

Sources of Data in Economics

- General description – economic data refer to pieces of information or collections of information that describe different aspects of economic processes
 - economic outcomes
 - end results of economic processes
 - usually examined as *endogenous* variables (variables determined by or inside the model)
 - examples include employment levels, price levels, quantities demanded, quantities supplied
 - economic constraints
 - recall that we describe economic decision-making in the context of making the most out of the resources available, i.e., maximizing subject to constraints
 - budget constraint information includes prices, incomes, wage rates, interest rates
 - institutional constraint information includes laws, regulations, practices, tax systems
 - technological constraints include production processes and time constraints
 - information constraints
 - constraints can be *exogenous* (determined outside or imposed on the model); however, they can also be endogenous (e.g., price levels are a constraint on individual behavior but also a market outcome)
 - economic objectives

- this information describes what is being maximized or how a decision is made
- individual objectives include maximizing preferences or wealth
- firm objectives are to maximize profits
- government objectives might be to maximize social welfare, stabilize prices, or maximize revenues
- objectives are typically assumed to be fixed or exogenous
- sometimes the objectives are not clear and thus become the focus of research (e.g., what are the goals of a non-profit organization like a university or a symphony)
- demographic information
 - information about births, deaths, family formation, population levels, age distributions, racial composition, etc.
 - sometimes examined as economic outcomes
 - sometimes considered as indicators for preferences (e.g., cultural or age differences) or constraints (e.g., potential labor supply, marriage opportunities)
 - for these and other “non-economic” data, we must be careful to specify how they fit into an economic model or contribute to an economic interpretation
- remainder of discussion will focus on other general categorizations of data

- Are the data measured at the micro or macro level? – do the data refer to individual decision-makers or an aggregation of decision-makers?
 - macro data are easier to work with
 - macro data have fewer observations; can typically be entered into a spreadsheet
 - with micro data, you often have to work directly with a large survey or administrative source
 - however, it is harder to answer certain questions with macro data (e.g., aggregation bias in demand and supply functions)

- Are the data quantitative or qualitative
 - quantitative data refer to measures of outcomes that can either be counted or mapped to the set of real numbers
 - characterizes most items that we consider to be pure economic data such as quantities transacted, prices charged, wages paid, employees hired, etc.
 - easily amenable to statistical analysis; that is, to use data from a sample (numerous cases) to make inferences about the world as a whole
 - focus is on testing numerical hypotheses
 - get more cases here
 - but have relatively little detail per case and little understanding of the individual cases
 - relatively easy to collect (just need to record a number); usually also collected in large quantities
 - qualitative data refer to other measures and information

- sometimes the measures can be *categorized* (mapped into a small number of classes)
 - examples include racial and ethnic measures or location measures
 - these measures can be treated using quantitative methods
- other measures are harder to categorize
 - examples include interview responses, writing passages, open-ended answers to questions
 - economists seldom work with these data
- one type of qualitative study that economists sometimes employ is the case study
 - use data from one or a few cases and examine these data in great detail; substitute depth for breadth
 - these data can come from just about anywhere and cover just about anything (companies, people, governments, countries, crises, etc.)
 - goals are
 - * to get a complete picture of the case, and
 - * to get data that address the predictions of the model
 - case studies are often used when situations aren't well understood or models aren't available; here the process works backwards;
 - * the data are collected
 - * theories are developed to explain patterns in the data

- other types of qualitative data that are sometimes collected or used by economists are
 - semi-structured interviews
 - open-ended questions
 - focus groups
- an important strain of qualitative research outside of economics is ethnographic research

- Are the data experimental or observational?
 - many statistical techniques assume that data is produced experimentally
 - in an experimental design, the experimenter
 - assigns the treatment to each case (treatments are beyond the control of the subjects)
 - can repeat the experiment
 - this does not characterize a lot of economic data
 - we don't assign a "depression" or "inflation" treatment to some countries or regions
 - we don't assign schooling levels to individuals
 - in some ways both of these examples involve subjects "selecting" their own treatments
 - we describe these data as observational
 - statistical analysis of observational data can be more complicated
 - this is not to say that there are not genuine economic experiments; it's just that most economic data is generated non-experimentally

- examples of experiments
 - small scale laboratory experiments are used to examine a variety of situations such as how people interact in simple games; these games are often very artificial and abstract
 - program experiments – make a new program available to some people but not to everyone and examine how outcomes differ across those who do and do not get into the program; used to examine welfare reform, job training, and adolescent interventions
 - social experiments – these are larger and more expensive; hence, they are not conducted often; best known are the income maintenance experiments from the 1960s and 1970s
- time series, cross-section or panel
 - descriptions usually applied to quantitative data
 - time series – one individual (or observational unit) followed across many time periods
 - examples: quarterly GDP, annual unemployment
 - disadvantages: typically small sample sizes; data may be highly aggregated; data generating process may change over time; observations may not be entirely independent from one time period to the next
 - cross-section – many individuals measured for one time period

- examples: political polls, consumer confidence surveys
- disadvantages: complicated to use (need to learn how to cope with non-response and invalid responses); may not have good measures of key economic variables like prices or local wage levels
- repeated cross-section – a series of cross-section samples across several time periods using different individuals in each period
 - examples: Current Population Survey
 - disadvantages: can be more complicated than simple cross-sections; may limit the number of variables that are collected
- longitudinal or panel data – follows a single cross-section of individuals over time
 - examples: Panel Study of Income Dynamics, Compustat
 - disadvantages: difficult and costly to track individuals over time; data sets are more complicated than other cross-section data sets; unavoidable random and non-random attrition
 - need to distinguish between data that are collected prospectively (as it happens) or retrospectively (recalled)
- Are the measures direct or indirect?
 - again, this is a classification that is usually applied to quantitative data

- direct measures come directly from the data
 - for example, we could look at the average income level for a county using the sample average from the Decennial Census
 - similarly, we could look at employment levels by aggregating the reports from the Current Population Survey
 - relatively straightforward to use and interpret
 - main questions with data quality are
 - is the sample representative?
 - do we understand how the question was answered or what the measure captured?
 - were the responses accurate?
- we use indirect measures (measures built up from other measures) when direct measures are not available or of very low quality
 - cannot obtain direct measures if necessary, specific data were not collected
 - sometimes only a few observations are available; so direct measures have intolerably large sampling variances
- common examples of indirect techniques are
 - interpolation – making an educated guess about a value when data on either side of the value are available (e.g., using data from 1998 and 2000 to estimate a value for 1999)
 - extrapolation – making an educated guess about a value when data are available only on one side (e.g.,

using data from 1998 and 1999 to make an estimate for 2000)

- more sophisticated measures possible
- What is the source of reporting – do the data come from self-reports, proxy reports, or administrative sources?
 - use surveys and interviews to obtain self-reports and proxy reports
 - in a proxy report someone reports for another person
 - examples are one person describing outcomes for other family or household members; an employer categorizing his or her employees; a person describing neighbors
 - the key advantage of self-reported data is that the instruments can be very flexible (you can ask people just about anything)
 - the disadvantages are
 - people might err in their answers (e.g., mistakes and recall errors)
 - people might choose not to provide information (item and unit non-response)
 - alternatively, sometimes economists turn to administrative records such as tax returns, employment records, program rolls to obtain data
 - quality of data is usually higher
 - exceptions are records where people have an incentive to misreport (e.g., tax records)

- also, sometimes the administrators do not care about particular items and either record them haphazardly or do not keep them up to date
 - data items are limited to those of interest to the particular agency; for example, demographic data often missing
 - universe is limited to people who participate in program or fill out a form
 - administrative records on welfare participants only cover the participants, not others who are eligible but don't participate or those who leave the rolls
 - certain types of tax information is limited to people who itemize
 - data are much more sensitive
 - people can choose whether or not to complete a survey; they might not have a choice about providing other types of information to the government
 - researchers need to be much more careful to preserve confidentiality
- Are the data primary or secondary?
 - data we collect ourselves is primary, while data others collect is secondary
 - flexibility
 - for primary data, we can collect what we want

- for secondary data, we are limited to what is available
- cost
 - it takes a lot of time and effort to collect primary data
 - obviously, you have to field the instrument and record/code the responses
 - prior to that, however, you have to design the instrument and create the sampling frame
 - using someone else's sample is much less time consuming; although, there is still a need to understand the data and possibly clean the data