19

Carrying Out an Empirical Project

In this chapter, we discuss the ingredients of a successful empirical analysis, with emphasis on completing a term project. In addition to reminding you of the important issues that have arisen throughout the text, we emphasize recurring themes that are important for applied research. We also provide suggestions for topics as a way of stimulating your imagination. Several sources of economic research and data are given as references.

19.1 Posing a Question

The importance of posing a very specific question cannot be overstated. Without being explicit about the goal of your analysis, you cannot know where to begin. The widespread availability of rich data sets makes it tempting to launch into data collection based on half-baked ideas, but this is often counterproductive. It is likely that, without carefully formulating your hypotheses and the kind of model you will need to estimate, you will forget to collect information on important variables, obtain a sample from the wrong population, or collect data for the wrong time period.

This does not mean that you should pose your question in a vacuum. Especially for a one-term project, you cannot be too ambitious. Therefore, when choosing a topic, you should be reasonably sure that data sources exist that will allow you to answer your question in the allotted time.

You need to decide what areas of economics or other social sciences interest you when selecting a topic. For example, if you have taken a course in labor economics, you have probably seen theories that can be tested empirically or relationships that have some policy relevance. Labor economists are constantly coming up with new variables that can explain wage differentials. Examples include quality of high school (Card and Krueger [1992] and Betts [1995]), amount of math and science taken in high school (Levine and Zimmerman [1995]), and physical appearance (Hamermesh and Biddle [1994], Averett and Korenman [1996], and Biddle and Hamermesh [1998]). Researchers in state and local public finance study how local economic activity depends on economic policy variables, such as property taxes, sales taxes, level and quality of services (such as schools, fire, and police), and so on. (See, for example, White [1986], Papke [1987], Bartik [1991], and Netzer [1992].]

Economists that study education issues are interested in determining how spending affects performance (Hanushek [1986]), whether attending certain kinds of schools improves performance (for example, Evans and Schwab [1995]), and what factors affect where private schools choose to locate (Downes and Greenstein [1996]).

Macroeconomists are interested in relationships between various aggregate time series, such as the link between growth in gross domestic product and growth in fixed investment or machinery (see De Long and Summers [1991]) or the effect of taxes on interest rates (for example, Peek [1982]).

There are certainly reasons for estimating models that are mostly descriptive. For example, property tax assessors use models (called *hedonic price models*) to estimate housing values for homes that have not been sold recently. This involves a regression model relating the price of a house to its characteristics (size, number of bedrooms, number of bathrooms, and so on). As a topic for a term paper, this is not very exciting: we are unlikely to learn much that is surprising, and such an analysis has no obvious policy implications. Adding the crime rate in the neighborhood as an explanatory variable would allow us to determine how important a factor crime is on housing prices, something that would be useful in estimating the costs of crime.

Several relationships have been estimated using macroeconomic data that are mostly descriptive. For example, an aggregate saving function can be used to estimate the aggregate marginal propensity to save, as well as the response of saving to asset returns (such as interest rates). Such an analysis could be made more interesting by using time series data on a country that has a history of political upheavals and determining whether savings rates decline during times of political uncertainty.

Once you decide on an area of research, there are a variety of ways to locate specific papers on the topic. The *Journal of Economic Literature* (JEL) has a detailed classification system in which each paper is given a set of identifying codes that places it within certain subfields of economics. The JEL also contains a list of articles published in a wide variety of journals, organized by topic, and it even contains short abstracts of some articles.

Especially convenient for finding published papers on various topics are **Internet** services, such as *EconLit*, which many universities subscribe to. *EconLit* allows users to do a comprehensive search of almost all economics journals by author, subject, words in the title, and so on. The *Social Sciences Citation Index* is useful for finding papers on a broad range of topics in the social sciences, including popular papers that have been cited often in other published works.

In thinking about a topic, you should keep some things in mind. First, for a question to be interesting, it does not need to have broad-based policy implications; rather, it can be of local interest. For example, you might be interested in knowing whether living in a fraternity at your university causes students to have lower or higher grade point averages. This may or may not be of interest to people outside your university, but it is probably of concern to at least some people within the university. On the other hand, you might study a problem that starts by being of local interest but turns out to have widespread interest, such as determining which factors affect, and which university policies can stem, alcohol abuse on college campuses.

Second, it is very difficult, especially for a quarter or semester project, to do truly original research using the standard macroeconomic aggregates on the U.S. economy. For example, the question of whether money growth, government spending growth, and

so on affect economic growth has been and continues to be studied by professional macroeconomists. The question of whether stock or other asset returns can be systematically predicted using known information has, for obvious reasons, been studied pretty carefully. This does not mean that you should avoid estimating macroeconomic or empirical finance models, as even just using more recent data can add constructively to a debate. In addition, you can sometimes find a new variable that has an important effect on economic aggregates or financial returns; such a discovery can be exciting.

The point is that exercises such as using a few additional years to estimate a standard Phillips curve or an aggregate consumption function for the U.S. economy, or some other large economy, are unlikely to yield additional insights, although they can be instructive for the student. Instead, you might use data on a smaller country to estimate a static or dynamic Phillips curve, or to test the efficient markets hypothesis, and so on.

At the nonmacroeconomic level, there are also plenty of questions that have been studied extensively. For example, labor economists have published many papers on estimating the return to education. This question is still studied because it is very important, and new data sets, as well as new econometric approaches, continue to be developed. For example, as we saw in Chapter 9, certain data sets have better proxy variables for unobserved ability than other data sets. (Compare WAGE1.RAW and WAGE2.RAW.) In other cases, we can obtain panel data or data from a natural experiment—see Chapter 13—that allow us to approach an old question from a different perspective.

As another example, criminologists are interested in studying the effects of various laws on crimes. The question of whether capital punishment has a deterrent effect has long been debated. Similarly, economists have been interested in whether taxes on cigarettes and alcohol reduce consumption (as always, in a ceteris paribus sense). As more years of data at the state level become available, a richer panel data set can be created, and this can help us better answer major policy questions. Plus, the effectiveness of fairly recent crime-fighting innovations—such as community policing—can be evaluated empirically.

While you are formulating your question, it is helpful to discuss your ideas with your classmates, instructor, and friends. You should be able to convince people that the answer to your question is of some interest. (Of course, whether you can persuasively answer your question is another issue, but you need to begin with an interesting question.) If someone asks you about your paper and you respond with "I'm doing my paper on crime" or "I'm doing my paper on interest rates," chances are you have only decided on a general area without formulating a true question. You should be able to say something like "I'm studying the effects of community policing on city crime rates in the United States" or "I'm looking at how inflation volatility affects short-term interest rates in Brazil."

19.2 Literature Review

All papers, even if they are relatively short, should contain a review of relevant literature. It is rare that one attempts an empirical project for which no published precedent exists. If you search through journals or use **online search services** such as *EconLit* to come up with a topic, you are already well on your way to a literature review. If you select a topic on your own—such as studying the effects of drug usage on college per-

formance at your university—then you will probably have to work a little harder. But online search services make that work a lot easier, as you can search by keywords, by words in the title, by author, and so on. You can then read abstracts of papers to see how relevant they are to your own work.

When doing your literature search, you should think of related topics that might not show up in a search using a handful of keywords. For example, if you are studying the effects of drug usage on wages or grade point average, you should probably look at the literature on how alcohol usage affects such factors. Knowing how to do a thorough literature search is an acquired skill, but you can get a long way by thinking before searching.

Researchers differ on how a literature review should be incorporated into a paper. Some like to have a separate section called "literature review," while others like to include the literature review as part of the introduction. This is largely a matter of taste, although an extensive literature review probably deserves its own section. If the term paper is the focus of the course—say, in a senior seminar or an advanced econometrics course—your literature review probably will be lengthy. Term papers at the end of a first course are typically shorter, and the literature reviews are briefer.

19.3 Data Collection

Deciding on the Appropriate Data Set

Collecting data for a term paper can be educational, exciting, and sometimes even frustrating. You must first decide on the kind of data needed to answer your posed question. As we discussed in the introduction and have covered throughout this text, data sets come in a variety of forms. The most common kinds are cross-sectional, time series, pooled cross sections, and panel data sets.

Many questions can be addressed using any of the data structures we have described. For example, to study whether more law enforcement lowers crime, we could use a cross section of cities, a time series for a given city, or a panel data set of cities—which consists of data on the same cities over two or more years.

Deciding on which kind of data to collect often depends on the nature of the analysis. To answer questions at the individual or family level, we often only have access to a single cross section; typically, these are obtained via surveys. Then, we must ask whether we can obtain a rich enough data set to do a convincing ceteris paribus analysis. For example, suppose we want to know whether families who save through individual retirement accounts (IRAs)—which have certain tax advantages—have less non-IRA savings. In other words, does IRA saving simply crowd out other forms of saving? There are data sets, such as the Survey of Consumer Finances, that contain information on various kinds of saving for a different sample of families each year. Several issues arise in using such a data set. Perhaps the most important is whether there are enough controls—including income, demographics, and proxies for saving tastes—to do a reasonable ceteris paribus analysis. If these are the only kinds of data available, we must do what we can with them.

The same issues arise with cross-sectional data on firms, cities, states, and so on. In most cases, it is not obvious that we will be able to do a ceteris paribus analysis with a

single cross section. For example, any study of the effects of law enforcement on crime must recognize the endogeneity of law enforcement expenditures. When using standard regression methods, it may be very hard to complete a convincing ceteris paribus analysis, no matter how many controls we have. (See Section 19.4 for more discussion.)

If you have read the advanced chapters on panel data methods, you know that having the same cross-sectional units at two or more different points in time can allow us to control for time-constant unobserved effects that would normally confound regression on a single cross section. Panel data sets are relatively hard to obtain for individuals or families—although some important ones exist, such as the Panel Study of Income Dynamics—but they can be used in very convincing ways. Panel data sets on firms also exist. For example, Compustat and the Center for Research in Security Prices (CRSP) manage very large panel data sets of financial information on firms. Easier to obtain are panel data sets on larger units, such as schools, cities, counties, and states, as these tend not to disappear over time, and government agencies are responsible for collecting information on the same variables each year. For example, the Federal Bureau of Investigation collects and reports detailed information on crime rates at the city level. Sources of data are listed at the end of this chapter.

Data come in a variety of forms. Some data sets, especially historical ones, are available only in printed form. For small data sets, entering the data yourself from the printed source is manageable and convenient. Sometimes, articles are published with small data sets—especially time series applications. These can be used in an empirical study, perhaps by supplementing the data with more recent years.

Many data sets are available on computer diskettes or magnetic tapes. The former are especially easy to work with. Currently, very large data sets can be put on small diskettes. Various government agencies sell data diskettes, as do private firms. Authors of papers are often willing to provide their data sets in diskette form or as e-mail attachments.

More and more data sets are available on the World Wide Web. The web is a vast resource of **online databases**. Numerous websites containing economic and related data sets have recently been created. Several other websites contain links to data sets that are of interest to economists; some of these are listed at the end of this chapter. Generally, searching the Internet for data sources is fairly easy and will become even more convenient in the future.

Entering and Storing Your Data

Once you have decided on a data type and have located a data source, you must put the data into usable form. If the data came on a diskette, they are already in some form, hopefully one in widespread use. The most flexible way to obtain data in diskette form is as a standard **text (ASCII) file**. All statistics and econometrics software packages allow raw data to be stored this way. Typically, it is straightforward to read a text file directly into an econometrics package, provided the file is properly structured. The data files we have used throughout the text provide several examples of how cross-sectional, time series, pooled cross sections, and panel data sets are usually stored. As a rule, the data should have a tabular form, with each observation representing a different row; the columns in the data set represent different variables. Occasionally, you might encounter a data set

stored with each column representing an observation and each row a different variable. This is not ideal, but most software packages allow data to be read in this form and then reshaped. Naturally, it is crucial to know how the data are organized before reading them into your econometrics package.

For time series data sets, there is only one sensible way to enter and store the data: namely, chronologically, with the earliest time period listed as the first observation and the most recent time period as the last observation. It is often useful to include variables indicating year and, if relevant, quarter or month. This facilitates estimation of a variety of models later on, including allowing for seasonality and breaks at different time periods. For cross sections pooled over time, it is usually best to have the cross section for the earliest year fill the first block of observations, followed by the cross section for the second year, and so on. (See FERTIL1.RAW as an example.) This arrangement is not crucial, but it is very important to have a variable stating the year attached to each observation.

For panel data, as we discussed in Section 13.5, it is best if all the years for each cross-sectional observation are adjacent and in chronological order. With this ordering, we can use all of the panel data methods from Chapters 13 and 14. With panel data, it is important to include a unique identifier for each cross-sectional unit, along with a year variable.

If you obtain your data in printed form, you have several options for entering it into a computer. First, you can create a text file using a standard **text editor**. (This is how several of the raw data sets included with the text were initially created.) Typically, it is required that each row starts a new observation, that each row contains the same ordering of the variables—in particular, each row should have the same number of entries—and that the values are separated by at least one space. Sometimes, a different separator, such as a comma, is better, but this depends on the software you are using. If you have missing observations on some variables, you must decide how to denote that; simply leaving a blank does not generally work. Many regression packages accept a period as the missing value symbol. Some people prefer to use a number—presumably an impossible value for the variable of interest—to denote missing values. If you are not careful, this can be dangerous; we discuss this further later.

If you have nonnumerical data—for example, you want to include the names in a sample of colleges or the names of cities—then you should check the econometrics package you will use to see the best way to enter such variables (often called *strings*). Typically, strings are put between double or single quotation marks. Or, the text file can follow a rigid formatting, which usually requires a small program to read in the text file. But you need to check your econometrics package for details.

Another generally available option is to use a **spreadsheet** to enter your data, such as Excel. This has a couple of advantages over a text file. First, because each observation on each variable is a cell, it is less likely that numbers will be run together (as would happen if you forget to enter a space in a text file). Second, spreadsheets allow manipulation of data, such as sorting or computing averages. This benefit is less important if you use a software package that allows sophisticated data management; many software packages, including EViews and Stata, fall into this category. If you use a spreadsheet for initial data entry, then you must often export the data in a form that can be read by your econometrics package. This is usually straightforward, as spreadsheets export to text files using a variety of formats.

A third alternative is to enter the data directly into your econometrics package. Although this obviates the need for a text editor or a spreadsheet, it can be more awkward if you cannot freely move across different observations to make corrections or additions.

Data downloaded from the Internet may come in a variety of forms. Often data come as text files, but different conventions are used for separating variables; for panel data sets, the conventions on how to order the data may differ. Some Internet data sets come as spreadsheet files, in which case you must use an appropriate spreadsheet to read them.

Inspecting, Cleaning, and Summarizing Your Data

It is extremely important to become familiar with any data set you will use in an empirical analysis. If you enter the data yourself, you will be forced to know everything about it. But if you obtain data from an outside source, you should still spend some time understanding its structure and conventions. Even data sets that are widely used and heavily documented can contain glitches. If you are using a data set obtained from the author of a paper, you must be aware that rules used for data set construction can be forgotten.

Earlier, we reviewed the standard ways that various data sets are stored. You also need to know how missing values are coded. Preferably, missing values are indicated with a nonnumeric character, such as a period. If a number is used as a missing value code, such as "999" or "−1", you must be very careful when using these observations in computing any statistics. Your econometrics package will probably not know that a certain number really represents a missing value: it is likely that such observations will be used as if they are valid, and this can produce rather misleading results. The best approach is to set any numerical codes for missing values to some other character (such as a period) that cannot be mistaken for real data.

You must also know the nature of the variables in the data set. Which are binary variables? Which are ordinal variables (such as a credit rating)? What are the units of measurement of the variables? For example, are monetary values expressed in dollars, thousands of dollars, millions of dollars, or so on? Are variables representing a rate—such as school dropout rates, inflation rates, unionization rates, or interest rates—measured as a percentage or a proportion?

Especially for time series data, it is crucial to know if monetary values are in nominal (current) or real (constant) dollars. If the values are in real terms, what is the base year or period?

If you receive a data set from an author, some variables may already be transformed in certain ways. For example, sometimes only the log of a variable (such as wage or salary) is reported in the data set.

Detecting mistakes in a data set is necessary for preserving the integrity of any data analysis. It is always useful to find minimums, maximums, means, and standard deviations of all, or at least the most important, variables in the analysis. For example, if you find that the minimum value of education in your sample is −99, you know that at least one entry on education needs to be set to a missing value. If, upon further inspection, you find that several observations have −99 as the level of education, you can be confident that you have stumbled onto the missing value code for education. As another example, if you find that an average murder conviction rate across a sample of cities is .632, you know

that conviction rate is measured as a proportion, not a percentage. Then, if the maximum value is above one, this is likely a typographical error. (It is not uncommon to find data sets where most of the entries on a rate variable were entered as a percentage, but where some were entered as a proportion, and vice versa. Such data coding errors can be difficult to detect, but it is important to try.)

We must also be careful in using time series data. If we are using monthly or quarterly data, we must know which variables, if any, have been seasonally adjusted. Transforming data also requires great care. Suppose we have a monthly data set and we want to create the change in a variable from one month to the next. To do this, we must be sure that the data are ordered chronologically, from earliest period to latest. If for some reason this is not the case, the differencing will result in garbage. To be sure the data are properly ordered, it is useful to have a time period indicator. With annual data, it is sufficient to know the year, but then we should know whether the year is entered as four digits or two digits (for example, 1998 versus 98). With monthly or quarterly data, it is also useful to have a variable or variables indicating month or quarter. With monthly data, we may have a set of dummy variables (11 or 12) or one variable indicating the month (1 through 12 or a string variable, such as *jan*, *feb*, and so on).

With or without yearly, monthly, or quarterly indicators, we can easily construct time trends in all econometrics software packages. Creating seasonal dummy variables is easy if the month or quarter is indicated; at a minimum, we need to know the month or quarter of the first observation.

Manipulating panel data can be even more challenging. In Chapter 13, we discussed pooled OLS on the differenced data as one general approach to controlling for unobserved effects. In constructing the differenced data, we must be careful not to create phantom observations. Suppose we have a balanced panel on cities from 1992 through 1997. Even if the data are ordered chronologically within each cross-sectional unit—something that should be done before proceeding—a mindless differencing will create an observation for 1992 for all cities except the first in the sample. This observation will be the 1992 value for city i, minus the 1997 value for city $i - 1$; this is clearly nonsense. Thus, we must make sure that 1992 is missing for all differenced variables.

19.4 Econometric Analysis

This text has focused on econometric analysis, and we are not about to provide a review of econometric methods in this section. Nevertheless, we can give some general guidelines about the sorts of issues that need to be considered in an empirical analysis.

As we discussed earlier, after deciding on a topic, we must collect an appropriate data set. Assuming that this has also been done, we must next decide on the appropriate econometric methods.

If your course has focused on ordinary least squares estimation of a multiple linear regression model, using either cross-sectional or time series data, the econometric approach has pretty much been decided for you. This is not necessarily a weakness, as OLS is still the most widely used econometric method. Of course, you still have to decide

whether any of the variants of OLS—such as weighted least squares or correcting for serial correlation in a time series regression—are required.

In order to justify OLS, you must also make a convincing case that the key OLS assumptions are satisfied for your model. As we have discussed at some length, the first issue is whether the error term is uncorrelated with the explanatory variables. Ideally, you have been able to control for enough other factors to assume that those that are left in the error are unrelated to the regressors. Especially when dealing with individual-, family-, or firm-level cross-sectional data, the self-selection problem—which we discussed in Chapters 7 and 15—is often relevant. For instance, in the IRA example from Section 19.3, it may be that families with unobserved taste for saving are also the ones that open IRAs. You should also be able to argue that the other potential sources of endogeneity—namely, measurement error and simultaneity—are not a serious problem.

When specifying your model you must also make functional form decisions. Should some variables appear in logarithmic form? (In econometric applications, the answer is often yes.) Should some variables be included in levels and squares, to possibly capture a diminishing effect? How should qualitative factors appear? Is it enough to just include binary variables for different attributes or groups? Or, do these need to be interacted with quantitative variables? (See Chapter 7 for details.)

A common mistake, especially among beginners, is to incorrectly include explanatory variables in a regression model that are listed as numerical values but have no quantitative meaning. For example, in an individual-level data set that contains information on wages, education, experience, and other variables, an "occupation" variable might be included. Typically, these are just arbitrary codes that have been assigned to different occupations; the fact that an elementary school teacher is given, say, the value 453 while a computer technician is, say, 751 is relevant only in that it allows us to distinguish between the two occupations. It makes no sense to include the raw occupational variable in a regression model. (What sense would it make to measure the effect of increasing *occupation* by one unit when the one-unit increase has no quantitative meaning?) Instead, different dummy variables should be defined for different occupations (or groups of occupations, if there are many occupations). Then, the dummy variables can be included in the regression model. A less egregious problem occurs when an ordered qualitative variable is included as an explanatory variable. Suppose that in a wage data set a variable is included measuring "job satisfaction," defined on a scale from 1 to 7, with 7 being the most satisfied. Provided we have enough data, we would want to define a set of six dummy variables for, say, job satisfaction levels of 2 through 7, leaving job satisfaction level 1 as the base group. By including the six job satisfaction dummies in the regression, we allow a completely flexible relationship between the response variable and job satisfaction. Putting in the job satisfaction variable in raw form implicitly assumes that a one-unit increase in the ordinal variable has quantitative meaning. While the direction of the effect will often be estimated appropriately, interpreting the coefficient on an ordinal variable is difficult. If an ordinal variable takes on many values, then we can define a set of dummy variables for ranges of values. See Section 7.3 for an example.

Sometimes, we want to explain a variable that is an ordinal response. For example, one could think of using a job satisfaction variable of the type described above as the dependent variable in a regression model, with both worker and employer characteristics

among the independent variables. Unfortunately, with the job satisfaction variable in its original form, the coefficients in the model are hard to interpret: each measures the change in job satisfaction given a unit increase in the independent variable. Certain models—*ordered probit* and *ordered logit* are the most common—are well suited for ordered responses. These models essentially extend the binary probit and logit models we discussed in Chapter 17. (See Wooldridge [2002, Chapter 15] for a treatment of ordered response models.) A simple solution is to turn any ordered response into a binary response. For example, we could define a variable equal to one if job satisfaction is at least 4, and zero otherwise. Unfortunately, creating a binary variable throws away information and requires us to use a somewhat arbitrary cutoff.

For cross-sectional analysis, a secondary, but nevertheless important, issue is whether there is heteroskedasticity. In Chapter 8, we explained how this can be dealt with. The simplest way is to compute heteroskedasticity-robust statistics.

As we emphasized in Chapters 10, 11, and 12, time series applications require additional care. Should the equation be estimated in levels? If levels are used, are time trends needed? Is differencing the data more appropriate? If the data are monthly or quarterly, does seasonality have to be accounted for? If you are allowing for dynamics—for example, distributed lag dynamics—how many lags should be included? You must start with some lags based on intuition or common sense, but eventually it is an empirical matter.

If your model has some potential misspecification, such as omitted variables, and you use OLS, you should attempt some sort of **misspecification analysis** of the kinds we discussed in Chapters 3 and 5. Can you determine, based on reasonable assumptions, the direction of any bias in the estimators?

If you have studied the method of instrumental variables, you know that it can be used to solve various forms of endogeneity, including omitted variables (Chapter 15), errors-in-variables (Chapter 15), and simultaneity (Chapter 16). Naturally, you need to think hard about whether the instrumental variables you are considering are likely to be valid.

Good papers in the empirical social sciences contain **sensitivity analysis**. Broadly, this means you estimate your original model and modify it in ways that seem reasonable. Hopefully, the important conclusions do not change. For example, if you use as an explanatory variable a measure of alcohol consumption (say, in a grade point average equation), do you get qualitatively similar results if you replace the quantitative measure with a dummy variable indicating alcohol usage? If the binary usage variable is significant but the alcohol quantity variable is not, it could be that usage reflects some unobserved attribute that affects GPA and is also correlated with alcohol usage. But this needs to be considered on a case-by-case basis.

If some observations are much different from the bulk of the sample—say, you have a few firms in a sample that are much larger than the other firms—do your results change much when those observations are excluded from the estimation? If so, you may have to alter functional forms to allow for these observations or argue that they follow a completely different model. The issue of outliers was discussed in Chapter 9.

Using panel data raises some additional econometric issues. Suppose you have collected two periods. There are at least four ways to use two periods of panel data without resorting to instrumental variables. You can pool the two years in a standard OLS

analysis, as discussed in Chapter 13. Although this might increase the sample size relative to a single cross section, it does not control for time-constant unobservables. In addition, the errors in such an equation are almost always serially correlated because of an unobserved effect. Random effects estimation corrects the serial correlation problem and produces asymptotically efficient estimators, provided the unobserved effect has zero mean given values of the explanatory variables in all time periods.

Another possibility is to include a lagged dependent variable in the equation for the second year. In Chapter 9, we presented this as a way to at least mitigate the omitted variables problem, as we are in any event holding fixed the initial outcome of the dependent variable. This often leads to similar results as differencing the data, as we covered in Chapter 13.

With more years of panel data, we have the same options, plus an additional choice. We can use the fixed effects transformation to eliminate the unobserved effect. (With two years of data, this is the same as differencing.) In Chapter 15, we showed how instrumental variables techniques can be combined with panel data transformations to relax exogeneity assumptions even more. As a rule, it is a good idea to apply several reasonable econometric methods and compare the results. This often allows us to determine which of our assumptions are likely to be false.

Even if you are very careful in devising your topic, postulating your model, collecting your data, and carrying out the econometrics, it is quite possible that you will obtain puzzling results—at least some of the time. When that happens, the natural inclination is to try different models, different estimation techniques, or perhaps different subsets of data until the results correspond more closely to what was expected. Virtually all applied researchers search over various models before finding the "best" model. Unfortunately, this practice of **data mining** violates the assumptions we have made in our econometric analysis. The results on unbiasedness of OLS and other estimators, as well as the t and F distributions we derived for hypothesis testing, assume that we observe a sample following the population model and we estimate that model once. Estimating models that are variants of our original model violates that assumption because we are using the same set of data in a *specification search*. In effect, we use the outcome of tests by using the data to respecify our model. The estimates and tests from different model specifications are not independent of one another.

Some specification searches have been programmed into standard software packages. A popular one is known as *stepwise regression*, where different combinations of explanatory variables are used in multiple regression analysis in an attempt to come up with the best model. There are various ways that stepwise regression can be used, and we have no intention of reviewing them here. The general idea is either to start with a large model and keep variables whose p-values are below a certain significance level or to start with a simple model and add variables that have significant p-values. Sometimes, groups of variables are tested with an F test. Unfortunately, the final model often depends on the order in which variables were dropped or added. (For more on stepwise regression, see Draper and Smith [1981].) In addition, this is a severe form of data mining, and it is difficult to interpret t and F statistics in the final model. One might argue that stepwise regression simply automates what researchers do anyway in searching over various models. However, in most applications, one or two explanatory variables are of primary interest, and then the goal is to see how robust the coefficients on those variables are to either adding or dropping other variables, or to changing functional form.

In principle, it is possible to incorporate the effects of data mining into our statistical inference; in practice, this is very difficult and is rarely done, especially in sophisticated empirical work. (See Leamer [1983] for an engaging discussion of this problem.) But we can try to minimize data mining by not searching over numerous models or estimation methods until a significant result is found and then reporting only that result. If a variable is statistically significant in only a small fraction of the models estimated, it is quite likely that the variable has no effect in the population.

19.5 Writing an Empirical Paper

Writing a paper that uses econometric analysis is very challenging, but it can also be rewarding. A successful paper combines a careful, convincing data analysis with good explanations and exposition. Therefore, you must have a good grasp of your topic, good understanding of econometric methods, and solid writing skills. Do not be discouraged if you find writing an empirical paper difficult; most professional researchers have spent many years learning how to craft an empirical analysis and to write the results in a convincing form.

While writing styles vary, many papers follow the same general outline. The following paragraphs include ideas for section headings and explanations about what each section should contain. These are only suggestions and hardly need to be strictly followed. In the final paper, each section would be given a number, usually starting with one for the introduction.

Introduction

The introduction states the basic objectives of the study and explains why it is important. It generally entails a review of the literature, indicating what has been done and how previous work can be improved upon. (As discussed in Section 19.2, an extensive literature review can be put in a separate section.) Presenting simple statistics or graphs that reveal a seemingly paradoxical relationship is a useful way to introduce the paper's topic. For example, suppose that you are writing a paper about factors affecting fertility in a developing country, with the focus on education levels of women. An appealing way to introduce the topic would be to produce a table or a graph showing that fertility has been falling (say) over time and a brief explanation of how you hope to examine the factors contributing to the decline. At this point, you may already know that, ceteris paribus, more highly educated women have fewer children and that average education levels have risen over time.

Most researchers like to summarize the findings of their paper in the introduction. This can be a useful device for grabbing the reader's attention. For example, you might state that your best estimate of the effect of missing 10 hours of lecture during a 30-hour term is about one-half a grade point. But the summary should not be too involved because neither the methods nor the data used to obtain the estimates have yet been introduced.

Conceptual (or Theoretical) Framework

In this section, you describe the general approach to answering the question you have posed. It can be formal economic theory, but in many cases, it is an intuitive discussion about what conceptual problems arise in answering your question.

As an example, suppose you are studying the effects of economic opportunities and severity of punishment on criminal behavior. One approach to explaining participation in crime is to specify a utility maximization problem where the individual chooses the amount of time spent in legal and illegal activities, given wage rates in both kinds of activities, as well as variable measuring probability and severity of punishment for criminal activity. The usefulness of such an exercise is that it suggests which variables should be included in the empirical analysis; it gives guidance (but rarely specifics) as to how the variables should appear in the econometric model.

Often, there is no need to write down an economic theory. For econometric policy analysis, common sense usually suffices for specifying a model. For example, suppose you are interested in estimating the effects of participation in Aid to Families with Dependent Children (AFDC) on the effects of child performance in school. AFDC provides supplemental income, but participation also makes it easier to receive Medicaid and other benefits. The hard part of such an analysis is deciding on the set of variables that should be controlled for. In this example, we could control for family income (including AFDC and any other welfare income), mother's education, whether the family lives in an urban area, and other variables. Then, the inclusion of an AFDC participation indicator (hopefully) measures the nonincome benefits of AFDC participation. A discussion of which factors should be controlled for and the mechanisms through which AFDC participation might improve school performance substitute for formal economic theory.

Econometric Models and Estimation Methods

It is very useful to have a section that contains a few equations of the sort you estimate and present in the results section of the paper. This allows you to fix ideas about what the key explanatory variable is and what other factors you will control for. Writing equations containing error terms allows you to discuss whether OLS is a suitable estimation method.

The distinction between a *model* and an estimation method should be made in this section. A model represents a *population* relationship (broadly defined to allow for time series equations). For example, we should write

$$colGPA = \beta_0 + \beta_1 alcohol + \beta_2 hsGPA + \beta_3 SAT + \beta_4 female + u \qquad (19.1)$$

to describe the relationship between college GPA and alcohol consumption, with some other controls in the equation. Presumably, this equation represents a population, such as all undergraduates at a particuar university. There are no "hats" ($\hat{}$) on the β_j or on *colGPA* because this is a model, not an estimated equation. We do not put in numbers for the β_j because we do not know (and never will know) these numbers. Later, we will *estimate* them. In this section, do not anticipate the presentation of your empirical results. In other words, do not start with a general model and then say that you omitted certain variables because they turned out to be insignificant. Such discussions should be left for the results section.

A time series model to relate city-level car thefts to the unemployment rate and conviction rates could look like

$$thefts_t = \beta_0 + \beta_1 unem_t + \beta_2 unem_{t-1} + \beta_3 cars_t + \beta_4 convrate_t + \beta_5 convrate_{t-1} + u_t, \qquad (19.2)$$

where the t subscript is useful for emphasizing any dynamics in the equation (in this case, allowing for unemployment and the automobile theft conviction rate to have lagged effects).

After specifying a model or models, it is appropriate to discuss estimation methods. In most cases, this will be OLS, but, for example, in a time series equation, you might use feasible GLS to do a serial correlation correction (as in Chapter 12). However, the method for estimating a model is quite distinct from the model itself. It is not meaningful, for instance, to talk about "an OLS model." Ordinary least squares is a method of estimation, and so are weighted least squares, Cochrane-Orcutt, and so on. There are usually several ways to estimate any model. You should explain why the method you are choosing is warranted.

Any assumptions that are used in obtaining an estimable econometric model from an underlying economic model should be clearly discussed. For example, in the quality of high school example mentioned in Section 19.1, the issue of how to measure school quality is central to the analysis. Should it be based on average SAT scores, percentage of graduates attending college, student-teacher ratios, average education level of teachers, some combination of these, or possibly other measures?

We always have to make assumptions about functional form whether or not a theoretical model has been presented. As you know, constant elasticity and constant semi-elasticity models are attractive because the coefficients are easy to interpret (as percentage effects). There are no hard rules on how to choose functional form, but the guidelines discussed in Section 6.2 seem to work well in practice. You do not need an extensive discussion of functional form, but it is useful to mention whether you will be estimating elasticities or a semi-elasticity. For example, if you are estimating the effect of some variable on wage or salary, the dependent variable will almost surely be in logarithmic form, and you might as well include this in any equations from the beginning. You do not have to present every one, or even most, of the functional form variations that you will report later in the results section.

Often, the data used in empirical economics are at the city or county level. For example, suppose that for the population of small to midsize cities, you wish to test the hypothesis that having a minor league baseball team causes a city to have a lower divorce rate. In this case, you must account for the fact that larger cities will have more divorces. One way to account for the size of the city is to scale divorces by the city or adult population. Thus, a reasonable model is

$$\log(div/pop) = \beta_0 + \beta_1 mlb + \beta_2 perCath + \beta_3 \log(inc/pop) + other\ factors, \tag{19.3}$$

where *mlb* is a dummy variable equal to one if the city has a minor league baseball team and *perCath* is the percentage of the population that is Catholic (so a number such as 34.6 means 34.6%). Note that *div/pop* is a divorce rate, which is generally easier to interpret than the absolute number of divorces.

Another way to control for population is to estimate the model

$$\log(div) = \gamma_0 + \gamma_1 mlb + \gamma_2 perCath + \gamma_3 \log(inc) + \gamma_4 \log(pop) + other\ factors. \tag{19.4}$$

The parameter of interest, γ_1, when multiplied by 100, gives the percentage difference between divorce rates, holding population, percent Catholic, income, and whatever else is

in "other factors" constant. In equation (19.3), β_1 measures the percentage effect of minor league baseball on *div/pop*, which can change either because the number of divorces or the population changes. Using the fact that $\log(div/pop) = \log(div) - \log(pop)$ and $\log(inc/pop) = \log(inc) - \log(pop)$, we can rewrite (19.3) as

$$\log(div) = \beta_0 + \beta_1 mlb + \beta_2 perCath + \beta_3 \log(inc) + (1 - \beta_3)\log(pop) + other\ factors,$$

which shows that (19.3) is a special case of (19.4) with $\gamma_4 = (1 - \beta_3)$ and $\gamma_j = \beta_j$, $j = 0,1,2,3$. Alternatively, (19.4) is equivalent to adding $\log(pop)$ as an additional explanatory variable to (19.3). This makes it easy to test for a separate population effect on the divorce rate.

If you are using a more advanced estimation method, such as two stage least squares, you need to provide some reasons for doing so. If you use 2SLS, you must provide a careful discussion on why your IV choices for the endogenous explanatory variable (or variables) are valid. As we mentioned in Chapter 15, there are two requirements for a variable to be considered a good IV. First, it must be omitted from and exogenous to the equation of interest (structural equation). This is something we must assume. Second, it must have some partial correlation with the endogenous explanatory variable. This we can test. For example, in equation (19.1), you might use a binary variable for whether a student lives in a dormitory (*dorm*) as an IV for alcohol consumption. This requires that living situation has no direct impact on *colGPA*—so that it is omitted from (19.1)—and that it is uncorrelated with unobserved factors in u that have an effect on *colGPA*. We would also have to verify that *dorm* is partially correlated with *alcohol* by regressing *alcohol* on *dorm*, *hsGPA*, *SAT*, and *female*. (See Chapter 15 for details.)

You might account for the omitted variable problem (or omitted heterogeneity) by using panel data. Again, this is easily described by writing an equation or two. In fact, it is useful to show how to difference the equations over time to remove time-constant unobservables; this gives an equation that can be estimated by OLS. Or, if you are using fixed effects estimation instead, you simply state so.

As a simple example, suppose you are testing whether higher county tax rates reduce economic activity, as measured by per capita manufacturing output. Suppose that for the years 1982, 1987, and 1992, the model is

$$\log(manuf_{it}) = \beta_0 + \delta_1 d87_t + \delta_2 d92_t + \beta_1 tax_{it} + \ldots + a_i + u_{it},$$

where $d87_t$ and $d92_t$ are year dummy variables and tax_{it} is the tax rate for county i at time t (in percent form). We would have other variables that change over time in the equation, including measures for costs of doing business (such as average wages), measures of worker productivity (as measured by average education), and so on. The term a_i is the fixed effect, containing all factors that do not vary over time, and u_{it} is the idiosyncratic error term. To remove a_i, we can either difference across the years or use time-demeaning (the fixed effects transformation).

The Data

You should always have a section that carefully describes the data used in the empirical analysis. This is particularly important if your data are nonstandard or have not been

widely used by other researchers. Enough information should be presented so that a reader could, in principle, obtain the data and redo your analysis. In particular, all applicable public data sources should be included in the references, and short data sets can be listed in an appendix. If you used your own survey to collect the data, a copy of the questionnaire should be presented in an appendix.

Along with a discussion of the data sources, be sure to discuss the units of each of the variables (for example, is income measured in hundreds or thousands of dollars?). Including a table of variable definitions is very useful to the reader. The names in the table should correspond to the names used in describing the econometric results in the following section.

It is also very informative to present a table of summary statistics, such as minimum and maximum values, means, and standard deviations for each variable. Having such a table makes it easier to interpret the coefficient estimates in the next section, and it emphasizes the units of measurement of the variables. For binary variables, the only necessary summary statistic is the fraction of ones in the sample (which is the same as the sample mean). For trending variables, things like means are less interesting. It is often useful to compute the average growth rate in a variable over the years in your sample.

You should always clearly state how many observations you have. For time series data sets, identify the years that you are using in the analysis, including a description of any special periods in history (such as World War II). If you use a pooled cross section or a panel data set, be sure to report how many cross-sectional units (people, cities, and so on) you have for each year.

Results

The results section should include your estimates of any models formulated in the models section. You might start with a very simple analysis. For example, suppose that percentage of students attending college from the graduating class (*percoll*) is used as a measure of the quality of the high school a person attended. Then, an equation to estimate is

$$\log(wage) = \beta_0 + \beta_1 percoll + u.$$

Of course, this does not control for several other factors that may determine wages and that may be correlated with *percoll*. But a simple analysis can draw the reader into the more sophisticated analysis and reveal the importance of controlling for other factors.

If only a few equations are estimated, you can present the results in equation form with standard errors in parentheses below estimated coefficients. If your model has several explanatory variables and you are presenting several variations on the general model, it is better to report the results in tabular rather than equation form. Most of your papers should have at least one table, which should always include at least the R-squared and the number of observations for each equation. Other statistics, such as the adjusted R-squared, can also be listed.

The most important thing is to discuss the interpretation and strength of your empirical results. Do the coefficients have the expected signs? Are they statistically significant? If a coefficient is statistically significant but has a counterintuitive sign, why might this be true? It might be revealing a problem with the data or the econometric method (for example, OLS may be inappropriate due to omitted variables problems).

Be sure to describe the *magnitudes* of the coefficients on the major explanatory variables. Often, one or two policy variables are central to the study. Their signs, magnitudes,

and statistical significance should be treated in detail. Remember to distinguish between economic and statistical significance. If a t statistic is small, is it because the coefficient is practically small or because its standard error is large?

In addition to discussing estimates from the most general model, you can provide interesting special cases, especially those needed to test certain multiple hypotheses. For example, in a study to determine wage differentials across industries, you might present the equation without the industry dummies; this allows the reader to easily test whether the industry differentials are statistically significant (using the R-squared form of the F test). Do not worry too much about dropping various variables to find the "best" combination of explanatory variables. As we mentioned earlier, this is a difficult and not even very well-defined task. Only if eliminating a set of variables substantially alters the magnitudes and/or significance of the coefficients of interest is this important. Dropping a group of variables to simplify the model—such as quadratics or interactions—can be justified via an F test.

If you have used at least two different methods—such as OLS and 2SLS, or levels and differencing for a time series, or pooled OLS versus differencing with a panel data set—then you should comment on any critical differences. If OLS gives counterintuitive results, did using 2SLS or panel data methods improve the estimates? Or, did the opposite happen?

Conclusions

This can be a short section that summarizes what you have learned. For example, you might want to present the magnitude of a coefficient that was of particular interest. The conclusion should also discuss caveats to the conclusions drawn, and it might even suggest directions for further research. It is useful to imagine readers turning first to the conclusion to decide whether to read the rest of the paper.

Style Hints

You should give your paper a title that reflects its topic. Papers should be typed and double-spaced. All equations should begin on a new line, and they should be centered and numbered consecutively, that is, (1), (2), (3), and so on. Large graphs and tables may be included after the main body. In the text, refer to papers by author and date, for example, White (1980). The reference section at the end of the paper should be done in standard format. Several examples are given in the references at the back of the text.

When you introduce an equation in the econometric models section, you should describe the important variables: the dependent variable and the key independent variable or variables. To focus on a single independent variable, you can write an equation, such as

$$GPA = \beta_0 + \beta_1 alcohol + \mathbf{x}\boldsymbol{\delta} + u$$

or

$$\log(wage) = \beta_0 + \beta_1 educ + \mathbf{x}\boldsymbol{\delta} + u,$$

where the notation $\mathbf{x}\boldsymbol{\delta}$ is shorthand for several other explanatory variables. At this point, you need only describe them generally; they can be described specifically in the data section in a table. For example, in a study of the factors affecting chief executive officer salaries, you might include the following table in the data section:

TABLE 1
Variable Descriptions

salary	annual salary (including bonuses) in 1990 (in thousands)
sales	firm sales in 1990 (in millions)
roe	average return on equity, 1988–1990 (in percent)
pcsal	percentage change in salary, 1988–1990
pcroe	percentage change in roe, 1988–1990
indust	= 1 if an industrial company, 0 otherwise
finance	= 1 if a financial company, 0 otherwise
consprod	= 1 if a consumer products company, 0 otherwise
util	= 1 if a utility company, 0 otherwise
ceoten	number of years as CEO of the company

A table of summary statistics using the data set 401K.RAW, which we used for studying the factors that affect participation in 401(k) pension plans, might be set up as follows:

TABLE 2
Summary Statistics

Variable	Mean	Standard Deviation	Minimum	Maximum
prate	.869	.167	.023	1
mrate	.746	.844	.011	5
employ	4,621.01	16,299.64	53	443,040
age	13.14	9.63	4	76
sole	.415	.493	0	1
Number of Observations = 3,784				

In the results section, you can either write the estimates in equation form, as we often have done, or in a table. Especially when several models have been estimated with different sets of explanatory variables, tables are very useful. If you write out the estimates as an equation, for example,

$$\widehat{\log(salary)} = \underset{(0.93)}{2.45} + \underset{(.115)}{.236}\log(sales) + \underset{(.003)}{.008}\,roe + \underset{(.028)}{.061}\,ceoten$$
$$n = 204,\ R^2 = .351,$$

be sure to state near the first equation that standard errors are in parentheses. It is acceptable to report the t statistics for testing H_0: $\beta_j = 0$, or their absolute values, but it is most important to state what you are doing.

If you report your results in tabular form, make sure the dependent and independent variables are clearly indicated. Again, state whether standard errors or t statistics are below the coefficients (with the former preferred). Some authors like to use asterisks to indicate statistical significance at different significance levels (for example, one star means significant at 5%, two stars mean significant at 10% but not 5%, and so on). This is not necessary if you carefully discuss the significance of the explanatory variables in the text.

A sample table of results follows:

TABLE 3

OLS Results. Dependent Variable: Participation Rate

Independent Variables	**(1)**	**(2)**	**(3)**
mrate	.156 (.012)	.239 (.042)	.218 (.342)
$mrate^2$	—	−.087 (.043)	−.096 (.073)
log(*emp*)	−.112 (.014)	−.112 (.014)	−.098 (.111)
$\log(emp)^2$	.0057 (.0009)	.0057 (.0009)	.0052 (.0007)
age	.0060 (.0010)	.0059 (.0010)	.0050 (.0021)
age^2	−.00007 (.00002)	−.00007 (.00002)	−.00006 (.00002)

(continued)

TABLE 3

OLS Results. Dependent Variable: Participation Rate (*Continued*)

Independent Variables	**(1)**	**(2)**	**(3)**
sole	−.0001 (.0058)	.0008 (.0058)	.0006 (.0061)
constant	1.213 (.051)	.198 (.052)	.085 (.041)
industry dummies?	no	no	yes
Observations *R*-Squared	3,784 .143	3,784 .152	3,784 .162

Note: The quantities in parentheses below the estimates are the standard errors.

Your results will be easier to read and interpret if you choose the units of both your dependent and independent variables so that coefficients are not too large or too small. You should never report numbers such as 1.051e−007 or 3.524e+006 for your coefficients or standard errors, and you should not use scientific notation. If coefficients are either extremely small or large, rescale the dependent or independent variables, as we discussed in Chapter 6. You should limit the number of digits reported after the decimal point. For example, if your regression package estimates a coefficient to be .54821059, you should report this as .548, or even .55, in the paper.

As a rule, the commands that your particular econometrics package uses to produce results should not appear in the paper; only the results are important. If some special command was used to carry out a certain estimation method, this can be given in an appendix. An appendix is also a good place to include extra results that support your analysis but are not central to it.

SUMMARY

In this chapter, we have discussed the ingredients of a successful empirical study and have provided hints that can improve the quality of an analysis. Ultimately, the success of any study depends crucially on the care and effort put into it.

KEY TERMS

Data Mining
Internet
Misspecification Analysis
Online Databases
Online Search Services
Sensitivity Analysis
Spreadsheet
Text Editor
Text (ASCII) File

SAMPLE EMPIRICAL PROJECTS

Throughout the text, we have seen examples of econometric analysis that either came from or were motivated by published works. We hope these have given you a good idea about the scope of empirical analysis. We include the following list as additional examples of questions that others have found or are likely to find interesting. These are intended to stimulate your imagination; no attempt is made to fill in all the details of specific models, data requirements, or alternative estimation methods. It should be possible to complete these projects in one term.

1. Do your own campus survey to answer a question of interest at your university. For example: What is the effect of working on college GPA? You can ask students about high school GPA, college GPA, ACT or SAT scores, hours worked per week, participation in athletics, major, gender, race, and so on. Then, use these variables to create a model that explains GPA. How much of an effect, if any, does another hour worked per week have on GPA? One issue of concern is that hours worked might be endogenous: it might be correlated with unobserved factors that affect college GPA, or lower GPAs might cause students to work more.

   A better approach would be to collect cumulative GPA prior to the semester and then to obtain GPA for the most recent semester, along with amount worked during that semester, and the other variables. Now, cumulative GPA could be used as a control (explanatory variable) in the equation.
2. There are many variants on the preceding topic. You can study the effects of drug or alcohol usage, or of living in a fraternity, on grade point average. You would want to control for many family background variables, as well as previous performance variables.
3. Do gun control laws at the city level reduce violent crimes? Such questions can be difficult to answer with a single cross section because city and state laws are often endogenous. (See Kleck and Patterson [1993] for an example. They used cross-sectional data and instrumental variables methods, but their IVs are questionable.) Panel data can be very useful for inferring causality in these contexts. At a minimum, you could control for a previous year's violent crime rate.
4. Low and McPheters (1983) used city cross-sectional data on wage rates and estimates of risk of death for police officers, along with other controls. The idea is to determine whether police officers are compensated for working in cities with a higher risk of on-the-job injury or death.
5. Do parental consent laws increase the teenage birthrate? You can use state level data for this: either a time series for a given state or, even better, a panel data set of states. Do the same laws reduce abortion rates among teenagers? The *Statistical Abstract of the United States* contains all kinds of state-level data. Levine, Trainor, and Zimmerman (1996) studied the effects of abortion funding restrictions on similar outcomes. Other factors, such as access to abortions, may affect teen birth and abortion rates.
6. Do changes in traffic laws affect traffic fatalities? McCarthy (1994) contains an analysis of monthly time series data for the state of California. A set of dummy

variables can be used to indicate the months in which certain laws were in effect. The file TRAFFIC2.RAW contains the data used by McCarthy. An alternative is to obtain a panel data set on states in the United States, where you can exploit variation in laws across states, as well as across time. (See the file TRAFFIC1.RAW.)

Mullahy and Sindelar (1994) used individual-level data matched with state laws and taxes on alcohol to estimate the effects of laws and taxes on the probability of driving drunk.

7. Are blacks discriminated against in the lending market? Hunter and Walker (1996) looked at this question; in fact, we used their data in Computer Exercises C7.8 and C17.2.
8. Is there a marriage premium for professional athletes? Korenman and Neumark (1991) found a significant wage premium for married men after using a variety of econometric methods, but their analysis is limited because they cannot directly observe productivity. (Plus, Korenman and Neumark used men in a variety of occupations.) Professional athletes provide an interesting group in which to study the marriage premium because we can easily collect data on various productivity measures, in addition to salary. The data set NBASAL.RAW, on players in the National Basketball Association (NBA), is one example. For each player, we have information on points scored, rebounds, assists, playing time, and demographics. As in Computer Exercise C6.9, we can use multiple regression analysis to test whether the productivity measures differ by marital status. We can also use this kind of data to test whether married men are paid more after we account for productivity differences. (For example, NBA owners may think that married men bring stability to the team, or are better for the team image.) For individual sports—such as golf and tennis—annual earnings directly reflect productivity. Such data, along with age and experience, are relatively easy to collect.
9. Answer this question: Are cigarette smokers less productive? A variant on this is: Do workers who smoke take more sick days (everything else being equal)? Mullahy and Portney (1990) use individual-level data to evaluate this question. You could use data at, say, the metropolitan level. Something like average productivity in manufacturing can be related to percentage of manufacturing workers who smoke. Other variables, such as average worker education, capital per worker, and size of the city (you can think of more), should be controlled for.
10. Do minimum wages alleviate poverty? You can use state or county data to answer this question. The idea is that the minimum wage varies across states because some states have higher minimums than the federal minimum. Further, there are changes over time in the nominal minimum within a state, some due to changes at the federal level and some because of changes at the state level. Neumark and Wascher (1995) used a panel data set on states to estimate the effects of the minimum wage on the employment rates of young workers, as well as on school enrollment rates.
11. What factors affect student performance at public schools? It is fairly easy to get school-level or at least district-level data in most states. Does spending per student matter? Do student-teacher ratios have any effects? It is difficult to estimate ceteris paribus effects because spending is related to other factors, such as family

incomes or poverty rates. The data set MEAP93.RAW, for Michigan high schools, contains a measure of the poverty rates. Another possibility is to use panel data, or to at least control for a previous year's performance measure (such as average test score or percentage of students passing an exam).

You can look at less obvious factors that affect student performance. For example, after controlling for income, does family structure matter? Perhaps families with two parents, but only one working for a wage, have a positive effect on performance. (There could be at least two channels: parents spend more time with the children, and they might also volunteer at school.) What about the effect of single-parent households, controlling for income and other factors? You can merge census data for one or two years with school district data.

Do public schools with more private schools nearby better educate their students because of competition? There is a tricky simultaneity issue here because private schools are probably located in areas where the public schools are already poor. Hoxby (1994) used an instrumental variables approach, where population proportions of various religions were IVs for the number of private schools.

Rouse (1998) studied a different question: Did students who were able to attend a private school due to the Milwaukee voucher program perform better than those who did not? She used panel data and was able to control for an unobserved student effect.

12. Can excess returns on a stock, or a stock index, be predicted by the lagged price/dividend ratio? Or, by lagged interest rates or weekly monetary policy? It would be interesting to pick a foreign stock index, or one of the less well-known U.S. indexes. Cochrane (1997) provides a nice survey of recent theories and empirical results for explaining excess stock returns.
13. Is there racial discrimination in the market for baseball cards? This involves relating the prices of baseball cards to factors that should affect their prices, such as career statistics, whether the player is in the Hall of Fame, and so on. Holding other factors fixed, do cards of black or Hispanic players sell at a discount?
14. You can test whether the market for gambling on sports is efficient. For example, does the spread on football or basketball games contain all usable information for picking against the spread? The data set PNTSPRD.RAW contains information on men's college basketball games. The outcome variable is binary. Was the spread covered or not? Then, you can try to find information that was known prior to each game's being played in order to predict whether the spread is covered. (Good luck!)
15. What effect, if any, does success in college athletics have on other aspects of the university (applications, quality of students, quality of nonathletic departments)? McCormick and Tinsley (1987) looked at the effects of athletic success at major colleges on changes in SAT scores of entering freshmen. Timing is important here: presumably, it is recent past success that affects current applications and student quality. One must control for many other factors—such as tuition and measures of school quality—to make the analysis convincing because, without controlling for other factors, there is a negative correlation between academics and athletic performance.

A variant is to match natural rivals in football or men's basketball and to look at differences across schools as a function of which school won the football game or one or more basketball games. ATHLET1.RAW and ATHLET2.RAW are small data sets that could be expanded and updated.

16. Collect murder rates for a sample of counties (say, from the FBI Uniform Crime Reports) for two years. Make the latter year such that economic and demographic variables are easy to obtain from the *County and City Data Book*. You can obtain the total number of people on death row plus executions for intervening years at the county level. If the years are 1990 and 1985, you might estimate

$$mrdrte_{90} = \beta_0 + \beta_1 mrdrte_{85} + \beta_2 executions + other\ factors,$$

where interest is in the coefficient on *executions*. The lagged murder rate and other factors serve as controls.

Other factors may also act as a deterrent to crime. For example, Cloninger (1991) presented a cross-sectional analysis of the effects of lethal police response on crime rates.

As a different twist, what factors affect crime rates on college campuses? Does the fraction of students living in fraternities or sororities have an effect? Does the size of the police force matter, or the kind of policing used? (Be careful about inferring causality here.) Does having an escort program help reduce crime? What about crime rates in nearby communities? Recently, colleges and universities have been required to report crime statistics; in previous years, reporting was voluntary.

17. What factors affect manufacturing productivity at the state level? In addition to levels of capital and worker education, you could look at degree of unionization. A panel data analysis would be most convincing here, using two census years (say, 1980 and 1990). Clark (1984) provides an analysis of how unionization affects firm performance and productivity. What other variables might explain productivity?

Firm-level data can be obtained from *Compustat*. For example, other factors being fixed, do changes in unionization affect stock price of a firm?

18. Use state- or county-level data or, if possible, school district-level data to look at the factors that affect education spending per pupil. An interesting question is: Other things being equal (such as income and education levels of residents), do districts with a larger percentage of elderly people spend less on schools? Census data can be matched with school district spending data to obtain a very large cross section. The U.S. Department of Education compiles such data.
19. What are the effects of state regulations, such as motorcycle helmet laws, on motorcycle fatalities? Or, do differences in boating laws—such as minimum operating age—help to explain boating accident rates? The U.S. Department of Transportation compiles such information. This can be merged with data from the *Statistical Abstract of the United States*. A panel data analysis seems to be warranted here.
20. What factors affect output growth? Two factors of interest are inflation and investment (for example, Blomström, Lipsey, and Zejan [1996]). You might use time series data on a country you find interesting. Or, you could use a cross section of

countries, as in De Long and Summers (1991). Friedman and Kuttner (1992) found evidence that, at least in the 1980s, the spread between the commercial paper rate and the Treasury bill rate affects real output.

21. What is the behavior of mergers in the U.S. economy (or some other economy)? Shughart and Tollison (1984) characterize (the log of) annual mergers in the U.S. economy as a random walk by showing that the difference in logs—roughly, the growth rate—is unpredictable given past growth rates. Does this still hold? Does it hold across various industries? What past measures of economic activity can be used to forecast mergers?
22. What factors might explain racial and gender differences in employment and wages? For example, Holzer (1991) reviewed the evidence on the "spatial mismatch hypothesis" to explain differences in employment rates between blacks and whites. Korenman and Neumark (1992) examined the effects of childbearing on women's wages, while Hersch and Stratton (1997) looked at the effects of household responsibilities on men's and women's wages.
23. Obtain monthly or quarterly data on teenage employment rates, the minimum wage, and factors that affect teen employment to estimate the effects of the minimum wage on teen employment. Solon (1985) used quarterly U.S. data, while Castillo-Freeman and Freeman (1992) used annual data on Puerto Rico. It might be informative to analyze time series data on a low-wage state in the United States—where changes in the minimum wage are likely to have the largest effect.
24. At the city level, estimate a time series model for crime. An example is Cloninger and Sartorius (1979). As a recent twist, you might estimate the effects of community policing or midnight basketball programs, relatively new innovations in fighting crime. Inferring causality is tricky. Including a lagged dependent variable might be helpful. Because you are using time series data, you should be aware of the spurious regression problem.

    Grogger (1990) used data on daily homicide counts to estimate the deterrent effects of capital punishment. Might there be other factors—such as news on lethal response by police—that have an effect on daily crime counts?
25. Are there aggregate productivity effects of computer usage? You would need to obtain time series data, perhaps at the national level, on productivity, percentage of employees using computers, and other factors. What about spending (probably as a fraction of total sales) on research and development? What sociological factors (for example, alcohol usage or divorce rates) might affect productivity?
26. What factors affect chief executive officer salaries? The files CEOSAL1.RAW and CEOSAL2.RAW are data sets that have various firm performance measures as well as information such as tenure and education. You can certainly update these data files and look for other interesting factors. Rose and Shepard (1997) considered firm diversification as one important determinant of CEO compensation.
27. Do differences in tax codes across states affect the amount of foreign direct investment? Hines (1996) studied the effects of state corporate taxes, along with the ability to apply foreign tax credits, on investment from outside the United States.

28. What factors affect election outcomes? Does spending matter? Do votes on specific issues matter? Does the state of the local economy matter? See, for example, Levitt (1994) and the data sets VOTE1.RAW and VOTE2.RAW. Fair (1996) performed a time series analysis of U.S. presidential elections.
29. Test whether stores or restaurants practice price discrimination based on race or ethnicity. Graddy (1997) used data on fast-food restaurants in New Jersey and Pennsylvania, along with zip code-level characteristics, to see whether prices vary by characteristics of the local population. She found that prices of standard items, such as sodas, increase when the fraction of black residents increases. (Her data are contained in the file DISCRIM.RAW.) You can collect similar data in your local area by surveying stores or restaurants for prices of common items and matching those with recent census data. See Graddy's paper for details of her analysis.
30. Do your own "audit" study to test for race or gender discrimination in hiring. (One such study is described in Example C.3 of Appendix C.) Have pairs of equally qualified friends, say, one male and one female, apply for job openings in local bars or restaurants. You can provide them with phony resumes that give each the same experience and background, where the only difference is gender (or race). Then, you can keep track of who gets the interviews and job offers. Neumark (1996) described one such study conducted in Philadelphia. A variant would be to test whether general physical attractiveness or a specific characteristic, such as being obese or having visible tattoos or body piercings, plays a role in hiring decisions. You would want to use the same gender in the matched pairs, and it may not be easy to get volunteers for such a study.

LIST OF JOURNALS

The following is a partial list of popular journals containing empirical research in business, economics, and other social sciences. A complete list of journals can be found on the Internet at http://www.econlit.org.

American Economic Review
American Journal of Agricultural Economics
American Political Science Review
Applied Economics
Brookings Papers on Economic Activity
Canadian Journal of Economics
Demography
Economic Development and Cultural Change
Economic Inquiry
Economica
Economics Letters
Empirical Economics