LIST OF FIGURES

PART I. THE MULTIPLE LINEAR REGRESSION 19

PART II. WORKSHOPS 139

LIST OF TABLES

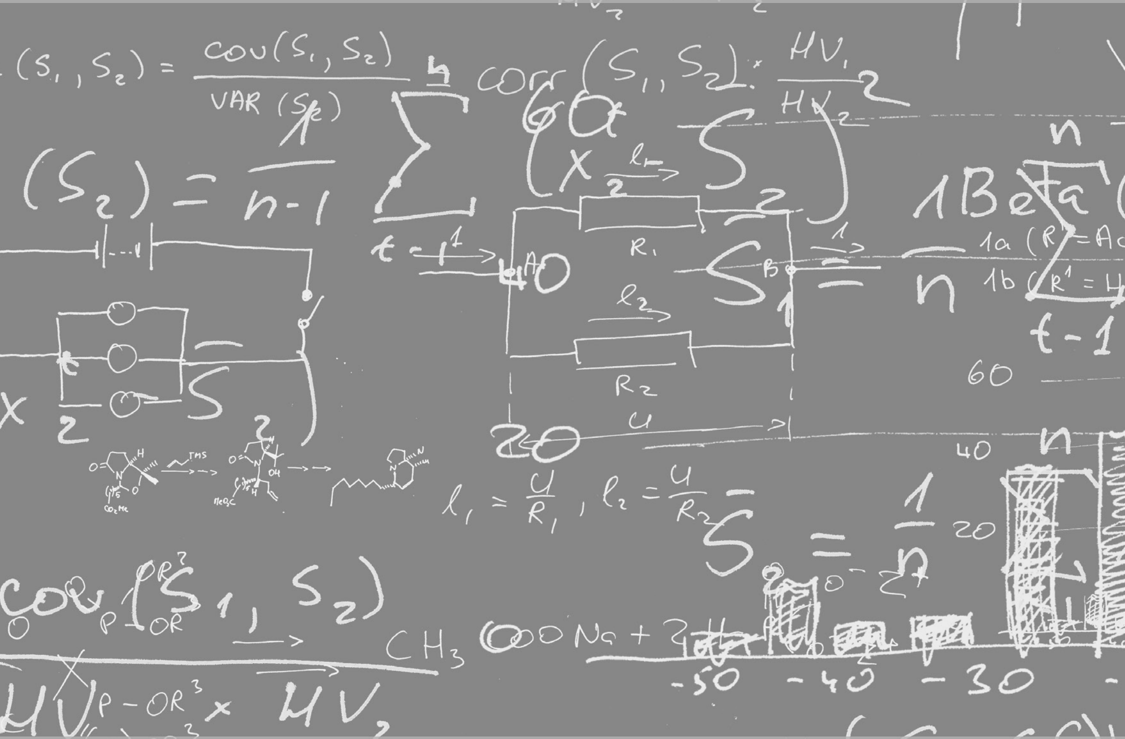
PART II. WORKSHOPS

ABSTRACT

Students of econometrics have access to a high number of excellent textbooks on the matter, such as the classic works of Damodar N. Gujarati [Gujarati, 2009], Jeffrey M. Wooldridge [Wooldridge, 2015], William H. Greene [Greene, 2003] or, for spanish readers, Alfonso Novales [Novales Cinca, 1988], which we totally recommend for further learning. However, these works include advanced econometric applications and, sometimes, complex mathematical notation, which exceeds the scope of an introductory subject. For this reason, we have decided to prepare this students oriented introductory work on the Multiple Linear Regression, one of the most basic and fundamental tools in econometrics.

This book contains the necessary notes for an introductory course of no longer than a semester, assuming only basic knowledge on statistics and probability from the student. The book is divided in two main parts. First, we introduce and explain the Multiple Linear Regression together with its main properties, applications and associated problems. In the second half, seven different exercises are proposed as applications for each one of the units in the first half. These exercises make use of different public access data bases and can be solved with help of the econometric software OxMetrics (version 7.0 or latter) and its internal package PcGive [Doornik, 2009].

Part I



The Multiple Linear Regression

INITIAL CONCEPTS

0

0.1 DEFINITION

Econometrics is the study of statistical structures that allow us to analyze an economic variable using other explicative economic variables. This is a process of modeling, of understanding a variable as a function of others, i.e., of explaining some set of observations as being related to some other observations.

The process of econometric modeling involves several steps and considerations:

1. Structure of the Model: explaining and formalizing the statistical relation between a set of variables implies assuming a mathematical relation between them. In this subject we will limit ourselves to linear (or linearizable) models.
2. Estimation: statistical relations imply fitting our previously defined model to the observed data. This process is known as model

estimation and completes the model expression while providing relevant economic information.

3. Uncertainty: all statistical models carry a certain degree of uncertainty due to the stochastic nature of economic data[1]. This implies that no perfect model can be achieved, as we would expect from a deterministic mathematical relation that estimates all observations with perfect numerical precision. Uncertainty introduces the error or residual term in the model.
4. Prediction: satisfactory econometric models can be used as predictors and give us a reliable estimation when we can assume the previous uncertainty.
5. Analysis tool: satisfactory econometric models can be used to understand economic realities and make decisions based on this generated knowledge.

   Information in an econometric model:

Data (the observations):	the reality
Selected Model (the structure):	the assumptions
Estimations (parameters and predictions):	the conclusions

0.2 THE DATA

Econometrics makes use of three different kinds of data:

— Cross section: several economic variables observed at the same moment over different individuals or places.

[1] Data is stochastic when it is produced by a random variable, i.e., it has a probability distribution or pattern that may be analyzed statistically but cannot be predicted precisely. It is the opposite to the deterministic property of mathematical objects.

- Time series: several economic variables observed over time.
- Panel datasets: several economic variables observed over time on different individuals or places.

The Multiple Linear Regression explained in this book is mostly used on cross section data.

0.3 THE MODEL

The process of econometrics modeling involves two different elements: a statistical relation and data. With these elements, we can analyze the observed economic reality:

$$\left.\begin{array}{r}\text{Data}\\[1em]\text{Selected Model}\end{array}\right\}\rightarrow\begin{array}{c}\text{Econometric}\\[1em]\text{methods}\end{array}\rightarrow\begin{array}{c}\text{Estimated}\\[1em]\text{values}\end{array}\rightarrow\left\{\begin{array}{l}\text{Prediction}\\[1em]\text{Economy description}\end{array}\right.$$

The general expression of a model always has to follow the structure

$$Y = \text{THEORY}(\beta, X) + \text{RESIDUALS}$$

where Y is the observed economic phenomena we want to explain, THEORY is the theoretical assumptions on the data, such as the model structure, β is the set of estimated parameters of the model, X is the matrix of explicative variables (the observed data we use to explain Y) and the RESIDUALS contain the uncertainty present in the model. We define the DATA as (Y, X).

The quality of the model depends on:

— Prediction horizon: interpolations are usually more reliable than extrapolations. Distant predictions are less reliable.

— Consistency of the estimators: we say an estimator is consistent when small variations on the explicative variables do not strongly affect their values.

— Quality of the estimators: how well do the model describes the data given our estimated parameters.

— Suitability of the model: the model has to be correctly stated, the specification of the model depends on the economic reality we are trying to describe.

These two last items can be understood as two sides of the same coin. When estimating a model our aim is to predict the observed data with the highest possible accuracy. This may lead us to very complex models adapting its shape to all subtleties and oscillations in your data. However, this is incorrect (see Fig. 0.1). A model should never pretend to describe the stochastic oscillations of data (those produced by inherent randomness of variables) since they are purely random, and therefore impossible to predict. On the contrary, the model should only describe the trend responsible of the observed data. Therefore, the model structure has to be stated before estimating it: this is a decision that we have to do based exclusively on theory, not in data. We use our observations to validate the estimated model. Once the model is correctly stated and estimated, the stochastic oscillations will generate the residuals.

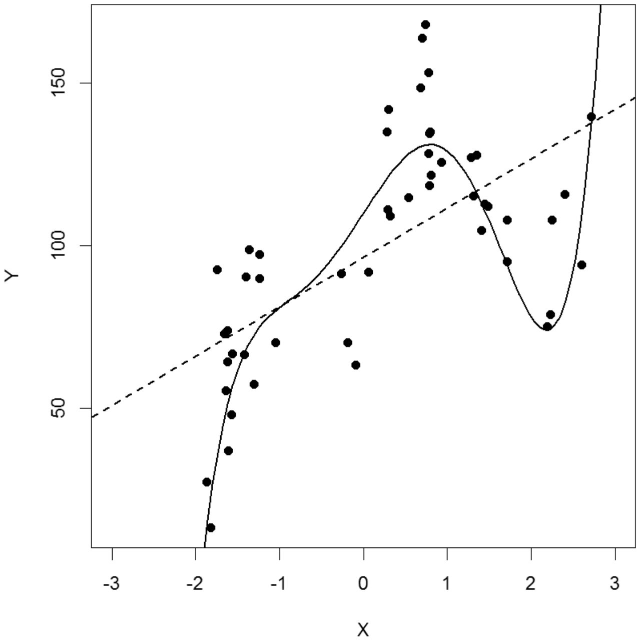


Figure 0.1
Dataset estimated with two different models. The solid line model is a quintic polynomial, while the dashed line model is a linear model. Even if the first one seems to be more precise in its estimations, such a complex model is uncommon in economics and we may prefer the lineal one.

0.4 LINEARITY

We say a model is linear when it can be expressed as a linear relation:

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k \tag{1}$$

In this model all variables X_i are different and the order of all of them is 1 (in principle, we have nothing like X_i^2 or $X_i X_j$). In two dimensions, this is the equation of a straight line: $Y = \alpha + \beta X$.

Sometimes non linear models can be linearized, this is, transformed in something that looks like eq. 1.

When the variables appear as functions of variables:

$$f(Y) = \beta_1 + \beta_2 g(X_2) + \ldots + \beta_k g(X_k) \tag{2}$$

we redefine these functions as: $Z = f(Y)$ and $W_i = g(X_i)$ and re-express:

$$Z = \beta_1 + \beta_2 W_2 + \ldots + \beta_k W_k \tag{3}$$

which satisfies the linear structure. This includes the case where $g(X_i) = X_i^2$, which allows us to work with variables of order greater than 1.

Another common example is the Cobb-Douglas production function [Douglas and Cobb, 1928]:

$$Y = AK^{\alpha} L^{\beta} \tag{4}$$

where K and L are our capital and labor variables. This equation can be linearized taking logarithms:

$$log(Y) = log(A) + \alpha \log(K) + \beta \log(L) \tag{5}$$

and then applying the previous substitutions. As we will see, logarithmic models are commonly used in economics.

However, not all models can be linearized. An example of this is:

$$Y = \alpha + \frac{\beta X}{Z - \gamma Z^2} \tag{6}$$

There is no mathematical way of expressing this formula into a linear shape. As said, we limit the scope of this work to linear models.

0.5 BASIC PROPERTIES AND DEFINITIONS

Here we define some basic properties and definitions that should be known by the students before starting the course:

VECTORS

A vector is defined as a column of values, and generally stated with capital letters:

$$V = \begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix}$$

therefore, a transposed vector (V^t) is a row, and products can be defined in two different ways:

$$V^t \cdot V = (v_1, \ldots, v_N) \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix} = \sum_{i=1}^{N} v_i^2$$

$$V \cdot V^t = \begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix} \cdot (v_1, \ldots, v_N) = \begin{pmatrix} v_1^2 & \ldots & v_1 \cdot v_N \\ \vdots & \ddots & \vdots \\ v_N \cdot v_1 & \ldots & v_N^2 \end{pmatrix}$$

TRANSPOSED OF A MATRIX PRODUCT

$$(AB)^t = B^t A^t \tag{7}$$

INVERSE OF A MATRIX PRODUCT

$$(AB)^{-1} = B^{-1} A^{-1} \tag{8}$$

EXPECTANCY AND ARITHMETIC MEAN

$$E(X) = \mu \simeq \bar{X} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{9}$$

VARIANCE AND SAMPLE VARIANCE

$$Var(X) = E((X - \mu)^2) = \sigma^2 \simeq s^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{X})^2 \tag{10}$$

The standard deviation of a variable is the square root of the variance: $\sigma = \sqrt{Var(X)}$.