Data Organization

A key part of empirical research is data collection and manipulation. In the final draft of your prospectus, you must locate a data source and provide a brief description of how these data fit into your research design. You want to make sure that you have a viable source of data before progressing too far into your project. Without data, you won't have a way to test your hypothesis.

It makes sense to look for data sources while conducting your literature review. The answer to the question, where can I find data?, is in the previous research on the topic. If you've done a comprehensive literature review, then you will be familiar with the data sources used in earlier research and this is where you should start. Most, if not all, of the papers you include in your literature review make use of data. This is the natural place to start looking for data sources. This information on data sources is often buried in the text, or even in footnotes attached to tables.

The Structure of Economic Data

It is important to distinguish between those organizations that collect or produce data and those that publish data. This is the same distinction one makes between a primary and secondary source.

Data comes in three forms: time series, cross section, and longitudinal (or panel) data.

- *Time series data*: different observations on the same variable at different points across time.
  Time series are organized by date (usually on the left-hand side of a table or spreadsheet). The date is the unit of observation. Time series data are available on different frequencies. The *frequency* refers to how often the data are observed. For example, the U.S. Census data are collected once every ten years. Financial market data like interest rates and stock prices are available daily (sometimes hourly). Time series data are most common in macroeconomics and research questions that use aggregate measures of economic variables.

- *Cross section data:* different observations of a comparable variable at the same point in time.
  Cross section data are organized by individual, as the individual is the unit of observation. Cross section data use a *unit of analysis*, such as nations, states, or individuals. For example, one could collect average personal disposable income in each of the 50 states in 2004. Or, one could collect average personal disposable income across individual households within a state in 2004. Cross section data are most common in applied microeconomics and research questions that use require the use of individual or "micro"-level data.

- *Panel data*: a cross section sample and follows this sample over time.
  This is a combination of the two forms above. For example, if you were to collect average personal disposable income in each of the 50 states in 2000, 2001, 2002, 2003, and 2004, this would constitute a panel. Panel data are used in both microeconomics and macroeconomics, but are more common in the former. In applied microeconomics, it is more common to see panels with a very large cross section of individuals over a relatively short period of time.

Data Sources

It is important to understand the different between a database and a data source. A database compiles data from one or more sources. The data source refers to the agency that collects the data. Some agencies both publish and collect the data. A list of data sources is provided on the course web site under Resources.

The sources above maintain databases for a collection of data they publish. Often, you will use a database because it is more convenient to do so. For example, FRED II reports much of the data from the BEA and BLS in easy-to-use Excel format. However, when referring to the data, you should always indicate the primary source in a Data Appendix. It is important to keep a clear distinction between primary versus secondary sources in your research. This applies to data sources, too.

Searching for Data

Before you search, consider the following:

- What are the desired variables you need to conduct your analysis?
  You should have a good idea of this based on your literature review. Having completed your literature review, you will start to see that different researchers often include similar types of explanatory variables (or "controls"), allowing them to isolate the relationship between the dependent variable and the explanatory variable of interest.

- How should each variable be defined?
  Often there are many different options for how to measure a theoretical concept. For example, there are several ways to measure unemployment, money supply, and the interest rate. Similarly, in applied microeconomics, there are many different ways to measure education level, race, and income. Look to the literature to guide you as to which measure is the most appropriate for your research.

- What data frequency and sample period?
  Are the data quarterly or monthly? In cross section, which day/month/year were the data collected?

- What is the unit of analysis?
  Are the data collected for individual persons, households, schools, or countries?

- What are the potential sources of the data for each variable?
  Again, you should rely on previous research to guide you.

As you search, consider the following:

- What data are available? Consider sample size (how many observations are available):
  - Time series: frequency and sample period available
  - Cross section: number of observations available
- Are their suitable proxy variables for variables that are not available?
- If not, how can the empirical model be modified to use the data available, but still test the hypothesis?

You will be expected to create a Data Appendix for your research paper. This includes

- primary source of your data,
- the frequency (time series only), sample period, and unit of analysis (cross section),
- the units in which in variable is measured in (e.g., dollars, percentage points, rates, billions of people, etc.).

When collecting your data, keep this in mind when keeping track of your data sources.