

Econometrics – Introductory Comments*

Economics of Education (ECO838)

September 2011

This note provides some introductory background to the fascinating subject of econometrics. It is intended to supplement our more detailed class discussion.¹ (If anything is unclear, please have a word.)

1 Variables vs. parameters

We distinguish variables from parameters in an econometric model. To underline this distinction:

- *Variables* are typically observed characteristics, such as a worker's education level or age, that can vary from individual to individual and also (potentially) over time. In a data set (which stores the information available to the researcher), we will typically have a list of variables associated with a given individual (the latter is referred to as an observation). The more observations we have in the data set, the more precise our estimates will tend to be, as the sampling variability gets smaller.
- *Parameters* are assumed to be fixed and unchanging, and are (usually) unknown to the researcher. (In this case, the challenge is to uncover the unknown parameters – that is, to develop systematic ways to estimate them, given the data at hand. More on this type of exercise below.)

*Author: Robert McMillan (copyright), mcmillan@chass.utoronto.ca

¹If you are interested in a much fuller treatment of the issues discussed here, please see Jeffrey M. Wooldridge's excellent book, *Introductory Econometrics: A Modern Approach*.

For example, to illustrate the distinction using a simple econometric model, namely

$$h_i = \alpha + \beta w_i + \epsilon_i, \tag{1}$$

the observed variables are given by h_i (hours worked per week by worker i) and w_i (the net wage that worker i receives); the unobserved variable is given by ϵ_i (random noise); and α and β are unknown parameters, to be estimated.

2 Estimation

The goal of estimation is, as mentioned, to recover the underlying parameters that helped to generate the data that we observe (e.g. hours worked for different workers, given their net wages).

For that purpose, econometricians have proposed a variety of *estimators*: recipes for computing parameter estimates from a given data set. The most widely used estimator is – without a shadow of a doubt – the *least squares estimator*, the basic mechanics of which we will discuss in class.

The estimates of the underlying parameters that we obtain by applying our least squares estimation recipe are themselves *random variables*.² This is because the particular estimates depend, in part, on the random errors in the econometric model: if the random errors were different, so our parameter estimates would be also. This is the consequence of sampling variability. These errors serve to obscure the true nature of the systematic part of the econometric model, as it is not clear what is due to noise versus systematic factors.

Aside 1: Parameter Estimates are Random Variables

For a slightly longer explanation of this point – a point about the source of sampling variability – one can conduct what is called a Monte-Carlo study, in which we carry out a series of experiments, each time drawing a set (a vector) of random errors, one for each observation, from a known distribution (for instance, a Normal $(0, 1)$ distribution).

²This means that the values they can take are governed by a probability distribution, and prior to realization, we cannot say which values they will take on – that is random.

For each set of errors, we are able to create a fresh scatterplot, relating (say) the net wage to hours worked for a set of workers – our sample of workers. In turn, we are then able to apply the least squares recipe to fit a line of best-fit through the data points.

Thus, for G experiments, we would have G estimated lines, each with an estimated intercept and slope. If we plotted, say, all the G slope estimates in a histogram, we would be tracing out the sampling distribution of the least squares slope estimator. Note that the different slope estimates we obtain are entirely due to the differences in the error draws from experiment-to-experiment.

Aside 2: Properties of Estimators

Monte-Carlo methods are useful in order to establish the properties of a given estimator.³ For instance, we can check to see whether a given estimator is unbiased for the true parameter – that is, on average is equal to the same value – in the somewhat artificial setting where we fix the parameter values, generate the data according to the econometric model in question, then apply the least squares recipe, seeing if we approximately ‘get back’ our underlying parameters.

In the example we have been considering (the one that relates the net wage to hours worked), suppose we plotted the sampling distribution of the slope estimator. It is then straightforward to compute the sample mean of this distribution, and this can be compared – the neat thing – with the *known* parameter value that was used to generate the data.

One of the attractive properties of the least squares estimator is that the parameter estimates are unbiased: on average, they equal the true parameter value. So we would expect our sample mean of the G slopes we obtained to be close to zero. How close depends on the sampling variance of the estimator – something we can also estimate from the Monte-Carlo exercise.

³This is especially the case when the properties cannot be established analytically

3 Choice of Specification

If we have a rich data set on a sample of workers (let's say), we will have a number of different pieces of information characterizing each worker. If we want to explain the variation in the dependent variable, such as hours worked, we should try to include relevant variables in our regression. Which variables are relevant and how is the issue of *model specification*. As we will discuss in class, we can use hypothesis testing to help determine which specification represents the data best – particularly, which variables do not appear to be relevant.

Hypothesis testing – intuition

We have noted that the parameter estimates obtained from estimating an econometric model are random variables, and so vary according to their sampling distributions. If we know the underlying distribution of the random error, we can estimate this sampling distribution using Monte-Carlo methods. We can also work it out statistically, as we will see.

The issue before us: Is a parameter estimate sufficiently ‘close’ to zero that we can effectively treat it as zero, and so ignore the influence of the variable it attaches to?

This will depend on both the particular parameter value we have estimated and some measure of the spread of the sampling distribution (the sampling variance).

Here is the standard testing procedure:

1. Estimate the parameter and its standard error (the square root of the sampling variance – a measure of spread).
2. Form the test statistic: $z = \frac{b^{ols}}{SE(b^{ols})}$. This follows a $N(0, 1)$ distribution under the null hypothesis that the true coefficient is zero, and so this distribution is centred around zero.
3. If the test statistic is far away from zero – concretely, if it is on the far side of the relevant critical value – then it is unlikely that pure chance could have generated it under the null hypothesis. This would lead us to reject the null hypothesis.⁴

⁴If we are not sure (under the alternative hypothesis) whether the true parameter value should in

Aside: Calculus

A slightly technical point: How can we justify ignoring the error term when working out the effect of incrementally changing a variable on the dependent variable using calculus?

Take the following equation:

$$h_i = \alpha_0 + \alpha_1 w_i + \alpha_2 E_i + \epsilon_i. \quad (2)$$

For instance, if we want to compute the effect of changing the net wage incrementally on hours worked, we would take the following derivative: $\frac{dh_i}{dw_i}$. Under the assumption that the error term is uncorrelated with the explanatory variables (a standard assumption), we can be sure that changes in net wages will not have an indirect effect working through the error – that is what the uncorrelatedness assumption implies. So we can just ignore the error in this case.

fact be above or below zero, then we would carry out a two-sided test, in which case the rejection region consists of critical values below and above zero (± 1.96 , if the test is at the 5-percent level). For a one-sided test, the critical value would be 1.645 (at the 5-percent level).