Primer on OLS and STATA

Urban Economics | Spring 2017

C. Luke Watson
Department of Economics
Michigan State University
East Lansing, MI
watso220@msu.edu

18 February 2017

Abstract

The purpose of this document is to introduce the basic theory of Ordinary Least Squares estimation and provide a tutorial on implementing and interpreting OLS in STATA to undergraduate economics students. After reading this document, students should know basic terminology and understand theory of OLS estimation, be able to import data and run OLS regressions in STATA, and interpret regression output information.

1 Motivation

Often we want to know why things happen. In political science, we might want to know why do certain laws get passed. In finance, we might want to know why do certain stocks have higher volitility. In urban economics, we might want to know some cities are growing or why some cities have high population densities. All of the above are examples of us wanting to know why some outcome happens.

Economic theory allows us to make predictions about outcomes using the logic of incentives and opportunity costs. However, theory only provides a prediction – a best guess – about why certain things happen. This is where **Econometrics** is used.

Econometrics is the combining of economic theory and statistics to test our predictions. First, we use economic theory to provide a clear and testible prediction. Next, we formulate an experiment that should allows us to test the prediction with collected data. Finaly, using statistics we analyze the results to see whether the prediction matches the data. In practice, the process is not as simple as the steps make it seem, but the spirit and guiding priciples are the same.

This lecture will first go over the theoretical support for the statisitical method **Ordinary Least Squares**. This is useful because it gives us confidence in our results and also shows us what we must be careful about when we do empirical econometric studies. Second, the lecture will provide an introductory tutorial on using **STATA** to conduct and evaluate an empirical study. This is useful for this class because you will be reading several academic economic articles that use these tools. Knowing these tools is also useful because being able to correctly use these econometric tools is a **skill** *that employers want and* **expect** *you to have*!

For this lesson, the example we will use is determining the size of a city based on economic variables. Suppose we have some economic model that predicts that the population of a city is increasing in the income of the city's residents; we can write this as

$$\uparrow I \implies P \uparrow$$

We will turn this prediction into a statistical model, then show how to derive the Ordinary Least Squares estimator to statistically test the significance of the estimate.

2 Ordinary Least Squares

The statistical method of OLS is essentially about finding how to make the least wrong prediction. For this lecture we are only going to consider a linear model of economic behavior. First, we will get some definitions and basic concepts. Next, we will derive the OLS estimator. Then, we will see why the OLS estimator is a good tool to use. Finally, we will see how we can evaluate the what the estimator tells us.

2.1 Basic Concepts

The estimator we will use calls for some basic knowledge of calculus and some important statistical concepts. The following definitions are meant to be a review to ensure we are all using the same terminology.

2.1.1 Calculus

Calculus is the mathematics of how functions change over their arguments. Recall that we are interested in how a city's population changes after a change in income, so calculus makes sense as a starting point.

Definition 2.1. Function: a function is a mathematical rule that links an independent variable to a dependent variable, where each IV is assigned to only one DV. Typically, we call the independent variable the argument of a function. Formally, $y = f(x)$ is a function as long as $y_1 = f(x) \wedge y_2 = f(x) \implies y_1 = y_2$. For our example, we are saying that Pop $= \mathbf{P}$(Income).

Definition 2.2. Multivariable Function: a multivariable function is a function with more than one argument. Formally, denote this as $y = f(x, z)$. For our example, we could specify Pop $= \mathbf{P}$(Income, Museums), where we think the number of museums also influences the population.

Definition 2.3. Derivative: the derivative of a function is a formula to determine the slope of the function. This tells us how the DV changes when the IV changes. Formally, we denote this is as $\frac{df(x)}{dx}$ or $f'(x)$.

Again, we are interested in how population changes as income changes. Once we create a formal model of the relationship, we can state our prediction in terms of a derivative. Our prediction $\uparrow I \implies P \uparrow$ can be stated generally as $\mathbf{P}'(\text{Income}) > 0$.

Definition 2.4. Partial Derivative: a partial derivative of a multivariable function is a formula for determining the change in the DV from a change from a single IV, while holding all the other IVs constant. Formally, this is denoted as $\frac{\partial f(x,z)}{\partial x} \equiv f_x(x, z)$ if we are taking the derivative with respect to x or $\frac{\partial f(x,z)}{\partial z} \equiv f_z(x, z)$ if the derivative is taken with respect to z.

Partial derivatives are typically taught in **Calculus 2**, but they are very easy. The idea is that we take a derivative with respect to the variable we care about, while holding all other variables constant. This is very similar to the *ceteris paribus* arguments typically heard in economics classes. Thus, if you want to find $\frac{\partial f(x,z)}{\partial x}$, simply treat z as a fixed number and take the derivative as normal.

Example 2.1. Consider the following example:
Let $\{x, z\}$ be variables, let $\{\alpha, \beta, \gamma, \delta, \epsilon\}$ be numbers (ex: $\alpha = 2,\ \beta = 77$).

$$f(x,z) = \alpha + \beta x^{\gamma} + \frac{\epsilon}{z^{\delta}} + xz$$
$$\frac{\partial f(x,z)}{\partial x} = \gamma\beta x^{\gamma-1} + z$$
$$\frac{\partial f(x,z)}{\partial z} = -\frac{\delta\epsilon}{z^{\delta-1}} + x$$

As you can see, in the second line z goes away since it is not related to x just like how α drops out; likewise, in the third line we treat x like α this time. ■

2.1.2 Statistics

We use statistics to assess theoretical predictions by testing whether the predictions are supported by the data. The most important concepts we need are the "error term" and "expected value" of a variable.

Definition 2.5. Statisitical Model: a statistical model is a formalization of theoretical relationship augmented by assumptions about probability distributions the variables.

Given our prediction, we construct the following statisitical model:

$$P = \alpha + \beta I + u$$

Note that this statistical model assumes that population (dependent variable) can be explained by a constant term α, a constant effect β from income (the independent variable), and a statistical error term u. Note: $u = P - \alpha - \beta I$.

The error term exists because we believe there are many factors that may explain a city's population other than the income of the residents – such as the number of museums. This invention is what allows us to bring in statistical theory for the estimation since specify/assume that u follows some probabily distribution – for example, the Normal(μ, σ^2) distribution. The statistical error term captures the randomness that we do not or cannot explain using our model. We can think of the error term as measuring 'how wrong our model is' compared to reality.

In almost all circumstances, we will make the assumption that on average the statistical error is zero – this amounts to the assumption that on average the model is correct. This assumption is mostly for convenience and does not affect most analysis. We will see later that our biggest worry is that the error term and the independent variable are correlated.

Definition 2.6. Estimator: an estimator is a rule for calculating an estimate once we have data from an experiment. A key distinction is that an estimator is a formula and an estimate is a specific number we get once we combine the formula and the data.

Example 2.2. The formula for a sample average of n observations is

$$\bar{x} = \frac{1}{n}\sum_{1=i}^{n}\{x_i\} = \frac{1}{n}(x_1 + x_2 + \cdots + x_{n-1} + x_n)$$

Once we have data – $\{2, 5, 1, 7, 100\}$ – we can apply the formula to the data and calculate our estimate.

$$\frac{1}{5}(1+2+5+7+100) = \frac{1}{5}(115) = 23$$

Thus $\bar{x}$ is our estimator that we use and 23 is the estimate that we calculate from the data. ■

Definition 2.7. Expected Value: the expected value of a random variable is the weighted sum of all possible values of the variable by the probability of each value's occurance; sometimes called the *mean* of a random variable. The expected value of a random variable is denoted mathematically as $\mathbf{E}[X]$

Example 2.3. Someone offers you a bet where a fair six-sided die is tossed. You win \$2 if the result is a "1," you win \$1 if the result is a "6," but otherwise you must pay \$1. What is the expected value of the bet? Given that the die is "fair" – meaning each number has an equal chance of occuring – you can calculate the expected value by caluclating the weighted average of the payments based on their probability of occuring – which is $\frac{1}{6}$ for each side.

$$\mathbf{E}[\text{Bet}] = \frac{1}{6}(\$2) + \frac{1}{6}(-\$1) + \frac{1}{6}(-\$1) + \frac{1}{6}(-\$1) + \frac{1}{6}(-\$1) + \frac{1}{6}(\$1) = \frac{1}{3} - \frac{2}{3} + \frac{1}{6} = -\frac{1}{6}$$

If someone offers you this bet then you expect to **lose money** on average! ■

We can use the expected value concept for any situation where there is a variable that has some randomness or uncertaintly about its value. This makes the expected value a valuable tool for the statistical model above that includes the statistical error term.

Definition 2.8. Variance: the variance of a random variable is a measure of how spread out the possible values of the variable are from its mean. The variance of a random variable is denoted mathematically as $\text{Var}(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2$.

A high variance essentially means that there is a greater likelihood that the variable will be far away from its mean, which makes the variable difficult to predict. In contrast, a low variance indicates that the variable stays close to its mean, so it will be easier to predict. We will use this concept in evaluating our estimator's statistical reliability.

Example 2.4. For the above example, we can calculate the varianc using the definition.

$$\begin{aligned}\text{Var}[\text{Bet}] &= \frac{1}{6}(\$2)^2 + \frac{1}{6}(-\$1)^2 + \frac{1}{6}(-\$1)^2 + \frac{1}{6}(-\$1)^2 + \frac{1}{6}(-\$1)^2 + \frac{1}{6}(\$1)^2 - (-\frac{1}{6})^2 \\ &= \frac{1}{6}(9) - \frac{1}{36} = \frac{53}{36} \approx 1.47\end{aligned}$$

This bet seems to have a large variance compared with its mean, which means its payout is somewhat unpredictable. Note that the probabilities are the same in the expected value and variance formulas – only the **values** of the variable are squared. ■

Definition 2.9. Covariance: the covariance of two random variables is a measure of how much the behavior of one random variable is related to the behavior of another. The covariance of two random variable is denoted mathematically as $\text{Cov}(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y]$.

Example 2.5. Some intuitive examples of covariance: (1) Positive Covariance: if the temperatute is high, then I am more likely to wear shorts, so $\text{Cov}(\text{Temperature}, \text{Wear Shorts}) > 0$; (2) Negative Covariance: if the temperatute is low, then I am more likely to wear a coat, so $\text{Cov}(\text{Temperature}, \text{Wear Coat}) < 0$; (3) Zero Covariance: regardless of the temperature, I am going to wear shoes, so $\text{Cov}(\text{Temperature}, \text{Wear Shoes}) = 0$. ■

Rules of $\mathbf{E}[]$, Var(), Cov()

1. $\mathbf{E}[a + bX + cY] = a + b\,\mathbf{E}[X] + c\,\mathbf{E}[Y]$
2. $\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y] + \mathsf{Cov}(X, Y)$
3. $\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X|Y]] \rightarrow$ this is the Law of Iterated Expectations
4. $\mathsf{Var}(a) = 0$
5. $\mathsf{Var}(bX + cY) = b^2\,\mathsf{Var}(X) + c^2\,\mathsf{Var}(Y) + 2bc\,\mathsf{Cov}(X, Y)$
6. $\mathsf{Var}(bX - cY) = b^2\,\mathsf{Var}(X) + c^2\,\mathsf{Var}(Y) - 2bc\,\mathsf{Cov}(X, Y)$
7. $\mathsf{Cov}(X, X) = \mathsf{Var}(X)$
8. $\mathsf{Cov}(X, a) = 0$
9. $\mathsf{Cov}(X, a + bY) = b\,\mathsf{Cov}(X, Y)$

Definition 2.10. t Test: This is a test to see if an estimate from a statistical model is statistically different from some theoretical predicted value. We can use this test to see if an estimate is equal, less than, or greater than some predicted value. Suppose $\hat{\theta}$ is our estimate and we want to test whether the estimate is significantly statstically different from a predicted value μ_0, then the test uses the following formula:

$$\hat{t} = \frac{\hat{\theta} - \mu_0}{\sqrt{\hat{\mathsf{Var}}(\hat{\theta})}}$$

This test is based on the **t-distribution** and by knowing the degrees of freedom (typically the number of observations minus the number of values we are estimating) which is very easily found online or in statistical tables, we can find out the "critical value" t_c. The critical value allows us to see if our prediction matches the data, such that if $\hat{t} < t_c$ then the estimate *is not* statisitically different from the hypothetical value, but if $\hat{t} > t_c$ then the estimate *is* statistically different.

2.2 OLS Estimator

Keeping our example, we believe that $\uparrow I \implies P \uparrow$ so we are using the statistical model $P = \alpha + \beta I + u$. We want to create estimators for (α, β) and we have a prediction that $\beta > 0$.

We are going to use the Ordinarly Least Squares estimator that is based on the "regression towards the mean" of a random variable – for this reason, we often say we are doing "regression analysis" or running "regressions."

We can derive the OLS estimator two ways: calculus and basic statistics. The calculus approach uses fewer assumptions and explains *why* we call it the "least squares" estimator. The statistical approach uses assumptions about how the data we observe is created (the statisitical model), but this method gives the most intuition about what the OLS estimator does.

2.2.1 Calculus Derivation of OLS

Assumption: $P = \alpha + \beta I + u$

Suppose we have n observations of populations and incomes of various randomly chosen cities. We want to minimize the total amount of errors – where the model 'misses' – from the model: $u = P - \alpha - \beta I$. Because some errors may be positive or negative, we cannot just add them up and be confident in our technique. Instead, we square the errors and then add them, since the square of a number is always positive: $a^2 \geq 0$. We want to find the estimator that minimizes the sum of squared errors – thus least squares estimator.

$$\max\{\mathbf{E}[(P-\alpha-\beta I)^2]\} \text{ w.r.t. } \{\alpha,\beta\}$$

$$\text{FOC}_\alpha : \frac{\partial}{\partial\alpha}\mathbf{E}[(P-\alpha-\beta I)^2] = \mathbf{E}[-2(P-\alpha-\beta I)] = 0$$
$$\implies \mathbf{E}[(P-\alpha-\beta I)] = \mathbf{E}[P] = \alpha + \beta\,\mathbf{E}[I]$$
$$\implies \alpha = \mathbf{E}[P] - \beta\,\mathbf{E}[I]$$

$$\text{FOC}_\beta : \frac{\partial}{\partial\beta}\mathbf{E}[(P-\alpha-\beta I)^2] = \mathbf{E}[-2(I)(P-\alpha-\beta I)] = 0$$
$$\implies \mathbf{E}[(I)(P)] = \alpha\,\mathbf{E}[I] + \beta\,\mathbf{E}[I^2]$$
$$\implies \mathbf{E}[(I)(P)] = (\mathbf{E}[P] - \beta\,\mathbf{E}[I])\,\mathbf{E}[I] + \beta\,\mathbf{E}[I^2] = \mathbf{E}[P]\,\mathbf{E}[I] - \beta\,\mathbf{E}[I]^2 + \beta\,\mathbf{E}[I^2]$$
$$\implies \mathbf{E}[(I)(P)] - \mathbf{E}[I]\,\mathbf{E}[P] = \beta(\mathbf{E}[I^2] - \mathbf{E}[I]^2)$$
$$\implies \mathsf{Cov}(I,P) = \beta\,\mathsf{Var}(I)$$
$$\implies \beta = \frac{\mathsf{Cov}(I,P)}{\mathsf{Var}(I)}$$

2.2.2 Statistical Derivation of OLS

Assumptions

1. $P = \alpha + \beta I + u$
2. $\mathbf{E}[u] = 0$
3. $\mathbf{E}[u|I] = \mathbf{E}[u]$

Consider taking the covariance of population and income:

$$\begin{aligned}\mathsf{Cov}(P,I) &= \mathsf{Cov}(\alpha+\beta I+u, I) = \beta\,\mathsf{Var}(I) + \mathsf{Cov}(u,I)\\ &= \beta\,\mathsf{Var}(I) + \mathbf{E}[Iu] - \mathbf{E}[u]\,\mathbf{E}[I] = \beta\,\mathsf{Var}(I) + \mathbf{E}[Iu]\\ &= \beta\,\mathsf{Var}(I) + \mathbf{E}[\mathbf{E}[Iu|I]] = \beta\,\mathsf{Var}(I) + \mathbf{E}[I\,\mathbf{E}[u|I]]\\ &= \beta\,\mathsf{Var}(I) + \mathbf{E}[I\,\mathbf{E}[u]] = \beta\,\mathsf{Var}(I) + \mathbf{E}[0]\\ &= \beta\,\mathsf{Var}(I)\end{aligned}$$
$$\implies \beta = \frac{\mathsf{Cov}(P,I)}{\mathsf{Var}(I)}$$

Consider taking the expected value of each side of the statistical model:

$$\begin{aligned}\mathbf{E}[P|I] &= \mathbf{E}[\alpha+\beta I+u]\\ &= \mathbf{E}[\alpha|I] + \mathbf{E}[\beta I|I] + \mathbf{E}[u|I]\\ &= \alpha + \beta\,\mathbf{E}[I] + \mathbf{E}[u]\\ &= \alpha + \beta\,\mathbf{E}[I]\end{aligned}$$
$$\implies \alpha = \mathbf{E}[P|I] - \beta\,\mathbf{E}[I]$$