CHAPTER 10 | # Conducting a Regression Study Using Economic Data

Conducting an empirical analysis of economic data can be rewarding and informative. If you follow some basic guidelines, it is possible to use your time efficiently and to avoid some potentially frustrating pitfalls. This chapter provides suggestions for how to conduct a regression study using economic data and how to report the results of that study.

One reason that this book repeatedly uses the California test score data is to illustrate the steps involved in undertaking a serious empirical application: becoming familiar with the data, developing and estimating a base regression specification, thinking through potential omitted variables, modeling relevant nonlinearities, performing sensitivity analysis, assessing the internal and external validity of the findings, and reporting the results and their limitations. This chapter steps back from the details of the test score application and describes the main steps in conducting an empirical analysis and reporting the results.

It is important to approach an empirical analysis with an open mind. It is tempting to think that your goal should be a high adjusted R^2 or an estimated coefficient of interest that is economically large and statistically significant. But this is not the purpose of a thoughtful empirical analysis; instead, the purpose is to answer a specific question while using your best judgment and being honest about what the data do and do not tell you. The coefficient of interest might be large and well estimated; it might be small and well estimated; or it might just be imprecisely estimated because of limitations of the data or because the question being asked is a very difficult one. Reaching any one of these conclusions— whether it confirms your prior suspicions or not—is interesting and helps you better to understand the topic you are researching.

In our analysis of the California test score data, if our objective had been to find a large coefficient, we might have stopped at the regression of test scores against the student–teacher ratio and never included any control variables. But, upon reflection, it became clear that that estimate was subject to considerable omitted variable bias, which was addressed by including the control variables. By the end of the analysis, we had concluded that the class size effect, while statistically significant, is economically small—a conclusion confirmed using a different observational data set (the Massachusetts data). The distinction is subtle between trying to measure an effect as reliably as possible and trying to prove that the effect is important, but it can make the difference between a study that is credible and one that is not.

10.1 Choosing a Topic

The first step in conducting an empirical analysis is choosing the topic you want to study and, within that topic, the specific question or questions you will investigate. Although there is not a single best way to choose a topic, the following suggestions might be useful.

1. *Pick a topic that you find personally interesting, ideally one about which you already have some knowledge*. The topic might be related to your career interests, summer work you did, employment experience of a family member, or something of intellectual interest to you. Often, a specific policy problem, a personal decision, or a business issue raises questions that can be addressed by an empirical study.

2. *Make the question that will be the main focus of your study as specific as possible*. The question, should the government increase school spending to reduce class sizes? is an important one but is too vague for immediate empirical study. The question, does a reduction in class size make students better off? is an improvement because it relates to a specific causal effect—the effect of class size reductions on student welfare—but it still is too vague (what does "better off" mean?). But, upon further narrowing, the question, does a reduction

in class size result in an increase in test scores? can in principle be answered using empirical analysis. The more narrowly the question relates to a measurable causal effect, the easier it will be to answer.

3. *Check the related literature.* You might find published studies on topics closely related to yours. Do not let this discourage you! Instead, use previous work to give you ideas about data sources and about what questions have not yet been answered. Web tools are very useful for finding related literature, and your instructor can give you additional suggestions about what journals to look in.

4. *Choose a question that can be answered using the available data.* Data sources and types of data are discussed in the next section. Although the question you originally pose might not be answerable using available data, the data might support the analysis of a related and equally interesting question.

5. *Discuss your topic with a classmate or your instructor.* If you find your topic interesting, then the odds are that others will too, and an instructor or classmate might suggest an angle that you have not thought of.

The choice of a question for analysis is linked to finding suitable data with which to answer the question.

10.2   Collecting Data

Finding a Data Set

The first step toward finding a data set is being as specific as possible about what data would help you to answer the question you are investigating. To do so, it helps to be clear about how you might actually conduct your analysis. What unit of observation would be most useful (individuals? firms? local governmental data? cross-country data? etc.)? What should you use as the dependent variable? What is the main independent variable of interest? What control variables would you consider to be most important, so as to address the main concerns about omitted variable bias?

With your list of variables in hand, the next step is to look for a data set. Your college or university might have a data librarian in the economics or social science library. If so, he or she might be able to suggest some data sets. An additional advantage of going to your college data librarian is that some data sets are only available to subscribers, and your college or university might subscribe to such data sets.

In addition to your data librarian, investigation on the Web can lead to some good data sources. Links to some of the main public sources for economic data are

available on this textbook's Web site (**www.aw-bc.com/stock_watson**). The data sources on this Web site include data in labor economics, law and economics, political economy, the economics of education, health economics, macroeconomics, and international economics.

With some creativity, the Web can be used to assemble interesting and new data sets. For example, one of our students collected data on the world top-100 times in the 100-meter dash each year since the early 1980s, along with control variables (such as wind speed and athlete age), and used regression methods to see if the usual annual improvement in these times was reduced when track and field authorities stepped up their enforcement of anti-doping rules (it was). All the data used in that analysis were downloaded from track and field Web sites. As a second example, Miles (2005) collected data from the Web sites of U.S. law enforcement agencies to see whether appearing on the TV show *America's Most Wanted* increased the chances of the subject being caught (it does; the researcher also collected data by watching the show).

In some cases, existing data sources might not suffice to answer the question you have in mind. Depending on your topic, you could consider conducting your own survey. However, this is not something you should embark on lightly. There are many pitfalls in survey design and survey administration, and conducting a survey can take a great deal of time. If you are contemplating conducting your own survey, you should consult a textbook on survey design.

Time Series Data and Panel Data

This book has focused on cross-sectional data. As discussed in Chapter 1, a cross-sectional data set consists of observations on n "entities" (individuals, firms, countries, etc.); typically these observations are recorded at a given point in time (for example, the same year).

Chapter 1 also described two other types of data sets, time series data and panel data. Time series data consist of observations on a single entity collected at T different points in time. For example, the price of a share of stock in a specific firm might be collected on the final day of each month for ten years, for a total of $T = 120$ observations. Panel data has both a time series and a cross-sectional dimension: in panel data, there are observations on n entities, where data for each entity is recorded over T time periods. For example, a data set consisting of the annual earnings of $n = 350$ individuals for each of $T = 8$ consecutive years would be a panel data set.

The techniques of regression analysis developed in this book for cross-sectional data can be applied to time series data and panel data; however, those

methods require some extensions and modifications. It is beyond the scope of this brief edition to go into those extensions in detail. However, if you have mastered the material in this book, it is not a big step to learn the necessary modifications to handle panel data and time series data. If you are interested in learning more about such data sets, see Chapter 10 (panel data) or Chapters 14 and 15 (time series data) in the full edition of this book.

Preparing the Data for Regression Analysis

A practical problem is that the data you collect needs to be prepared in an electronic form suitable for regression software. Often data come in text (ASCII) form, in which case it can be convenient to read or to paste the data into a spreadsheet in which columns denote the variable and rows denote the observation. The spreadsheet can then be saved in a format (such as comma separated value) that can be read by your statistical software.

10.3   Conducting Your Regression Analysis

Students who have worked through the empirical exercises in this textbook will be familiar with the steps involved in conducting a regression analysis using a given data set. These steps involve familiarizing yourself with the data, estimating one or more base specifications, thinking through potential nonlinearities, conducting a sensitivity analysis by estimating alternative specifications, and assessing (and addressing when possible) potential threats to the internal validity of the study.

The starting point for your empirical analysis is getting acquainted with the data. Plot the data, using histograms and/or scatterplots. Are there big outliers, and if so are those observations accurately recorded or are they typographical or data manipulation errors? Typographical or computer errors should be corrected in your master data spreadsheet. Once you are confident that the data are free from such errors, you can turn to specific relations. Are the units of the data the ones you expected, and are they the ones you want to use? Do the relations you see in the scatterplots make sense? Do relationships look linear, or do they look nonlinear?

Once you are familiar with your data set, you can begin your regression analysis. This is the point at which all the preparatory work you have done thinking through your study begins to pay off. Because you have thought hard about your problem and the data, you should already have in mind a base specification, along with some possible alternative specifications. The process of determining a base specification and alternative specifications is discussed in Section 7.5, and the

guidelines for whether to include a variable in a regression are given in Key Concept 9.2. Some of the alternative specifications might investigate possible nonlinearities, as illustrated in the analysis in Section 8.4 of the California test score data.

After you have some regression results, it is useful to go through the checklist of threats to internal validity in Key Concept 9.7. Are there arguably important threats to the internal validity of your study? If so, can you address them using multiple regression analysis of your data?

At this point, it is useful to share your findings with a classmate or instructor. The process of explaining what you have done and what you have found will help you think through any shortcomings of the analysis—your classmate or instructor can help with this too—and this in turn will point to additional specifications and additional sensitivity analysis to undertake. In this way, conducting an empirical analysis is a process that goes through multiple iterations.

10.4 Writing Up Your Results

Successful papers describing an empirical study often have five sections.

1. *Introduction.* The introduction succinctly states the problem you are interested in, briefly describes your data and the method of analysis, and summarizes your main conclusions.

2. *Discussion of Relevant Literature and Economic Theory.* This section describes closely related previous studies on your topic and summarizes any relevant economic theory. The length of this section depends on the scope of the paper; for a senior thesis, this section, especially the literature review, might be lengthy, but for a term paper this section might be short.

3. *Data Description.* This section provides the details of the data sources, any transformations you have done to the data (for example, changing the units of some variables), gives a table of summary statistics (means and standard deviations) of the variables, and provides scatterplots and/or other relevant plots of the data. If there are outliers other than those arising from corrected typographical or computer errors, this is the place to point them out.

4. *Empirical Results.* This section provides the main empirical results in the paper. Conventionally, regression results are presented in tabular form, with footnotes clearly explaining the entries; Tables 7.1, 8.2, 8.3, 9.1, and 9.2 can serve as templates. The initial table of results should present the main results; sensitivity analysis using alternative specifications can be presented in additional columns in that table or in subsequent tables. The text should provide

a careful discussion of the results, including assessments both of statistical significance and of economic significance, that is, the magnitude of the estimated relations in a real-world sense. Present "full disclosure" results: report results that you consider to be an honest and complete summary of what the data say concerning your question of interest, including results that raise doubts about or suggest limitations of your interpretation.

The empirical analyses of the test score data in Sections 7.6 and 8.4 provide examples of discussions of base and alternative specifications. This section of your paper should also contain a discussion of the potential threats to the validity of your analysis. Key Concept 9.7 provides a list of five potential threats to internal validity of regression studies using observational data. Some of those threats might not be relevant to your study, and this section should focus on the most salient threats. All empirical analyses have limitations, and it is important to provide a concise statement of what you consider to be the most substantial limitations of your analysis.

5. *Summary and Discussion.* This section summarizes your main empirical findings and discusses their implications for the original question of interest.

The guidelines in this chapter for conducting an empirical study are summarized in Key Concept 10.1.

GUIDELINES FOR CONDUCTING AN EMPIRICAL ECONOMIC STUDY

The following guidelines can help you be efficient when you undertake an empirical study.

1. Choose a topic that interests you personally.

2. Develop a few narrow questions and think through an empirical analysis that would answer them. For each question, what base specification would you use? What is the key regressor and what is the regression coefficient of interest? What might be important sources of omitted variable bias?

3. Learn about relevant data sets by consulting a data librarian or the Web (see **www.aw-bc.com/stock_watson**).

4. Narrow your question further. Will your candidate data set plausibly help you to estimate the parameter of interest?

5. Format the data so that they can be read into your statistical software.

6. Compute summary statistics, scatterplots, and other data diagnostics. Correct or discard outliers arising from data entry or computer errors.

7. Conduct your regression analysis:
   a. Estimate your base regression.
   b. Estimate alternative specifications that address potential nonlinearity and omitted variable bias.
   c. Assess the threats to the internal validity of your analysis using the list in Key Concept 9.7.
   d. Explain to a classmate or instructor what you have done, why you have done it, and what you have found.
   e. Repeat steps a–d until you are satisfied that you have addressed, as best you can, the main threats to the internal validity of your analysis.

8. Write up your results using the outline in Section 10.4. Discuss the statistical and economic (real-world) significance of your results, report "full disclosure" results, and discuss any remaining threats to internal and external validity.