption that the value of Y for unit u when exposed to treatment t will be the

(p. 137) same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive" (Morgan and Winship 2014: 48). SUTVA thus means that any individual is not affected by the treatment assigned to other individuals.

Figure 10.1 Visualizing average treatment effect

Note: ATE = 2.

The independence assumption states that knowing that a unit was assigned to the treatment group tells you nothing about the counterfactual outcome for the control group, and equivalently, assigning a unit to the control group tells you nothing about its counterfactual treatment. Formally:

$$(Y^0, Y^1) \perp\!\!\!\perp D$$

where $\perp\!\!\!\perp$ denotes independence, and D denotes treatment, the counterfactual is independent of the treatment (Morgan and Winship 2014). The key is that the treatment only "flows" forward to the outcome, that the outcome does not tell you anything about the treatment, and that all variation in D is completely random. This is referred to by Rubin (1978: 42) as "ignorability."

10.2.2 Accounting for Nonrandom Treatment

Using only these assumptions, we can engage in causal inference, though as Manski (1990) demonstrated, our inferences will typically be very imprecise. To make progress, we must account for whether observations select into treatment—are the characteristics of observations *endogenous to*, or caused in part by, their propensity for receiving the treatment D? If these characteristics exist and are not accounted for, they can bias our estimate of the average treatment effect. Fortunately, as Rubin (1978) noted, if we can *condition* on those characteristics S that affect treatment, we can refer to the remaining variation as "ignorable."

To understand what is meant by "conditioning," we return briefly to the example of regression. In its simplest form, regression takes the form

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \epsilon$$

Y represents an outcome, D represents the proposed cause of that outcome, X represents a set of confounders, β_0 – β_2 are coefficients to be estimated, and ϵ represents the error term. In order to believe that β_1 accurately captures the causal effect of D on Y in the context of the potential outcomes framework—that is, that increasing D by 1 will cause a β_1-unit change in Y—we must also believe the following:[3]

1. D causes Y

2. ϵ causes Y
3. ϵ does not cause D
4. Y does not cause D
5. Nothing that causes ϵ also causes D.

(p. 138) Assumptions 1 and 2 are typically unproblematic and 4 is typically justified by theory. To believe assumptions 3 and 5, however, we have to believe that the proposed cause D is not caused either by an omitted variable in the error term or by a cause of an omitted variable in the error term. In other words, D must either be truly exogenous or its causes must be both distinct from and uncorrelated with ϵ.

In the densely interconnected and interactive world of human behavior, these are heroic assumptions. In the case of democratic peace, for example, it is entirely possible that unmodeled variation in peaceful and conflictual behavior influences the probability that a state will become or remain democratic (see, e.g., Reuveny and Li 2003), violating assumption 3. Similarly, an unmodeled confounder, like liberal political norms, could influence both conflict behavior and democracy (Weart 2000), violating assumption 5. In either case the treatment, democracy, is nonrandom with respect to the characteristics of the units.

Conditioning on the variables that influence treatment, broadly speaking, involves introducing information about those variables into the analysis. Done sensibly, conditioning makes identification plausible in observational data. Conditioning cannot typically be done without making *some* assumptions about the process of treatment, however, and the plausibility of identification hinges critically on the plausibility of the assumptions necessary to produce it.

10.2.2.1 Matching

Perhaps the most well-known technique for causal inference in political science is matching, originally developed by Rubin (1973). The goal of matching is to approximate the randomization in an experiment, to obtain the best possible estimate of the effect of the treatment on the outcome (Rubin 2008). Treatment is viewed as a probabilistic (usually binary) outcome, for which one estimates the *propensity score*, the probability of receiving treatment. After the propensity score is estimated, treated observations are "matched" to untreated ones, optimizing balance to find those control observations which are most similar to the treated observations.

The key identifying assumption for causal inference is *selection on observables*: all variables that influence nonrandom selection must be observed and incorporated into the analysis. Once they have been, it becomes possible to compare treated and untreated groups without leveraging many additional assumptions—a substantial advantage. However, selection on observables can be a challenging assumption to meet even under ideal circumstances (Sekhon 2009; Keele 2015).

The main advantage to matching is that it enables the analyst to take into account stratification in their data. Using the example above, if X perfectly predicts the probability of receiving treatment D so that there is a "selection effect," by matching on X the analyst can compare treated and control groups, recovering the Average Treatment Effect on the Treated (the ATT). In this case, given a binary treatment and outcome, we are interested in $E[Y_i^1|D_i = 1] - E[Y^0|D_i = 1]$, since we believe the treated and untreated groups are drawn from different populations (Sekhon 2009). While we cannot observe (p. 139) this, by conditioning on the variables X in the population which lead to selection into treatment, $E[Y_i^1|D_i = 1, X] - E[Y^0|D_i = 0, X]$, we can estimate the "observational equivalent" of the ATE (Sekhon 2009). Modeling this using regression returns a biased estimate of β_1, however, since this selection effect violates assumption 3 something in ϵ is causing D. Matching helps to account for this.

There are a variety of ways to estimate the propensity score. Traditionally, the simplest way was to use predicted probabilities from a logistic regression (Rosenbaum and Rubin 1983). Once the propensity score is estimated and a balanced set of cases is selected, the analyst subtracts the control outcomes from the treated outcomes to find the ATE and conducts a simple t-test to see if the difference in means is significant. If the difference in means is significant, we assume that the treatment had an effect on the outcome.

However, the traditional propensity-score approach is sensitive to misspecification (Imai and Ratkovic 2014), scholars should use other techniques for calculating propensity scores, such as optimization using genetic algorithms, known as "genetic matching" (Diamond and Sekhon 2013) or a method of moments estimator which also optimizes balance, the Covariate Balancing Propensity Score (CBPS) (Imai and Ratkovic 2014).

10.2.2.2 Instrumental Variables

Instrumental variables (IVs) have a long lineage in econometrics and in the 1990s were adopted for explicitly identifying causal effects (Angrist et al. 1996). Instruments were originally developed to address issues of endogenous predictors, where in

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \epsilon$$

D and Y are correlated through ϵ, violating either assumption 3 or assumption 5 at the start of Section 10.2.1 or both, depending on the source of the correlation. The goal of instrumental variables modeling is to find a source of exogenous variation (Z) that causes variation in Y *only* through , so that if we look at the relationship between Y and the variation in D we can attribute to Z to infer causality. The key identifying assumption, therefore, is that Z does not cause Y except via D.

IV modeling requires estimating a multiple-equation model, either a two-stage least squares or a simultaneous likelihood model. In the first equation, an instrument Z is estimated for X, where Z only affects Y through D, so that $D = Z\pi + on$, where π is the effect of Z on X. The predicted values (X^*) are then used in place of D in the second equation for Y.

$Y = \alpha + D^{*}\beta + \epsilon$ then returns an unbiased estimate of β, as long as the covariance between Z and ϵ is asymptotically 0, that the instrument Z only effects the outcome Y through the mediating variable D, known as the "exclusion restriction" (Sovey and Green 2011: 190). This relationship can never truly be tested; the analyst must provide a plausible theoretical reason for their instrument choice, such that asymptotically, the exclusion restriction might be met. The analyst must also avoid the "weak instrument problem": asymptotically, the covariance between Z and D asymptotically must not equal zero, otherwise, substantial bias will result (Sovey and Green 2011).

(p. 140)

10.2.2.3 Binary Outcomes and Simultaneous Equations

Another complication with instrumental variables crops up in the case of the nonlinear models, like logit and probit, that are the bread and butter of quantitative international security studies. Because the theory behind instrumental variables was developed in the context of the linear model, the IV method does not translate especially well to the nonlinear case (Greene 2009).

The solution to nonrandom treatment in nonlinear models involves modeling the sources of both treatment and outcome in simultaneous equations. The model that captures the logic of endogenous binary treatment most intuitively is the *recursive bivariate probit* model, which has seen surprisingly little use in political science:

$$Y^{*} = \beta_0 + \beta_1 D + \beta_2 \mathrm{X} + \epsilon$$

$$D^{*} = \gamma_0 + \gamma_1 \mathrm{W} + \upsilon$$

$$-1 \leq \rho(\epsilon, \upsilon) \leq 1$$

Simply put, treatment is a function of W and the outcome is a function of treatment and X. The impact of residual unobserved confounders is captured in the correlation (ρ) between ϵ and υ.

In this model, identification can take place even if all variables that influence selection are not observed and measured, so one need not assume selection on observables as in matching. Instead, identification is obtained both via the exclusion restriction (at least one variable in W must be legitimately excluded from X) and via functional-form assumptions (e.g. that the link functions for the two equations are probits and the distribution of the error terms follows a cumulative normal distribution). To reduce reliance on strong functional-form assumptions, Braumoeller et al. (2017) have introduced a flexible recursive bivariate model to political science that allows the user to relax many of the parametric assumptions of the standard model.

10.2.2.4 Other Tools for Causal Inference

We have only scratched the surface of a vast literature on tools for causal inference. Among those techniques we did not address were regression discontinuities, difference-in-differences (Angrist and Pischke 2008), inverse propensity score weighting and doubly robust estimation (Morgan and Winship 2014), and the role of time in causal inference (Blackwell 2013). Mediation analysis is a rapidly growing sub-field of the causal inference literature, with a focus on the mechanisms by which the treatment affects the outcome (Imai et al. 2011). The development of sensitivity analyses has provided another flourishing literature in causal inference: building on the work of Manski (1990), scholars build tests to examine how strong assumptions have to be to make a particular design work (Mebane and Poast 2013).

(p. 141) Finally, there is the graphical approach to causality, developed by Pearl (2009) and Spirtes et al. (1993), which expresses causality in the form of Directed Acyclic Graphs (DAGs) and conditional probabilities. Though the graphical and potential outcomes frameworks are mathematically equivalent (Morgan and Winship 2014), the clean representation of the graphical model can clarify the math and intuitions of the equivalent potential outcomes model. Despite their origins in computer science rather than the social sciences and the relative recency of their development, DAGs have begun to make some inroads into political science (see, e.g., Blackwell 2013; Imai et al. 2014).

10.2.2.5 The Next Frontier: Interference

As discussed at the start of Section 10.2.1, one of the key assumptions in causal inference is SUTVA, the stable unit treatment value assumption, in particular, a lack of interference, where subject *i*'s treatment does not affect subject *j*'s outcome, and vice versa. Network effects, which are often present in International Relations data, raise challenging issues for causal inference: how does one (for example) match on dyads? Not accounting for network structures can bias our inferences: we are no longer measuring $Y = DY^1 + (1 - D)Y^0$, but rather $Y = (DY^1 + \text{other stuff}) + ((1 - D)Y^0 + \text{other stuff})$; not accounting for this "other stuff" biases our estimate of the average treatment effect.

Some solutions are available, but no consensus exists regarding a "best" or "default" solution. If observations are grouped and interference occurs within but not between groups, a multilevel model is appropriate. By fitting a model with slopes and intercepts which vary across groups, one can isolate the effect of treatment (Gelman and Hill 2006). However, if the analyst suspects that there is a complex network structure driving the observed outcomes (Cranmer et al. 2012), the modeling strategy will need to involve these network structures. The analyst could make use of a variety of approaches to network causal inference developed in epidemiology and biostatistics (Ogburn and VanderWeele 2014), statistics (Athey 2016), or political science (Bowers et al. 2013). The downside to these methods is that they often require a far larger *N* (more observations) than International Relations data can provide and also usually fail to take time (*T*) into account.

An open challenge for the field is developing a method to take advantage of the data structures common in international security data: a small N ($\approx$ 150 countries or 22,350 dyads per year), but a large T, with the Correlates of War extending from 1816 to 2010 (Palmer et al. 2015). Future work could attempt to leverage this longitudinal structure to estimate causal effects over time in the presence of interference.

10.2.3 What Can Causal Inference Offer IR?

At first, the promise of causal inference for students of IR seems great: simply switching from ordinary probit to recursive bivariate probit would not be much of a challenge for most practitioners, and the results would more plausibly capture causal effects. (p. 142) However, since we assume causation requires manipulation, simply changing the model does not mean we are suddenly "doing causal inference." We must think about the design underlying our models.

The key insight from the study design literature is that the design of a causal study should be robust, relying on several tests of the underlying theoretical cause, rather than just a single test (Keele and Minozzi 2013). Rosenbaum (2010) refers to this as testing for the "reasons for effects," while causal inference is usually interested in the "direct causes." We might also ask what else a theory implies, to assess the strength of our theorized cause.

For example, hypothesis testing in the Democratic Peace literature is often reduced to one or two dependent variables (MIDs or wars from Palmer et al. 2015), and an independent variable (regime, usually coded by Polity score; see Marshall and Jaggers 2002). Scholars run a handful of regressions, tweaking "control variables," but rarely looking at alternative measures of the outcomes or treatments upon which their theory depends.

What might they do instead? In asking what their particular theory of the Democratic Peace implies, they might ask if they would expect to see similar patterns if they operationalize regime differently (as Weeks 2014 did), or if they operationalize conflict differently. For example, rather than examining wars, if a scholar expects that democratic dyads are more pacific, they might use event data (Schrodt et al. 2008)[4] and investigate whether there are fewer "conflictual" events and more "cooperative" events between democracies, as compared to autocracies. They could also, following Keele and Minozzi (2013), utilize an array of different research designs and statistical models designed for causal inference (matching, instrumental variables, natural experiments, and the like) in an attempt to triangulate an answer.

Regardless, the challenges of causal inference do not justify inaction. If the goal is to move beyond simplistic regression models and make causal claims, scholars, even in the presence of data which make standard causal inference techniques difficult, can use the principles of causal inference to better test the causal claims in many IR theories.

10.3 Prediction

Prediction gets a bad rap in International Relations. Kenneth Waltz (1979) dismissed it and later argued that "tests [of theory] are always problematic" and that prediction should not be the goal of theory (Waltz 1997, 2016). This approach has been contested by Ward et al. (2010) and Schrodt (2014) among others, who argue that predictive modeling is the ultimate test for the validity of theory. We argue, following Shmueli (2010), that explanation (causal inference) is fundamentally different than prediction. Both have utility, but scholars should not confuse causal models for predictive ones or vice versa.

What is prediction? Following Shmueli (2010), we define prediction as using models or algorithms to predict unobserved values of Y, given some set of predictors X. Models "trained" on some existing data where both X and Y are known are then "tested" on (p. 143) some new data, where one might only have X. The scholar then "predicts" the new Y values. In the remainder of this section, we will review the process of prediction, how to assess predictive models, and conclude with a sample of various techniques that IR scholars can use for prediction.

10.3.1 The Process of Prediction

Unlike causal inference, prediction is usually not interested in coefficient values, statistical significance, or "the exact role of each variable in terms of an underlying causal structure," (Shmueli 2010: 300) but, rather, in modeling the data-generating process to best approximate Y given X. Many of the approaches used for prediction are "algorithmic:" The model is a "black box" for approximating the best model of Y, given X (Breiman 2001b). The goal of these models is to choose some combination of variables which optimizes some function (mean-squared error, Akaike Information Criterion (AIC), etc.) to find the best fit. Unlike GLMs, where the data-generating process is (more or less) linear and additive, many algorithmic predictive models can capture nonlinear, non-additive aspects of a data-generating process.

How does the analyst determine if a model is predicting well? The common technique for testing predictive models in the absence of new data is cross-validation. The data are partitioned into "training" and "test" sets, with models fitted on the training data and not allowed to "see" the test data, and then x-values from the test data are used to predict new y-values, which are then compared to the observed y-values in the test data. This *out-of-sample* prediction, has become more common in political science, especially since the publication of King and Zeng (2010).

10.3.2 Assessing Predictive Models

Once the analyst has fit a variety of models and wants to find the model that best captures their data-generating process, they need to decide on a criterion to assess their predictive model. If the analyst fits a model by maximum likelihood, they can use the AIC, which takes the model deviance and adds a penalty for number of predictors, penalizing predictors which only add noise (McElreath 2016 provides an easy-to-understand derivation of the AIC). The result is a model fit criterion which asymptotically measures out-of-sample prediction (Gelman and Hill 2006). However, there is no guarantee the assessment holds in finite samples, especially if the coefficient distributions are not Gaussian (McElreath 2016). Instead, while scholars can (and should!) use AIC in their models, they should also choose among metrics that are designed to work with observed out-of-sample predictions.

The simplest of these is mean-squared error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

(p. 144) MSE squares the difference between actual (y_i) and predicted ($\hat{f}(x_i)$) outcomes and then takes the average (James et al. 2013). A smaller MSE means better out-of-sample fit, the squared term penalizes outliers more severely. The MSE is most useful for assessing non-binary outcomes; the squared term simply provides the absolute value in the binary case: $(0-1)^2 = (1-0)^2$.

For binary outcomes, like MID/war onset, a broad literature on assessment has developed in the machine learning community. One tool in particular, the Receiver Operating Characteristic (ROC) curve, has become a standard tool for assessing out-of-sample fit. Originally developed in signal processing during the Second World War, ROC curves were imported to political science by King and Zeng (2001). Generating the ROC curve requires taking the predicted probabilities generated by a binary outcome model, using these probabilities to predict a binary outcome, and then comparing the correctly predicted outcomes to the incorrectly predicted outcomes (James et al. 2013), for the entire probability range $[0, 1]$ generated by the model. This creates a curve, and "area under the curve" (AUC) can be used to assess binary predictive model fit; a higher AUC indicates a larger proportion of true positives compared to false positives, the model is predicting more "correct" than "incorrect" outcomes. If a model has an AUC below $.5$, the number of true positives is less than the number of false positives, and the model is doing worse than flipping a coin to predict the outcome (Davis and Goadrich 2006). The ROC curve can be used both in- and out-of-sample; in our examples, we calculate it for out-of-sample prediction.

It is worth noting that if we have far more 0s than 1s, an ROC curve will not always find the best model for predicting 1s, since it will treat predicting true positives for both 0 and 1 as equally important. If we care more about 1 than 0, this is not helpful. The precision-recall (PR) curve, built for data with highly-imbalanced dependent variables will be far more useful. Using the same probabilities generated by a binary model that ROC curves

use, PR curves divide the number of correctly predicted ones (precision) by the total number of ones predicted at each point on the predicted probability (recall). The PR curve, like the ROC curve, provides a useful metric to evaluate models and penalizes models that do well on AUC by generating mostly zeroes. Indeed, models that perform the best on precision-recall perform the best on AUC, but the converse does not hold (Davis and Goadrich 2006); security scholars interested in predicting conflict would be wellserved by validating their models using precision-recall curves.

10.3.3 Varieties of Predictive Models

A broad variety of models for prediction exist in the statistics and machine learning literature, and entire books have been written on them (see Hastie et al. 2009; James et al. 2013)[5]

In this section we will focus on three common models—the generalized linear model, the elastic net, and the random forest—and illustrate these models on a simulated dataset with a rare outcome.[6] The GLM in this case is a logistic regression. The elastic net, (p. 145) implemented in the "glmnet" package in R, comes from Zou and Hastie (2005) by way of Hastie et al. (2009), and grows out of Tibshirani's (1996) work on the LASSO (an acronym for "Least Absolute Shrinkage and Selection Operator"). The elastic net penalizes overly complex models, regularizing coefficients by forcing them toward zero to prevent overfitting, unless they add predictive power. A key advantage to the elastic net for security scholars is that, unlike traditional regression, it works when there are more variables than observations, even when these variables are highly correlated. Since the elastic net is built around a GLM, it runs relatively quickly, even with large datasets.

Random forests, originally developed by Breiman (2001a), are a black box tool for regression and prediction which improve upon regression trees by creating an ensemble of trees that split outcomes based on predictors and, by iterating this process many times (1000 in our simulation), reduce overfitting, bias, and variance. The key advantage to random forests is that they make no functional form assumptions about the data-generating process and so are more flexible than a GLM-based approach.

In the following simulation, we generated 1000 observations with 100 possible variables. The outcome is highly skewed, with ≈10% 1s. The outcome data are generated from the first 35 of these variables; models are trained on 80 percent of the data and tested on the remaining 20 percent. ROC and precision-recall curves are used to measure model fit in out-of-sample prediction (Figures 10.2 and 10.3).

Examining this outcome, the elastic net outperforms both logistic regression and the random forest. In the ROC curve, the best-predicting model should hug the upper-left corner: the elastic net makes a sizable predictive improvement on both the random forest and the logistic regression. In the precision-recall curve, the best-predicting model should hug the top right corner, again the elastic net does best. While the logistic regression appears to perform better than the random forest, its recall is zero below this $\approx .55$ cutoff, so, while it may predict these zeroes (peace) correctly, as scholars, we are not interested in this, since it cannot tell us when there is a one (war). When the analyst uses these plots to check their analyses, they should keep in mind potential differences in the two, and check both, to make sure their results are not the results of particular model specifications and/or quirks in the data.

10.4 Combining Prediction and Causal Inference?

As we hope to have demonstrated, prediction and causal inference are different tasks, and those tasks are typically best accomplished with different modeling approaches. However, several machine learning techniques for prediction are also contributing to causal inference, as datasets become so large that traditional methods no longer work.

A new and exciting set of developments for security scholars is the use of machine learning approaches for causal inference in high-dimensional data (see Athey 2015 for a general overview) which use predictive methods in the service of causal inference. (p. 146) (p. 147) Propensity scores are essentially a prediction task, and when dozens (or hundreds) of variables can potentially play a role in the propensity score, traditional methods like logistic regression are overwhelmed. Machine learning methods can be used to build propensity scores from large collections of variables using ensembles of models, and then standard causal inference techniques are used to estimate effects. Van der Laan and Rose (2011) provide a suite of methods for this approach, and Samii et al. (2016) provide a political science example. Similarly, Hill (2012) suggests using Bayesian Additive Regression Trees, a Bayesian variation on random forests, to flexibly model response to treatment, without making parametric assumptions about the distribution of the outcome.

Figure 10.2 ROC plot for three models

Figure 10.3 Precision-recall curve plot for three models

Recent work by economists using machine learning techniques for causal inference holds promise for security scholars as well. As Belloni et al. (2014) and Athey and Imbens (2016) discuss, an analyst facing hundreds of possible variables to include in a causal analysis can either select a handful in an ad hoc manner, or use machine learning tools like the LASSO or random forests to select the subset of variables which matter for analysis, and are the asymptotically best choice. Particularly in an era of “big data,” where scholars will have access to hundreds of variables and hundreds of thousands of observations, tools like these are a promising path for dealing with large datasets.

10.5 Conclusion

In this chapter, we have discussed the difference between prediction and causal inference in statistics, and what this means for scholars of international security. Following Shmueli (2010), we sought to highlight that these are two different activities, with different goals, modeling strategies, and methods to analyze data. We do not set the two up as competing approaches in a zero-sum empirical game, but rather as different ways to evaluate the empirics underlying theories, each of which is well-suited to a different set of statistical tools.

We leave with the hope that our discussion of the distinction between prediction and causation and our call for better quantitative research will lead to improvements in the field of international security and bring our everyday methods more in line with our traditions of strong theorizing and effective data-gathering.

References

Achen, Christopher H. 2002. Toward a New Political Methodology: Microfoundations and ART. *Annual Review of Political Science*, 5(1): 423–50.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434): 444–55.

Athey, Susan. 2015. Machine Learning and Causal Inference for Policy Evaluation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 5–6. Syndey: ACM.

Athey, Susan and Guido Imbens. 2016. The State of Applied Econometrics—Causality and Policy Evaluation. *arXiv preprint arXiv:1607.00699*.

Athey, Susan, Dean Eckles, and Guido W. Imbens. 2016. Exact P-values for Network Interference. *Journal of the American Statistical Association.* Doi/ 10.1080/01621459.2016.1241178.

Barber, David. 2012. *Bayesian Reasoning and Machine Learning*. Cambridge: Cambridge University Press.

Beck, Nathaniel, Gary King, and Langche Zeng. 1999. Improving Quantitative Studies of International Conflict: A Conjecture. *American Political Science Review*, 94(1): 21–36.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. High-Dimensional Methods and Inference on Structural and Treatment Effects. *The Journal of Economic Perspectives*, 28(2): 29–50.

Blackwell, Matthew. 2013. A Framework for Dynamic Causal Inference in Political Science. *American Journal of Political Science*, 57(2): 504–20.

Bowers, Jake, Mark M. Fredrickson, and Costas Panagopoulos. 2013. Reasoning about Interference between Units: A General Framework. *Political Analysis*, 21(1): 97–124.

Braumoeller, Bear F. 2003. Causal Complexity and the Study of Politics. *Political Analysis*, 11(3): 209–233.

Braumoeller, Bear F., Giampiero Marra, Rosalba Radice, and Aisha Bradshaw. 2017. Flexible Causal Inference for Political Science. *Political Analysis*, forthcoming.

Breiman, Leo. 2001a. Random Forests. *Machine Learning*, 45(1): 5–32.

(p. 149) Breiman, Leo. 2001b. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3): 199–231.

Clarke, Kevin A. and David M. Primo. 2007. Modernizing Political Science: A Model-Based Approach. *Perspectives on Politics*, 5(4): 741–53.

Cranmer, Skyler J., Bruce A. Desmarais, and Elizabeth J. Menninga. 2012. Complex Dependencies in the Alliance Network. *Conflict Management and Peace Science*, 29(3): 279–313.

Davis, Jesse and Mark Goadrich. 2006. The Relationship between Precision-Recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–40. Pittsburgh, PA: ACM.

Diamond, Alexis and Jasjeet S Sekhon. 2013. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics*, 95(3): 932–45.

Driscoll, Jesse and Daniel Maliniak. 2016. Did Georgian Voters Desire Military Escalation in 2008? Experiments and Observations. *The Journal of Politics*, 78(1): 265–80.

Gelman, Andrew and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gelman, Andrew and Eric Loken. 2014. The Statistical Crisis in Science: Data-Dependent Analysis—A “Garden of Forking Paths”—Explains Why Many Statistically Significant Comparisons Don’t Hold Up. *American Scientist* 102(6): 460.

Greene, William H. 2009. Discrete Choice Modeling. In Terence C. Mills and Kerry Patterson (eds), *Palgrave Handbook of Econometrics*, Vol. 2 (Applied Econometrics), pp. 473–556. Basingstoke: Palgrave Macmillan.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Vol. 1 Springer series in statistics. Berlin: Springer.

Hill, Jennifer L. 2012. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*. Doi/10.1198/jcgs.2010.08162

Holland, Paul W. 1986. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396): 945–60.

Hyde, Susan D. 2015. Experiments in International Relations: Lab, Survey, and Field. *Annual Review of Political Science*, 18: 403–24.

Imai, Kosuke and Marc Ratkovic. 2014. Covariate Balancing Propensity Score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1): 243–63.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies. *American Political Science Review*, 105(4): 765–89.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 6. Berlin: Springer.

Keele, Luke. 2015. The Statistics of Causal Inference: A View from Political Methodology. *Political Analysis*, 23(3): 313–35.

Keele, Luke and William Minozzi. 2013. How Much is Minnesota Like Wisconsin? Assumptions and Counterfactuals in Causal Inference with Observational Data. *Political Analysis*, 21(2): 193–216.

King, Gary and Langche Zeng. 2001. Improving Forecasts of State Failure. *World Politics*, 53(4): 623–58.

McElreath, Richard. 2016. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Vol. 122 Boca Raton, FL: CRC Press.

(p. 150) MacKay, David J. C. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press.

Manski, Charles F. 1990. Nonparametric Bounds on Treatment Effects. *The American Economic Review*, 80(2): 319–23.

Marshall, Monty G. and Keith Jaggers. 2002. Polity IV project: Political regime characteristics and transitions, 1800–2002. *Available at:* **http://www.systemicpeace.org/polityproject.html**, accessed August 30, 2107.

Mebane, Walter, R Jr. and Paul Poast. 2013. Causal Inference without Ignorability: Identification with Nonrandom Assignment and Missing Treatment Data. *Political Analysis*, 21(2): 233–51.

Morgan, Stephen L. and Christopher Winship. 2014. *Counterfactuals and Causal Inference*. Cambridge: Cambridge University Press.

Ogburn, Elizabeth L., and Tyler J. VanderWeele. 2014. Vaccines, Contagion, and Social Networks. *arXiv preprint arXiv:1403.1241*.

Palmer, Glenn, Vito d'Orazio, Michael Kenwick, and Matthew Lane. 2015. The MID4 Dataset, 2002–2010: Procedures, Coding Rules and Description. *Conflict Management and Peace Science*, 32(2): 222–42.

Pearl, Judea. 2009. *Causality*. Cambridge: Cambridge University Press.

Reuveny, Rafael and Quan Li. 2003. The Joint Democracy–Dyadic Conflict Nexus: A Simultaneous Equations Mode." *International Studies Quarterly*, 47: 325–46.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. Vol. 10. Berlin: Springer.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70(1): 41–55.

Rubin, Donald B. 1973. Matching to Remove Bias in Observational Studies. *Biometrics*, 159–83.

Rubin, Donald B. 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5): 688.

Rubin, Donald B. 1978. Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics*, 6(1): 34–58.

Rubin, Donald B. 2008. For Objective Causal Inference, Design Trumps Analysis. *The Annals of Applied Statistics*, 2(3): 808–40.

Samii, Cyrus, Laura Paler, and Sarah Zukerman Daly. 2016. Retrospective Causal Inference with Machine Learning Ensembles: An Application to Anti-recidivism Policies in Colombia. *Political Analysis*, 24(4): 434–56.

Schrodt, Philip A. 2014. Seven Deadly Sins of Contemporary Quantitative Political Analysis. *Journal of Peace Research*, 51(2): 287–300.

Schrodt, Philip A., Omür Yilmaz, Deborah J. Gerner, and Dennis Hermreck. 2008. The CAMEO (Conflict And Mediation Event Observations) Actor Coding Framework. In *Annual Meeting of the International Studies Association*.

Sekhon, Jasjeet S. 2004. Quality Meets Quantity: Case Studies, Conditional Probability, and Counterfactuals. *Perspectives on Politics*, 2(2): 281–93.

Sekhon, Jasjeet S. 2009. Opiates for the Matches: Matching Methods for Causal Inference. *Annual Review of Political Science*, 12: 487–508.

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

Sovey, Allison J. and Donald P Green. 2011. Instrumental Variables Estimation in Political Science: A Readers' Guide. *American Journal of Political Science*, 55(1): 188–200.

Spirtes, Peter, Clark N. Glymour, and Richard Scheines. 1993. *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

(p. 151) Tibshirani, Robert. 1996. Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–88.

Van der Laan, Mark J. and Sherri Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Berlin: Springer Science & Business Media.

Waltz, Kenneth N. 1979. *Theory of International Politics*. Long Grove, IL: Waveland Press.

Waltz, Kenneth N. 1997. Evaluating Theories. *American Political Science Review*, 91(4): 913–17.

Ward, Michael D. 2016. Can We Predict Politics? Toward What End? *Journal of Global Security Studies*, 1(1): 80–91.

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.

Weart, Spencer R. 2000. *Never At War: Why Democracies Will Not Fight One Another*. New Haven, CT: Yale University Press.

Weeks, Jessica L. P. 2014. *Dictators at War and Peace*. Ithaca, NY: Cornell University Press.

Wright, Quincy and Louise Leonard Wright. 1942. *A Study of War*. Chicago: University of Chicago Press.

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.

Notes:

(1.) The junior author of this piece should be held blameless for the blunt tone of this and the remaining paragraphs in this section.

(2.) God only knows how many times (Gelman and Loken 2014).

(3.) Although we have read many definitions of endogeneity, user Bill's is the clearest and most succint: http://stats.stackexchange.com/questions/59588/what-do-endogeneity-and-exogeneity-mean-substantively. Because we cannot improve on Bill's list of five assumptions, we have reproduced them here almost verbatim, with notation slightly tweaked to be in line with Morgan and Winship (2014).

(4.) These are daily-level events data that record a variety of concrete actions, from "made a speech" to "employ aerial weapons."

(5.) Available online from the authors:

http://www-bcf.usc.edu/gareth/ISL/ISLRSixthPrinting.pdf, http://statweb.stanford.edu/tibs/ElemStatLearn/printings/ESLIIprint10.pdf, http://www.inference.phy.cam.ac.uk/itprnn/book.pdf,http://web4.cs.ucl.ac.uk/staff/D.Barber/pmwiki.php?n=Brml.online Barber (2012) and MacKay (2003).

(6.) All code is available for replication at https://github.com/adamlauretig/Statisticsandinternationalsecurity.

Adam M. Lauretig

Adam M. Lauretig is a PhD candidate in Political Science at The Ohio State University.

Bear F. Braumoeller

Bear F. Braumoeller is Associate Professor of Political Science at The Ohio State University.