

---

# 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.

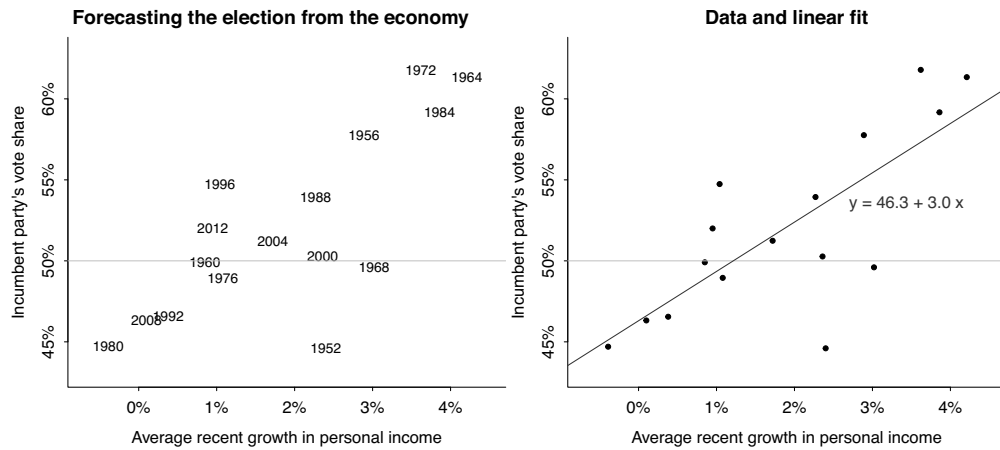


Figure 1.1: Predicting elections from the economy: (a) the data, (b) the linear fit,  $y = 46.3 + 3.0x$ .

## 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:<sup>1</sup>

```
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,  $y = a + bx + \text{error}$ :<sup>2</sup>

```
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
```

<sup>1</sup>Data and code for all examples in this book are at [www.stat.columbia.edu/~gelman/regression/](http://www.stat.columbia.edu/~gelman/regression/). Information for this particular example is in the folder ElectionsEconomy at this website.

<sup>2</sup>Section 1.6 introduces R code for least squares and Bayesian regression.

### 1.3. SOME EXAMPLES OF REGRESSION

5

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  $\sigma$ , 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.

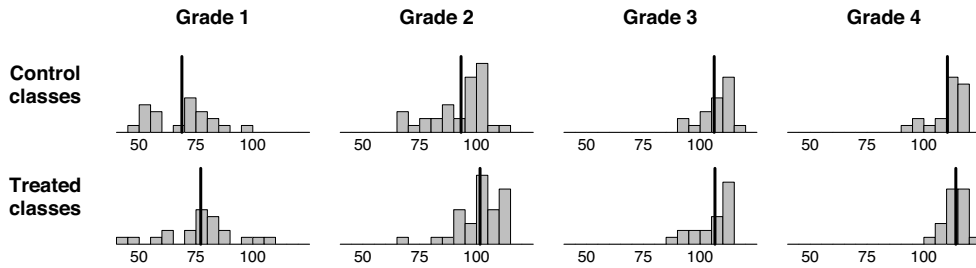


Figure 1.2 Post-treatment classroom-average test scores from an experiment measuring the effect of an educational television program, *The Electric Company*, on children's reading abilities. The dark vertical line in each histogram shows the average for the corresponding group of classrooms.

### 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.<sup>3</sup> 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.

<sup>3</sup>Data and code for this example are in the folder `ElectricCompany`.

## 1.3. SOME EXAMPLES OF REGRESSION

7



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.<sup>4</sup>

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

<sup>4</sup>Data and code for this example are in the folder Peacekeeping.

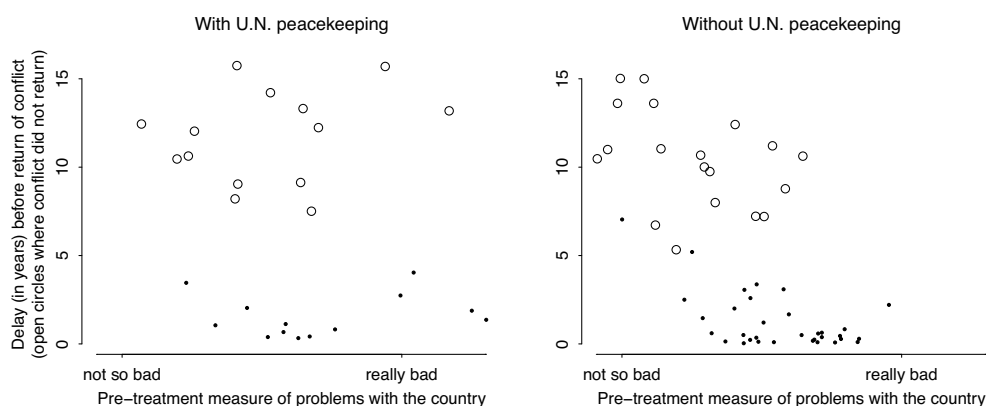


Figure 1.4 Outcomes after civil war in countries with and without United Nations peacekeeping, plotted vs. a measure of how bad the situation was in the country. After adjusting for this pre-treatment variable, peacekeeping remains associated with longer periods without war.

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

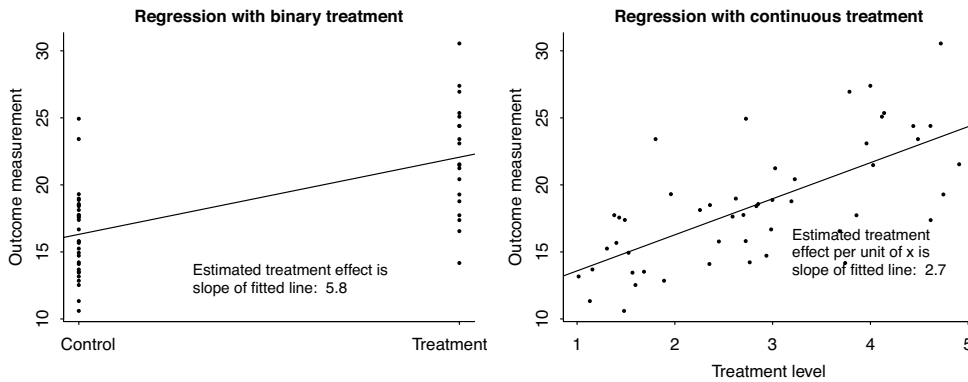


Figure 1.5 Regression to estimate a causal effect with (a) simple comparison of treatment and control, or (b) a range of treatment levels.

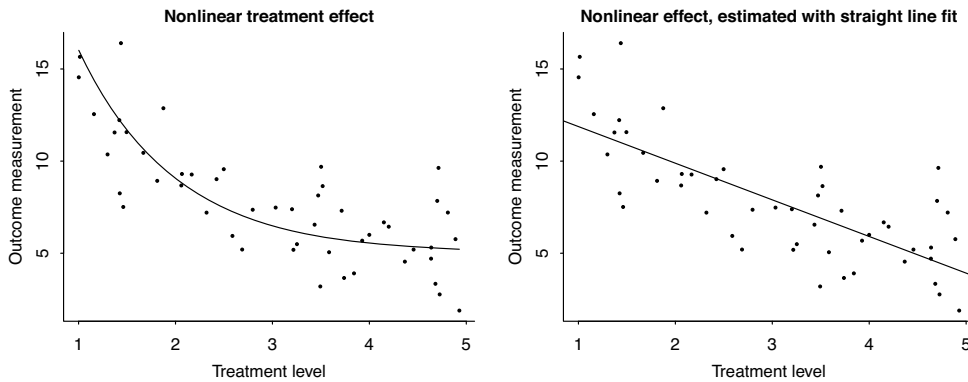


Figure 1.6 (a) Hypothetical data in which the causal effect is a nonlinear function of the treatment level; (b) same data with a linear effect estimated. It is always possible to estimate a linear model, even if it does not fit the data.

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.<sup>5</sup>

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

<sup>5</sup>Code for this example is in the folder SimpleCausal.



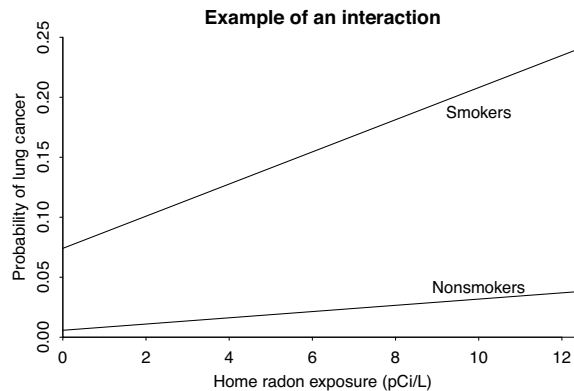


Figure 1.7 Lifetime added risk of lung cancer for men, as a function of average radon exposure in picocuries per liter (pCi/L). The relation between cancer rate and radon is different for smokers and nonsmokers.

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.<sup>6</sup> 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.

<sup>6</sup>Code for this example is in the folder SimpleCausal.

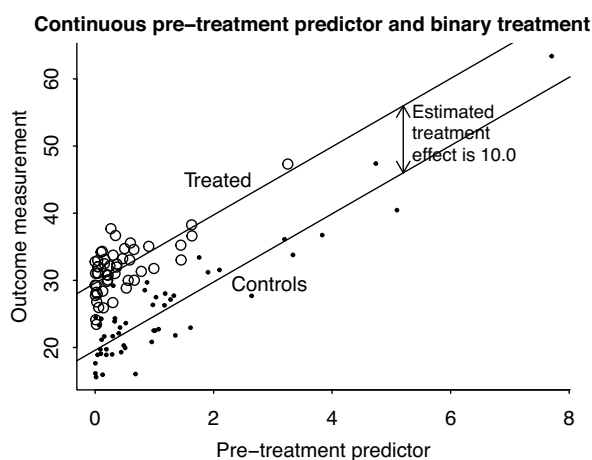


Figure 1.8 Hypothetical data with a binary treatment and a continuous pre-treatment variable. Treated units are displayed with circles on the scatterplot, and controls are shown with dots. Overlaid is a fitted regression predicting the outcome given treatment and background variable, with the estimated 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:  $\text{earnings} = 11\,000 + 1500 * (\text{height} - 60) + \text{error}$ , where annual earnings are measured in dollars, height is measured in inches, and the errors are mostly in the range  $\pm 22\,000$  (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,  $y = a + bx + \text{error}$  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.