5

Model Evaluation and Selection

Building statistical models in terms of systematic and stochastic components implies that, when the estimation is completed, the researcher has constructed a working model of the data-generating process. Such models embody more than just decisions about covariates. The researcher also makes decisions about the functional form linking covariates to the systematic component, the process governing random variation, and the degree to which we believe that different observations are independent of one another. Before worrying about particular parameters or other estimates, we must first convince both ourselves and our audiences that the model itself is a useful one. Models, as we all know, should be evaluated based on their relative usefulness for specific, well-defined purposes.

We begin this chapter with a caricature of current social science practice in evaluating and disseminating the results of statistical modeling. We do this to highlight common pitfalls and motivate what continues to be a shockingly underutilized model evaluation tool: out-of-sample prediction. We then describe the mechanics of predicting out-of-sample and demonstrate its use in real research problems.

5.1 CURRENT PRACTICE

Unfortunately, current publication practice demands output that often fails to fully exploit the underlying statistical models while also failing to display results in the most memorable and easily understood formats. Rather, readers are confronted with the Big Ugly Table of Numbers (BUTON), presenting estimated coefficients and their associated standard errors for a handful of models, perhaps fit to differing subsets of the available data. That is, most scholarly output consists of coefficient point estimates along with some measure of the uncertainty in these estimates. Scholars proceed apace to draw inferences by

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

calculating *t*- or *z*-statistics and comparing these values to arbitrary thresholds that are, nevertheless, imbued with near-magical importance. Based on the values of these statistical tests, scholars adorn the BUTON with stars, crosses, dots, accents, and other decoration.

5.1.1 BUTON

BUTONs are ubiquitous. Several appear in this book; Table 5.2 later in this chapter is a good example. This table satisfies at least one basic idea about presenting your research: science should be transparent; procedures and results should be widely available. Of course, it is important to also share the data, so that these results may be replicated by other scholars in different laboratories around the world, using different computers and different programs. In addition to these kinds of standard numerical displays, many scholars will include a variety of useful model diagnostics and "goodness-of-fit" statistics such as likelihood ratios, R^2, BIC, *F*- and other Wald tests. All these pieces of information are conditional on the sample used to fit the model; they tell us little about the extent to which the model is highly tuned to a particular data set.

BUTON are rarely compelling when trying to convince readers of a particular model's benefits. Why? Here are three easy exercises to illustrate the problem:

1. Think of your favorite empirical study. Write down a coefficient and a standard error from that model. On paper. Without looking. Put the answers here: $\boxed{\hat{\beta}: \qquad\qquad ; \sigma_{\hat{\beta}}: \qquad\qquad}$
2. Okay, try this one. Write down the estimated coefficient from any article you read in the past week. Put your answer here: $\boxed{\hat{\beta}: \qquad\qquad}$
3. What was the estimated intercept for the model from last week's homework? Answer: $\boxed{\hat{\beta}_0: \qquad\qquad}$

All those coefficients, all those standard errors. Like so many through the years, they are forgotten. This suggests that having a table of numbers somewhere for the careful scholar to review is important, but presenting a large table of numbers is unlikely to be compelling or memorable. The current publishing norm is to present tables of regression output, explaining it as you go. Nevertheless, this is frequently not the best option for making your analysis stick in your readers' minds. Nor is it necessarily a good strategy for your own model checking. It seems a sad waste of effort to reduce hard modeling work to simple tabular summaries, especially when the estimated models contain all the necessary components to build a simulation of the process you began studying in the first place. We want to take advantage of the models' richness to explore how we expect our dependent variables to behave in different situations of substantive relevance.

Since tables will continue to be produced – and rightly so – some care in their production seems merited. Some basic principles for constructing tables are:

- Tables are best for cataloging and documenting your results. Fill them with details for those carefully studying your work. Think of them as entries in the scientific record.
- BUTON are not well suited for quickly transmitting the crux of your findings in the body of an article or in a presentation to a wide audience.
- Tables should facilitate precise, analytical comparisons.
- Comparisons should flow from left to right.
- There should be enough white space to allow the eyes to construct focused comparisons easily.
- There should be no unnecessary rules (i.e., lines) that separate columns or rows within the table.
- Tables are most useful for reporting data: counts of things, or percentages.
- Numbers should be right justified to a common number of decimal places to facilitate comparisons.
- Tables should present information only as precisely as necessary; entries should reflect a reasonable degree of realism in the accuracy of measurement. This means 8% or 8.3% is generally better than 8.34214%. Items that yield extremely large (`9.061e+57`) or small (`5.061e-57`) quantities should be rescaled or, in the latter case, simply called 0.
- Rows and columns should be informatively and adequately labeled with meaningful English-language text, in groups or hierarchies if necessary.
- Entries in the table should be organized in substantively meaningful ways, not alphabetically.
- Give the table an informative title and footnotes to detail information inside the tabular display.

5.1.2 Graphical Displays

Fortunately things have begun to improve and many scholars create detailed appendices containing many BUTON that then appear on the Web alongside data archives. For example, Prorok's article in the 2016 *American Journal of Political Science* includes no big table of numbers, except in the supporting information, which is found on the publisher's website. This is becoming more common and is entirely sensible.

There is also an old and growing movement to display our data and results graphically whenever possible (Cleveland and McGill, 1984; Gelman et al., 2002; Kastellec and Leoni, 2007). The coefficient or ropeladder plots displayed in Figure 3.4 are but one example of such tools. Graphical displays, when done well, can convey much more information in a smaller area than tables. Readers

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

tend to be faster and more accurate in making comparisons when using graphs compared to tables (Feliciano et al., 1963). Subsequent recall of relationships or trends tends to be better when presented graphically. Maps (or map-like visualizations) are particularly memorable (Saket et al., 2015).

But presenting figures alone is not the answer, either (Gelman, 2011). Academic researchers (much less lay audiences or policy makers) routinely misinterpret confidence intervals of the sort displayed in Figure 3.7. Researchers can also produce graphics that are information-sparse, communicating relatively little compared to a table containing the same information. For example, Mutz and Reeves (2005) use experiments to study the impact of televised incivility upon political trust. They use bar charts to display their findings. These charts appear as Figure 3 in their article, commanding a prominent place and appearing in color in the journal's digital edition. The figure's number is apt because their display contains just three pieces of information. Three comparisons of interest appear in 31 square inches; the entire page is 62 square inches. Compared to displays such as Figure 3.3 this is a low content-to-space ratio.

In designing compelling graphical displays Tufte (1992) has excellent advice, including

- Show the data,
- Encourage comparisons,
- Maximize the data-to-ink ratio,
- Erase non-data and redundant data-ink,

to which we add,

- Give the graphical display an informative caption to detail information inside the display.

5.1.3 What's Left Out?

The dominant modes of scholarly communication – journals and books – impose constraints. Journal articles and books have page limits, printing costs, and black-and-white images. The constraints imposed by publication imply that only a relatively small subset of all the analysis we do for a particular project is ever disseminated. As a result, readers usually have no idea what models or findings were *omitted* from the reported tables. When combined with the fact that traditional outlets also tend not to publish "null" findings, we wind up with well-documented publication biases (Gerber and Malhotra, 2008; Gerber et al., 2001), among other problems.

Even if there were no constraints, and we all published arbitrarily long (and colorful) blog entries detailing all our analysis, we would still face the cognitive and resource limitations of our audiences. How many readers are really willing to dig through an online appendix of many models and terse equations? Are

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

the paper's findings really sufficiently important that it is reasonable to expect audiences and reviewers to spend that much time and effort? Astute consumers of our analysis recognize this fact and understand the incentives we all face, leading them to be understandably skeptical of our results. How can a reader be reassured that the models we choose to present aren't simply the result of a specification search? Credible, reproducible, and transparent methods for model selection help.

Reproducibility, Transparency, and Credibility

Over the last decade, several incidents drew attention to the reproducibility problems in social scientific research. In one example, enterprising PhD students discovered an error in calculations conducted directly in the Excel spreadsheet program (Herndon et al., 2014), leading to the revision of an influential study in economics. In perhaps the most widely reported incident, another team of then-graduate students discovered scientific fraud by reanalyzing study data (Broockman et al., 2015), resulting in the retraction of a study published in *Science*. This came as quite a shock to much of the social science community, but the so-called crisis of reproducibility has been well-known for at least a decade (Ioannidis, 2005).

Several organized responses to doubts about reproducibility have emerged. Recently http://retractionwatch.com/ appeared as a resource for monitoring publications that have been (or should be) withdrawn from the published body of knowledge. In the world of cancer biology, there is an ongoing project to independently replicate a subset of experimental results from a number of high-profile papers published between 2010 and 2012 (https://osf.io/) while building a framework for reproducible research. A team of over 270 researchers in psychology has come together in a systematic effort to reproduce the results of 100 published studies in major psychology journals, with sobering results to date (Open Science Collaboration, 2015).

The solution to this broad problem is complicated, but researchers can take concrete steps to make sure that their own research is reproducible and that they don't wind up reading about their research on retractionwatch.com. First, we need to keep a record of everything. In today's world, this means version control. Second, we need to keep a backup of every change, ideally off-site or in the cloud. Third, we want a work flow that will minimize error by being repeatable so that we can redo calculations and graphics if data or ideas about which model is most appropriate should change. This means working from files with scripts, not clicking buttons. Storing data and performing analysis in spreadsheet programs is a recipe for disaster. Fourth, we want a way to integrate statistical analysis and writing. Cutting and pasting is not the solution, nor is manual entry of results into your textual report. Fifth, it is important to have a way to annotate our procedures and our data, for others and for yourself. Sixth, it is important to share our data – not just our conclusions – with others.

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

5.1.4 Sign, Significance, and the Perils of *p*-Values

Social science theories are still generally too coarse to generate strong predictions about, say, the functional form a relationship should take, or what exactly the "null" hypothesis should be in a particular application. As a result, we generally state our hypotheses as something like "The conditional mean of Y should be increasing in X_f," where f denotes the covariate favored by the theoretical argument. We then fit models, including X_f, and if β_f, the regression parameter for X_f, is the right sign and the standard error is small enough, we declare the relationship "significant," scarcely acknowledging the fact that this "significance" is entirely conditional on the model.

In case you were wondering ... 5.1 Frequentist *p*-values

Given a statistical model with parameter θ and an hypothesized value of the parameter, θ_h, the p-value is the probability of observing a value at least as extreme as the one obtained with the actual data at hand if θ_h were true. Researchers view small p-values as reason to reject the hypothesis that $\theta = \theta_h$.

Researchers claim *statistical significance* (at the α level) if the p-value for a favored parameter in a particular model is less than some arbitrary threshold, α. If α is unstated, then social science convention holds $\alpha = 0.05 = \frac{1}{20}$.

Current practice takes as given that a small p-value for β_f is a sufficient reason to prefer a model that includes X_f. This claim to statistical significance is usually based on the (unstated) null hypothesis that the "true" value of the regression coefficient is 0.0 and some statistical theory telling us that the estimates are normally distributed in sufficiently large samples. In many instances, model selection is implicit, i.e., authors present several models but then discuss only one in any detail.

Authors sometimes proceed to describe the "substantive significance" of their findings, i.e., the size of the marginal effect implied by their chosen model for the variable of interest. Substantive significance is hard to evaluate. The now-standard approach is for an author to pick a model that she likes and compare the change in the expected value of the outcome over some range of the covariate of interest. If this difference in expected values is "big," then the author claims to have identified an important relationship. Authors may even present a whole collection of models as "robustness checks," presumably as evidence that no matter how one fits the model, a similar relationship between the covariate of interest and the response obtains.

Notwithstanding all the claims about the sign, significance, and magnitude of a relationship, explicit justification for the selected model is rare. There are myriad examples in the literature that discuss in detail the statistical significance of estimated regression coefficients in different models but fail to comment on the comparative fit of these same models. Indeed, we observe authors arguing for or against models that are practically indistinguishable from one another *as models*.

This practice of fitting models based on theory (perhaps post hoc) and then filtering results based on the p-values for some covariates is problematic. There is a well-established literature that points away from the uncritical use of p-values in model selection for three reasons. First, p-values do not have the same interpretation after measurement and model selection as they do ex ante. Second, any statement of magnitude is conditional on the model employed, as is any claim to statistical significance. *Identifying "significance," substantive or otherwise, therefore requires that authors first justify their preferred models and the comparison scenarios.* Only then can we meaningfully turn to the magnitude of the difference implied by the chosen models, covariates, and scenarios. Third, using p-values ignores the issue of *model* uncertainty. Any statement about parameter values is conditional on the model and its underlying assumptions. If we are uncertain about which of a variety of possible models is most appropriate, then using p-values to justify selection is nonsensical.

Theory, by itself, is a weak justification for preferring a particular model specification, especially if one of the researchers' goals is to "test" that exact theory. Rather, evaluating model performance can be viewed as an integral part of the research enterprise: a good and useful theory should lead to better prediction. If the data support the theoretical claim, then the theoretically motivated model specification should outperform feasible (and simpler) competitors. Unfortunately, standard practice often provides no explicit reason to prefer the models presented compared to competitors; p-values on regression coefficients do not help in making this determination. We argue that out-of-sample prediction is a sensible, flexible, and powerful method for adjudicating between competing models and justifying model selection.

An Example: Trade and the World Trade Organization, Part I

As an example, consider the vigorous empirical debate over whether the World Trade Organization (WTO) and its precursor, the General Agreement on Trade and Tariffs (GATT), alter countries' patterns of trade in goods and services. Rose (2004) kicked off the debate. He pools annual dyadic trade flows for the post–World War II period and regresses it on the dyad-year's WTO status, along with numerous other covariates. He finds no evidence for a consistent relationship between a dyad's GATT/WTO status and trade flows. Subramanian and Wei (2007), Tomz et al. (2007), and Goldstein et al. (2007) claim to overturn Rose's null findings by arguing for more-nuanced measurement of trade and GATT/WTO participation. Rose's bilateral trade data set, pooled over 1949–99, has 234,597 observations, while Goldstein et al. (2007)'s expanded data set has $n = 381,656$. In such a context, statistical

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

significance for specific coefficients is a weak criterion by which to evaluate a model; it is highly unlikely that any relationship in these data is exactly 0.0. The danger of overfitting here is not trivial; these authors use models with country, dyad and/or year effects along with a slate of covariates running into the double digits. Nevertheless, the bulk of the debate revolves around whether the GATT/WTO variables are statistically significant. None of the authors challenging Rose present evidence for why we should prefer a model with GATT/WTO variables included; presumably the existence of significant coefficient estimates is justification enough. The substantive debate is much impoverished by the participants' focus on *p*-values, failing to engage the models in a predictive sense.

5.2 THE LOGIC OF PREDICTION, OUT-OF-SAMPLE

The logic of out-of-sample prediction is simple: to the extent our statistical models capture underlying social processes, they should be able to predict instances not used to calibrate the model in the first place, assuming the underlying DGP has not changed. Two comparisons are relevant for evaluating the model. The first involves comparing the model's in-sample fit to its ability to predict new data. This is a way of guarding against overfitting. The second compares models against each other as a way of justifying model selection.

Overfitting refers to an estimated model that is sculpted to the data in hand where that data is not necessarily characteristic of all the data that might be observed. Overfitting is a common threat to models that are reasonably complex, especially in observational studies. With enough parameters we can perfectly fit the observed data. But such a model is not likely a very useful one nor a correct description of the data-generating process. To anneal our description of the data-generating process against the threat that the estimated model is tuned entirely to the observed data, we can test the model's ability to predict data not used to fit the model in the first place.[1] These data for validating the model can literally be "new" in the sense that they were observed after fitting the model. But it is more common that we consciously hold back some of our data for later use in the model evaluation phase.[2]

A model that fits about as well in-sample as out-of-sample indicates that the model can generate predictions that are in line with the data-generating process (which is almost never observed in nonexperimental studies). It also enables greater confidence that the estimated model was not overly influenced by some particular features of the data used to calibrate the model. We can

[1] This is also related to Type-I error: testing an hypothesis using the data that suggested the hypothesis in the first place is unlikely to reject a null.

[2] What constitutes "new" data poses thorny questions in environments where agents are observing and learning from each other while acting strategically. How best to implement and evaluate out-of-sample prediction in such circumstances requires particular care. See, for example, Gleditsch and Ward (2013) for the case of interstate conflict.

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

think of this second point, broadly, as a generalization of the concept of influence diagnostics (leverage statistics), including "hat" quantities such as DF-fit, DF-β, and others that decompose the fit into the individual observations' contributions to summary statistics.

Simply performing equivalently in- and out-of-sample does not mean a model is "good." It merely means that it has not been overly sculpted to a particular set of observations. In the second set of comparisons we seek to make judgments about how useful a particular model is by evaluating predictive performance (out-of-sample) relative to other models. If our favored model is no better than feasible and simpler alternatives at predicting new data, then we have little reason to prefer that model, regardless of whether our theoretically inspired specification has "significant" coefficients for special covariates. If we have little reason to believe that the favored covariate is an important part of the underlying data-generating process, then it makes little difference that its regression coefficient conforms to theoretical expectations in an overfit model. Even with well-designed and executed experiments, where the researcher partly controls the DGP, and we can recover estimates of causal relationships, out-of-sample prediction can be helpful. One of the commonly cited weaknesses of experimental interventions is the difficulty in sorting out how well findings generalize to other contexts where exposure is nonrandom. Advance in this domain typically proceeds through a process of comparing competing models, ideally using new data (Clarke, 2006).

A direct consequence of the possibility of overfitting is that statistical significance for a particular parameter does *not* imply that that a model including that parameter predicts better out-of-sample than one without. In fact, the more complicated model may even perform more poorly (Lo et al., 2015; Ward et al., 2010). This issue becomes more salient as the size of the data set increases. In an observational study with twenty observations, statistical significance at five percent, for example, requires systematic patterning and very little error; in a binomial model you could be wrong one time in twenty. The same study with 20,000 observations imposes fewer requirements, since it is unlikely that any regression parameter has a true value of exactly 0. In the binomial case, one in twenty with 20,000 as the number of observations yields 1,000 plausible errors within the five-percent range. With 200,000 observations, it is difficult for an estimated parameter not to achieve "significance," largely because we may expect variation sufficient for precise estimates.

5.2.1 The Process of Evaluating Out-of-Sample Predictions

In out-of-sample prediction, interest centers on evaluating a model's prediction error, often referred to as *generalization error* or *generalization performance*. To estimate this quantity, we must identify a *training set* (the data used to fit the model), a *test set* (the data used to evaluate model's predictions), a model, and a loss function, which measures the model prediction's deviation from the

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

actual value in the test set. Good models will have good performance in the training *and* test data sets.

More formally, suppose we observe our outcome of interest, $\mathbf{y}$, and covariates $\mathbf{X}$ for a set of n units. We can partition our data into our training set, S, and our test set, V, such that $n = S \cup V$ and $S \cap V = \emptyset$. We also have a model for Y_i, denoted $M(\mathbf{x}_i; \boldsymbol{\theta})$, and is a function of the covariates and parameters, $\boldsymbol{\theta}$. The term $\hat{\boldsymbol{\theta}}_S$ is the parameter estimate based on S, the observations in the training set.

For continuous Y we denote a prediction based on $M(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_S)$ as $\hat{y}_i$. The most commonly used loss function for continuous Y is squared error loss:

$$\text{Loss}(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 = (y_i - M(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_S))^2.$$

Less commonly used is absolute error:

$$\text{Loss}(y_i, \hat{y}_i) = |y_i - M(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_S)|.$$

In the case that Y is categorical, falling into one of G possible categories, there are several frequently employed loss functions. In situations where the models generate predicted probabilities for being in a category g we can sum across categories in a manner analogous to squared error or absolute loss. For the absolute error case, in which $\mathbb{1}_g(y_i)$ is the indicator function for category g and $M_g(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_S) = \hat{M}_g$ is the predicted probability of being in category g, we get

$$\text{Loss}(y_i, M(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_S)) = \sum_{g=1}^{G} |(\hat{M}_g - \mathbb{1}_g(y_i))|.$$

Also commonly used is the bounded loss function, which takes on a value of unity whenever $\hat{y}_i \neq y_i$ and 0 otherwise. The predicted category, $\hat{y}_i$ is typically the category with the largest predicted probability: $\hat{y}_i = \arg\max_g \hat{M}_g$. Finally, there is deviance given as

$$\text{Loss}(y_i, M(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_S)) = -2 \sum_{g=1}^{G} \mathbb{1}_g(y_i) \log \hat{M}_g.$$

The quantity of interest is the expected prediction error, or Err in the notation of Hastie et al. (2008):

$$\text{Err} = \text{E}\left[\text{Loss}\left(Y, M(X, \hat{\boldsymbol{\theta}}_S)\right)\right],$$

where expectations are taken over all that is random (partition of training and test set, etc.).

Cross-Validation

The question naturally arises as to where the test and training sets come from in real-world applications. In a data-rich environment we might consider actually

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

withholding some of the data for later use as the test set. Or we might expect a new sample to arrive later in time. But both of these are relatively uncommon in the social sciences. Cross-validation is a way to use the data we do have as both training and test sets, just not at the same time.

Initial work on cross-validation followed the thinking of Seymour Geisser, who believed that the inferential framework of hypothesis testing was misleading. Instead, he believed that using prediction-based tools would lead to the selection of better, i.e., more useful models, even if these models were not the "true" models.

More specifically, *k*-fold cross validation divides the data randomly into *k* disjoint subsets (or folds) of approximately equal size.[3] The division into subsets must be independent of all covariates for estimating generalization error. Each of the *k* subsets will serve as the test set for the model fit using the data in the remaining $k - 1$ subsets as the training set. For each observation, we calculate the prediction error based on the predicted value generated by the model fit to the training data. More formally, if $\kappa(i)$ is the fold containing observation i and $-\kappa(i)$ is its complement, then the *k*-fold cross-validation estimate of *Err* is given by

$$\mathrm{Err}_{\mathrm{CV}}(M, \mathbf{y}, \mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Loss}(y_i, M(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_{-\kappa(i)})).$$

Alternatively we can write the *k*-fold cross validation estimate as the average of prediction errors within each of the *k* folds:

$$\mathrm{CV}_j(M, \mathbf{y}, \mathbf{X}) = \frac{1}{|\kappa_j|} \sum_{i \in \kappa_j} \mathrm{Loss}(y_i, M(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_{-\kappa(i)}))$$

$$\mathrm{Err}_{\mathrm{CV}}(M, \mathbf{y}, \mathbf{X}) = \frac{1}{k} \sum_{j=1}^{k} \mathrm{CV}_j(M, \mathbf{y}, \mathbf{X}),$$

where κ_j denotes the set of observations i in each fold $j \in \{1, 2, \ldots, k\}$, and $|\kappa_j|$ denotes the cardinality, or number of observations, of this set.

What about *k*? One choice is "leave-one-out" cross-validation, in which $k = n$. Leave-one-out has some nice properties but has higher variance and is computationally more expensive than setting *k* to some smaller value. Shao (1993) shows that leave-one-out cross-validation will not lead to the selection of the "true" model as $n \to \infty$ but leaving out a larger number of observations will – if we are willing to entertain the notion of a "true" model. As a result, cross-validation setting $k = 5$ or 10 is fairly common. Under such a decision there may be some bias in our estimate of Err, but the bias is upwards. As *n* grows large the distinction becomes less relevant. In comparing models,

[3] Note that we are temporarily departing from our notation convention in other chapters, where *k* was used to denote the number of covariates in a model. Here we follow the literature on cross-validation and use *k* to denote the number of "folds."

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

the important consideration is using the same k for all the cross-validation estimates. The exact value of k is less important.

We clearly prefer models with lower prediction error. How big of a difference in prediction error is "big enough?" Currently, a general description of the distribution cross-validation estimator does not exist. But we can estimate the empirical variance of the cross-validation Err:

$$\text{var}\left[\text{Err}_{\text{CV}}\right] = \frac{1}{k}\text{var}\left[\text{CV}_1, \ldots, \text{CV}_k\right].$$

The square root of this quantity is the standard error of the cross-validation estimate of prediction error. Hastie et al. (2008) suggest the "one standard error rule" in which we select the most parsimonious model whose cross-validation-estimated prediction error is within one standard deviation of the model with the smallest prediction error. Put another way, if a simpler model's prediction error falls within one standard deviation of a more complicated model's prediction error then the simpler model is to be preferred.

When should cross-validation occur, and how does it relate to model building more generally? Hastie et al. (2008) argue that "In general, with a multi-step modeling procedure, cross-validation must be applied to the entire sequence of modeling steps. In particular, samples must be 'left out' before any selection or filtering steps are applied." (p. 245–249). It is important to note, however, that they are discussing cross-validation in the context of machine learning, where there are a very large number of predictors and little to inform model selection (e.g., some genomics applications). Most social science applications, on the other hand, present arguments justifying the inclusion of specific predictors or decisions about functional form (e.g., "interaction terms"). In such situations, some model selection has already been accomplished. Cross-validation can then be used to compare competing models and justify model choices without necessarily building models from scratch for each of the k folds.

Example of Cross-Validation

We look to the Fearon and Laitin (2003) classic article on civil wars for an example. We display the BUTON for a reestimation of their logistic regression model in Table 5.1.

For exposition we conduct a two-fold cross-validation. We randomly split the data in two sets denoted κ_1 and κ_2. We first fit the model using just the data in κ_1. Using those estimated coefficients, along with covariate data, we produce an in-sample predicted probability for κ_1 and an out-of-sample predicted probability for the cases in κ_2. We then refit the model, reversing the roles for κ_1 and κ_2, producing both in-sample and out-of-sample predicted probabilities for each observation. From here we can construct a variety of displays and undertake various calculations summarizing the differences, if any. A ROC plot summarizing in-sample and out-of-sample predictive performance is one example, shown in Figure 5.1. The model's out-of-sample predictive

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

TABLE 5.1 *Logistic regression of civil war onset, replicating Fearon and Laitin (2003).*

	β	$\sigma_{\hat{\beta}}$
Intercept	−6.75	0.73
Prior war	−0.90	0.31
GDP per capita	−0.34	0.07
Population	0.26	0.07
% mountainous	0.23	0.08
Noncontiguous state	0.35	0.28
Oil exporter	0.91	0.28
New state	1.59	0.34
Instability	0.60	0.24
Democracy	0.02	0.02
Ethnic fractionalization	0.07	0.37
Religious fractionalization	0.40	0.51
n	6,402	
$\log \mathcal{L}$	−483	
AIC	990	
BIC	1,071	

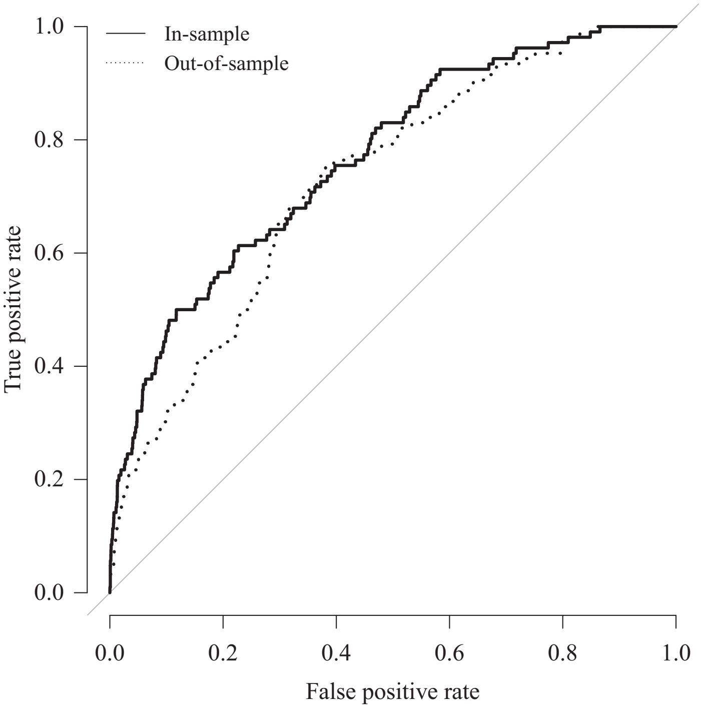

FIGURE 5.1 ROC plot of twofold cross-validation of the Fearon and Laitin replication from Table 5.1.

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

$\mathcal{R}$ Code Example 5.1 *Twofold CV*

```
library(pROC)
flmdw <- read.csv("flmdw.csv")
infitset<-sample(rownames(flmdw),size=dim(flmdw)[[1]]/2) #divide the sample
totalset<-rownames(flmdw)
intestset<-setdiff(totalset,infitset)
fl.trainset<-flmdw[infitset,]
fl.testset<-flmdw[intestset,]
#estimate
training.fit<- glm(as.factor(onset) ~ warl + gdpenl + lpopl1 +
    lmtnest + ncontig + Oil + nwstate + instab + polity2l +
    ethfrac + relfrac, family = binomial(link = logit),
    data = fl.trainset)
pi.train <- predict(training.fit, type="response") #in-sample predictions
pi.test <- predict(training.fit, type="response",
    newdata = fl.testset) #out-of-sample predictions

flmdw$is.fit<-flmdw$oos.fit<-NA
flmdw[infitset,"is.fit"]<-pi.train
flmdw[intestset,"oos.fit"]<-pi.test
training.fit2<- glm(as.factor(onset) ~ warl + gdpenl + lpopl1 +
    lmtnest + ncontig + Oil + nwstate + instab + polity2l +
    ethfrac + relfrac, family = binomial(link = logit),
    data = fl.testset)
pi.train2 <- predict(training.fit2, type="response")
pi.test2 <- predict(training.fit2, type="response", newdata = fl.trainset)
flmdw[intestset,"is.fit"]<-pi.train2
flmdw[infitset,"oos.fit"]<-pi.test2

#plot
par(bty="n", las=1)
plot.roc(flmdw$onset,flmdw$is.fit, xlab="False positive rate",
   ylab="True positive rate", lwd=3,
   legacy.axes=T, se=T)
plot.roc(flmdw$onset,flmdw$oos.fit,add=T, lty=3, lwd=3)
legend("topleft", legend=c("In-sample", "Out-of-sample"),
    lty=c(1,3), bty="n")
```

performance is marginally worse than in-sample. Code Example 5.1 demonstrates how we undertake two-fold cross-validation “by hand.”

For the sake of simplicity in this two-fold cross-validation example, we did not utilize the estimated models’ descriptions of our estimation uncertainty. To do this we could redraw many coefficient vectors from the multivariate normal distribution to produce many predictions, both in- and out-of sample, building up distributions of predictions. Doing so, we would discover that, in this case, the model’s in- and out-of-sample predictions overlap considerably.

5.2.2 Variations on a Theme

Cross-validation is part of a class of *resampling* tools that take advantage of modern computing power to repeatedly analyze subsamples of a common data set in order to build up information about a quantity or distribution of interest. The bootstrap, jackknife, and randomization tests are other tools in this class.

Bootstrap Estimation

The bootstrap differs from cross-validation, although both rely on randomly sampling the data we have to build up distributional insights for complicated problems. Bootstrapping is an estimation method, whereas cross-validation is typically a post-estimation tool for formal evaluation.

Bootstrap methods involve repeatedly estimating a statistic of interest on a series of random subsamples from the observed data. Importantly, bootstrapping relies on sampling *with replacement*, whereas cross-validation partitions the data into disjoint subsets. Thus, bootstrap methods aim to build up distributional information about the statistic of interest (calculating standard errors, for example).

Suppose we have a data set with 1,000 observations, and we are interested in the model $Y \sim f(y; \boldsymbol{\theta})$. We could find bootstrap estimates of the standard errors around $\hat{\boldsymbol{\theta}}$ by randomly selecting 100 observations (with replacement), estimating the regression model, and storing the estimate of $\hat{\boldsymbol{\theta}}$. Repeating this procedure many times builds up a distribution for $\hat{\boldsymbol{\theta}}$. This distribution is treated as the sampling distribution, because it is. The goal is to derive robust estimates of population parameters when inference to the population is especially problematical or complicated.

Jackknife

The jackknife is an even older resampling technique designed to build up sampling information about a statistic. The jackknife involves iteratively removing each data point and calculating the statistic on the $n-1$ remaining observations and then averaging across all these values. The jackknife is similar to leave-one-out cross-validation in that it serially leaves a single observation out of the estimation and then computes an average of the statistics being examined from the $n-1$ jackknifed subsample statistics. The variance of the averages calculated on the resamples is an estimate of the variance of the original sample mean but robust in small samples as well.

Permutation and Randomization Tests

Permutation and randomization tests are yet other reanalysis techniques designed to test whether differences between groups of observations "matter" under the maintained hypothesis that they should not. As an example, consider a researcher-controlled experiment with n subjects in which $s < n$ receive a stimulus, while the other $n - s$ do not. The researcher then measures the outcome of interest and proceeds to make comparisons. The so-called sharp null hypothesis maintains that the effect of the treatment is exactly 0.0 for all subjects. Under this assumption, we should be able to shuffle which subjects

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

are labeled as having received the stimulus and still retrieve approximately the same outcome distributions as what was observed. In "shuffling" the label, the same proportion of the subject pool is labeled as "stimulated," we merely alter which subjects those are. In other words, we are sampling without replacement. In a permutation test, the researcher actually generates all $\binom{n}{s}$ distinct ways of assigning stimulus to s subjects out of n, constructing the distribution of the statistic of interest under the sharp null. We can then determine exactly where in this distribution our observed data fall, generating exact p-values.

Permutation tests can become cumbersome in even moderate samples. For example, a study with 30 subjects, half of whom receive a stimulus, has 155,117,520 different combinations. In order to determine whether our observed data are "surprising," we do not necessarily need to construct all possible permutations. We can simply do many of them. When inference is based on such a random sample of all possible combinations, we refer to a *randomization test*.

5.2.3 Out-of-Sample Prediction: Trade and the WTO, Act II

The participants in the trade-and-WTO debate introduced in Section 5.1.4 rely on the augmented "gravity model," estimated in the following log-linear form:

$$\begin{aligned} \log y_{ijt} &\sim f_{\mathcal{N}}(\mu_{ijt}, \sigma^2_{ij}) \qquad \forall \quad j \neq i \\ \mu_{ijt} &= \mathbf{x}^{\mathsf{T}}_{ijt}\boldsymbol{\beta}, \end{aligned}$$

where i and j index countries and t indexes time. Interest centers on the trading pair or "dyad," ij. In this example, all models treat dyad-years as conditionally independent. The dependent variable, Y, is averaged bilateral trade; $\mathbf{X}$ is a matrix of possibly time-varying covariates, including year indicator variables. The parameters $\boldsymbol{\beta}$ and σ^2_{ij} are to-be estimated quantities. The subscripts on σ^2 indicate the commonly recognized problem that repeated measurements on the same dyad should show some dependence; this is most commonly addressed by using a sandwich estimator and "clustering" standard errors at the dyad level.

We begin by replicating the Rose (2004) and Tomz et al. (2007) (TGR) analysis. For the sake of direct comparison, we use the same slate of covariates as Rose and TGR and the data set provided on Tomz's website. In addition to indicators for dyad-level GATT/WTO involvement, $\mathbf{X}$ includes the log great circle distances between countries, the log product of real GDP, the log product of real GDP per capita, the log product of land area, and indicators for colonial ties, involvement in regional trade agreements, shared currency, and the Generalized System of Preferences, shared language, shared borders, shared colonizing powers, whether i ever colonized j, whether i was ever a territory of j, the number of landlocked countries in the dyad, and the number of island nations in the dyad.

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

TGR measure GATT/WTO participation based on whether countries were formal members, “non-member participants,” or out of the system entirely. An “F-F” dyad is one where both countries are formal GATT/WTO members; “F-N” dyads are those where one is a formal member and one is a non-member participant; “F-O” dyads have one formal member and one country out of the system. “N-N,” “N-O,” and “O-O” dyads are defined analogously.

We will base our comparisons on four models. In Model 1 we fit a benchmark gravity model using only GDP, per capita GDP, and distance as covariates. Model 2 augments the benchmark model with TGR’s five variables encoding the dyad’s GATT/WTO status. Model 3 is Rose’s “default” gravity model reported in Rose (2004, 104) but excluding any variables identifying GATT/WTO participation. In the fourth model we replicate TGR’s model 3 (Tomz et al., 2007, 2012), which includes a series of variables accounting for the GATT/WTO status of the dyad.[4] We refer to this model as TGR-3.

To evaluate these competing models we compare cross-validated out-of-sample predictive performance. Table 5.3 displays the mean-squared prediction error for each model under two prediction scenarios. First, we conducted five-fold cross validation using all the data; these results are the first column in Table 5.3. As a substantive matter, TGR argue that the GATT effect is most pronounced in the pre-1967 period. We therefore repeat the cross-validation but using only data prior to 1967. These results are in the second column. We see that including GATT/WTO variables buys very little in terms of predictive performance out-of-sample. Comparing Models 1 and 2, we see that the addition of the GATT/WTO covariates improves model performance by less than 1%, with a similar result when comparing Model 3 with TGR-3. In the pre-1967 period, the inclusion of the GATT/WTO covariates improves model performance by about 1%.

Table 5.3 also reports the standard error of the cross-validation for the model with the smallest MSE in each column. Using the one standard error rule we have no reason to prefer the TGR-3 model over Model 3 in the full data set. In the pre-1967 period, TGR-3’s predictive performance is one standard error better than Model 3, yet still represents a 1% improvement.

Participants in the trade-GATT/WTO debate fit models to similar data and then based their knowledge claims on the (non)significance of GATT/WTO covariates without first comparing model fit. In our reestimations, we recover TGR’s “significant” GATT/WTO coefficients given a null hypothesis of $\beta = 0.00$ and $\alpha = 0.05$. But this is weak evidence in such a huge data set. Using an out-of-sample prediction heuristic, we find that the more-complicated TGR-3 model is nearly indistinguishable from simpler alternatives in terms of its ability to predict trade flows not already included in fitting the model. We see

[4] From our estimation and based on TGR’s discussion of their findings, there appears to be a typographical error in the original table; estimates and standard errors for “Formal member and non-member participant” and “Both non-member participants” were transposed. Correct estimates are reported in Table 5.2.

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

TABLE 5.2 *Linear regression of bilateral trade flows 1948–99, replicating Tomz et al. (2007).*

	(1)	(2)	(3)	(4-TGR3)
Distance	−1.237	−1.251	−1.120	−1.129
	(0.020)	(0.005)	(0.022)	(0.022)
GDP	0.815	0.837	0.916	0.926
	(0.007)	(0.002)	(0.009)	(0.010)
GDPpc	0.506	0.479	0.321	0.312
	(0.012)	(0.003)	(0.014)	(0.014)
GATT/WTO: F-F		0.469		0.173
		(0.022)		(0.067)
GATT/WTO: F-N		0.712		0.410
		(0.024)		(0.071)
GATT/WTO: N-N		1.305		0.796
		(0.042)		(0.142)
GATT/WTO: F-O		0.136		0.064
		(0.023)		(0.065)
GATT/WTO: N-O		0.304		0.327
		(0.030)		(0.090)
GSP			0.857	0.851
			(0.032)	(0.032)
Regional FTA			1.200	1.187
			(0.106)	(0.110)
Currency union			1.116	1.114
			(0.122)	(0.123)
Common language			0.315	0.312
			(0.040)	(0.040)
Shared border			0.528	0.517
			(0.111)	(0.110)
Num. landlocked			−0.271	−0.269
			(0.031)	(0.031)
Num. island			0.041	0.018
			(0.036)	(0.036)
Land area			−0.097	−0.093
			(0.008)	(0.008)
Common colonizer			0.584	0.523
			(0.067)	(0.067)
Currently colonized			1.078	0.937
			(0.234)	(0.234)
Ever had colonial rel.			1.162	1.153
			(0.117)	(0.115)
Common country			−0.015	−0.019
			(1.081)	(1.071)
$\bar{R}^2$	0.62	0.62	0.65	0.65
AIC	1,005,451	1,002,665	986,246	985,288
BIC	1,006,031	1,003,298	986,951	986,044

Note: All models include a constant, year dummies, and report dyad-clustered standard errors in parentheses. $n = 234{,}597$.

TABLE 5.3 *Out-of-sample predictive performance for models of international trade. Cell entries are mean-squared prediction errors, and quantities in parentheses are standard errors for models with the smallest MSE in each column.*

	CV (all years)	CV (pre-1967)
Model 1	4.35	1.79
Model 2	4.31	1.74
Model 3	4.05	1.59
Model 4 (TGR-3)	4.04 (0.02)	1.57 (0.01)

little reason to believe that, among these competitors, the models including GATT/WTO covariates should be privileged over those that do not.

5.2.4 Benefits of Cross-Validation

In-sample fit statistics are not necessarily informative about how well the model describes the data-generating process. Cross-validation is a general approach that helps overcome these shortcomings.

Clearly, using real test sets of data are preferable to constructed test sets. But we go through a lot of trouble and expense to collect the data we do have. It is a shame to spend it all in one place. Cross-validation, retrodiction, and forecasting help the researcher to determine how well something is fitting outside of the null hypothesis testing framework for inference. Indeed, you can use cross-validation heuristically, to examine which variables, for example, will most degrade your ability to accurately use your estimated model to generate precise predictions. It is sometimes the case that deleting a highly significant variable may have little impact on the predictive power of a model, while another variable is an extremely powerful predictor.

Cross-validation uses the power of resampling to address problems in analyzing data that are not a random sample from some larger population. Much of the data analyzed in the social sciences falls into this bucket, all the more so in "big data" applications. The standard inferential framework may be less informative than a predictive one if we are dealing with all the existing data anyway.

Finally, cross-validation allows us to frame the results in terms of the substantive questions driving the results, rather than in terms of a BUTON containing numbers that no one will remember tomorrow.

5.3 CONCLUSION

This chapter provided some tools for model selection and evaluation, in contrast to current practice, in which model selection is often implicit and small p-values are prized. Model selection is a broad topic, and we have only

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

scratched the surface. But the key point is that we cannot declare victory simply because we fit a model with "significant" results. It is easy to overfit the data and build a model that *only* describes the data already in hand. That is almost never the goal, because the researcher already has that data and can make models arbitrarily close to perfect with those data. The issue is whether the estimated model will be useful for additional data.

We must also compare models to one another and justify our preferred specifications *prior to* making inference about any parameters. Out-of-sample prediction is a powerful tool for annealing results against overfitting. Cross-validation is the most common way of conducting out-of-sample tests.

Most statistical software packages contain functionality for out-of-sample prediction, but it is often worthwhile to design predictive exercises to speak directly to questions of substantive interest. Existing software's default settings for loss function or method of dividing the data may not be best for a particular application. In certain circumstances we may want to see how a model works in specific kinds of cases, not necessarily in all of them. By dividing your sample into different sets and using a cross-validation strategy, we can probe the dependencies between the model and the data. If it is possible to keep some data isolated from the estimation process altogether, then all the better.

We are not, however, arguing for a pure data-mining approach, absent substantive knowledge and theoretical reflections. A good theory should lead us to specify models that predict better, but better-predicting models do not necessarily reflect a "true" or even causal set of relationships. We must also be careful in constructing out-of-sample prediction exercises when there are dependencies in the data (e.g., temporal correlation) or when we have reason to believe that the fundamental processes at work may have changed. As a result, the prediction heuristic is general, powerful, simple to understand, and relatively easy to build in a computer. But it will not solve all our model-building problems nor will it obviate the need for careful reflection on how the data were obtained.

5.4 FURTHER READING

Applications

Hill and Jones (2014) use cross-validation to systematically evaluate a variety of competing empirical models of government repression. Grimmer and Stewart (2013) show how cross-validation and other out-of-sample methods are critical to the burgeoning field of machine learning, especially in the context of text analysis. Titiunik and Feher (2017) use randomization inference in examining the effects of term limits in the Arkansas legislature. Ho and Imai (2006) use randomization inference in the context of California's complicated candidate

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

randomization procedure to determine whether appearing on the first page of a ballot affected vote share in the 2003 recall election.

Previous Work

On the misinterpretation of confidence intervals and *p*-values, see Belia et al. (2005); Cumming et al. (2004); Hoekstra et al. (2014).

See Singal (2015) for a detailed description of the *Science* retraction of the LaCour and Greene study. Important recent articles about reproducibility in the social sciences have begun to appear (Benoit et al., 2016; Laitin and Reich, 2017; Miguel et al., 2014). Many political science journals now require publicly visible data repositories and more (Bueno de Mesquita et al., 2003; DA-RT, 2015; Gleditsch et al., 2003; King, 1995).

Many current recommendations on research reproducibility stem from Knuth's invention of *literate programming* (1984), which was a way of integrating textual documentation with computer programs. Gentleman and Temple Lang (2007) expanded this idea to statistical programming, and recently Xie (2015) further updated these ideas with the use of *markdown* and *pandoc* (MacFarlane, 2013).

Regarding the WTO-trade dispute, Park (2012) and Imai and Tingley (2012) revisit the Goldstein et al. (2007) findings in the context of other methodological discussions. Both show the GATT/WTO finding to be fragile. Ward et al. (2013) dispute the assumption of dyadic conditional independence in gravity models of international trade, arguing for models that incorporate higher-order network dependencies in the data.

Advanced Study

On the interpretation and (mis)use of *p*-values in model selection, see Freedman (1983); Gill (1999); Raftery (1995); Rozeboom (1960).

Model selection need not imply that we choose one "winner." Rather, in a Bayesian framework, we can average across models (Bartels, 1997; Raftery, 1995). This approach has received renewed interest in political science (Montgomery et al., 2012a,b; Nyhan and Montgomery, 2010).

Hastie et al. (2016) is the canonical text for cross-validation and applied machine learning. Arlot and Celisse (2010) provide a recent review of the state of the art. Stone (1977) shows that choosing models based on leave-one-out cross validation is asymptotically equivalent to minimizing the AIC, whereas Shao (1997) links *k-fold* cross-validation to the BIC. Hastie et al. (2008) observe that leave-one-out cross validation is approximately unbiased as an estimator of the expected prediction error. Markatou et al. (2005) presents some inferential approaches for cross-validation results in the linear regression case. Efron and Tibshirani (1998) gives a detailed treatment of the bootstrap

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press

and jackknife procedures; see also Davison and Hinkley (1997). Gerber and Green (2012) provide extensive discussion of randomization and permutation inference in political science.

Software Notes

There are many cross-validation and related routines in $\mathcal{R}$. The `cv.glm` function in the `boot` library (Canty and Ripley, 2016) produces cross-validation estimates for many of the models explored in this book. One disadvantage of this implementation is that it simply returns another single number summary, which, while informative, can be improved upon. The `crossval` function in the `bootstrap` package (Tibshirani and Leisch, 2017) requires some user manipulation but has more flexibility. Both `boot` and `bootstrap` enable bootstrap and jackknife resampling. `cvTools` (Alfons, 2012) and `caret` (Kuhn, 2016) contain cross-validation functionality as well. The `ri` (Aronow and Samii, 2012) package enables randomization and permutation inference.

https://doi.org/10.1017/9781316888544.008 Published online by Cambridge University Press