Appendix B

10 quick tips to improve your regression modeling

B.1 Think about variation and replication

Variation is central to regression modeling, and not just in the error term. If a regression is fit to different datasets, we can expect the relations between variables, and thus coefficients and causal effects, to vary. Fitting the same model to different datasets—the technique called the "secret weapon" in Section 10.9—can give a sense of variation across problems, which in many settings is more relevant to applications than the standard errors from a single study.

Replication ideally implies performing all the steps of a study from the start, not just increasing the sample size and collecting more data within an existing setting. Repeating an entire experiment can be seen as a way of capturing the variation corresponding to various aspects of data collection and measurement, not just the variation seen within a single study. And this is all in addition to the advantages of a fresh perspective and an avoidance of forking paths in data coding and analysis.

In some fields, such as psychology and cell biology, it can be easy and inexpensive to replicate an experiment from scratch. In mostly observational sciences, such as economics and political science, replication can be more difficult—we cannot re-run the international economy and political system so as to observe 10 more recessions or 20 more civil wars. For such problems, replication will need to be more indirect, for example analyzing local economic or political activity within different countries.

B.2 Forget about statistical significance

Forget about p-values, and forget about whether your confidence intervals exclude zero.

We say this for three reasons. First, if you discretize results based on significance tests, you are throwing away information. Measures of significance such as p-values are noisy, and it is misleading to treat an experiment as a success or failure based on a significance test. Second, in the sorts of problems we work on, there are no true zeroes. For example, religious attendance is associated with attitudes on economic as well as social issues, and both these correlations vary by state. And it does not interest us, for example, to test a model in which social class affects vote choice through party identification but not along a direct path. More generally, no true populations are identical, and anything that plausibly could have an effect will not have an effect that is exactly zero. Third, comparisons and effects vary by context, so there is typically little reason to focus on whether a confidence interval excludes zero, as if that would tell us something useful about future differences.

B.3 Graph the relevant and not the irrelevant

Graphing the fitted model

Graphing the data is fine (see Chapter 2), but it is also useful to graph the estimated model itself; see lots of examples of regression lines and curves throughout this book. A table of regression coefficients

does not give you the same sense as graphs of the model. This point should seem obvious but can be obscured in statistics textbooks that focus so strongly on plots for raw data and for regression diagnostics, forgetting the simple plots that help us understand a model.

Make many graphs

Real data are full of complexity, and regression models can be hard to understand. Try different visualizations of your data, and look at your model from different angles. Grids of plots can be helpful in visualizing many dimensions, as in Figure 2.10, and a series of graphs can tell a story in a way that would not be possible with a single image; see Section 2.4 for an example. Letting go of the search for the single perfect graph liberates you to learn more from your data and to understand and explain your findings better.

Don't graph the irrelevant

Are you sure you really want to make those quantile-quantile plots, influence diagrams, and all the other things that spew out of a standard regression package? What are you going to do with all that? Just forget about it and focus on something more important. A quick rule: any graph you show, be prepared to explain.

B.4 Interpret regression coefficients as comparisons

Regression coefficients are commonly called "effects," but this terminology can be misleading. From the data alone, a regression only tells us about comparisons between individuals, not about changes within individuals.

Taken as a data description, a linear regression coefficient is the modeled average difference in the outcome, comparing two individuals that differ in one predictor, while being at the same levels of all the other predictors. In the special case of a single binary predictor, the coefficient is a simple difference: the average of y for individuals with $x = 1$, minus the average with $x = 0$. For a continuous predictor, we should either scale it so that a difference of 1 unit is of interest, or we should multiply the coefficient by a reasonable change in that predictor.

There are several benefits to thinking of regressions as comparisons. First, the interpretation as a comparison is always available: it is a description of the model and does not require any causal assumptions. Second, we can consider more complicated regressions as built up from simpler models, starting with simple comparisons and adding adjustments. Third, the comparative interpretation also works in the special case of causal inference, where we can consider comparisons between the same individual receiving two different levels of a treatment.

Causal inference can be considered as an application of statistical modeling in which predictions are being made about potential outcomes and where we often summarize inferences as average causal effects, which represent average predictive comparisons under the model. In the special case of an ignorable treatment assignment with no interaction, the average treatment effect is the same as the coefficient of the treatment indicator; more generally, we need to work through the model's predictions to construct an average causal effect.

B.5 Understand statistical methods using fake-data simulation

Simulating fake data can take more effort than fitting models to real data. For example, to simulate from a linear model, you need to pick reasonable values of the coefficients and residual standard deviation, and you also need to specify all the predictors x for your fake data. Depending on the context, you might want some structure in these predictors, as demonstrated in the simulated midterm and final exams in Sections 6.5 and 16.6 and the simulated poststratification in Section 17.2.

The effort of creating and simulating from a fake world has several payoffs. First, the decisions made in constructing this world can be clarifying: how large a treatment effect could we realistically expect to see, how large are the interactions we want to consider, what might be the correlation between pre-test and post-test, and so forth. Asking these questions requires contact with the application in a way that can increase our understanding. Second, fake-data simulation is a general way to study the properties of statistical methods under repeated sampling. Put the simulation and inference into a loop and you can see how close the model's estimates and predictions are to the assumed true values. Here it can make sense to simulate from a process that includes features not included in the model you will use to fit the data—but, again, this can be a good thing in that it forces you to consider assumptions that might be violated. Third, fake-data simulation is a way to debug your code. With large samples or small data variance, your fitted model should be able to recover the true parameters; if it can't, you may have a coding problem or a conceptual problem, where your model is not doing what you think it is doing. It can help in such settings to plot the simulated data overlaid with the assumed and fitted models. Finally, fake-data simulation, or its analytical equivalent, is necessary if you want to design a new study and collect new data with some reasonable expectation of what you might find.

B.6 Fit many models

Think of a series of models, starting with the too-simple and continuing through to the hopelessly messy. Generally it's a good idea to start simple. Or start complex if you'd like, but prepare to quickly drop things out and move to the simpler model to help understand what's going on. Working with simple models is not a research goal—in the problems we work on, we usually find complicated models more believable—but rather a technique to help understand the fitting process.

A corollary is the need to be able to fit models relatively quickly. Realistically, you don't know what model you want, so it's rarely a good idea to run the computer overnight fitting a single model. At least, wait until you've developed some understanding by fitting many models.

When fitting multiple models, you should keep track of what models you have fit. This is important for the purpose of understanding your data and also to protect yourself from the biases that can arise when you have many possible ways of analyzing your data (many researcher degrees of freedom or forking paths). In such settings, it's best to record all that you've done and report results from all relevant models, rather than to pick just one and then overinterpret a story from it.

B.7 Set up a computational workflow

With a bit of work, you can make your computations faster and more reliable. This sounds like computational advice but is really about statistical workflow: if you can fit models faster, you can fit more models and better understand both data and model.

Data subsetting

Related to the "fit many models" approach are simple approximations that speed the computations. Computers are getting faster and faster—but models are getting more and more complicated! And so these general tricks might remain important. A simple and general trick is to break a large dataset into subsets and analyze each subset separately. For example, perform separate analyses within each region of a country, and then display in one plot the estimates and uncertainties corresponding to the different regions.

An advantage of working with data subsets is that computation is faster, allowing you to explore the data by trying out more models. In addition, separate analyses, when well chosen, can reveal variation that is of interest.

There are two disadvantages of working with data subsets. First, it can be inconvenient to partition

the data, perform separate analyses, and summarize the results. Second, the separate analyses may not be as accurate as would be obtained by putting all the data together in a single analysis. Moving forward, one can use multilevel modeling to get some of the advantages of subsetting without losing inferential efficiency or computational stability.

Fake-data and predictive simulation

When computations get stuck, or a model does not fit the data, it is usually not clear at first if this is a problem with the data, the model, or the computation. Fake-data and predictive simulation are effective ways of diagnosing problems. First use fake-data simulation to check that your computer program does what it is supposed to do, then use predictive simulation to compare the data to the fitted model's predictions.

B.8 Use transformations

Consider transforming just about every variable in sight:

- Logarithms of all-positive variables (primarily because this leads to multiplicative models on the original scale, which often makes sense).
- Standardizing based on the scale or potential range of the data (so that coefficients can be more directly interpreted and scaled); an alternative is to present results in scaled and unscaled forms.

Plots of raw data and residuals can also be informative when considering transformations, as with the log transformation for arsenic levels in Section 14.5.

In addition to univariate transformations, consider interactions and predictors created by combining inputs (for example, adding several related survey responses to create a total score). The goal is to create models that *could* make sense (and can then be fit and compared to data) and that include all relevant information.

B.9 Do causal inference in a targeted way, not as a byproduct of a large regression

Don't assume that a comparison or regression coefficient can be interpreted causally. If you are interested in causal inference, consider your treatment variable carefully and use the tools of Chapters 18–21 to address the challenges of balance and overlap when comparing treated and control units to estimate a causal effect and its variation across the population. Even if you are using a natural experiment or identification strategy, it is important to compare treatment and control groups and adjust for pre-treatment differences between them.

When considering several causal questions, it can be tempting to set up a single large regression to answer them all at once; however, in observational settings (including experiments in which certain conditions of interest are observational) this is not appropriate, as we discuss in Sections 19.5–19.7.

B.10 Learn methods through live examples

We have demonstrated the concepts, methods, and tools in this book through examples that are of interest to us, many of which came from our applied research. Consider these as a starting point: when learning these and more complicated ideas yourself, apply them to problems that you care about, gather data on these examples, and develop statistical understanding by simulating and graphing data from models that make sense to you. Know your data, your measurements, and your data-collection procedure. Be aware of your population of interest and the larger goals of your data collection and analysis. Understand the magnitudes of your regression coefficients, not just their signs. You will need this understanding to interpret your findings and catch things that go wrong.