Chapter 2

Data and measurement

In this book, we'll be fitting lines (and some curves) to data, making comparisons and predictions and assessing our uncertainties in the resulting inferences. We'll discuss the assumptions underlying regression models, methods for checking these assumptions, and directions for improving fitted models. We'll discuss the challenges of extrapolating from available data to make causal inferences and predictions for new data, and we'll use computer simulations to summarize the uncertainties in our estimates and predictions.

Before fitting a model, though, it is a good idea to understand where your numbers are coming from. The present chapter demonstrates through examples how to use graphical tools to explore and understand data and measurements.

2.1 Examining where data come from

Example: Human Development Index

Figure 2.1 went viral on the web a few years ago. The map compares the 50 states and Washington, D.C., in something called the Human Development Index (HDI), which had previously been used to compare different countries in public health measures. The coding of the map is kind of goofy: the states with the three lowest values are Louisiana at 0.801, West Virginia at 0.800, and Mississippi at 0.799, but their shading scheme makes Mississippi stand out.

But we have bigger concerns than that. Is Alaska really so developed as all that? And what's up with Washington, D.C., which, according to the report, is ranked at #4, behind only Connecticut, Massachusetts, and New Jersey?

Time to look behind the numbers. From the published report, the HDI combines three basic dimensions:

- Life expectancy at birth, as an index of population health and longevity.
- Knowledge and education, as measured by the adult literacy rate (with two-thirds weighting) and the combined primary, secondary, and tertiary gross enrollment ratio (with one-third weighting).
- Standard of living, as measured by the natural logarithm of gross domestic product (GDP) per capita at purchasing power parity (PPP) in U.S. dollars.

Now we can see what's going on. There is not much variation by state in life expectancy, literacy, or school enrollment. Sure, Hawaiians live a few years longer than Mississippians, and there are some differences in who stays in school, but by far the biggest differences between states, from these measures, are in GDP. The average income in Connecticut is twice that of Mississippi. And Washington, D.C., ranks high because its residents have a high average income.

To check out the relation between HDI and income, we loaded in the tabulated HDI numbers and plotted them versus some historical data on average income by state.[1] Figure 2.2a shows the result. The pattern is strong but nonlinear. Figure 2.2b plots the ranks and reveals a clear pattern, with most of the states falling right on the 45-degree line and a high correlation between the two rankings. We were surprised the correlation isn't higher—and surprised the first scatterplot above is

[1] Data and code for this example are in the folder `HDI`.

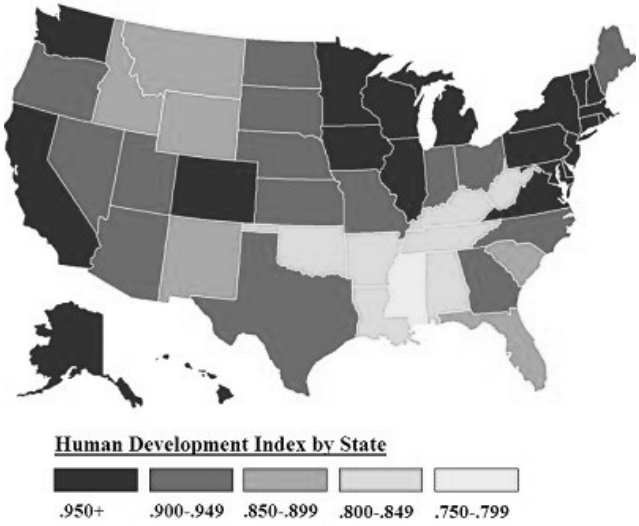

Figure 2.1 *Map that appeared on the internet of the so-called "Human Development Index," ranking the 50 states and Washington, D.C., from PlatypeanArchcow (2009).*

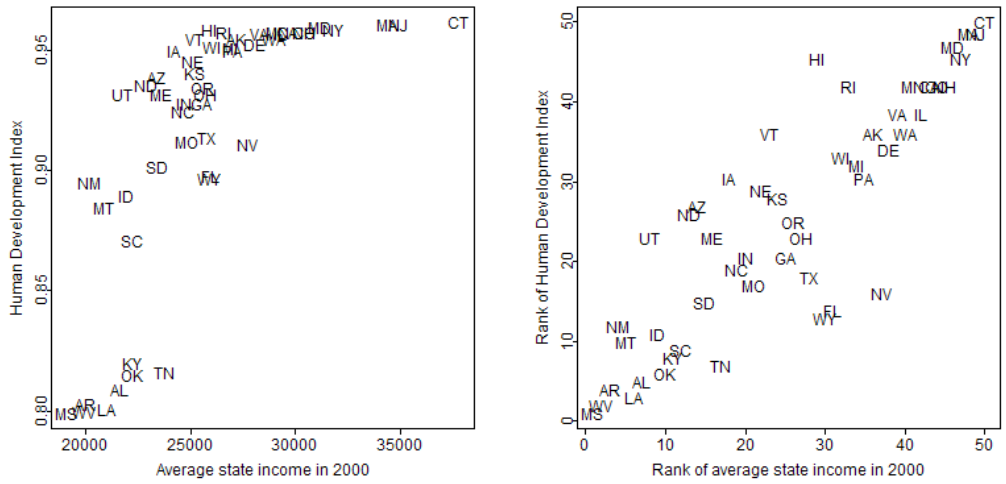

Figure 2.2 *Graphing the Human Development Index versus average income by state: (a) scatterplot of the data, (b) scatterplot of ranks.*

so nonlinear—but, then again, we're using state income rather than GDP, so maybe there's something going on with that. No, the logarithmic transformation is not what's doing this, at least not if you're logging income as is stated in the report. Logging stretches out the lower end of the scale a bit but does not change the overall pattern of the plot. The income values don't have enough dynamic range for the log transformation to have much effect.

Or maybe more is going on than we realize with those other components. If anyone is interested in following up on this, we suggest looking into South Carolina and Kentucky, which are so close in average income and so far apart on the HDI; see Figure 2.2a.

In any case, the map in Figure 2.1 is pretty much a map of state income with a mysterious transformation and a catchy name. The relevance of this example is that we were better able to understand the data by plotting them in different ways.

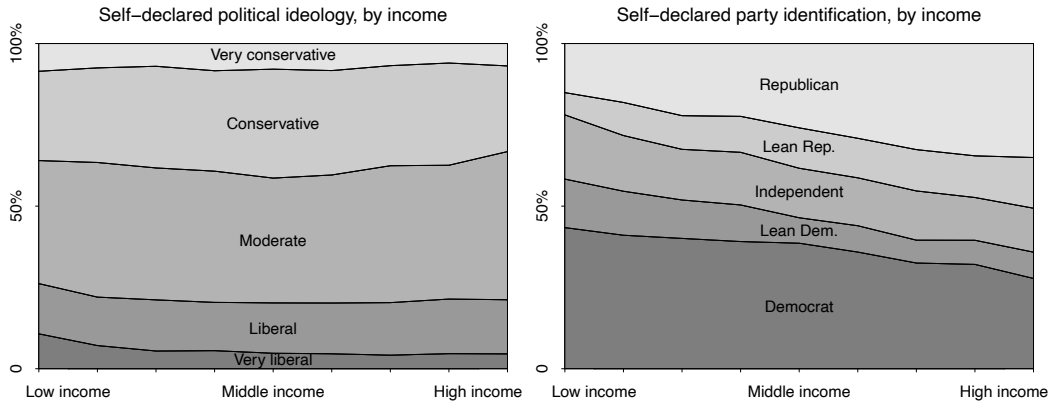

Figure 2.3 *Distribution of (a) political ideology and (b) party identification, by income, from a survey conducted during the 2008 U.S. election campaign.*

Details of measurement can be important

Example: Political ideology and party identification

In American politics, there are two major parties, and most voters fall in an ideological spectrum ranging from left, or liberal, to right, or conservative. The split between Democrats and Republicans roughly aligns with the division between liberal and conservative.

But these two scales of partisanship and ideology are not identical.[2] Figure 2.3a shows that the proportion of political liberals, moderates, and conservatives is about the same for all income levels. In contrast, Figure 2.3b shows a strong relation between income and Republican partisanship, at least as of 2008, when these survey data were gathered. Party identification and political ideology were each measured on a five-point scale running from left to right, but, as the graphs show, there are clear differences between the two variables.

How does Figure 2.3 relate to the general themes of our book? Regression is a way to summarize and draw inferences from data. As such, conclusions from regressions will depend on the quality of the data being analyzed and the relevance of these data to questions of interest. The partisanship and ideology example is a reminder that even very similar measures can answer quite different questions.

Unfortunately, gaps between measurement and reality are a general problem in scientific research and communication. For example, Temple University's medical school issued a press release entitled "Extra-virgin olive oil preserves memory & protects brain against Alzheimer's"—but the actual study was performed on mice and had no direct connection with dementia or Alzheimer's disease. The claim thus lacks external validity (see Section 2.2). This sort of leap happens all the time. In some sense it is necessary—lab experimentation precedes clinical trials—but we should be open and aware of what we know.

2.2 Validity and reliability

We discuss the important issue of measurement for two reasons. First, we need to understand what our data actually mean. We have looked at ways to visualize data and extract information. But if we do not know what the data actually represent, then we cannot extract the right information.

Data analysis reaches a dead end if we have poor data. There are some measurement problems that no amount of fixing and adjusting can solve. In Section 1.3 we discussed how we made adjustments to the Xbox polling data to account for differences between sample and population. But if we had asked our respondents the wrong question, or had nor recorded key background variables that could be used for the adjustment, then there would have been no easy fix.

[2] Data and code for this example are in the folder Pew.

The second reason for discussing measurement is that learning about accuracy, reliability, and validity will set a foundation for understanding variance, correlation, and error, which will all be useful in setting up linear models in the forthcoming chapters.

Most of us don't think very much about measurement on a day-to-day basis, primarily because we take for granted the measures we work with, and even where we know there are some issues with precision, the precision we have is usually good enough for our purposes. So we have no trouble talking about the temperature outside, the weight of groceries, the speed of a car, etc. We take for granted the correspondence between the numbers and the "thing" that we are measuring. And we're usually not worried about the precision—we don't need temperature to the nearest half degree, or our car speed to six decimal places.

This is all dependent on what we are measuring and what our proposed inferences are. A scale that measures weight to an accuracy of 1 kilogram is fine for most purposes of weighing people, great for weighing elephants, and terrible for weighing medicine at a pharmacy. The property of being precise enough is a combination of the properties of the scale and what we are trying to use it for.

In social science, the way to measure what we are trying to measure is not as transparent as it is in everyday life. Sometimes this is because what we want to measure is "real" and well defined, but difficult to actually count. Examples include counting the number of immigrants, or measuring daily intake of food in uncontrolled conditions.

Other times, the thing we are trying to measure is pretty straightforward, but a little bit fuzzy, and the ways to tally it up aren't obvious, for example, counting the number of people in your neighborhood you know or trust, or counting the number of vocabulary words you know.

Sometimes we are trying to measure something that we all agree has meaning, but which is subjective for every person and does not correspond to a "thing" we can count or measure with a ruler. Examples include attitudes, beliefs, intentions to vote, and customer satisfaction. In all these cases, we share an understanding of what we are talking about; it is deeply embedded in our language and understanding that people have opinions about things and feelings. But attitudes are private; you can't just weigh them or measure their widths. And that also means that to probe them you have to invent some kind of measure such as, "Tell us on a scale of 0 to 100 how much you enjoyed the service you got today?" The relative answer matters, but we could have asked on a scale of 1 to 3, or for that matter 300 too 500. We just hope that people can be sincere when they answer and that they use the scale the same way. These concerns arise if you are designing your own study or when analyzing data collected by others.

It can be helpful to take multiple measurements on an underlying construct of interest. For example, in a class evaluation survey, students are typically asked several questions about the quality of an instructor and a course. And various health conditions are measured using standard batteries of questions. For example, the Beck Depression Inventory includes 21 items, each of which is given a score from 0 to 3, and then these are added to get a total from 0 to 63.

A measure can be useful for some purposes but not others. For example, in public health studies, a "never smoker" is typically defined as someone who has smoked fewer than 100 cigarettes in his or her lifetime, which generally seems like a reasonable definition when studying adult behavior and health. But in a study of adolescents, it would be mistaken to put a youth who has smoked 90 cigarettes in the same "never smoker" category as a youth who has smoked zero or one or two cigarettes.

Validity

A measure is *valid* to the degree that it represents what you are trying to measure. It's easy to come up with negative examples. A written test is not a valid measure of musical ability. There is a vast gap between the evidence and what we want to make inferences about.

Similarly, asking people how satisfied they are with some government service might not be considered a valid measure of the effectiveness of that service. Valid measures are ones in which there is general agreement that the observations are closely related to the intended construct.

We can define the *validity* of a measuring process as the property of giving the right answer on

average across a wide range of plausible scenarios. To study validity in an empirical way, ideally you want settings in which there is an observable true value and multiple measurements can be taken.

In social science, validity can be difficult to assess. When the truth is not available, measurements can be compared to expert opinion or another "gold standard" measurement. For instance, a set of survey questions designed to measure depression in a new population could be compared to the opinion of an experienced psychiatrist for a set of patients, and it can also be compared to a well-established depression inventory.

Reliability

A *reliable* measure is one that is precise and stable. If we make a measurement, and then we have occasion to do it again, we would hope that the value would not move (much). Put another way, the variability in our sample is due to real differences among people or things, and not due to random error incurred during the measurement process.

For example, consider a test that is given twice to the same group of students. We could use the correlation between the scores across the two administrations of the test to help understand the extent to which the test *reliably* measures the given construct.

Another approach would be to have different raters administer the same measure in the same context. For instance, we could compare the responses on a measure of classroom quality across raters who observed the same classroom at the same time. Or we could compare judges' ratings of proficiency of gymnasts' performance of a given skill based on the same demonstration of that skill. This is referred to as inter-rater reliability.

Sample selection

Yet another feature of data quality is *selection*, the idea that the data you see can be a nonrepresentative sample of a larger population that you will not see. For example, suppose you are interested in satisfaction with a city's public transit system, so you interview people who ride the buses and trains; maybe you even take some measurements such as travel times or percentage of time spent sitting or standing. But there is selection: you only include people who have chosen to ride the bus or train. Among those excluded are those who have chosen not to ride the bus or the train because they are unhappy with those services.

In addition to this sort of selection bias based on who is included in the dataset, there are also biases from nonresponse to particular survey items, partially observed measurements, and choices in coding and interpretation of data. We prefer to think about all these measurement issues, including validity, reliability, and selection, in the context of larger models connecting measurements to underlying relationships of interest.

2.3 All graphs are comparisons

As demonstrated throughout this book, we can learn a lot by looking at data with an open mind. We present three quick examples here. In the larger context of workflow, we go back and forth between data exploration, modeling, inference, and model building, and each step requires its own tools.

Simple scatterplots

Example: Health spending and lifespan

Figure 2.4 shows some data on health spending and life expectancy, revealing that the United States spends much more per person than any other country without seeing any apparent benefit in lifespan.[3]

Here is R code to plot the data:

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

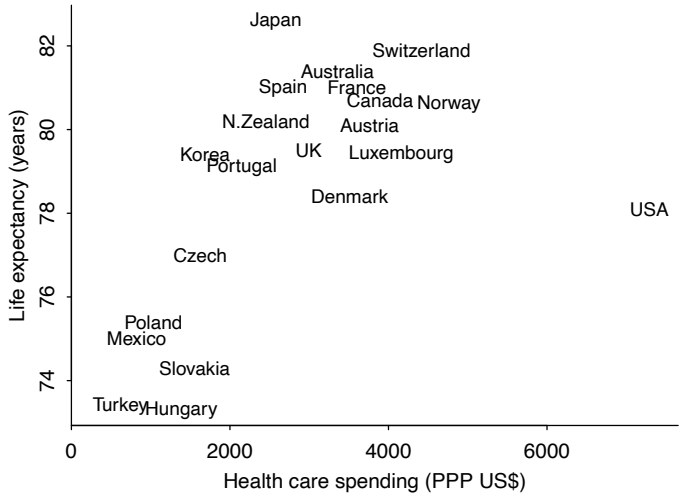

Figure 2.4 *Health care spending and life expectancy in several countries. This scatterplot shows two things: the generally positive correlation between spending and lifespan, and the extreme position of the United States.*

```
health <- read.table("healthdata.txt", header=TRUE)
country <- rownames(health)
plot(health$spending, health$lifespan, type="n")
text(health$spending, health$lifespan, country)
```

To make the graph just as displayed in Figure 2.4, further commands are required, and these are available on our website, but the code here gives the basic idea.

The graph shows the exceptional position of the United States and also shows the relation between spending and lifespan in the other countries.

Displaying more information on a graph

You can make as many plots as you want (or as your patience allows), but it is useful to think a bit about each plot, just as it is useful to think a bit about each model you fit.

The points within a scatterplot correspond to the unit of analysis in your study. At least in theory, you can display five variables easily with a scatterplot: x position, y position, symbol, symbol size, and symbol color. A two-way grid of plots allows two more dimensions, bringing the total number of variables potentially displayed to seven.

Example: Redistricting and partisan bias

We demonstrate some of the virtues of a rich visual description of data and estimates with Figure 2.5, a graph from our applied research that was central to the discovery and presentation of our key finding. The scatterplot in question displays three variables, conveyed by x position, y position, and symbol, a comparison of treatments to control with a before and after measurement. In this case, the units are state legislative elections, and the plot displays estimated partisan bias (a measure of the extent to which the drawing of district boundaries favors one party or the other) in two successive election years. The "treatments" are different kinds of redistricting plans, and the "control" points (indicated by dots on the figure) represent pairs of elections with no intervening redistricting. We display all the data and also show the regression lines on the same scale. As a matter of fact, we did not at first think of fitting nonparallel regression lines; it was only after making the figure and displaying parallel lines that we realized that nonparallel lines (that is, an interaction between the treatment and the "before" measurement) are appropriate. The interaction is crucial to the interpretation of these data: (1) when there is no redistricting, partisan bias is not systematically changed; (2) the largest effect of any kind of redistricting is typically to bring partisan bias closer to zero. The lines and points together show this much more clearly than any numerical summary.

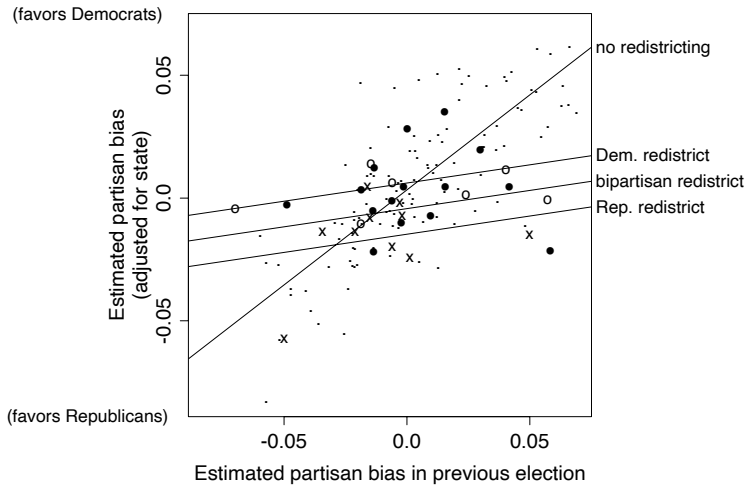

Figure 2.5 *Effect of redistricting on partisan bias in U.S. state legislative elections. Each symbol represents a state and election year, with solid circles, open circles, and crosses representing Democratic, bipartisan, and Republican redistricting, respectively. The small dots are the control cases—state-years that did not immediately follow a redistricting. Lines show fit from a regression model. Redistricting tends to make elections less biased, but small partisan biases remain based on the party controlling the redistricting.*

We sometimes have had success using descriptive symbol names such as two-letter state abbreviations. But if there are only two or three categories, we're happier with visually distinct symbols. For example, to distinguish men and women, we would not use M and W or M and F. In genealogical charts, men and women are often indicated by open squares and open circles, respectively, but even these symbols are hard to tell apart in a group. We prefer clearly distinguishable colors or symbols such as the open circles, solid circles, crosses, and dots in Figure 2.5. When a graph has multiple lines, label them directly, as in Figure 1.7.

These suggestions are based on our experience and attempts at logical reasoning; as far as we know, they have not been validated (or disproved) in any systematic study.

Multiple plots

Example: Last letters of names

Looking at data in unexpected ways can lead to discovery. For example, Figure 2.6 displays the distribution of the last letters of boys' names in the United States in 1906. The most common names in that year included John, James, George, and Edward, for example.

We can learn by putting multiple related graphs in a single display. Figures 2.6 and 2.7 show the dramatic change in the distribution of last letters of boys' names during the twentieth century. In recent years, over a third of boys have been given names that end in "n," with the most common being Ethan, Jayden, Aiden, Mason, and Logan.

There is no single best way to display a dataset. For another view of the data just discussed, we created Figure 2.8, which shows time series of the percentage of boys' names recorded each year ending in each letter.[4] The graph has 26 lines, and we have labeled three of them. We played around with different representations but found the graphs hard to read when more than three lines were highlighted. There has been a steady increase in boys' names ending in "n" during the past 60 years.

Looking at names data another way, Figure 2.9 plots the proportion of boys' and girls' names each year that were in the top 10 names for each sex. Traditionally, boys' names were chosen from a narrower range than girls, with the top 10 names representing 30–40% of all boys, but in recent years,

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

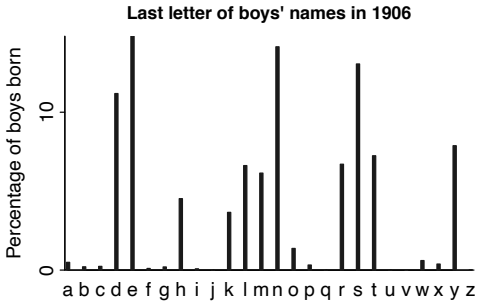

Figure 2.6 *Distribution of last letters of boys' names from a database of American babies born in 1906. Redrawn from a graph by Laura Wattenberg.*

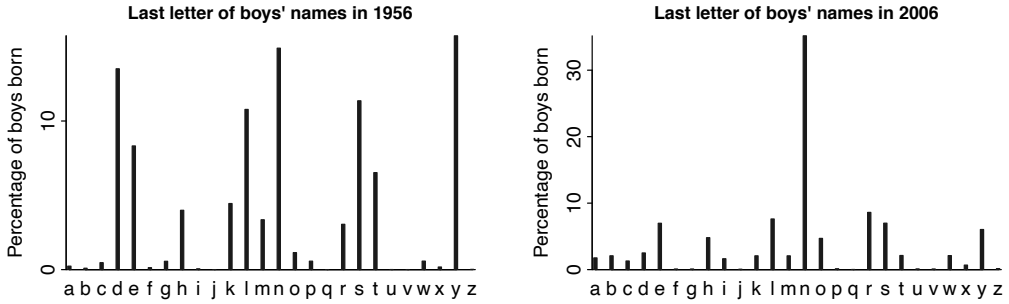

Figure 2.7 *Distribution of last letters of boys' names of American babies born in 1956 and 2006. Redrawn from graphs by Laura Wattenberg. Putting these plots together with the 1906 graph (Figure 2.6) shows a striking trend.*

name choice in the United States has become much more diverse. There are many more patterns to be found in this rich dataset.

Grids of plots

A scatterplot displays two continuous variables, say y vs. x_1. Coloring the dots enables us to plot a third variable, x_2, with some small number of discrete levels. Realistically it can be difficult to read a plot with more than two colors. We can then include two more discrete variables by constructing a two-way grid of plots representing discrete variables x_3 and x_4. This approach of *small multiples* can be more effective than trying to cram five variables onto a single plot.

Example: Swings in congressional elections

Figure 2.10 demonstrates with a grid relating to incumbency in U.S. congressional elections.[5] Each graph plots the swing toward the Democrats from one election to another, vs. the Democratic candidate's share of the vote in the first election, where each dot represents a different seat in the House of Representatives, colored gray for elections where incumbents are running for reelection, or black for open seats. Each row of the graph shows a different pair of national election years, and the four columns show data from different regions of the country.

Breaking up the data in this way allows us to see some patterns, such as increasing political polarization (going from the 1940s through the 1960s to the 1980s, we see a decreasing number of elections with vote shares near 50%), increasing volatility of elections (larger swings in the later periods than before), and a change in the South, which in the 1940s was overwhelmingly Democratic but by the 1980s had a more symmetric range of election results. It would be difficult to see all this in a single plot; in addition, the graph could be easily extended to additional rows (more years of data) or columns (smaller geographic subdivisions).

More generally, we can plot a continuous outcome y vs. a continuous predictor x_1 and discrete

[5]Data and code for this example are in the folder `Congress`.

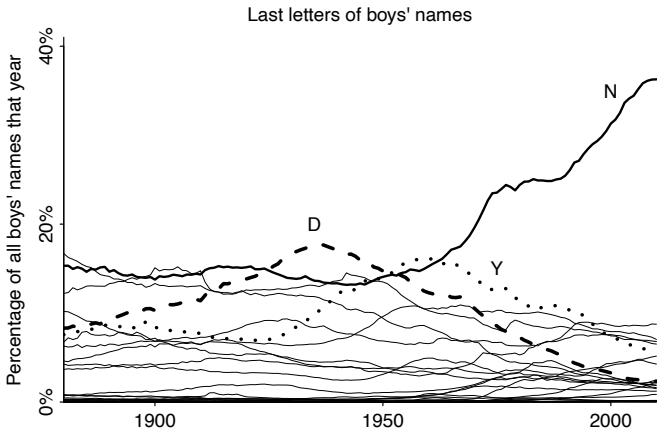

Figure 2.8 *Trends in percentage of boys' names ending in each letter. This graph has 26 lines, with the lines for N, D, and Y in bold to show the different trends in different-sounding names. Compare to Figures 2.6 and 2.7, which show snapshots of the last-letter distribution in 1906, 1956, and 2006.*

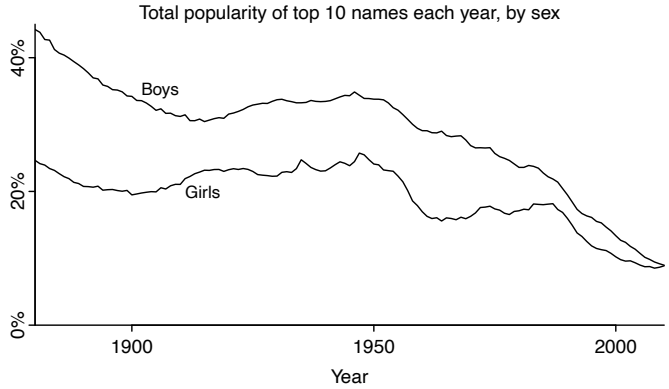

Figure 2.9 *Trends in concentration of boys' and girls' names. In the late 1800s, and then again at different periods since 1950, there have been steep declines in the percentage of babies given the most popular names, so that now the top 10 names of each sex represent only about 10% of baby names. Thus, even as the sounds of boys' names have become more uniform (as indicated by the pattern of last letters shown in Figure 2.6), the particular names chosen have become more varied.*

predictors x_2, x_3, and x_4. If there is interest, we can also plot fitted lines within each plot, showing the expected value of y as a function of x_1 for different fixed values of the other three predictors.

The discrete variables can also represent continuous bins. For example, to display data from an experiment on blood-pressure medication, we could plot after vs. before measurements with different colors for treated and control students, with top and bottom rows of plots showing data from men and women, and rows corresponding to different age categories of patients. Age is a continuous variable, but it could be binned into categories for the graph.

Applying graphical principles to numerical displays and communication more generally

When reporting data and analysis, you should always imagine yourself in the position of the reader of the report. Avoid overwhelming the reader with irrelevant material. For the simplest (but still important) example, consider the reporting of numerical results, either alone or in tables.

Do not report numbers to too many decimal places. There is no absolute standard for significant digits; rather, you should display precision in a way that respects the uncertainty and variability in the

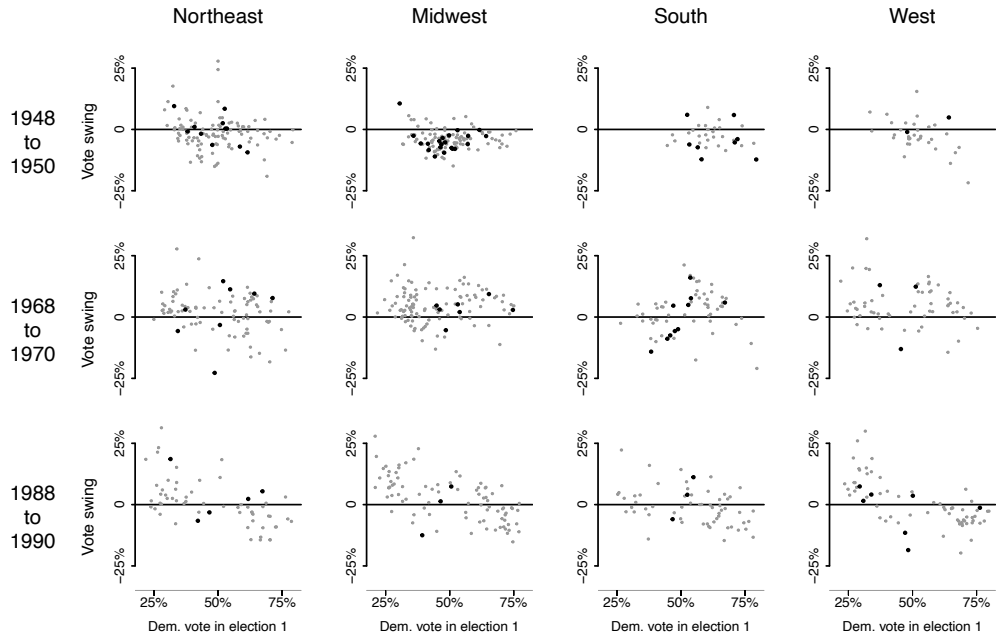

Figure 2.10 *Swings in U.S. congressional elections in three different periods. This grid of plots demonstrates how we can display an outcome (in this case, the swing toward the Democrats or Republicans between two elections in a congressional district) as a function of four predictors: previous Democratic vote share, incumbency status (gray for incumbents running for reelection, black for open seats), region of the country, and time period. Uncontested and landslide elections have been excluded.*

numbers being presented. For example, the uncertainty interval [3.276, 6.410] would be more clearly written as [3.3, 6.4]. (An exception is that it makes sense to save lots of extra digits for intermediate steps in computations, for example, 51.7643 − 51.7581.) A related issue is that you can often make a list or table of numbers more clear by first subtracting out the average (or for a table, row and column averages). The appropriate number of significant digits depends on the uncertainty. But in practice, three digits are usually enough because if more were necessary, we would subtract out the mean first.

The biggest source of too many significant digits may be default computer output. One solution is to set the rounding in the computer program (for example in R, `options(digits=2)`).

A graph can almost always be made smaller than you think and still be readable. This then leaves room for more plots on a grid, which then allows more patterns to be seen at once and compared.

Never display a graph you can't talk about. Give a full caption for every graph, as we try to do in this book. This explains, to yourself and others, what you are trying to show and what you have learned from each plot. Avoid displaying graphs that have been made simply because they are conventional.

Graphics for understanding statistical models

We can consider three uses of graphics in statistical analysis:

1. Displays of raw data, often called "exploratory analysis." These don't have to look pretty; the goal is to see things you did not expect or even know to look for.
2. Graphs of fitted models and inferences, sometimes overlaying data plots in order to understand model fit, sometimes structuring or summarizing inference for many parameters to see a larger pattern. In addition, we can plot simulations of replicated data from fitted models and compare them to comparable plots of raw data.

3. Graphs presenting your final results—a communication tool. Often your most important audience here is yourself—in presenting all of your results clearly on the page, you'll suddenly understand the big picture.

The goal of any graph is communication to self or others. More immediately, graphs are comparisons: to zero, to other graphs, to horizontal lines, and so forth. We "read" a graph both by pulling out the expected (for example, the slope of a fitted regression line, the comparison of a series of uncertainty intervals to zero and each other) and the unexpected. In our experience, the unexpected is usually not an "outlier" or aberrant point but rather a systematic pattern in some part of the data.

Some of the most effective graphs simply show us what a fitted model is doing. See Figure 15.6 for an example.

Graphs as comparisons

All graphical displays can be considered as comparisons. When making a graph, line things up so that the most important comparisons are clearest. Comparisons are clearest when scales are lined up. Creative thinking might be needed to display numerical data effectively, but your creativity can sometimes be enhanced by carefully considering your goals. Just as in writing, you sometimes have to rearrange your sentences to make yourself clear.

Graphs of fitted models

It can be helpful to graph a fitted model and data on the same plot, as we do throughout the book. We also like to graph sets of estimated parameters; see, for example, in Figure 10.9. Graphs of parameter estimates can be thought of as proto-models in that the graph suggests a relation between the y-axis (the parameter estimates being displayed) and the x-axis (often time, or some other index of the different data subsets being fit by a model). These graphs contain an implicit model, or a comparison to an implicit model, the same way that any scatterplot contains the seed of a prediction model.

Another use of graphics with fitted models is to plot predicted datasets and compare them visually to actual data, as we discuss in Sections 11.4 and 11.5. For data structures more complicated than simple exchangeable batches or time series, plots can be tailored to specific aspects of the models being checked.

2.4 Data and adjustment: trends in mortality rates

Even when there are no questions of data quality or modeling, it can make sense to adjust measurements to answer real-world questions.

Example: Trends in mortality rates

In late 2015, economists Anne Case and Angus Deaton published a graph illustrating "a marked increase in the all-cause mortality of middle-aged white non-Hispanic men and women in the United States between 1999 and 2013." The authors stated that their numbers "are not age-adjusted within the 10-y 45–54 age group." They calculated the mortality rate each year by dividing the total number of deaths for the age group by the population as a whole, and they focused on this particular subgroup because it stood out with its increase: the death rates for other age and ethnic groups were declining during this period.

Suspecting an aggregation bias, we examined whether much of the increase in aggregate mortality rates for this age group could be due to the changing composition of the 45-to-54-year-old age group over the 1990 to 2013 time period. If this were the case, the change in the group mortality rate over time may not reflect a change in age-specific mortality rates. Adjusting for age confirmed this suspicion. Contrary to the original claim from the raw numbers, we find there is no longer a steady increase in mortality rates for this age group after adjusting for age composition. Instead, there is an increasing trend from 1999 to 2005 and a constant trend thereafter. Moreover, stratifying age-adjusted mortality rates by sex shows a marked increase only for women and not men.

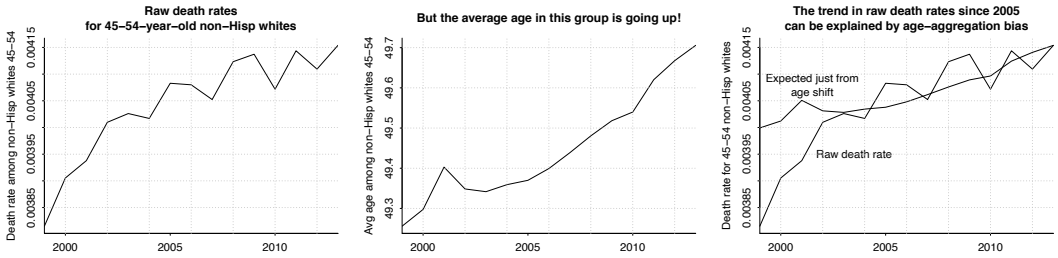

Figure 2.11 *(a) Observed increase in raw mortality rate among 45-to-54-year-old non-Hispanic whites, unadjusted for age; (b) increase in average age of this group as the baby boom generation moves through; (c) raw death rate, along with trend in death rate attributable by change in age distribution alone, had age-specific mortality rates been at the 2013 level throughout.*

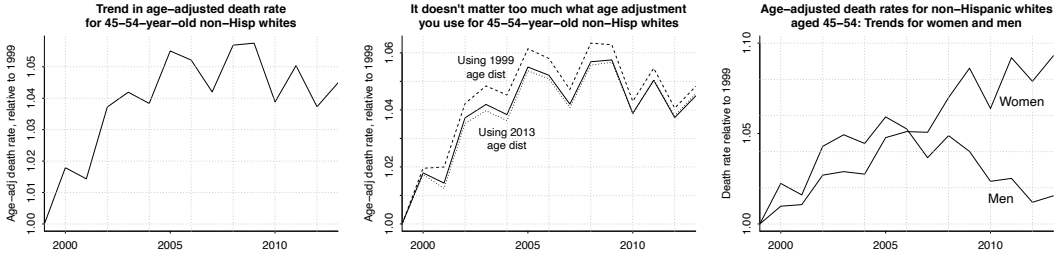

Figure 2.12 *(a) Age-adjusted death rates among 45-to-54-year-old non-Hispanic whites, showing an increase from 1999 to 2005 and a steady pattern since 2005; (b) comparison of two different age adjustments; (c) trends in age-adjusted death rates broken down by sex. The three graphs are on different scales.*

We demonstrate the necessity of the age adjustment in Figure 2.11.[6] The unadjusted numbers in Figure 2.11a show a steady increase in the mortality rate of 45-to-54-year-old non-Hispanic whites. During this period, however, the average age in this group increased as the baby boom generation passed through. Figure 2.11b shows this increase.

Suppose for the moment that mortality rates did not change for individuals in this age group from 1999 to 2013. In this case, we could calculate the change in the group mortality rate due solely to the change in the underlying age of the population. We do this by taking the 2013 mortality rates for each age and computing a weighted average rate each year using the number of individuals in each age group. Figure 2.11c shows the result. The changing composition in age explains about half the change in the mortality rate of this group since 1999 and all the change since 2005.

Having demonstrated the importance of age adjustment, we now perform an adjustment for the changing age composition. We ask what the data would look like if the age groups remained the same each year and only the individual mortality rates changed. Figure 2.12a shows the simplest such adjustment, normalizing each year to a hypothetical uniformly distributed population in which the number of people is equal at each age from 45 through 54. That is, we calculate the mortality rate each year by dividing the number of deaths for each age between 45 and 54 by the population of that age and then taking the average. This allows us to compare mortality rates across years. Consistent with Figure 2.11c, the resulting mortality rate increased from 1999 to 2005 and then stopped increasing.

We could just as easily use another age distribution to make valid comparisons across years. Checking, we find that age-adjusted trend is not sensitive to the age distribution used to normalize the mortality rates. Figure 2.12b shows the estimated changes in mortality rate under three options: first assuming a uniform distribution of ages 45–54; second using the distribution of ages that existed in 1999, which is skewed toward the younger end of the 45–54 group; and third using the 2013 age distribution, which is skewed older. The general pattern does not change.

[6]Data and code for this example are in the folder `AgePeriodCohort`.

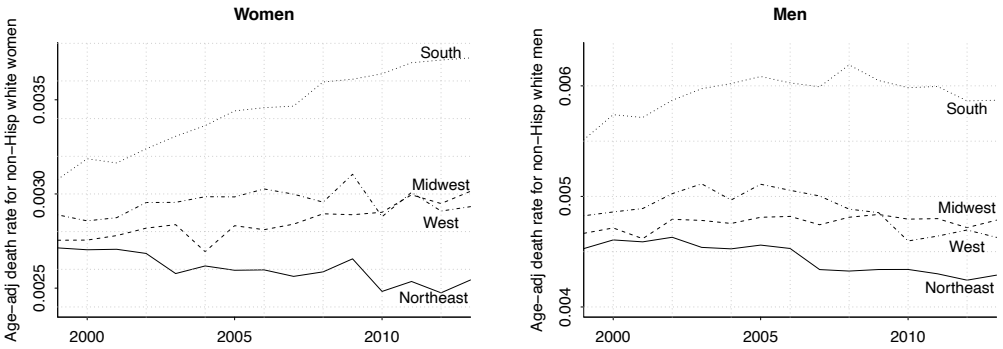

Figure 2.13 *Age-adjusted death rates among 45-to-54-year-old non-Hispanic white men and women, broken down by region of the country. The most notable pattern has been an increase in death rates among women in the South. In contrast, death rates for both sexes have been declining in the Northeast. The graphs are on different scales; as can be seen from the y-axes, death rates are lower for women than for men.*

Calculating the age-adjusted rates separately for each sex reveals a crucial result, which we display in Figure 2.12c. The mortality rate among white non-Hispanic American women increased from 1999 to 2013. Among the corresponding group of men, however, the mortality rate increase from 1999 to 2005 was nearly reversed from 2005 to 2013.

In summary, age adjustment is not merely an academic exercise. Due to the changing composition of the 45-to-54-year-old age group, adjusting for age changes the interpretation of the data in important ways. This does not change a key finding that had been seen in the unadjusted data: the comparison of non-Hispanic U.S. middle-aged whites to other countries and other ethnic groups. These comparisons hold up after our age adjustment. The aggregation bias in the published unadjusted numbers is on the order of 5% in the trend from 1999 to 2003, while mortality rates in other countries and other groups fell by around 20% during that period.

One can learn more by further decomposing these data. For example, Figure 2.13 breaks down the age-adjusted death rates in that group by U.S. region. The most notable pattern has been an increase in death rates among women in the South. In contrast, death rates for both sexes have been declining in the Northeast, the region where mortality rates were lowest to begin with. These graphs demonstrate the value of this sort of data exploration.

2.5 Bibliographic note

Some general references on data display and exploration include Cleveland (1985, 1993), Friendly and Kwan (2003), Chambers et al. (1983), Tukey (1977), Mosteller and Tukey (1977), Tufte (1983, 1990), Bertin (1967), and Wainer (1984, 1997). Gelman and Unwin (2013) discuss different goals of information visualization and statistical graphics.

For statistical graphics in R, the books by Healy (2018), Wickham (2016), and Murrell (2005) are good starting points. Fox (2002) is also helpful in that it focuses on regression models. An important topic not discussed in the present book is dynamic graphics; see Buja et al. (1988).

There are various systematic ways of studying statistical graphics. One useful approach is to interpret exploratory visualization as checks of explicit or implicit models. Another approach is to perform experiments to find out how well people can gather information from various graphical displays; see Hullman, Resnick, and Adar (2015) for an example of such research. More work is needed on both these approaches: relating to probability models is important for allowing us to understand graphs and devise graphs for new problems, and effective display is important for communicating to ourselves as well as others.

For some ideas on the connections between statistical theory, modeling, and graphics, see Buja et

al. (2009), Wilkinson (2005), and, for our own perspective, Gelman (2004a). Unwin, Volinsky, and Winkler (2003), Urbanek (2004), and Wickham (2006) discuss exploratory model analysis, that is, visualization of different models fit to the same data.

For different perspectives on tabular displays, compare Tukey (1977); Ehrenberg (1978); Gelman, Pasarica, and Dodhia; and Wickham and Grolemund (2017, chapter 10).

For background on the Human Development Index, see Gelman (2009a). The graphs of political ideology, party identification, and income come from Gelman (2009b). The graph of health spending and life expectancy appears in Gelman (2009c). The graphs of baby names are adapted from Wattenberg (2007).

Validity and reliability are discussed in textbooks on psychometrics but have unfortunately been underemphasized in applied statistics; see Gelman (2015b). Rodu and Plurphanswat (2018) discuss a problem with the "never smoker" definition in a study of adolescent behavior. The middle-aged mortality rate example appears in Gelman (2015c) and Gelman and Auerbach (2016); see also Schmid (2016), Case and Deaton (2015, 2016), and Gelman (2017).

2.6 Exercises

2.1 *Composite measures*: Following the example of the Human Development Index in Section 2.1, find a composite measure on a topic of interest to you. Track down the individual components of the measure and use scatterplots to understand how the measure works, as was done for that example in the book.

2.2 *Significant digits*:

(a) Find a published article in a statistics or social science journal in which too many significant digits are used, that is, where numbers are presented or displayed to an inappropriate level of precision. Explain.

(b) Find an example of a published article in a statistics or social science journal in which there is *not* a problem with too many significant digits being used.

2.3 *Data processing*: Go to the folder `Names` and make a graph similar to Figure 2.8, but for girls.

2.4 *Data visualization*: Take any data analysis exercise from this book and present the *raw data* in several different ways. Discuss the advantages and disadvantages of each presentation.

2.5 *Visualization of fitted models*: Take any data analysis exercise from this book and present the *fitted model* in several different ways. Discuss the advantages and disadvantages of each presentation.

2.6 *Data visualization*: Take data from some problem of interest to you and make several plots to highlight different aspects of the data, as was done in Figures 2.6–2.8.

2.7 *Reliability and validity*:

(a) Give an example of a scenario of measurements that have *validity* but not *reliability*.

(b) Give an example of a scenario of measurements that have *reliability* but not *validity*.

2.8 *Reliability and validity*: Discuss validity, reliability, and selection in the context of measurements on a topic of interest to you. Be specific: make a pen-on-paper sketch of data from multiple measurements to demonstrate reliability, sketch true and measured values to demonstrate validity, and sketch observed and complete data to demonstrate selection.

2.9 *Graphing parallel time series*: The mortality data in Section 2.4 are accessible from this site at the U.S. Centers for Disease Control and Prevention: `wonder.cdc.gov`. Download mortality data from this source but choose just one particular cause of death, and then make graphs similar to those in Section 2.4, breaking down trends in death rate by age, sex, and region of the country.

2.10 *Working through your own example*: Continuing the example from Exercise 1.10, graph your data and discuss issues of validity and reliability. How could you gather additional data, at least in theory, to address these issues?