Takashi Yamano

Lecture Notes on Advanced Econometrics

Lecture 1: Empirical Research and Sampling

Why do we study Econometrics? For almost all of us in this class, we study econometrics because we need to conduct sold empirical analyses, instead of advancing econometric theory. More precisely for you, you study econometrics because you need it as a tool to write a thesis or dissertation to complete your program.

Some research papers do not include econometric analyses but use conceptual frameworks and subject-matter knowledge to support their arguments. They may add descriptive data with some basic tables and figures. In most cases in academic studies, however, conceptual and descriptive data analyses may not be enough. Recent years, data are becoming available more than before and powerful econometric packages are easier to use than ever. Thus, it is much easier for researchers to test their hypotheses empirically with good data and sophisticated econometrics methods. Thus, there are fewer excuses for not conducting econometric-analyses.

Although econometrics can be a powerful tool to test hypotheses, it is quite easy to misuse it. To convince readers that you have interesting and reliable results, you first need to convince them that you have arrived robust conclusions based on sound data and well-thought-out methods.

I hope this course will help you on this point.

This course uses two textbooks: **"Introductory Econometrics" (2nd edition) by Jeffrey Wooldrige** and **"Econometrics Analysis" (5th edition) by William H. Greene**. The first book is written for empirical researchers who want to learn how to conduct econometric analyses. As the sub-title of the first book --"A Modern Approach"-- suggests, this book also covers recent developments in applied econometrics.

The field of Applied Econometrics is developing rapidly due to improvements in personal computers and expanding availability of data. By using econometric-packages that are

currently available in the market, it does not take much time to estimate complicated models. Large data are available through Internet or are sold in CD-ROMs. This means that a wide variety of methods are available to you.

Previously, those advanced methods are only available to people who had advanced knowledge on econometrics. Thus there was little need for introductory textbooks to explain advanced methods. However, because those advanced methods are currently available from software packages, such as STATA, there is a need for an introductory textbook to explain advanced methods for people who do not have advanced knowledge on econometrics. Wooldridge realizes this and has written the textbook.

However, because this textbook covers many methods in a single volume, it does not explain theory in details. Thus, I use the second textbook (Greene's "Econometric Analysis") to study econometric theory in details. Especially, the second book explains econometric theory in matrix. Econometric theory can be explained clearly with matrix as you will find in the lectures. I will provide you a basic training in matrix calculation in this course.

I start my lecture on how to conduct empirical projects because the main purpose of this course is to provide you a tool to write empirical papers. I think it is important for you to have a clear idea on how you are going to use econometrics in your research.

Carrying out an Empirical Project

Posing a question

There are some ways to pose a research question:

New Theory
New Issues
New Data
New Methods
Changed or Different Environments

New Theory: You may pose this type of questions when you come up with a new theory to look at an issue. You may provide a formal theory to an old issue. For instance,

people knew that educated people gained some returns from education even before Chicago economists, such as T.W. Shultz and Gary Becker. However, Chicago economists created a new theory of human capital by conceptualizing education as investment in human capital.

New Issues: You may pose this type of questions when you recognize emerging issues. These issues may have existed before, but you may argue that these issues have gain importance. International trade and finance were not important when it was difficult to transfer goods or money across countries. But today, it is very difficult to find a place without internationally-traded goods in the world. As a result, international trade and finance issues have gained their importance in economic theory.

New Data: Although many data sets have become available recently, this is a relatively new phenomenon. There are many empirical issues that have not been examined empirically because appropriate data were not available before. Good data are hard to come by. So when you come across a newly available data with high quality, you should write a paper by saying "This paper uses a newly available data set which enables us to overcome limitations in previous studies…"

New Methods: you can advance the knowledge among researchers by applying a new and *more appropriate* method on old issues. By studying econometrics, you will be able to apply advanced methods on old issues.

Changed or Different Environments: his is the most common practice. This is to apply an existing research question and method in a changed (over time) environment or different regions or countries. You may conduct an old research when you suspect that the environment has changed over time and that results would be different from previous ones. Or you may conduct an existing research in a different place or country.

But how do you know if your idea is really new? The answer: Dig in the literature.

Literature Review

The most common place to look for research questions is papers in journals (we call them *literature*). It turns out that you are not the only one who needs to conduct economic

analyses. Many people have done or are doing exactly what you are doing, and you can find other people's efforts in their published papers in journals, books, and on Internet.

Although you may want to come up with a brand-new-question that nobody has thought about it before, there are few questions that people have not thought about. Chances are that some economists have thought about your seemingly-brand-new-ideas many years ago, if not many decades ago, and have done fine analyses.

As you review literature, try to narrow down your focus and find a line of research that is on the same research question. For instance, you may want to study education in developing countries and start reading about it. As you go through previous studies you will find studies on (a) the effects of education on farm productivity, (b) the effects of mothers' education on child health, (c) the economic returns of education in labor markets, (d) the effects of school quality on test-scoring, and so on. Each one of them could be called a line of research.

At the end, along the line of research, you should look for "a gap in the literature." And in your paper, you should be able to say: "Previous studies have found these results before, but we still do not know the answer to this question. The purpose of this paper, therefore, is to find an answer to this question."

In your analysis, you do not need to cover all of the previous studies on the issue on the line of research that you are interested in. If you are writing a journal paper, your readers are usually experts on the subject. Probably they know about previous studies better than you do. Thus, you just need to mention relevant studies to identify "a gap in the literature" that you are going to fill.

Data Collection

High quality data can lead researchers to high quality empirical research. Econometrics can help you to produce high quality empirical research from high quality data. But econometrics can not help you to produce high quality empirical research from low quality data. The quality of the data draws the upper limit on the quality of research. An empirical analysis using low quality data with very advanced econometric techniques is

not useful practically, even if the technique itself is impressive. On the other hand, a simple analysis with high quality data can provide useful information. (I am a strong believer of the latter method.) Large data sets are useful but not necessarily better than small data because often large data sets lack detailed information, although large observations can help narrowing standard errors.

More and more data are becoming available from Internet. This is great. But when you use such data, you first need to understand what you get. It is dangerous to use data without a full knowledge on how the data were collected and arranged.

When you collect data (yourself from fields or from secondary data sources), you should think very carefully if you will be able to answer your research questions with the data you collect. If not, you need to search for different data or change your research questions. Otherwise after many months (or years) of efforts, you would be told by your advisers or reviewers of your paper that "Well, it seems you can't answer your questions with your data …"

Econometric Analysis

The purpose of your research is to find answers to your research questions and convince people (readers) that you have reliable results. To achieve your goal, you may not need the most advanced econometric methods. For the most cases, you probably do not need the most advanced econometric methods. Simple methods are usually sufficient and powerful.

You should, however, study advanced methods for following reasons. First, you should be able to choose the most appropriate method among all methods. Second, you should have a full understanding on the method of your choice. Third, you should recognize the limitations of the method.

Writing an Empirical Paper

After your analyses, you need to summarize your findings in a concise manner. A short paper is desirable. Based on my limited experiences, I would suggest:

(a) Start with a short paper with a specific question
(b) Work with people who have publications
(c) Mimic other papers
(d) Keep on writing; if you have been stuck in one section of your paper, you should start writing other sections

Writing an empirical paper is very difficult. But you are not alone. Don't be discouraged!

Reference

Brorsen, B. Wade. (1987) "Observations on the Journal Publication Process," *North Central Journal of Agricultural Economics*, Vol. 9, No. 2.

On English Writing, I suggest:

Strunk and White (2000) *The Elements of Style*, new edition, Massachusetts: Alan and Bacon. [This is the classic.]

O'Conner, P.T. (1996) *Woe Is I: The grammarphobe's guide to better English in plain English*, New York: Puttnam's Sons Publishers. [This is fun to read.]

Data (Ch. 1 in Wooldridge; Ch. 1 in Lohr, 1999)

Once you have research questions, you need to collect data. You may collect data from secondary data sources such as official statistics from governments or other institutions, or you may collect data yourself with your colleagues. There are several types of data:

Time Series Data
Cross-Sectional Data
Pooled Cross Sections
Panel (or Longitudinal) Data

When you consider using a data set, you need to understand the following:

Observation Unit: an object on which a measurement is taken
Target population: the complete collection of observations we want to study
Sample: a subset of population
Sampled population: the population from which the sample was taken

Sampling unit: the unit of actual sample
Sampling frame: the list of sampling units

You need to find the data that will enable you to find answers to your research questions. For instance, if you are interested in individual behavior, you want to have individual-level data. You may think this is obvious but in many cases you may not find individual-level data available to you. When you cannot find individual-level data to study individual behavior, you may be forced to use aggregated data.

Suppose, for instance, that you are interested in writing a research paper on the impacts of a drought on farm production but only have access to aggregated data. To carry out an analysis, you need to know how the data were collected from farmers. When was the survey conducted? (Before or after the drought?) Who were the target population? (Country, some regions, or some districts?) Were hard-hit areas included in the sampled population? (If not, there is no point of using the data.) Were hard-hit areas combined with not-so-hard-hit areas? (Farmers in not-so-hard-hit areas might have enjoyed high output prices, created by the drought.)

Without a good understanding on data, econometric analyses cannot provide good analyses.

Some Important Concepts

Causality and Ceteris Paribus

In scientific research, researchers create experimental environment in labs. In such environments, researchers can carefully control factors, such as temperature or humidity. Under carefully controlled environments, it is relatively easy to change one variable, holding other variables constant, and examine a reaction in an object. Because all of the other variables are held constant, *ceteris paribus*, it is easy and reliable to find *a causal effect* of the variable, which was changed, on the object.

In social science, such as economics, it is impossible to control the social environment. What we can do is to observe changes in people's behavior under simultaneous changes in many factors. When many factors change at the same time, we may find two variables changing in the same direction. This could be just a coincidence (an association). Or this

could be a causal effect of one factor to another (a causal effect). For instance, we find that the demand for air-conditioners was high during the past summer and that the household income was also high at the same time. Thus, you find an association between the demand for air-conditioners and the household income. But you also find that the temperature was very high in the past summer. Thus, it is possible that the demand for air-conditioner was high not because of the high income but because of the high temperature. Thus, the household income did not have a causal effect on the demand for the air-conditioners but the high temperature had a causal effect. Distinguishing an association from a causal effect is one of the most difficult issues in empirical research.

Omitted Variables

With econometrics, what we hope to do is to isolate an effect of one variable from other effects of other variables. By including all of the variables that affect an outcome, you are hoping to control for other factors.

A major problem is that you often do not have information on some important variables. We call those variables **unobserved variables**. Some of unobserved variables are "observable but not observed," while other variables could simply be **unobservable**. People's ability and taste are examples of unobservable variables.

When there are unobserved variables, those variables will be omitted from econometrics analyses. When important variables are omitted from econometric models, omitted variables may cause biases in estimated coefficients (**the omitted variables problem**).

For instance, suppose that you want to measure a causal effect of agricultural credit on farm productivity. You have information on farm productivity and agricultural credit, but you do not have information on the education of farmers. Suppose you find a positive **association** between farm productivity and agricultural credit: the higher the amount of agricultural credit, the higher the farm productivity. But it is possible that the credit is given to educated farmers whose farm productivity is high even without the credit. In this case, we would find a positive association between credit and farm productivity even when credit does not have any positive impacts on farm productivity.

In this case, what we would like to do is to measure the impacts of credit on farm productivity among farmers with the same level of education. But we will not be able to do this without the education information of farmers. Even when education information is available, we do not have the information on farmers' ability (which can not be measured fully by formal education) in general.

The omitted variable problem is one of the most common and serious problems in cross-section analyses. We will discuss this problem extensively in this course. And later in this course, we will learn methods to overcome this problem.

Simultaneity

The direction of the causal effect may not be one-way. Two factors may influence each other. In this sense the direction of causality is not clear. For instance, the previous example of farm productivity and agricultural credit can be considered as a simultaneous problem. One could argue that agricultural credit increases farm productivity, while other would argue that agricultural credit is given to highly productive farmers.

In many cases, we can consider a simultaneous problem as an omitted variables problem. For instance, the above simultaneous problem can be considered as an omitted variables problem: if we have the information on all of the characteristics that represent farmers' ability to repay credit, then we can compare two farmers who have the same ability to repay but one is given credit while the other is not. If we find that the farmer who is given credit has higher farm productivity than the other farmer, without credit, who otherwise has the same repayment ability, then we may conclude that there is a causal effect of credit on farm production.

The problem, however, is that it is difficult to observe all of the characteristics that represent farmers' ability as data. Thus the simultaneous problem between agricultural credit and farm productivity remain as an omitted variables problem of farmers' ability. For now, I would just want you to be aware of differences between an association and a causal effect. What we usually want to know is a causal effect of a variable, which is often an important policy variable, not an association.

Unbiasedness and Preciseness

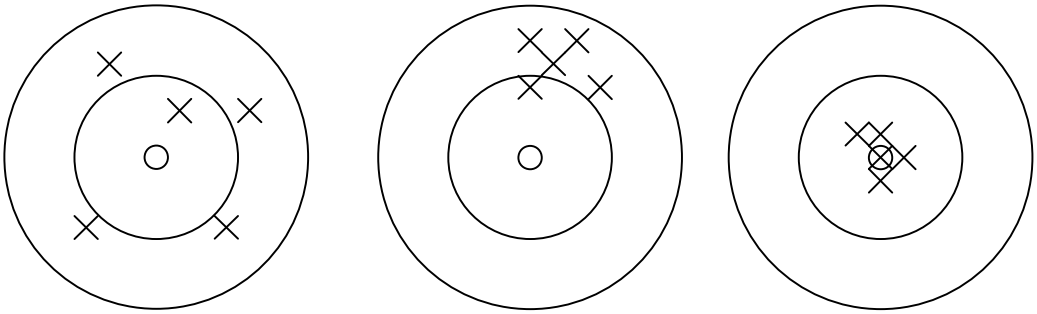
In Econometrics (and Statistics), it is important to clearly distinguish unbiasedness and preciseness. An estimate could be Biased and Imprecise, Biased but Precise, Unbiased but Imprecise, and Unbiased and Precise (Accurate). See Figure 2.2 in Lohr (1999).

In Econometrics (and Statistics), we suppose that there is one true value, **b**. This could be an average number (e.g., the average income, expenditure, age) or a size of an impact of one variable on another (e.g., the impact of smoking one pack of tobacco on the probability of having the lung cancer). By using data and econometric methods, we estimate the average number or the size of the impact. Let's denote the estimated number as $\boldsymbol{\beta}$.

When the estimated number (estimator), $\boldsymbol{\beta}$, is very different from the true number, **b**, for some systematic reasons, we call the estimator **biased**. For instance, if we sample people from a telephone book, we miss people who are not registered in the telephone book. If the people who are not on the telephone book are systematically different from the people on the telephone book, then an estimator by using data from the telephone-book sample would be biased from the true value of the entire population, including both people who are on the telephone book and who are not.

The estimator from the telephone-book sample, however, could be **precise**. For instance, as the number of sample from the telephone book increases, the estimator becomes more precisely estimated. But because the sample is not representing the entire population, the precisely estimated estimator is still biased. We will discuss more on this issue throughout this course.

Figure 1 summarizes the discussion on biasness and preciseness.



A: unbiased but not precise B: precise but biased C: unbiased and precise

Figure 1. Biasness and Preciseness

Source: Figure 2.2, page 29, in Lohr (1999)

Lecture 2: The Simple Regression Model

In this lecture we review the simple (bivariate) linear regression model. We focus on statistical assumptions to obtain unbiased estimators. A good understanding on the assumptions of the single linear regression model will help you to understand the importance of each assumption of the multivariate regression model.

Define the simple linear regression (SLR) model:

$$y = \beta_0 + \beta_1 x + u$$

where y is a dependent variable,
x is an independent variable (or a covariate), and
u is the error term (or disturbance).

We want to obtain unknown two parameters: β_0 *and* β_1. These two parameters are called *true values* that can not be observed. What we can do is to estimate these two parameters. We denote the estimated parameters as $\hat{\beta}_0$ *and* $\hat{\beta}_1$. We may call them as estimated coefficients or estimators. There are several ways to estimate the two parameters. The most common method is the least squared method.

Least Squared (Residuals) Method:

Define β_0 *and* β_1: (unknown) parameters or coefficients

$\hat{\beta}_0$ *and* $\hat{\beta}_1$: Least Squared estimates

b_0 and b_1: coefficients that are used to obtain $\hat{\beta}_0$ *and* $\hat{\beta}_1$

b_0 and b_1 are two variables that could take any values. We use these two variables while we are searching for $\hat{\beta}_0$ *and* $\hat{\beta}_1$. When the sum of squared residuals is minimized, b_0 and b_1 would be equal to $\hat{\beta}_0$ *and* $\hat{\beta}_1$, respectively. Define

$$y_i = b_0 + b_1 x_i + u_i$$

u_i is the residual.

Minimizing Problem: Find a pair of b_0 and b_1 that minimizes the sum of squared residuals:

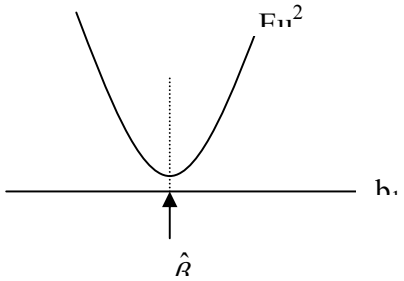
$$min \qquad \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

First order conditions (F.O.C.) with respect to (w.r.t) b_0 and b_1 are

$$\text{w.r.t. } b_0 \qquad -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad (1)$$

$$\text{w.r.t. } b_1 \qquad -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad (2)$$

Notice that b_0 and b_1 are replaced by $\hat{\beta}_0$ *and* $\hat{\beta}_1$ because the first order derivatives are set to be zero in (1) and (2).



From (1), we have

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} y_i - \hat{\beta}_0 \sum_{i=1}^{n} 1 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$n\bar{y} - n\hat{\beta}_0 - n\hat{\beta}_1\bar{x} = 0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \qquad (1')$$

From (2),

$$\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

By using (1´), we have

$$\sum_{i=1}^{n} x_i y_i - (\bar{y} - \hat{\beta}_1\bar{x}) n\bar{x} - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}) - \hat{\beta}_1 (\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) = 0$$

We can organize this to find $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Thus, the least squared estimates are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \qquad (3)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (4)$$

Unbiasedness of SLR

In this sub-section, we examine the unbaiasedness of the simple liner regression (**SLR**) estimates. To show the unbaiasedness, we need four assumptions, which are specified below. We need to understand why we need each assumption and how the SLR estimates will be biased when one of the four assumptions is violated.

Assumptions:

SLR 1 (Linear in parameters):	$y = \beta_0 + \beta_1 x + u$
SLR 2 (Random sampling)	x and y are random
SLR 3 (Zero conditional mean)	$E(u\|x) = 0$
SLR 4 (Sample variation in x)	$E(x_i - \bar{x})^2 > 0$

From (4), we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{(Here, we have used **SLR1.**)}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})\beta_0 + \sum_{i=1}^{n}(x_i - \bar{x})\beta_1 x_i + \sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\beta_0(n\bar{x} - n\bar{x}) + \beta_1\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (5)$$

At this point, we can see that: if the expected value of the second term is zero, then the expected value of $\hat{\beta}_1$ will be equal to β_1, the unknown true value.

By taking the expectation of the both sides of (5), we have

$$E(\hat{\beta}_1) = \beta_1 + \frac{\sum_{i=1}^{n}(E(x_i u_i) - E(\bar{x}u_i))}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (6)$$

Next, we have important steps:

(i) $E(\bar{x}u_i) = \bar{x}\,E(u_i)$ because $\bar{x}$ is just a constant.
(ii) By the definition, we know that $E(x_i u_i) = E(x_i)E(u_i) + Cov(x_i\, u_i)$.
Under **SLR3**, we assume $Cov(x_i\, u_i) = 0$, thus we have $E(x_i u_i) = E(x_i)E(u_i)$.

Thus, we have

$$E(\hat{\beta}_1) = \beta_1 + \frac{\sum_{i=1}^{n}(\bar{x}E(u_i) - \bar{x}E(u_i))}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \beta_1$$

Therefore under assumptions **SLR1-4**, the least squared estimate $\hat{\beta}_1$ is unbiased. When

$\hat{\beta}_1$ is unbiased, $\hat{\beta}_0$ ($= \bar{y} - \hat{\beta}_1 \bar{x}$) is also unbiased.

However, **SLR 3** is a very strong assumption especially in a single linear regression model because the disturbance, u_i, includes so many important omitted variables. When **SLR 3** is violated, we have an omitted variable problem. We will discuss about the omitted variables problem in the next lecture. For now, we simply assume that the four assumptions are satisfied.

Variances of OLS Estimators

For the OLS estimates to be the most efficient estimates among many other estimates, we need to add one more assumption:

SLR 5 (Homoskedasticity): $\text{Var}(u|x) = \Phi^2$

The homoskedasticity assumption indicates that the size of the variance of u is constant (or does not depend on x). When this assumption is violated, then we say the error term exhibits heteroskedasticity. Note that even under the heteroskedasticity, the least squared estimates are not biased as we have shown previously.

For now, we assume the homoskedasticity and obtain the variances of estimates. By using (4), we have

$$Var(\hat{\beta}_1) = E(\hat{\beta}_1 - \beta_1)^2$$

$$Var(\hat{\beta}_1) = E[\beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \beta_1]^2$$

From **SLR5**, $\text{Var}(u|x) = \Phi^2$

$$Var(\hat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^{n} (x_i{}^2 - \bar{x}^2)}{\left[\sum_{i=1}^{n} (x_i - \bar{x}\ )^2\right]^2}$$

$$= \frac{\sigma^2 \sum_{i=1}^{n} (x_i - \bar{x}\ )^2}{\left[\sum_{i=1}^{n} (x_i - \bar{x}\ )^2\right]^2}$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x}\ )^2} \qquad (7)$$

Note: (i) the larger the Φ^2, the larger the $Var(\hat{\beta}_1)$, and

(ii) the larger the variance in x_i, the smaller the $Var(\hat{\beta}_1)$.

Check the textbook page 56 for the variance of the estimated constant term, $Var(\hat{\beta}_0)$.

The Standard Error of $\hat{\beta}_1$

We can get the standard deviation of $\hat{\beta}_1$ by taking the (positive) squared root of the variance: $\text{sd}(\hat{\beta}_1) = +\sqrt{Var(\hat{\beta}_1)} = \dfrac{\sigma}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x}\ )^2}}$.

However, Φ is unknown. Thus, we have to estimate Φ. The unbiased estimator of Φ^2 is

$$\hat{\sigma}^2 = \frac{1}{(n-2)} \sum_{i=1}^{n} \hat{u}_i^2 = \frac{SSR}{(n-2)}.$$

This is adjusted by the (n-2) degrees of freedom. The degree of freedom is (n-2) because we have n observations with two estimators, which include an intercept.

Thus the standard error of $\hat{\beta}_1$ is

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{SSR/(n-2)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

The standard error for $\hat{\beta}_0$ can be obtained by replacing $\sigma^2 with\ \hat{\sigma}^2$ in equation (2.58) in the text.

R-squared

The R-squared is the fraction of the sample variation in y that is explained by x.

$$R = SSE / SST = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= 1 - SSR / SST = 1 - \frac{\sum_{i=1}^{n}\hat{u}_i^{\,2}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

SSE: the explained sum of squares; SSR: the residual sum of squares; SST: the total sum of squares. SST = SSE + SSR

Regression Through the Origin

Suppose we have a model such as

$$y = \beta_1 x + u.$$

Least Squared Approach:

$$min \qquad \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - b_1 x_i)^2$$

First order conditions (F.O.C.) with respect to (w.r.t) b_0 and b_1 are

$$\text{w.r.t. } b_1 \qquad -2 \sum_{i=1}^{n} x_i (y_i - \tilde{\beta}_1 x_i) = 0$$

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{8}$$

Lecture 3: Simple Omitted Variables Problem

In this lecture, we expand the simple linear regression model to multivariate liner model. The main motivation to include more than one variable into a regression model is that it is very difficult to identify the effect of one independent variable, x, on the dependent variable, y, without considering the other factors.

For instance, we can measure the effect of schooling more precisely if we compare people who have the exactly the same characteristics, except the schooling. (There have been many studies on twins.) If we find a difference in wage rates between the two identical people except the schooling, then we can conclude that the difference is due to the difference in the schooling. Unfortunately, in social science, it is impossible to find even two persons exactly the same except one factor. (Even twins are different in many ways.)

The problem of omitting important variables in regression analyses is called the omitted variables problem, and this is the core of many problems in econometrics. I start this lecture with the simplest case of the omitted variables problem.

Omitted Variables Problem: a simple case

Suppose that the true mode should be:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u. \qquad (3\text{-}1)$$

But suppose that we have information only on x_1 but not x_2 so that we can only estimate a simple bivariate model with x_1 but not x_2. According to the least squared approach from the previous lecture, we know that the least squared estimators are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 \qquad (3\text{-}2)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)y_i}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2} \tag{3-3}$$

However, the true model is (3-1). Thus, we can replace y_i in (3-3) by using the true model in (3-1),

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+u_i)}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}$$

As we did before, we can simplify the right hand side of the equation:

$$\hat{\beta}_1 = \frac{\beta_0\sum_{i=1}^{n}(x_{i1}-\bar{x}_1) + \beta_1\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)x_{i1} + \beta_2\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)x_{i2} + \sum_{i=1}^{n}(x_{i1}-\bar{x}_1)u_i}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}$$

$$= \frac{\beta_0 n(\bar{x}_1-\bar{x}_1) + \beta_1\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2 + \beta_2\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2) + \sum_{i=1}^{n}(x_{i1}-\bar{x}_1)u_i}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}$$

$$= \beta_1 + \frac{\beta_2\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2) + \sum_{i=1}^{n}(x_{i1}-\bar{x}_1)u_i}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}$$

Thus, by taking the expectation of both sides of the equation, the last term becomes zero under assumptions SLR 1-4, while the second term remains:

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2\frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2} \tag{3-4}$$

This indicates that the least squared estimate of β_1, will be biased if the second term of (3-4) is not zero, i.e., x_1 and x_2 are not correlated.

Note, further, that the last part of the second term is simply a least squared estimator of a simple bivariate regression model between x_1 and x_2:

$$x_2 = \delta_1 + \delta_2 x_1 + e$$

$$E(\tilde{\delta}_2) = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}$$

Thus, (3-4) can be written as

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \tilde{\delta}_2$$

Therefore the bias is $\beta_2 \tilde{\delta}_2$. Thus, the signs of β_2 *and* $\tilde{\delta}_2$ determine **the direction of the bias** (upward or downward). If the two parameters, β_2 *and* $\tilde{\delta}_2$, have the same signs (both positive or both negative), then the direction of bias is positive. But if the parameters have the opposite signs, then the direction of bias is negative.

For instance, let's say that we want to estimate the effect of agricultural credit on farm productivity but do not have information on the education of farmers. Thus, the error term of a bivariate model includes education as an omitted variable:

$$farm\_prod = \beta_0 + \beta_1 credit + (educ + u)$$

Because *educ* is correlated with *farm_prod* and *credit* positively, β_2 *and* $\tilde{\delta}$ are both positive. The direction of a potential bias is positive.

The magnitude of the bias depends on the absolute values of β_2 *and* $\tilde{\delta}$. If the correlations between y and x_2 and between x_1 and x_2 are strong, then the size of the bias becomes bigger. On the other hand, if one of the two correlations is weak the size of the bias becomes small.

Example 3-1: Individual Non-farm Income in Uganda

In rural areas of Uganda, non-farm income provides much needed cash to farm households. Most educated men and women in rural areas have opportunities to hold regular jobs, making a constant monthly wage throughout a year. Other people who are not fortunate enough to hold regular jobs are likely to earn non-farm income from small self-employed businesses such as making baskets or trading goods.

Suppose that we are interested in the gender differences in non-farm income and want to test a hypothesis that women make less non-farm income than men. But a simple comparison of non-farm income between men and women does not provide a reliable test for this hypothesis if characteristics of men and women are not similar.

One major factor in non-farm income is education. If men are better educated than women and men make more non-farm income than women, we can not be sure if the high non-farm income is a result of gender or education.

For example, let's use the data from Uganda. The data are collected by FASID in collaboration with Makerere University in Uganda in 2003. The data come from 940 households. Among them, we find 648 people who earned some income from non-farm activities. Table 1 indicates average non-farm income in US$, average schooling years, age, and observations for men and women.

Table 1. Individual Non-farm Income in Uganda

```
table female, c(mean nonincUS mean edu mean age n age) f(%8.1f) row;
------------------------------------------------------------------------------
   female | mean(nonincUS)       mean(edu)       mean(age)          N(age)
```

```
----------+-----------------------------------------------------------
        0 |          519.9             6.6            37.5             504
        1 |          387.1             6.1            35.9             144
          |
    Total |          490.4             6.5            37.1             648
-------------------------------------------------------------------------
```

As you can see in Table 1, men are slightly better educated and make more non-farm income. To compare non-farm income between men and women while holding education levels constant, we have created categories for education: no education (category 0), 1-4 years of schooling (1), 5-7 years of schooling (2), 8-11 years of schooling (3), and more than 12 years of schooling (4). Then we have calculated average non-farm income for each category for men and women separately. Although women make less non-farm income than men in general, t-tests indicate that the difference is statistically significant at the 5 percent level only for Category 3.

Table 2. Non-farm Income by Gender and Education

```
. table educat female, c(mean nonincUS) f(%8.0f) row;
----------------------
          |   female
   educat |    0     1    Difference (p-value)
----------+-----------
        0 |  200   139      60.7  (0.29)
        1 |  237   207      30.1  (0.72)
        2 |  510   430       79.9  (0.61)
        3 |  695   313      381.2  (0.05)
        4 |  981  1098     -116.9  (0.66)
    Total |  520   387      132.8  (0.09)
----------------------
```

Now, let's consider in which direction the bias would be if we estimate the non-farm income equation with a female dummy but not education. Education (the dependent variable) is positively correlated with education (the omitted variable), and education (the omitted variable) is negatively correlated with the female dummy (the independent

variable). Thus, according to the theory above, the direction of the bias should be negative.

$$ln(non\text{-}farm\ income) = \exists_0 + \exists_1 female + (educ + u)$$

```
. reg lnnoninc female

      Source |       SS       df       MS              Number of obs =     648
-------------+------------------------------           F(  1,   646) =   19.90
       Model |  43.8293159     1  43.8293159           Prob > F      =  0.0000
    Residual |  1423.03517   646  2.20284082           R-squared     =  0.0299
-------------+------------------------------           Adj R-squared =  0.0284
       Total |  1466.86449   647   2.2671785           Root MSE      =  1.4842

------------------------------------------------------------------------------
    lnnoninc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.6255663   .1402434    -4.46   0.000    -.9009543   -.3501782
       _cons |   12.95637   .0661114   195.98   0.000     12.82656    13.08619
------------------------------------------------------------------------------


. reg lnnoninc female edu

      Source |       SS       df       MS              Number of obs =     648
-------------+------------------------------           F(  2,   645) =   76.59
       Model |  281.510361     2  140.755181           Prob > F      =  0.0000
    Residual |  1185.35413   645  1.83775834           R-squared     =  0.1919
-------------+------------------------------           Adj R-squared =  0.1894
       Total |  1466.86449   647   2.2671785           Root MSE      =  1.3556

------------------------------------------------------------------------------
    lnnoninc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
      female |  -.5463488   .1282851    -4.26   0.000    -.7982557   -.2944418
         edu |   .1478726   .0130027    11.37   0.000     .1223398    .1734055
       _cons |   11.97554   .1052843   113.74   0.000      11.7688    12.18229
------------------------------------------------------------------------------
```

The results in the first model indicate that women make about 63 percent less non-farm income than men, but the results in the second model indicate that the difference is about 55 percent. Thus, the results indicate that the estimated coefficient **was biased downward** when *edu* was not included. Of course, there are probably many other variables that should be included in this model to estimate the gender difference more precisely. We will come back to this model later in this course.

End of Example