

Data

Economics is an empirically-oriented research discipline. Its models and analyses are intended to be applicable to real-world situations and especially real-world data. Empirical research in economics is made easier by the fact that so many of the variables of interest—prices of goods, quantities transacted, amounts earned, sums lent, etc.—are not only quantifiable but also potentially observable. Still, the relative ease with which these variables *can be* measured, sometimes leads economists to take for granted the methods by which the variables are *actually* measured. As data are the fundamental building blocks in our empirical analyses, it is important that we carefully consider their sources, properties, and uses.

The word “data” is used ubiquitously inside and outside of economics. Despite our exposure to the word, it still helps to review what the word means. “Data” is the plural of the word “datum,” which means a piece of information, such as a fact, measurement, or observation. Thus, the word “data” describes multiple pieces or a collection of information.¹ They can be analyzed by themselves, such as when we examine trends in national output or personal incomes, or joined together to form more complex pieces of information, such as when we combine an income series with a price index to construct real, or inflation-adjusted, income measures. We reserve the term “economic data” to refer to pieces of information that describe different aspects of economic processes, including economic outcomes and constraints.

The data that economists use in their empirical research do not just magically spring, fully-formed, out of the nowhere (although students might be excused for suspecting this, given the small amount of attention given to the subject in most economics textbooks). All data have some source and are created by some process. To use and interpret data properly, we need to know how the information was generated and collected. This chapter discusses these processes as well as general properties of data. It begins by describing the ideal use of data and then moves on to practical concerns of what data actually represent, how they are recorded, how they are organized, and how they are manipulated before entering an analysis.

A motivation behind writing this chapter is that students are often surprised and occasionally frustrated at the time and effort that it takes to identify and prepare data for an analysis. Their project planning gives a lot of thought to the theoretical models and statistical models that they might use but relatively little thought to how they will get the proper data for those analyses. New researchers invariably discover that data are seldom in a form that can be immediately analyzed in any serious way and that data preparation consumes the lion’s share of the time of an empirical investigation. The data tasks of a project are made easier by some advance knowledge, by planning and thought, and by practice.

Moving between theory and data

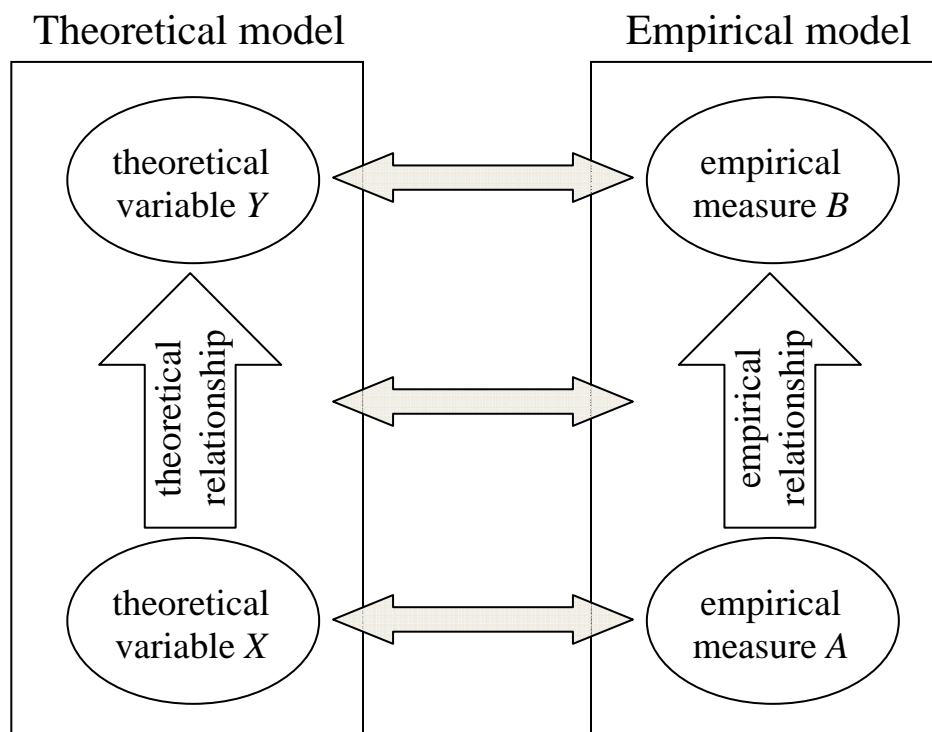
In the previous chapter, we discussed how theoretical models are central to social science

¹This is a good place to interrupt with the grammatical point that the word “data” is plural and, hence, should always be matched with plural forms of modifiers and verbs. For example, we write “these data” rather than “this data” and “the data are” rather than “the data is.”

research. Recall that theoretical models help us to understand how and why people—either alone or in groups such as households, firms, communities, or countries—behave the way they do. They also help us to understand how people might behave under different or yet-to-be-observed circumstances. Recall also that theoretical models are written in terms of a set of variables and that they describe relationships between these variables. To the extent that these variables and relationships have counterparts that can be observed among actual people, the theoretical models can be “taken to the data” to be tested, calibrated, or used to make predictions.

It is helpful to bear in mind the distinction between theories, which describe things as they might exist or are hypothesized to exist, and empirical observations, which describe things as they do exist. To emphasize this distinction, we use one set of terms for the theoretical elements in our analyses and another set for the empirical elements. Specifically, we refer to the elements from our theories as *theoretical constructs*; these include the *theoretical variables* and *theoretical relationships* from our conceptual models. In principle, each of these theoretical constructs has an empirical counterpart, or *empirical construct*, that we might be able to observe and record. The empirical counterparts to theoretical variables and relationships are called *measures* and *empirical relationships*, respectively. We would describe the whole group of empirical constructs that are associated with a particular theoretical model as the *empirical model*, or the empirical representation of a model.

Developing an empirical representation of a theoretical model



The diagram on the preceding page illustrates the process of developing empirical counterparts to a simple theoretical model with two theoretical variables, X and Y , and a single theoretical relationship. The theoretical model, with its three constructs, is shown in the box on the left side of the diagram. In the box, the two variables appear as ovals, while the relationship appears as a thick vertical arrow. An empirical representation of this theoretical model is shown on the right side of the diagram. Notice that each of the theoretical constructs has an empirical counterpart. There is a measure A that corresponds to the theoretical variable X , a measure B that corresponds to the theoretical variable Y , and an empirical relationship that corresponds to the theoretical relationship. In the diagram, the correspondence between each of the theoretical and empirical constructs is indicated by the horizontal arrows.

An empirical analysis uses all of the elements—the measures and the relationships—from the right side of the diagram. In this chapter, however, we will focus on the measures. In subsequent chapters, we will discuss ways of investigating the relationships between the measures.

Demand example. As a concrete example of how to develop an empirical representation of a model, consider a simple theoretical demand model. Recall that an economic demand function relates changes in the quantity of a good demanded to changes in the price of the good, holding other characteristics such as consumers' incomes and tastes constant. To describe these elements in terms of the left side of the diagram, we would consider the amount of the good to be the outcome, or Y , variable, the price of the good to be the X variable, and the demand function itself to be the theoretical relationship.

Suppose that we wanted to apply this model to examine household energy demand. An empirical analysis might use information from the State Energy Data System (SEDS) of the U.S. Energy Information Administration (EIA) on the British Thermal Units (BTUs) of total per-capita residential consumption of energy in each state in a particular year as an empirical measure of the “goods” outcome and information from the same source on the nominal dollars spent per BTU in each state as an empirical measure of the price variable. Residential consumption and price data for 2005 are shown in the table below. With these measures, a simple correlation or ordinary least squares regression could be estimated to establish the empirical relationship.

2005 Residential Energy Consumption and Prices

State	Good: Per-capita residential energy consumption (trillion BTUs)	Price: Dollars per million BTU
Alabama	466.8	20.84
Alaska	1193.9	14.75
Arizona	248.6	22.50
Arkansas	409.5	19.74
California	232.3	20.80
Colorado	305.1	15.20
Connecticut	258.2	21.58
Delaware	371.9	20.17
Dist. of Columbia	327.0	18.83
Florida	257.3	27.59
Georgia	348.4	21.36
Hawaii	263.0	57.01

Idaho	352.9	14.89
Illinois	324.0	14.96
Indiana	464.2	16.26
Iowa	415.4	17.65
Kansas	376.4	16.37
Kentucky	472.3	16.70
Louisiana	803.7	21.81
Maine	367.6	19.88
Maryland	279.1	19.20
Massachusetts	242.9	20.78
Michigan	313.3	14.39
Minnesota	362.2	15.82
Mississippi	407.6	21.69
Missouri	330.8	16.59
Montana	448.2	15.87
Nebraska	373.4	15.16
Nevada	302.1	20.63
New Hampshire	257.4	21.02
New Jersey	315.2	17.16
New Mexico	352.3	17.10
New York	217.0	20.92
North Carolina	314.8	21.61
North Dakota	648.1	15.64
Ohio	356.2	16.82
Oklahoma	438.6	17.64
Oregon	301.8	17.25
Pennsylvania	327.5	18.79
Rhode Island	213.3	19.37
South Carolina	398.2	22.35
South Dakota	350.6	16.97
Tennessee	390.5	17.99
Texas	506.0	25.62
Utah	302.1	13.19
Vermont	269.5	20.99
Virginia	345.4	19.75
Washington	328.3	15.95
West Virginia	439.7	15.56
Wisconsin	336.0	17.43
Wyoming	911.9	15.10

Sources: Energy Information Administration. *State Energy Consumption Estimates 1960 to 2005* <http://www.eia.doe.gov/emeu/states/sep_use/notes/use_print2005.pdf> and *State Energy Price and Expenditure Estimates 1970 to 2005* <http://www.eia.doe.gov/emeu/states/sep_prices/notes/pr_print2005.pdf>, accessed Aug. 23, 2008.

The preceding discussion and example imply a one-way relationship from theory to data, but this need not be the case. Research can move in the other direction—from an analysis of data to the development of hypotheses and conceptual models. Empirical investigations that are used to motivate theoretical models are called *exploratory analyses*, while empirical

investigations that test or confirm hypotheses are called *confirmatory analyses*. In practice, the distinction between these types of analyses can be blurred. Research often operates in both directions with a given set of hypotheses motivating a series of empirical analyses and the empirical results then leading to revisions and refinements in those initial hypotheses.

The figure also gives us an opportunity to define “economic” data and measures more formally. Economic data are empirical measures that correspond to theoretical economic constructs. These may be measures of the outcomes of economic processes, such as employment levels, quantities demand or supplied, and the like. They may also be measures of the constraint variables from consumers’ utility maximization or firms’ profit maximization problems. These would include measures of the things that enter budget constraints, like prices and incomes; things that enter production constraints, like technological characteristics; and things that constrain markets and behaviors themselves, like laws, regulations, and tax systems. Economic data can also extend to measures of the objectives that people or firms have and of the information or expectations that they hold.

Finally, the figure provides some implicit guidance about how to manage and focus our data gathering. When we go out looking for data, it is not unusual to come across many more interesting measures than we can possibly use. How do we decide which measures to include in our analyses and which measures to leave aside? A good initial rule of thumb is to keep the measures that relate somehow to our conceptual model, that is, keep the measures that we can interpret within the framework of the model.

Are you collecting data yourself or using somebody else’s data?

All data describe *something* and are collected *somehow* by *someone*, helpful things in to keep in mind as we consider data that might be appropriate for our specific research. Let us start with the *someone* doing the collecting. Are you going to collect the data or will you rely on somebody else’s efforts?

Social scientists distinguish between *primary* and *secondary* data. Primary data are pieces of information that are directly observed or recorded by you as either an individual researcher or as a member of a research team, while secondary data are pieces of information that have been observed or recorded by someone else. At one level, the distinction is artificial because all data are initially primary data. From this perspective, secondary data are just someone else’s primary data. Indeed, if all data were perfectly suited to answering our research questions, there would be no practical import to these designations beyond acknowledging the efforts involved in obtaining the data. Unfortunately, for a variety of reasons, data are often imperfect, and the designations do matter.

Primary data collectors will obviously be more familiar than secondary users with the precise methods used to gather and record the data. From this, primary researchers will also usually be more aware of the flaws and limitations of their data. All empirical researchers need to understand the properties of the data that they examine; the use of secondary data does not absolve the researcher of this responsibility. While secondary users avoid the effort of actually collecting the data, they often need to devote extra effort understanding the strengths and weaknesses of the data. Even if they share the same appreciation for the data, primary and secondary researchers may still differ in the research questions that they trying to answer. To the

extent that such questions guide the selection of items and the collection methodology, the data may be less suitable in their secondary uses than their primary use. Finally, the complexities of processing and working with data introduce opportunities for error. As a secondary analysis involves another round of data handling, there is an added chance of a processing mistake sneaking in.²

Most empirical research by economics students involves analyses of secondary data, rather than primary data collection. In the discussion that follows, we will talk about many general properties of data but will give examples mostly in terms of secondary data.

What are the data describing?

The next things to take up are the characteristics of the processes that the data are describing.

Observational, experimental, and artificial processes. Most data that economists study are *observational* in the sense that they describe real-world behavior that can be observed, at least in principle. Thus, these data describe actual people engaging in actual activities in actual settings without the researcher intervening in those activities or settings. For instance, observational data might describe what goods people bought last month, how much workers earned last year, or how many children couples had during their marriages. The energy consumption and price data that we have already considered are observational.

We can contrast these with *experimental* data, which describe people's behavior in situations where researchers do intervene. We think of experiments as taking place in laboratories, and indeed, economists do conduct these types of experiments (see, e.g., Smith 1962, Plott & Smith 2008). Researchers have also conducted social experiments. Social experiments occur in real world settings and involve researchers intervening in things like public services that people might receive or the shopping choices that they might face. For example, in the late 1960s and early 1970s, social scientists conducted income maintenance experiments in New Jersey, Pennsylvania, rural Iowa and North Carolina, and the cities of Gary, Denver, and Seattle. In these experiments, some welfare families were randomly selected to receive special welfare benefits, while other continued to receive regular payments. The experiments were designed to evaluate whether a negative income tax might induce people to work more and leave welfare sooner (Munnell 1987).

Whether they take place in a laboratory, a welfare or employment office, or even in an entire state or community, experiments all involve researchers manipulating people's environments independent of those people's control. Experimental data are easier to analyze statistically, because the intervention or manipulation that is being studied is not confounded with other characteristics of the subjects. This is less likely to be the case in observational settings where people choose the programs that they participate in or the communities that they

²To be fair, a careful secondary analysis can sometimes spot weaknesses in the data and either alert researchers to these problems or even make corrections. An example of a study calling attention to problems is the 1986 analysis by Lillard et al. that described the Census Bureau's procedures for assigning values for employment and earnings in its Current Population Survey when survey subjects were unwilling or unable to provide these data themselves.

live in. At the same time, experiments take additional effort to perform. They also occur under limited and sometimes very artificial circumstances, making it difficult to generalize the results into other situations.

Yet another source of data is wholly artificial processes that researchers themselves create. Artificial data are created by computer programs that either replicate or extrapolate from existing data or that create entirely new data from processes like random number generators. Artificial data are useful because the researcher controls their creation and thus knows their exact properties. Artificial data are used to test the properties of statistical routines, for instance, to see if the routines can recover the true parameters of the underlying processes. They are also used to approximate sampling distributions of statistical estimates (probability distributions of the estimates). Finally, they are used in simulation analyses, where researchers create artificial populations that behave and interact according to a predefined set of rules.

Individual or aggregate observations. The processes that economists study can also be categorized in terms of the number of units—people, firms, institutions, etc.—that contribute to each observation. Measures that describe individual units are referred to as *micro*, or individual, data, while measures that come from aggregations of individuals are referred to as *macro*, or aggregate, data. The state-level energy consumption and price data that we discussed earlier are macro data, because they represent the behavior of all of the households in a state. Alternatively, we could have examined micro data from the Residential Energy Consumption Survey, which is conducted by the EIA about every four years and interviews households about their energy use.

Micro and macro data each have their advantages. Macro data are easier to work with and assemble than micro data. Macro data typically have fewer observations and are usually already tabulated, so they can be readily entered into a spreadsheet. State- and country-level data are especially convenient. As we have already seen, the EIA produces state-level estimates of energy consumption, expenditures, and prices. The U.S. Census Bureau produces a compendium called the *State and Metropolitan Area Data Book* with an enormous number of state-level statistics. The U.S. Bureau of Economic Analysis produces the *Regional Economic Information System* with state-level data on incomes, output, and employment. The U.S. Department of Education produces its *Digest of Education Statistics* with many tables of state-level educational information. Numerous other agencies produce state-by-state tables.

While micro data typically require more work and processing than macro data, it is easier to use them to condition on particular characteristics. For example, if we were interested in the differences in energy use between home owners and renters or between couples and single adults, we could directly examine each of these different types of residential consumers using the Residential Energy Consumption Survey. We could not examine this using the tabulated figures from the SEDS, because the SEDS data do not make these distinctions. In general, micro data give you more flexibility to organize or group the data the way that you want.

Static and dynamic measures. Another way to categorize data is by whether they represent net outcomes at a point in time or changes in those outcomes over time. The first type of measure is *static*, while the second type is *dynamic*. Economists generally think of static measures as representing equilibrium outcomes, such as the quantities traded in a market or the levels of output produced by a firm or an economy. These measures are also referred to as stock measures, levels, or incidence measures. Dynamic measures can represent equilibrium outcomes; however, they are also used to describe adjustments in behavior toward equilibrium.

Dynamic measures are often expressed in terms of changes, flows, or rates.

Static outcomes are generally easier to measure than dynamic outcomes, because a static measure requires a single observation, while a dynamic measure sometimes requires repeated observations (e.g., before and after observations to measure changes). This is not always the case, however. Some types of flow measures come from single observations. Good examples are the number of people entering unemployment by filing initial unemployment insurance claims, the number of young people transitioning out of school by graduating, the number of people coming onto the welfare rolls, and the amount of funds flowing into financial accounts through deposits.

One analytical advantage of change measures is that they net (difference) out any characteristics of the underlying observational units that are constant over time. In an analysis, this can reduce spurious correlations from confounding characteristics. Consider our energy demand example. Even if we think that a static demand model provides a sound basis for analyzing residential energy consumption, we might still wish to examine change data. In the theoretical model, the amounts demanded depend not only on prices but also on tastes and preferences, which can be hard to measure. If we believe that people's tastes and preferences are fixed over short periods of time, the use of change data would allow us to control for these characteristics by netting them out and free us from having to actually measure them.

Time series, cross-section, and panel observations. An issue related to the level versus change distinction is the organization of the data across time and space. A *time series* of data follows one observational unit (person, firm, country, etc.) across multiple time periods. For example, a time series of a country's macroeconomic performance might include quarterly or annual measures of its gross domestic product, inflation rate, money supply growth, and fiscal balance. Time series are most often used in macroeconomic analyses. The resulting data sets tend to be small and manageable. However, the observations also tend to be closely related from one period to the next (serially correlated), reducing the effective variation in the data set. Because the observational units are followed over time, it is straightforward to construct change measures.

In contrast to a time series with its single observational unit and multiple time periods, a *cross-section* of data contains information for multiple observational units within a single time period. The state energy data that we selected from the SEDS are a cross-section. They have consumption and price measures for 50 states and the District of Columbia for a single year. Other examples of cross-section data sets are polls and surveys of individuals. Some cross-section data sets can be very large. Government surveys commonly include thousands of subjects. To analyze this many observations, a researcher would have to turn to specialized statistical software, such as SAS, SPSS, or Stata, instead of a spreadsheet package.

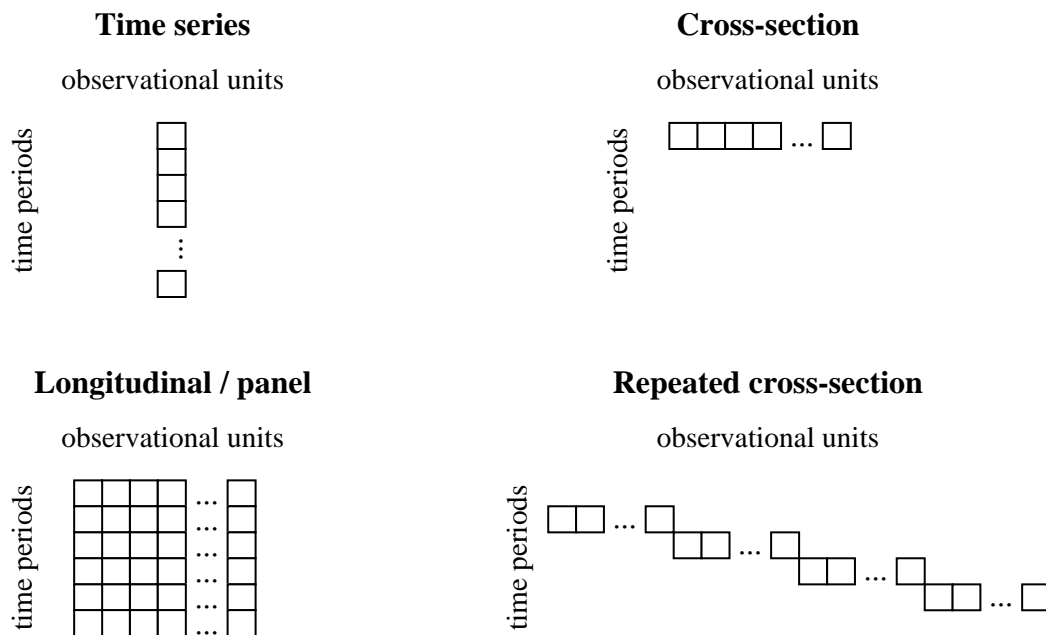
Two other data arrangements combine features of time series and cross-section data sets. The first of these is a longitudinal, or panel, data set in which a cross-section of subjects is followed over time. The underlying source for our energy data, the SEDS, is actually a panel. It contains annual consumption data for all states extending back to 1960 and annual price and expenditure information extending back to 1970. In the example analysis, we selected a single cross-section, but we could have also used the entire panel. Besides the SEDS, some other examples of longitudinal data sets are the Panel Study of Income Dynamics, a survey that has followed members of 5,000 U.S. households since 1968; the European Community Household

Panel, a yearly survey that has asked common questions of households across 15 European countries since 1994; Compustat, an annual data set with financial information for public firms whose stocks are traded on major U.S. exchanges, and the National Center for Education Statistics “Common Core of Data,” a data set with financial and institutional information for public schools and school districts across the U.S. Longitudinal data sets are especially valuable for examining changes in outcomes across individuals. The downsides of these data are that they are more complicated to use than cross-section data; for example, an analyst needs to be able to link the information for each observational unit over time. Longitudinal data sometimes also suffer from special problems, such as attrition (subjects leaving the sample over time).

The second “combination” data arrangement is the repeated cross-section in which different sets of subjects are studied over time. The Residential Energy Consumption Survey is a repeated cross-section. It asks similar questions of different samples of individuals every four or so years. Because the questions are similar or identical over time, the answers can be compared, even though the subjects themselves change. Repeated cross-section data support the analysis of changes over time in group outcomes but not changes in individual outcomes. Other repeated cross-section data sets include the General Social Survey, an annual survey of attitudes and opinions in the U.S.; the Census Bureau’s annual economic surveys (e.g., the Survey of Manufacturers), which gather information of firms in different industries; and the American Community Survey, a large economic and demographic survey fielded annually in all communities in the U.S.

The organization of time series, cross-section, panel, and repeated cross-section data are illustrated in the figure below.

Organization of data



Methods of obtaining data

Economists rely on two primary methods for obtaining data on subjects: questionnaire surveys and administrative systems. As the name suggests, questionnaire surveys ask people or representatives of organizations questions about the behaviors and characteristics that researchers are interested in. Questionnaires can be fielded in a number of ways. An interviewer can ask the questions, either in person or over the phone, and record the subject's responses. Using an interviewer helps to boost response rates, improve the consistency of responses, and reduce misunderstandings regarding the questions. Interviewers are especially helpful when the questionnaire is complex, such as when it asks lots of conditional questions. In addition, the interviewer can provide independent assessments of the subject and his or her co-operation.

Questionnaires can also be mailed to subjects. This is less expensive than using an interviewer. However, response rates for mailed questionnaires tend to be lower. Also, the researcher is depending on the subject to read and comprehend the questions (and to record answers correctly). Successful mailed questionnaires tend to be relatively simple. One advantage of mailed questionnaires is that subjects may feel more comfortable answering personal questions on paper than in an interview. Because of this, researchers sometimes combine survey modes by using interviewers for most of their questions but allowing subjects to answer sensitive questions on paper.

Questionnaires can also be fielded over the internet or on computers. Computers are an improvement over paper-and-pencil surveys because they can be programmed to provide prompts and explanations for questions and to screen for unreasonable responses. A disadvantage, though, with web-based systems is that researchers might miss poor households and older households that lack or are uncomfortable with computers. Computers can also help interviewers; increasingly, the standard in survey research is to use CAPI, or computer-assisted personal interviewing.

Whether interviewers, mailed questionnaires, or web questionnaires are used, there is an additional issue of whom to interview. If the researcher wants to know information about the subject herself, the survey can be directed toward that person. Sometimes, however, we also want to know about other members of the subjects' household, other employees at a firm, or neighbors in a community. One inexpensive way to collect information about other people is to have a single respondent report results for them, that is, to have the respondent provide *proxy reports* on other people. The shortcoming of this approach is that the information from proxy reports is usually less accurate than from self-reports. Of course, obtaining reports directly from other people in a household, firm, or community increases the costs of a survey.

Administrative systems capture data about people as they interact with institutions. For example, tax data are recorded by the Internal Revenue Service as people pay their taxes and complete their tax forms. Earnings data are recorded by state Unemployment Insurance (UI) systems as these systems track people's eligibility for UI benefits. Public assistance programs, such as Temporary Assistance for Needy Families, Food Stamps, and Medicaid, maintain records on clients' applications, continued eligibility, and program use. Administrative systems contain information on all of the participants in a given program. They also overcome some of the recall and reporting problems associated with questionnaires, especially when it comes to detailed program data like benefits and participation dates. The main drawback to using these systems is that they only begin tracking people when those people come onto a program and stop

tracking them after they leave. The data also have to be cleaned by the administering agencies to remove identifying information.

Representativeness. One key concern regarding data collection is that we want the information to be representative of the group being studied and possibly of other groups. Representativeness can be compromised many different ways. For example, in survey research, busy people tend to be harder to reach than less active people. Similarly, poor people, who sometimes lack telephones or move frequently, can be harder to interview than more affluent people. Depending on the design of the survey, cooperation among different groups might vary, leading to an unrepresentative response pattern. Administrative data cover the universe of people in a particular program but can be unrepresentative of more general populations, such as the group of people who are eligible or potentially eligible for a program.

When we begin working with a data set, we usually compare characteristics of the subjects to known population averages to determine the data set's representativeness. If we were working with data on peoples, we would compare the gender, ages, ethnicities, and other characteristics to distributions for the entire population. If we were working with data on firms, we might compare the sizes or industrial composition to published totals. A data set in which 90 percent of the respondents were male, or in which all of the firms had 10,000 or more employees would be representative of certain populations but not the entire population.

Reliability and validity of measures. Another obvious concern involves whether the measures in our data sets are accurately capturing what they are supposed to measure. Suppose that there is an underlying characteristic that we want to measure through responses to questions or some other approach. *Reliability* refers to repeated or related questions about the characteristic yielding similar responses. For example, if we began a survey with a question about how much a household spent last month on electricity and later asked how much the household's last monthly electricity bill was, we would expect similar answers. *Validity* refers to our measures actually capturing the characteristic they are supposed to. In the above example, we could compare the household's answers about its electricity spending to actual records on payments from a utility or to the household's bank statements. We would like survey questions to be reliable (have little noise) and be valid (capture the underlying information). Most major surveys go through extensive pre-testing to improve the reliability and validity of their measures. Nevertheless, problems can remain with some measures.

For example, people provide better information about dates in their life histories if they are initially prompted to think about a few significant dates, such as their marriage dates or first child's birth date. People provide better income and expenditure information if they are given clear dates and if they are reminded about various sources of income and types of expenditures. People can provide more accurate descriptions of activities that have taken place recently than events that took place some time ago (recall bias). Finally, people will sometimes skew their responses to avoid potential embarrassment (social desirability bias).

It is always important to review the survey instrument or administrative forms to see exactly how questions were asked or information was obtained. In some cases, reliability or internal consistency scores are available for measures. Similarly, studies of validity are sometimes available for measures. Researchers should also perform consistency checks of their own. The first thing to check is whether all of our measures are within their expected ranges. For example, hours of work should be non-negative, and answers to categorical questions should

fall within the available categories. The next thing is to do perform conditional reasonableness checks. For example in a data set with earnings and employment information, we would check whether earnings are positive for workers and only for workers. Lastly, we would check whether the data exhibit standard correlation patterns, such as earnings being positively associated with schooling and experience.

Preparing an analysis data set

Once the data are in hand, we still usually have a substantial amount of work to do. One task is to select the observations that will actually enter our analyses. For example, we might drop some observations that appear to be outliers or that are missing subsets of information. Alternatively, we might select observations because they are functionally relevant to our analyses. For example, in an analysis of countries' economic performance, we might restrict the data set to just developing or just developed countries. In an analysis of individual employment outcomes, we might restrict the analysis to adults who have completed school but not yet reached retirement age.

The other task is to construct the actual measures that we will analyze. If we were examining the energy data over several years, we would most likely want to adjust the price data for inflation. If we were working with data on people's pay rates, we would want to form a consistent hourly, weekly, or annual rate. From categorical data on people's educational attainment, we might need to form indicators for high school or college completion. If longitudinal data are available, we may wish to difference level measures across periods to form change measures. In general, some manipulation is required to convert the data as reported into a form that corresponds with our theoretical constructs.

Some suggestions for finding data

As you consider possible data sources for your own research, the first place to look is in the articles, reports, and books that you have been reading on your research question. Research publications list their sources, including their data sources. The beauty of using your literature review to uncover data is that you will already know that the data are relevant. Also, the major strengths and flaws of the data will have been discussed. For course-related research, following up on the sources used in published research is a great way to go. For dissertation or potentially publishable research, existing, frequently-used data sources are also valuable, though you might also want to search out newer and fresher sources of information.

Another way to start the data search is to work through one of the many available compendia of data. For the U.S., the best compendium is the *Statistical Abstract of the United States*, which is published annually by the Census Bureau (and is available on-line). The *Statistical Abstract* contains hundreds of demographic, economic, and political tabulations, so it is a great immediate resource in its own right. As importantly, the tables in the *Abstract* contain notes referencing the original sources of the data. The *Abstract* also has an appendix with a "Guide to Sources," so it provides a jumping off place for further searches. In addition to the *Statistical Abstract*, the Census Bureau also publishes the *State and Metropolitan Area Data Book* and the *County and City Data Book*, which contain many of the same measures but tabulated for states, cities, and counties. Most states also prepare statistical abstracts and other

tabulations; these are referenced in the Census Bureau's abstract. Along the same lines, the statistical agencies of other governments also prepare abstracts. The *Statistical Abstract of the United States* references these and contains many international statistics.

Before we leave our discussion of the Census Bureau, it is also important to mention that it is the country's premier agency for conducting survey research. We think of the Bureau fielding the decennial population census, which is used for political reapportionment, but the organization also conducts many other censuses and surveys. A *census* is an exhaustive enumeration of some population. The Census Bureau conducts an *Economic Census* of all businesses in the U.S. every five years; the last was conducted in 2007. It also conducts a *Census of Governments*, covering all governments from the federal level down to the local level, on the same time schedule. Micro data from all of these censuses are available to researchers.

Two of the large population surveys conducted by the Census Bureau are the *Current Population Survey* and the *American Community Survey*. These contain a wealth of economic and demographic information about people, families, and households in the U.S. and are workhorse data sets for researchers. The Bureau also has a major longitudinal population survey, the *Survey of Income and Program Participation*, which interviews households every four months over a four-year period. The Bureau also conducts annual surveys of businesses and governments, such as the *Annual Survey of Manufacturers*, the *Annual Survey of Retail Trade*, and the *Annual Survey of Government Finances*. Again, micro data from all of these surveys are available to researchers.

Another handy compendium of economic figures for the U.S. is the statistical appendix to the annual *Economic Report of the President*, which is published by the Council of Economic Advisors early each year around the time that the President submits budget recommendations to the Congress. The statistical appendix has detailed income, output, employment, productivity, price, trade, money supply, and fiscal information going back over long periods of time. The tables also reference the source agencies for these data. The *Economic Report* provides "one-stop shopping" for many U.S. macroeconomic time series.

The other logical places to search for U.S. statistics are the governmental agencies themselves. All federal departments collect and report statistics and information related to their missions. For example, the Bureau of Labor Statistics collects employment, wage, productivity, and price information on behalf of the Department of Labor. The National Center for Education Statistics collects schooling data for the Department of Education. The National Center for Health Statistics maintains vital records (births and deaths), disease, and health information for the Department of Health and Human Services. As we have already discussed, the Energy Information Administration collects data on behalf of the Department of Energy. The FedStats web-site (<http://www.fedstats.gov>) can lead you to these and other sources.

For international data, researchers can go to the statistics portal of the Organization for Economic Co-operation and Development (OECD) and its *SourceOECD* database, which requires a subscription. For statistics specific to the European Union, researchers can go to the Eurostat site of the European Commission. More generally, researchers can also visit the United Nations' *UNdata* statistical site at <http://data.un.org/>.

There are also two large data archives that can assist social science researchers. The world's largest archive is the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan. Its web address is <http://www.icpsr.org>. The ICPSR

warehouses thousands of data sets along with documentation. The data sets include major governmental surveys as well as studies by private organizations, such as polls conducted by news organizations, and by individual researchers. The ICPSR web-site has excellent search tools to help researchers find and compare data sets.

Economists will find the data sets in the Data Collection at the National Bureau of Economic Research (NBER) to be especially useful. The web address for this site is <http://www.nber.org/data/>. The data sets were contributed by NBER researchers. They include business cycle dates (the NBER's Business Cycle Dating Committee is the official source for these dates); links to the Penn-World tables, a longitudinal database of international output, price, and finance statistics; monthly Current Population Surveys and supplements since the 1960s, including a harmonized file with consistent measures from 1964 to 1992; and many other data sets.

Obviously, no list of potential data sources will be exhaustive. However, the above list should give young researchers a good start on their projects.

References

- Lillard, Lee, James P. Smith, and Finis Welch. “What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation.” *Journal of Political Economy* 94:3 part 1 (June 1986), 489-506.
- Munnell, Alicia H., ed. *Lessons from the Income Maintenance Experiments: Proceedings of a Conference Held in September 1986*. Boston: Federal Reserve Bank of Boston, 1987.
- Plott, Charles, and Vernon L. Smith, eds. *Handbook of Experimental Economics Results*. Amsterdam: North-Holland, 2008.
- Smith, Vernon L. “An Experimental Study of Competitive Market Behavior.” *Journal of Political Economy* 70:2 (April 1962), 111-137.
- U.S. Census Bureau. *Statistical Abstract of the United States: 2008*. Washington, DC: U.S. Government Printing Office, 2008.