INTRODUCTION AND OVERVIEW

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

- Important to understand the approach we will be using in this course. The notion of a population is very important – just as in a basic probability and statistics course.
- This term, we will assume that our data are obtained as a random sample from a specified population. All methods through Chapter 18 assume random sampling from an appropriate population. Chapter 19 begins the study of data problems, such as censoring, missing data, stratified sampling, and clustering.

- Given random sampling, issues of model specification are purely population issues. In particular, the notion of **identification** of parameters – determining when can we consistently estimate population parameters – can be studied entirely in terms of the population.
- We will cover panel data, which has a time series dimension. Because variables are usually correlated over time, we should not assume independence over time. The random sampling assumption will apply to units (people, firms, and so on) sampled from a cross section; correlation across time largely will be unrestricted.

- The setup here is a bit different than in traditional approaches to multiple regression. We will not assume explanatory variables are nonrandom or “fixed in repeated samples.” Assuming fixed regressors simplifies some algebra but is not realistic.
- Assuming fixed regressors ignores what is probably the most important consideration: how is the error term related to the regressors?
- If we do not have to worry about the relationship between unobservables and the explanatory variables, the most interesting issues go away.

• As an example of our general approach, we might start with a model for a conditional density, $f(y|\mathbf{x};\boldsymbol{\theta})$, where $\mathbf{x}$ is allowed to be a vector and so is $\boldsymbol{\theta}$ – the parameter vector. Suppose we know the form of $f(y|\mathbf{x};\boldsymbol{\theta})$ up to the unknown parameters, $\boldsymbol{\theta}$. Then we can discuss whether $\boldsymbol{\theta}$ is identified based on population considerations, including how the density depends on $\boldsymbol{\theta}$ and the distribution of $\mathbf{x}$ in the population.

• We then assume we have access to a random sample of size N, $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, N\}$.

- Often we are only interested in some feature the conditional distribution, or some feature of a joint distribution. (Regression analysis is the leading case.) We can still investigate identification without fully specifying a distribution.
- A simple example is to define $\mu = E(y)$ (the population mean) as the parameter of interest. For any distribution such that μ is well defined, μ is identified under random sampling.
- Technical Note: Unlike in a course in probability and statistics, we will not always be careful in using different notations for random variables and a particular outcome.