2 An Introduction to Data Analysis

CHAPTER OUTLINE

LEARNING OBJECTIVES

- Discuss the goals and scope of the book.
- Explain the importance of acquiring skills in data analysis.
- List the components of data analysis and how they fit together.
- Form hypotheses from descriptions of data.
- Explain the connection between hypotheses, models, and estimates.
- Define diagnostics and explain their role in data analysis.
- Formulate new questions.

Overview

Scholars, practitioners, and policymakers interested in explaining human behavior are drowning in data. This book is designed for those who want to swim safely to shore. The goal is to introduce the method, logic, art, and practice of data analysis. Specifically, the book provides the essential skills and tools necessary to examine data in the service of solving problems. Toward that end, you will learn to marry the art and practice of data analysis with how experienced practitioners approach problems, formulate hypotheses, estimate models, and present their results. The underlying philosophy is learning by doing. Learning statistics along with the art and practice of data analysis is best achieved by *doing* data analysis.

This introductory chapter is organized as follows. First, I discuss some of the motivations behind analyzing quantitative data. Second, I provide a short exploration of an important question in the social sciences to illustrate the main components of data analysis. The example does not fully explain each component. The goal here is to provide a road map for the rest of the book.

The book is organized according to the process of data analysis laid out in this chapter: describing data and formulating hypotheses, building and estimating models, diagnostics, and generating the next question. Presenting your results also forms an important part of this book. The process of data analysis is iterative: it is an ongoing conversation that goes back and forth between all of the different components. For example, the estimates we generate from our statistical models rarely settle the matter. New and better questions are often the result.

Motivating Data Analysis

The motivations are many, but here are three of the most rehearsed. First, more things are being measured and quantified. Not only are more data available on demographic and financial trends, our daily activity (driving, dating, shopping, and listening to music) is available for analysis. Second, and partly related to the first, data analysis is in demand. Consequently, the skills associated with analyzing quantitative data are marketable. Third, data analysis helps us separate facts from fiction, a defining characteristic of successful democratic societies.

Big Data Are Getting Bigger

A significant part of our lives is spent on the Internet. The bread crumbs left behind are used by government, private industry, and others. Politicians are driven (in part) by survey results. CEOs make decisions based on data collected from their employees and their customers. Here are just a few examples of data analysis that reveal both its ubiquity and effectiveness.

1. Noticing an increasing amount of junk mail from Target advertising diapers, baby formula, and onesies, the inundated recipient calls Target to ask why they're sending him so much baby-oriented advertising (it had been years since there was a baby in the house). Target explained that recent purchasing data indicated there was a pregnant woman in his household. One week after calling Target, he discovers his daughter is pregnant.
2. By constructing a meta-analysis of polls, Nate Silver (in his FiveThirtyEight column) correctly predicted the outcome of the presidential race in 2012 in all 50 states.
3. When diseases spread (e.g., the Ebola outbreaks in Africa), the U.S. Centers for Disease Control and Prevention and other authorities previously relied on 2-week-old data collected from hospitals and clinics. In a quickly developing scenario like the H1N1 (bird flu), Ebola, Zika, or COVID-19 virus outbreaks, 2 weeks is too late. A team from Google let the Internet figure out where outbreaks occur. To develop their model, they traced the spread of H1N1 and correlated it with search terms on the Internet: high fever, cough, and aches. They eventually found 45 phrases that allowed them to track the outbreak, informing officials when and where, exactly, the newest flu outbreak was occurring.[1]
4. In an article entitled "China Invents the Digital Totalitarian State" (2016), *The Economist* details the plans of a new Chinese program called the social-credit project designed to increase trust in society by making social-credit scores available. This example, more than the others, illustrates how big data can be abused. In what is called the "judgment defaulter's list" (a tally of citizens who have defied a court order), Chinese authorities keep track of who they consider to be trustworthy in society to help spur economic cooperation and trust among its citizens.

[1] The examples above can be found in *Big Data: A Revolution That Will Transform How We Live, Work, and Think* by Kenneth Cukier and Viktor Mayer-Schönberger (2013). Google's attempts to track flu outbreaks have not come without criticism. For a nice discussion of the shortcomings associated with big data, see the discussion in Lazer et al. (2014).

5. To help reduce crime in Chicago, a city plagued by high homicide rates, new data and data analysis software help to reduce crime ("Violent Crime Is Down in Chicago," 2018). Equipped with sensors that locate gunshots throughout the city, along with maps of liquor stores and freeway on-ramps, analysts identify areas where crime will most likely occur. That information, combined with data on televised sporting events, increases the accuracy of locating potential problems. The new data allow police in Chicago's inner-city areas to patrol more effectively.

Whether in commerce, politics, government, health, or crime, data are the key to solving important problems. These examples show that data are used increasingly to understand and influence human behavior or in the Chinese case exert social control. Those who know how to collect, analyze, and explain data can influence important decisions.

Data Analysis Is a Marketable Skill

Quantitative data analysis is a marketable skill. Confronted by an ocean of data, companies, government offices, and nongovernmental organizations recognize the need for accurate and timely analysis. One need only to Google the phrase "jobs in R" to see the possibilities.

In the private sector, an increasing number of companies rely on surveys to understand their market, employees, and customers. The expanded use of Survey Monkey and Qualtrics adds to that reliance. Conducting surveys is an art. Done poorly, surveys either mislead or say nothing about the problem at hand. Understanding the basics of quantitative data analysis develops an appreciation and facility for survey research.

Government agencies need well-trained social scientists. Ten or 15 years ago, government agencies would only pay lip service to their need for them. Over the last several years, they've actually started to hire social scientists. Data skills will bolster a career in local, state, and federal government agencies; organizations know that a work force staffed with effective analysts produces better policy.

Postgraduate programs want students with a quantitative background. A firm grasp of quantitative skills can help gain entrance into many top 20 departments, no matter the discipline. Law school candidates with a specialty in engineering, computer science, or quantitative analysis have a unique skill that puts them head and shoulders above others, an important consideration given the competition.

In addition to law school, quantitative skills can open doors to policy schools and graduate programs in the social sciences. The number of subjects we study in the social sciences that involve examining quantitative evidence is growing. Over the last few decades, there's just too much quantitative data and evidence to be ignored. Applicants to graduate programs in these areas increase the likelihood of gaining entrance to top-ranked schools if they can demonstrate a facility with data analysis.

Data Analysis Is a Public Good

Perhaps more important than these purely instrumental reasons, society depends on good analysis. As wonderfully illustrated in the book *A Mathematician Reads the Newspaper*, we deserve the kind of public policy we get (Paulos, 1995). In one example, Paulos recounts a contested state senate race in Pennsylvania where fraud was detected by fitting a regression line to previous election results and determining that the race in question was simply too anomalous to be credible. Given the current divisiveness that characterizes U.S. politics, understanding sampling and the inferences we can draw from it can either bolster our faith in the results or suggest a new election is necessary.

Another characteristic of our political age is the growth of conspiracy theories. With so much data floating around, it is easy to connect dots that are completely unrelated yet convince many that dark forces are at work. In a prescient example, Paulos recounts an exercise undertaken by John Leavy, a computer programmer from the University of Texas. Leavy fed data into a computer to come up with similarities between different pairs of U.S. presidents. In that exercise, he found some very interesting connections between Presidents William McKinley and James Garfield who were both assassinated. Observe the similarities:

> It turns out that both of these presidents were Republicans who were born and bred in Ohio. They were both Civil War veterans, and both served in the House of Representatives. Both were ardent supporters of protective tariffs and the gold standard, and both of their last names contained eight letters. After their assassinations they were replaced by their vice presidents, Theodore Roosevelt and Chester Alan Arthur, who were both from New York City, who both sported mustaches, and who both had names containing seventeen letters. (Paulos, 1995, p. 91)

Whether citizens become data scientists or sophisticated consumers of quantitative data, a more informed public is better able to determine fact from fiction. An important goal of this book is to demonstrate how statistics are used and abused. Distinguishing between fact and fiction depends on individuals with the skill and knowledge this book hopes to provide.

KNOWLEDGE CHECK: Explain the importance of acquiring skills in data analysis.

1. Given big data's growing influence on our lives, why is data analysis an important skill?
   a. It helps in our careers.
   b. We get better policy.
   c. It defines our roles as citizens and consumers.
   d. With more skills we can realize higher salaries.
2. Data analysis as a skill (a private good) is helpful with which of the following?
   a. We can realize a higher salary.
   b. Both government and the private sector need skilled individuals.
   c. It helps with gaining entrance into postgraduate degree programs (e.g., law school, business school, graduate school in general).
   d. It helps society.
3. As a public good, which of the following describe the benefits of having data analytic skills.
   a. It leads to better government policy.
   b. It helps society determine fact from fiction.
   c. It aids democracy.
   d. It helps businesses understand their customers.

The Main Components of Data Analysis

There is no shortage of good textbooks on how to conduct social science research in which the focus is on the process of hypothesis formation, theory testing, and drawing inferences. Here, the goal is the same but the approach is different. In this book we learn those foundations of inquiry through data visualization and by working on real problems with real data.

In that spirit, this section outlines four components of data analysis in the order they should occur: (1) describing data and formulating hypotheses, (2) building and estimating models, (3) diagnostics, and (4) generating the next question. There are concepts and techniques (i.e., model building and estimation, transforming variables, diagnostics, etc.) that will be new. The purpose of this chapter is to introduce the broad outlines of good data analysis with an example. When new techniques and concepts are introduced, I will indicate where they are covered in detail later in the book rather than dwell on them here. Let me define the main components of data analysis.

1. Describing data and formulating hypotheses

We describe data to better understand the problem and to ask better questions. At its base, describing data focuses primarily on identifying the typical case (central tendency) and understanding how typical that typical case is (dispersion). Describing data, however, should go much deeper than that. Observing where specific cases or entire classes of cases lie in relation to others is an important part of the enterprise. The more we know about our data, the better questions we'll generate and the better hypotheses we'll formulate. The concepts and tools necessary for describing data are found in Chapters 3–7.

Before we continue, here is a word about **theories** and **hypotheses**. Typically, hypotheses are guesses and theories are interconnected hypotheses that form a greater whole and have been tested with some success. The distinction is often fuzzy since in the vernacular theories and hypotheses are often used interchangeably. In this book, a hypothesis will refer to a specific guess about how two things are related (e.g., religiosity and political ideology). A theory will refer to a set of related hypotheses that explain successfully some empirical phenomenon.

2. Building and estimating models

Once we have some familiarity with our data and some possible explanations have been forwarded, we move on to building and estimating models. **Models** are simplified versions of reality that help us understand our complex world. Models can be thought of as arguments or explanations. They are arguments we make to explain an empirical problem or puzzle. For example, if we want to explain why some countries have high rates of homicide, we construct a model or argument that might include income, age of the population, number of police, and efficacy of the judicial system. There are a multitude of other possible causes we could include, but it helps to keep things simple. We don't want to recreate reality; we merely want to approximate it. With a good model in hand (i.e., a model that contains the main causes but not every single possible one), we can begin to understand how important each cause is and we can estimate its impact.

Estimates can be very sensitive to the model we choose. How we look at the data influences what we see. Consequently, model building and estimation is a process that should be performed as a back-and-forth between theory and evidence. The practice of model building is treated in Chapters 10–14.

3. Diagnostics

After we've constructed models and obtained some estimates, we turn to **diagnostics**. Diagnostics are a set of tools we use to determine whether we're using the right kind of model. To ascertain whether our model is appropriate, we examine how well the predictions from our model match reality. The difference between our prediction and reality is called the residual. For example, if our model does a good job of predicting infant mortality rates in all countries except for the oil-rich states of the Middle East, the resulting diagnostics will say so. That is, the residuals for these cases will be relatively large. Perhaps our model estimates are overly influenced by those Middle Eastern countries. Diagnostics help us determine whether our estimates provide a good sense of how the world really works, are the product of some strange cases, or are the result of a poorly chosen model.

It is important to keep in mind that diagnostics can both detect problems and help uncover interesting relationships, generating additional explanations or hypotheses. Diagnostics are explained in Chapters 15 and 16.

4. Generating the next question

Finally, armed with our estimates and a sense of how well our predictions fit reality, additional questions and ideas inevitably surface. A useful way to construct or identify these additional questions and ideas is to use a simple if-then statement: *if* the estimates we obtained are correct, *then* we would expect to see *x*. Following each set of estimates with that statement helps to unearth possible explanations and additional hypotheses to test. Since it is impossible to prove anything with complete certainty, the exercise of generating additional hypotheses to test is extremely important. Much like a prosecutor in a court of law, we must provide as much circumstantial evidence as possible to convince the jury. The more evidence we provide, the more likely the jury will find the defendant guilty. The Appendix presents a way to approach generating additional hypotheses with a few tips and tricks.

Now that we have a sense of what the main components of data analysis are, let's dive into an example of what they look like in real life. The next section serves as an introduction to data analysis by providing a very brief exploration of a problem, a puzzle that has occupied the attention of the social sciences for decades.

KNOWLEDGE CHECK: List the components of data analysis and how they fit together.

4. Which are components of data analysis?
   a. Generating and forming hypotheses
   b. Diagnostics
   c. Describing the data
   d. Collecting the data
5. What is the correct order of the components of data analysis?
   a. Estimating models, forming hypotheses, diagnostics, forming new questions
   b. Describing data, forming hypotheses, estimating models, diagnostics, forming new questions

   c. Forming hypotheses, describing data, estimating models, diagnostics, forming new questions

   d. Describing data, forming hypotheses, forming new questions, diagnostics

6. Which of the following are accurate statements about data analysis?

   a. There is an agreed-upon method everyone follows.

   b. Hypothesis formation always comes before describing data.

   c. Describing data always comes before hypothesis formation.

   d. There is an agreed-upon order of analysis.

7. What is the difference between a theory and hypothesis?

   a. There really is no difference.

   b. Hypotheses are guesses, theories are certain.

   c. Theories are simply hypotheses that have stood the test of time.

   d. Theories define an integrated approach or set of hypotheses that have accumulated corroborating evidence over time.

8. Which statements best describe models?

   a. A model replicates reality.

   b. A model is a simplification of reality.

   c. A model is best judged by its ability to predict.

   d. A model's use is defined both by its ability to predict and its simplicity.

Developing Hypotheses by Describing Data

The best analysis starts with a puzzle or a question. To illustrate the broad outlines of data analysis, let's begin with a tried-and-true example. Why are some countries richer than others? Work drawn from development economics, political science, and sociology emphasizes education, religion, ethnic fractionalization, gender equality, and political stability.[2] These explanations represent only some of the more common hypotheses scholars have forwarded to account for the huge disparity we observe between countries. Let's explore these hypotheses in a little more detail.

Art and Practice of Data Visualization

DESCRIBING DATA AND FORMULATING HYPOTHESES

Describing data and formulating hypotheses is an interactive process. Data analysis never starts without some guess about what we'll find. An honest analysis, however, changes what we know. At its best, data analysis leads to new questions we neglected to ask in the first place.

[2] For an informative treatment of growth theory and the current scholarship on the empirics of growth, see Jones (2015). For a recent review of the literature on long-run growth trends, see Lloyd and Lee (2018).

Hypothesis I: Ethnolinguistic Fractionalization

What is the relationship between a country’s ethnic landscape and economic growth? Some have argued there is a direct relationship between economic development and the number of ethnic groups in a country. Endowed with a multitude of different languages, religions, and beliefs, individuals in heterogeneous societies may find it more difficult to organize production and markets. Rivalries between different ethnic groups (ethnic conflict) can also inhibit economic development.

To explore the first hypothesis, we now move to more involved R code. Remember that the best way to work with R is to save chunks of code so that they can be copied and pasted when needed. Once you’ve established how you want a scatter plot or any other figure to look, it’s easy to switch out variables, choose colors, or change a figure’s underlying features.

So far, you’ve been introduced to just a few simple base R functions, *hist()* and *plot()*. To get you up and running with publishable figures and tables, let’s dive in. Throughout the book, I use a package in R called ‘ggplot2.’ The “gg” in **ggplot** stands for the grammar of graphics, a way of thinking about graphs based on layers. Let me explain.

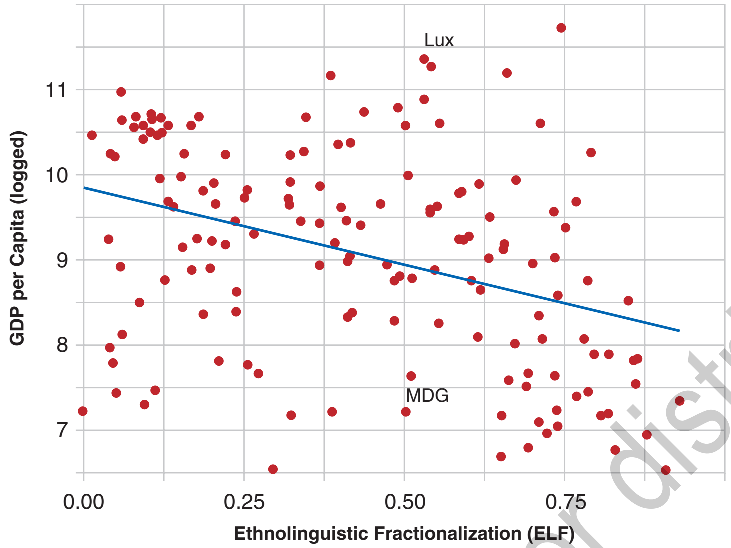
The code in Code Chunk 2-1 starts with the *ggplot()* function. Within the parentheses, we first indicate which data set to use: *world*. The next feature of the code is the *aes()* function or “aesthetics.” The aesthetics indicate which variables to use and any shapes, colors, or labels to add. Once that first layer is established, note that the rest of the code is simply a series of additional layers. First, we add a layer of points with *geom_point()*. We then add a straight line with *geom_smooth()*, a title with *ggtitle()*, x- and y-axis labels with *xlab* and *ylab*, a style with *theme_minimal()* and *theme()*, and some text with *geom_text_repel()*. The code within the *geom_text_repel()* function is a little complex because we want R to only label the two data points that represent Luxembourg (LUX) and Madagascar (MDG). Notice how each layer is added by placing an addition symbol between each added layer. I encourage you to start with the *ggplot()* function, execute it, and observe the result. Then, add each additional layer one by one, observing how the plot evolves into the final product. That exercise will demonstrate the logic of the grammar of graphics: the simple addition of layer upon layer.

The code in Code Chunk 2-1 draws a nice scatter plot that illustrates the relationship between ethnic heterogeneity and the logged value of gross domestic product (GDP) per capita.

Code Chunk 2-1

```
ggplot(world, aes(ethfrac, log(gdppc), label = iso3c)) +
  geom_point(color=“#bf0000”) +
  geom_smooth(method=“lm”, se=FALSE, color=“#0000bf”) +
  ggtitle(“Figure 2-1: ELF Lowers Income”) +
  ylab(“GDP per Capita (logged)”) +
  xlab(“Ethnolinguistic Fractionalization (ELF)”) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = “bold”),
        axis.title = element_text(size = 8, face = “bold”)) +
  geom_text_repel(size = 2.8,
       aes(label=ifelse(iso3c==“MDG” |
                        iso3c==“LUX”,
                        as.character(iso3c),’’),
           hjust = 0, vjust=-1), show.legend=FALSE)
```

FIGURE 2-1 ELF Lowers Income



In Figure 2-1, I use GDP per capita to measure wealth (income). GDP is simply the total amount of all goods and services in a country produced in a given year. Put simply, ethnolinguistic fractionalization (ELF) indicates how likely your neighbor speaks a different language. The higher the score, the more languages spoken over a country's population. The **scatter plot** of GDP per capita and the degree of ethnic heterogeneity[3] suggests that as ethnic heterogeneity increases, GDP per capita decreases. Scatter plots show where a single observation (a country, in this case) is situated on a two-dimensional grid delineated by two variables (GDP per capita [logged] and ethnolinguistic fractionalization). Scatter plots will be discussed at greater length in Chapter 5.

Note that while there is an overall negative relationship (the line slopes downward), knowing the degree of ethnic heterogeneity in a country does little to help us predict the income level with precision. For example, Madagascar (MDG) and Luxembourg (LUX) have roughly the same ethnolinguistic heterogeneity score (ELF) but extremely different levels of per capita income. Even though knowing the ELF score can give us a rough guess, there is still considerable variation in GDP per capita left to explain.

Although there does seem to be a negative relationship between the ELF score and GDP per capita, a downward sloping line may not be the best way to summarize the relationship. A closer look at the plot indicates there is a cluster of countries that might be influencing the line (the upper-left corner). If those cases were not there, it's possible the line would have a flatter slope. That cluster of countries might be pulling the left part of the line upward, resulting in a negative slope. It turns out that those countries fall into a particular category labeled the

[3] Per capita income is measured by GDP per capita. The GDP per capita measure is presented in its logged form for reasons that will be explained in the chapter on transforming variables (Chapter 6). The measure of ethnicity used is the ethnolinguistic fractionalization score constructed by Alesina et al. (2003). In earlier formulations, the score was constructed to represent the probability that two randomly selected individuals in a country were from the same ethnolinguistic group. Alesina and coauthors use the same underlying construct but combine racial and linguistic characteristics. These data are from Pollock (2014).

industrial countries, a collection of countries that includes Europe, North America, Australia, New Zealand, and Japan.

In the following code, note how some changes to Code Chunk 2-1 allow me to color the industrial countries blue, turn the original line into a dashed line, and fit a separate line (using the *smooth()* function) to the nonindustrialized countries (those colored red). I can also place some words in the figure with the *annotate()* function. We're not going to get bogged down with all of the code at this point, since we want to concentrate our efforts on the components of data analysis. As the chapters unfold, you will become well acquainted with all of the code in Code Chunk 2-2.

Code Chunk 2-2

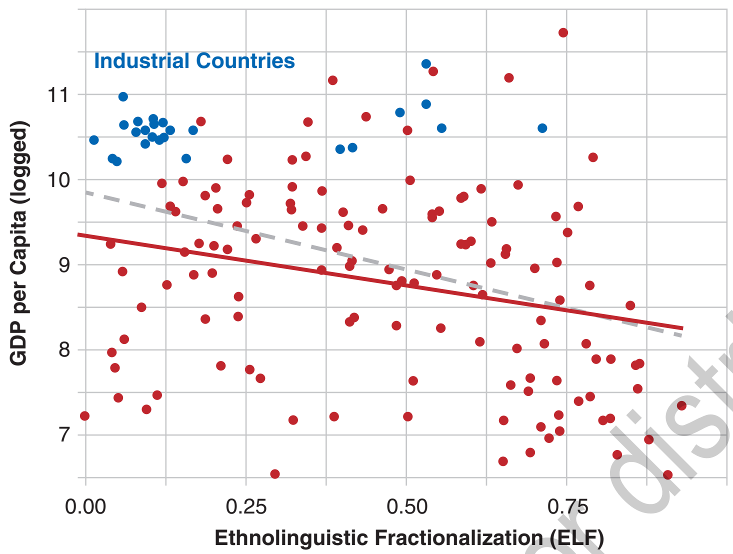
```
ggplot(world, aes(ethfrac, log(gdppc))) +
  geom_point(color=ifelse(world$aclpregion==”Industrial Countries”,
                          “#0000bf”, “#bf0000”)) +
  geom_smooth(method=“lm”, col=“grey”, se=FALSE,
              linetype = “dashed”) +
  ggtitle(“Figure 2-2: Industrialized Countries Are Clustered”) +
  ylab(“GDP per Capita (logged)”) +
  xlab(“Ethnolinguistic Fractionalization”) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = “bold”),
        axis.title = element_text(size = 8, face = “bold”)) +
  geom_smooth(data=subset(world,
                          aclpregion != “Industrial Countries”),
                          color=“#bf0000”, se=FALSE, method = “lm”) +
  annotate(“text”, x = .15, y = 11.5,
           label = “Industrial Countries”, col=“#0000bf”)
```

Figure 2-2 demonstrates how sensitive the line is to the influence of the industrial countries. Note how the line fit for the rest of the world (the red line) is relatively flat compared to the gray dashed line that accounts for the industrial countries.

To further emphasize the influence of the cluster, I asked R to fit a curved line to all of the countries in the world that are not classified “industrial.” Smooths will be discussed and defined in detail later (Chapter 5). For now, think of a **smooth** as simply a line or curve that R fits to the data. The curve demonstrates there may be another pattern lurking. In Figure 2-3, the black line shows that when we describe the data with a nonlinear smooth (the curve), an interesting pattern emerges. While the overall trend might be roughly negative (gray dashed line), a majority of the data (the nonindustrialized countries) tell a different story. As we move from left to right along the x axis, GDP per capita increases, reaches a maximum, then decreases. Among nonindustrialized countries, there may be an optimal degree of ethnolinguistic heterogeneity that encourages economic dynamism.

In Code Chunk 2-3, I've only replaced the straight red line in Figure 2-2 with a curved black line. The curved line is the default for *geom_smooth()*, so I simply removed the expression *method=“lm”* that tells R to draw a straight line.

■ **FIGURE 2-2** Industrialized Countries Are Clustered



Code Chunk 2-3

```
ggplot(world, aes(ethfrac, log(gdppc))) +
  geom_point(color=ifelse(world$aclpregion==”Industrial Countries”,
                         “#0000bf”, “#bf0000”)) +
  geom_smooth(method=“lm”, col=“grey”, se=FALSE,
              linetype = “dashed”) +
  ggtitle(“Figure 2-3: The Relationship Could Be Nonlinear”) +
  ylab(“GDP per Capita (logged)”) +
  xlab(“Ethnolinguistic Fractionalization”) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = “bold”),
        axis.title = element_text(size = 8, face = “bold”)) +
  geom_smooth(data=subset(world,
                         aclpregion != “Industrial Countries”),
                         color=“black”, se=FALSE) +
  annotate(“text”, x = .15, y = 11.5,
           label = “Industrial Countries”, col=“blue”)
```

The curved pattern depicted by the black line might lead us to conclude that if there is a severe lack of ethnic heterogeneity or if there is too much, economic development suffers.

While the observation has some merit, an even closer look indicates the upside-down U-shape pattern may be somewhat overblown. Note in Figure 2-4 that the four most impoverished and ethnolinguistically homogeneous countries are all poor, very small island

FIGURE 2-3 The Relationship Could Be Nonlinear

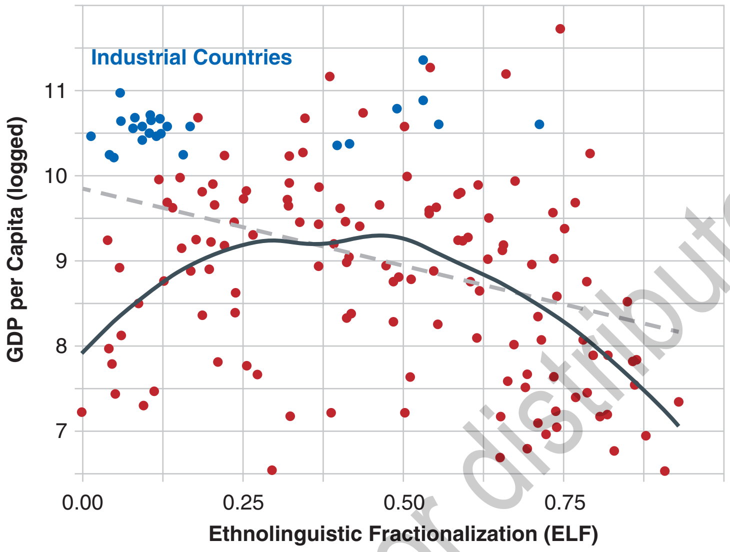
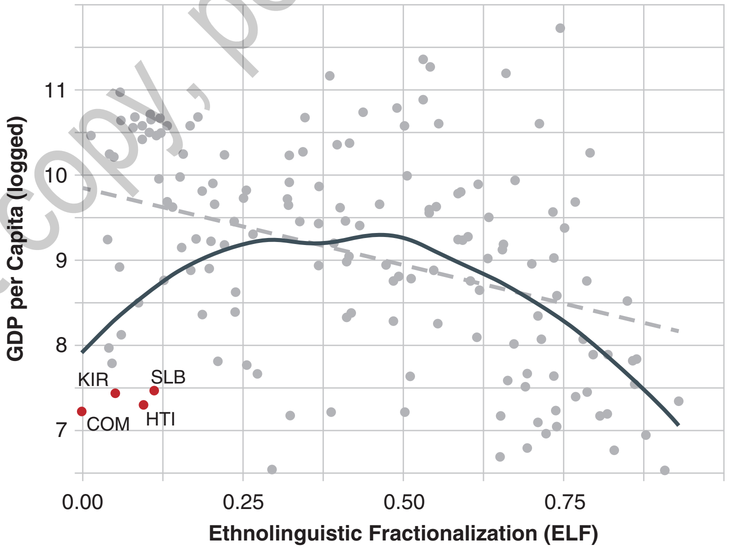


FIGURE 2-4 Small Countries Have Low GDP per Capita



nations: Comoros (COM), Haiti (HTI), Kiribata (KIR), and the Solomon Islands (SLB). Given the small size of these countries and those in close proximity, perhaps size matters.[4]

[4] There is a literature in economics that examines whether there is an optimal state size for economic performance. For one contribution in that debate, see Alesina (2003).

Among the developing countries, the smallest (regardless of heterogeneity) all have relatively low per capita incomes.

To highlight the four cases, I use the *ifelse()* function both to color them differently and to include their labels.

Code Chunk 2-4

```
ggplot(world, aes(ethfrac, log(gdppc))) +
  geom_point(color=ifelse(world$iso3c==“COM” |
                            world$iso3c==“HTI” |
                            world$iso3c==“KIR” |
                            world$iso3c==“SLB”, “#bf0000”, “grey”)) +
  geom_smooth(method=“lm”, col=“grey”, se=FALSE,
              linetype = “dashed”) +
  ggtitle(“Figure 2-4: Small Countries Have Low GDP per Capita”) +
  ylab(“GDP per Capita (logged)”) +
  xlab(“Ethnolinguistic Fractionalization”) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = “bold”),
        axis.title = element_text(size = 8, face = “bold”)) +
  geom_smooth(data=subset(world,
                          aclpregion != “Industrial Countries”),
                          color=“black”, se=FALSE) +
  geom_text_repel(size = 2.8, aes(label=ifelse(iso3c==“COM” |
                                                 iso3c==“HTI” |
                                                 iso3c==“KIR” |
                                                 iso3c==“SLB”,
                                          as.character(iso3c),’’)))
```

Since we may want to avoid summarizing the relationship between GDP per capita and ELF based on the small island nations of Comoros, Kiribata, Haiti, and the Solomon Islands, there does seem to be a negative relationship between GDP per capita (logged) and the ELF score as we travel from left to right across the x axis. There is, in other words, some preliminary evidence that the hypothesis has some merit. Nevertheless, a great deal of variation remains and the relationship is by no means perfectly **linear**. By linear, I mean a straight line with a single slope.

Before we move on to exploring the next hypothesis, let's take stock to highlight the process of formulating hypotheses and describing data. We started the section with a hunch about the ethnic makeup of a country and its economy. A simple scatter plot showed there might be a negative relationship between the two. We identified a cluster of countries that explains the downward sloping line. We then fit a curve to the remaining data and found there was an upside-down U-shaped pattern: countries with low and high levels of ethnolinguistic heterogeneity had lower levels of income compared to countries with medium levels. Further investigation revealed that the countries with low levels of income and low heterogeneity were small island nations. That observation suggested the geographic size of a country might matter.

Throughout this section, we saw how identifying different clusters of data led to new discoveries and suggested additional hypotheses to test.

Hypothesis II: Women’s Suffrage

Another hypothesis concerns women and their role in both government and the economy. Are women being educated? Do they have equal rights? We might hypothesize that women’s participation in society is essential to economic development. As a very crude measure of women’s participation, we have data—a variable in the *world* data set—that record the year women obtained the right to vote. What is the relationship between the year women gained suffrage and GDP per capita? To answer the question, consider the following scatter plot that plots GDP per capita against the year women gained suffrage in the country (Figure 2-5).

Code Chunk 2-5

```
ggplot(world, aes(womyear, log(gdppc))) +
  geom_point(col=“#bf0000”) +
  geom_smooth(method=“lm”, se=FALSE, color=“#0000bf”) +
  ggtitle(“Figure 2-5: The Later the Suffrage,
              the Poorer the Country”) +
  ylab(“GDP per Capita (logged)”) +
  xlab(“Year Women Gained Suffrage”) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = “bold”),
         axis.title = element_text(size = 8, face = “bold”))
```

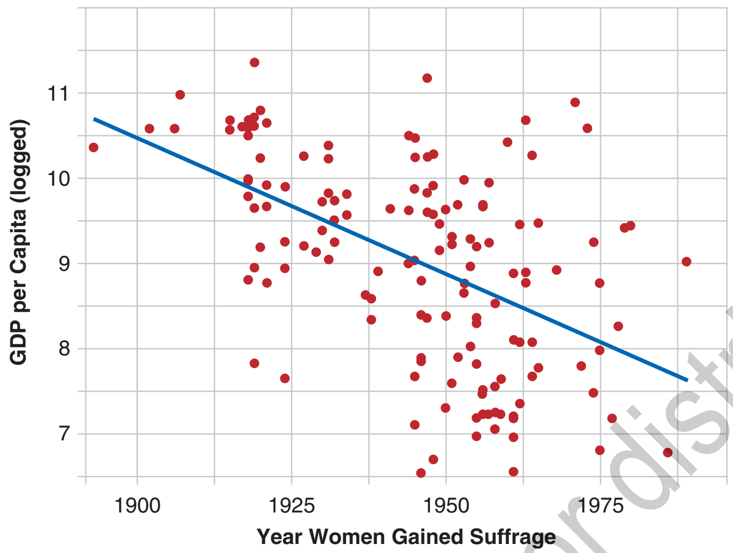
Although GDP per capita can take on very different values at each level of women’s suffrage (there is considerable variation), there is a downward sloping trend. In other words, as the year when women were granted suffrage increases (suffrage was granted later), it appears that GDP per capita decreases. What is driving the result? Perhaps the downward sloping line merely describes the difference between the industrialized West and Africa. The vote was not extended to anyone in Africa, after all, until independence was achieved in the 1960s.

In Code Chunk 2-6, only a few refinements are added to the code in Code Chunk 2-5. First, we use the *ifelse()* command to color the industrial countries blue and the sub-Saharan countries red. Second, we use the *annotate()* function to place the “Industrial Countries” and “Sub-Saharan Africa” labels in the figure.

If the relationship is simply the difference between the industrial countries[5] and sub-Saharan Africa, there may be things other than women’s participation that distinguish the industrial

[5] There is no good way to name this group of countries: the West, the industrialized West, developed countries, postindustrialized, high-income countries, and so forth. Rather than enter into a debate on nomenclature, I simply use the classification given by Przeworski et al. (2000) whose data are used here. The countries considered industrial are Australia, Austria, Belgium, Canada, Switzerland, Cyprus, Germany, Denmark, Spain, Finland, France, United Kingdom, Greece, Ireland, Italy, Japan, Lichtenstein, Luxembourg, Malta, Netherlands, Norway, New Zealand, Portugal, San Marino, Sweden, and the United States.

FIGURE 2-5 The Later the Suffrage, the Poorer the Country

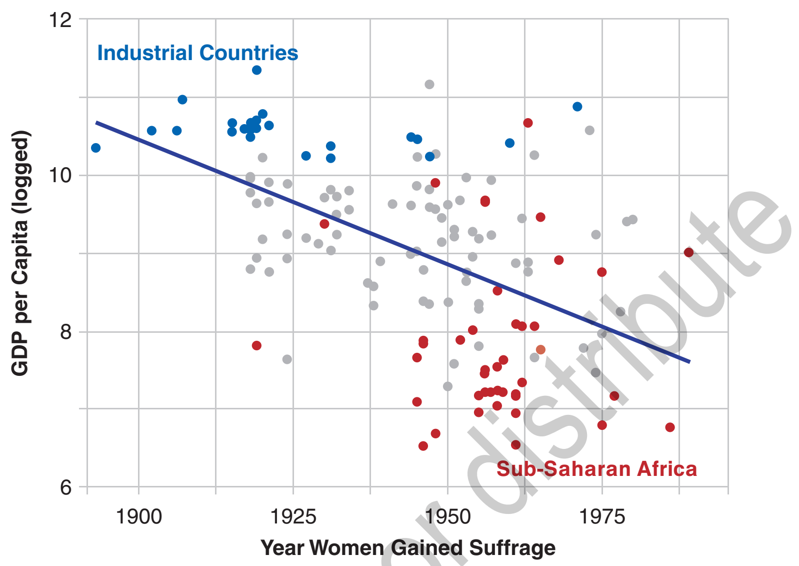


Code Chunk 2-6

```
ggplot(world, aes(womyear, log(gdppc))) +
  geom_point(color=ifelse(world$aclpregion==
                          “Industrial Countries”, “#0000bf”,
                          ifelse(world$aclpregion==
                          “Sub-Saharan Africa”, “#bf0000”, “grey”))) +
  annotate(“text”, x = 1905, y = 11.75,
           label = “Industrial Countries”,
           col=“#0000bf”) +
  annotate(“text”, x = 1975, y = 6.2,
           label = “Sub-Saharan Africa”, col=“#bf0000”) +
  geom_smooth(method=“lm”, se=FALSE) +
  ggtitle(“Figure 2-6: The Industrial Countries and
          Sub-Saharan Africa Form Clusters”) +
  ylab(“GDP per Capita (logged)”) +
  xlab(“Year Women Gained Suffrage”) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = “bold”),
        axis.title = element_text(size = 8, face = “bold”))
```

countries and sub-Saharan Africa from each other (e.g., a colonial past). The cluster of industrial countries in the upper-left quadrant of the plot and the cluster of sub-Saharan African countries in the lower-right quadrant certainly suggests that the date of independence or colonialism is part of this story (see Figure 2-6).

■ FIGURE 2-6 The Industrial Countries and Sub-Saharan Africa Form Clusters



Although we've made some important discoveries so far, let's not stop here. Part of the challenge and fun of data analysis is to develop a keen eye that notices every nook and cranny. Many times, that's how the most interesting discoveries are made. Note, for example, there are relatively few observations (dots) between 1935 and 1945 on the x axis. This implies that countries stopped granting suffrage during those years.

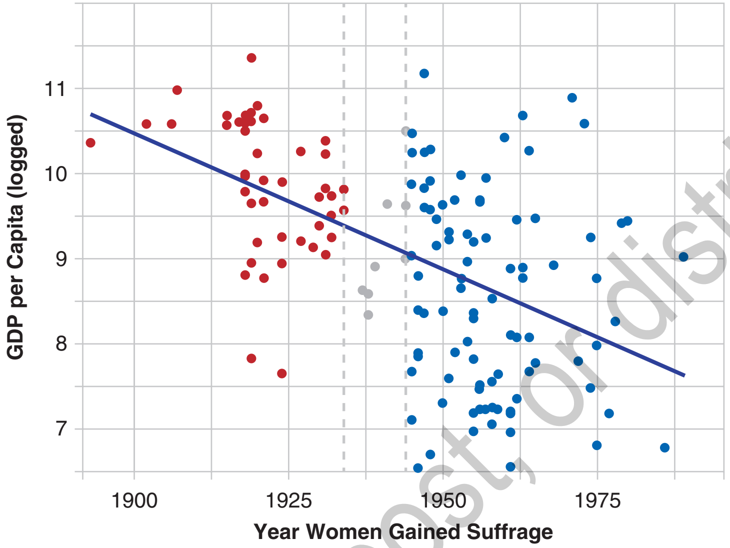
In Code Chunk 2-7, I introduce the *geom_vline()* function, which creates a vertical line on a specific part of the x axis. This is helpful when we want to highlight a very specific value.

Code Chunk 2-7

```
ggplot(world, aes(womyear, log(gdppc))) +
  geom_point(color=ifelse(world$womyear < 1935, “#bf0000”,
                          ifelse(world$womyear > 1944,
                                 “#0000bf”, “grey”))) +
  geom_smooth(method=“lm”, se=FALSE) +
  geom_vline(xintercept=1934, col=“grey”, linetype = “dashed”) +
  geom_vline(xintercept=1944, col=“grey”, linetype = “dashed”) +
  ggtitle(“Figure 2-7: Gap During World War II”) +
  ylab(“GDP per Capita (logged)”) +
  xlab(“Year Women Gained Suffrage”) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = “bold”),
        axis.title = element_text(size = 8, face = “bold”))
```

The pause between 1935 and 1945 might be a feature of the data worth exploring. The gap during the period suggests that the fight for women's rights took a back seat to fighting World War II (see Figure 2-7). To illustrate, I colored the points before 1935 red and the points after 1945 blue. After the war, a number of countries joined the community of nations

FIGURE 2-7 Gap During World War II



Code Chunk 2-8

```
ggplot(world, aes(womyear, log(gdppc))) +
  geom_point(color=“#bf0000”) +
  geom_smooth(method=“lm”, se=FALSE, color=“#0000bf”) +
  ggtitle(“Figure 2-8: Large Variation in GDP per Capita”) +
  ylab(“GDP per Capita (logged)”) +
  xlab(“Year Women Gained Suffrage”) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = “bold”),
        axis.title = element_text(size = 8, face = “bold”)) +
  geom_text_repel(size = 2.8,
                  aes(label=ifelse(iso3c==“ETH” |
                                      iso3c==“BDI” |
                                     iso3c==“SLE” |
                                     iso3c==“HUN” |
                                      iso3c==“CYP” |
                                      iso3c==“LBY”,
                                   as.character(iso3c),’’),
                      hjust = 0, vjust=-1), show.legend=FALSE)
```

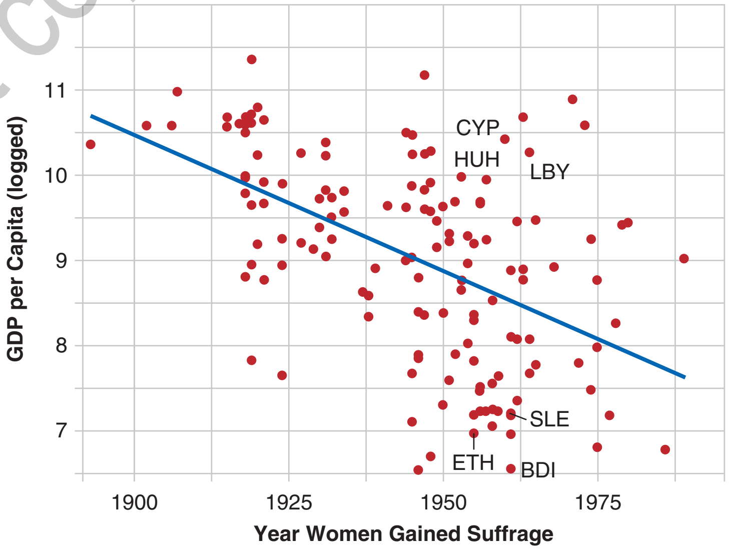
as independent entities. At their birth, many granted voting rights to women. Clearly, GDP per capita could be driven by the political autonomy of a country rather than women's rights. Both the two clusters of countries and the gap separating pre- and postwar cases suggest that colonialism and independence explains levels of GDP per capita.

Another feature of the relationship involves the huge variation in per capita income for countries that granted suffrage at around the same time. By huge variation, I mean that at very similar dates when suffrage was granted, levels of GDP per capita range from being fairly high to fairly low. Note that Ethiopia (ETH), Burundi (BDI), and Sierra Leone (SLE) granted suffrage at around the same time as Hungary (HUN), Cypress (CYP), and Libya (LBY), yet the two groups of countries have very different levels of per capita income (Figure 2-8). Clearly, women's suffrage alone does not explain everything. If it did, all of the countries plotted in the graph would be much closer to the line.

Refer back to Figure 2-6, which had the industrial countries in blue, the sub-Saharan African countries in red, and the rest of the world's regions in gray. Given the pattern we observed between the industrial countries and Africa in Figure 2-6, a very useful test might be to fit a line to the sub-Saharan countries separately to see if a relationship exists between women's suffrage and income per capita for sub-Saharan Africa. By doing so, we're asking whether the negative relationship between GDP per capita and the year of suffrage is indeed negative or if it is purely generated by the differences between the industrial countries and sub-Saharan Africa. If there is a strong negative relationship between GDP per capita and women's suffrage, then we would expect to find it among the sub-Saharan African countries themselves.

To distinguish different groups in a figure, it's often helpful to create a variable that indicates a 1 for a specific group and a 0 for all of the rest. The two first lines in Code Chunk 2-9 create such a variable for sub-Saharan Africa. First, I use the *ifelse()* function to state that if the variable *world$region* equals sub-Saharan Africa, then generate a 1; otherwise generate a 0. I then take that new variable *world$SSA* and convert it to a factor. I'll go into more detail on factors in Chapter 5.

■ **FIGURE 2-8** Large Variation in GDP per Capita



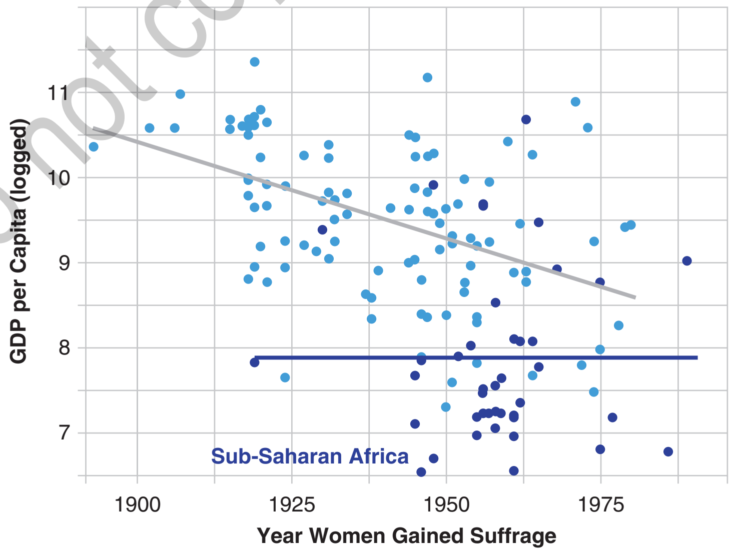
Code Chunk 2-9

```
world$SSA <- ifelse(world$region==“Sub-Saharan Africa”, 1, 0)
world$SSA <- as.factor(world$SSA)

ggplot(world, aes(womyear, log(gdppc), col=SSA)) +
  geom_point() +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = “bold”),
          axis.title = element_text(size = 8, face = “bold”)) +
  geom_smooth(method=“lm”, se=FALSE) +
  ggtitle(“Figure 2-9: No Relationship in Sub-Saharan Africa”) +
  ylab(“GDP per Capita (logged)”) +
  xlab(“Year Women Gained Suffrage”) +
  scale_color_manual(values=c(“lightblue”, “#0000bf”)) +
  theme(legend.position = “none”) +
  annotate(“text”, x = 1928, y = 6.8,
             label = “Sub-Saharan Africa”, col=“#0000bf”)
```

In Figure 2-9, the sub-Saharan countries are colored dark blue. As the figure demonstrates, there does not seem to be a relationship between women's suffrage and economic development in sub-Saharan Africa (the line is flat). In the rest of the world, however, the relationship seems strongly negative. Sub-Saharan Africa is different in this respect and accounting for it helps clarify the strong relationship that exists between GDP per capita and women's suffrage in the other countries.

■ **FIGURE 2-9** No Relationship in Sub-Saharan Africa



Think about what this implies for our investigation. Women's suffrage occurred simultaneously with independence and democratization in many countries. To the degree that other things are intertwined with women's suffrage (independence, democratization, or both), a measure that merely captures the year women were allowed to vote may only be the result of a larger process. Consequently, these other important events—independence and democratization—need to be accounted for in our explanation.

Before we examine the next hypothesis, let's pause here again to take stock of how the process of formulating hypotheses and describing data unfolds. Careful observation of simple scatter plots provides considerable information. In this example, we again took advantage of coloring the observations by region to make additional discoveries. We first explored whether the negative relationship between income and the year women were granted suffrage could be explained by the differences between industrial countries and sub-Saharan Africa. We also noticed a slight gap in the data during World War II. This suggested that other things might explain the relationship between women's suffrage and income. Finally, we examined whether the negative relationship among all the observations existed within the sub-Saharan cases. Throughout the process, we can see how describing data and formulating hypotheses interact.

Good descriptions of the data can help us evaluate hypotheses as well as formulate new ones. With one scatter plot, we went from evaluating the hypothesis that women's suffrage influences economic growth to discovering how a country's independence or move toward democracy should be considered as well. Simple views of the data may provide what might seem to be an endless number of patterns, discrepancies, or oddities that are worth investigating. As the last few figures demonstrate, important insights can be gained by thoroughly examining each plot. I hope to have demonstrated here that a simple scatter plot can tell us a lot about a relationship if we simply take the time to investigate, coloring groups of data differently to learn what cases determine the slope of a line or the shape of a curve. Spotting interesting aspects of the data through visualization in different ways is a useful skill worth developing.

Hypothesis III: Human Capital

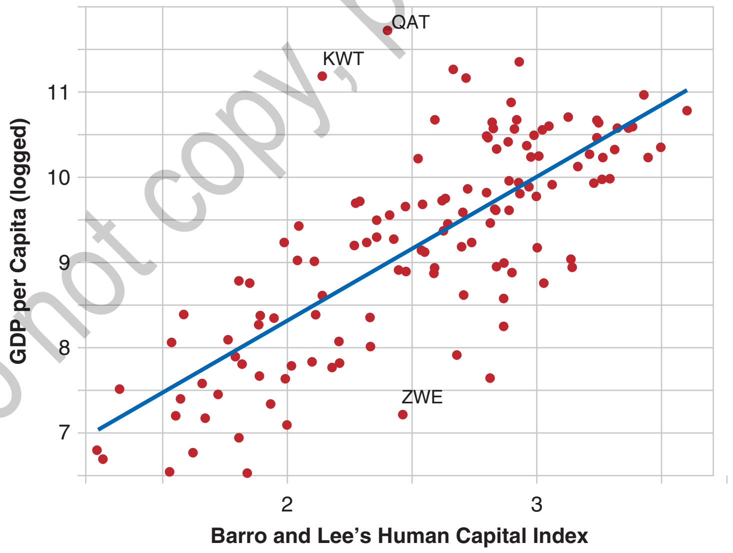
Continuing on, an important component of economic growth is human capital: education. As a population accumulates knowledge, economists argue economic development follows. How does a measure of human capital (an index of human capital constructed by Robert Barro and Jong-Wha Lee) relate to GDP per capita? According to Figure 2-10, it appears the relationship is strong and positive: the more human capital per worker, the higher GDP per capita. Of course, a developed economy may allow more people to attend school: the direction of causality may run the other way (from wealth to education). Nevertheless, we can say there's a strong association between the two variables. Note how the observations in this figure are much closer to the line, which indicates our predictions of GDP per capita at each level of the human capital index. Also observe how the shape of this plot is better **summarized** by a line than the previous two plots. Note my use of the word "summarize." We use summaries to characterize the data. When we use a line, we summarize the data with two numbers: the intercept of a line and its slope.

The code in Code Chunk 2-10 should look familiar by now. Along with the *ggplot()* function that establishes a grid with two dimensions (GDP per capita and human capital), it specifies a layer of dots with the function *geom_point()* and a line with the function *geom_smooth()*. It also uses the *ifelse()* function to label the cases of Qatar (QAT), Kuwait (KWT), and Zimbabwe (ZWE).

Code Chunk 2-10

```
ggplot(world, aes(pwthc, log(gdppc))) +
  geom_point(col="#bf0000") +
  geom_smooth(method="lm", se=FALSE, col="#0000bf") +
  ggtitle("Figure 2-10: Strong Relationship Between
          GDP and Education") +
  ylab("GDP per Capita (logged)") +
  xlab("Barro and Lee’s Human Capital Index") +
  geom_text_repel(size = 2.8,
       aes(label=ifelse(iso3c=="QAT" |
                          iso3c=="KWT" |
                          iso3c=="ZWE",
                       as.character(iso3c),’’),
                       hjust = 0, vjust=-1),
      show.legend=FALSE) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold"))
```

■ FIGURE 2-10 Strong Relationship Between GDP and Education



Besides the distinct linear pattern, a noticeable feature of the plot is the outliers, the cases far removed from the rest of the data. For example, Qatar (QAT) and Kuwait (KWT) stand out as having been severely underpredicted—they lie considerably above the line (the line represents our prediction). Given the levels of the human capital index, we would expect GDP per capita to be much lower in those countries. Of course, both have significant proven oil reserves, perhaps accounting for our inaccurate prediction. Note also one of the cases lying below the line, Zimbabwe (ZWE). At Zimbabwe's level of human capital, we would have predicted a higher level of per capita income. In Zimbabwe's case (a primarily agricultural country), growing corruption in the 1990s and 2000s along with unstable property rights—disputes over land—might help explain its underachievement. Even though the human capital index does fairly well in explaining GDP per capita, the outliers help identify other possible explanations: natural resources, property rights, and politics.

In this section, we learned that identifying specific cases suggests possible hypotheses we might want to test in the future. Outliers alert us to important causal factors that we hadn't previously considered. In this example, while investigating the importance of human capital's role in determining a country's income, we found that oil (Qatar and Kuwait) and property rights (Zimbabwe) might contribute to income levels as well. Again, we see how describing data generates additional hypotheses.

Hypothesis IV: Political Stability

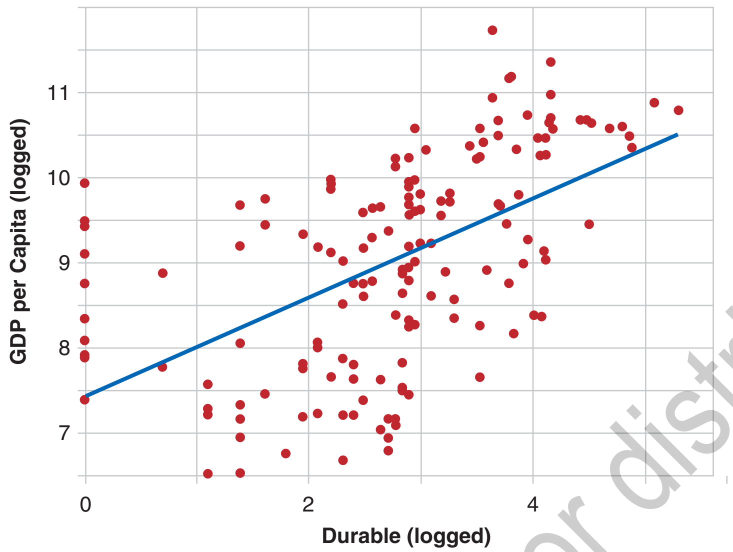
What role does politics play in the development of national economies? Countries with unstable regimes are probably not the best venues for trade and investment. Is political instability associated with lower levels of wealth? To consider that possibility, GDP per capita is plotted (Figure 2-11) against the number of years since a significant change in the regime (the variable *durable*). The *durable* variable is logged for the same reasons we logged GDP per capita: transforming variables by taking logs will be explained in detail when I cover transforming data in Chapter 6. There does seem to be a positive relationship between regime stability and GDP per capita—countries with relatively stable regimes experience high levels of income (Figure 2-11). As we move from left to right along the x axis, the level of income appears to increase.

Code Chunk 2-11 lists the set of commands used previously to draw a scatter plot. The only wrinkle involves logging the *durable* variable. Since the *durable* variable contains some zeros, we add a 1 to the variable before logging since the log of 0 is undefined. There'll be more on logging and transforming variables in Chapter 6.

Code Chunk 2-11

```
ggplot(world, aes(log(durable + 1), log(gdppc))) +
  geom_point(col="#bf0000") +
  theme_minimal() +
  geom_smooth(se=FALSE, col="#0000bf", method="lm") +
  ggtitle("Figure 2-11: Political Stability Increases Wealth") +
  ylab("GDP per Capita (logged)") +
  xlab("Durable (logged)") +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold"))
```

FIGURE 2-11 Political Stability Increases Wealth



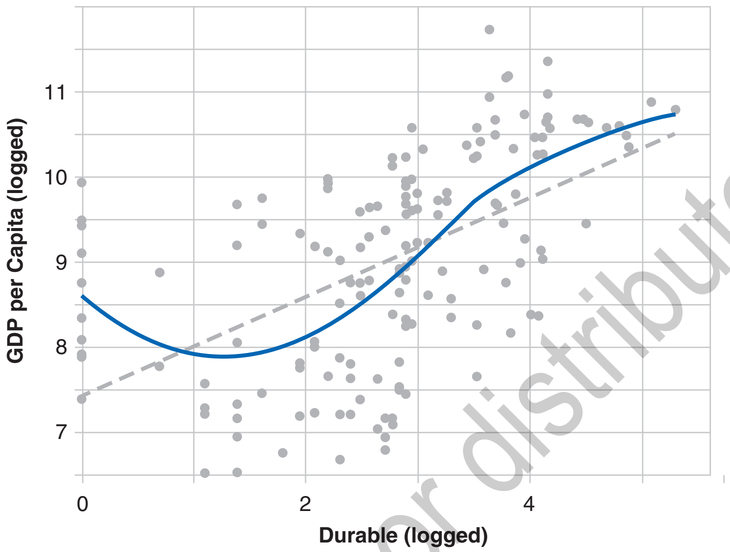
While there does seem to be a positive relationship between stability and income, the blue line may not be the best way to summarize the relationship. In particular, notice the points at 0 on the x axis. Save for one, they are all located above the line. Also notice that as we move from left to right along the x axis between 0 and 2, most of the observations are located below the line. Then, above 3, most of the points can be found above the line.

To compare a straight line with a curved line, I added two lines to the figure with the *geom_smooth()* function. One of the layers simply draws a curved line to the data (the default), while the other specifies *method=“lm”*, which tells R to fit a straight line to the data (Code Chunk 2-12).

Code Chunk 2-12

```
ggplot(world, aes(log(durable + 1), log(gdppc))) +
  geom_point(col=“grey”) +
  geom_smooth(se=FALSE, col=“#0000bf”) +
  geom_smooth(method=“lm”, se=FALSE,
                linetype = “dashed”, col = “grey”) +
  ggtitle(“Figure 2-12: The Nonlinear Relationship Between
           GDP per Capita and Political Stability”) +
  ylab(“GDP per Capita (logged)”) +
  xlab(“Durable (logged)”) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = “bold”),
        axis.title = element_text(size = 8, face = “bold”))
```

■ FIGURE 2-12 The Nonlinear Relationship Between GDP per Capita and Political Stability



In this example, the shape of the relationship may be best described by an s-shaped curve. At the low end of *durable*, income is high (higher than our line predicts).[6] As we move from left to right on the x axis, income per capita increases sharply and we begin to systematically underpredict the level of GDP per capita (the majority of cases lie above our prediction on the right portion of the figure).

Although the curvature of the s-shape is not severe, patterns like this suggest there are thresholds. In this example, the s-shape suggests that economic development does not take off until several years of political stability are achieved. At the high end of political stability, the s-shape indicates that each additional year of political stability change does not improve economic development appreciably. In other words, once the same set of constitutional rules have been in force for a number of years, each additional year does not have the same positive influence.

While the relationship between regime stability and GDP per capita may not be strictly linear (it may not be a straight line) and there remains plenty of variation, there is a **positive association** between the two variables. In other words, as regime stability increases, so does GDP per capita: the two variables increase together. Regime stability might be another variable we should consider when trying to explain why some countries are rich and some are poor.

To summarize, in this section we saw another example of how describing data can inform formulating hypotheses and vice versa. In earlier examples, we saw that certain techniques, such as identifying clumps of data as well as individual outliers, help generate additional hypotheses we might want to test. This example demonstrates that examining the functional form of the

[6] Since the log of 0 is undefined, I added 1 to the durable variable before taking logs since some of the cases registered 0, a common trick when transforming data. Notice the cases at 0 along the x axis; they represent the cases that were 0 in raw form but then registered 1 when I added the constant 1 to the variable. The log of 1 is 0.

relationship—whether the observations form a straight line or s-shaped curve—also informs our investigation. Now that we have a better sense of what the data look like and we've interrogated a few likely suspects, it's time to move on to model building and estimation.

KNOWLEDGE CHECK: Form hypotheses from descriptions of data.

9. Which of the following accurately describe the process of forming hypotheses?
   a. We always enter into the exercise with some preconceived notions or priors.
   b. Never enter into the exercise with some preconceived notions or priors.
   c. Adding different colors and labeling cases can help us generate hypotheses.
   d. Don't start describing the data until you have generated a hypothesis to test.
10. When evaluating the hypothesis that ethnolinguistic heterogeneity explains income levels, what features of the data generated new hypotheses?
    a. Incomes are high in Europe.
    b. The industrialized countries of the world are clustered together.
    c. The relationship may not be linear.
    d. There are some poor island nations that have low heterogeneity.
11. What features of the data generated the hypothesis that independence has an important role to play in explaining GDP per capita?
    a. There is a gap in the data during World War II.
    b. There is a clear linear relationship between when women were allowed to vote and GDP per capita.
    c. Small island nations were early in allowing women to vote.
    d. There is a linear relationship between income and human capital.
12. What visualization techniques were used to make discoveries in this section?
    a. Representing groups of data with different shapes
    b. Representing groups of data with different colors
    c. Singling out specific cases with labels
    d. Singling out specific cases with color
13. What did you notice in the preceding section in terms of describing data and generating hypotheses?
    a. Viewing the data can prove hypotheses.
    b. Describing data usually settles the matter.
    c. Describing the data often amends the hypothesis or leads to new questions.
    d. Hypotheses are rarely proven true or completely dismissed by useful descriptions of the data.

Model Building and Estimation

Once the data are described and hypotheses have been generated, it's time to develop our argument or to construct a model. We're now moving into material that will be covered in Chapters 10–14. Regression tables will appear in this section without an in-depth discussion of their mechanics. They are presented here to help identify some important tools you will develop that figure prominently in the process of data analysis.

The scatter plots we've examined so far suggest that there are multiple causes of economic development. Ethnolinguistic heterogeneity, women's suffrage, human capital, and regime stability all seem to matter. To obtain the best estimate of their relationship with GDP per capita, we need a good model. For our purposes, let's define a model as simply a list of factors we want to include in an explanation. Well-reasoned hypotheses and a familiarity with the data should be combined toward that end. A good model identifies only the essential elements that help accurately predict outcomes—the outcome in this example is a country's GDP per capita.

Before we continue, it is important to acknowledge that both scholars and practitioners disagree on how to build models. Important philosophical debates lurk beneath the surface. Put simply, the debate concerns whether theory or the data should be our guide. One side doesn't trust the data, arguing that knowing all of the facts in the world does us no good unless we know how things fit together—the theory. The other side doesn't trust the theory, arguing that the best theory comes from knowing as much about the world as possible—the data. Devotion to one approach to the exclusion of the other, in my opinion, often misleads. Relying on a conversation between the two is more fruitful. So far, we've seen how looking at the data can inform theory. Next we find how theory can influence the conclusions drawn from the data. Recognizing how both approaches work in practice is a primary concern of this book.

Start with a simple model: a model that uses one factor—the year women were granted suffrage—to explain GDP per capita (Table 2-1). Models with two variables, a dependent variable (GDP per capita) and an independent variable (women's suffrage; *womyear*), are called **bivariate regression models**. Models with more than two variables—a dependent variable (GDP per capita) and more than one independent variable—are called **multiple regression models**. In the bivariate case, the model uses a line to summarize the relationship between GDP per capita and women's suffrage without accounting for anything else.

The code in Code Chunk 2-13 defines an object called *mod1* and then presents the results from *mod1* in a regression table using the *stargazer()* function from the 'stargazer' package. The *mod1* object is defined as a simple bivariate regression model using the *lm()* command. The *lm* in the *lm()* function stands for *linear model*. In the regression, the log of GDP per capita (*gdppc*) is regressed on the year women were granted suffrage in a country (*womyear*). An important convention in statistics is to say that the dependent variable *is regressed on* the independent variable.

Code Chunk 2-13

```
mod1 <- lm(log(gdppc)~womyear, data=world)

stargazer(mod1, header=FALSE,
          title = “Table 2-1: Estimates for a Bivariate Model”,
          type = “html”, out = “table3.htm”)
```

TABLE 2-1 Estimates for a Bivariate Model

	DEPENDENT VARIABLE:
	log(gdppc)
womyear	–0.032***
	(0.005)
Constant	70.750***
	(8.831)
Observations	143
R^2	0.257
Adjusted R^2	0.252
Residual SE	1.046 (*df* = 141)
F statistic	48.874*** (*df* = 1; 141)

Note: $^{***}p < 0.01$.

The stargazer function requires the right kind of an object as an argument. In this case, the object *mod1*—defined by a linear model—is appropriate. We can also specify the title of the table, the type of output (HTML in this case), and a file where the output will appear. Note that when you knit to an HTML file or PDF, a nice table will appear. Unfortunately, as of this writing, stargazer does not work when knitting a Microsoft Word document.

With the bivariate model, we find there is a **negative association** between GDP per capita and women's suffrage—the number (–.032). In other words, as the women's suffrage variable increases (the later the date suffrage was granted), GDP per capita decreases. The number (–.032) represents the slope of the line that describes the relationship between women's suffrage and GDP per capita.

Before moving to the next section, let's recap. After describing the data and exploring a few hypotheses, we constructed a model to explain GDP per capita. We first constructed a very simple model that fit a line summarizing the relationship between GDP per capita and the year women were granted suffrage in a country. We found that the slope of that line (our estimate) was –.032, implying that women's participation in politics does play an important role in economic development.

KNOWLEDGE CHECK: Explain the connection between hypotheses, models, and estimates.

14. Which of the following statements about model building are true?
    a. Theory should always come before the facts.
    b. The facts should always come before the theory.
    c. There should be a conversation, a back-and-forth between theory and facts.
    d. In the author's opinion, there should be a conversation, a back-and-forth between theory and facts.

15. Indicate which statements are true.

    a. Models with two variables are called bivariate regression models.

    b. Models with more than two variables are called bivariate models.

    c. To obtain the best estimate of a relationship between two variables, we need a good model.

    d. Models with more than two variables are called multiple regression models.

Diagnostics

Now that we've obtained estimates, we need to ascertain how much trust to place in our results. Tables with model estimates can evince an air of confidence and authority. Those estimates aligned so neatly in nice-looking columns and rows could be based on some questionable assumptions, outlying cases, or very poor measures. Much like when being confronted by a pushy salesman who leads you over to a bright, shiny, Corvette on a used-car lot, it might be good to check under the hood. Diagnostics provide the tools to determine whether our estimates are reasonable or highly misleading. Will the Corvette go from 0 to 60 in less than 4 seconds or will it die once we've driven it off the lot?

Stability of the Results

Checking the **stability** of the results and whether we've met certain assumptions are only two possible ways to perform diagnostics. Chapters 15 and 16 are devoted to the enterprise. Here, I simply want to illustrate how our choice of model (theory) can influence our estimates of the relationship (the data).

To illustrate, I estimate a multiple regression model with the *lm()* function and then construct a table displaying the results using the *stargazer()* command (Code Chunk 2-14).

Code Chunk 2-14

```
mod2 <- lm(log(gdppc)~womyear + ethfrac + durable + pwthc,
          data=world)

stargazer(mod2, header=FALSE,
          title = “Table 2-2: Estimates Are Unstable”,
          type = “html”, out = “table4.htm”)
```

How stable are the results from the bivariate model? By stability, I mean how much the results change when we alter the model slightly. If small changes in the model (adding a few variables) produce noticeable changes, we say the results are unstable. Remember that in the bivariate regression, the slope was –.032. Once other variables are added (see Table 2-2), the slope on women's suffrage changed to –.001, decreasing in magnitude by a factor of 32! Clearly, our estimate of women's suffrage depends on the particular model we use.

To see how this translates visually, I generate a plot that shows how women's suffrage is related to GDP per capita when we account for human capital, political instability, and ethnolinguistic fractionalization. This is called an **added variable plot**, which shows the relationship

TABLE 2-2 Estimates Are Unstable

	DEPENDENT VARIABLE:
	log(gdppc)
womyear	−0.001
	(0.004)
ethfrac	−0.698**
	(0.291)
durable	0.007***
	(0.002)
pwthc	1.416***
	(0.163)
Constant	6.682
	(8.201)
Observations	105
R^2	0.725
Adjusted R^2	0.714
Residual SE	0.657 (*df* = 100)
F statistic	65.813*** (*df* = 4; 100)

Note: $^{**}p < .05$; $^{***}p < 0.01$.

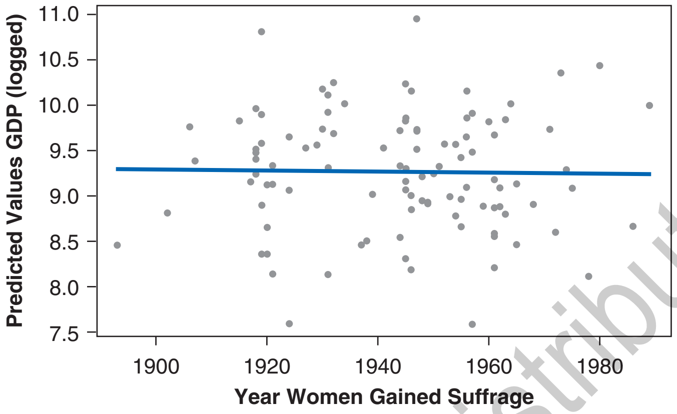
between two variables when other variables are accounted for. Again, you'll learn how to generate this very useful view of the data in Chapters 15 and 16. Until then, consider the relationship between GDP per capita (logged) and women's suffrage when all of the variables are included in the model (Figure 2-13).

An easy way to visualize results from regression analysis involves using an R package called 'visreg.' In Code Chunk 2-15, I use the *visreg()* function to plot the predicted values from the multiple regression model I defined in Code Chunk 2-14. In this example, I want to visualize the relationship between women's suffrage and GDP per capita, accounting for the other variables in the model. The 'visreg' package makes this easy. I simply indicate which model I want to use (*mod2*, in this case) and specify the independent variable I'm interested in (*womyear*).

Code Chunk 2-15

```
visreg(mod2, “womyear”, ylab=“Predicted Values GDP (logged)”,
       xlab=“Year Women Gained Suffrage”,
       main=“Figure 2-13: Women’s Suffrage Has No Effect”,
       band=FALSE)
```

FIGURE 2-13 Women's Suffrage Has No Effect



The blue line in the figure represents what level of GDP per capita model 2 predicts based on women's suffrage. The dots represent the actual values as they exist for each country.

While this doesn't necessarily settle the issue, the added variable plot suggests there is not much of a relationship between model 2's prediction of income (GDP per capita) and women's suffrage when controlling for the other variables. Once human capital, political stability, and ethnolinguistic fractionalization have been accounted for, knowing when women were granted suffrage doesn't really tell us much about the level of GDP per capita. This exercise suggests that the relationship between GDP per capita and women's suffrage is *unstable with respect to* the inclusion of political stability, ethnolinguistic fractionalization, and human capital in our model. It illustrates an important lesson: our estimates depend on the model we choose.

Residual Plots

Finally, we can check whether our predictions get better or worse as we range over our predictions, from low to high levels of GDP per capita. If our predictions are systematically better or worse for rich or poor countries, we know our model could be improved. A **residual plot** (Figure 2-14) tells us very quickly if the difference between each case and our prediction (what we call the residual) follows any pattern. Figure 2-14 plots the residuals against the predictions. The horizontal dashed line represents perfect predictions. If a case rests on that line, our model has correctly predicted the actual level of income for that country. Cases further away from the dashed line, such as Singapore (SGP), are those that the model predicted poorly. This residual plot indicates that the model severely underpredicted the level of income in Singapore and Luxembourg.

To produce residuals, we first need to estimate the model and collect the residuals. In Code Chunk 2-16, you can see I use the *lm()* function again to estimate a linear model. This time it includes more than one independent variable; it includes four: *womyear*, *ethfrac*, *durable*, and *pwtch*. Once the model is estimated, I create two variables (*world$res* and *world$pred*) using the *resid()* function and the *predict()* function.

Code Chunk 2-16

```
mod2 <- lm(log(gdppc)~womyear + ethfrac +
             durable + pwthc, na.action = na.exclude,
           data=world)

world$res <- resid(mod2)
world$pred <- predict(mod2)
```

Now that I have a variable with the predicted values from the model and the residuals, I'm ready to plot the two against each other. I do this using the code in Code Chunk 2-17. Code Chunk 2-17 defines a scatter plot that colors the points red, labels the points of five countries using the *ifelse()* function, and draws a horizontal line that represents the predicted values (where the residuals equal 0).

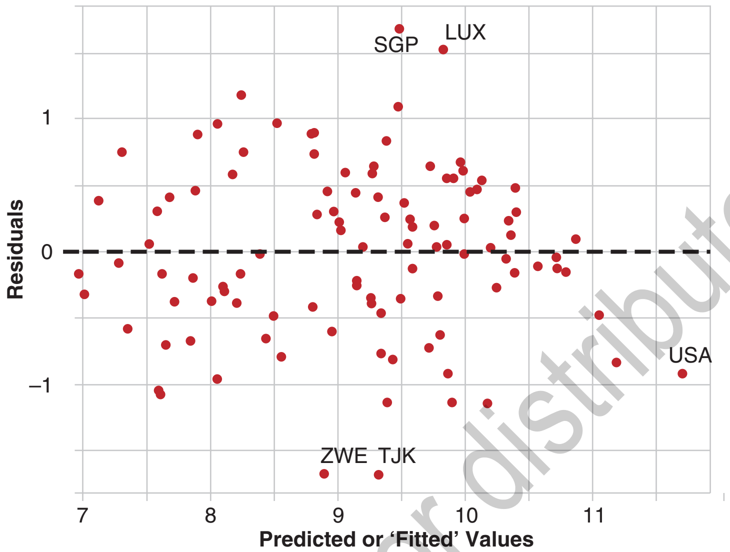
Code Chunk 2-17

```
ggplot(world, aes(pred, res)) +
  geom_point(col=“#bf0000”) +
  geom_text_repel(size = 2.8,
       aes(label=ifelse(iso3c==“SGP” |
                           iso3c==“TJK” |
                           iso3c==“ZWE” |
                           iso3c==“USA” |
                           iso3c==“LUX”,
                        as.character(iso3c),’’),
        hjust = 0, vjust=-1), show.legend=FALSE) +
  geom_hline(yintercept = 0, linetype = 2) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = “bold”),
        axis.title = element_text(size = 8, face = “bold”)) +
  xlab(“Predicted or ‘Fitted’ Values”) +
  ylab(“Residuals”) +
  ggtitle(“Figure 2-14: Some Possible Outliers”)
```

Here we look for patterns. As we'll learn in subsequent chapters, an important assumption in model estimation is that our predictions are just as good or bad for low or high values. In this example we're making the assumption that the model's accuracy does not change when we're making predictions for relatively poor countries or relatively rich ones. The plot in Figure 2-14 can be characterized as a formless cloud: our predictions are just as good or bad as we make predictions for rich and poor countries. There are no obvious observable patterns. From this we would conclude that an important assumption of our analysis is met.[7]

[7] An explanation of the assumptions associated with ordinary least squares (OLS) regression analysis, which is used in this example, will be presented in the chapter on diagnostics (Chapter 15).

■ FIGURE 2-14 Some Possible Outliers



The residual plot also indicates if any outliers exist (cases far removed from the rest of the data). With this regression model, Singapore (SGP), Luxembourg (LUX), Tajikistan (TJK), the United States (USA) and Zimbabwe (ZWB) are potential outliers. These outlying cases are ones our model did not predict very well. Why? As we learned with the scatter plots in the previous section, investigating those cases may provide clues as to what other factors might be involved.

KNOWLEDGE CHECK: Define diagnostics and explain their role in data analysis.

16. Which are good analogies for diagnostics?
    a. Checking underneath the hood of a car
    b. Examining the X-ray of a patient
    c. Cross-examining a witness in a trial
    d. Questioning a politician at a press conference
17. What are we looking for when we perform diagnostics?
    a. Do the results show a strong relationship?
    b. Is the hypothesis confirmed?
    c. Do the results change with small changes to the model?
    d. Are there odd cases that are unduly influencing the results?

18. What are we looking for in a residual plot?
    a. Odd cases.
    b. Odd patterns
    c. Any patterns
    d. The typical case

Next Questions

One of the most important aspects of data analysis involves formulating new questions. What do the estimates imply and how can that be tested? In the regression analysis described earlier, including human capital, ethnolinguistic fractionalization, and political stability in the regression changed the estimate of women's suffrage. Why is that? Perhaps women's political rights and the human capital index are closely related (we can test that). Women's influence on GDP might also depend on human capital. The right to vote may have little impact in countries with low levels of education. When women are educated, the right to vote could be extremely powerful. In addition to these possibilities, the model should probably contain other variables (e.g., date of independence, colonial past, and democracy). As we include those variables in our model, the estimates will change, leading to new discoveries.

Asking the next question performs two functions. First, it can help confirm our hypotheses. Generating the next question, forming a hypothesis around it, and testing it can bring new evidence to bear on our original hypothesis. Second, it can lead to new discoveries. Often, we find that the original question we asked was not the most interesting, or that it sent us barking up the wrong tree. In either case, asking the next question is a fundamental aspect of data analysis.

How do we formulate new questions? I find that a helpful way to spark new questions starts with an if-then statement: "if" the results are true, "then" what else would we expect to see in the data? This exercise tests our knowledge of the problem, our ability to think creatively, and our ability to think logically. This is surely part of what we mean by "critical thinking." As with anything, it requires practice. It is a muscle that, if used often, will grow big and strong. If seldom used, it will atrophy.

A final feature of new questions that I'll mention is that they often help us "think outside the box." While this is an overused and ill-defined directive often muttered when the boss is completely out of ideas, thinking outside the box comes directly from asking new questions. In other words, asking new questions *is* thinking outside the box. New questions force us to consider "what if"? When we think outside the box, we're simply exploring beyond our current model—the proverbial box.

KNOWLEDGE CHECK: Formulate new questions.

19. Why do we formulate new questions?
    a. They challenge our old questions.
    b. They force us to reconsider our model.
    c. They provide additional evidence.
    d. They help us think "outside the box."

20. How do we formulate new questions?
    a. Get to know the subject.
    b. Use an if-then statement.
    c. Practice.
    d. Exercise.

SUMMARY

In this chapter, the goal was to motivate the enterprise and to introduce the process of data analysis. The growing amount of information and its use in our everyday lives requires we know how to use and understand quantitative data. In particular, I hope to have shown the back-and-forth between theory and evidence that undergirds the approach taken in the following pages. Although all scholars and practitioners take slightly different approaches, the process outlined in this chapter is a good way to start: (1) describing data and formulating hypotheses, (2) building and estimating models, (3) diagnostics, and (4) generating the next question.

This book is less about probability theory and statistics and more about ways to discover patterns that exist in our data and the inferences we can draw from them. There are many things we want to explain (wealth, inequality, violence, voting, etc.), and there are a lot of data to explore.

We have many tools at our disposal to understand the many characteristics of our data. For example, if our problem is to reduce homicide rates in the United States, it would be helpful to know where the lowest and highest rates are found. Do most states have similar rates or do they vary dramatically? Are similar rates clustered in the South, Northeast, and West? Answering these questions involves describing the data. The more we know about our data, the better questions we'll ask and the better models we'll construct. In Chapter 3, I discuss the importance of describing data and provide examples that illustrate why it is so important.

COMMON PROBLEMS

- *Exploration versus presentation.* Understanding the difference between exploration and presentation is usually not articulated in most books on statistics. Much of what was presented in this chapter shows what goes on behind the scenes. Following paths that ultimately lead nowhere is not usually presented. Exploring data for the purpose of discovery and presenting data for the purpose of persuasion are two very different things.

- *Understanding the residual plot.* Spend time understanding the analytics of the residual plot. The most difficult aspect is to understand that the flat horizontal line represents residuals with a value of 0, indicating that the cases were perfectly predicted. Dots above the line represent instances where the model underpredicted and dots below the horizontal line represent instances where the model overpredicted.

- *Theory versus hypothesis.* Students and practitioners often fall into using these two terms interchangeably. Theories are frameworks or systems of relationships used to explain a variety of different phenomenon. Hypotheses are more specific and tentative. Hypotheses represent our best guesses about how the world works. Those guesses are tested to confirm or challenge existing theory. As evidence accumulates indicating that the theory is incorrect, it's time to change our theory.

- *Patience.* Patience, or rather the lack thereof, is a common theme in data analysis. Students and practitioners get impatient. Their impatience manifests itself in three primary ways during the process of data analysis. First, they often jump past providing a good description of the data right to estimating models. Second, students and practitioners report their results without spending adequate time with diagnostics. Finally, relatively little time is spent thinking about how to present data effectively. As you'll see in Chapter 7, considerations concerning data presentation are crucial.

REVIEW QUESTIONS

1. Describe the struggle between theory and evidence.
2. What are the main components of data analysis?
3. Explain how colors, shapes, and labels aid in data analysis.
4. Why is generating the next question such a useful exercise?
5. What are we looking for when we examine a residual plot?
6. What is meant by stability of the results?
7. How would you describe the grammar of graphics used in ggplot?
8. What is a model?
9. What are diagnostics?
10. What purpose do diagnostics serve?

PRACTICE ON ANALYSIS AND VISUALIZATION

1. What does the *ifelse()* function below say in English?

   ```
   ifelse(world, region ==”Europe”, 1, 0)
   ```

   a. If the *world* variable equals “region,” then assign 1.
   b. If the *world* variable equals “region,” then assign 0.
   c. If *world$region* equals Europe, assign 1 otherwise 0.
   d. If *world$region* equals Europe, assign 0 otherwise 1.

2. Which of the following statements accurately describe the process of data analysis?

   a. Model estimation benefits from describing data.
   b. Diagnostics should always represent the end of our analysis.
   c. Discovery comes once model estimates have been generated.
   d. If a model is properly formulated, diagnostics are not necessary.

3. Which of the following statements accurately describe the purpose of diagnostics?

   a. They help identify important cases.
   b. They indicate whether we are using the appropriate model.
   c. They generate additional questions.
   d. They can help generate additional hypotheses.

4. When we examined the relationship between women's suffrage and income levels, what if-then statement did we generate?

   a. If the relationship between women's suffrage and GDP per capita is just the result of the important differences between Europe and sub-Saharan Africa, then we would expect to see the same relationship exist among the countries of sub-Saharan Africa.
   b. If the relationship between women's suffrage and GDP per capita is not just the result of the important differences between Europe and sub-Saharan Africa, then we would expect to see

the same relationship exist among the countries of sub-Saharan Africa.

   c. If the relationship between suffrage and income is strong, the relationship should exist between Europe and Africa.

   d. If the relationship between women's suffrage and GDP per capita is not just the result of the important differences between Europe and sub-Saharan Africa, then we would not expect to see the same relationship exist among the countries of sub-Saharan Africa.

5. What did the outlying cases of Kuwait, Qatar, and Zimbabwe illustrate when looking at the relationship between human capital accumulation and GDP per capita?

   a. They indicated the possible importance of natural resources, property rights, and politics.

   b. They should be removed from the data since they are not representative cases.

   c. They all point to the importance of oil as an explanation of GDP per capita.

   d. Corruption is clearly an important variable that needs to be considered.

6. What did exploring the stability of our results reveal?

   a. There are important thresholds we need to account for in our model estimation.

   b. The wrong estimation technique is being used.

   c. Regardless of the variables we include in this model, the results always indicate the same thing.

   d. Our estimates for a variable in the model can depend heavily on how we specify the model—what variables we include in the model.

7. What is an added variable plot?

   a. A plot that shows the residuals against the predicted values

   b. A plot that shows the predicted values against the residuals

   c. A plot showing the relationship between one variable, accounting for all of the others

   d. A plot showing the relationship between the dependent and independent variables

8. What observation led us to ask whether the relationship between ethnic heterogeneity and GDP per capita was a linear pattern?

   a. There were many poor island nations.

   b. The clump of countries we identified as being "industrialized" may be generating the linear pattern.

   c. The slope of the line declined when ignoring the European cases.

   d. There seems to be a nonlinear relationship between income and ethnolinguistic heterogeneity when all of the cases are considered.

9. If there is large variation in y at different levels of x in a scatter plot, what does that indicate?

   a. The relationship is not linear.

   b. There is no relationship between the x and y variables.

   c. The x variable does a good job of explaining the variation in y.

   d. There could be other variables that might help explain the variation.

10. Which of the following are true in the context of residual plots?

   a. Points above the horizontal line represent cases that are underpredicted.

   b. Points above the horizontal line represent cases that are overpredicted.

   c. Patterns in the data suggest there is a problem with the model.

   d. Patterns in the data suggest the model is appropriate.

ANNOTATED R FUNCTIONS

The following functions appear in this chapter. They are listed in order of their first appearance (with the code chunk number in parentheses) and annotated here to give a very brief description of their use. Some are not standalone functions and only work in combination with other commands. As a reminder, the code in every chapter will work properly if executed in the order it appears. Proper execution also depends on typing the author-defined *libraries()* command, which loads the required R packages.

ggplot(): defines the basic structure of a plot (usually the x and y variables). (2-1)

aes(): the aes (called "aesthetics") function is used in ggplot to define the basic structure of the plot, which often includes the variables you want to use and any shapes or colors. (2-1)

ylab(): labels the y axis in ggplot. (2-1)

xlab(): labels the x axis in ggplot. (2-1)

theme_minimal(): specifies a minimalist style for ggplot. (2-1)

theme(): specifies font, size, and so forth in a ggplot. (2-1)

annotate(): allows the placement of text in the figure. (2-2)

geom_vline(): places a horizontal line in the figure. (2-7)

scale_colour_manual: a function that allows the user to specify exactly what colors to use in the figure. (2-9)

stargazer(): a function from the 'stargazer' package that helps create professional-looking tables. (2-13)

lm(): a function that specifies a linear regression. The "lm" stands for linear model. (2-13)

ANSWERS

KNOWLEDGE CHECK

1. a
2. a, b, c
3. a, b, c
4. a, b, c
5. b
6. c
7. d
8. b, d
9. a, c
10. b
11. a
12. b, c, d
13. c, d
14. d
15. a, c, d
16. a, b, c, d
17. c, d
18. a, c
19. a, b, c, d
20. a, b, c, d

PRACTICE ON ANALYSIS AND VISUALIZATION

1. c
2. a
3. a, b, c, d
4. b
5. a
6. d
7. c
8. b
9. d
10. a, c

online resources

Access digital resources, including datasets, at **http://edge.sagepub.com/brownstats1e**.