EC221 Lecture 0 - what are we doing?

Tom Glinnan

January 2022

The LT of EC221 is fundamentally different to the MT. The point of the MT is to get you thinking the right way about causality, and how to approach causal questions. However, this can only take you so far - in order to use the tools well, we have to be able to understand when exactly we should use them. The LT of EC221 is all about doing this

The game works like this: whenever we see data, what we're really seeing is the outcome of some 'machine' which spits data out according to a rule[1]. If we see data x_i and y_i then the equation

$$y_i = 7 + 3x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0,1)$$

represents such a 'machine'. Namely, we pretend what is going on under the surface is that God or 'nature' is taking some value of x_i, drawing ε_i from a $N(0,1)$, and then putting them together in this way to get y_i. We then see x_i and y_i for many people i, but nothing about the machine itself. The real name for this 'machine' is the *data generating process* (DGP), and statistics is all about working out properties of the DGP from data, or as you have probably also heard it put, learning about the population[2] from samples. So much of what we know about the world can come down to solving this one problem

Assumptions

Before we dive in, we have to speak about one of the most important concepts: the need for assumptions. If we don't assume anything about the DGP, then there's essentially nothing we can say about it from our data. This is because the space of *all possible DGPs* is so huge that we're realistically never going to be able to pin down which sort of one it is without crazily high amounts of data[3]. As such we need to make assumptions. A typical one that will look familiar to you is

[1] We don't have to believe this is literally happening, but just that it's good way to think about randomness. The philosophy of probability is a deep (but fun) rabbit hole to go down if you want to feel confused!

[2] The concept of the population is another way to think about what's going on: instead of a DGP, we think of there being infinitely many data points, but we only see some of them. They're equivalent ways of describing the same thing - the machine effectively just picks which of the population to give to us in our sample; and the population is described by the equation of the machine. The population idea is more intuitive to some people, but the DGP idea more naturally leads to talking about assumptions, which is what we do here

[3] For those interested: 'big data' and machine learning more generally *do* try to do estimation with very few assumptions, as they have so much data (such as every button every person has ever pressed on Facebook). The world would be a lot simpler place if Economists had the same benefits for things we usually want to study

something like this:

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad \mathbb{E}(\varepsilon_i | x_i) = 0$$

You might hear the equation defining a DGP being referred to as a 'model'. If you've ever wondered what this equation **actually is** - it's expressing a property we think that the real world DGP has: namely that it can be written in the above form for some α, β and some distribution over the ε_i which satisfies $\mathbb{E}(\varepsilon_i|x_i) = 0$. To put it another way, we're assuming that the DGP that generated our data lives inside some set - namely, the set of DGPs that can be written this way - which is far smaller than the set of 'all' DGPs

Estimators

All an estimator is is a function that takes in some data, and gives you a real number. However complicated the function is, and whatever it's meant to do, they all can be thought of in this way. For example, if we see data x_i and y_i then we can define the *Bivariate OLS* estimator:

$$\hat{\beta}_{ols} = \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i^2} = \frac{\hat{Cov}(x_i, y_i)}{\hat{Var}(x_i)}$$

The expression is just a complicated function of x_i and y_i. The reason why we give it the symbol $\hat{\beta}_{ols}$ and talk about it as an estimator *for* β is because of a theorem: that if the DGP has the above form, $\hat{\beta}_{ols}$ is a 'good approximation' of β. We formalise this by defining several properties that an estimator can have that make it 'good' - two of which are unbiasedness and consistency[4]. The reason why we care about $\hat{\beta}_{ols}$ is the theorem is that if the DGP satisfies the above assumption, then $\hat{\beta}_{ols}$ is an unbiased and consistent estimator of β. To clarify: an *estimator* is the name of the *function* from data to $\mathbb{R}$; an *estimate* is this function evaluated at the data that we actually have. This is why we can talk about an estimator having a variance - we're effectively asking the question of how much our estimate would vary if we had sampled different data from the DGP than we actually did[5]

The Fundamental Trade-off

There's nothing at all stopping us from choosing any weird function of data and calling it an estimator, but this will nearly always be a terrible idea. To find out which estimators are good and bad in certain situations, we rely on theorems which have this logic:

> **If** the DGP satisfies assumptions ..., **then** our estimator has the properties of being ...

Then, these theorems help tell us which estimators are good and bad to use in a certain situation: **if** we think that we live in a world where we can write the DGP as $y_i = \alpha + \beta x_i + \varepsilon_i$ where

[4] Which we will explain later in the course

[5] This is related to the idea of statistical significance. The reason it exists as a concept is because we only see a sample, not the whole population (ie some data from the machine, not the machine itself). Intuitively, our estimate for a regression coefficient is statistically significant if we think it's far away from 0, and if we think that getting different data wouldn't have changed our estimate that much: giving us the test statistic $(\hat{\beta} - 0)/\hat{Var}(\hat{\beta})$. The larger this is, the more statistically significant $\hat{\beta}$ is

$\mathbb{E}(\varepsilon_i|x_i) = 0$, **then** OLS gives a 'good' estimate of α and β (where 'good' is formalised into the properties of unbiasedness and consistency)

In addition, we can rank assumptions by how unrealistic they are: if A and B are two assumptions and A implies B, then we say that A is a stronger assumption than B, and that B is weaker than A: whenever A holds, so must B, but B can hold sometimes when A doesn't. For example, as we will see later in the course, for OLS to be consistent we require that $\mathbb{E}(x_i\varepsilon_i) = 0$, which is a weaker assumption than $\mathbb{E}(\varepsilon_i|x_i) = 0$ (needed for unbiasedness)[6]. This comes from the fact that if $\mathbb{E}(\varepsilon_i|x_i) = 0$, then we must have that $\mathbb{E}(x_i\varepsilon_i) = 0$ - something we can prove. If it helps, think of 'stronger' as meaning 'more unrealistic': we can be less sure that it's going to hold. In the same vein, we will generally talk about how 'strong' or 'weak' assumptions are: strong means that we think they're probably not likely to hold in our given situation, and weak means we think they're probably likely to.

The past two paragraphs hint at the fundamental trade-off at the core of statistics: **if we choose a method which relies on stronger assumptions to work well, it will give us better estimation if those assumptions are true, but often worse estimation if the assumptions are false**. Methods which rely on weaker assumptions work alright in a wide range of situations, but never really give us the estimating power we would get if we used the other method in a situation it's designed for. For example, if the DGP satisfies the above assumption and also homoskedasticity (that the variance of ε_i is a constant for all i), then OLS has the bonus property of *efficiency*, so it will do very good estimation for β. Alternatively, if we break the assumption that $\mathbb{E}(\varepsilon_i|x_i) = 0$ then OLS may fail to be unbiased and consistent. The art of Econometrics is all about working out what we can realistically assume about the real world, then deciding based on that which method we should use.[7]

Something to note: when you do a regression in Stata, all you're doing is telling the computer to calculate some estimate based on the data you have. The command reg y x does bivariate OLS, which works well if the DGP can be written in the above form; reg y x z does multivariate OLS, which works well if the data has the form $y_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i$, with $\mathbb{E}(\varepsilon_i|x_i, z_i) = 0$. You're not

[6] As such, it is very possible that we have OLS being consistent, but biased, if $\mathbb{E}(x_i\varepsilon_i) = 0$ but $\mathbb{E}(\varepsilon_i|x_i) \neq 0$

[7] For those interested: assumptions have another, more subtle, role than just helping you get good estimation. If you think about it, β isn't defined at all in the case where we make no assumptions on the DGP, as we just have $y_i = f_i(x_i)$ - you can't really point to an object and say that that's the β. However, if we assume the DGP above, then we *can* talk about β as an object to be estimated, and explain why we might care if we found out its value. Some assumptions are therefore needed to help define and interpret the objects we want to estimate in the first place - before we even start to talk about estimation. This is known as the *identification problem*, and although it's very important in Econometrics it's not always relevant for many of the models we consider in LT (apart from the lecture on simultaneous equations), so we don't talk about it much. However, you've already seen me use it in class. I've said before about 'ignoring the error term' in defining how to interpret β in the above equation: formally, the assumption of $\mathbb{E}(\varepsilon_i|x_i) = 0$ allows me to write $\mathbb{E}(y_i|x_i) = \alpha + \beta x_i + \mathbb{E}(\varepsilon_i|x_i) = \alpha + \beta x_i$, and so $\beta = \partial\mathbb{E}(y_i|x_i)/\partial x_i$. Without an assumption such as this, it's difficult to even interpret what β means, let alone estimate it! Similarly, assumptions on a treatment being randomised allow us to interpret the coefficient in front of it as a causal effect, but without this assumption we can't do that

changing the DGP when you run a different regression - just doing methods which sometimes have nice properties and sometimes don't. Stata has no idea what you're assuming about the DGP - it just gives you a number. The regressions you run should be based on what you want to assume!

How does this link to the MT?

Though we didn't talk about it in this language, this is exactly what we've been doing in the course so far. In MT we said that OLS was a good estimator to use to analyse a randomised experiment, or if we think we've controlled for all of the confounders. In addition, we said that IV was a good method if the exclusion restriction and the relevancy condition hold. These are intuitive statements designed to get you to think about when you should use the methods of OLS and IV - in the LT we formalise these into proofs with the above form: under certain assumptions on the DGP (roughly corresponding to either being in a randomised experiment, or controlling for all confounders), OLS has certain nice properties, but under certain other assumptions OLS has bad properties. IV can have good properties when OLS doesn't, but equally it can be bad in other cases

We analyse lots of different estimation methods this way, and this provides a formal understanding of when certain methods are good and bad. However don't think that the MT material is unimportant: it's very easy to get lost in the math and the formalism of LT if you don't have a good feel for what certain assumptions mean, and when they're likely to hold. That is what the MT was all about, and it's something we'll spend a lot more time talking about in LT. To master Econometrics, you really need to be able to do both