A new framework for evaluating theory: shifting from null-hypothesis testing to predictive analytics in political science

The standard approach to political science research is null-hypothesis statistical testing (NHST). This framework depends on correctly specifying empirical models and meeting very restrictive assumptions, many of which are rarely met, impossible to satisfy in practice, or practically ignored in the review process. Whether these assumptions are treated with due dilligence or not has no bearing on the consequence when these assumptions are not satisfied. A NHST approach that does not satisfy a model's assumption makes the building blocks of hypothesis testing (e.g. standard errors, t-tests, and p-values) unreliable, and thus not safe to draw inferences from. Moreover, NHST is prone to overfitting because models are fit on full samples and predictive accuracy is not considered as a metric of model performance there is no way to test how well a model predicts out-of-sample, and thus this approach is detached from usefulness in real life. I propose a shift toward evaluating the models generated to represent theories and hypotheses by considering their added value to a base model's predictive accuracy. This approach is more versatile, less dependent on assumptions, and commonly more easy to implement than NHST. Furthermore, it puts political science research in a better place to make policy prescriptions because it considers out-of-sample prediction accuracy as a paramount metric of a theories validity.

Introduction

> "Anticipating the future is both a social obligation and intellectual challenge that no scientific discipline can escape."
> Schneider et al. (2010), p. 1

> "No matter how I turn it over in my mind, the number one task of peace research always turns out to be that of prediction"
> Singer (1973)

> " . . . explanation in the absence of prediction is not scientifically superior to predictive analysis, it isn't scientific at all! It is, instead, 'pre-scientific'."
> Schrodt (2014), p. 290

Do modern political science theories effectively explain and therefor predict the phenomena that are being studied? How can political scientists assess the validity of their theories? I believe these two questions capture the driving force of today's research agendas. Each researcher wants to understand the phenomena they have dedicated themselves to, and as a result they also want to provide the best evidence that their theory and hypotheses are correct and predict best (Muchlinski et al. 2016; Singer 1973).

Null-hypothesis statistical testing (NHST) is the standard quantitative analytic approach in modern political science. Since the advent of personal computers, and the subsequent availability of empirical modeling software such as Stata, R, Python, etc. it has become increasingly easy to estimate both simple and complex models to assess the validity of theories. This increased ease of access has been coupled with an increased expectation of quantitative analysis. This general process adheres to the following steps:

1. Theorize and hypothesize about some phenomena.

2. Establish some a priori heuristic to judge empirical analysis of the theory. Most commonly this is setting an alpha level of .05.

3. Identify relevant data and collect it.

4. Estimate models to assess the relevant null hypotheses (H_0) identified in step one.

5. Using the heuristic established in step two infer what the models say about the validity of the theory and hypothesis of step one. Most commonly, if the p-value is less than or equal to the alpha level then the corresponding coefficient is considered statistically significant.

This dissertation will focus on proposing a new methodological framework for quantitative analysis in political science that focuses on evaluating how well theories explain political events. The NHST approach over past decades offers only weak tools for analyzing how effectively variables of interest explain the outcome, and inference based on coefficients related to hypotheses is the central focus. Instead, I propose a paradigmatic shift in what political science should consider paramount. Specifically, the ultimate litmus test for whether a theory or hypothesis is good or not should be whether or not it improves out-of-sample prediction. Any variable that does not improve out-of-sample prediction is very likely not theoretically relevant to the event that occurs in its current operationalization. My framework, framework for out-of-sample prediction (FOOSP), argues that it is imperative to move theoretical evaluation toward prediction, and away from NHST.

There are two central problems with NHST: first, p-values are limited in the information that they provide, and difficult to have confidence in; second, goodness of fit tests and NHST in general provide no guard against overfitting, and provide no tools for adjudicating between which theories are better at predicting.

NHST depends on p-values and statistical significance to determine the validity of theories and hypotheses. The p-value is the proportion of the distribution of the test statistic (typically t-test or z statistic) that remains in the tails beyond the observed test statistc. For an explicit definition see Gailmard (2014, p.244): "The p-value is the probability, assuming H_0 is true, that the random test statistic is at least as extreme (relative to the expected value under H_0) as its observed value." Put into a mathematical expression a p-value is simply $P(data|H_0)$ (Levine et al. 2008). Thus, a p-value represents strictly the probability of observing the test statistic of a coefficient produced from the model estimated on a sample given H_0 is true.

A p-value has very limited information to infer from when used correctly. At its best, it tells the researcher if the observed relationship is likely or unlikely given the null hypothesis. However, this interpretation is often stretched beyond this meaning and is interpreted in ways that are not accurate or scientific. First, it is commonplace for no strict alpha level to be set, and thus the criterion for statistical significance becomes ad hoc. Most authors use multiple levels of significance and stars to denote different p-values (e.g. Piazza 2012). This practice is so ingrained and common that it is the default of the 'esttab' package in Stata, which outputs regression coefficient tables. Second, it is becoming common to include relatively large p-values as significant such as .1 (e.g. Daxecker and Prins 2013; Berrebi and Klor 2006; Young 2013) although it is questionable at this point whether one should feel safe rejecting the null hypothesis. Third, the interpretation of these different alpha levels as more or less evidence for a theory is also common. Take for example Bleek and Lorber (2014, p. 442, emphasis added)): "... security guarantees remained insignificant, although they came very close to the 10 percent threshold. For the reproduced logit model, security guarantees remained *extremely significantly* negatively correlated with all the three stages of proliferation behavior." In this example a small p-value is "extremely significant" in the authors' interpretation despite there being no such thing as an extremely significant p-value. In truth, they can only conclude with a high degree of certainty that it is unlikely to observe that relationship given that the null is true.

P-values offer very little in the way of scientific study of a phenomena, because, at best, they offer a probability that the relationship observed is true given the null hypothesis. This is distinct from offering the probability that the hypothesized alternative relationship is true and works as theorized. This will be discussed in greater detail later, but I am not alone in my staunchness against p-values. Discussing his home field of psychology Meehl (1979, p.817) wrote: "I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology." While his sentiments are stronger than mine, I agree in principal and will show that a fundamental analytic shift is needed. Statistical significance does not tell a researcher anything about

how well a variable explains the outcome, and this is fundamentally the concern of scientific research (Schrodt 2014).

Notably, there are goodness of fit tests for each model type, and different ways to analyze variance explained. In OLS there is R^2, the percent of variance explained by the model, and $\bar{R}^2$, the same as R^2 but penalized for the addition of variables to the model. Logistic regression has receiver operating characteristics (ROC) curves and the area under the curve (AUC). Each of these tools is useful in understanding how a model is fit to the sample used to train it, but none of them provide any information of how a model predicts or explains outside of the sample nor do they provide specific evidence to caution against overfitting. These measurements also provide no tools for discerning which models are better, because multiple different combinations of variables can produce similar goodness of fit statistics and there is no way to compare the importance of coefficients (Breiman 2001). Moreover, these residual-based methods for explanatory assessment are not very useful in a multi-dimensional space (Breiman 2001; Landwehr et al. 1984; Bickel et al. 2006)

Also notable is that political science research almost always uses the full sample to train the model, and does not take into consideration overfitting the data. Overfitting occurs when a model "follows the errors, or noise, too closely (James et al. 2017, p. 22)." When a model is fit on the entire sample, and nothing is withheld to test a model on, overfitting is near impossible to detect. Models that are overfit produce coefficients that are partially describing randomness, and thus are not generalizable. Inferences that are not generalizable are not useful outfside of the given sample.

One further broad critique of NHST is the general lack of discussion of prediction. Given the central goal of explaining political phenomena, it is not possible to evaluate the explanatory power of a theory effectively without considering out-of-sample prediction. Any models not tested using out-of-sample prediction are prone to unidentified overfitting, and offer no evidence of a theories predictive power. A simple solution to this problem is to consider a model's predictive ability out-of-sample both with and without the variables of interest to the theory at hand. Then, the base model that excludes the new theoretical contribution can be used as a control for whether

The implementation of even simple machine learning approaches can resolve these problems. My framework is proposed as a set of increasingly more difficult tools to adopt, proposing that all research adopts the minimally invasive strategy in all contexts, and suggesting that all research strive to adopt as many strategies as possible. This framework also poses a fundamental shift from variable-by-variable significance testing to entire model (and thus theory) predictive assessment. This framework is based on the belief that theories should fundamentally be judged by their ability to predict and forecast out-of-sample, and theories that cannot predict, or improve prediction, are not good theories (Singer 1973; Schrodt 2014; Hegre et al. 2019; Schneider et al. 2010; Hegre et al. 2017).

Problems with NHST

The abuse of p-values as a metric for whether or not an idea or variable merits publication in a scientific setting is a sufficient problem in all quantiative fields that the American Statistical Assocaition recently released a statement on the use of p-values (Wasserstein and Lazar 2016). In this there are six principles, which are important to denote at the outset of the discussion of why NHST is problematic (pp. 131-132):

1. "P-values can indicate how incompatible data are with a specified statistical model."

2. "P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone."

3. "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specified threshold."

4. "Proper inference requires full reporting and transparency."

5. "A p-value, or statistical significance, does not measure the size of an effect or the importance of a result."

6. "By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis."

The first point suggests that p-values can indicate whether or not data is compatible with the null hypothesis. Small p-values indicate incompatability. Point two is elaborated to explain that the p-value contains information only about the relationship between the data and the specified null and alternative hypotheses, and that is it. Point three argues that binary decision-points are arbitrary, and adjudicating a coefficient's importance based on a p-value is misguided and incomplete. Point four encourages the sharing of all specifications, which guards against the practice of 'p-hacking', when a researcher estimates multiple models and only reports the most favorable ones. Points five and six are self-explanatory.

The ASA statement affirms my argument above that p-values are a measurement regarding the probability distribution of a test statistic relevant to a null hypothesis. The use of a p-value outside of that setting is irresponsible and misleading. Nevertheless, the publication of insignificant, or 'null,' results is rare. This practice suggests that a significant coefficient is indicative of something important, and also indicates a belief as a field that significant results illuminate explanatory or predictive mechanisms. P-values provide no such mechanism, as will be discussed in detail below.

Breiman (2001) further illuminates general problems with the NHST approach: what he calls data modeling (as opposed to algorithmic modeling) focuses on fitting models to full samples, and assessing their inferential value based on goodness of fit and p-values. As will be discussed further this approach is flawed because goodness of fit metrics are poor determinants of model fit in high-dimensional models (Breiman 2001; Landwehr et al. 1984; Bickel et al. 2006), and goodness of fits do not serve as good adjudicators or which model is 'correct' in a set of similarly well-fit models.

Together, the faults of p-values and the limited nature of goodness-of-fit (GoF) analysis serve as two of the stronger arguments against NHST. Below, I enumerate many arguments against NHST, and expand on these.

Null and alternative hypotheses

The null hypothesis in most political science research is that an explanatory variable has no relationship with a dependent variable: $H_0 : \beta = 0$ where β represents the coefficient on a variable of interest from some model type. The alternative hypothesis in this context is almost always $H_A : \beta \neq 0$. In almost every setting it is likely not true that any explanatory variable and a dependent variable have truly 0 relationship (Levine et al. 2008, p. 177). Even if a relationship is small or mostly random it is not 0. Thus, the null hypothesis is never truly false, so the NHST framework is predicated on a strawman concept (Levine et al. 2008; Meehl 1986).

Furthermore, it is important to further discuss the implications of a rejected null hypothesis. Given a sufficiently small p-value that one feels confident rejecting H_0 then the alternative hypothesis remains. In practice, the alternative hypothesis is purported to be of whatever formulation the author's theory leads to and is built on mechanisms sourced from a theory.

Consider the timeless piece of Fearon and Laitin (2003), it is foundational in the civil war literature and has been cited 2118 times (according to the Cambridge University Press website, 6/28/2019). Their first hypothesis (p. 78) argues that more ethnically and religiously diverse countries will experience greater risk for civil war. Formally, $H_0 : P(Civil\ War|Diversity) \leq P(Civil\ War|\neg Diversity)$ $H_A : P(Civil\ War|Diversity) > P(Civil\ War|\neg Diversity)$. They use Stata and their replication materials indicate they run a simple logistic regression and estimate two-tailed z-tests to get p-values because that is the defualt in the Stata logistic command. Thus, despite their hypotheses being unidirectional, and one-tailed, the real stated null is $H_0 : P(Civil\ War|Diversity) = P(Civil\ War|\neg Diversity)$ and the alternative is $H_A : P(Civil\ War|Diversity) \neq P(Civil\ War|\neg Diversity)$. The significance of the coefficients offers only sufficient information to reject the proposed relationship given the null of no difference between diverse and not diverse settings, and does not give sufficient information to confirm the proposed alternative hypothesis. Inasmuch as one has confidence in the model's specification and lack of biased coefficients and correct standard errors, one can extend interpretation *independently of the p-values* to the coefficients, while still only concluding in terms of statistical significance that the relationship between civil war and diversity is not zero.

Despite the rigidity of p-values, their analysis still shows signs of their readiness to misinterpret coefficients based on p-values. In discussion of the per capita income coefficients they say "... is strongly significant in both a statistical and substantive sense ..."(p. 83). The notion of 'strong statistical significance' is completely misguided. A small p-value on the per capita income variable denotes that they are relatively confident in their ability to reject the null hypothesis that per capita income has no relationship with civil war onset, it does support the claim that per capita income definitely has a relationship with civil war onset. However, it is reasonable for them to interpret the coefficient, just not to conlcude that the p-value offers evidence for its truthfulness.

In sum, the NHST setup severely limits the intrinsic connection between what the null hypothesis being tested is and what the authors propose. In the seminal work of Fearon and Laitin there are two manifestations. First, they hypothesize a one-tail hypothesis, but test a two-tail hypothesis. Second, the interpretation of coefficients as strongly significant and rejecting the null hypothesis as validating the alternative hypothesis is incorrect. In a following section I discuss how $P(Data|H_0) \neq P(H_0|Data)$.

The impossibility of meeting a model's assumptions

The valid interpretation of NHST models is dependent on the correct specification of the model, which are dependent on meeting a model's assumptions. Consider simple linear regression (OLS) assumptions: (1) a model is linear and correctly specified, (2) there is not too much multicollinearity, (3) the disturbance term has zero expectation, (4) the disturbance term is homoscedastic, (5) the values of the disturbance term have independent distributions, (6) the disturbance term has a normal distribution (Dougherty 2011b, pp. 159-160). If each of these assumptions is not met then either the coefficients or the standard errors will be biased or wrong, and inference based on that model is wrong.

As a simple proof of concept for why NHST with OLS (and all of its variants) are poorly suited for the study of international relations consider a setting in which the unit of analysis is country-year and the dependent variable is trade outflow or some continuous economic indicator.

First, it is practically impossible on multiple fronts to satisfy the first assumption because it implies that all relevant variables are in the model. The only way the true model can be specified, and thus omitted variable bias is eliminated, is if there is a way to know the true model. If there is no randomness in the dependent variable at all then it is possible to have a model with perfect fit, but that leads to the second near impossibility: identifying and gathering sufficient data and accurately measuring all of the relevant variables. Even given a list of what the perfect model is, obtaining complete data that is perfectly measured in political science is impossible at the country-year unit of analysis level. Especially with economic data, many countries are unwilling to share. Furthermore, it is impossible to tell if these assumptions are met and therefore impossible to give confidence to an audience in these results.

Second, and more damning to NHST, it is not practical or possible to perfectly account for the interdependence of observations on at least three fronts: spatial, temporal, and relational. One can technically model temporal correlation, spatial correlation, and relational dynamics (via network models), but doing all three together is certainly not practical. Not only would the model be incredibly complex and difficult to interpret, but identifying that model correctly is very difficult. From a temporal standpoint, one must identify the lag structure of the dependent variable and the lag structure of the independent variables necessary to get white noise residuals. From a spatial standpoint, one must account for spatial correlation in the residuals according to a similar set of potential lag structures but in a spatial plane rather than temporal one.

Most importantly, without correctly accounting for problems with the residuals that include spatial and temporal correlation the standard errors of the coefficients are incorrect and the t or z-statistics used to calculate p-values are then incorrect. In this setting, even if evaluating hypotheses based on p-values was a great idea it is very difficult to be confident that the p-values are correct in settings of complex data structures. This is just one of the many reasons that a common critique of p-values is that they are fickle to model specification (Basuchoudhary and Bang 2018; Colaresi and Mahmood 2017). Transition toward predictive models in machine learning and focusing on adding to predictive power of models avoids these problems because they are not reliant on strict assumptions because they do not have concern with NHST.

P-values are specific and rigid

As briefly discussed above, p-values are designed and structured to offer very specific and limited information. A p-value is the probability that the following conditional probability is true: $P(data|H_0)$ (Gill 1999). In words, this means that a p-value represents *exclusively* the probability of observing the observed data, or data more extreme, given that H_0 is true. In the context of a regression, 'data' is the sample observed and the relationships within that sample, so that conditional can be rewritten to be $P(\beta|H_0)$, where β is the coefficient representing some relationship. Thus, the p-value represents the probability of observing a coefficient given that the null hypothesis (usually $H_0 : \beta = 0$): $p - value = P(\hat{\beta}|\beta = 0)$.

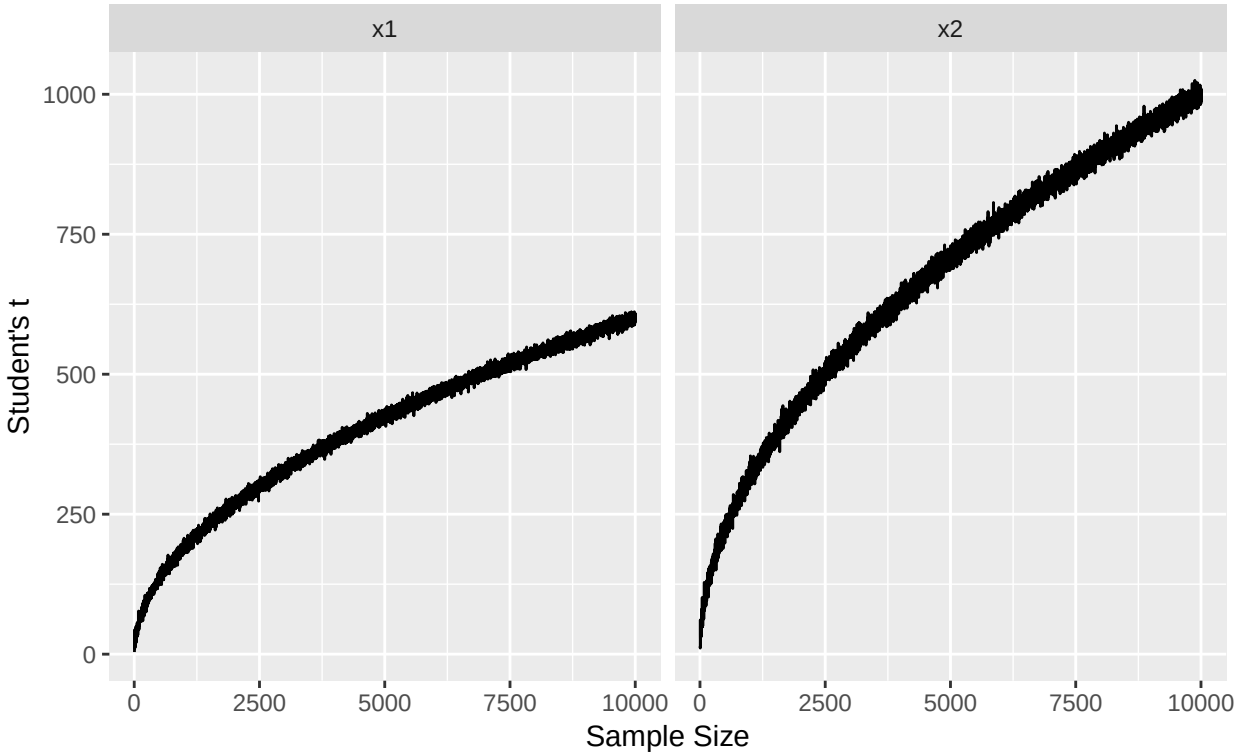
This is not interesting information relative to the conditional $P(H_0|data)$ which would be easier to infer from. Given the discussion above, this can be re-written as $P(\beta = 0|\hat{\beta})$. If the p-value provides this information then the conclusions and inferences from NHST would indeed be much more interesting. Given Bayes law, $P(A|B) = \frac{P(A)}{P(B)}P(B|A)$, it is clear that $P(H_0|data) = P(data|H_0) \iff P(data) = P(H_0)$. Written alternatively in the context of the null and alternative hypotheses: $P(\beta = 0|\hat{\beta}) = P(\hat{\beta}|\beta = 0) \iff P(\hat{\beta}) = P(\beta = 0)$. There is no way have have any information regarding the true probability of either independent probability $P(\beta = 0)$ or $P(\hat{\beta})$ and thus it is impossible to ever conclude that $P(H_0|data) = P(data|H_0)$.

P-values are dependent on sample sizes in ways that coefficients are not

Apart from understanding exactly which conditional probability a p-value represents, consider the equations that obtain the p-values. The student's t for a β coefficient in a two variable OLS equation(Dougherty 2011a): $t = \frac{\hat{\beta_1} - \beta_1^0}{\frac{S_{\hat{\beta}_1}}{\sqrt{n}}}$. The equation for $\hat{\beta}_1$ is: $\hat{\beta}_1 = \frac{\sum(X_{1i}-\bar{X}_1)(Y_i-\bar{Y})\sum(X_{2i}-\bar{X}_2)^2 - \sum(X_{2i}-\bar{X}_2)(Y_i-\bar{Y})\sum(X_{1i}-\bar{X}_1)^2}{\sum(X_{1i}-\bar{X}_1)^2\sum(X_{2i}-\bar{X}_2)^2 - [\sum(X_{1i}-\bar{X}_1)\sum(X_{2i}-\bar{X}_2)]^2}$ Immediately obvious from these two equations that makeup the student's t value is that the equation for $\hat{\beta}$ is not impacted by n, the number of observations, while the t-statistics is downwardly weighted by an increased n. So, all else equal, as $n \to \infty$ the t-statistics increases. Thus, as samples in analyses that use p-values and t-statistics (and z-scores) grow, test statistics also grow and low p-values become more common and thus less meaningful because it is purely a function of large datasets (Gill 1999; Colaresi and Mahmood 2017; Basuchoudhary and Bang 2018; Levine et al. 2008; Meehl 1986).

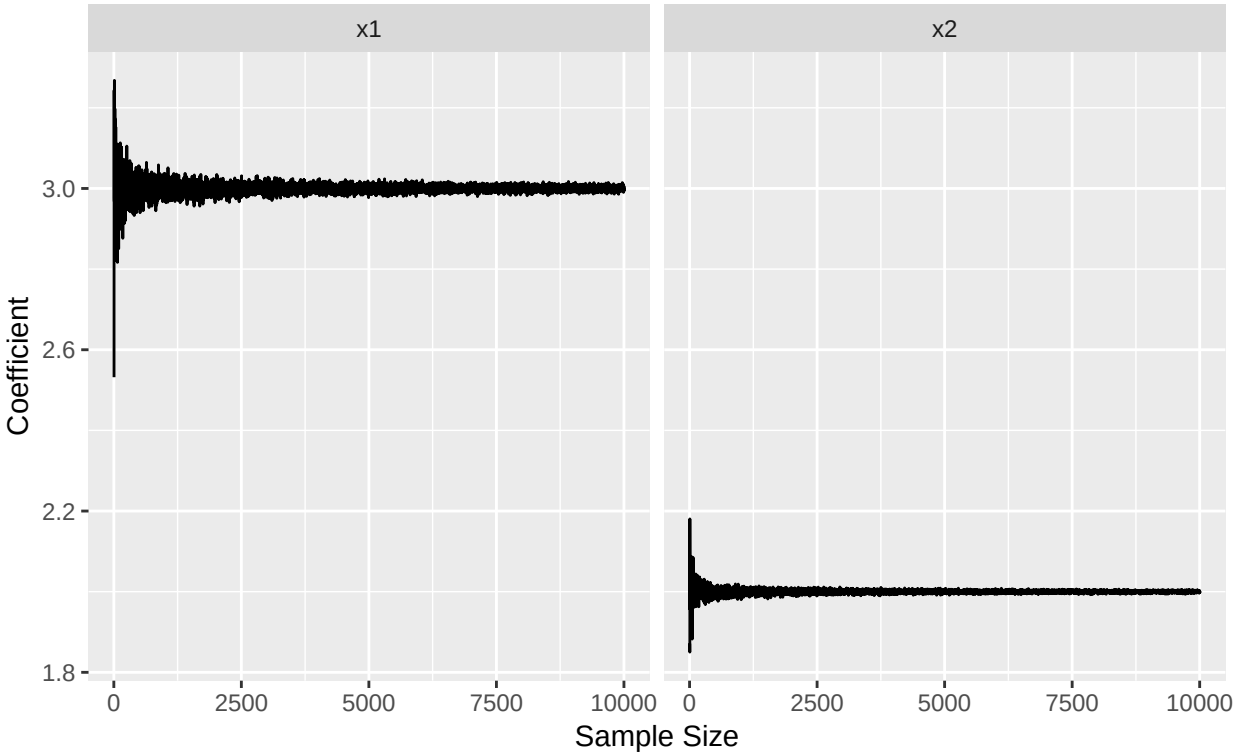
If it is easy to show mathematically that p-values are not reliable indicators of the truth of the null hypothesis with large sample sizes, then there is ample evidence to consider them a moot point in big-data analyses. For further visual and asymptotic evidence, consider the plots below generated from simulated data. Over 10,000 iterations I generated the same equation ($y = 1+3(x1)+2(x2)+\epsilon$ where $\epsilon \sim N(0,1)$) increasing the sample size by one for each iteration (starting at n = 4). After generating this data I refit the linear OLS model $Y = \alpha+\beta_1(x1)+\beta_2(x2)+\epsilon$ and plotted the results from the estimates. As you can see in Figure 1, student's t increases as n increases, with deviations because of the variance in the randomly generated data. Figure 2 shows the estimated coefficients for each iteration. These figures together show that as n increases, as long as the relationship between the independent variables and the dependent variable remain the same and the general distribution of those variables remain the same, p-values will get progressively smaller because student's t gets progressively larger.

Figure 1: T−Statistics for simple OLS as sample size increases



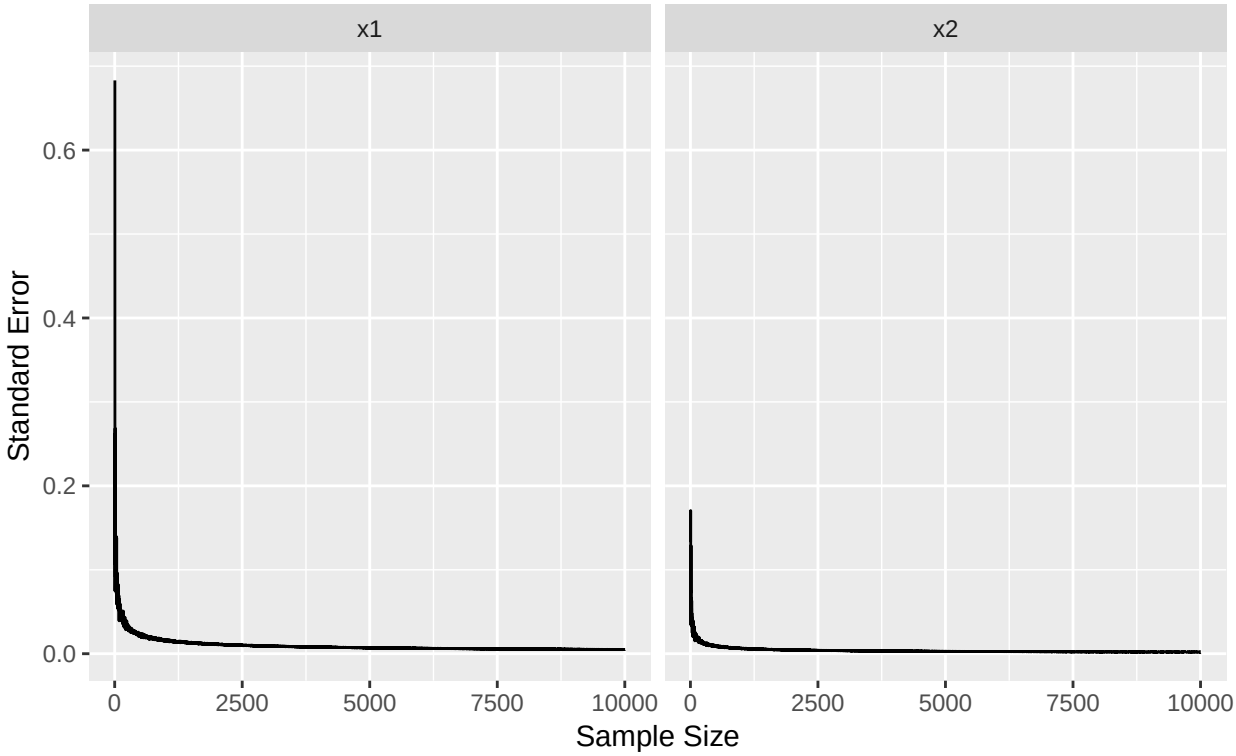
Generative equation: y = 1 + 3(x1) + 2(x2) + e. e is distributed with mean 0 and SD 1.

Figure 2: Coefficients for simple OLS as sample size increases



Generative equation: y = 1 + 3(x1) + 2(x2) + e. e is distributed with mean 0 and SD 1.

Figure 3: Standard errors for simple OLS as sample size increases



Generative equation: y = 1 + 3(x1) + 2(x2) + e. e is distributed with mean 0 and SD 1.

Dangers of Goodness of Fit tests

Goodness of fit tests (GoF) only consider how well a model fits the data, not how well it predicts the outcome (Breiman 2001). Particularly important to this section is the danger of overfitting, and how goodness of fit tests may not detect this. Beck et al. (2000) argues that focusing on GoF instead of out-of-sample prediction results in focusing on fitting models to idiosyncrasies rather than uncovering truly important structural dynamics.

GoF tests are also problematic in that in high-dimensional spaces they do not always provide reliable inferences (Breiman 2001; Landwehr et al. 1984). Furthermore, multiple different variable combinations can produce the same GoF values, thus making it impossible to determine which model is the best given the dependent variable, making it impossible to discern between theories.

Ward et al. (2010) makes a clear inferential distinction between assessing models by their GoF and their out-of-sample prediction ability: full-sample models cannot offer any validation for causal arguments becauase the models explanatory ability within sample is merely a description of the data (Beck et al. 2000). My argument here is not that predictive modeling has any causal analysis component. Rather, if one purports to understand the relationship between a set of predictors, and an outcome of interest, then considering a description of the training data (i.e. the model fit on the full sample) is not sufficient because of the potential of undetected overfitting of the model. Furthermore, GoF tests, p-values, and the coefficients themselves provide no means for discerning relative importance of predictors in NHST approaches, while some machine learning approaches do consider this.

The only way to truly avoid over fitting and adjudicate between models is to test models out-of-sample (Beck et al. 2000; Colaresi and Mahmood 2017; Breiman 2001; Ward et al. 2010; Hindman 2015). Testings out-of-sample can easily identify when a model is overfit or underfit: in both cases it will predict poorly. An underfit model (which will also be identifiable by GoF metrics) will simply not have enough information to predict well out-of-sample. An overfit model follows random movement that is atheoretical in the dependent variable, and thus inferences drawn from those models are not useful or accurate. Worse yet, it is not possible

to reliably detect overfit models without out-of-sample prediction. The best bet is to use $\bar{R}^2$, which penalizes for lack of parsimony.

Fafchamps and Labonne (2017) offer one of the most restrictive models to make out-of-sample testing a required facet of the publication process. They suggest that researchers should submit their collected data sample before analysis and a third-party splits the sample, only returning part of the sample to be used as a a training set for the project until publication. Once the project is accepted for publication the same model is then used to fit on the withheld portion of the sample to evaluate the theory's performance. This is distinct from predictive modeling because it still suggests using NHST on the test set, but the intuition of training and testing on separate datasets is similar.

In conclusion, the faults listed above of NHST cast significant doubt on its ability to meet the two fundamental goals of political science research: explanation of events, and assessment of and between theories. Strictly speaking, NHST does provide measures of explanation of variance *within the fit model*, but provides no such tool to test explanation out-of-sample. An overfit model can explain artificially high levels of variance because it tries to fit itself to randomnesss in the DV, which hampers generalizability and leads to incorrect inferences based on coefficients (Hill et al. 2013; Colaresi and Mahmood 2017). Aside from the problem of model fit and predictive performance, NHST provides no direct way to asses whether or not a model's variables actually add value to predicting the outcome of interest (Ward et al. 2010; Gelman and Loken 2014). Coefficients offer insight into the correlational relationship between independent and dependent variables, and give some indication of the effect size within sample. Within sample effect size does not translate to out-of-sample prediction, and a model that only fits its training data well while performing poorly out of sample is not effective at identifying good theoretical arguments. NHST provides no methodology for assessing predictive power of the model, and thus provide no mechanism for validating a theory's usefulness.

Given the conclusion that NHST fails to meet the needs of political scientists, I propose that whenever possible a new machine-learning based framework for out-of-sample prediction (FOOSP) supplants NHST. This framework meets the needs of political scientists for assessing theoretical validity and is actually simpler in understanding and implimentation than NHST because of it's weaker focus on standard errors and p-values and stronger focus on predictive accuracy. Ultimately, theories that predict poorly, or add no value to existing theoretically derived models, are bad theories. FOOSP provides a framework to more effectively identify predictive versus nonpredictive theories allowing the field in general to move forward in a more Lakatosian direction, and away from degenerative theory testing under NHST.

Next, I discuss examples of work already moving in this direction and utilizing machine learning of all kinds on political science and other social science problems. Then, I explain my framework for analysis moving forward.

Machine Learning in Political Science: examples of FOOSP in practice

Two examples that are the best representations of the framework that I am proposing the field switches to are Blair and Sambanis (2019) and Ward et al. (2010). Blair and Sambanis (2019) tries two alternative approaches to predicting civil war escalation: one which treats theory as obsolete and irrelevant, thus considering all possible predictors, another which develops a theory based predictive model and they pit the models' predictive ability against each other as a way to discern whether or not theoretical foundations improve prediction. They conclude that the theoretically informed model does improve prediction, indicating that the era of big data and machine learning does not signal the end of theorizing.

Ward et al. (2010) argue as I have that p-values are not useful for evaluating theory and they consider two popular models of civil war onset, Fearon and Laitin (2003) and Collier and Hoeffler (2004), in order to discern which theory predicts civil war onset better. They find that despite both papers validating their theories via p-values and the NHST approach, Collier and Hoeffler (2004) proved to be a much better predictive model.

Basuchoudhary and Bang (2018) does a short literature review of the NHST literature on terrorism at the country-year level and estimates predictive models using all variables to identify how good the existing theories predict terrorism. Their starting point is similar to mine: NHST is an untenable approach because of the lofty assumptions that p-values depend on for correct interpretation. Moreover, they offer some evidence that machine learning approaches need not be black boxes by identifying the top drivers of terrorism occurences at the country-year level: assissinations, guerilla war, military personnel, religious fractionaliza-

tion, and military politics. Moreover, they use partial dependence plots to show the nonlinear relationship these variables have with terrorism.

Gleditsch and Ward (2013) uses a logit model to predict interestate conflict conditional on the type of Issue Correlates of War (ICOW) disputes that exist. Their approach is markedly less explicit in interpretation than Basuchoudhary and Bang (2018) because Gleditsch and Ward (2013) simply test predictive accuracy between entire models rather than assessing feature importance within a model. Nevertheless, this approach does consider the model, and thus theory's, ability to predict out of sample which they also argue is central to validating theory given the faults of p-values.

Blair et al. (2014) uses surveys across two different waves in more than 240 towns in Liberia to predict local violence in those towns. They use survey and other collected data from 2008 to predict violence in 2010 (the next survey year) and use the 2010 data to predict 2012. The authors find that once prediction is considered as the heuristic by which models are evaluated their results do not necessary line up with the NHST literature. Ethnic heterogeneity remains a good predictor of violence, but economic shocks do not. They also find that local violence is more likely with minority-majority ethnic power sharing agreements, which is not consistent with the NHST literature.

Schutte (2017) use poisson point process models to predict the Geo-Referenced Event Dataset (GED) (Sundberg and Melander 2013) violence. They use spatial models to try to predict where poltiical violence occurs and use dirichilet tesselation to randomly grid data based on observed locations of violence. For out-of-sample prediction they use a very computationally complex approach: leave-one-out cross-validation. This approach leaves one observation out of the training data to fit the model, and then predicts the outcome on that left out observation using the trained model. Then, iterature across all observations and aggregate your predictive measure based on each prediction and iteration. In a spatial context, this is a good approach because it can be difficult to sensibly split a sample into a test and training sample because this will break the fundamental structural characteristic of space in a way that biases the model. Using leave-one-out cross-validation is beneficial becaus it minimizes the lost information on each training iteration.

Witmer et al. (2017) have perhaps the most ambitious example of out-of-sample forecasting attempts. They used the ACLED data in 1-degree grid cells and simulated future climate data to predict subnational violence all the way into 2065. Obviously, this sort of work is impossible to validate, but it is a good example of researchers focussing on prediction as a heuristic. It would be ideal if the authors also included some out-of-sample prediction that could have its accuracy measured, because predicting 40-plus years into the future is only helpful if there is some semblence of the model's accruacy in predicting a context we already know the outcome of.

Hill et al. (2013) use random forests to predict terrorist group culpability of attacks in the Philippines (where previously multinomial logit and naive bayes were used). They use cross-validation via out-of-bag prediction to guard against overfitting of the model and find that Random forests predict as well or better than the others as long as there is no class imbalance problem. This paper serves as a great example where political

Ding et al. (2017) use a variety of geo-coded data to predict terrorist events around the world in 2015. They achieve about a 96% accuracy rate using .1x.1 (latitude x longitude) grids around the world. They used neural networks, support vector machines, and random forests. Unfortunately, this model is a good example of when predictive modeling can be uninteresting because their primary predictor is a history of terrorism and their goal is to maximize predictive accuracy. They also include variables such as drug areas, and geocoded ethnic composition. However, their goal and their paper is wholly divorced from theory, and only concerned with predictive accuracy. While helpful in a sense that it can identify when and where terrorism occurs, it does not help understand why, which is the central goal of science.

Mueller and Rauh (2018) use Latent Dirichlet Allocation (LDA) models to asses topics in newspaper articles to predict the onset of civil war. They use LASSO machine learning techniques to predict the onset of civil conflict using the results of unsupervised machine learning. Specifically, they find that increased media reporting on conflict is a good predictor of domestic conflict onset. This is a great example of the multifaceted utility of machine learning in general, using it both as a modeling and theory-evaluating tool and as a coding tool via text analysis. However, this is also an example of where the focus on good prediction outweighs the focus on meaningful theory: it is obvious that when media reports on conflic then conflict is likely to occur. It is not common for there to be no precursor conflict leading up to a large-scale conflict, and so this paper basically identifies precursor conflicts, and says conflict begets conflict, which is not that

interesting or novel.

Some proposals in the literature to alter the framework for analysis

There exist two important proposals in the literature that I think are critical to highlight as proposing alternative frameworks for analysis. One by Gross (2015) which proposes a shift to a focus on hypothesized substantive effects and a focus on confidence intervals and another by Colaresi and Mahmood (2017) which is very similar but less demanding than mine.

Gross (2015) approaches the problem of p-value dependency by proposing a shift away from NHST in a relatively minor way. His proposal involves roughly the same empirical approach and modeling, but changes what the null and alternative hypothesis say. Instead of the null hypothesis being zero relationship, and the alternative being a nonzero relationship, he proposes that the field shifts to substantively interesting hypotheses. An alternative hypothesis should identify a minimum threshold at which a coefficient becomes substantively interesting and only considers the null hypothesis rejected if the coefficient lands within the substantively interesting alternative hypothesis range. For example, consider a hypothesis about the effect of years of education on income. The NHST H_0 is that education has zero relationship, and a sufficiently small p-value provides evidence against the null hypothesis. Under Gross's method the researcher's focus would be on validating the alternative hypothesis by identifying a range of substantively interesting values. For example, I expect each additional year of education to increase income by at least 2,000 USD a year. In this case H_0 is that the effect of education is anything less than 2,000 USD.

Under this approach the resaearcher considers the confidence interval of the coefficient estimate in relation to the null hypothesis. If the range of the null is included within the confidence interval, then it is not possible to conclude with confidence that the null hypothesis is not true, thus providing no evidence for the hypothesis. This is not incredibly different from p-values in that it still depends on standard errors to generate confidence intervals, and these are still asymptotically decreasing with n. However, it does put more focus on identifying meaningful effect sizes and more specific hypotheses. This method does not escape the fundamental problem of econometric modeling that is satisfying impossible assumptions. I have more confidence in conclusions drawn in this framework than I do in conclusions drawn from p-values alone, but only marginally.

Colaresi and Mahmood (2017) propose a version of Box's loop: build, compute, critique, and think. These steps take place in order on loop. The focus of the framework is to shift focus from models dependent on asumptions of well fit model toward validating theory by examing predictive accuracy in a modeling process that builds on previous models in an iterative fashion that improves prediction. This is similar enough to what I propose below that I will defer discussion to below, but it is important to note that Colaresi and Mahmood (2017) is far from the first to suggest the use of machine learning in political science as a paradigmatic shift in focus (Beck et al. 2000).

Framework for out-of-sample prediction

1. Theorize and hypothesize about some phenomena.

2. Identify some a priori base model to which predictive accuracy will be compared.
3. Identify and collect relevant data. Split this data into a training set and a test set.
4. Fit a model(s) best suited for this data to a training set first without the new variables identified in step one, then with. Where practical and applicable, apply cross-validation strategies within the training set.
5. Use the resulting models from step four to predict the values of the test set. Evaluate the accuracy of the test set using appropriate predictive accuracy assessment tools (e.g. point metrics (root mean squared error, mean absolute error) or forecast density evaluation).
6. Compare the base model's predictive accuracy with the new model's predictive accuracy. Based on the change in predictive accuracy, infer about the validity of the proposed theory.

Step one in this process is the same as NHST: identify some explanation of a phenomena that is interesting to political science that can be evaluated empirically. Step two is where NHST and FOOSP diverge: under FOOSP it is critical that a base model is identified. Ideally, this base model is the best predicting model

the literature has to offer to date. In practice, this model is very difficult to identify given the variance in methods, samples, and concept operationalization. Thus, evidence in this context should only be interpreted inasmuch as the sample is generalizable. Ideally, one can identify the broadest sample, and the most general model that the literature has some moderate level of agreement on. Over time, this framework will be able to parse down and remove variables that add no predictive value. Furthermore, methods such as penalized regression models (e.g. ridge regression, and LASSO) or random forests can be used to evaluate variable importance for prediction. Variables that are not important for prediction, should not be include in favor of parsimony and avoiding overfitting.

Step three is also similar in most ways to NHST, with the caveat of splitting the sample. A good rule of thumb is to keep anywhere from 33% (Colaresi and Mahmood 2017) to 10% of the data as a test set and use the rest to train the model. Within the training set, cross-validate where possible and practical. Cross-validation is the process in which a training set is split into k parts, and a model is fit on k-1 parts, while the remaining part of the training set is used to evaluate; this process iteratures within the training set until each remaining k has been used as an evaluation set (Colaresi and Mahmood 2017; Stone 1974; Breiman 2001; James et al. 2017, p 176). This process iteratively improves models by identifying flaws on each iteration, and generating a composite model after each iteration. This is generally referred to as k-fold cross-validation, and it is difficult to implement in contexts where spatio-temporal dynamics are part of the modeling process because the way you split the data fundamentally affects the modeling process.

Other methods of cross-validating data are bootstrapping and bagging. Bootstrapping samples with replacement from the existing training set and estimates a model on this new bootstrapped training set. Bagging is best for high variance data, and its process involves bootstrapping several samples from the training data and averaging the resultant models' predictions to account for high variation (James et al. 2017, pp. 316-317). This is a common approach in decision-trees.

Step four is simple enough: just fit the two models: base and updated model. If there are multiple hypotheses to be tested, perhaps multiple updated models should be estimated, carefully considering when to add more than one variable at a time.

Step five is where FOOSP shines as a staunch improvement over NHST. Simply take the cross-validated models and test their predictive accuracy on the left out test set. There are a variety of methods this can take, the simplest of which was described above: leaving aside about 1/3rd of the data for a test set. Other methods include leave-one-out cross-validation (LOOCV); during step three fit the model to the entire sample minus one, and test prediction on the left-out observation, repeat this process for all observations and average the mean-squared error across all iterations (James et al. 2017). This is obviously computationally intensive, but provides extensive cross-validation to learn and improve models from. In the cases of bootstrapping and bagging, it is normal to test on the training set, or 'out-of-bag.'

Step six simply involves evaluating the information retreived from the previous five steps. In an ideal case one finds that adding the variables measured for the theory and hypothesis have substantively significantly improved the model's predictive accuracy. In this setting statistical significance is *never* part of the discussion because standard errors and residuals in the GoF sense are completely irrelevant to the central concerns of FOOSP. Below, I elaborate on specific methods that can and should be used such as ridge regression, LASSO, k-nearest neighbors (KNN), and Random Forests. Each of these models, as well as standard OLS and other generalized linear models (e.g. logistic regression) and much more also fit nicely into this general framework. Also below, I elaborate further on exactly why freedom from the tyranny of GoF, standard errors, and p-values makes science easier. In the final dissertation even more alternatives will be described and each model will be demonstrated on simulated data in comparison to NHST methods.

The tyranny of residuals in NHST and FOOSP's liberation

Simply put, the foundation of NHST as discussed above is that models meet their strict assumptions. Under these met strict assumptions standard errors are correct and p-values can be calculated and null hypotheses can be rejected, but alternative hypotheses cannot be confirmed. NHST also does not allow the examination of predictive accuracy out-of-sample. NHST provides almost no value for scientific advancement of theories, despite science's best effort for decades to extract value.

FOOSP, alternatively, does not have any concern, or even discussion of standard errors, or p-values, and thus does not have to meet those strict assumptions. Predictive models either predict well out of sample,

or they do not. Theories, hypotheses, and variables either add predicted value, or they do not. Whereas under NHST not properly modeling temporal or spatial dynamics can result in a virtually uninterpretable model from an inferential perspective, predictive-based modeling is largely unaffected. If space and time are important, and unmodelled, then the model predicts poorly, but it still predicts. Moreover, the predictions that it makes, while predicting poorly, are still interpretable even if they model performs poorly.

Thus, whereas under NHST researchers can fail to model important correlation in the residuals and fallaciously interpret the results when they contain little truth, FOOSP directly punishes for poorly modeling real dynamics by predicting poorly. In NHST, this factor is never even considered, and many journals and referees do not ask for spatio-temporal models on average.

Bias-variance trade-off

One theme with machine learning models and a general shift to FOOSP that must be discussed before moving forward is the bias-variance trade-off of predictive machine learning methods. Assume that there is some true model predicting Y, $Y = f(X)$. In trying to uncover this relationship, I estimate the relationship $\hat{Y} = \hat{f(x)} + \epsilon$ on a training set and intend to test this estimate relationship on a withheld test set as described above. Ideally, the best model at predicting Y assuming $f(x)$ is linear will minimize mean squared error (MSE) $\frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n}$ also written $\frac{\sum_{i=1}^{n}(y_i - \hat{f(x_i)})^2}{n}$ (James et al. 2017, p. 29). In the test set, the expected value of MSE is $E(y_0 - \hat{f(x_0)})^2 = Var(\hat{f(x_0)}) + Bias(\hat{f(x_0)})^2 + Var(\epsilon)$ where y_0 and x_0 denote observations from a previously unseen test set. $Var(\hat{f(x_0)})$ is the variance of the specified model if it were estimated using a different training set, and $Bias(\hat{f(x_0)})$ is the error "introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model (James et al. 2017, pp. 34-35)", or how wrong it is in the test set.

The bias-variance trade-off rewards parsimonious models with lower variance but punishes with higher bias on average, and it punishes complicated models with higher variance but lower bias. Intuitively, this mimics real world processes that political scientists try to model: a simple description of events likely does not predict when that event will occur well, but it will also not predict a large variety of conditions where it may occur. Alternatively, a complex description may predict many cases where the event may occur, and many of them will be correct. An ideal description is as simple as possible while predicting as correctly as possible, i.e. having low model variance and low model bias. Machine learning models, and out-of-sample predictive accuracy heuristics, are constantly combating this concept as a central tenet of what makes a model good or bad. The researchers must decide if they prefer variance of bias when both cannot be minimized. Some methods discused below are optimized to be lower in either bias or variance, but not always both.

Explanation of models

Penalized Linear Regression

Penalized linear regression is one of the simplest and most easily executable alternatives to standard OLS regression. Penalized regression models optimize for prediction and for parsimony. OLS is incapable of producing a parsimonious model because all coefficients will be nonzero (Hindman 2015, p. 56). This can be particularly problematic in the context of modeling nonlinear or conditional relationships with linear model via quadratic/polynomial terms and interaction effects. Also, OLS is a poorer predictor in comparison to penalized regression models. this is because the model variance of OLS is relatively high, and thus dependent on the training data. A change in the training data can have a large impact on the prediction resulting from the model (James et al. 2017, p. 218). There are two common alternatives to OLS, ridge regression and least absolute shrinkage and selection operators (LASSO). Both are linear models that operate under similar principals.

When one estimates OLS the resultant model minimizes $RSS = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$. Ridge regression and LASSO regression operate in a similar framework that minimizes RSS in addition to a tuning (shrinkage) parameter and some linear combination of the resultant coefficients. For ridge regression the value being minimized is $RSS + \lambda\sum_{j=1}^{p}\beta_j^2$ where p is the number of parameters in the model. LASSO minimizes $RSS + \lambda\sum_{j=1}^{p}|\beta_j|$. In each value λ is a tuning parameter set by the user to increase parsimony as its distance from zero increases. The tuning parameter is positive and ranges from zero to infinity. This tuning

parameter should be selected via cross-validation methods and minimizing prediction error in a test set. Note that when the tuning parameter is set to zero the result is identical to OLS.

The major difference between ridge regression and LASSO approaches is that LASSO will set coefficients on variables that are not very helpful in prediction to zero, whereas ridge regression is less parsimonious because its regularization method in the value to be minimized cannot set coefficients to zero, only very, very small. Both approaches have higher variance than OLS, because they will penalize the addition of variables that do not add much predicted value on the assumption that they are likely only predicting noise.

Solving LASSO and ridge regression is notably more complex than OLS. One approach advocated for by Hindman (2015) is least angle regression (LARS) algorithm, which iteratively increases the coefficient of the most-correlated variable from zero in the direction of the correlation until moving the coefficient away from zero does not improve model fit in a way that outways its added penalty.

This approach, and these types of regression generally, do not rely on hypothesis testing and do not produce standard errors because their goals are not hypothesis testing. However, these models are very useful for identifying which variables are still important for predicting outcomes. First, if a variable's coefficient is nonzero in LASSO, or sufficiently far from zero that it is interesting in ridge regression, then a researcher can conlude that these variables are useful for predicting the outcome. In the context of testing new theoretical relationships, this can be helpful because an added variable is helpful in predicting empirically if and only if its coefficient is nonzero, or not close to zero. In other words, if a variable's added value in predicting the outcome (i.e. its reduction of RSS) is greater than its added penalty then it is theoretically relevant and useful. This approach can also be used to discern between a large set of variables to decide which are and are not relevant predictors. Furthermore, the shift toward prediction-based model evaluation merits a shift toward models that predict better. Penalized regression approaches generally predict better than OLS because of their favorability of parsimony (Hindman 2015). This decreases training accuracy but improves test accuracy by guarding against overfitting.

K-nearest neighbors

K-nearest neighbors (KNN) is a very simple classification method that thrives in identifying non-linear effects. It is quite similar to matching in some senses. Given a sample and a categorical outcome variable, k-nearest neighbors uses Bayes rule to classify an observations based on the K other most similar observations. Given K (the number of comparison observations to select) and a test observation x_0, KNN selects K closest points in a training set (this set of K points is denoted N_0). Then, it estimates the conditional probability that x_0 belongs category j via $P(Y = j|X = x_0) = \frac{1}{k} \sum_{i \in N_0} I(y_i = j)$ (James et al. 2017, p. 39).

This approach can adjudicate whether or not a variable and its related theory is useful for improving prediction by estimating a KNN model both with and without the new variable. If the model with the new variable improves prediction (either increases true positives/negatives or decreases false positives/negatives) then an argument can be made that this variable improves prediction and thus understanding of the outcome. Admittedly, this is a very 'black box' approach because it does not offer any output information on direction or effect size, but it is also not dependent on meeting virtually any assumptions unlike NHST.

KNN is sensitive to the selection of K. As K increaes, the model becomes less flexible and varaince decreases thus typically increasing bias. The benefit of KNN is that it is nonparametric and nonlinear, thus allowing one to uncover a good predictive model without having to consider the shape of the relationship between those variables. Iferentially, though, this is a definite disadvantage.

Tree methods

Tree methods are particularly useful for political science research because they do not make parametrics assumptions about the form of the relationship between the features of the model and the outcome variable. Moreover, they can be used on a wide variety of outcome variables: continuous, binary, ordinal, or categorical. The simplest form of tree modelling takes a top-down approach (James et al. 2017). This approach starts with the entire dataset in one node and selects the variable that best partitions the data according to the RSS for regression trees or the Gini coefficient (for binary outcomes) first, and then divides the dataset into nodes depending on values of that data. Then, at each new node the variable that most evenly divides the dataset is again selected and the data is again partitioned into new nodes. This approach repeats itself partitioning the outcome variable into smaller and smaller groups in order to minimize the measure of predictive accuracy

conditional on values of the predictors in the training set. Then, this tree can be validated on a test set. Simple tress like this increase in rigidity and inflexibility (thus decreasing variance) as the number of features increase, but decrease bias in the training data but increasing bias in the test set because of overfitting the data.

Several approaches exist to avoid overfitting the data. Bagging is a method in which a sample is bootstrapped from the training set, and each a variable is selected at random from the set of available variables for each node. The resulting tree is then tested out-of-bag (on the test set) and through iteration the results of many trees are averaged for the final prediction of the model based on bagged trees.

Another more common approach to political science is random forests. This starts with a random sample of the training set, and instead of considering the full number of predictor at each node, each node of each tree considers a random sample of size $\sqrt{p}$ predictors where p is the number of predictors available. The researcher specifies the number of iterations and trees to estimate, and the results from these trees are averaged for each observation as the final prediction of the model. Restricting the potential set of predictors at each node prevents the correlation of trees within the 'forest' because not all nodes in all trees have the same set of predictors to choose from (James et al. 2017).

Summary of FOOSP

In sum, FOOSP offers an alternative to scientific research in political science guided by NHST. Its foundation is guided by the validation principle of out-of-sample prediction. FOOSP uses machine learning principles to shift the focus of political science research away from standard errors and p-values that are central to NHST and toward predictive accuracy. Good theory improves predictive accuracy, and the current state of research in political science does not reflect the importance of predictive accuracy with few exceptions. The methods discussed above are a small subset of the available machine learning approaches outside of the common models to political science. However, it is important to note that almost every model (e.g. OLS, logistic regression, time series models, spatial models, survival models) common to political science research can benefit from two central components of FOOSP: cross-validation and out-of-sample prediction.

Cross-validation allows researchers to fit models on a variety of samples from a given dataset and avoid a overfitting models by identifying when one model is fitting to primarily noise. Overfit models pose particular threat to theoretical advancement because they are not generalizable and in NHST they are almost impossible to identify. Out-of-sample prediction focuses on the effectiveness of the model as a whole to predict when the outcomes of interest will occur. This is then used as a metric to evaluate the model in its entirety. A hypothesis is considered useful if the variable related to the hypothesis improves prediction relative to a base model that excludes that variable. This shift requires no consideration of standard errors and p-values, only a consideration how well the model actually predicts the outcome it was built to predict.

It is important to note that in theory building it is necessary to identify and estimate a model that predicts *well*, it is however crucial to factors that *improve prediction*. In other words, if the best theoretically informed model of terrorism in some context can only predict correctly 40% of the time, then that metric itself is not enough to criticize the theory. Rather, there must always be a base model established for a point of comparison to decide which theories add value to predictive modeling. The dissertation itself will evolve this idea in conjunction with how added value alone must be considered in the context of loss parsimony, and whether or not an added variable adds enough value to merit its cost to parsimony. This is because cross-validation and out-of-sample prediction are not completely impervious to overfitting, although they are much more resilient than NHST.

Chapters of the dissertation

Each chapter of the dissertation will build on the ideas laid out in this proposal and show how in practice there are real world differences. The first chapter will be mostly what is presented here, but also including monte carlo analysis of a variety of data problems that political science faces. The first chapter will also use monte carlo analysis to show how some models predict better than others, and how this difference in predictive accuracy is important if political science wants to help provide policy prescriptions. The second chapter will look at existing examples of foundational findings according to NHST approaches and show how NHST makes it almost impossible to adjudicate between competing explanations of the same outcome. The

goal of this section is to demonstrate using real world data and already-published work that the difference in inferences being drawn from NHST and FOOSP are substantively important, and more importantly that NHST is misleading and counterproductive. The third chapter will present a novel project that implements the FOOSP framework from beginning to end. Specifically, I will use the case of Colombia and assess the validity of the terrorism literature's theories on predicting terrorism locally and across time. Below, I give a little more detail on each of these chapters.

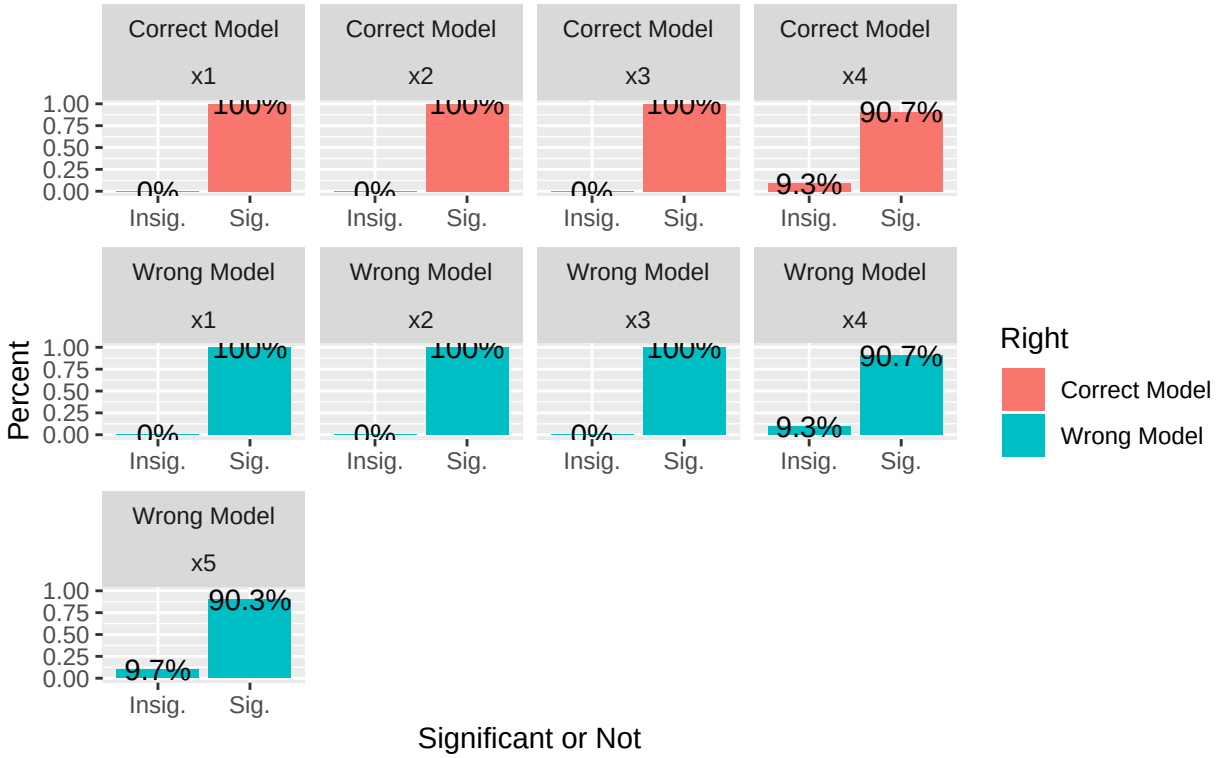
Chapter 1: NHST v. FOOSP

In addition to what is above, monte carlo simulations of the problems identified with NHST will be fronot and center in this chapter. Specifically, there will be three things to demonstrate with simulated data: asymptotic properties of standard errors and p-values and their sensitivity to correct identification, the flaws of goodness of fit tests, and the benefits of cross-validation on model prediction.

First, I will show as I have above that p-values decrease with sample size, so the larger the dataset the more likely one is to encounter type I errors, and thus draw false inferences. This will be demonstrated to be true even in relatively small samples (n = 10,000) that are very common in international relations literature. There is an example of what will be shown below: in standard OLS, even if a variable is specifically specified to have zero relationship with Y, or it is not part of the generative equation of Y, there is still an absurdly high rate of statistical significance when the effect is truly nothing.

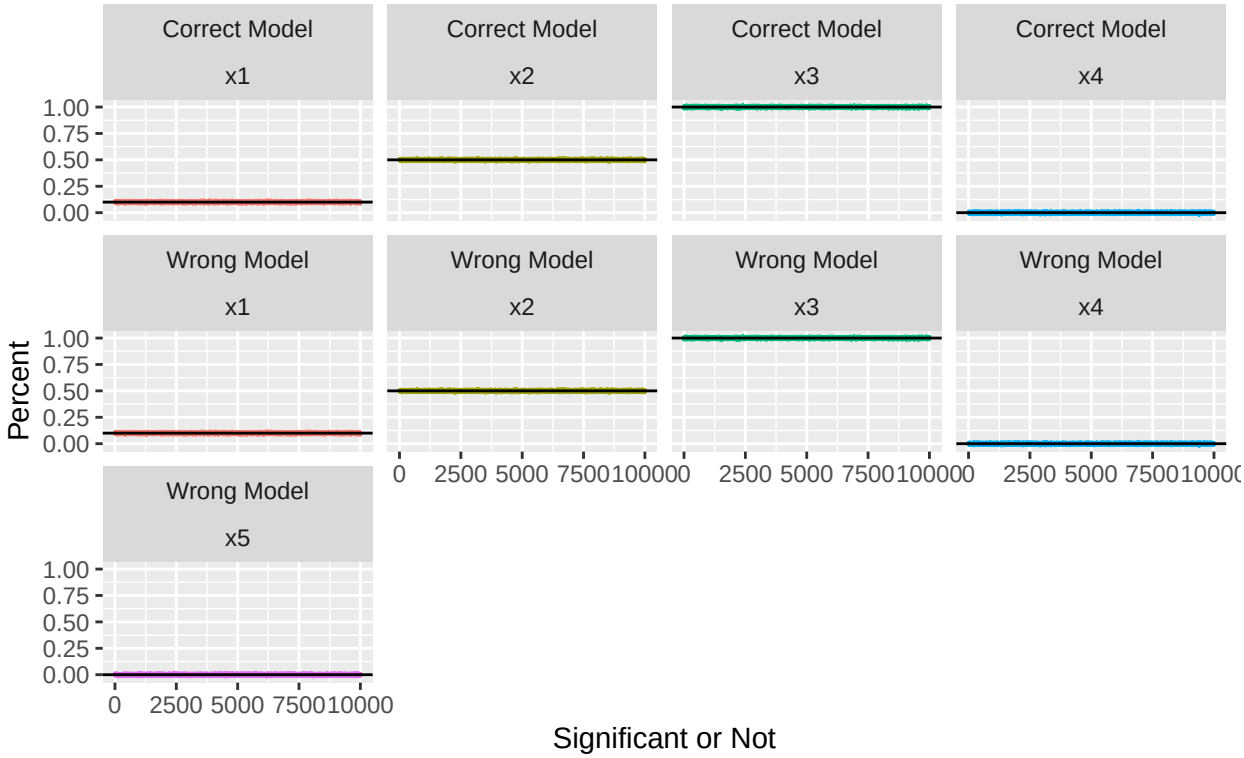
Figure 4 below shows that even in a sample size of only 10,000 across 10,000 iterations variables that are specifically specified to have zero relationship with Y (x4) or that are just not part of the generative equation (x5) are still significant about 90% of the time. The problem here is that there comes a point when it really does not matter how small the coefficient is, because the standard errors asymptotically decrease while the true relationship between an x and y are independent of sample size. As you can see in Figure 5, the estimation of the coefficients is still pretty accurate across each iteration even when including irrelevant variables, but irrelevant variables are statisticall significant. This emphasizes the need for a shift toward at the very minimum substantive interpretation of variables, rather than using p-values as a distinction for importance.

ɟure 4: Percent of significant coefficients for OLS across 10,000 iterations



ation: y = 1 + .1(x1) + .5(x2) + 1(x3) + 0(x4) + e. e ~ N(0,1). All x ~ N(x#,2). N = 10,000 for each model.

Figure 5: Coefficient estimates for OLS across 10,000 iterations



+ .5(x2) + 1(x3) + 0(x4) + e. e ~ N(0,1). All x ~ N(x#,2). N = 10,000 for each model. Horizontal black line represents true coefficient.

After discussing the properties of p-values and their faults in general, I will move on to demonstrate the importance of modeling the dynamics of the data when it comes to drawing correct inferences. Specifically, I will show that in the context of time series data, spatial data, and spatio-temporal data not correctly modeling these relationships results in plainly incorrect standard errors and thus unreliable hypothesis testing and inferences. Currently, I need to learn more about how to structure these simulations in order to avoid creating straw-man simulations. Ultimately, my goal is to show that unless you model the spatio-temporal dynamics perfectly, your models are unreliable and in practice it is virtually impossible to model them correctly. Importantly, in the FOOSP framework the lack of reliance on standard errors means that not modelling these dynamics correctly results simply in poorer prediction, but not invalid models.

Further contributions of this chapter will be a deeper discussion of alternative machine learning models that are not currently used in political science and a shift and a deeper discussion of cross-validation strategies in complex data structures such as time-series cross-sectional data.

Chapter 2: Different inferences from NHST and FOOSP

Similar to some of the work done by Blair and Sambanis (2019) and Colaresi and Mahmood (2017), this chapter will show that a shift in analytic framework will result in different inferences. I will identify a small number of the most cited papers in international relations pertaining to civil wars, interstate wars, MID onset, and repression in recent years and assess whether or not their theoretical contribution to these topics merits the attention that it has gotten. If these papers do not add to the predictive accuracy in a way that outweighs its cost to parsimony, then that is grounds for reconsidering their place in the development of theory.

The goal of this section is to show current strategies of testing theories are prone to all of the problems identifiedin chapter one, and once considering this problems and adjusting the theory evaluation metric different inferences will be drawn. Furthermore, if a paper tested in this section withstands the shift in analytic framework then FOOSP will provide even more evidence in favor of the theory by demonstrating that it is not an artifact of overfitting the data and that it improves out-of-sample prediction. This section

provides evidence that FOOSP not only evaluates theories better for the sake of political scientists improving their own theories, but it also makes the deliverable to policymakers better because evidence can be provided for out-of-sample prediction. Alternatively, this can also identify areas where political science is poor at predicting events, and thus perhaps policymakers should be cautious heading the advice of political scientists.

It is important once again to note in this section that a theory under the FOOSP framework *must not predict well in an absolute sense* rather it must *improve prediction* from the base model. This dissertation and this chapter is not an indictment of political science as a field of poorly-predicting theory, rather it is a push in the direction of focusing on improving prediction.

Chapter 3: FOOSP start to finish

This chapter will use a novel research project that I have already begun the data collection for that will try to predict the occurence of terrorism in Colombia at the most disaggregated level of analysis possible. This chapter has two central purposes and takeaways: first, it shows an implementation of FOOSP in a project as an example of what this framework looks like in action, and second, it serves as a contribution to the terrorism literature in proposing that existing theories must disaggregate their analysis in order to better understand terrorism.

Most terrorism research that exists theorizes about country-year level factors despite the fact that terrorism is primarily a local occurrence and event Krueger (2007) it is studied at the country-year level. Moreover, the vast majority of terrorism is domestic, and there is dramatic variance within a year and within a country on when and where terrorism occurs Enders et al. (2011). Despite the fact that the country-year unit of analysis is completely inappropriate for terrorism analysis nearly all of the research on this topic takes place at this level, and the most universal machine learning project on this topic also takes place at this level (Basuchoudhary and Bang 2018).

The result of this level of anlaysis problem is a poor understanding of when and where terrorism will occur which results in poor policy prescriptions. The goal of this chapter is to collect all possible variables at the lowest level of spatial and temporal aggregation possible in Colombia, group these variables into theories that exist in the literature, and iteratively build the best possible predictive model to adjudicate which branches of theory in the terrorism literature are helpful for predicting terrorism. Extreme bounds analysis has been used in some cases within the NHST framework to determine which variabels are robustly significant and which are not (Hegre and Sambanis 2006). My approach here is similar in that its goal is to identify which components of existing theory are productive in improving prediction, but I do not try all possible permutations of variables.

Two approaches I will impliment that are useful for feature selection will be penalized regression and random forests. Both approaches offer tools that allow me to discern which features of the model are most important for prediction, and in the case of LASSO I can even gain some evidence that some features have no effect relative to their cost in parsimony.

Conclusion

There are two central themes to this dissertation: p-value based hypothesis testing is not reliable, and NHST provides no tools to adjudicate between theories' ability to predict outcomes. A central component of these arguments is the assumption that political scientists care about having good theoretical explanations of why an event occurs, and that these explanations can be used to predict when and where an event will occur.

P-values are dependent on satisfying myriad assumptions of whatever model is being fit. In even the simplest linear context, it is nearly impossible to correctly identify a model and produce standard errors that are accurate and thus it is nearly impossible to produce p-values that are reliable. But, given a scenario where there is 100% confidence in p-values, the information the provide is limited: $P(data|H_0)$. That is, the likelihood of the observed relationship given the null hypothesis is true. This provides no explicit evidenced for theorized alternative hypotheses, only evidence against a non-relationship. Furthermore it is almost never true that two variables have truy zero relationship to begin with, so the null hypothesis is a straw man.

NHST is completely unconcerned with out-of-sample prediction, and thus in-sample explanation of variance is the best tool reesarchers have to make claims in this framework about how well a model predicts. In truth, in sample explanation of variance is heavily dependent on model specification and prone to overfitting.

Overfitting is particularly problematic because it hinders generalizability, and it is undetectable without out-of-sample prediction testing. Moreover, goodness of fit tests are not sufficient to adjudicate between models because several combinations of variables can provide the same goodness of fit statistic while providing no evidence for or against any variable's added value in prediction. Given a central goal is to demonstrate which theories have more evidence than others, a method that provides no tools to identify which theory does and does not predict better than a base model provides no tools to test a theory's real-world applicability.

In order to improve theory-building in political science I propose a shift in analytic frameworks toward a focus on out-of-sample prediction. If a hypothesis' proposed relationship is true then it should improve out-of-sample prediction because it aids in understanding why, when, and where an event will occur. Thus, in order to substantiate a new hypothesis one must show that it adds predictive value out-of-sample, because in sample prediction is prone to overfitting. This requires identifying a base model that includes all relevant variables in the literature, adding the new variable of interest, and assessing the added predictive accuracy. If predictive accuracy is improved, this is some evidence that the hypothesis is true.

Ultimately, this framework provides evidence baseed in the phenomena that are being studied, and it absolves researchers of crimes against standard errors. This approach is in general simpler and easier to implement than NHST because standard errors are not part of the consideration. The consequences of not modeling important dynamics is a poor predictor, but not a completely defunct model.

References

Atin Basuchoudhary and James T. Bang. Predicting Terrorism with Machine Learning: Lessons from "Predicting Terrorism: A Machine Learning Approach". *Peace Economics, Peace Science and Public Policy*, 24(4), 2018. ISSN 1554-8597. doi: 10.1515/peps-2018-0040. URL http://www.degruyter.com/view/j/peps.2018.24.issue-4/peps-2018-0040/peps-2018-0040.xml.

Nathaniel Beck, Gary King, and Langche Zeng. Improving Quantitative Studies of International Conflict: A Conjecture. *The American Political Science Review*, 94(1):21–35, 2000. ISSN 0003-0554. doi: 10.2307/2586378. URL http://www.jstor.org/stable/2586378.

Claude Berrebi and Esteban F. Klor. On Terrorism and Electoral Outcomes: Theory and Evidence from the Israeli-Palestinian Conflict. *The Journal of Conflict Resolution*, 50(6):899–925, 2006. ISSN 0022-0027. URL http://www.jstor.org/stable/27638530.

Peter J. Bickel, Ya'acov Ritov, and Thomas M. Stoker. Tailor-Made Tests for Goodness of Fit to Semiparametric Hypotheses. *The Annals of Statistics*, 34(2):721–741, 2006. ISSN 0090-5364. URL http://www.jstor.org/stable/25463434.

Robert Blair and Nicholas Sambanis. Forecasting Civil Wars: Theory and Structure in an Age of 'Big Data' and Machine Learning. *Working Paper.*, 2019. URL http://egap.org/registration/1760.

Robert A. Blair, Christopher Blattman, and Alexandra Hartman. Predicting Local Violence. *SSRN Electronic Journal*, 2014. ISSN 1556-5068. doi: 10.2139/ssrn.2497153. URL http://www.ssrn.com/abstract=2497153.

Philipp C. Bleek and Eric B. Lorber. Security Guarantees and Allied Nuclear Proliferation. *Journal of Conflict Resolution*, 58(3):429–454, 2014. ISSN 0022-0027. doi: 10.1177/0022002713509050. URL https://doi.org/10.1177/0022002713509050.

Leo Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199–215, 2001. ISSN 0883-4237. URL http://www.jstor.org/stable/2676681.

Michael Colaresi and Zuhaib Mahmood. Do the robot: Lessons from machine learning to improve conflict forecasting. *Journal of Peace Research*, 54(2):193–214, March 2017. ISSN 0022-3433. doi: 10.1177/0022343316682065. URL https://doi.org/10.1177/0022343316682065.

Paul Collier and Anke Hoeffler. Greed and Grievance in Civil War. *Oxford Economic Papers*, 56(4):563–595, 2004. ISSN 0030-7653. URL http://www.jstor.org/stable/3488799.

Ursula Daxecker and Brandon Prins. Insurgents of the Sea: Institutional and Economic Opportunities for Maritime Piracy. *Journal of Conflict Resolution*, 57(6):940–965, December 2013. ISSN 0022-0027. doi: 10.1177/0022002712453709. URL https://doi.org/10.1177/0022002712453709.

Fangyu Ding, Quansheng Ge, Dong Jiang, Jingying Fu, and Mengmeng Hao. Understanding the dynamics of terrorism events with multiple-discipline datasets and machine learning approach. *PLOS ONE*, 12(6):e0179057, June 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0179057. URL http://dx.plos.org/10.1371/journal.pone.0179057.

Christian Dougherty. *Introduction to Econometrics*. Oxford University Press, Oxford, fourth edition, 2011a.

Christopher Dougherty. *Introduction to Econometrics*. Oxford University Press, Oxford ; New York, 4 edition edition, April 2011b. ISBN 978-0-19-956708-9.

Walter Enders, Todd Sandler, and Khusrav Gaibulloev. Domestic versus Transnational Terrorism: Data, Decomposition, and Dynamics. *Journal of Peace Research*, 48(3):319–337, 2011.

Marcel Fafchamps and Julien Labonne. Using Split Samples to Improve Inference on Causal Effects. *Political Analysis*, 25(4):465–482, October 2017. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2017.22. URL http://www.cambridge.org/core/journals/political-analysis/article/using-split-samples-to-improve-inference-on-causal-effects/6727EF63175DF517BF56EC64EFF3E807.

James D. Fearon and David D. Laitin. Ethnicity, Insurgency, and Civil War. *American Political Science Review*, 97(1):75–90, February 2003. ISSN 1537-5943, 0003-0554. doi: 10.1017/S0003055403000534. URL http://www.cambridge.org/core/journals/american-political-science-review/article/ethnicity-insurgency-and-civil-war/B1D5D0E7C782483C5D7E102A61AD6605.

Sean Gailmard. *Statistical Modeling and Inference for Social Science*. Cambridge University Press, New York, 2014.

Andrew Gelman and Eric Loken. The Statistical Crisis in Science. *American Scientist*, 102(6):460–465, 2014. ISSN 0003-0996. URL http://search.proquest.com/docview/1627946025?rfr_id=info%3Axri%2Fsid%3Aprimo.

Jeff Gill. The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly*, 52(3):647–674, September 1999. ISSN 1065-9129. doi: 10.1177/106591299905200309. URL https://doi.org/10.1177/106591299905200309.

Kristian Skrede Gleditsch and Michael D Ward. Forecasting is difficult, especially about the future: Using contentious issues to forecast interstate disputes. *Journal of Peace Research*, 50(1):17–31, January 2013. ISSN 0022-3433, 1460-3578. doi: 10.1177/0022343312449033. URL http://journals.sagepub.com/doi/10.1177/0022343312449033.

Justin H. Gross. Testing What Matters (If You Must Test at All): A Context-Driven Approach to Substantive and Statistical Significance: TESTING WHAT MATTERS. *American Journal of Political Science*, 59(3):775–788, July 2015. ISSN 00925853. doi: 10.1111/ajps.12149. URL http://doi.wiley.com/10.1111/ajps.12149.

Håvard Hegre and Nicholas Sambanis. Sensitivity Analysis of Empirical Results on Civil War Onset. *Journal of Conflict Resolution*, 50(4):508–535, August 2006. ISSN 0022-0027, 1552-8766. doi: 10.1177/0022002706289303. URL http://journals.sagepub.com/doi/10.1177/0022002706289303.

Håvard Hegre, Nils W Metternich, Håvard Mokleiv Nygård, and Julian Wucherpfennig. Introduction: Forecasting in peace research. *Journal of Peace Research*, 54(2):113–124, March 2017. ISSN 0022-3433. doi: 10.1177/0022343317691330. URL https://doi.org/10.1177/0022343317691330.

Håvard Hegre, Marie Allansson, Matthias Basedau, Michael Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyles, Lisa Hultman, Stina Högbladh, Remco Jansen, Naima Mouhleb, Sayyed Auwn Muhammad, Desirée Nilsson, Håvard Mokleiv Nygård, Gudlaug Olafsdottir, Kristina Petrova, David Randahl, Espen Geelmuyden Rød, Gerald Schneider, Nina von Uexkull, and Jonas Vestby. ViEWS: A political violence early-warning system. *Journal of Peace Research*, 56(2):155–174, March 2019. ISSN 0022-3433. doi: 10.1177/0022343319823860. URL https://doi.org/10.1177/0022343319823860.

Joshua B Hill, Daniel J Mabrey, and John M Miller. Modeling terrorism culpability: An event-based approach. *The Journal of Defense Modeling and Simulation*, 10(2):181–191, April 2013. ISSN 1548-5129. doi: 10.1177/1548512912455470. URL https://doi.org/10.1177/1548512912455470.

Matthew Hindman. Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences. *The ANNALS of the American Academy of Political and Social Science*, 659(1):48–62, May 2015. ISSN 0002-7162. doi: 10.1177/0002716215570279. URL https://doi.org/10.1177/0002716215570279.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, 1st ed. 2013, corr. 7th printing 2017 edition edition, September 2017. ISBN 978-1-4614-7137-0.

Alan B. Krueger. *What makes a terrorist: economics and the roots of terrorism*. Lionel Robbins lectures. Princeton University Press, Princeton, N.J., 2007. ISBN 978-0-691-13438-3.

James M. Landwehr, Daryl Pregibon, and Anne C. Shoemaker. Graphical Methods for Assessing Logistic Regression Models. *Journal of the American Statistical Association*, 79(385):61–71, March 1984. ISSN 0162-1459. doi: 10.1080/01621459.1984.10477062. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1984.10477062.

Timothy R. Levine, René Weber, Craig Hullett, Hee Sun Park, and Lisa L. Massi Lindsey. A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research. *Human Communication Research*, 34(2):171–187, April 2008. ISSN 0360-3989, 1468-2958. doi: 10.1111/j.1468-2958.2008.00317.x. URL https://academic.oup.com/hcr/article/34/2/171-187/4210725.

P E Meehl. What Social Scientists Don't Understand. In *Metatheory in social science: Pluralisms and subjectivities*, page 24. University of Chicago Press, Chicago, 1986.

Paul E. Meehl. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4):806, 1979. ISSN 1939-2117. doi: 10.1037/0022-006X.46.4.806. URL http://psycnet.apa.org/fulltext/1979-25042-001.pdf.

David Muchlinski, David Siroky, Jingrui He, and Matthew Kocher. Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis*, 24(1):87–103, 2016. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpv024. URL http://www.cambridge.org/core/journals/political-analysis/article/comparing-random-forest-with-logistic-regression-for-predicting-classimbalanced-civil-war-onset-data/109E1511378A38BB4B41F721E6017FB1.

Hannes Mueller and Christopher Rauh. Reading Between the Lines: Prediction of Political Violence Using Newspaper Text. *American Political Science Review*, 112(2):358–375, May 2018. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055417000570. URL http://www.cambridge.org/core/journals/american-political-science-review/article/reading-between-the-lines-prediction-of-political-violence-using-newspaper-text/4EABB473AFE18F157EEDE4339F34ABB0.

James A. Piazza. Types of Minority Discrimination and Terrorism. *Conflict Management and Peace Science*, 29(5):521–546, 2012.

Gerald Schneider, Nils Petter Gleditsch, and Sabine C. Carey. Exploring the Past, Anticipating the Future: A Symposium. *International Studies Review*, 12(1):1–7, March 2010. ISSN 15219488, 14682486. doi: 10.1111/j.1468-2486.2009.00909.x. URL https://academic.oup.com/isr/article-lookup/doi/10.1111/j.1468-2486.2009.00909.x.

Philip A Schrodt. Seven deadly sins of contemporary quantitative political analysis. *Journal of Peace Research*, 51(2):287–300, March 2014. ISSN 0022-3433. doi: 10.1177/0022343313499597. URL https://doi.org/10.1177/0022343313499597.

Sebastian Schutte. Regions at Risk: Predicting Conflict Zones in African Insurgencies*. *Political Science Research and Methods*, 5(3):447–465, July 2017. ISSN 2049-8470, 2049-8489. doi: 10.1017/psrm.2015.84. URL http://www.cambridge.org/core/journals/political-science-research-and-methods/article/regions-at-risk-predicting-conflict-zones-in-african-insurgencies/4DCDBA2BCC8B4E3D5057A2C37DDB2BD6.

J. David Singer. The peace researcher and foreign policy prediction. *Papers of the Peace Science Society*, 21:1:13, 1973. doi: 10.4324/9780203128398-18. URL https://www.taylorfrancis.com/.

M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974. ISSN 0035-9246. URL http://www.jstor.org/stable/2984809.

Ralph Sundberg and Erik Melander. Introducing the UCDP Georeferenced Event Dataset. *Journal of Peace Research*, 50(4):523–532, July 2013. ISSN 0022-3433, 1460-3578. doi: 10.1177/0022343313484347. URL http://journals.sagepub.com/doi/10.1177/0022343313484347.

Michael D Ward, Brian D Greenhill, and Kristin M Bakke. The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4):363–375, July 2010. ISSN 0022-3433. doi: 10.1177/0022343309356491. URL https://doi.org/10.1177/0022343309356491.

Ronald L. Wasserstein and Nicole A. Lazar. The ASA Statement on p-values: Context, Process, and Purpose. *The American Statistician*, 70(2):129–133, April 2016. ISSN 0003-1305, 1537-2731. doi: 10.1080/00031305.2016.1154108. URL https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108.

Frank DW Witmer, Andrew M Linke, John O'Loughlin, Andrew Gettelman, and Arlene Laing. Subnational violent conflict forecasts for sub-Saharan Africa, 2015–65, using climate-sensitive models. *Journal of Peace Research*, 54(2):175–192, March 2017. ISSN 0022-3433. doi: 10.1177/0022343316682064. URL https://doi.org/10.1177/0022343316682064.

Joseph K. Young. Repression, Dissent, and the Onset of Civil War. *Political Research Quarterly*, 66(3): 516–532, 2013. ISSN 1065-9129. URL http://www.jstor.org/stable/23563162.