

2. Review of Econometrics

Erick F Molina

Summer, 2024

1 Introduction

Once we have reviewed certain concepts from probability and inference, we can move on to another type of analysis. Now we are interested in the relationship between two random variables Y and X . In this section we will review how we can establish this relationship and some other questions that will arise as we try to understand it.

2 The linear regression model

To try to understand the relationship between Y and X , we will assume that we have a sample where both are i.i.d. Let us begin by stating a basic model. In this model, the relationship is linear and we have a term that represents the estimation error or the unexplained part of the model:

$$Y_i = \beta_0 + \beta_1 X_i + U_i \tag{1}$$

The parameters of the model are β_0 and β_1 . We will call the first coefficient (β_0) the *intercept*, and the second (β_1) the *slope*. As we can see, this model is a linear equation of the form $y = mx + b$, just as we learned in calculus classes. These coefficients describe a population and our objective is to estimate their values. We must remember that we never observe these parameters, because they are not data. The only thing we can do, then, is estimate them using the data and certain assumptions. The assumptions we use have to be credible for the model to try to reflect reality.

One of the assumptions we will make involves the expected value of the error term. Specifically, $\mathbb{E}[U \mid X] = \mathbb{E}[U]$ for all X . In other words, U is independent of X . In addition, we will assume that $\mathbb{E}[U] = 0$ for all U . Combining both assumptions, we obtain $\mathbb{E}[U \mid X] = 0$, which in turn implies that

$$\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X \quad (2)$$

This result is called the Conditional Expectation Function (CEF), and we will be working with it.

3 The identification problem

An important problem in econometrics is the problem of *identification*. An identified parameter is one that is determined uniquely by the distribution of the observed variables. Relatively, it is straightforward in the context of conditional and unconditional expectation, but it is important to explore it in order to be clear about it.

Suppose that F denotes the distribution of the observed data, for example, the distribution of the pair (Y, X) . Suppose that $\mathcal{F}$ is a set of distributions F and θ is a parameter of interest, for example the expected value $\mathbb{E}[Y]$.

Definition 1. A parameter $\theta \in \mathbb{R}$ is identified in $\mathcal{F}$ if for every $F \in \mathcal{F}$, there is a uniquely determined value of θ . This means that, given a dataset, it is possible to deduce the value of the parameter uniquely and without ambiguity.

We can illustrate it as follows: Suppose we have a model where the data y are distributed as $N(\mu, \sigma^2)$. If we only consider the mean μ and keep the standard deviation fixed, different values of μ produce different distributions. Therefore, we can uniquely identify μ from the observed data.

By contrast, imagine a model in which the data y are generated as the product of two parameters θ_1 and θ_2 , that is, $y = \theta_1 \cdot \theta_2$. Here, if we have a single value of y , we cannot distinguish among different combinations of θ_1 and θ_2 that produce the same product (for example, $y = 6$ could be the result of $\theta_1 = 2$ and $\theta_2 = 3$, or $\theta_1 = 1.5$ and $\theta_2 = 4$). Therefore, the parameters are not identified.

It would seem that identification is a general property of parameters as long as we exclude *degenerate* cases. In other words, if there are no data points that are very different from the rest, then we will be able to identify the parameters. This is true for moments of observed data, but not necessarily for more complex models.

4 Ordinary Least Squares (OLS)

Once we have the model, the next question is how we estimate the coefficients. To do so, we will use the ordinary least squares method, OLS. This method consists of minimizing the squared estimation errors U_i .

The results of the minimization problem yield the following:

$$\begin{aligned}\beta_0 &= \mathbb{E}[Y] - \beta_1 \mathbb{E}[X] \\ \beta_1 &= \frac{\mathbb{E}[YX] - \mathbb{E}[Y]\mathbb{E}[X]}{\mathbb{E}[X^2] - \mathbb{E}[X]^2} = \frac{Cov(X, Y)}{Var(X)}\end{aligned}\quad (3)$$

From here we can obtain the estimators using the sample analog estimators that we learned in the inference review.

$$\begin{aligned}\hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n X_i \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n Y_i X_i - \sum_{i=1}^n Y_i \sum_{i=1}^n X_i}{\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2}\end{aligned}\quad (4)$$

Now that we have obtained the results, what happens when we want to understand the relationship between Y and a set of variables X , instead of a single variable? The OLS problem extends to a linear algebra problem. Now we have to minimize the following:

$$\min_{\beta} \mathbb{E}[Y - X'\beta]^2 \quad (5)$$

Where X is a set of variables. The minimization problem gives us as a result that

$$\beta = \mathbb{E}[XX']^{-1} \mathbb{E}[XY] \quad (6)$$

To estimate the parameter, we use its sample analog:

$$\hat{\beta} = \sum_{i=1}^n (X_i X_i')^{-1} \sum_{i=1}^n (X_i Y_i) \quad (7)$$