

Intro to Data Analysis, Economic Statistics and Econometrics

- Statistics deals with the techniques for collecting and analyzing data that arise in many different contexts.
- Econometrics involves the development and use of special statistical methods within a framework that is consistent with the ways of economic analysis.
- Econometric analysis starts from a statement about some behavioral relationship, which might come from economic theory, or simple observation of reality. Then it is developed into an equation that specifies how one variable is determined by the values of other variables.

Let's start with a simple example

- It is coherent to believe that there is some relationship between consumption and income, in economic theory is widely believed that this relationship is fairly stable at the aggregate level. We can write this relationship as follows:

$$C = \beta_0 + \beta_1 Y + u \quad (1)$$

- C represents consumption, Y represents income, and they are the *variables* in this equation. β_0 and β_1 are the *coefficients*, and they are unknown constants or parameters.
- The *disturbance* term, u represents all the factors on top of income, that could help determine consumption that we choose not to include or think they are not relevant enough. This term can also include pure chance and error in the measurement of C .
- This behavioral equation is a model of the economic process that determines the level of consumption in the economy. It is obviously abstract and simplified given the complexity of reality.
- In most cases in econometrics we are going to assume our models are correctly specified, in some cases we will be able to test them. If the model is incorrect in its essence all our conclusions would be wrong.
- Econometricians face the challenging task of blending knowledge of economic theory and statistics to produce models that reflect, as correctly as possible, the complex economic reality.

- The main task of the econometrician when faced with a model like (1) is to try to estimate the unknown parameters of interest, β_0 and β_1 . For example we could use data on these variables and statistical techniques we will describe later on to get that $\beta_0 = 0.568$ and $\beta_1 = 0.907$. Then we can write the estimated model as

$$C = 0.568 + 0.907 Y. \quad (2)$$

- The interesting thing here is to tie this back to the economic model. So we interpret the estimated value of β_1 as the marginal propensity to consume, which theoretically is believed to be less than 1.
- Can we conclude from the result in (2) that the theory is correct? We have to remember that we have only computed an estimate of β_1 . In order to confirm or reject the hypothesis implied by the theory we need to study questions regarding errors associated to the estimation process, we need to study more to be able to answer this question.

Some useful terminology

- *Data* are the quantitative facts or pieces of information that we deal with in any statistical analysis. When taken together with all the other data we use in a particular analysis they constitute a *data set*.
- When we study an economic process, the data present information about a set of cases of the occurrence of that process, these cases are called *observations*, and the nature of these cases define what we call *unit of observation*.
- For example in our initial example, if we understand the relationship at the macroeconomic level, the observations might be different years of a particular economy, or they might be information about different countries.
- If we understand the relationship between consumption and income as occurring at the microeconomic level, then the unit of observation is going to be the family, or the individual. The observations will consist of data for the same family over time, or for different families.

- A *variable* is a measurable characteristic about which we collect data. For example when studying consumption at the micro level this might be: consumption, income, family size, the number of children, county of residence, whether all adults in the household are working, etc. Notice that some of them are discrete (take on a limited number of values) and others are continuous (can assume any value in a given range).
- All these concepts come together to form a *rectangular data set*, where we have a relatively large number of observations, n , which are arranged in the rows, and a number of variables measuring some characteristic of each of our units of observations, arranged in the columns.
- A given set of observations is often thought of as being a *sample* drawn from some larger population. In many applications it is important for the researcher to have access to a representative sample, so the conclusions can be considered valid for everyone in that population.

*It is important to distinguish between **Cross-section data** and **Time-Series data**.*

- Cross-section data arise when the observations are data for different entities (persons, firms, countries) for which a common set of variables is measured at about the same point in time. For example, the data collected in the U.S. decennial Census are a cross-section, with information on income, age, family size, etc. The data on unemployment rates, inflation rates and labor force participation rates of individuals 15 to 64, in all the U.S. States in a given year are also a Cross-section.
- Time-series data arise when the observations are data for the same agent in different periods of time. For example, records of a person's employment and earnings in each year of his life. Data on unemployment rates in a particular state since the 1960s.

Before moving into specific measures used to describe data a couple of additional issues should be discussed.

- **Change and Growth:** When trying to describe how a variable changes over time we can use two different concepts. The *absolute change* or growth, which is simply the difference between a particular period's value and the previous one. For a time-series variable Y the absolute growth in the i^{th} period is given by the difference

$$\Delta Y_i = Y_i - Y_{i-1}, \quad (3)$$

notice that the result might be negative. Then we can compute the *relative change*, which relates the absolute change in a variable to its previous level. Relative change or growth in the i^{th} period is measured by the *rate of change* or the *rate of growth*

$$r_i = \frac{\Delta Y_i}{Y_{i-1}} = \frac{Y_i - Y_{i-1}}{Y_{i-1}} = \frac{Y_i}{Y_{i-1}} - 1. \quad (4)$$

Notice that rate here has the meaning of a measure of relative change. Although the absolute growth concept is easier we usually rely more on rate of change or rates of growth to describe economic data.

- **Constructed Variables:** It is extremely common in economics to construct variables as a combination of other variables. We construct indexes to reflect the average of an underlying set of prices or quantity variables in each year. For example the Dow Jones Index, the Consumer Price Index (CPI), the Index of Industrial Output, the Index of Consumer Confidence, etc.

Data Analysis: Measures of Central Tendency

- The *Mean* or “Average”: by far the most common measure of central tendency. It is defined as the sum of the values of the variable divided by the number of observations:

$$\sum_{i=1}^N \frac{x_i}{N} = \mu \quad (5)$$

when we have the whole population N . In a given sample:

$$\sum_{i=1}^n \frac{x_i}{n} = \bar{X} \text{ (sometimes denoted as } \hat{\mu}\text{)}, \quad (6)$$

where usually $n \leq N$.

In a simple example, imagine we have the incomes of five individuals on campus: \$15,000; \$35,000; \$45,000; \$79,000; \$120,000. The mean of these incomes is \$58,800. It is easy to compute and easy to understand.

However, it has disadvantages: for example is very sensitive to extreme values or outliers. In our example if the highest salary is replaced with say (do you know whose salary this is?) \$225,000, then the average is \$79,800, which is much larger although only one observation has changed.

Notice that the mean is in general a value that is not in the data. However, one advantage is that it has a clear mathematical formulation, and it is very stable across different samples of a given population.

- The *Median*: it is less common but in many cases a better way to describe a variable. It is the value of the middle observation on X , after the observations have been ordered from smallest to largest. In other words, 50% of the observations are above this number, and 50% are below, that is why it is also called the 50th percentile. In our example above the median is \$45,000 (notice that this is regardless of whether the highest number is \$120,000, or \$ 225,000. If we have an even number of observations, then we have to compute the average of the two values in the middle. In the example above if we add the \$225,000 to the original 5 values, the median would be \$62,000.
- Notice that the median is not affected by extreme values, but it is less stable when comparing different samples of the same population. Notice also the fact that in this example the median is considerably smaller than the mean, given by the fact that the set of higher salaries are comparatively further away from the median than the lower salaries. The distribution is not symmetric.
- The *Mode*: It is the most frequently occurring value. The mode is sometimes a useful measure for identifying the typical value of a discrete variable, such as family size, but it is not too useful for continuous variables because every observation's value may be unique. For our example, all values happen equally often. It has the advantage of actually occurring in the data.