Tratto da: Econometric Analysis of Cross Section and Panel Data, Jeffrey M. Wooldridge, MIT Press 2002

1 Introduction

1.1 Causal Relationships and Ceteris Paribus Analysis

The goal of most empirical studies in economics and other social sciences is to determine whether a change in one variable, say w, causes a change in another variable, say y. For example, does having another year of education cause an increase in monthly salary? Does reducing class size cause an improvement in student performance? Does lowering the business property tax rate cause an increase in city economic activity? Because economic variables are properly interpreted as random variables, we should use ideas from probability to formalize the sense in which a change in w causes a change in y.

The notion of **ceteris paribus**—that is, holding all other (relevant) factors fixed—is at the crux of establishing a **causal relationship**. Simply finding that two variables are correlated is rarely enough to conclude that a change in one variable causes a change in another. This result is due to the nature of economic data: rarely can we run a controlled experiment that allows a simple correlation analysis to uncover causality. Instead, we can use econometric methods to effectively hold other factors fixed.

If we focus on the average, or expected, response, a ceteris paribus analysis entails estimating $\mathrm{E}(y \mid w, \mathbf{c})$, the expected value of y conditional on w and $\mathbf{c}$. The vector $\mathbf{c}$—whose dimension is not important for this discussion—denotes a set of **control variables** that we would like to explicitly hold fixed when studying the effect of w on the expected value of y. The reason we control for these variables is that we think w is correlated with other factors that also influence y. If w is continuous, interest centers on $\partial \mathrm{E}(y \mid w, \mathbf{c})/\partial w$, which is usually called the **partial effect** of w on $\mathrm{E}(y \mid w, \mathbf{c})$. If w is discrete, we are interested in $\mathrm{E}(y \mid w, \mathbf{c})$ evaluated at different values of w, with the elements of $\mathbf{c}$ fixed at the same specified values.

Deciding on the list of proper controls is not always straightforward, and using different controls can lead to different conclusions about a causal relationship between y and w. This is where establishing causality gets tricky: it is up to us to decide which factors need to be held fixed. If we settle on a list of controls, and if all elements of $\mathbf{c}$ can be observed, then estimating the partial effect of w on $\mathrm{E}(y \mid w, \mathbf{c})$ is relatively straightforward. Unfortunately, in economics and other social sciences, many elements of $\mathbf{c}$ are not observed. For example, in estimating the causal effect of education on wage, we might focus on $\mathrm{E}(wage \mid educ, exper, abil)$ where *educ* is years of schooling, *exper* is years of workforce experience, and *abil* is innate ability. In this case, $\mathbf{c} = (exper, abil)$, where *exper* is observed but *abil* is not. (It is widely agreed among labor economists that experience and ability are two factors we should hold fixed to obtain the causal effect of education on wages. Other factors, such as years

with the current employer, might belong as well. We can all agree that something such as the last digit of one's social security number need not be included as a control, as it has nothing to do with wage or education.)

As a second example, consider establishing a causal relationship between student attendance and performance on a final exam in a principles of economics class. We might be interested in $\mathrm{E}(score \mid attend, SAT, priGPA)$, where *score* is the final exam score, *attend* is the attendance rate, SAT is score on the scholastic aptitude test, and *priGPA* is grade point average at the beginning of the term. We can reasonably collect data on all of these variables for a large group of students. Is this setup enough to decide whether attendance has a causal effect on performance? Maybe not. While SAT and *priGPA* are general measures reflecting student ability and study habits, they do not necessarily measure one's interest in or aptitude for econonomics. Such attributes, which are difficult to quantify, may nevertheless belong in the list of controls if we are going to be able to infer that attendance rate has a causal effect on performance.

In addition to not being able to obtain data on all desired controls, other problems can interfere with estimating causal relationships. For example, even if we have good measures of the elements of $\mathbf{c}$, we might not have very good measures of y or w. A more subtle problem—which we study in detail in Chapter 9—is that we may only observe equilibrium values of y and w when these variables are simultaneously determined. An example is determining the causal effect of conviction rates (w) on city crime rates (y).

A first course in econometrics teaches students how to apply multiple regression analysis to estimate ceteris paribus effects of explanatory variables on a response variable. In the rest of this book, we will study how to estimate such effects in a variety of situations. Unlike most introductory treatments, we rely heavily on conditional expectations. In Chapter 2 we provide a detailed summary of properties of conditional expectations.