Chapter 1

Overview

This book explores the challenges of building, understanding, and using predictive models. It turns out there are many subtleties involved even with simple linear regression—straight-line fitting. After a review of fundamental ideas of data, measurement, and statistics in the first five chapters of the book, we cover linear regression with one predictor and multiple predictors, and then logistic regression and other generalized linear models. We next consider various applications of regression involving generalization from the data at hand to larger questions involving sampling and causal inference. The book concludes with a taste of more advanced modeling ideas and appendixes on quick tips and getting started with computing.

This introductory chapter lays out the key challenges of statistical inference in general and regression modeling in particular. We present a series of applied examples to show how complex and subtle regression can be, and why a book-length treatment is needed, not just on the mathematics of regression modeling but on how to apply and understand these methods.

1.1 The three challenges of statistics

The three challenges of statistical inference are:

1. *Generalizing from sample to population*, a problem that is associated with survey sampling but actually arises in nearly every application of statistical inference;
2. *Generalizing from treatment to control group*, a problem that is associated with causal inference, which is implicitly or explicitly part of the interpretation of most regressions we have seen; and
3. *Generalizing from observed measurements to the underlying constructs of interest*, as most of the time our data do not record exactly what we would ideally like to study.

All three of these challenges can be framed as problems of prediction (for new people or new items that are not in the sample, future outcomes under different potentially assigned treatments, and underlying constructs of interest, if they could be measured exactly).

The key skills you should learn from this book are:

- *Understanding regression models*. These are mathematical models for predicting an outcome variable from a set of predictors, starting with straight-line fits and moving to various nonlinear generalizations.
- *Constructing regression models*. The regression framework is open-ended, with many options involving the choice of what variables to include and how to transform and constrain them.
- *Fitting regression models to data*, which we do using the open-source software R and Stan.
- *Displaying and interpreting the results*, which requires additional programming skills and mathematical understanding.

A central subject of this book, as with most statistics books, is *inference*: using mathematical models to make general claims from particular data.

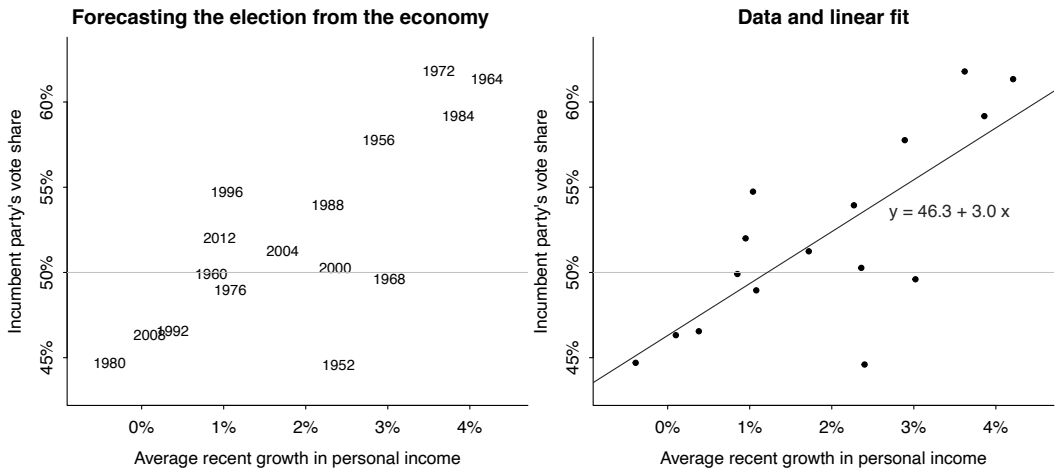


Figure 1.1: *Predicting elections from the economy: (a) the data, (b) the linear fit,* $y = 46.3 + 3.0x$.

1.2 Why learn regression?

Example: Elections and the economy

Regression is a method that allows researchers to summarize how predictions or average values of an *outcome* vary across individuals defined by a set of *predictors*. For example, Figure 1.1a shows the incumbent party's vote share in a series of U.S. presidential elections, plotted vs. a measure of economic growth in the period leading up to each election year. Figure 1.1b shows a linear regression fit to these data. The model allows us to predict the vote—with some uncertainty—given the economy and under the assumption that future elections are in some way like the past.

We do our computing in the open-source software R; see Appendix A for how to set up and use R on your computer. For this example, we first load in the data:[1]

```
hibbs <- read.table("hibbs.dat", header=TRUE)
```

Then we make a scatterplot:

```
plot(hibbs$growth, hibbs$vote, xlab="Average recent growth in personal income",
  ylab="Incumbent party's vote share")
```

Then we estimate the regression, $y = a + bx +$ error:[2]

```
M1 <- stan_glm(vote ~ growth, data=hibbs)
```

And then we add the fitted line to our graph:

```
abline(coef(M1), col="gray")
```

This produces something similar to Figure 1.1b.

To display the fitted model, we type `print(M1)`, which gives the following output:

```
            Median MAD_SD
(Intercept) 46.3    1.7
growth       3.0    0.7

Auxiliary parameter(s):
      Median MAD_SD
sigma 3.9    0.7
```

[1] Data and code for all examples in this book are at `www.stat.columbia.edu/~gelman/regression/`. Information for this particular example is in the folder `ElectionsEconomy` at this website.

[2] Section 1.6 introduces R code for least squares and Bayesian regression.

The first column shows estimates: 46.3 and 3.0 are the coefficients in the fitted line, $y = 46.3 + 3.0x$ (see Figure 1.1b). The second column displays uncertainties in the estimates using median absolute deviations (see Section 5.3). The last line of output shows the estimate and uncertainty of σ, the scale of the variation in the data unexplained by the regression model (the scatter of the points above and below from the regression line). In Figure 1.1b, the linear model predicts vote share to roughly an accuracy of 3.9 percentage points. We explain all the above code and output starting in Chapter 6.

If desired we can also summarize the fit in different ways, such as plotting residuals (differences between data and fitted model) and computing R^2, the proportion of variance explained by the model, as discussed in Chapter 11.

Some of the most important uses of regression are:

- *Prediction*: Modeling existing observations or forecasting new data. Examples with continuous or approximately continuous outcomes include vote shares in an upcoming election, future sales of a product, and health status in a medical study. Examples with discrete or categorical outcomes (sometimes referred to as classification) include disease diagnosis, victory or defeat in a sporting event, and individual voting decisions.
- *Exploring associations*: Summarizing how well one variable, or set of variables, predicts the outcome. Examples include identifying risk factors for a disease, attitudes that predict voting, and characteristics that make someone more likely to be successful in a job. More generally, one can use a model to explore associations, stratifications, or structural relationships between variables. Examples include associations between pollution levels and disease incidence, differential police stop rates of suspects by ethnicity, and growth rates of different parts of the body.
- *Extrapolation*: Adjusting for known differences between the *sample* (that is, observed data) and a population of interest. A familiar example is polling: real-world samples are not completely representative and so it is necessary to perform some adjustment to extrapolate to the general population. Another example is the use of data from a self-selected sample of schools to make conclusions about all the schools in a state. Another example would be using experimental data from a drug trial, along with background characteristics from the full population, to estimate the average effect of the drug in the population.
- *Causal inference*: Perhaps the most important use of regression is for estimating *treatment effects*. We define causal inference more carefully in Part 5 of this book; for now we'll just talk about comparing outcomes under treatment or control, or under different levels of a treatment. For example, in an education study, the outcome could be scores on a standardized test, the control could be an existing method of teaching, and the treatment could be some new innovation. Or in public health, the outcome could be incidence of asthma and the continuous treatment could be exposure to some pollutant. A key challenge of causal inference is ensuring that treatment and control groups are similar, on average, before exposure to the treatment, or else adjusting for differences between these groups.

In all these settings, it is crucial that the regression model have enough complexity to carry the required information. For example, if most of the participants in a drug trial are healthy and under the age of 70, but there is interest in estimating an average effect among the general elderly population, then it is important to include age and prior health condition as predictors in the regression model. If these predictors are not included, the model will simply not have enough information to allow the adjustment that we want to do.

1.3 Some examples of regression

To give a sense of the difficulties involved in applied regression, we briefly discuss some examples involving sampling, prediction, and causal inference.

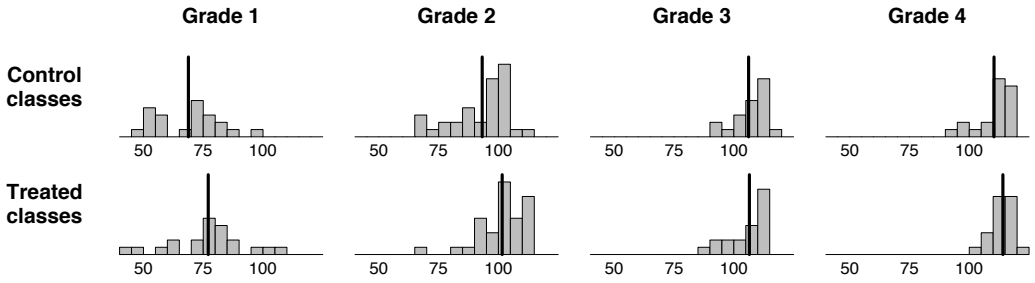


Figure 1.2 *Post-treatment classroom-average test scores from an experiment measuring the effect of an educational television program, The Electric Company, on children's reading abilities. The dark vertical line in each histogram shows the average for the corresponding group of classrooms.*

Estimating public opinion from an opt-in internet survey

Example: Xbox survey

In a research project with colleagues at Microsoft Research, we used a regression model to adjust a convenience sample to obtain accurate opinion monitoring, at a sharper time scale and at less expense than traditional survey methods. The data were from a novel and highly non-representative survey dataset: a series of daily voter intention polls for the 2012 presidential election conducted on the Xbox gaming platform with a total sample size of 750 148 interviews from 345 858 unique respondents. This is a characteristic problem of big data: a very large sample, relatively inexpensive to collect, but not immediately representative of the larger population. After adjusting the Xbox responses via multilevel regression and poststratification (MRP), we obtained estimates in line with forecasts from leading poll analysts, which were based on aggregating hundreds of traditional polls conducted during the election cycle.

The purpose of the Xbox project was not to forecast individual survey responses, nor was it to identify important predictors or causal inference. Rather, the goal was to learn about nationwide trends in public opinion, and regression allowed us to adjust for differences between sample and population, as we describe in Section 17.1; this required *extrapolation*.

A randomized experiment on the effect of an educational television program

Example: Electric Company experiment

A study was performed around 1970 to measure the effect of a new educational television program, The Electric Company, on children's reading abilities. An experiment was performed on children in grades 1–4 in two small cities in the United States. For each city and grade, the experimenters selected 10 to 20 schools, within each school selecting the two classes in the grade whose average reading test scores were lowest. For each pair, one of these classes was randomly assigned to continue with their regular reading course and the other was assigned to view the TV program. Each student was given a pre-test at the beginning of the school year and a post-test at the end.

Figure 1.2 shows post-test data for the control and treated classrooms in each grade.[3] Comparing the top and bottom row of graphs, we see what appears to be large beneficial effects in grades 1 and 2 with smaller effects for the higher grades, a plausible result given that most children in grades 3 and 4 already know how to read.

Further statistical analysis is required to adjust for differences in pre-treatment test scores between the two groups, and to assess uncertainty in the estimates. We return to this example in Section 19.2.

[3]Data and code for this example are in the folder `ElectricCompany`.

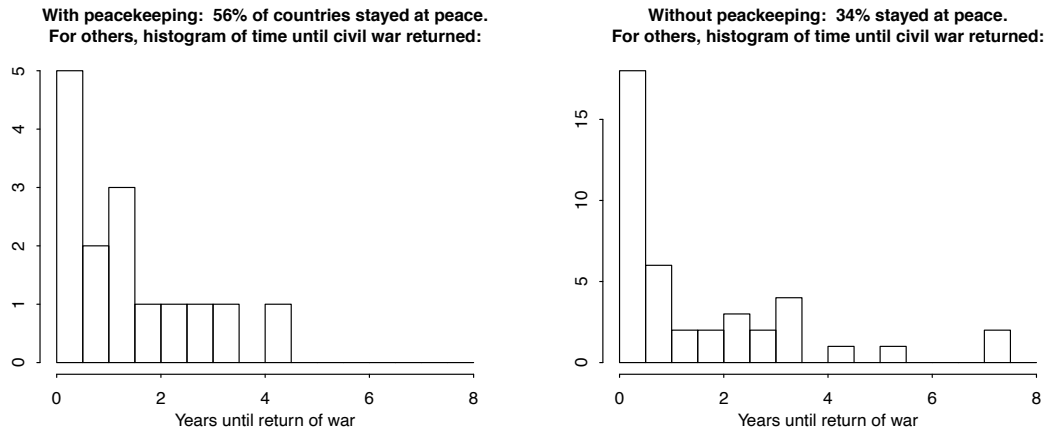


Figure 1.3 *Outcomes after civil war in countries with and without United Nations peacekeeping. The countries with peacekeeping were more likely to stay at peace and took on average about the same amount of time to return to war when that happened. However, there is a concern that countries with and without peacekeeping may differ in their pre-existing conditions; see Figure 1.4.*

Estimating the effects of United Nations peacekeeping, using pre-treatment variables to adjust for differences between treatment and control groups

Example: United Nations peace-keeping

Several years ago, political scientist Page Fortna conducted a study on the effectiveness of international peacekeeping. She analyzed data from countries that had been involved in civil wars between 1989 and 1999, comparing countries with and without United Nations peacekeeping. The outcome measure was whether there was a return to civil war in the country and, if so, the length of time until that happened. Data collection ended in 2004, so any countries that had not returned to civil war by the end of that year were characterized as being still at peace. The subset of the data summarized here contains 96 ceasefires, corresponding to 64 different wars.[4]

A quick comparison found better outcomes after peacekeeping: 56% stayed at peace, compared to 34% of countries without peacekeeping. When civil war did return, it typically came soon: the average lag between ceasefire and revival of the fighting was 17 months in the presence of peacekeeping and 18 months without. Figure 1.3 shows the results.

There is, however, a concern about *selection bias*: perhaps peacekeepers chose the easy cases. Maybe the really bad civil wars were so dangerous that peacekeepers didn't go to those places, which would explain the difference in outcomes.

To put this in more general terms: in this study, the "treatment"—peacekeeping—was not randomly assigned. In statistics jargon, Fortna had an *observational study* rather than an *experiment*, and in an observational study we must do our best to adjust for pre-treatment differences between the treatment and control groups.

Fortna adjusted for how bad off the country was before the peacekeeping-or-no-peacekeeping decision was made, using some objective measures of conditions within the country. The analysis was further complicated because in some countries we know the time until return to civil war, whereas in other countries all we can say is that civil war had not yet returned during the period of data collection. In statistics, this sort of incomplete data process is called "censoring," which does not mean that someone has refused to provide the data but rather that, due to the process of data collection, certain ranges of data cannot be observed: in this case, the length of time until resumption of civil war is inherently unknowable for the countries that remained at peace through the date at which data collection had concluded. Fortna addressed this using a "survival model," a complexity that we will ignore here. For our purposes here we summarize the combination of pre-treatment predictors as a

[4]Data and code for this example are in the folder `Peacekeeping`.

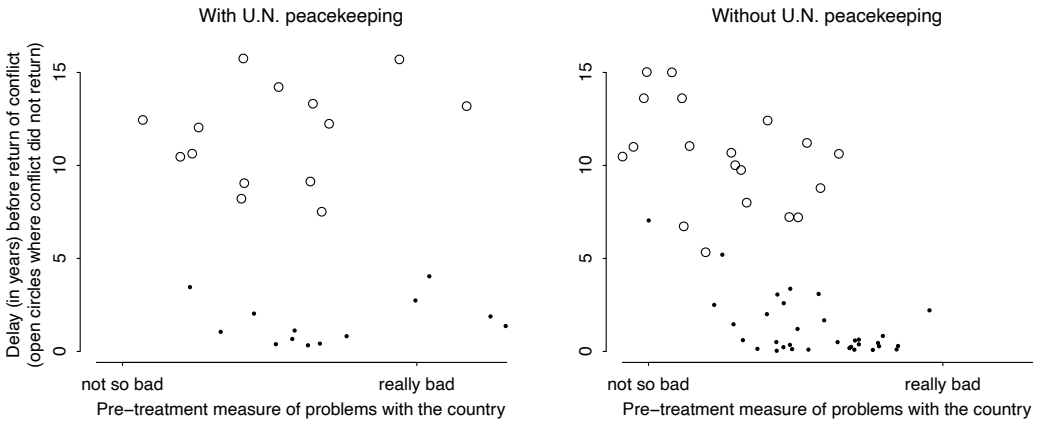


Figure 1.4 *Outcomes after civil war in countries with and without United Nations peacekeeping, plotted vs. a measure of how bad the situation was in the country. After adjusting for this pre-treatment variable, peacekeeping remains associated with longer periods without war.*

scalar "badness score," which ranges from 1.9 for the Yemen civil war in 1994 and 2.0 for India's Sikh rebellion in 1993, to the cases with the highest badness scores, 6.9 for Angola in 1991 and 6.5 for Liberia in 1990.

Figure 1.4 shows outcomes for treated and control countries as a function of badness score, with some missing cases where not all the variables were available to make that assessment. According to these data, peacekeeping was actually performed in tougher conditions, on average. As a result, adjusting for badness in the analysis (while recognizing that this adjustment is only as good as the data and model used to perform it) *increases* the estimated beneficial effects of peacekeeping, at least during the period of this study.

Estimating the effect of gun laws, and the difficulty of inference using regression with a large number of predictors

Example: gun-control policies

A leading medical journal published an article purporting to estimate the effects of a set of gun-control policies:

> Of 25 firearm laws, nine were associated with reduced firearm mortality, nine were associated with increased firearm mortality, and seven had an inconclusive association. . . . Projected federal-level implementation of universal background checks for firearm purchase could reduce national firearm mortality from 10.35 to 4.46 deaths per 100 000 people, background checks for ammunition purchase could reduce it to 1.99 per 100 000, and firearm identification to 1.81 per 100 000.

This study attempted causal inference using regression on the treatment variables, adjusting for background variables to account for differences between treatment and control groups. The model was also used to make forecasts conditional on different values of the predictors corresponding to various hypothetical policy implementations.

But we believe these results are essentially useless, for two reasons: First, in this sort of regression with 50 data points and 30 predictors and no prior information to guide the inference, the coefficient estimates will be hopelessly noisy and compromised by dependence among the predictors. Second, the treatments were observational, not externally applied. To put it another way, there are systematic differences between states that have implemented different gun-control policies, differences which will not be captured in the model's other predictors (state-level covariates or background variables), and there is no reason to think that the big differences in gun-related deaths between states are mostly attributable to these particular policies.

Comparing the peacekeeping and gun-control studies

Why do we feel satisfied with the conclusions drawn from the peacekeeping study but not with the gun-control study? In both cases, policy conclusions have been drawn from observational data, using regression modeling to adjust for differences between treatment and control groups. So what distinguishes these two projects?

One difference is that the peacekeeping study is focused, whereas the gun-control study is diffuse. It is more practical to perform adjustments when there is a single goal. In particular, in the peacekeeping study there was a particular concern that the United Nations might be more likely to step in when the situation on the ground was not so bad. The data analysis found the opposite, that peacekeeping appeared to be performed in slightly worse settings, on average. This conclusion is not airtight—in particular, the measure of badness is constructed based on particular measured variables and so it is possible that there are important unmeasured characteristics that would cause the adjustment to go the other way. Still, the pattern we see based on observed variables makes the larger story more convincing.

In contrast, it is hard to make much sense of the gun-control regression, for two reasons. First, the model adjusts for many potential causal variables at once: the effect of each law is estimated conditional on all the others being held constant, which is not realistic given that multiple laws can be changed at once, and there is no particular reason for their effects to add up in a simple manner. Second, the comparisons are between states, but states vary in many systematic ways, and it is not at all clear that a simple model can hope to adjust for the relevant differences. Yes, the comparisons in the peacekeeping project vary between countries, but the constructed badness measure seems more clearly relevant for the question being asked in that study.

We don't want to make too much of the differences between these studies, which ultimately are of degree and not of kind. Policies need to be evaluated in peacekeeping, gun control, and other areas, and it makes sense to use data and statistical analysis to aid in decision making. We see the peacekeeping study, for all its potential flaws, as a good example in that it starts with a direct comparison of data and then addresses a potential threat to validity in a focused way. In contrast, in the gun-control study the adjustments for pre-treatment variables seem less convincing, indeed fatally dependent on implausible model assumptions, which can happen when data are modeled in an unstructured way.

Indeed, statistical methods are part of the problem in that the gun-control claims would never have been publishable without the false sense of confidence supplied by regression analysis and statistical statements of uncertainty. Regression analysis was taken naively to be able to control for variation and give valid causal inference from observational data; and statistical significance and confidence intervals were taken naively to be able to screen out noise and deliver replicable statements about the world outside the data at hand. Put these together, and the result was that a respected medical journal was induced to publish strong and poorly supported conclusions taken from a messy set of aggregate trend data.

1.4 Challenges in building, understanding, and interpreting regressions

We can distinguish two different ways in which regression is used for causal inference: estimating a relationship and adjusting for background variables.

Regression to estimate a relationship of interest

Start with the simplest scenario of comparability of treatment and control groups. This condition can be approximated by *randomization*, a design in which people—or, more generally, experimental units—are randomly assigned to treatment or control groups, or through some more complicated

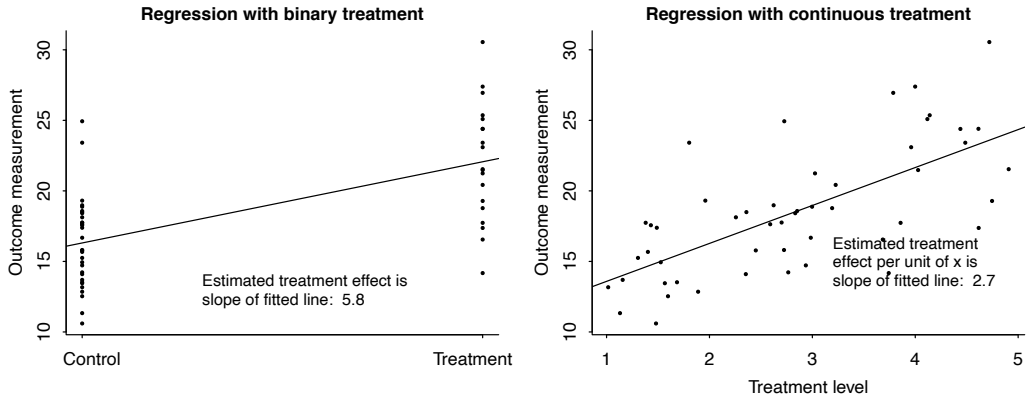


Figure 1.5 *Regression to estimate a causal effect with (a) simple comparison of treatment and control, or (b) a range of treatment levels.*

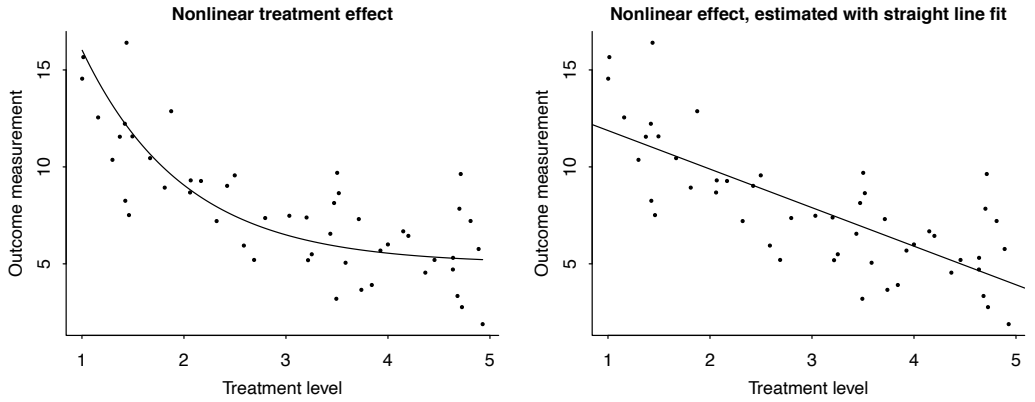


Figure 1.6 *(a) Hypothetical data in which the causal effect is a nonlinear function of the treatment level; (b) same data with a linear effect estimated. It is always possible to estimate a linear model, even if it does not fit the data.*

design that assures balance between the groups. In Part 5 of this book we discuss in detail the connections between treatment assignments, balance, and statistical analysis. For now, we just note that there are various ways to attain approximate comparability of treatment and control groups, and to adjust for known or modeled differences between the groups.

If we are interested in the effect of some treatment x on an outcome y, and our data come from a randomized or otherwise balanced experiment, we can fit a regression—that is, a model that predicts y from x, allowing for uncertainty.

Example: Hypothetical linear and nonlinear models

If x is binary ($x = 0$ for control or $x = 1$ for treatment), then the regression is particularly simple; see Figure 1.5a. But the same idea holds for a continuous predictor, as shown in Figure 1.5b.[5]

In this setting, we are assuming comparability of the groups assigned to different treatments, so that a regression analysis predicting the outcome given the treatment gives us a direct estimate of the causal effect. Again, we defer to Part 5 a discussion of what assumptions, both mathematical and practical, are required for this simple model to make sense for causal inference.

But setting those qualms aside, we can continue by elaborating the model in various ways to better fit the data and make more accurate predictions. One direction is to consider nonlinear modeling of a continuous treatment effect. Figure 1.5b shows a linear estimate, Figure 1.6a shows an example of an underlying nonlinear effect, and Figure 1.6b shows what happens if this curve is fit by a straight line.

Another important direction is to model *interactions*—treatment effects that vary as a function of

[5]Code for this example is in the folder `SimpleCausal`.

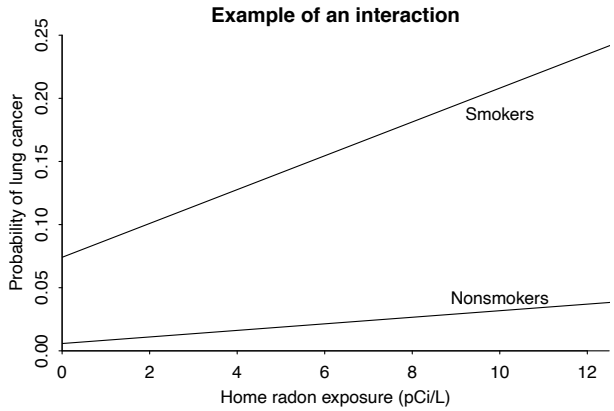


Figure 1.7 *Lifetime added risk of lung cancer for men, as a function of average radon exposure in picocuries per liter (pCi/L). The relation between cancer rate and radon is different for smokers and nonsmokers.*

Example: Radon, smoking, and lung cancer

other predictors in the model. For example, Figure 1.7 shows the estimated effects of radon gas on lung cancer rates for men. Radon causes cancer (or, to be more precise, it increases the probability of cancer), with this effect being larger among smokers than nonsmokers. In this model (which is a summary of the literature and is not the result of fitting to any single dataset), the effect of radon is assumed to be linear but with an interaction with smoking.

Interactions can be important and we discuss them throughout the book. If we care about the effect of a treatment, then we also care about how this effect varies. Such variation can be important for practical reasons—for example, in deciding how to allocate some expensive medical procedure, or who is at most risk from some environmental hazard—or for the goal of scientific understanding.

Regression to adjust for differences between treatment and control groups

In most real-world causal inference problems, there are systematic differences between experimental units that receive treatment and control. Perhaps the treated patients were sicker, on average, than those who received the control. Or, in an educational experiment, perhaps the classrooms that received the new teaching method had more highly motivated teachers than those that stuck with the old program. In such settings it is important to *adjust* for pre-treatment differences between the groups, and we can use regression to do this.

Example: Hypothetical causal adjustment

Figure 1.8 shows some hypothetical data with a fitted linear regression.[6] A key difference compared to Figures 1.5 and 1.6 is that in this case the variable on the x-axis is a pre-treatment predictor, *not* the treatment level.

Adjusting for background variables is particularly important when there is *imbalance* so that the treated and control groups differ on key pre-treatment predictors. Such an adjustment will depend on some model—in the example of Figure 1.8, the key assumptions are linearity and additivity—and a good analysis will follow up with a clear explanation of the consequences of any adjustments.

For example, the hypothetical analysis of Figure 1.8 could be summarized as follows:

> On average, the treated units were 4.8 points higher than the controls, $\bar{y} = 31.7$ under the treatment and $\bar{y} = 25.5$ for the controls. But the two groups differed in their pre-treatment predictor: $\bar{x} = 0.4$ for the treated units and $\bar{x} = 1.2$ for the controls. After adjusting for this difference, we obtained an estimated treatment effect of 10.0.

This estimated effect is necessarily model based, but the point of this example is that when there is imbalance between treated and controls on a key predictor, some adjustment should be done.

[6] Code for this example is in the folder `SimpleCausal`.

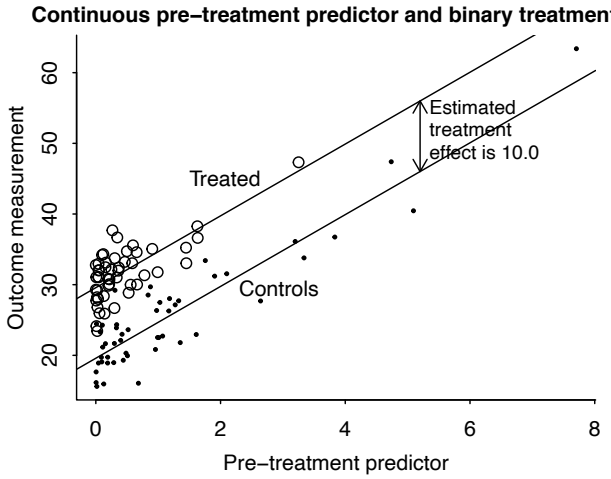


Figure 1.8 *Hypothetical data with a binary treatment and a continuous pre-treatment variable. Treated units are displayed with circles on the scatterplot, and controls are shown with dots. Overlaid is a fitted regression predicting the outcome given treatment and background variable, with the estimated treatment effect being the difference between the two lines.*

Interpreting coefficients in a predictive model

There can be challenges in interpreting regression models, even in the simplest case of pure prediction.

Example: Earnings and height

Consider the following model fit to survey data: earnings = 11 000 + 1500 ∗ (height − 60) + error, where annual earnings are measured in dollars, height is measured in inches, and the errors are mostly in the range ±22 000 (in mathematical terms, the errors have mean 0 and standard deviation 22 000). This is a prediction model, but it is close to useless for *forecasting* because the errors from the model are so large: it is not particularly helpful to predict someone's earnings as 25 000 with uncertainty 22 000. The regression is, however, somewhat useful for *exploring an association* in that it shows that the estimated slope is positive (with an associated standard error conveying uncertainty in that slope). As *sampling inference*, the regression coefficients can be interpreted directly to the extent that the people in the survey are a representative sample of the population of interest (adult residents of the United States in 1990); otherwise, it is best to include additional predictors in the model to bridge the gap from sample to population. Interpreting the regression as a *causal inference*—each additional inch of height gives you another $1500 a year in earnings—may feel natural. But such an interpretation is questionable because tall people and short people may differ in many other ways: height is not a randomly assigned treatment. Moreover, height is a problematic variable to consider causally in other ways that will be discussed later in the book. Rather, the best fit for this example might be the *exploring associations* category. Observing a pattern in data might prompt a researcher to perform further research to study reasons that taller people earn more than shorter people.

Building, interpreting, and checking regression models

Statistical analysis cycles through four steps:

- Model building, starting with simple linear models of the form, $y = a + bx +$ error and expanding through additional predictors, interactions, and transformations.
- Model fitting, which includes data manipulation, programming, and the use of algorithms to estimate regression coefficients and their uncertainties and to make probabilistic predictions.
- Understanding model fits, which involves graphics, more programming, and an active investigation of the (imperfect) connections between measurements, parameters, and the underlying objects of study.

- Criticism, which is not just about finding flaws and identifying questionable assumptions, but is also about considering directions for improvement of models. Or, if nothing else, limiting the claims that might be made by a naive reading of a fitted model.

The next step is return to the model-building step, possibly incorporating new data in this effort.

A challenge in serious applied work is how to be critical without being nihilistic, to accept that we can learn from statistical analysis—we can generalize from sample to population, from treatment to control, and from observed measurements to underlying constructs of interest—even while these inferences can be flawed.

A key step in criticizing research claims—and in understanding the limits to such criticisms—is to follow the steps that link the larger claims to the data and the statistical analysis. One weakness of the gun-control study discussed on page 8 is that conclusions were made regarding proposed changes in laws, but the comparisons were done across states, with no direct data on laws being implemented or removed. In contrast, the analysis of height and earnings was more clearly descriptive, not claiming or implying effects of policy changes. Another concern with the gun-control study was that the estimated effects were so large, up to fivefold reductions of the death rate. This is a sign of overinterpretation of noisy data, in this case taking existing variation among states and too eagerly attributing it to available factors. One might just as well try correlating firearm mortality with various laws on poultry processing and find similar correlations that could be given causal attributions. In contrast, the study of peacekeeping is more controlled—looking at one intervention rather than trying to consider 25 possibilities at once—and is more open about variation. The point of Figure 1.4 is not to claim that peacekeeping has some particular effect but rather to reveal that it was associated with a delay in return to civil war, in comparison to comparable situations in countries that did not have United Nations intervention.

No study is perfect. In the Xbox analysis, we used a non-representative sample to draw inference about the general population of voters. The Electric Company study was a controlled experiment, so that we have little worry about differences between treatment and control group, but one can be concerned about generalizing from an experimental setting to make claims about the effects of a national release of the television show. The common theme is that we should *recognize* challenges in extrapolation and then work to *adjust* for them. For the Xbox survey we used regression to model opinion as a function of demographic variables such as age, sex, and education where the sample differed from the population; the Electric Company data were analyzed separately for each grade, which gives some sense of variation in the treatment effect.

1.5 Classical and Bayesian inference

As statisticians, we spend much of our effort fitting models to data and using those models to make predictions. These steps can be performed under various methodological and philosophical frameworks. Common to all these approaches are three concerns: (1) what *information* is being used in the estimation process, (2) what *assumptions* are being made, and (3) how estimates and predictions are *interpreted*, in a classical or Bayesian framework. We investigate these in turn.

Information

The starting point for any regression problem is data on an outcome variable y and one or more predictors x. When data are continuous and there is a single predictor, the data can be displayed as a scatterplot, as in Figures 1.5 and 1.6. When there is one continuous predictor and one binary predictor, the data can be displayed as a scatterplot with two different symbols, as in Figure 1.8. More generally, it is not always possible to present all the data in a single display.

In addition to the data themselves, we typically know something about how they were collected. For example, in a survey, we can look at the survey questions, and we might know something about

how they were asked and where and when the interviews took place. If data are laboratory assays, we might have knowledge of the biases and variation of the measurements, and so on.

Information should also be available on what data were observed at all. In a survey, respondents may be a random sample from a well-defined population (for example, sampled by extracting random names from a list) or they could be a convenience sample, in which case we should have some idea which sorts of people were more or less likely to be reached. In an experiment, treatments might be assigned at random or not, in which case we will typically have some information on how assignment was done. For example, if doctors are choosing which therapies to assign to individual patients, we might be able to find out which therapies were considered by each doctor, and which patient characteristics were relevant in the assignment decisions.

Finally, we typically have *prior knowledge* coming from sources other than the data at hand, based on experience with previous, similar studies. We have to be careful about how to include such information. For example, the published literature tends to overestimate effect sizes, as there is a selection by which researchers are under pressure to find large and "statistically significant" results; see Section 4.5. There are settings, however, where local data are weak and it would be foolish to draw conclusions without using prior knowledge. We give an example in Section 9.4 of the association between parental characteristics and the sexes of their children.

Assumptions

There are three sorts of assumptions that are essential to any regression model of an outcome y given predictors x. First is the functional form of the relation between x and y: we typically assume linearity, but this is more flexible than it might seem, as we can perform transformations of predictors or outcomes, and we can also combine predictors in linear or nonlinear ways, as discussed in Chapter 12 and elsewhere in this book. Still, the choices of transformations, as well as the choice of which variables to include in the model in the first place, correspond to assumptions about the relations between the different variables being studied.

The second set of assumptions involves where the data came from: which potential observations are seen and which are not, who is surveyed and who does not respond, who gets which experimental treatment, and so on. These assumptions might be simple and strong—assuming random sampling or random treatment assignment—or weaker, for example allowing the probability of response in a survey to be different for men and women and to vary by ethnicity and education, or allowing the probability of assignment of a medical treatment to vary by age and previous health status. The strongest assumptions such as random assignment tend to be simple and easy to understand, whereas weaker assumptions, being more general, can also be more complicated.

The third set of assumptions required in any statistical model involves the real-world relevance of the measured data: are survey responses accurate, can behavior in a lab experiment be generalized to the outside world, are today's measurements predictive of what might happen tomorrow? These questions can be studied statistically by comparing the stability of observations conducted in different ways or at different times, but in the context of regression they are typically taken for granted. The interpretation of a regression of y on x depends also on the relation between the measured x and the underlying predictors of interest, and on the relation between the measured y and the underlying outcomes of interest.

Classical inference

The traditional approach to statistical analysis is based on summarizing the information in the data, not using prior information, but getting estimates and predictions that have well-understood statistical properties, low bias and low variance. This attitude is sometimes called "frequentist," in that the classical statistician is interested in the long-run expectations of his or her methods—estimates should be correct on average (unbiasedness), confidence intervals should cover the true parameter value 95% of the time (coverage). An important principle of classical statistics is *conservatism*: sometimes data

are weak and we can't make strong statements, but we'd like to be able to say, at least approximately, that our estimates are unbiased and our intervals have the advertised coverage. In classical statistics there should be a clear and unambiguous ("objective") path from data to inferences, which in turn should be checkable, at least in theory, based on their frequency properties.

Classical statistics has a lot to offer, and there's an appeal to summarizing the information from the data alone. The weaknesses of the classical approach arise when studies are small and data are indirect or highly variable. We illustrate with an example.

Example: Jamaica childhood intervention

In 2013, a study was released by a team of economists, reporting "large effects on the earnings of participants from a randomized intervention that gave psychosocial stimulation to stunted Jamaican toddlers living in poverty. The intervention consisted of one-hour weekly visits from community Jamaican health workers over a 2-year period . . . We re-interviewed the study participants 20 years after the intervention." The researchers estimated the intervention to have increased earnings by 42%, with a 95% confidence interval for the treatment effect which we reconstruct as [+2%, +98%]. That is, the estimate based on the data alone is that the treatment multiplies average earnings by a factor of 1.42, with a 95% interval of [1.02, 1.98] for this multiplicative factor; see Exercise 3.8.

The uncertainty here is wide, which is unavoidable given that the estimate is based on comparing earnings of only 127 children, who when they grow up have earnings that are highly variable. From the standpoint of classical inference, there's nothing wrong with that wide interval—if this same statistical procedure were applied over and over, to many different problems, the resulting 95% confidence intervals would contain the true parameter values 95% of the time (setting aside any imperfections in the data and experimental protocols). However, we know realistically that these intervals are more likely to be reported when they exclude zero, and therefore we would *not* expect them to have 95% coverage in the real world; see Exercises 5.8 and 5.9. And, perhaps more to the point, certain values in this interval are much more plausible than others: the treatment might well have an effect of 2% or even 0%, but it is highly unlikely for it to have an benefit of 98% and actually double people's earnings. We say this from prior knowledge, or general understanding. Indeed, we do not trust the estimate of 42%: if the study were to be replicated and we were offered a bet on whether the result would be greater or less than 42%, we would confidently bet on the "less than" side. This is not to say that the study is useless, just that not much can be learned about the effects of early childhood intervention from these data alone.

Bayesian inference

Bayesian inference is an approach to statistics which incorporates prior information into inferences, going beyond the goal of merely summarizing existing data. In the early childhood intervention example, for instance, one might start with the assumption that the treatment could make a difference but that the average effect would most likely be less than 10% in a positive or negative direction. We can use this information as a *prior* distribution that the multiplicative treatment effect is likely to be in the range [0.9, 1.1]; combining this with the data and using the rules of Bayesian inference, we get a 95% posterior interval of [0.92, 1.28], which ranges from an 8% negative effect of the intervention to a possible 28% positive effect; see Exercise 9.6 for details. Based on this Bayesian analysis, our best guess of the observed difference in a future replication study is much lower than 42%.

This simple example illustrates both the strength and the weaknesses of Bayesian inference. On the plus side, the analysis gives more reasonable results and can be used to make direct predictions about future outcomes and about the results of future experiments. On the minus side, an additional piece of information is required—the "prior distribution," which in this case represents the perhaps contentious claim that the effect of the treatment on earnings is probably less than 10%. For better or worse, we can't have one without the other: in Bayesian inference, the prior distribution represents the arena over which any predictions will be evaluated. In a world in which the treatment could plausibly double average earnings, the raw estimate of 1.42 and interval of [1.02, 1.98] yield reasonable predictions. But in a world in which such huge effects are implausible, we must adjust our expectations and predictions accordingly.

So, in that sense, we have a choice: classical inference, leading to pure summaries of data which can have limited value as predictions; or Bayesian inference, which in theory can yield valid predictions even with weak data, but relies on additional assumptions. There is no universally correct answer here; we should just be aware of our options.

There is also a practical advantage of the Bayesian approach, which is that all its inferences are probabilistic and thus can be represented by random simulations. For this reason, whenever we want to summarize uncertainty in estimation beyond simple confidence intervals, and whenever we want to use regression models for predictions, we go Bayesian. As we discuss in Chapter 9, we can perform Bayesian inference using noninformative or weakly informative priors and obtain results similar to classical estimates, along with simulation draws that can be used to express predictive uncertainty, or we can use informative priors if so desired.

To the extent that we have relevant information that is *not* in our model (for example, awareness of bias, selection on unmeasured characteristics, prior information on effect sizes, etc), then we have a duty to account for this as well as we can when interpreting our data summaries.

1.6 Computing least squares and Bayesian regression

We write R code to make graphs and compute data summaries, fit statistical models, and simulate fake data from theoretical or fitted models. We introduce code in the book as needed, with background on R in Appendix A.

In general we recommend using Bayesian inference for regression: if prior information is available, you can use it, and, if not, Bayesian regression with weakly informative default priors still has the advantage of yielding stable estimates and producing simulations that enable you to express inferential and predictive uncertainty (that is, estimates with uncertainties and probabilistic predictions or forecasts). For example, in the election model presented in Section 1.2, to which we return in Chapter 7, simulations from the fitted Bayesian model capture uncertainty in the estimated regression coefficients and allow us to compute probabilistic predictions for future elections conditional on assumptions about the election-year economy.

You can fit Bayesian regression in R using commands of the form,

```
fit <- stan_glm(y ~ x, data=mydata)
```

But some users of statistics will be unfamiliar or uncomfortable with Bayesian inference. If you are one of these people, or if you need to communicate with people who are more comfortable with classical statistics, you can fit least squares regression:

```
fit <- lm(y ~ x, data=mydata)
```

Finally, another concern about `stan_glm` is that it can go slowly for large problems. We can make it faster by running it in optimizing mode:

```
fit <- stan_glm(y ~ x, data=mydata, algorithm="optimizing")
```

For the examples in this book, datasets are small and speed is not really a concern, but it is good to be aware of this option in larger applications. When run in optimizing mode, `stan_glm` performs an approximate fit, but it still produces simulations that can again be used to summarize inferential and predictive uncertainty.

In summary, if you would prefer to avoid Bayesian inference, you can replace most of the instances of `stan_glm` in this book with `lm` for linear regression or `glm` for logistic and generalized linear models and get nearly identical results. Differences show up in an example in Section 9.5 with a strong prior distribution, examples of logistic and ordered logistic regressions with complete separation in Sections 14.6 and 15.5, our implementation of cross validation in Section 11.8, and various examples throughout the book where we use simulations to express uncertainty in estimates or predictions. It is also possible to fit least squares regression and get Bayesian uncertainties by running `stan_glm` with flat prior distributions, as we discuss in Section 8.4.

Bayesian and simulation approaches become more important when fitting regularized regression and multilevel models. These topics are beyond the scope of the present book, but once you are comfortable using simulations to handle uncertainty, you will be well situated to learn and work with those more advanced models.

1.7 Bibliographic note

We return to the election forecasting example in Chapter 7. We discuss the height/earnings regression, the Xbox survey, and the Electric Company experiment in detail in Chapters 6, 17, and 19, respectively. Further references on these topics appear in the bibliographic notes to those chapters.

Fortna (2008) discusses the United Nations peacekeeping study and its implications. The gun-control study appears in Kalesan et al. (2016); see also Gelman (2016a, b); the last of these references links to a reply by the authors to criticisms of that controversial paper.

For more on radon and cancer risk, see Lin et al. (1999). The Jamaica childhood intervention experiment comes from Gertler et al. (2013) and is discussed further by Gelman (2013, 2018).

1.8 Exercises

Data for examples and assignments in this and other chapters are at `www.stat.columbia.edu/~gelman/regression/`. See Appendix A for an introduction to R, the software you will use for computing.

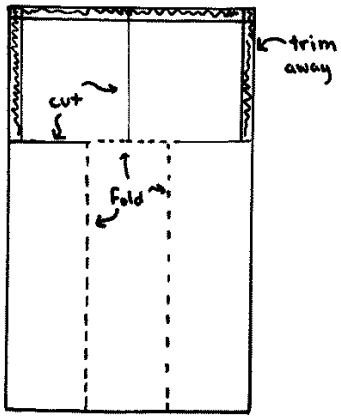
Example: Helicopter design

1.1 *From design to decision*: Figure 1.9 displays the prototype for a paper "helicopter." The goal of this assignment is to design a helicopter that takes as long as possible to reach the floor when dropped from a fixed height, for example 8 feet. The helicopters are restricted to have the general form shown in the sketch. No additional folds, creases, or perforations are allowed. The wing length and the wing width of the helicopter are the only two design parameters, that is, the only two aspects of the helicopter that can be changed. The body width and length must remain the same for all helicopters. A metal paper clip is attached to the bottom of the helicopter.
Here are some comments from previous students who were given this assignment:

> Rich creased the wings too much and the helicopters dropped like a rock, turned upside down, turned sideways, etc.
>
> Helis seem to react very positively to added length. Too much width seems to make the helis unstable. They flip-flop during flight.
>
> Andy proposes to use an index card to make a template for folding the base into thirds.
>
> After practicing, we decided to switch jobs. It worked better with Yee timing and John dropping. 3 – 2 – 1 – GO.

Your instructor will hand out 25 half-sheets of paper and 2 paper clips to each group of students. The body width will be one-third of the width of the sheets, so the wing width can be anywhere from $\frac{1}{6}$ to $\frac{1}{2}$ of the body width; see Figure 1.9a. The body length will be specified by the instructor. For example, if the sheets are U.S.-sized (8.5×5.5 inches) and the body length is set to 3 inches, then the wing width could be anywhere from 0.91 to 2.75 inches and the wing length could be anywhere from 0 to 5.5 inches.
In this assignment you can experiment using your 25 half-sheets and 10 paper clips. You can make each half-sheet into only one helicopter. But you are allowed to design sequentially, setting the wing width and body length for each helicopter given the data you have already recorded. Take a few measurements using each helicopter, each time dropping it from the required height and timing how long it takes to land.

(a) Record the wing width and body length for each of your 25 helicopters along with your time measurements, all in a file in which each observation is in its own row, following the pattern of `helicopters.txt` in the folder `Helicopters`, also shown in Figure 1.9b.



Helicopter_ID	width	length	time
1	1.8	3.2	1.64
1	1.8	3.2	1.74
1	1.8	3.2	1.68
1	1.8	3.2	1.62
1	1.8	3.2	1.68
2	1.8	3.2	1.62
2	1.8	3.2	1.65
2	1.8	3.2	1.66
2	1.8	3.2	1.63
2	1.8	3.2	1.66

Figure 1.9 *(a) Diagram for making a "helicopter" from half a sheet of paper and a paper clip. The long segments on the left and right are folded toward the middle, and the resulting long 3-ply strip is held together by a paper clip. One of the two segments at the top is folded forward and the other backward. The helicopter spins in the air when dropped. (b) Data file showing flight times, in seconds, for 5 flights each of two identical helicopters with wing width 1.8 inches and wing length 3.2 inches dropped from a height of approximately 8 feet. From Gelman and Nolan (2017).*

(b) Graph your data in a way that seems reasonable to you.

(c) Given your results, propose a design (wing width and length) that you think will maximize the helicopter's expected time aloft. It is not necessary for you to fit a formal regression model here, but you should think about the general concerns of regression.

The above description is adapted from Gelman and Nolan (2017, section 20.4). See Box, Hunter, and Hunter (2005) for a more advanced statistical treatment of this sort of problem.

1.2 *Sketching a regression model and data*: Figure 1.1b shows data corresponding to the fitted line $y = 46.3 + 3.0x$ with residual standard deviation 3.9, and values of x ranging roughly from 0 to 4%.

(a) Sketch hypothetical data with the same range of x but corresponding to the line $y = 30 + 10x$ with residual standard deviation 3.9.

(b) Sketch hypothetical data with the same range of x but corresponding to the line $y = 30 + 10x$ with residual standard deviation 10.

1.3 *Goals of regression*: Download some data on a topic of interest to you. Without graphing the data or performing any statistical analysis, discuss how you might use these data to do the following things:

(a) Fit a regression to estimate a relationship of interest.

(b) Use regression to adjust for differences between treatment and control groups.

(c) Use a regression to make predictions.

1.4 *Problems of statistics*: Give examples of applied statistics problems of interest to you in which there are challenges in:

(a) Generalizing from sample to population.

(b) Generalizing from treatment to control group.

(c) Generalizing from observed measurements to the underlying constructs of interest.

Explain your answers.

1.5 *Goals of regression*: Give examples of applied statistics problems of interest to you in which the goals are:

(a) Forecasting/classification.

(b) Exploring associations.

(c) Extrapolation.

(d) Causal inference.

Explain your answers.

1.6 *Causal inference*: Find a real-world example of interest with a treatment group, control group, a pre-treatment predictor, and a post-treatment predictor. Make a graph like Figure 1.8 using the data from this example.

1.7 *Statistics as generalization*: Find a published paper on a topic of interest where you feel there has been insufficient attention to:

(a) Generalizing from sample to population.

(b) Generalizing from treatment to control group.

(c) Generalizing from observed measurements to the underlying constructs of interest.

Explain your answers.

1.8 *Statistics as generalization*: Find a published paper on a topic of interest where you feel the following issues *have* been addressed well:

(a) Generalizing from sample to population.

(b) Generalizing from treatment to control group.

(c) Generalizing from observed measurements to the underlying constructs of interest.

Explain your answers.

1.9 *A problem with linear models*: Consider the helicopter design experiment in Exercise 1.1. Suppose you were to construct 25 helicopters, measure their falling times, fit a linear model predicting that outcome given wing width and body length:

$$\text{time} = \beta_0 + \beta_1 * \text{width} + \beta_2 * \text{length} + \text{error},$$

and then use the fitted model time $= \beta_0 + \beta_1 * \text{width} + \beta_2 * \text{length}$ to estimate the values of wing width and body length that will maximize expected time aloft.

(a) Why will this approach fail?

(b) Suggest a better model to fit that would not have this problem.

1.10 *Working through your own example*: Download or collect some data on a topic of interest of to you. You can use this example to work though the concepts and methods covered in the book, so the example should be worth your time and should have some complexity. This assignment continues throughout the book as the final exercise of each chapter. For this first exercise, discuss your applied goals in studying this example and how the data can address these goals.