Econometrics book

Ben Lambert

January 7, 2015

Contents

Part I

The basics

Chapter 1

How to best use this book

Chapter 2

What is econometrics?

Chapter 3

Estimators and their purpose

3.1 Chapter mission statement

After reading this chapter the student will understand the concept of an estimator, and its use in statistics.

3.2 The goal of this chapter

Statistical inference is the process of drawing conclusions about a population from a sample of observations. Estimators are the tools which statisticians use on samples of data; resulting in estimates of population-wide quantities which can be used to test hypotheses about the wider world. In this chapter the reader will be introduced to the concept of a *sampling distribution*, and how these can be examined to gauge the quality of an estimator.

3.3 What is an estimator, and why should we care?

We rarely in life have all relevant data available to us before we make a decision. When we decide where to go on holiday, we don't travel to a country, speak first-hand with locals, and taste the local food prior to deciding on a final destination for the family. We don't know perfectly

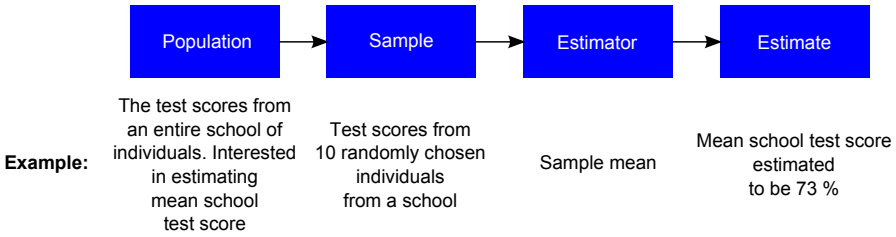

Figure 3.1: The estimation process.

what the weather will be like this afternoon when we get dressed for work. Instead, we extrapolate based on samples of data, and use these as predictive windows into the unknown. Statistical inference is the logical framework that allows us to make reasonable decisions based on our limited access to the facts.

These limited facts are what is known as a *sample*; by definition a subset of data from a *population* of interest. A population of interest might be the UK 18-30 females. However, it needn't just be applied to people, and could just as easily be defined to be the countries within the EU. All that matters is that the population, as a concept, is a wider entity on which we would like to make well-informed statements.

Estimators are tools that we use on samples that enable us to make *estimates* of quantities within the wider population being investigated. We then use these estimates to test hypotheses about the population, which allow us to better understand the functioning of the world around us.

Before we choose an estimator to use, we need to specify a *model* of the population which is proposed as a simplified and tractable representation of real life. Estimators can then be used to estimate components of these models, which when complete, can be deployed to help us explain some key features of a process we observe. Fortunately, statistical inference allows us to test the assumptions on which models are built, as well as interrogate their implications.

3.4 Models

> Essentially, all models are wrong, but some are useful.
>
> George Box

Real life in complex. It contains so many, seemingly-independently-moving things, that it can seem difficult/impossible to explain or predict anything effectively. To make life simpler for ourselves, we frequently build *models*. When we are planning how much to budget for electricity this month, we might think that electricity spend in the corresponding month last year might be a reasonable guide. When we are determining how long to spend cramming for an exam, we implicitly have an internal model of the trade off between test score and effort. When we set our alarm for the morning after, we believe that it will roughly take 10 minutes for us to snooze, and another 20 minutes to get showered and ready before leaving the house. These are all examples of models.

All models are abstractions from reality, which allow us to isolate, and concentrate on what we believe are the important parts of a system of interest. They are necessarily simplifications, and hence are not exactly representative of reality. However, as George Box's quote suggests, they can *sometimes* be useful.

The aforementioned are models which we most likely generate internally, but nonetheless are *implicitly* used to help us make decisions. Often for self-betterment, and to allow interrogation of thought however, we want to *explicitly* describe our models. The language of mathematics provides us with a logical framework with which to adequately describe these abstractions.

Before we talk through examples, it is worth reflecting on the various purposes for building models in the first place. Joshua Epstein in his article 'Why model?' lists amongst others, the following motivations for writing down *explicit* models:

- Prediction
- Explanation

- Guide data collection
- Discover new questions
- Bound outcomes to plausible ranges
- Illuminate uncertainties
- Challenge the robustness of prevailing theory through perturbations
- Reveal the apparently simple (complex) to be complex (simple)

There are no doubt other reasons, but I believe that the above covers the majority of rationales.

Imagine we are interested in determining the average test score for a particular standardised exam within the US. We don't have access to test scores for all individuals who take the exam in a specific year however. Instead we might suppose that test scores for individuals in the population are normally distributed about this theoretic mean, μ, with some theoretic variance σ^2:

$$IQ_i \sim \mathcal{N}\left(\mu, \sigma^2\right) \tag{3.1}$$

In (3.1), the i subscript on IQ represents the individuals in our population. So if we have a population size of 10,000, i runs from 1 to 10,000. It is important to stress that, whilst in this example, we can suppose that μ exists, this is not generally the case for quantities of interest in statistics/econometrics. The assumption inherent in the above model is that test scores are normally distributed around this tangible quantity with some theoretic variance. We do not actually believe that test scores are exactly normally distributed[1], however to make life easier and more palatable, and thus easier to deal with, we make this approximation.

Another model which we might choose to state is that there is a linear relationship the number of years of experience and the wage which an individual commands, on average:

$$wage_i = \alpha + \beta experience_i + \epsilon_i \tag{3.2}$$

[1]However, the central limit theorem provides some justification for making this approximation. See section 3.7.

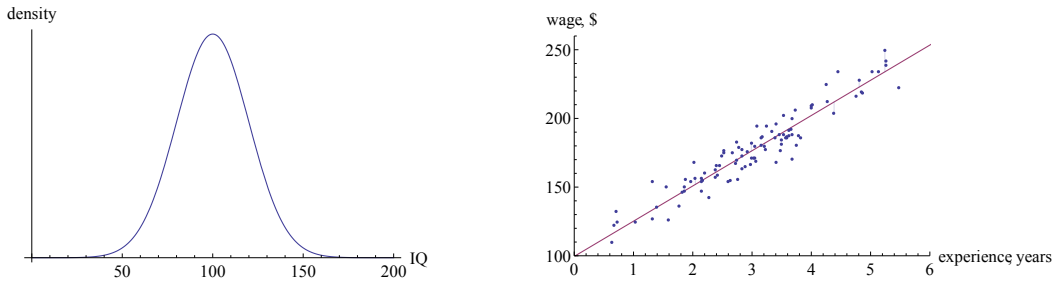

Figure 3.2: Left: the normal model for IQ. Right: the linear model between experience and wages, with the error terms ϵ_i indicated as vertical deviations from the straight line.

In (3.2), we have chosen a straight-line[2] relationship between the experience and wages, apart from a catch-all variable which encapsulates the various other idiosyncratic factors[3] which might impact the wage an individual obtains. For example, the type of job undertaken, the number of hours worked, or their level of education. We assume here that conditional on the level of experience for individuals, the average effect of these other factors is zero.

A model is only as good as the assumptions on which it rests. It might be that IQ test scores are more variable in the US population, than a normal distribution allows. It could be that there are diminishing returns to experience, meaning that the increment to wage for an extra year of work diminishes, dependent on the stage in the particular individual's career; invalidating the assumption of *linearity*.

Much like models assumptions are necessarily simplifying, and therefore *wrong*. However, there is a spectrum of *wrong*. An assumption is good so long as it captures the essence of reality sufficiently to allow the model to be used as is required. If the assumption is too *wrong*, then the model will cease to be useful, and we are forced to go back and examine its foundations.

[2]The model stated in (3.2) is, (with the exception of ϵ_i term), of the form $y = mx + c$, which is taught in high school.

[3]Individual attributes.

3.5 Sampling distributions

3.6 Good properties of an estimator

3.7 The central limit theorem

Part II

Cross sectional data: useful and important

Chapter 4

When to use Ordinary Least Squares?

Chapter 5

How to make conclusions - an introduction to hypothesis testing

Chapter 6

How to interpret regression results

Chapter 7

Testing the Gauss-Markov assumptions, and what to do if they are violated

Chapter 8

Instrumental variables: allowing inference in difficult circumstances

Chapter 9

Monte Carlo: How to test the quality of an estimator

Part III

Time series: harder to master, but necessary

Chapter 10

Why and how do we need to think about time series differently to cross sectional?

Chapter 11

The basic building blocks of time series models: autoregressive and moving averages

Chapter 12

Testing for stationarity and what to do with non-stationary data

Chapter 13

Cointegration: allowing for realism in time series models

Chapter 14

An introduction to models for real processes: partial adjustment and error-correction models

Part IV

Panel data: the best of both worlds

Chapter 15

The benefits of panel data

Chapter 16

Why do we need more estimators? An introduction to First Differences and Fixed Effects

Chapter 17

The poor relation: Random Effects

Part V

A simple new paradigm in estimation: Maximum Likelihood

Chapter 18

The flaws in the Linear Probability Model

Chapter 19

Beautifully simple: An introduction to Maximum Likelihood

Chapter 20

Draw conclusions by likelihood: the Wald, the Score and the LM tests