ECO 270: Econometric Methods[1]

All economics majors at Bates must take considerable coursework in econometrics. This topic is therefore redundant with coursework that students have taken, are taking, or will take later. So I will keep this material brief. My goal is to provide the bare minimum of background needed to talk about empirical results related to both short and long run macroeconomics. I will leave most of the technical details to your other classes.

Empirical work is that which relies on observations. In economics, this usually entails collecting data on observed behavior. In many fields, this is often the result of experiments. Experimental economics, however is a relatively small field so most empirical economics is based off of observing agents in the actual economy. *Econometrics* refers to the statistical analysis of such data.

Suppose that we wish to understand the relationship between two variables: the average tax rate (hereafter denoted as T_t), and GDP growth (hereafter Y_t). Specifically, suppose that we want to understand how changes to taxes affect GDP growth. For both variables, the t subscript indicates the time period.

The simplest approach that we can take is to look at *unconditional correlations.* Such correlations measure how two variables move together. Formally:

$$r_{T,Y} = \frac{1}{t-1} \sum \frac{T - \bar{T}}{s_T} \frac{Y - \bar{Y}}{s_Y} \tag{1}$$

where $\bar{T}$ is the average tax rate, and s_T is the standard deviation of taxes. Suppose that we observe a series of tax ratios :$T_t = [.1, .2, .1, .3]$, and a series of GDP growth rates: $Y_t = [.05, .1, .05, .15]$.

If we plug these series into (1), we obtain a correlation of 1. This implies that the two variables move together perfectly in the same direction, note that when the tax rate doubles, so does GDP growth.

Now suppose that the tax rate has the same series, but that GDP growth has the following: $Y_t = [.3, .15, .3, .1]$. Once again the variables move together perfectly, but in the negative direction. When the tax ratio doubles, GDP growth declines by 50%. This implies a correlation

[1] These are undergraduate lecture notes. They do not represent academic work. Expect typos, sloppy formatting, and occasional (possibly stupefying) errors.

of -1. A correlation of zero indicates that there is no tendency for two variables to move together.

Correlations represent one way that economists judge theoretical models. Theoretical models often yield predictions about how two variables are correlated. Macroeconomists are particularly interested in how variables such as consumption, investment, etc. are correlated with aggregate output. When this correlation is positive, a variable is said to be *procyclical*, when it is negatively correlated, it is said to be *countercyclical.* All else equal, a model is better if its predicted correlations match those of the actual data.

Suppose that we find a strong negative correlation between T_t and Y_t. It may be tempting to use this result to establish a causal relationship between the two variables. In this case, such a conclusion would be unwise. There are two related issues:

1. It is not possible to establish which variable causes the other. Perhaps higher taxes reduce either the incentive to supply labor or aggregate demand. In this case, higher taxes cause lower growth. But it is also possible that lower GDP growth cause budget deficits which then incentivize policy makers to raise taxes. in this case, lower growth causes higher taxes. A simple correlation offers us no insight into which way, or both, the causation may run.

2. A simple correlation does not control for other factors. Consider the example of government spending. Higher spending, depending on the conditions and which theory you believe, can either increase or decrease GDP growth. At the same time, higher spending creates an incentive for fiscal policymakers to raise taxes.[2] It is thus possible that spending causes both variables and there may be no causation between Y_t, and T_t or that the true causation may be very different than that suggested by the simple correlation. Government spending is just one example with these data, there are an infinite number of possible variables that could affect results. Simple correlations neglect these. This is known as *omitted variable bias.*

The best way to establish causation would be to run an experiment. In this case, we might the following:

1. Take 1,000,000 countries and randomly assign each of them a tax rate. By making a random assignment, we can be sure that no other variable is causing the tax rate.

[2] This claim does appear dubious based on watching the last 30 years of fiscal policy. But suspend your disbelief for the purpose of the example.

2. If the sample is big enough, then we can be confident that differences in spending, technology, and anything else we can think of average out across the sample.

3. Calculate simple correlations.

Now the correlation can establish causation with extreme confidence. Such experiments are often possible in fields such as medicine. But they are rarely possible in macroeconomics. So empirical macroeconomists usually turn to another tool, *linear regression analysis*. Regression analysis attempts to minimize omitted variable bias by controlling for other factors when experiments are not feasible.

The first step is to collect data on the other factors that we believe might have a significant effect on GDP growth. Suppose for example, that we gather data on the ratio of government spending to GDP (G_t), and the level of education (E_t). We can then run the following regression:

$$Y_{i,t} = \alpha + \beta_0 T_{i,t} + \beta_1 G_{i,t} + \beta_2 E_{i,t} + \mu_{i,t} \tag{2}$$

Note that I have changed the subscripts to include both i and t. This is to allow for variation over both time and places, such as countries. The parameters α and $\beta_0 - \beta_2$ are known as the *regression coefficients*. The term $\mu_{i,t}$ is the error of the regression. For any observation, we do not expect a perfect fit. The error is difference between the predicted level of GDP growth (left side), and the actual level (right side). To obtain the regression coefficients, we must have define some measure of fit. Usually, we choose them in order to minimize the sum of squared errors: $\sum \mu_{i,t}^2$.

We could obtain the regression coefficients by simply trying out different combination of regression coefficients and choosing the set with the smallest squared errors. Often, however, a simple mathematical formula will yield the correct regression coefficients. So suppose that we do this and get the following result:[3]

$$Y_{i,t} = 0.06 - 0.5T_{i,t} + 0.02G_{i,t} + 0.8E_{i,t} + \mu_{i,t} \tag{3}$$

We can interpret this result as follows: All else equal, a 1% increase in the tax to GDP ratio corresponds to a 0.5% reduction in GDP growth. Here, “all else equal” means that we are leaving the other variables on the right hand side (spending and education) unchanged. We have thus attempted to control for these two factors in a way that simple correlations do not. Now consider the result for a specific observation, say France in 1985:

[3] Note that these results are entirely fake.

$$Y_{Fra,85} = 0.06 - 0.5T_{i,t} + 0.02G_{i,t} + 0.8E_{i,t} + 0.01 \tag{4}$$

We chose our regression coefficients to provide the best fit. This does not mean that they yield perfect fit. For the fake French data, plugging in the French tax, spending, and education data for 1985 yields a predicted level of GDP growth that is 0.01 less than the actual growth rate. This is the error of of the regression for this observation. But we have chosen our regression coefficients in order to make these (squared) errors as small as possible.

Our regression helps us establish causation by controlling for other factors. In this case, however, we probably have not controlled for all of the factors that could affect GDP growth by including them on the right hand side. We thus still have omitted variable bias. For this example, it is severe enough so that this result is not of much interest. Among other factors, we have omitted the quality of institutions, and the type of government.

For observational data, it is not possible to eliminate omitted variable bias. Any variable of interest is affected by countless factors and it is impossible to control for all of them. This problem greatly limits the appeal of atheoretical econometric work. If we run a regression without carefully thinking about the economic mechanisms behind the data, then it is impossible to be confident that we have effectively minimized omitted variable bias.

For a variety of factors, including omitted variable bias, it is impossible to use econometrics to establish causality with 100% confidence. But if we are able to use good economic theory/intuition to carefully control for the other things that matter, and if we deal with the many additional complications that these notes leave for your other coursework, then econometrics can help make a convincing case for causation.

These notes barely scratch the surface of empirical macroeconomics. But I hope that they are adequate to allow us to meaningfully interpret econometric results without blindly accepting them. After this topic, the class will examine the topic of economic growth. We will begin by examining some empirical results, and we will follow it by comparing two theoretical models: the Solow Model and the Endogenous Growth Model.

To conclude this topic, we will discuss two other pitfalls of empirical work:

#1: The Kitchen Sink

Many undergrads in econometrics courses are guilty of making this mistake. For our results to be valid, we must have significantly more observations than right hand side variables. To see this, suppose that I have two years worth of data on labor supply. In 2008, labor supply =10, and in 2009, labor supply =20. Now suppose that I randomly make up two variables. Call the first random variable A. In 2008, A=3, and in 2009, A=2. Call the second B. In 2008, B=1, and in 2009, B=0. Because A and B are random variables, there obviously is no real world relationship among them and labor supply. Our regression, however, would look like:

$$LaborSupply = 10 * A - 20 * B \tag{5}$$

There is no error in this regression, it perfectly explains the data ($R^2 = 1$, for those of you who have taken Econometrics). It may thus be tempting to conclude that the results are especially good. But these results are, of course, are nonsensical. There is no reason to expect this same relationship to persist in the future. It is purely the result of our including too many variables in our regression. This problem ties into the earlier discussion on omitted variable bias. Unless we include everything, we have omitted variable bias. But if we do include everything, we have this problem. The challenge is to find a happy medium.

#2: Excessive Reliance on Extrapolation

This mistake may or may not be made as part of regression analysis. Consider the following egregious example:

In the journal *Obesity*[4], five authors note that obesity rates are increasing among almost all subgroups in the American population. The authors note that if the rate of increase continues:

"By 2048, all American adults would become overweight or obese"

While extrapolation may be a useful period over short periods of time in a stable environment, assuming that current growth rates will stay the same for 40 years is transparently absurd. Using the authors' technique, we could also reach the following, equally stupid, conclusions:

a. At some point after 2048, 110% of Americans will be obese.

b. Average tax rates have been decreasing in the United States. At some point in the future, they will become negative.

c. Between August 2012, and October 2013, average daily temperatures in Lewiston have been falling. Within ten years, the average daily temperature will hit absolute zero.

[4] Youfa Wang, May A. Beydoun, Lan Liang, Benjamin Caballero and Shiriki K. Kumanyika, "Will All Americans Become Overweight or Obese? Estimating the Progression and Cost of the US Obesity Epidemic" *Obesity* (2008) 16 10, 2323-2330.

d. NFL players have become faster over the past 20 years. Eventually, they will all run at the speed of light.

A prominent example of this is the level of GDP in the United States and Japan in the 1980s. At the time, U.S. GDP was substantially higher than that of Japan, but GDP growth was higher in Japan than in the United States. Many people extrapolated this result and concluded that the level of GDP in Japan would surpass that of the United States several decades in the future. But this extrapolation was too far in the future to be valid (and it has not yet happened and it seems unlikely to occur in the foreseeable future). This example is useful to keep in mind when comparing the recent macroeconomic performances of the United States and China.

There are, of course, many other ways to screw up empirical work. This list is in no way intended to be exhaustive.