CONTENTS

- The Logic of Large-n Studies: a Means Comparison 150
- Multivariate Linear Regression 154
- Tools for Categorical Data: Cross-Tabulation and Logistic Regression 165
- Reverse-Causality and Longitudinal Analysis 170
- Conducting Your Own Large-n Study 172
- Conclusion 175
- Key Terms 176

Large-n Observational Studies

Large-n observational studies have been a mainstay of political science research since the mid-1900s. Such studies became increasingly more attractive to political scientists as (1) new quantitative sources of data became available and (2) computers made it easier for social scientists to analyze those data. More to the point, large-n research offers an invaluable tool in drawing both descriptive and causal inferences. Without large-n observational data, we would not know what percentage of Americans favor a certain policy, political view, or candidate. We would not be able to compare measures of literacy, infant mortality, corruption, inequality, or economic development across countries. To understand why the political world works the way it does, we must first understand how it works, and, without large-n studies, we would not know nearly enough about the *how*. But the value of the large-n observational approach is not limited to making descriptive inferences. For many research questions, we might not be content with determining if one country has greater corruption than another or if corruption is increasing or decreasing with time. We might want to go further and ask *why*. Why is corruption higher in some countries than in others, and why is it increasing or decreasing?

In Chapter 5, we discussed how experiments are powerful tools for generating causal inferences because of their high internal validity. By using random assignment to hold constant other potential explanatory factors between control and condition groups, experiments allow us to make reliable inferences about the impact of an independent variable on a dependent variable. Instead of relying on control and experimental groups, large-n studies use statistical methods to mathematically control for alternative explanations.

In this chapter, we will explore several statistical methods that allow researchers to control for potential rival causes and reduce the risk of omitted variable bias. We will begin with a basic means comparison, which we will use to explain the logic of statistical control. From there we will graduate to more sophisticated techniques, including linear and logistic regression. While most of the chapter focuses on overcoming the challenges posed by systematic and random sampling, and measurement error and omitted variable bias, the chapter also delves into the problem of reverse causality and explores how longitudinal techniques, or comparing data over time, can be used to address this challenge. The chapter concludes by considering some ways that you can incorporate large-n approaches into your own research.

THE LOGIC OF LARGE-N STUDIES: A MEANS COMPARISON

Large-n observational studies can to some extent attempt to replicate the logic of control and experimental groups. Say, for example, that a researcher is interested in explaining why some countries have high levels of corruption while other countries have low levels. She hypothesizes that democracies will have lower levels of corruption than authoritarian regimes. She bases this hypothesis on the theory that elections allow voters the means to punish corrupt officials and that democracy is conducive to a free press, which can expose corruption. It would certainly be possible to divide countries into two groups—democracies and autocracies—and then see which group has a higher level of corruption.

While this method is similar to an experiment in that we can divide countries into groups, it is very different because the researcher does not control which countries are in the control group (autocracies) and which are in the condition group (democracies); there is no random assignment. This produces several important challenges to making causal inferences, one of which is omitted variable bias, which happens when the relationship between two variables is spurious. Given that the groups are not randomly assigned, there is a risk that those countries that are democracies also have other factors in common *vis-à-vis* those countries that are autocracies. For example, many of the democracies with low levels of corruption, such as Denmark and New Zealand, also have high levels of economic development; whereas many authoritarian countries, such as Zimbabwe, have low levels of economic development. It could be that economic development, not democracy, is driving differences in corruption. By failing to take economic development into account, we might be erroneously attributing explanatory power to "democracy" when the actual factor driving variation in the dependent variable is "economic development."[1] We might have omitted an important explanatory variable and biased our conclusion: hence the term **omitted variable bias**.

Large-n studies offer a solution to this problem. Because there are many observations, the researcher can further divide her sample into those with low, medium, and high levels of economic development. If she continues to observe that democracies have lower levels of corruption, regardless of whether they are at low, medium, or high levels of economic development, then she will be more confident in her findings. So while experiments use random assignment to ensure that their control and experimental groups are the same across all other potential explanatory factors, large-n studies use statistical techniques to simulate such control.

Let's see how this works with real data. As we know from Chapter 4, there are several different ways to conceptualize and measure both democracy and corruption. For illustrative purposes, let's use Cheibub, Gandhi, and Vreeland's measurement of regime type and group countries into democratic and authoritarian regimes based on the presence or absence of free and fair elections.[2] For corruption, we can use Transparency International's Corruption Perception Index (CPI), whose strengths and weaknesses were discussed in Chapter 4.

In Chapter 4, we also discussed how the nature of data determines what statistical techniques we use. We divided quantitative data into interval and categorical data, the latter of which we further divided into ordinal- (able to be put in an order) and nominal- (unable to be put in an order) level data. In this case the CPI score, which ranges from 1–10, can be treated as an interval-level variable and democracy as a categorical-level variable. As we know, for interval-level data, we are able to calculate a mean, or average score. Thus, it is very easy to divide countries between democracies and autocracies, and calculate and compare the average CPI score for each group, a process known simply as a **means comparison**.

The means comparison pictured in Figure 6.1 shows the average corruption score among countries that were democratic and countries that were authoritarian in 2008. As evident in the figure, democracies on average have a higher CPI (4.7) than autocracies (3.0), which means they have lower levels of corruption than their authoritarian counterparts. The data appear to support our researcher's hypothesis!

Of course we are not done there. We have to consider if we can draw valid inferences from this information. As we did in Chapter 4, we have to first ask ourselves if there are any concerns related to conceptualization, measurement, or sampling. Since we already covered these issues, we will not go into them again in detail here. Nonetheless, we would want to ask ourselves if our measures of democracy and corruption are consistent with our conceptualizations and if there are any systematic problems in the way that these variables are measured. Because we are examining nearly every country, sampling error is not a concern. Even so, we still would want to calculate confidence intervals around our two sample statistics as a means to take random measurement error into account. If you will recall from Chapter 4, there is considerable random measurement error in the CPI. The I-shaped lines at the top of the bars in Figure 6.1 represent the 95 percent confidence intervals for the two means. For example, we calculate an average CPI score among authoritarian countries of 3.0 (on a 1–10 scale) and a margin of error of approximately 0.3.[3] Assuming no systematic error in measurement (a big assumption), once we take random error into account, we are 95 percent confident that the true average CPI is between 2.7 and 3.3.

These error bars help provide a better test of our hypothesis that democratic countries are more likely to have low levels of corruption than authoritarian countries. We observe a difference of 1.7 between the two means (4.7–3.0), but are we confident in this difference once random error is taken into account? The foil of our hypothesis, what is referred to as the **null hypothesis**, posits that there is *no difference* in the CPI scores between democracies and authoritarian countries. If the error in the data is greater than the observed difference of 1.7, then we would *not* be able to conclude that democracies and autocracies have different levels of corruption. Nonetheless, we can see in the bar chart that the upper bound of the authoritarian error bar is about 3.3 and the lower bound of the democracy error bar is 4.3. Because there is no overlap, we are very confident that the difference observed is not due to random chance. We call this a **statistically significant relationship**, because the relationship observed is greater than the random error in the data.[4]

---

[1] Consider how this is similar to the process we described in Panel D of Figure 1.1.

[2] José Antonio Cheibub, Jennifer Gandhi, and James Raymond Vreeland, "Democracy and dictatorship revisited," *Public Choice* 143 (2010): 67–101.

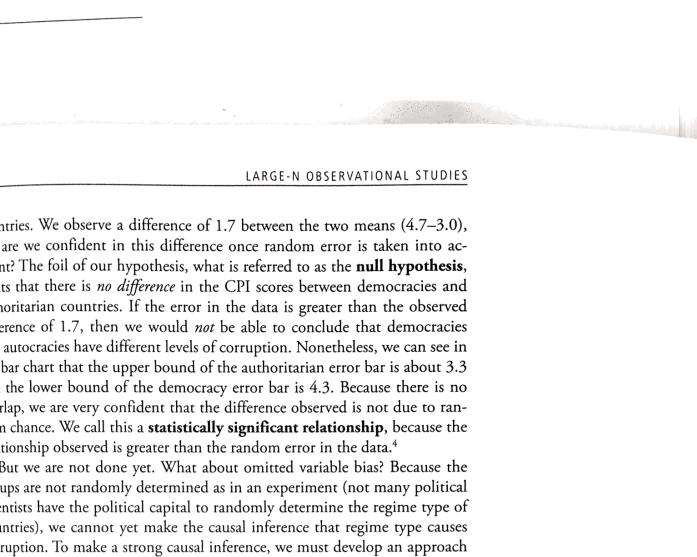


FIGURE 6.1 Average Corruption Perception Index Score among Democracies and Dictatorships (10 = Low Corruption)

Source: Author's analysis using Cheibub, Gandhi, and Vreeland's 2008 classification of regime type and Transparency International's CPI for 2008.

Note: n = 189.

[3] In this case the standard error would be calculated as the standard deviation (1.3) divided by the square root of the total number of authoritarian countries (69). This yields a standard error of .16. Doubling this produces a margin of error of 3.2.

But we are not done yet. What about omitted variable bias? Because the groups are not randomly determined as in an experiment (not many political scientists have the political capital to randomly determine the regime type of countries), we cannot yet make the causal inference that regime type causes corruption. To make a strong causal inference, we must develop an approach that allows us to account for potential omitted variables. As mentioned above, it might be economic development—not regime type—that is driving the difference in corruption scores.

As a result, Figure 6.2 controls for the level of economic development using an ordinal-level variable of low, medium, and high Gross Domestic Product (GDP) per capita. This is simply referred to as a **controlled means comparison** because it adds a control variable: economic development. The controlled means comparison clearly shows that the level of economic development is an important predictor of corruption. Low-income countries have much lower CPI scores (higher corruption) than their high-income counterparts. But what does the figure say about our democracy hypothesis? Among the low-income and middle-income countries, the average corruption scores for democracies are only slightly higher than for dictatorships. Further testing (not presented here) reveals that the relationship between regime type and corruption is still statistically significant, which means that we are confident that there is a relationship even once random error is taken into account; however, the impact is small. Only among the high-income countries does democracy appear to have a substantial effect on corruption levels. The impact of regime type appears to matter, but the magnitude of the effect depends on the level of economic development, and it matters far less than suggested by Figure 6.1. Clearly, by omitting economic development from our study in Figure 6.1, we biased our results.

[4] The fact that there is no overlap between the bars means that we are much more than 95 percent confident. Because of the way a standard error of the difference between two means is calculated, material that is not discussed here, it is possible for there to be some overlap in the error bars and still be 95 percent confident in the difference.

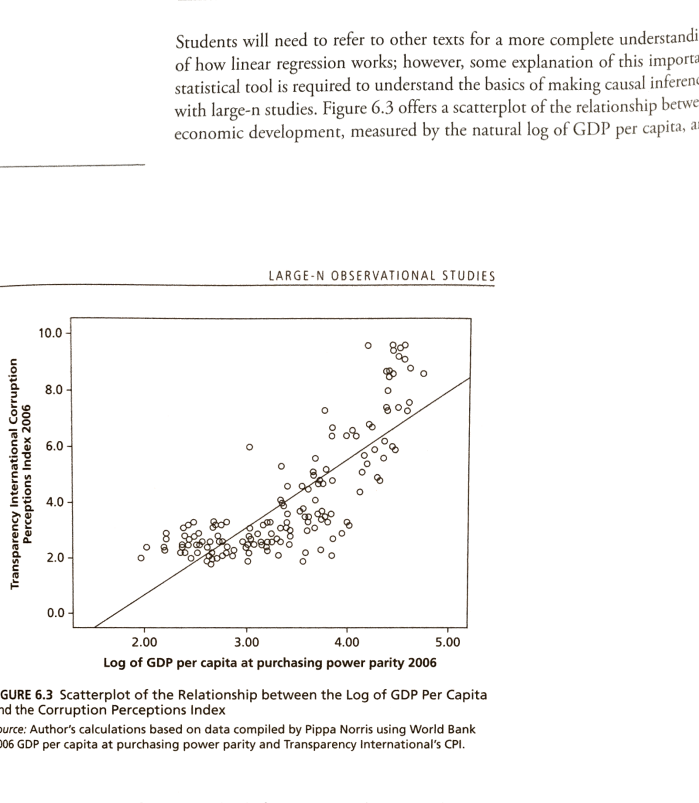


FIGURE 6.2 Average Corruption Perception Index Score among Democracies and Dictatorships Controlling for Low, Middle, and High Levels of Economic Development (10 = Low Corruption)

Source: Author's analysis using Cheibub, Gandhi, and Vreeland's 2008 classification of regime type, Transparency International's CPI Index, and the World Bank's GDP per capita.

In summary, the controlled comparison presented in Figure 6.2 attempts to replicate the logic of control used in an experiment. We cannot ignore other factors like income levels, but, because we have many observations, we are able to divide the sample into groups and control for income mathematically. Controlled means comparisons are, nonetheless, limited in that we can really only control for one factor at a time. There might be other factors that could explain variation in corruption and whose exclusion might bias our results. For example, some scholars have argued that religious tradition, colonial tradition, natural resource endowments, and the size of government all help explain corruption. The bad news is that our means comparison cannot control for all of these factors at the same time; however, the good news is that a more sophisticated statistical tools, known as regression analysis, allows us to do just that.

MULTIVARIATE LINEAR REGRESSION

Students will need to refer to other texts for a more complete understanding of how linear regression works; however, some explanation of this important statistical tool is required to understand the basics of making causal inferences with large-n studies. Figure 6.3 offers a scatterplot of the relationship between economic development, measured by the natural log of GDP per capita, and corruption perceptions.[5] Again, the level of measurement determines the type of statistical tool that we use. In this case, a scatterplot is preferable to a means comparison because both of the variables are interval level. Because each country has a unique GDP per capita, the average CPI for each value of GDP would not tell us very much.

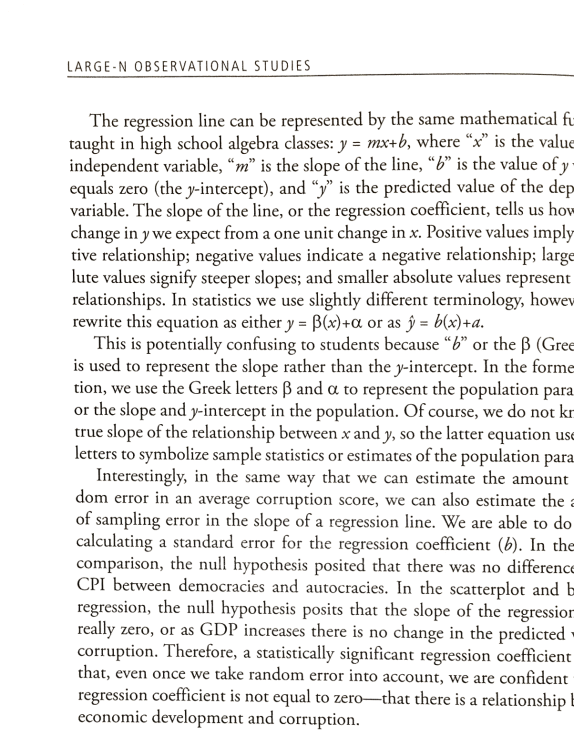


FIGURE 6.3 Scatterplot of the Relationship between the Log of GDP Per Capita and the Corruption Perceptions Index

Source: Author's calculations based on data compiled by Pippa Norris using World Bank 2006 GDP per capita at purchasing power parity and Transparency International's CPI.

Each of the circles in the scatterplot represents an individual country at its particular level of development and level of perceived corruption. The relationship is not perfectly linear, but one can see that high-income countries almost always have low levels of corruption (a high CPI score) and most of the low-income countries have high levels of corruption (a low CPI score). Clearly, there is a correlation, or relationship, between these two variables. The scatterplot also includes a line running through the data known as a **bivariate regression line**, or a line that estimates the relationship between two variables. The slope of this line is summarized by a **regression coefficient**, which provides researchers with an estimate of the precise nature of that relationship.

[5] In this case, the log of GDP per capita is used to produce a linear relationship between the two variables. Because of the uneven distribution of GDP per capita, this variable would otherwise have a curvilinear relationship with the CPI.

The regression line can be represented by the same mathematical function taught in high school algebra classes: $y = mx+b$, where "x" is the value of the independent variable, "m" is the slope of the line, "b" is the value of y when x equals zero (the y-intercept), and "y" is the predicted value of the dependent variable. The slope of the line, or the regression coefficient, tells us how much change in y we expect from a one unit change in x. Positive values imply a positive relationship; negative values indicate a negative relationship; larger absolute values signify steeper slopes; and smaller absolute values represent weaker relationships. In statistics we use slightly different terminology, however, and rewrite this equation as either $y = \beta(x)+\alpha$ or as $\hat{y} = b(x)+a$.

This is potentially confusing to students because "b" or the β (Greek beta) is used to represent the slope rather than the y-intercept. In the former equation, we use the Greek letters β and α to represent the population parameters, or the slope and y-intercept in the population. Of course, we do not know the true slope of the relationship between x and y, so the latter equation uses Latin letters to symbolize sample statistics or estimates of the population parameters.

Interestingly, in the same way that we can estimate the amount of random error in an average corruption score, we can also estimate the amount of sampling error in the slope of a regression line. We are able to do this by calculating a standard error for the regression coefficient (b). In the means comparison, the null hypothesis posited that there was no difference in the CPI between democracies and autocracies. In the scatterplot and bivariate regression, the null hypothesis posits that the slope of the regression line is really zero, or as GDP increases there is no change in the predicted value of corruption. Therefore, a statistically significant regression coefficient implies that, even once we take random error into account, we are confident that the regression coefficient is not equal to zero—that there is a relationship between economic development and corruption.

Scatterplots with a regression line make an excellent graphical tool to present data. Even without any statistical background, a reader can look at the scatterplot in Figure 6.3 and quickly understand the bivariate relationship between economic development and corruption. Slight modifications such as shading plots from a region of interest, labeling countries, or highlighting interesting outliers provide additional tools to portray important information to readers (see Figure 6.4). Nonetheless, while the scatterplot and bivariate regression line appear to be more sophisticated tools than a means comparison, they suffer from the same basic limitation. Because the scatterplot only looks at the relationship between two variables, it could suffer from omitted variable bias.

Multivariate linear regression, or regression with more than one independent variable, continues the logic of the regression line, but instead of describing the relationship between just the independent and dependent variable, what is known as a bivariate relationship, it estimates the independent effect of each variable on the dependent variable while controlling for other factors. Peter Kennedy uses Venn diagrams to describe the logic behind multivariate regression analysis.[6] The Venn diagram in Figure 6.5 summarizes the hypothetical effect of democracy and economic development on corruption. As can be seen in the figure, the democracy and economic development circles together cover more than half of the corruption circle. This implies that by knowing the level of economic development of countries and whether or not they are democracies, we can explain more than half of the variation in corruption. Basic multivariate linear regression analysis (there are various other regression techniques) produces a measure of the association called an R^2, which tells us the percentage of variation in the dependent variable that is explained by all the independent variables in the analysis. In this case it would be over .5 or 50 percent.

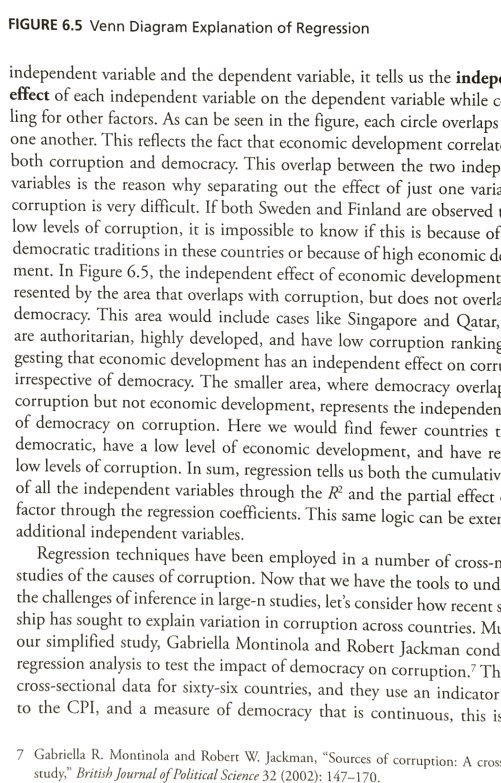


FIGURE 6.4 Scatterplot of the Relationship between the Log of GDP Per Capita and the Corruption Perceptions Index with Regional Shading

Source: Author's calculations based on data compiled by Pippa Norris using World Bank 2006 GDP per capita at purchasing power parity and Transparency International's CPI.

Multivariate regression, like a bivariate regression, also produces a regression coefficient; however, rather than describe the bivariate relationship between an independent variable and the dependent variable, it tells us the **independent effect** of each independent variable on the dependent variable while controlling for other factors. As can be seen in the figure, each circle overlaps within one another. This reflects the fact that economic development correlates with both corruption and democracy. This overlap between the two independent variables is the reason why separating out the effect of just one variable on corruption is very difficult. If both Sweden and Finland are observed to have low levels of corruption, it is impossible to know if this is because of strong democratic traditions in these countries or because of high economic development. In Figure 6.5, the independent effect of economic development is represented by the area that overlaps with corruption, but does not overlap with democracy. This area would include cases like Singapore and Qatar, which are authoritarian, highly developed, and have low corruption rankings, suggesting that economic development has an independent effect on corruption, irrespective of democracy. The smaller area, where democracy overlaps with corruption but not economic development, represents the independent effect of democracy on corruption. Here we would find fewer countries that are democratic, have a low level of economic development, and have relatively low levels of corruption. In sum, regression tells us both the cumulative effect of all the independent variables through the R^2 and the partial effect of each factor through the regression coefficients. This same logic can be extended to additional independent variables.

[6] Peter Kennedy, *Guide to econometrics* (New York: Wiley-Blackwell, 2008).

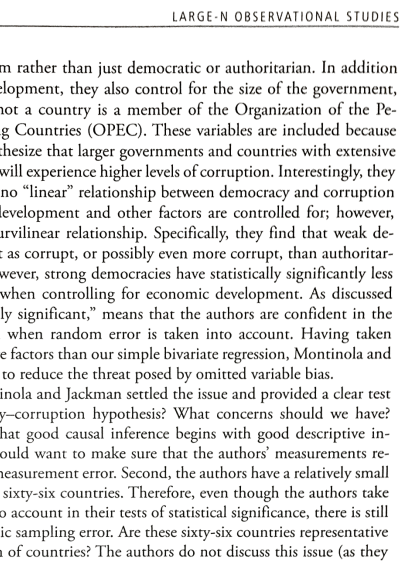


FIGURE 6.5 Venn Diagram Explanation of Regression

Regression techniques have been employed in a number of cross-national studies of the causes of corruption. Now that we have the tools to understand the challenges of inference in large-n studies, let's consider how recent scholarship has sought to explain variation in corruption across countries. Much like our simplified study, Gabriella Montinola and Robert Jackman conducted a regression analysis to test the impact of democracy on corruption.[7] They have cross-sectional data for sixty-six countries, and they use an indicator similar to the CPI, and a measure of democracy that is continuous, this is to say along a continuum rather than just democratic or authoritarian. In addition to economic development, they also control for the size of the government, and whether or not a country is a member of the Organization of the Petroleum Exporting Countries (OPEC). These variables are included because the authors hypothesize that larger governments and countries with extensive natural resources will experience higher levels of corruption. Interestingly, they find that there is no "linear" relationship between democracy and corruption once economic development and other factors are controlled for; however, they do find a curvilinear relationship. Specifically, they find that weak democracies are just as corrupt, or possibly even more corrupt, than authoritarian countries; however, strong democracies have statistically significantly less corruption even when controlling for economic development. As discussed above, "statistically significant," means that the authors are confident in the relationship even when random error is taken into account. Having taken into account more factors than our simple bivariate regression, Montinola and Jackman are able to reduce the threat posed by omitted variable bias.

[7] Gabriella R. Montinola and Robert W. Jackman, "Sources of corruption: A cross-country study," *British Journal of Political Science* 32 (2002): 147–170.

So have Montinola and Jackman settled the issue and provided a clear test of the democracy–corruption hypothesis? What concerns should we have? First, we know that good causal inference begins with good descriptive inference. So we would want to make sure that the authors' measurements reduce the risk of measurement error. Second, the authors have a relatively small sample size: only sixty-six countries. Therefore, even though the authors take random error into account in their tests of statistical significance, there is still a risk of systematic sampling error. Are these sixty-six countries representative of the population of countries? The authors do not discuss this issue (as they should), but the answer is probably not. As in many cross-national studies, the authors were limited to the countries they had data for. Missing data are a serious problem in cross-national research, producing an effect akin to the non-response bias discussed in Chapter 4. Third, although the authors include three control variables, they do not control for a host of other factors that researchers have argued are important to explaining corruption. As a result, there is still a risk of omitted variable bias.

In his 2000 study, Daniel Treisman tested a larger set of variables against corruption.[8] While he still confronts problems of measurement error and sampling error, he runs less of a risk of omitted variable bias. He too is concerned with the relationship between democracy and corruption, and he finds evidence that "uninterrupted democracy" corresponds with reduced corruption levels. He also finds that economic development, Protestant religious culture, and British colonial heritage have statistically significant, independent, and negative relationships with corruption.

[8] Daniel Treisman, "The causes of corruption: A cross-national study," *Journal of Public Economics* 76 (2000): 399–457.

You and Khagram are concerned with how omitted variable bias inhibited studies about the relationship between inequality and corruption.[9] Prior to their research, several studies had found that inequality did not have an independent impact on corruption.[10] This surprised You and Khagram, who theorized that in unequal societies the rich have a great deal to lose from fair political, administrative, and judicial processes, particularly given that the demand for redistribution is likely to be high in an unequal economy. Under a corrupted system, however, the wealthy are able to buy influence. The authors argue that previous studies found that inequality had no impact because prior research had failed to control for a legacy of a socialist legal system. You and Kagram contend that socialist systems bred both corruption and greater equality, but, in the rest of the world, inequality exacerbated corruption. Failing to control for a legacy of a socialist legal system masked this effect of inequality, and, in fact, once this factor is controlled for, the authors find that inequality has an independent, statistically significant effect on corruption. Their analysis highlights that *omitted variable bias can work in either of two directions*. Absent adequate controls, an observed relationship might be spurious or an unobserved relationship might be overlooked.

The authors also point out another potential flaw in many regression analyses. Linear regression assumes additive relationships between the independent variables and the dependent variable. Consider the independent variables of democracy and inequality, and the dependent variable of corruption. An additive relationship would suggest that (1) democracy has an independent effect on corruption and (2) inequality has a separate, independent effect. An interactive relationship, however, would suggest that the relationship between inequality and corruption *depends* on the level of democracy in the country, and vice versa. This interactive relationship is similar to the means comparison we showed in Figure 6.2, where regime type only had an effect on corruption in nations with high levels of economic development.

Regression techniques can be used to explore interactive relationships; however, it is necessary for analysts to create and test interaction variables. You and Khagram explain the surprisingly inconsistent findings regarding both democracy and inequality by testing for just such an interaction. They argue that, in an authoritarian regime, elites can maintain power through repression. In a democracy, elites have fewer repressive tools at their disposal and corruption offers an alternative means to influence policy decisions in their favor. While the method that they use to test this interactive hypothesis is somewhat different, one could imagine dividing the sample of countries between democracies and authoritarian countries. Then, as presented hypothetically in Figure 6.6, one could generate a scatterplot of the relationship between inequality and corruption among democracies and a separate scatterplot of this relationship among autocracies. You and Khagram find evidence consistent with their interactive theory: that inequality is an important determinant of corruption among democracies but not among autocracies.

[9] Jong-Sung You and Sanjeev Khagram, "A comparative study of inequality and corruption," *American Sociological Review* 70 (2005): 136–157.

[10] For example, see Bryan W. Husted, "Wealth, culture, and corruption," *Journal of International Business Studies* 30 (1999): 339–360 and Martin Paldam, "Corruption and religion: Adding to the economic model," *Kyklos* 54 (2001): 383–413.

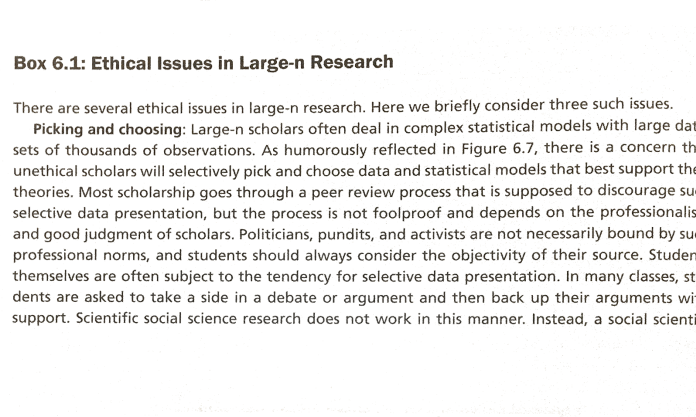


FIGURE 6.6 Hypothesized Interactive Relationship

One final large-n corruption analysis worth profiling is Danila Serra's study, which attempts to address a problem heretofore not discussed.[11] After two decades of large-n statistical analyses of the causes of corruption, Serra reviews seven such studies and finds that their results are somewhat contradictory, even though the different authors use many of the same variables. She's particularly concerned about a very worrisome problem for large-n quantitative research: robustness. **Robust** findings mean that statistical analyses consistently reveal the same relationship between the independent and dependent variables regardless of what control variables are included in the analysis, or, in the terminology of many social scientists, how the model is specified. There is a concern that unethical scholars will selectively present regression models to best support their theories. Serra quotes Ronald Coase as saying "If you torture data long enough, Nature will confess."[12] Most scholarship goes through a peer review process that is supposed to discourage such selective data presentation, but the process is not foolproof, and, of course, sometimes researchers just make mistakes. Regardless, of the cause, when different scholars run slightly different models and come up with divergent conclusions, this suggests that the findings are not robust.

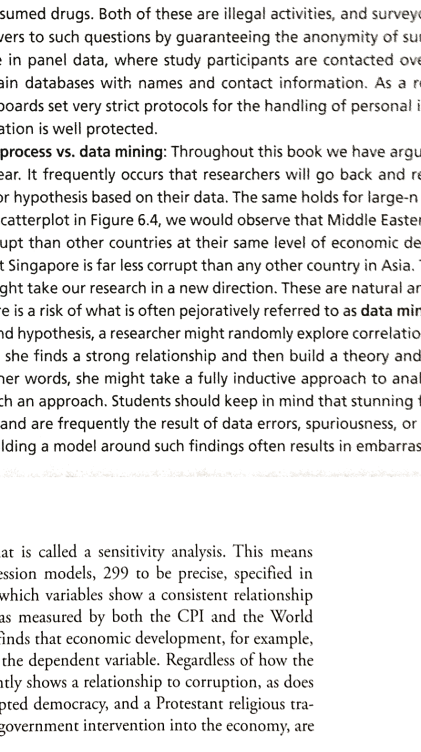
[11] Danila Serra, "Empirical determinants of corruption: A sensitivity analysis," *Public Choice* 126 (2006): 225–256.

[12] Serra, "Empirical determinants of corruption," p. 229.

Box 6.1: Ethical Issues in Large-n Research

There are several ethical issues in large-n research. Here we briefly consider three such issues.

Picking and choosing: Large-n scholars often deal in complex statistical models with large datasets of thousands of observations. As humorously reflected in Figure 6.7, there is a concern that unethical scholars will selectively pick and choose data and statistical models that best support their theories. Most scholarship goes through a peer review process that is supposed to discourage such selective data presentation, but the process is not foolproof and depends on the professionalism and good judgment of scholars. Politicians, pundits, and activists are not necessarily bound by such professional norms, and students should always consider the objectivity of their source. Students themselves are often subject to the tendency for selective data presentation. In many classes, students are asked to take a side in a debate or argument and then back up their arguments with support. Scientific social science research does not work in this manner. Instead, a social scientist should view his role as objectively testing hypotheses—not trying to support an argument. Complex statistical models may make political science appear more "scientific"; however, model specification and interpretation still depends on the good judgment of the researcher. To avoid disappointing null results, it is usually a good idea to select and design research projects whose results will be interesting regardless of whether or not the tested hypothesis is supported.



"That's the gist of what I want to say. Now get me some statistics to base it on."

FIGURE 6.7 Ethical Issues in Data Presentation

Source: Joseph Mirachi/The New Yorker Collection/www.cartoonbank.com

Anonymity: Another ethical issue in survey research is ensuring the anonymity of study participants. In the course of this book we have discussed surveys that ask respondents if they have paid bribes and if they have consumed drugs. Both of these are illegal activities, and surveyors can only reliably obtain honest answers to such questions by guaranteeing the anonymity of survey results. This is an even larger issue in panel data, where study participants are contacted over time and researchers have to maintain databases with names and contact information. As a result, many university research review boards set very strict protocols for the handling of personal information to ensure that such information is well protected.

The non-linear research process vs. data mining: Throughout this book we have argued that the research process is non-linear. It frequently occurs that researchers will go back and rethink their research question, theory, or hypothesis based on their data. The same holds for large-n studies. For example, in observing the scatterplot in Figure 6.4, we would observe that Middle Eastern countries are comparatively less corrupt than other countries at their same level of economic development. We would also observe that Singapore is far less corrupt than any other country in Asia. These interesting empirical puzzles might take our research in a new direction. These are natural and desirable adjustments; however, there is a risk of what is often pejoratively referred to as **data mining**. Rather than begin with a theory and hypothesis, a researcher might randomly explore correlations between variables in a dataset until she finds a strong relationship and then build a theory and hypothesis around the findings. In other words, she might take a fully inductive approach to analyzing data. We recommend against such an approach. Students should keep in mind that stunning findings are often too good to be true and are frequently the result of data errors, spuriousness, or some other challenge to inference. Building a model around such findings often results in embarrassing errors.

As a result, Serra runs what is called a sensitivity analysis. This means that she runs numerous regression models, 299 to be precise, specified in slightly different ways to see which variables show a consistent relationship with corruption perceptions as measured by both the CPI and the World Bank's corruption index. She finds that economic development, for example, has a *robust* relationship with the dependent variable. Regardless of how the model is specified, it consistently shows a relationship to corruption, as does political instability, uninterrupted democracy, and a Protestant religious tradition. Other factors, such as government intervention into the economy, are not robust.

Box 6.2: Interpreting Linear Regression Results

Regression tables are often very intimidating; however, once you know what you are looking at, they are actually fairly simply to interpret. Table 6.1 below presents one of many regression outputs from Treisman's article on corruption. In the left-hand column we are presented with all of the independent variables in the study. The dependent variable, the CPI, is mentioned in the title. The top number to the right of the independent variable name is the regression coefficient. If the number is negative, the model predicts a negative relationship with the dependent variable. This number can be understood as the amount of change in "y" that is predicted to occur with a one unit change in "x." In the case of the variable "uninterrupted democracy," the model predicts that, controlling for other factors, a country with uninterrupted democracy will have .81 less corruption on a 1–10 scale than a country that has had interrupted democracy.

TABLE 6.1 One of Treisman's (2000) Many Regression Models Explaining Transparency International's Corruption Perception Index in 1997

Uninterrupted democracy (1950–1995)	−.81* (.47)
Common law system	.98* (.51)
Former British colony or the UK	−2.31*** (.62)
Never a colony	−.27 (.47)
Percentage Protestant 1980	−.003*** (.01)
Ethnolinguistic division	.01 (.01)
Fuel, metal, and mineral exports	.01 (.01)
Log GDP per capita	−3.63*** (.66)
Federal	.66** (.29)
Imports/GNP (%)	−.01** (.01)
R^2	.8987
N	42

note: *p<0.10; **p<0.05; ***p<0.01.

Source: Daniel Treisman, "The causes of corruption: A cross-national study," *Journal of Public Economics* 76 (2000): 399–457.

We have to be somewhat careful in interpreting this number, however, as this is what the model predicts based on the sample. This number does not take into account the random error in data. As a result, below this number in parentheses is the standard error of the regression coefficient; this is our measure of random error in our data. What we want to know is that, once we take this error into account, do we still observe a relationship between the independent variables and the dependent variable? Or is it possible that the regression coefficient is really 0: that there is no relationship? The stars to the right of the coefficient provide us with the answer to this question. If there is no star, it means that we are less than 90 percent confident that there is a relationship; one star means that we are 90 percent confident in the relationship; two stars mean that we are 95 percent confident in the relationship, and three stars mean that we are 99 percent confident in the relationship. At the bottom of the table we are presented with the R^2, which tells us that together all of the variables explain 89 percent of the variation in the CPI. We are also presented with the sample size of 42 countries, which should raise some alarms about potential sampling error.

TOOLS FOR CATEGORICAL DATA: CROSS-TABULATION AND LOGISTIC REGRESSION

Let's consider the role of large-n studies in addressing some of this book's other research questions. In Chapter 5, we discussed the use of experiments to test the impact of advertisements and outreach on voting and vote choice. While experiments are increasingly more common, the primary means to test the effects of advertisements and outreach on voter behavior has traditionally been through large-n observational studies. For example, since 1996, the Wisconsin Advertising Project has been collecting an impressive supply of data on the campaign advertisements aired in U.S.'s largest television markets. Each advertisement is coded for its tone (positive, negative, or otherwise), the issues mentioned in the ad, and the types of appeals and images seen in the ad. Additionally, the dataset includes a measure of each date and time that the advertisement aired and on which stations and programs it aired. These data have revolutionized the study of advertising effects by providing scholars with an independent measure of which ads individuals were likely exposed to depending on where they live.

In addition, surveys such as the American National Election Study (ANES) ask survey respondents if they have been contacted by someone urging them to vote or to register to vote, and it also asks respondents if they voted in the previous election. Unlike the previous dependent variables that we have examined in this chapter, the decision to vote is not an interval-level variable but a two-category nominal-level variable. Therefore, a means comparison and linear regression would be inappropriate statistical methods; one cannot really find an average score for a dependent variable or create a scatterplot when there are only two possible values. Instead, we require different tools to analyze categorical-dependent variables. One simple tool for categorical data is a cross-tabulation. A **cross-tabulation** merely divides the sample into those who were urged to vote and those who were not. Then, it compares the percentage of those who were contacted and voted with the percentage of those who were not contacted but voted anyway. If the former percentage is higher, it would suggest that outreach matters.

Such a cross-tabulation is presented in Table 6.2 using data from the ANES 2010. The cross-tabulation shows that, of those who were not contacted, 72.4 percent reported voting, and, of those who were contacted, 80.1 percent reported voting: an observed difference of 7.6 percent. Again, the data appear to support our hypothesis, although the relationship is not particularly strong.

As in the previous examples, however, we have to consider our challenges to inference. Again we would want to consider conceptualization, measurement, and sampling concerns. There is a negative stigma attached to not voting, and, as mentioned earlier, survey respondents tend to state that they voted when in fact they did not, a form of systematic error that we called a social desirability bias. Furthermore, given that the question about outreach preceded the question about voting in the survey instrument, individuals who experienced outreach efforts might be more inclined to be dishonest about their actual voting behavior. Furthermore, unlike the above comparison of countries, data in Table 6.2 are based on a sample, and as such it is subject to sampling error.

Again, we can take random error into account with a test of statistical significance. The null hypothesis would predict that the percentage who voted is the same regardless of whether a respondent was urged to vote or not. It is possible that the small difference observed, only 7.6 percent, is smaller than the amount of error in the data. In other words, it is possible that the difference we observe in the sample does not exist in the population. In this case, we would use a different statistical test than we used with the mean comparison and the regression coefficient: a Chi-square (x^2) test. Again, we will not go into the specifics of how the x^2 is calculated, but essentially the x^2 asks what the cross-tabulation would look like if the null hypothesis were correct—that is, if there was no relationship between the variables—and then compares that hypothetical cross-tabulation with our actual cross-tabulation. While we do not present the calculations here, the result would tell us that we can be at least 95 percent confident that the observed difference is not due to random chance. Nonetheless, given that the true difference between these groups might be less than 7.6 percent, we have to conclude that there is possibly only a very weak relationship between outreach and voting.

TABLE 6.2 Cross-Tabulation between Contacted to Vote or Register and Vote Behavior

	Contacted	Not contacted	Total
Did not vote	210 19.9%	289 27.6%	499 23.8%
Voted	843 **80.1%**	759 **72.4%**	1,602 76.2%
Total	1,053 100%	1,048 100%	2,101 100%

Source: American National Election Survey 2010.

An additional concern is of course omitted variable bias. Because survey respondents were not randomly assigned to receive an outreach message (as in a field experiment) and because we are not controlling for other factors, the observed relationship could be spurious. In fact, there are many factors that we might want to control for, including interest in politics, the type of outreach method, the message of outreach efforts, education levels, the nature of the political race, etc. . . . As with a controlled means comparison, we are able to add one control variable and conduct a **controlled cross-tabulation**. For example, we could divide our sample into high- and low-education respondents, and then examine the relationship between outreach efforts and voter turnout among each of these two groups. Nonetheless, as with the means comparison, we are limited to one control variable, and there are many factors we might want to control for. Fortunately, just as multivariate linear regression allowed us to control for the effect of numerous independent variables on an *interval*-level dependent variable, we have another regression tool that will allow us to control for the effect of additional factors on our *categorical* dependent variable: **logistic regression**.

Logistic regression offers a similar tool to linear regression that can be used for examining **dichotomous** dependent variables, or dependent variables with just two values. Of course, a dependent variable with only two values does not permit us to develop a scatterplot and run a regression line through the data points. What logistic regression does, therefore, is transform this dichotomous dependent variable into an interval-level variable. It does this by transforming the choice to vote or not vote into a probability of voting, which is in fact interval. It actually goes a couple of steps further and transforms the probability of voting into the odds of voting and then the logged odds of voting. The term "logistic" regression comes from this process of studying the logged odds.

Again, the math is not particularly important for our purposes here. While we certainly hope that you will learn these techniques in detail, as you might imagine each technique requires considerable time and effort to understand and master. Unfortunately, in this process, it is often easy to lose sense of the forest in examining the trees. Our goal here is to convey the big picture. Your take away should be the following:

> Different types of data require different statistical tools. If we know a little bit about our data, such as whether they are interval, ordinal, or nominal, then we can select the appropriate tool.
>
> Regardless of the tool at hand, all of these tests attempt to isolate the independent effect of an independent variable on a dependent variable while controlling for other factors.
>
> They then go a step further and tell us (1) if we can be confident that the relationship is not zero when we take random error into account and (2) the strength of the observed relationship.

In this case, we use logistic regression because we are attempting to explain a dichotomous dependent variable. There are also several other regression techniques that have been developed for studying diverse forms of categorical data.[13]

In contrast with the results of large-n observational studies, in a series of laboratory experiments discussed in Chapter 5, Ansolabehere and Iyengar found that individuals who viewed negative ads were less likely to report that they will vote.[a] The experiments were innovative because the negative campaign ads were imbedded in news broadcasts, much like they would be in real life, and respondents thought that they were participating in an experiment about the news. Nonetheless, there were still limits to external validity. In particular, the authors were only able to measure a study participant's intention to vote—not his or her actual behavior.

Wattenberg and Brians's study, discussed below, overcame this external validity problem by asking participants if they had seen campaign ads and if they had voted in the past election.[b] Nonetheless, while Wattenberg and Brians had overcome the external validity challenge to inference, their research was subject to other internal validity challenges. Ansolabehere, Iyengar, and Simon, in a response to Wattenberg and Brians, argued that relying on individuals to remember advertisements they had seen was not only unreliable, but also potentially biased.[c] In other words, Wattenberg and Brians's study suffered from measurement error. They noted that, when they conducted their initial experiments on negative advertising, over 50 percent of their experimental subjects could not recall having seen a campaign advertisement they had been shown just a half hour earlier. Using a sophisticated regression procedure called two-stage least squares to attempt to address this bias, the authors re-analyze the same data as Wattenberg and Brians and conclude that negative advertising does in fact have a negative impact (albeit small) on the probability of voting. They argue that their experiments and their analysis of survey data point in the same direction: negative advertising suppresses the vote.

Nonetheless, the debate is not over. In addition to other scholars who jumped into the fray, Krasno and Green sought to answer the research question using a natural experiment.[d] The U.S. electoral college system creates incentives for presidential campaigns to focus their advertising resources on swing states. However, the television markets are not entirely fixed at the state lines. As a result, New Jersey voters living in eight counties near the Pennsylvania border were inundated with ads while those near the New York border barely saw any ads. By comparing aggregate voter turnout across media markets (their unit of analysis) in states like New Jersey in the 2000 presidential election, the authors sought to measure the impact of negative campaign ads on voter turnout. They argue that focusing on aggregate voter turnout is preferable because of the aforementioned tendency of survey respondents to say they had voted when in fact they had not, a serious measurement problem. The result of their study: negative ads were found to have no effect.

Krasno and Green probably do not have the final word, however. Although they agree with the conclusions, Franz, Freedman, Goldstein, and Ridout raised a number of concerns with the Krasno and Green study, including issues of measurement error and concerns over the statistical techniques that they employed.[e] The issue is hardly settled and scholars will continue to develop research designs to gain a better understanding of the effects of negative advertising on citizen participation in elections.

a. Stephen Ansolabehere and Shanto Iyengar, *Going negative* (New York: Free Press, 1995).

b. Wattenberg and Brians, "Negative campaign advertising: demobilizer or mobilizer?"

c. Stephen D. Ansolabehere, Shanto Iyengar, and Adam Simon, "Replicating experiments using aggregate and survey data: The case of negative advertising and turnout," *American Political Science Review* 93 (1999): 901–909.

d. Jonathan S. Krasno and Donald P. Green, "Do televised presidential ads increase voter turnout? Evidence from a natural experiment," *Journal of Politics* 70 (2008): 245–261.

e. Franz et al., "Understanding the effect of political advertising on voter turnout."

Martin Wattenberg and Craig Brians used logistic regression to study the effects of advertisements on voter turnout using data from the 1992 and 1996 ANES.[14] Wattenberg and Brians examined whether individuals who recalled seeing a negative ad were more or less likely to vote compared to those who recalled seeing no advertisements or a positive advertisement. Because recalling an advertisement and reporting voting could be correlated with many other factors that might also explain voter turnout, the authors controlled for as many possible explanations as possible, including how much attention the respondents paid to news about the campaign, the respondents' levels of education and income, and a variety of other political and demographic indicators. After controlling for these factors, Wattenberg and Brians found that when respondents recalled a negative advertisement they were *more* likely to vote in the 1992 election but no more likely to vote in the 1996 election. Thus, the results were inconclusive. In a subsequent study, Kenneth Goldstein and his colleagues used data from the Wisconsin Advertising Project and found that negative advertising has little effect on making individuals either more or less likely to vote.[15]

[13] Ordered logit or ordered probit is used to study ordinal dependent variables, and multinomial logit or multinomial probit is used with nominal-level dependent variables.

[14] Martin P. Wattenberg and Craig Leonard Brians, "Negative campaign advertising: Demobilizer or mobilizer?" *American Political Science Review* 93 (1999): 891–899.

REVERSE-CAUSALITY AND LONGITUDINAL ANALYSIS

Thus far we have discussed systematic and random sampling and measurement error as well as omitted variable bias at length. We have also addressed the issue of robustness. There is, however, another *major* challenge to inference in large-n studies: **reverse-causality** (see Panel B in Figure 1.1). Let's return to our example from earlier in the chapter to illustrate this concept. Above we hypothesized that democratic regimes would have lower levels of corruption than authoritarian regimes. In other words, regime type *affects* corruption. But couldn't the reverse also be true? Couldn't corruption help determine regime type? In fact, several military coups in recent history have been partially justified as a response to corruption in the political system. Take for example military takeovers in Pakistan in 1999, Thailand in 2006, and Bangladesh in 2007.

Consider another research example from above: whether being urged to vote has an impact on voting. In this case, we asked if a "get out the vote" campaign influences vote behavior, but, again, couldn't the reverse also be true? Wouldn't it make sense for a vote campaign to focus on communities with low voter turnout? In other words, voter turnout might determine who is reached by a campaign, rather than vice versa.

Just as with omitted variable bias, the problem of reversal-causation stems from the inability of large-n observational researchers to randomly assign study participants to control and condition groups as in experimental research. While large-n researchers can divide between those who have been contacted and urged to vote and those who have not, they do not determine who falls into which group. With reverse-causality, the intended "dependent variable" in a given research project might actually be a factor in determining who ends up in which group.

How then can we solve this dilemma? There are different potential answers to this question, but a simple one is to look at data over time. To use the example above, if we saw that countries tend to become authoritarian first and then become more corrupt later, then this would support our hypothesis that authoritarian regimes foster corruption. In other words, for an independent variable to cause a dependent variable, a change in the independent variable must have preceded a change in the dependent variable. Temporal ordering is easy to establish in experiments, where the researcher assigns a value of the independent variable first and then measures for the dependent variable, but such temporal ordering is generally not possible with cross-sectional data, since all variables are measured at the same point in time.

[15] Michael M. Franz, Paul Freedman, Ken Goldstein, *et al.*, "Understanding the effect of political advertising on voter turnout: A response to Krasno and Green," *Journal of Politics* 70 (2008): 262–268.

Studies on the influence of gender on the behavior of elected officials are illustrative of both this problem and the longitudinal solution to the problem. Several studies have argued that female legislators bring different experiences and viewpoints to the policymaking process than their male counterparts. A simple way to test this hypothesis using cross-sectional data would be to compare the behavior of male and female legislators. In fact, such a comparison would show that women legislators introduce more bills related to women's issues and that they do more to support such bills throughout the legislative process. Nonetheless, such cross-sectional studies suffer from our reverse-causality problem. It might be that women legislators are brought into office and do more to support women's issues because they represent districts that prioritize women's issues. Cross-sectional studies can attempt to control for district-level differences by including variables for demographic factors, attitudinal characteristics measured in survey data, and party identification, but they cannot control entirely for district-level differences or preferences.

Gerrity, Osborn, and Morehouse Mendez address this problem by using longitudinal data of congressional behavior over several legislative sessions.[16] Rather than study legislators as their unit of analysis, they study seats that have turned over and compare differences in seats where (1) a woman has replaced a man with seats where (2) a woman has replaced a woman with seats where (3) a man has replaced a man. They limit their sample to seat changes within the same party to control for any partisan differences. They find that women replacing men are in fact more likely to support bills dealing with women's issues but that they are no more likely to discuss women's issues in floor speeches.

The reverse-causality problem can be found in a great deal of cross-national research or any research where a geographical division is the unit of analysis. Consider for example the strong cross-sectional relationship that we can observe between democracy and economic development. Wealthy countries are almost all democratic and many lower-income countries are authoritarian. Observing this relationship, however, doesn't tell us whether democracy drives economic development or whether economic development causes democracy. To address this problem, several scholars, such as Acemoglu and his colleagues, have used a method known as fixed-effects estimation. If country is the unit of analysis, rather than look at differences *between* countries, **fixed-effects estimations** look at differences *within* countries over time. The authors explain:

> Consider, for example, the comparison of the United States and Colombia. The United States is both richer and more democratic, so a simple cross-country comparison . . . would suggest that higher per capita income cause democracy. The idea of fixed effects is to move beyond this comparison and investigate the "within-country variation," that is, to ask whether Colombia is more likely to become (relatively) more democratic as it *becomes* (relatively) richer.[17]

[16] Jessica C. Gerrity, Tracy Osborn, and Jeanette Morehouse Mendez, "Women and representation: A different view of the district?" *Politics and Gender* 3 (2007): 179–200.

Longitudinal analyses using fixed-effects models are also better able to address omitted variable-bias problems. For example, in cross-national research, there are so many differences between countries that no study can control for all of them. While there are general variables that can be included in such an analysis, there are also a number of country-specific factors that cannot be included, such as historical factors or cultural norms. Without taking these into account, however, cross-sectional studies might suffer from omitted variable bias. By comparing within-country data over time, fixed-effects models automatically control for idiosyncratic factors specific to a given country.

Fixed-effects models are so powerful because they do not just perform this analysis for one country, but they do it for all countries simultaneously. The estimation then combines all the separate in-country analyses together to determine if—generally speaking—countries become more democratic as they become wealthier. Of course, using this method requires having panel data, or data that can be analyzed both longitudinally and cross-sectionally. Fortunately, there are periodic measures of democracy and economic development for almost all countries, so in this case such a study is feasible. Perhaps surprisingly, and despite the strong cross-sectional relationship, Acemoglu and his colleagues looked at changes within countries over a hundred-year period and found that increases in economic development *do not* correspond with increases in democracy.

CONDUCTING YOUR OWN LARGE-N STUDY

Large-n studies offer an attractive tool for student-led research. Unlike experiments, which are often expensive for students to conduct on their own, and small-n studies, which might require extensive field research, students can get started with a large-n study simply by downloading a dataset. Rather than design their own survey or collect their own data, students can take advantage of the detailed and expensive work of the many researchers that have come before them. The amount of data that are readily available is astounding, and yet most of it goes completely unused by university students. The largest obstacle is often inadequate statistical training to understand techniques like regression and a lack of familiarity with data management tools such as Stata, SPSS, SAS or R. On the other hand, even students who have been trained in these techniques and software packages are so used to thinking qualitatively that they are often hesitant to take advantage of such quantitative resources.

[17] Daron Acemoglu, Simon Johnson, James A. Robinson, *et al.*, "Income and democracy," *American Economic Review* 98 (2008): 808–842, p. 810.

While we encourage students to pursue more detailed statistical training than this book allows for, even students without a strong statistical background can take advantage of quantitative resources. Even if one does not feel comfortable with some of the sophisticated tools of causal inference, he or she can certainly use large-n quantitative data for descriptive inferences. How does a country of interest compare with others in terms of corruption perception? What is the president's approval rating and how has it changed over time? What do Americans think about the 2010 health care law? These are all questions of descriptive inference that thoughtful students without an extensive background in statistics can use quantitative data to answer.

Even in terms of causal inference, cross-tabulations, means comparisons, and scatterplots offer relatively easy-to-use tools that do not require detailed statistical or mathematical knowledge beyond a thorough understanding of the logic of inference. One particularly easy-to-use resource for cross-national data available to students without any statistical background is Gapminder. Available at www.gapminder.org, the Gapminder software allows students without knowledge of statistical software packages to develop scatterplots using an enormous quantity of cross-national indicators. Consider the scatterplot presented in Figure 6.8 from Gapminder. Not only does the figure clearly show the bivariate relationship between GDP per capita and life expectancy, but individual countries are also labeled, the size of the plot shows population size, and the shading of the plot indicates region. This one picture communicates an enormous amount of information and is an excellent example of good data presentation. The software even allows users to animate data and see changes over time.

For students who do have statistical training, there is a host of additional resources. There are numerous survey datasets that are available with a simple click of a button. The World Values Survey, for example, is publicly available and offers students access to numerous survey questions tapping into culture, politics, and economics in close to 100 countries (www.worldvaluessurvey.org). Pew's Global Attitudes Project offers access to similar surveys conducted by Pew (www.pewglobal.org). Gallup provides a searchable database on their website where one can search for results from all of their public opinion polling and articles (http://brain.gallup.com). Consortiums of scholars produce regular public opinion surveys across many regions of the world, including the Latinobarómetro (www.latinobarometro.org), AmericasBarometer (www.vanderbilt.edu/lapop), Eurobarometer (ec.europa.eu/public_opinion/index_en.htm), Afrobarometer (www.afrobarometer.org), and Asian Barometer (www.asianbarometer.org).

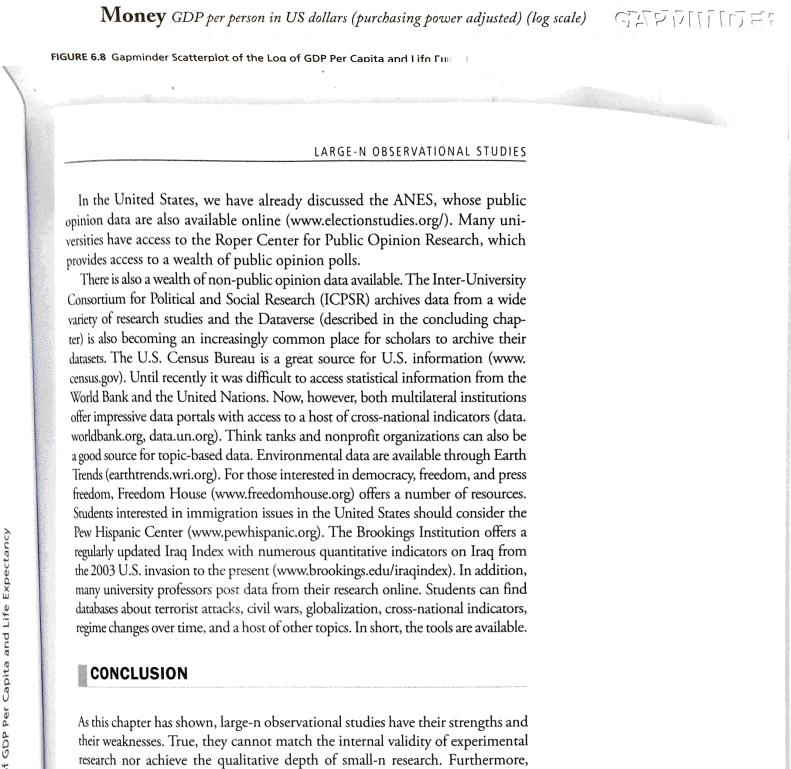


FIGURE 6.8 Gapminder Scatterplot of the Log of GDP Per Capita and Life Expectancy

Source: Gapminder.org

In the United States, we have already discussed the ANES, whose public opinion data are also available online (www.electionstudies.org/). Many universities have access to the Roper Center for Public Opinion Research, which provides access to a wealth of public opinion polls.

There is also a wealth of non-public opinion data available. The Inter-University Consortium for Political and Social Research (ICPSR) archives data from a wide variety of research studies and the Dataverse (described in the concluding chapter) is also becoming an increasingly common place for scholars to archive their datasets. The U.S. Census Bureau is a great source for U.S. information (www.census.gov). Until recently it was difficult to access statistical information from the World Bank and the United Nations. Now, however, both multilateral institutions offer impressive data portals with access to a host of cross-national indicators (data.worldbank.org, data.un.org). Think tanks and nonprofit organizations can also be a good source for topic-based data. Environmental data are available through Earth Trends (earthtrends.wri.org). For those interested in democracy, freedom, and press freedom, Freedom House (www.freedomhouse.org) offers a number of resources. Students interested in immigration issues in the United States should consider the Pew Hispanic Center (www.pewhispanic.org). The Brookings Institution offers a regularly updated Iraq Index with numerous quantitative indicators on Iraq from the 2003 U.S. invasion to the present (www.brookings.edu/iraqindex). In addition, many university professors post data from their research online. Students can find databases about terrorist attacks, civil wars, globalization, cross-national indicators, regime changes over time, and a host of other topics. In short, the tools are available.

CONCLUSION

As this chapter has shown, large-n observational studies have their strengths and their weaknesses. True, they cannot match the internal validity of experimental research nor achieve the qualitative depth of small-n research. Furthermore, they run the risk of measurement and sampling error, omitted variable bias, reverse-causality, and generalizing from non-robust results. Nonetheless, scholars conducting large-n analyses have numerous tools at their disposal to reduce (although not eliminate) these challenges to inference. Random measurement and sampling error can be taken into account in tests of statistical significance. Systematic measurement error can be reduced by selecting accurate measurements and systematic sampling error through random sampling. In some cases, both can be addressed through weighting. The threat of omitted variable bias can be reduced by controlling for additional factors and by using fixed-effects models where longitudinal data are available. The problem of reverse-causality can also be addressed by examining changes over time. The problem of generalizing non-robust findings can be minimized through running multiple models, ethical scholarship, and the replication of results by other scholars. The flexibility offered by a growing body of increasingly sophisticated statistical techniques and the ever increasing availability of datasets make and will continue to make large-n observational studies an attractive research method in political science.

The increasing sophistication of statistical techniques is, however, something of a double edged sword. Students of political science run the risk of falling into two camps, those who blindly accept quantitative work as inherently authoritative and those who reject it entirely. Hopefully this chapter has illustrated that both approaches are incorrect. As with other research methods, large-n observational studies confront numerous challenges to inference. By recognizing these challenges and taking steps to confront them, scholars can make valid and reliable inferences. By developing an understanding of the logic of inference, students of political science can become more knowledgeable consumers of research as well as scholars in their own right. Admittedly, it is difficult for students not to become bogged down or confused by complex statistical analyses; however, by focusing on the logic of inference, students will at least know the right questions to ask in evaluating a piece of scholarship. Rather than try to understand all the complexities of a nuanced statistical argument, students should ask: Did the authors address problems of sampling and measurement error, omitted variable bias, and reverse-causality?

In his book *Damned Lies and Statistics*, Joel Best sums up both the problem and the solution when he writes:

> The solution to the problem of bad statistics is not to ignore all statistics, or to assume that every number is false. Some statistics are bad, but others are pretty good, and we need statistics—good statistics—to talk sensibly about social problems. The solution, then, is not to give up on statistics, but to become better judges of the numbers we encounter. We need to think critically about statistics. . . .[18]

KEY TERMS

bivariate regression line	155
controlled cross-tabulation	167
controlled means comparison	153
cross-tabulation	166
data mining	163
dichotomous	167
fixed-effects estimations	172
independent effect	158
logistic regression	167
means comparison	151
multivariate linear regression	156
null hypothesis	153
omitted variable bias	151
regression coefficient	155
reverse-causality	170
robust	161
statistically significant relationship	153

[18] Joel Best, *Damned lies and statistics: Untangling numbers from the media, politicians, and activists* (Berkeley: University of California Press, 2012), p. 6.