EC338 Pre-reading

Neil Lloyd

2023-09-18

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

1 Introduction

This e-book provides some pre-reading material for the *EC338: Econometrics 2 - Microeconometrics* module. In line with the Applied Microeconomics and Microeconometrics literatures, EC338 has evolved over time to focus on causal inference, the identification strategies that underpin various casual estimands, and their corresponding estimators.

As a result of this shift in focus, less time is given to more advanced material relating to the linear estimators used in this literature. As the relevant material is covered in *EC226: Econometrics 1*, you should be able to follow EC338 with a little revision and study. However, the change in notation may 'trip you up' and certain topics will be easier to understand with a richer understanding of linear estimators. For this reason, I have prepared this material as a brief (re-)introduction to ordinary least squares (OLS).

In EC338 we work extensively with dummy variables. As you know from EC226, estimating a linear regression model with discrete independent variables is relatively simple and the interpretation of the coefficient is *typically* straight-forward. However, the use of dummy variables requires a careful consideration of (perfect) collinearity and to understand collinearity (or rank conditions) it helps to think of data as a matrix, or system of column vectors. This becomes even more important when we start to consider models with various dimensions of fixed-effects.

These notes begin by revisiting the basic linear regression model and OLS estimator using vector notation. Next we revisit the properties of OLS, using this same notation, but without extensive proofs. Proofs should be available from a range of textbooks, including Wooldridge (2011). The material that will be least familiar to you will be the geometry of OLS. Here we will cover projection matrices and how they can be used to understand partitioned regression. Finally, we discuss dummy variables, their projections, and issues of colinearity.

Remember, this is still the same material covered in EC226, just using vector notation. In many instances this simplifies the notation, as summations over n and/or t can be replaced with a simple inner product of vectors. For example, consider the average of random variable Y_i,

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

If we define two $n \times 1$ column vectors,

$$\ell = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad \text{and} \qquad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

then the inner produce of these vectors is,

$$\langle \ell, Y \rangle = \ell' Y = 1 \cdot Y_1 + 1 \cdot Y_2 + \cdots + 1 \cdot Y_n = \sum_{i=1}^{n} Y_i$$

In addition,

$$\langle \ell, \ell \rangle = \ell' \ell = 1 \cdot 1 + 1 \cdot 1 + \cdots + 1 \cdot 1 = \sum_{i=1}^{n} 1 = n$$

Thus, the average can be expressed as,

$$(\ell' \ell)^{-1} \ell' Y = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

It turns out, this linear transformation of Y,

$$AY = (\ell' \ell)^{-1} \ell' Y$$

has a structure that is common to many linear estimators; a structure that derives from the projection matrix of the ℓ-vector: $P_\ell = \ell(\ell' \ell)^{-1} \ell'$. We will discuss this in Chapter Chapter **??** when we explore the geometry of OLS.

At the end of this e-book is a compendium on linear algebra basics. I may reference these during the e-book. The compendium is not exhaustive and does not include proofs. Please consult a linear algebra text book for further reading if you require.

The material in this e-book will not be examined directly, but will assist with your understanding of the material covered during term.

2 The Linear Regression Model

2.1 Vector notation

Most undergraduate textbooks discuss data in terms of random variables: the dependent or outcome variable (typically denoted Y_i) and various independent or explanatory variables $(X_{i1}, X_{i2}, \dots, X_{ik})$. There's nothing wrong with this language, but to understand the geometry of OLS we will need to think in terms of random *vectors*.

When working with cross-sectional data, we think of a random sample as a collection of n realizations of the same random variable. Each observation represents a different unit (e.g., individual, firm, classroom, etc.) and we typically add the assumption that the data is *i.i.d.* (independently and identically distributed) across units of observation. It makes no difference then if we think of this random sample as a collection of $n \times 1$ random vectors, where each *row* represents a different unit and the unit of observation is maintained across random vector.

For each unit you observe the outcome (Y_i) and a vector of explanatory variables (or regressors):

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ik} \end{bmatrix}$$

Take note of the ordering of the subscripts: the first denotes the unit of observation (i) and the second the number of the regressor $(j = 1 \dots k)$. The pair (Y_i, X_i) represents an *observation*, where Y_i is a single random variable and X_i a random (column) vector.[1] A collection of observations forms a sample.

You are, no doubt, familiar with the linear regression model. A simple univariate model is typically written as,

$$Y_i = \beta_1 + \beta_2 X_{i2} + \varepsilon_i$$

[1] In these notes, as in the remainder of EC338, I treat X_i as a column vector. Some texts, including Wooldridge (2011), will treat X_i as a row vector. This distinction is not significant, but will affect your notation. I will point this out at a later stage.