Lecture Notes in Introductory Econometrics

Academic year 2017-2018

Prof. Arsen Palestini

MEMOTEF, Sapienza University of Rome

Arsen.Palestini@uniroma1.it

Contents

Chapter 1

Introduction

The present lecture notes introduce some preliminary and simple notions of Econometrics for undergraduate students. They can be viewed as a helpful contribution for very short courses in Econometrics, where the basic topics are presented, endowed with some theoretical insights and some worked examples. To lighten the treatment, the basic notions of linear algebra and statistical inference and the mathematical optimization methods will be omitted. The basic (first year) courses of Mathematics and Statistics contain the necessary preliminary notions to be known. Furthermore, the overall level is not advanced: for any student (either undergraduate or graduate) or scholar willing to proceed with the study of these intriguing subjects, my clear advice is to read and study a more complete textbook.

There are several accurate and exhaustive textbooks, at different difficulty levels, among which I will cite especially [4], [3] and the most exhaustive one, *Econometric Analysis* by William H. Greene [1]. For a more macroeconomic approach, see Wooldridge [5, 6].

For all those who approach this discipline, it would be interesting to 'define it' somehow. In his world famous textbook [1], Greene quotes the first issue of *Econometrica* (1933), where Ragnar Frisch made an attempt to characterize Econometrics. In his own words, the Econometric Society should ' promote studies that aim at a unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems'. Moreover: 'Experience has shown that each of these three viewpoints, that of Statistics, Economic Theory, and Mathematics, is a necessary, but not a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the *unification* of all three that is powerful. And it is this unification that constitutes Econometrics.'.

Although this opinion is 85 years old, it is perfectly shareable. Econometrics relies upon mathematical techniques, statistical methods and financial and economic expertise and knowledge. I hope that these lecture notes will be useful to clarify the nature of this discipline and to ease comprehension and solutions of some basic problems.

Chapter 2

The regression model

When we have to fit a sample regression to a scatter of points, it makes sense to determine a line such that the residuals, i.e. the differences between each actual value of y_i and the correspondent predicted value $\hat{y}_i$ are as small as possible. We will treat separately the easiest case, when only 2 parameters are involved and the regression line can be drawn in the 2-dimensional space, and the multivariate case, where $N > 2$ variables appear, and N regression parameters have to be estimated. In the latter case, some Linear Algebra will be necessary to derive the basic formula. Note that sometimes the independent variables such as x_i are called **covariates** (especially by statisticians), **regressors** or **explanatory variables**, whereas the dependent ones such as y_i are called **regressands** or **explained variables**.

Basically, the most generic form of the linear regression model is

$$y = f(x_1, x_2, \ldots, x_N) + \epsilon = \beta_1 + \beta_2 x_2 + \cdots + \beta_N x_N + \epsilon. \tag{2.0.1}$$

We will use α and β in the easiest case with 2 variables. It is important to briefly discuss the role of ϵ, which is a **disturbance**. A disturbance is a further term which 'disturbs' the stability of the relation. There can be several reasons for the presence of a disturbance: errors of measurement, effects caused by some indeterminate economic variable or simply by something which cannot be captured by the model.

2.1 Ordinary least squares (OLS) estimation method: two-variable case

In the bivariate case, suppose that we have a dataset on variable y and on variable x. The data are collected in a sample of observations, say N different observations, on units indexed by $i = 1, \ldots, N$. Our aim is to approximate the value of y by a linear combination $\hat{y} = \alpha + \beta x$, where α and β are real constants to be determined. The i-th **square residual** e_i is given by

$$e_i = y_i - \hat{y}_i = y_i - \alpha - \beta x_i,$$

and the procedure consists in the minimization of the sum of squared residuals. Call $S(\alpha, \beta)$ the function of the parameters indicating such a sum of squares, i.e.

$$S(\alpha, \beta) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \alpha - \beta x_i)^2 . \tag{2.1.1}$$

The related minimization problem is unconstrained. It reads as

$$\min_{\alpha,\ \beta} S(\alpha, \beta), \tag{2.1.2}$$

and the solution procedure obviously involves the calculation of the first order derivatives. The first order conditions (FOCs) are:

$$-2 \sum_{i=1}^{N} (y_i - \alpha - \beta x_i) = 0 \qquad \Longrightarrow \qquad \sum_{i=1}^{N} y_i - N\alpha - \beta \sum_{i=1}^{N} x_i = 0.$$

$$-2 \sum_{i=1}^{N} (y_i - \alpha - \beta x_i)\, x_i = 0 \qquad \Longrightarrow \qquad \sum_{i=1}^{N} x_i y_i - \alpha \sum_{i=1}^{N} x_i - \beta \sum_{i=1}^{N} x_i^2 = 0.$$

After a rearrangement, these 2 equations are typically referred to as **normal equations** of the 2-variable regression model:

$$\sum_{i=1}^{N} y_i = N\alpha + \beta \sum_{i=1}^{N} x_i, \tag{2.1.3}$$

$$\sum_{i=1}^{N} x_i y_i = \alpha \sum_{i=1}^{N} x_i + \beta \sum_{i=1}^{N} x_i^2. \tag{2.1.4}$$

Solving (2.1.3) for α yields:

$$\alpha = \frac{\sum_{i=1}^{N} y_i - \beta \sum_{i=1}^{N} x_i}{N} = \overline{y} - \beta \overline{x}, \tag{2.1.5}$$

after introducing the arithmetic means: $\overline{x} = \dfrac{\sum_{i=1}^{N} x_i}{N}$, $\overline{y} = \dfrac{\sum_{i=1}^{N} y_i}{N}$.

Plugging (2.1.5) into (2.1.4) amounts to:

$$\sum_{i=1}^{N} x_i y_i - (\overline{y} - \beta\overline{x})N\overline{x} - \beta\sum_{i=1}^{N} x_i^2 = 0,$$

hence β can be easily determined:

$$\sum_{i=1}^{N} x_i y_i - N\overline{x}\cdot\overline{y} + \beta\left(N\overline{x}^2 - \sum_{i=1}^{N} x_i^2\right) = 0 \qquad \Longrightarrow$$

$$\Longrightarrow \qquad \beta = \frac{\sum_{i=1}^{N} x_i y_i - N\overline{x}\cdot\overline{y}}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2}, \tag{2.1.6}$$

and consequently, inserting (2.1.6) into (2.1.5), we achieve:

$$\alpha = \overline{y} - \frac{\overline{x}\sum_{i=1}^{N} x_i y_i - N\overline{x}^2\cdot\overline{y}}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2}. \tag{2.1.7}$$

The regression line is given by:

$$\hat{y} = \alpha + \beta x, \tag{2.1.8}$$

meaning that for each value of x, taken from a sample, $\hat{y}$ predicts the corresponding value of y. The residuals can be evaluated as well, by comparing the given values of y with the ones that would be predicted by taking the given values of x.

It is important to note that β can be also interpreted from the viewpoint of probability, when looking upon both x and y as random variables. Dividing numerator and denominator of (2.1.6) by N yields:

$$\Longrightarrow \qquad \beta = \frac{\dfrac{\sum_{i=1}^{N} x_i y_i}{N} - \overline{x}\cdot\overline{y}}{\dfrac{\sum_{i=1}^{N} x_i^2}{N} - \overline{x}^2} = \frac{Cov(x,y)}{Var(x)}, \tag{2.1.9}$$

after applying the 2 well-known formulas:

$$Cov(x,y) = \mathbb{E}[x\cdot y] - \mathbb{E}[x]\mathbb{E}[y], \qquad Var(x) = \mathbb{E}[x^2] - (\mathbb{E}[x])^2\,.$$

There exists another way to indicate β, by further manipulating (2.1.6). Since

$$\sum_{i=1}^{N} x_i y_i - N\overline{x}\cdot\overline{y} = \sum_{i=1}^{N} x_i y_i - N\overline{x}\cdot\overline{y} + N\overline{x}\cdot\overline{y} - N\overline{x}\cdot\overline{y} =$$

$$= \sum_{i=1}^{N} x_i y_i - \overline{x} \sum_{i=1}^{N} y_i - \overline{y} \sum_{i=1}^{N} x_i + N\overline{x} \cdot \overline{y} = \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})$$

and

$$\sum_{i=1}^{N} x_i^2 - N\overline{x}^2 = \sum_{i=1}^{N} x_i^2 + N\overline{x}^2 - 2N\overline{x} \cdot \overline{x} = \sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} \overline{x}^2 - 2\overline{x} \sum_{i=1}^{N} x_i =$$

$$= \sum_{i=1}^{N} (x_i - \overline{x})^2,$$

β can also be reformulated as follows:

$$\beta = \frac{\sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N} (x_i - \overline{x})^2}. \tag{2.1.10}$$

The following Example illustrates an OLS and the related assessment of the residuals.

Example 1. *Consider the following* 6 *points in the* (x, y) *plane, which correspond to* 2 *samples of variables* x *and* y*:*

$$P_1 = (0.3, 0.5), \qquad P_2 = (0.5, 0.7), \qquad P_3 = (1, 0.5),$$

$$P_4 = (1.5, 0.8), \qquad P_5 = (0.8, 1), \qquad P_6 = (0.7, 1.5).$$

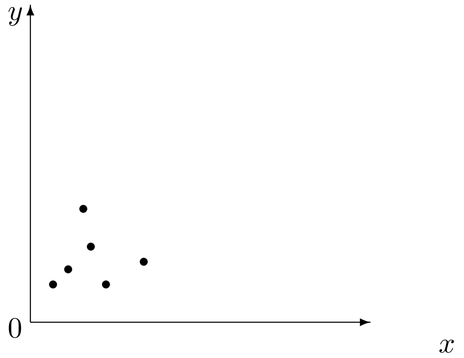

Figure 1. *The given scatter of points.*

Let us calculate the regression parameters α *and* β *with the help of formulas (2.1.7) and (2.1.6) to determine the regression line: Since*

$$\overline{x} = \frac{0.3 + 0.5 + 1 + 1.5 + 0.8 + 0.7}{6} = 0.8, \ \overline{y} = \frac{0.5 + 0.7 + 0.5 + 0.8 + 1 + 1.5}{6} = 0.8\overline{3},$$

we obtain:

$$\alpha = 0.8\overline{3}-$$

$$\frac{0.8(0.3\cdot 0.5+0.5\cdot 0.7+1\cdot 0.5+1.5\cdot 0.8+0.8\cdot 1+0.7\cdot 1.5)-6\cdot(0.8)^2\cdot 0.8\overline{3}}{(0.3)^2+(0.5)^2+1^2+(1.5)^2+(0.8)^2+(0.7)^2-6\cdot(0.8)^2}=$$

$$=0.7877.$$

$$\beta=\frac{0.3\cdot 0.5+0.5\cdot 0.7+1\cdot 0.5+1.5\cdot 0.8+0.8\cdot 1+0.7\cdot 1.5-6\cdot 0.8\cdot 0.8\overline{3}}{(0.3)^2+(0.5)^2+1^2+(1.5)^2+(0.8)^2+(0.7)^2-6\cdot(0.8)^2}=0.057,$$

hence the regression line is:

$$y = 0.057x + 0.7877.$$

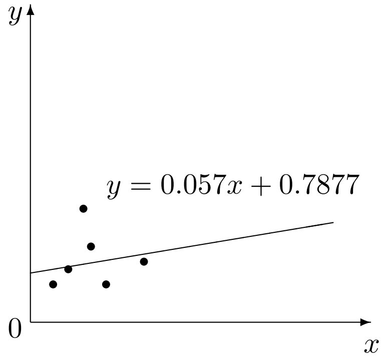

Figure 2. *The regression line.*

We can also calculate all the residuals e_i, i.e. the differences between y_i and $\hat{y}_i$, and their squares e_i^2 as well.

i	y_i	$\hat{y}_i$	e_i	e_i^2
1	0.5	0.8048	−0.3048	0.0929
2	0.7	0.8162	−0.1162	0.0135
3	0.5	0.8447	−0.3447	0.1188
4	0.8	0.8732	−0.0732	0.0053
5	1	0.8333	0.1667	0.0277
6	1.5	0.8276	0.6724	0.4521

Note that the sum of the squares of the residuals is $\sum_{i=1}^{6} e_i^2 = 0.7103$. Moreover, the larger contribution comes from point P_6, as can be seen from Figure 2, whereas P_2 and P_4 are 'almost' on the regression line.

2.2 Assessment of the goodness of fit

Every time we carry out a regression, we need a measure of the fit of the obtained regression line to the data. We are going to provide the definitions of some quantities that will be useful for this purpose:

- **Total Sum of Squares**:
$$SST = \sum_{i=1}^{N} (y_i - \overline{y})^2.$$
- **Regression Sum of Squares**:
$$SSR = \sum_{i=1}^{N} (\hat{y}_i - \overline{y})^2.$$
- **Error Sum of Squares**:
$$SSE = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2.$$

The 3 above quantities are linked by the straightforward relation we are going to derive. Since we have:

$$y_i - \overline{y} = y_i - \hat{y}_i + \hat{y}_i - \overline{y} \qquad \Longrightarrow \qquad (y_i - \overline{y})^2 = (y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2 =$$
$$= (y_i - \hat{y}_i)^2 + (\hat{y}_i - \overline{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}).$$

Summing over N terms yields:

$$\sum_{i=1}^{N} (y_i - \overline{y})^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{N} (\hat{y}_i - \overline{y})^2 + 2\sum_{i=1}^{N} (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}).$$

Now, let us take the last term in the right-hand side into account. Relying on the OLS procedure, we know that:

$$2\sum_{i=1}^{N} (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = 2\sum_{i=1}^{N} (y_i - \hat{y}_i)(\alpha + \beta x_i - \alpha - \beta\overline{x}) = 2\beta\sum_{i=1}^{N} (y_i - \hat{y}_i)(x_i - \overline{x}) =$$

$$= 2\beta\sum_{i=1}^{N} (y_i - \hat{y}_i + \overline{y} - \overline{y})(x_i - \overline{x}) = 2\beta\sum_{i=1}^{N} (y_i - \alpha - \beta x_i + \alpha + \beta\overline{x} - \overline{y})(x_i - \overline{x})$$

$$= 2\beta\sum_{i=1}^{N}(y_i-\overline{y}-\beta(x_i-\overline{x}))(x_i-\overline{x}) = 2\beta\left[\sum_{i=1}^{N}(y_i-\overline{y})(x_i-\overline{x}) - \beta\sum_{i=1}^{N}(x_i-\overline{x})^2\right] =$$

$$= 2\beta\left[\sum_{i=1}^{N}(y_i-\overline{y})(x_i-\overline{x}) - \frac{\sum_{j=1}^{N}(x_j-\overline{x})(y_j-\overline{y})}{\sum_{j=1}^{N}(x_j-\overline{x})^2}\cdot\sum_{i=1}^{N}(x_i-\overline{x})^2\right] = 0,$$

after employing expression (2.1.10) to indicate β. Since the above term vanishes, we obtain:

$$\sum_{i=1}^{N}(y_i-\overline{y})^2 = \sum_{i=1}^{N}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{N}(\hat{y}_i-\overline{y})^2,$$

then the following relation holds:

$$SST = SSE + SSR.$$

Now we can introduce a coefficient which is helpful to assess the closeness of fit: the **coefficient of determination** $R^2 \in (0,1)$.

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{N}(\hat{y}_i-\overline{y})^2}{\sum_{i=1}^{N}(y_i-\overline{y})^2} = \beta^2\frac{\sum_{i=1}^{N}(x_i-\overline{x})^2}{\sum_{i=1}^{N}(y_i-\overline{y})^2}.$$

An equivalent formulation of R^2 is the following one:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{N}(y_i-\overline{y})^2}.$$

The regression line fits the scatter of points better as close as R^2 is to 1. We can calculate R^2 in the previous Example, obtaining the value: $R^2 = 0.004$.

2.3 Ordinary least squares (OLS) estimation method: multiple variable case

When $N > 2$, we are in a standard scenario, because typically more than 2 variables are involved in an economic relationship. The standard linear equation that we are faced with reads as:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_N x_N + \epsilon, \tag{2.3.1}$$

where we chose not to use x_1 to leave the intercept alone, and ϵ represents the above-mentioned disturbance. Another possible expression of the same equation is:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_N x_N + \epsilon. \tag{2.3.2}$$

In (2.3.1) there are N regression parameters to be estimated. Taking the expectations and assuming $\mathbb{E}(\epsilon) = 0$, we have:

$$\mathbb{E}[y] = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_N x_N, \tag{2.3.3}$$

which is usually indicated as the **population regression equation**. In (2.3.3) β_1 is the **intercept** and $\beta_2, \ldots, \beta_N$ are the **regression slope parameters**. Suppose that our sample is composed of M observations for the explanatory variables x_i. We can write the values in the i-th observations as:

$$y_i,\ x_{2i},\ x_{3i}, \ldots, x_{Ni}.$$

For all $i = 1, \ldots, M$, we have:

$$y_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_N x_{Ni} + \epsilon_i,$$

or, in simple matrix form:

$$Y = X\beta + \epsilon, \tag{2.3.4}$$

where Y, β and ϵ are the following vectors:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_M \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_M \end{pmatrix}.$$

On the other hand, X is the following $M \times N$ matrix:

$$X = \begin{pmatrix} 1 & x_{21} & x_{31} & \cdots & x_{N1} \\ \\ 1 & x_{22} & \cdots & \cdots & x_{N2} \\ \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2M} & \cdots & \cdots & x_{NM} \end{pmatrix}.$$

If $\hat{\beta}_1, \ldots, \hat{\beta}_N$ are estimated values of the regression parameters, then $\hat{y}$ is the predicted value of y. Also here residuals are $e_i = y_i - \hat{y}_i$, and e is the vector collecting all the residuals. We have

$$Y = \hat{Y} + e \qquad \Longleftrightarrow \qquad e = Y - X\hat{\beta}.$$

Also in this case we use OLS, so we are supposed to minimize the sum of the squares of the residuals $S = \sum_{i=1}^{N} e_i^2$. We can employ the standard properties of Linear Algebra to achieve the following form (^{T} indicates transpose):

$$S = e^T e = \left(Y - X\hat{\beta}\right)^T \left(Y - X\hat{\beta}\right) = \left(Y^T - \hat{\beta}^T X^T\right)\left(Y - X\hat{\beta}\right) =$$

$$= Y^T Y - \hat{\beta}^T X^T Y - Y^T X\hat{\beta} + \hat{\beta}^T X^T X\hat{\beta} =$$

$$= Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta},$$

because they are all scalars, as is simple to check ($e^T e$ is a scalar product). The 2 negative terms have been added because $\hat{\beta}^T X^T Y = \left(Y^T X\hat{\beta}\right)^T$.

As in the 2-variables case, the next step is the differentiation of S with respect to $\hat{\beta}$, i.e. N distinct FOCs which can be collected in a unique vector of **normal equations**:

$$\frac{\partial S}{\partial \hat{\beta}} = -2X^T Y + 2X^T X\hat{\beta} = 0. \tag{2.3.5}$$

The relation (2.3.5) can be rearranged to become:

$$X^T X\hat{\beta} = X^T Y \qquad \Longrightarrow \qquad \left(X^T X\right)^{-1} X^T X\hat{\beta} = \left(X^T X\right)^{-1} X^T Y,$$

which can be solved for $\hat{\beta}$ to achieve the formula which is perhaps the most famous identity in Econometrics:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y. \tag{2.3.6}$$

Clearly, the matrix $X^T X$ must be non-singular, and the determination of its inverse may need a long and computationally costly procedure.

In the next Example, we are going to employ (2.3.6) in a simple case where the regression parameters are only 3 and the observations are 3 as well, to avoid excessive calculations.

Example 2. *Suppose that we have* 3 *observations of the* 2 *explanatory variables* X_2 *and* X_3*. The samples are collected in the following column vectors: where* Y*,* β *and* ϵ *are the following vectors:*

$$Y = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}, \qquad X_2 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}, \qquad X_3 = \begin{pmatrix} 1 \\ -2 \\ 0 \end{pmatrix},$$

hence the matrix X *is:*

$$X = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & -2 \\ 1 & -1 & 0 \end{pmatrix}.$$

The regression line will have the following form:

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3.$$

By formula (2.3.6), the column vector $\hat{\beta}$ *is determined by:*

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y =$$

$$= \left(\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 1 & -2 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & -2 \\ 1 & -1 & 0 \end{pmatrix} \right)^{-1} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 1 & -2 & 0 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 3 & 0 & -1 \\ 0 & 2 & -2 \\ -1 & -2 & 5 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 1 & -2 & 0 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}.$$

Now the calculation of the inverse of the above matrix must be carried out (it is invertible because its determinant is 16*). There are some methods that can be found in any basic Linear Algebra textbook*[1]*. When the dimensions of the involved matrices are higher and the regressions are run by a software such as Matlab or Stata, there are built-in packages or add-ons that can do the task. However, after the calculation, we find that*

$$\begin{pmatrix} 3 & 0 & -1 \\ 0 & 2 & -2 \\ -1 & -2 & 5 \end{pmatrix}^{-1} = \begin{pmatrix} 3/8 & 1/8 & 1/8 \\ 1/8 & 7/8 & 3/8 \\ 1/8 & 3/8 & 3/8 \end{pmatrix},$$

as is immediate to verify.

Finally, we can identify the regression parameters:

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} 3/8 & 1/8 & 1/8 \\ 1/8 & 7/8 & 3/8 \\ 1/8 & 3/8 & 3/8 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 1 & -2 & 0 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 1/4 & -3/4 \\ 1/2 & -1/4 & -1/4 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3/2 \\ 1/2 \\ 1/2 \end{pmatrix},$$

consequently the regression equation turns out to be:

$$\hat{Y} = 1.5 + 0.5X_2 + 0.5X_3.$$

[1] Otherwise, I suggest to take a look at the clear and simple Notes by Prof. Paul Smith:

$https : //sites.math.washington.edu/\ smith/Teaching/308/308_notes.pdf$

2.4 Assumptions for classical regression models

Typically, some assumptions are made on the explanatory variables and on the disturbances in the regression models. Such assumptions are not always the same, as can be seen by comparing different approaches. We are going to refer to the list of assumptions proposed by Greene (see [1], p. 56), augmenting it with a brief explanation on their meaning and importance. The first trivial assumption, which is not listed generally, concerns the values of x in the sample. We assume that there are some variations in each sample, meaning that for all $h = 2, \dots, N$, there exist at least 2 different values, i.e. $i \neq j$ such that $x_{hi} \neq x_{hj}$. If this assumption is not verified, there are some variables which are actually constant.

Assumptions involving the explanatory variables

- **(1A) - Linearity:** *a linear relationship is specified between explained and explanatory variables, i.e. (2.3.1) or (2.3.2);*
- **(1B) - Full rank:** *no exact linear relationship exists among any of the model's explanatory variables;*
- **(1C) - Data generation:** *the data collected in the independent variables can be either constants or random variables or a mixture of both;*

Assumption (1A) intends to establish the validity of the regression equation, whereas assumption (1B) means that no further constraints have to be taken into account (clearly, any linear relation among explanatory variables would be equivalent to the redundancy of some variables, so the system should be reduced.

On the other hand, assumption (1C) states that analysis is carried out conditionally on the observed values of X, so hence the outcome will not be influenced by the specific nature of the values (either fixed constants or random draws from a stochastic process).

We also have to consider 4 assumptions on all the disturbances, that are listed as follows.

Assumptions involving the disturbances
It is assumed that for all $i = 1, \dots, N$, the disturbances ϵ_i:

- **(2A) - Exogeneity of the independent variables:** $\mathbb{E}[\epsilon_i] = 0$ *and* $\mathbb{E}[\epsilon_i \mid X] = 0$*;*
- **(2B) - Homoscedasticity**[2]**:** $Var(\epsilon_i) = \mathbb{E}[\epsilon_i - \mathbb{E}[\epsilon_i]]^2 = \sigma^2 =$ *constant; moreover* $Var[\epsilon_i|X] = \sigma^2 =$ *constant;*

[2] Sometimes the word **Homoskedasticity** is used too.

- **(2C) - Non-autocorrelation:** $Cov(\epsilon_i, \epsilon_j) = \mathbb{E}\left\{[\epsilon_i - \mathbb{E}[\epsilon_i]][\epsilon_j - \mathbb{E}[\epsilon_j]]\right\} = 0$ *for all* $i \neq j$*;*
- **(2D) - Normal distribution:** *each* ϵ_i *is normally distributed with zero mean.*

$(2A)$ refers to the mean values of the disturbances, either conditional on X or not. This property denotes **exogeneity** of X (in other words, X is an exogenous variable), which has great importance in economic models, because it corresponds to the fact that X is really an external variable, so its effect on Y is 'pure'. On the other hand, assumption $(2B)$ is called **Homoscedasticity**, and it means that conditional variance is constant. When this assumption does not hold, there is **Heteroscedasticity** (or **Heteroskedasticity**), which is a definitely more complex case.

Here we are going to state some results to prove the correctness, or **unbiasedness**[3] of the regression parameters under some of the above assumptions. Basically, why do we use estimators achieved from the OLS method? We will see that estimators α and β have very relevant properties. Suppose that the following linear equation:

$$y = \alpha^* + \beta^* x + \epsilon$$

contains the best parameters to fit the scatter of points. The following results are stated in the 2-variables case, but they can be easily extended to N variables.

Proposition 3. *If assumptions* $(2A)$ *and* $(2C)$ *hold, then the estimators* α *given by (2.1.7) and* β *given by (2.1.6) are unbiased, i.e.*

$$\mathbb{E}[\alpha] = \alpha^*, \qquad \mathbb{E}[\beta] = \beta^*.$$

Proof. Firstly, let us calculate the expected value of β, with the help of the linear regression equation:

$$\mathbb{E}[\beta] = \mathbb{E}\left[\frac{\sum_{i=1}^N x_i y_i - N\overline{x}\cdot\overline{y}}{\sum_{i=1}^N x_i^2 - N\overline{x}^2}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^N x_i(\alpha^* + \beta^* x_i + \epsilon_i) - N\overline{x}\cdot(\alpha^* + \beta^*\overline{x})}{\sum_{i=1}^N x_i^2 - N\overline{x}^2}\right] =$$

$$= \mathbb{E}\left[\frac{\alpha^*\sum_{i=1}^N x_i}{\sum_{i=1}^N x_i^2 - N\overline{x}^2} + \frac{\beta^*\sum_{i=1}^N x_i^2}{\sum_{i=1}^N x_i^2 - N\overline{x}^2} + \frac{\sum_{i=1}^N x_i\epsilon_i}{\sum_{i=1}^N x_i^2 - N\overline{x}^2} - N\frac{\overline{x}\cdot(\alpha^* + \beta^*\overline{x})}{\sum_{i=1}^N x_i^2 - N\overline{x}^2}\right] =$$

$$= \mathbb{E}\left[\alpha^*\frac{\sum_{i=1}^N x_i - N\overline{x}}{\sum_{i=1}^N x_i^2 - N\overline{x}^2}\right] + \mathbb{E}\left[\beta^*\frac{\sum_{i=1}^N x_i^2 - N\overline{x}^2}{\sum_{i=1}^N x_i^2 - N\overline{x}^2}\right] + \mathbb{E}\left[\frac{\sum_{i=1}^N x_i\epsilon_i}{\sum_{i=1}^N x_i^2 - N\overline{x}^2}\right].$$

[3] The word **unbiased** refers to an estimator which is 'on average' equal to the real parameter we are looking for, not systematically too high or too low.

The second term is not random, hence $\mathbb{E}[\beta^*] = \beta^*$, whereas the first one vanishes because $\sum_{i=1}^N x_i = N\overline{x}$. Consequently, we have:

$$\mathbb{E}[\beta] = \beta^* + \mathbb{E}\left[\frac{\sum_{i=1}^N x_i\epsilon_i}{\sum_{i=1}^N x_i^2 - N\overline{x}^2}\right].$$

Now, since the numerator of the second term is equal to $\sum_{i=1}^N x_i\mathbb{E}[\epsilon_i \;\; x]$, by the *Law of Iterated Expectations* (see [1], Appendix B), it vanishes by assumption $(2A)$, hence $\mathbb{E}[\beta] = \beta^*$.

Turning to α, we know from (2.1.7) that:

$$\mathbb{E}[\alpha] = \mathbb{E}[\overline{y} - \beta^*\overline{x}] = \mathbb{E}[\overline{y}] - \beta^*\mathbb{E}[\overline{x}] = \mathbb{E}[\alpha^* + \beta^*\overline{x}] - \beta^*\mathbb{E}[\overline{x}] =$$

$$= \mathbb{E}[\alpha^*] + \beta^*\mathbb{E}[\overline{x}] - \beta^*\mathbb{E}[\overline{x}] = \mathbb{E}[\alpha^*] = \alpha^*.$$

□

Clearly, the mean value is not the only important characteristic of the regression parameters: as usually happens with random variables, the variance is crucial as well. This means that in addition to being a correct estimator, parameter β must also have a low variance. We are going to introduce the following result, also known as the **Gauss-Markov Theorem**, under the homoscedaticity assumption:

Theorem 4. *If the above assumptions hold, then β given by (2.1.6) is the estimator which has the minimal variance in the class of linear and unbiased estimators of β^*.*

Proof. Suppose that another estimator b exists as a linear function of y_i with weights c_i:

$$b = \sum_{i=1}^N c_i y_i = \sum_{i=1}^N c_i(\alpha + \beta x_i + \epsilon_i) = \alpha\sum_{i=1}^N c_i + \beta\sum_{i=1}^N c_i x_i + \sum_{i=1}^N c_i\epsilon_i.$$

In this case, since $\mathbb{E}[b] = \beta$, necessarily

$$\sum_{i=1}^N c_i = 0, \qquad \sum_{i=1}^N c_i x_i = \sum_{i=1}^N c_i(x_i - \overline{x}) = 1.$$

Hence, we have that

$$b = \beta + \sum_{i=1}^N c_i\epsilon_i \qquad \Longrightarrow \qquad Var(b\ |x) = Var\left(\beta + \sum_{i=1}^N c_i\epsilon_i \mid x\right).$$

We already know that

$$Var(b \mid x) = \sigma^2 \sum_{i=1}^{N} c_i^2, \qquad Var(\beta \mid x) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

(see [1], p. 99). Consider now the following sums:

$$\sum_{i=1}^{N} w_i^2 = \frac{1}{\sum_{i=1}^{N}(x_i - \overline{x})^2}, \sum_{i=1}^{N} w_i c_i = \frac{1}{\sum_{i=1}^{N}(x_i - \overline{x})^2}.$$

We can note that:

$$Var(b \mid x) = \sigma^2 \sum_{i=1}^{N} c_i^2 = \sigma^2 \sum_{i=1}^{N}(w_i + c_i - w_i)^2 =$$

$$= \sigma^2 \left[\sum_{i=1}^{N} w_i^2 + \sum_{i=1}^{N}(c_i - w_i)^2 + 2\sum_{i=1}^{N} w_i(c_i - w_i) \right] =$$

$$= \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2} + \sigma^2 \sum_{i=1}^{N}(c_i - w_i)^2 = Var(\beta \mid x) + \sigma^2 \sum_{i=1}^{N}(c_i - w_i)^2,$$

consequently $Var(b \mid x) > Var(\beta \mid x)$, i.e. β has the minimum variance. □

Sometimes estimator β in (2.1.6) is indicated as the BLUE (Best Linear Unbiased Estimator).

After discussing the properties of β, we should examine the distribution of errors. Assumption $(2D)$ establishes that each disturbance ϵ_i is normally distributed with 0 mean. At the present stage, we do not have any information on σ^2, i.e. the variance of β is still to be estimated. First, we have to come back to the expression of the least squares residuals e_i in N variables and apply (2.3.6):

$$e = Y - X\hat{\beta} = Y - X\left(X^T X\right)^{-1} X^T Y = (I_M - X\left(X^T X\right)^{-1} X^T)Y,$$

where I_M is the usual $M \times M$ identity matrix.

Now, call $\mathcal{M} = I_M - X\left(X^T X\right)^{-1} X^T$ the $M \times M$ **residual maker** (see [1], p. 71). We have that $e = \mathcal{M}Y$, furthermore, by construction:

$$\mathcal{M}X = (I_M - X\left(X^T X\right)^{-1} X^T)X = X - X\left(X^T X\right)^{-1}(X^T X) = X - XI_N = 0,$$

i.e the null $M \times N$ matrix. We know from the above identity that the residual maker is also useful because

$$e = \mathcal{M}Y = \mathcal{M}(X\beta + \epsilon) = \mathcal{M}\epsilon.$$

So an estimator of σ^2 can be obtained from the sum of squared residuals:

$$e^T e = (\mathcal{M}\epsilon)^T \mathcal{M}\epsilon = \epsilon^T \mathcal{M}^T \mathcal{M}\epsilon.$$

Before proceeding, we prove another key property of the residual maker:

$$\mathcal{M}^T\mathcal{M} = \left(I_M - X\left(X^TX\right)^{-1}X^T\right)^T\left(I_M - X\left(X^TX\right)^{-1}X^T\right) =$$

$$= \left(I_M - X\left(\left(X^TX\right)^{-1}\right)^T X^T\right)\left(I_M - X\left(X^TX\right)^{-1}X^T\right) =$$

$$= I_M - X\left(\left(X^TX\right)^{-1}\right)^T X^T - X\left(X^TX\right)^{-1}X^T +$$

$$+ X\left(\left(X^TX\right)^{-1}\right)^T X^TX\left(X^TX\right)^{-1}X^T =$$

$$= I_M - X\left(\left(X^TX\right)^{-1}\right)^T X^T - X\left(X^TX\right)^{-1}X^T + X\left(\left(X^TX\right)^{-1}\right)^T X^T =$$

$$= I_M - X\left(X^TX\right)^{-1}X^T = \mathcal{M}.$$

Since $\mathcal{M}^T\mathcal{M} = \mathcal{M}$, we have that:

$$e^T e = \epsilon^T \mathcal{M}\epsilon.$$

Borrowing a property of the trace of a matrix from Linear Algebra, we have:

$$tr(\epsilon^T\mathcal{M}\epsilon) = tr(\mathcal{M}\epsilon\epsilon^T) \implies \mathbb{E}[tr(\epsilon^T\mathcal{M}\epsilon) \mid X] = \mathbb{E}[tr(\mathcal{M}\epsilon\epsilon^T) \mid X].$$

Now we note that $\mathcal{M}$ can be taken out of the expectation, so that:

$$\mathbb{E}[tr(\mathcal{M}\epsilon\epsilon^T) \mid X] = tr(\mathcal{M}\mathbb{E}[\epsilon\epsilon^T \mid X] = tr(\mathcal{M}\sigma^2 I_M) = \sigma^2 tr(\mathcal{M}).$$

The trace of $\mathcal{M}$ can be calculated easily, using its properties:

$$tr(\mathcal{M}) = tr(I_M - X\left(X^TX\right)^{-1}X^T) = tr(I_M) - tr(\left(X^TX\right)^{-1}X^TX) = M - N.$$

Finally, we obtain that $\mathbb{E}[e^t e \mid X] = (M-N)\sigma^2$, and we are able to define an unbiased estimator of σ^2, which is called s^2:

$$s^2 = \frac{e^T e}{M-N}. \tag{2.4.1}$$

Note that $\mathbb{E}[s^2] = \sigma^2$. The quantity (2.4.1) will be very useful in the testing procedures. We will also call s the **standard error of regression**.

To conclude this preliminary discussion on parameters, given the previous assumptions and results, we can state that the distribution of β is the following:

$$\beta|x \ \sim \ N\left(\beta^*, \sigma^2(X^TX)^{-1}\right), \tag{2.4.2}$$

i.e. a multivariate normal distribution, meaning that each component of β is normally distributed:

$$\beta_k|x \ \sim \ N\left(\beta^*_k, \sigma^2(X^TX)^{-1}_{kk}\right). \tag{2.4.3}$$

Finally, as far as s^2 is concerned, we must remember that

$$\mathbb{E}[s^2 \mid x] = \mathbb{E}[s^2] = \sigma^2.$$

Chapter 3

Maximum likelihood estimation

The **maximum likelihood estimation** is one of the most important estimation methods in Econometrics. It can be shown to be consistent and asymptotically efficient under general conditions. Namely, the **Maximum Likelihood Estimator (MLE, from now on)** of a value is the number that is 'most likely' or has the maximum likelihood of generating that specific value.

An MLE must be found by first deriving a **likelihood function**, for example in a form such as $L = L(\theta, x_1, \dots, x_N)$, where θ is the variable which characterizes the population under consideration.

Let's take into account the following worked example.

Example 5. *Suppose that our population involves values of a discrete random variable X having geometric probability distributions:*

$$p(x_i) = (1-\theta)\theta^{x_i},$$

where x_i is a random observation on X. Since observations are independent for a random sample, we can write the probability of obtaining our N observations as

$$L = p(x_1) \cdot p(x_2) \cdot \dots \cdot p(x_N) = (1-\theta)^N \theta^{x_1+\dots+x_N}.$$

Typically, we prefer to take the logarithmic function of L rather than L itself, to ease the subsequent calculations. Call $l(\cdot) = \ln(L(\cdot)$ the ***log-likelihood function****. We have:*

$$l(\theta) = \ln(L(\theta)) = \ln(1-\theta) + x_1 \ln(\theta) + \dots + \ln(1-\theta) + x_N \ln(\theta) =$$

$$= N \ln(1-\theta) + \ln(\theta) \sum_{i=1}^{N} x_i.$$

Since $l(\theta)$ is maximized at the same value θ^* as $L(\theta)$, we can take the FOC:

$$\frac{\partial l}{\partial \theta} = -\frac{N}{1-\theta} + \frac{\sum_{i=1}^N x_i}{\theta} = 0 \qquad \Longrightarrow$$

$$\Longrightarrow \cdots \Longrightarrow \theta \left(\frac{1}{\sum_{i=1}^N x_i} + \frac{1}{N} \right) = \frac{1}{N} \qquad \Longrightarrow$$

$$\Longrightarrow \qquad \theta^* = \frac{\frac{1}{N}}{\frac{1}{\sum_{i=1}^N x_i} + \frac{1}{N}} = \frac{\sum_{i=1}^N x_i}{\sum_{i=1}^N x_i + N}. \tag{3.0.1}$$

On the other hand, the following Example describes an analogous derivation, when X is a continuous random variable distributed according to a Poisson density having θ as its parameter.

Example 6. *Call X a continuous random variable whose probability density function is a Poisson distribution of the kind:*

$$p(x) = \theta e^{-\theta x},$$

meaning that, as usual:

$$Pr\{X \leq x\} = F(x) = \int_{-\infty}^{x} \theta e^{-\theta t} dt.$$

Also in this case, the likelihood function is

$$L(\theta) = \prod_{i=1}^{N} p(x_i) = \theta^N e^{-\theta(x_1 + \cdots + x_N)},$$

whereas the corresponding log-likelihood function is

$$l(\theta) = \ln(L(\theta)) = \sum_{i=1}^{N} \ln(\theta e^{-\theta x_i}) = N \ln \theta - \theta \sum_{i=1}^{N} x_i. \tag{3.0.2}$$

Differentiating (3.0.2) with respect to θ, we obtain:

$$l'(\theta) = \frac{N}{\theta} - \sum_{i=1}^{N} x_i = 0 \implies \theta^* = \frac{N}{\sum_{i=1}^N x_i} = \frac{1}{\overline{x}},$$

where $\overline{x}$ is the usual arithmetic mean.

3.1 Maximum likelihood estimation of regression parameters

Consider a regression line

$$y_i = \alpha + \beta x_i + \epsilon_i,$$

where the disturbances ϵ_i have null mean values and their variance is the same as y_i, i.e. $Var(y_i) = Var(\epsilon_i) = \sigma^2$, for all $i = 1, \dots, N$. The explained variables y_i are normally distributed and their mean values are given by

$$\mathbb{E}[y_i] = \alpha + \beta x_i.$$

Hence, we can write the probability density function (p.d.f.) of each variable y_i as follows:

$$p(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\alpha-\beta x_i)^2}{2\sigma^2}}.$$

Due to the classical assumptions, the disturbances ϵ_i are uncorrelated, normally distributed, and independent of each other. This implies independence for y_i as well. Hence, the likelihood function will be the usual product of p.d. functions:

$$L(\alpha, \beta, \sigma^2) = p(y_1) \cdot \dots \cdot p(y_N).$$

Taking the logarithm yields:

$$l(\alpha, \beta, \sigma^2) = \ln(L(\alpha, \beta, \sigma^2)) = -\frac{N \ln(2\pi)}{2} - \frac{N \ln(\sigma^2)}{2} - \frac{1}{2\sigma^2} \sum_{i=1}^{N} [y_i - \alpha - \beta x_i]^2.$$

Now the standard procedure to find $\tilde{\alpha}$ and $\tilde{\beta}$ so as to minimize the sum of the squares must be implemented, given the negative sign of the above expression. Hence, the maximum likelihood estimators α and β are exactly the same as in the OLS procedure. However, we have to calculate the FOC with respect to the variance to determine the third parameter:

$$\frac{\partial l}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{N} [y_i - \alpha - \beta x_i]^2 = 0,$$

leading to:

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{N} [y_i - \alpha - \beta x_i]^2}{N} = \frac{\sum_{i=1}^{N} e_i^2}{N}. \qquad (3.1.1)$$

It is interesting to note that the same happens in the multivariate case.

3.2 Confidence intervals for coefficients

Interval estimation is fundamental to identify the best estimate of a parameter involving an explicit expression of its uncertainty. If we are to select an interval for a parameter θ, we typically assume that it must be symmetric. This means that, if we found the value $\hat{\theta}$, a suitable interval might be $[\hat{\theta} - \delta,\ \hat{\theta} + \delta]$, where δ may be chosen equal to 0.01 or to 0.05, conventionally.

Remembering that by (2.4.2) and (2.4.3) we have that:

$$\beta_k | x \ \sim \ N\left(\beta_k^*, \sigma^2 S_{kk}\right),$$

where S_{kk} is the k-th diagonal element of the matrix $(X^T X)^{-1}$. Therefore, taking 95 percent (i.e., $\alpha_c = 0.05$) as the selected confidence level, we have that

$$-1.96 \leq \frac{\beta_k - \beta_k^*}{\sqrt{\sigma^2 S_{kk}}} \leq 1.96$$

$$\Downarrow$$

$$Pr\left\{\beta_k - 1.96\sqrt{\sigma^2 S_{kk}} \leq \beta_k^* \leq \beta_k + 1.96\sqrt{\sigma^2 S_{kk}}\right\} = 0.95,$$

which is a statement about the probability that the above interval contains β_k^*.

If we choose to use s^2 instead of σ^2, we typically use the t distribution. In that case, given the level α_c:

$$Pr\left\{\beta_k - t^*_{(1-\alpha_c/2),[M-N]}\sqrt{s^2 S_{kk}} \leq \beta_k^* \leq \beta_k + t^*_{(1-\alpha_c/2),[M-N]}\sqrt{s^2 S_{kk}}\right\} = 1 - \alpha_c,$$

where $t^*_{(1-\alpha_c/2),[M-N]}$ is the appropriate quantile taken from t distribution. If $1 - \alpha_c = 0.95$, we obtain the confidence interval for each β_k^*, i.e.

$$\beta_k^* \in \ \left(\beta_k - 1.96\sqrt{s^2 S_{kk}},\ \beta_k + 1.96\sqrt{s^2 S_{kk}}\right).$$

Chapter 4

Approaches to testing hypotheses

There are several possible tests that are usually carried out in regression models, in particular testing hypotheses is quite an important task to assess the validity of an economic model. This Section is essentially based on Chapter 5 of Greene's book [1], which is suggested for a much more detailed discussion of this topic.

The first example proposed by Greene where a null hypothesis is tested concerns a simple economic model, describing price of paintings in an auction. Its regression equation is

$$\ln P = \beta_1 + \beta_2 \ln S + \beta_3 AR + \epsilon, \tag{4.0.1}$$

where P is the price of a painting, S is its size, AR is its 'aspect ratio'. Namely, we are not sure that this model is correct, because it is questionable whether the size of a painting affects its price (Greene proposes some examples of extraordinary artworks such as *Mona Lisa* by Leonardo da Vinci which is very small-sized). This means that this is an appropriate case where we can test a **null hypothesis**, i.e. a hypothesis such that one coefficient (β_2) is equal to 0. If we call H_0 the null hypothesis on β_2, we also formulate the related **alternative hypothesis**, H_1, which assumes that $\beta_2 \neq 0$.

The null hypothesis will be subsequently tested, or measured, against the data, and finally:

- **if the data are inconsistent with H_0 with a reasonable degree of certainty, H_0 will be rejected.**

- **Otherwise, provided the data are consistent with H_0, H_0 will not be rejected.**

Note that rejecting the null hypothesis means ruling it out conclusively, whereas not rejecting it does not mean its acceptance, but it may involve further investigation and tests.

The first testing procedure was introduced by Neyman and Pearson (1933), where the observed data were divided in an **acceptance region** and in a **rejection region**.

The so-called **general linear hypothesis** is a set of restrictions on the basic linear regression model, which are linear equations involving parameters β_i. We are going to examine some simple cases, as are listed in [1] (Section 5.3):

- one coefficient is 0, i.e. there exists $j = 1, \dots, N$ such that $\beta_j = 0$;
- two coefficients are equal, i.e. there exist j and k, $j \neq k$, such that $\beta_j = \beta_k$;
- some coefficients sum to 1, i.e. (for example) $\beta_2 + \beta_5 + \beta_6 + \beta_8 = 1$;
- more than one coefficient is 0, i.e. (for example) $\beta_3 = \beta_5 = \beta_9 = 0$.

Then there may be a combination of the above restrictions, for example we can have that 2 coefficients are equal to 0 and other 2 coefficients are equal, and so on. There can also be some non-linear restrictions, in more complex cases.

We will discuss the 3 main tests in the following Sections: the Wald Test, the Likelihood Ratio (LR) Test, the Lagrange Multipliers (LM) test.

4.1 Hints on the main distributions in Statistics

We are going to recall two major distributions which are particularly helpful when implementing the tests, especially regression analysis, analysis of variance, and so on: the **chi-squared or χ^2 distribution with k degrees of freedom** and the **Student's $t-$distribution** (or **t-distribution**).

Given $k > 1$ independent, normally distributed random variables $Z_1, \dots, Z_p$, the sum of their squares is distributed according to the χ^2 distribution with k degrees of freedom, i.e

$$\sum_{j=1}^{k} Z_j^2 \;\sim\; \chi^2(k).$$

The p.d.f. of $\chi^2(k)$ is the following one:

$$f(x;k) = \begin{cases} \dfrac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{\sqrt{2^k} \cdot \Gamma\left(\dfrac{k}{2}\right)} & \text{if } x > 0 \\ & \\ 0 & \text{otherwise} \end{cases} , \tag{4.1.1}$$

where $\Gamma(\cdot)$ is **Euler's (and Legendre's) Gamma function**, i.e.

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx,$$

for all $z \in \mathbb{C} \setminus \mathbb{Z}_-$, and in particular, it is defined on positive integers as follows:

$$\Gamma(n) = (n-1)!$$

for all $n \in \mathbb{N}$.

The properties of $\chi^2(k)$ are many, and they can be found on any Statistics textbook. A particularly meaningful one is:

if $X_1, \ldots, X_k$ are independent normally distributed random variables, such that $X_i \sim N(\mu, \sigma^2)$, then:

$$\sum_{i=1}^{k}(X_i - \overline{X})^2 \sim \sigma^2\chi^2_{k-1},$$

where $\overline{X} = \dfrac{X_1 + \cdots X_k}{k}$.

A key use of χ^2 distributions concerns the F statistic, especially the construction of the **Fisher - Snedecor distribution** F, which will be treated in the last Section of the present Chapter.

On the other hand, Student's t-distribution is quite important when the sample size at hand is small and when the standard deviation of the population is unknown. The t-distribution is widely employed in a lot of statistical frameworks, for example the Student's t-test to assess the statistical significance of the difference between two sample means, or in the linear regression analysis.

Basically, we assume to take a sample of p observations from a normal distribution. We already know that a true mean value exists, but we can only calculate the sample mean. Defining $\nu = p - 1$ as the number of degrees of freedom of the t-distribution, we can assess the confidence with which a given range would contain the true mean by constructing the distribution with the following p.d.f.:

$$f(x;\nu) = \begin{cases} \dfrac{\Gamma\left(\frac{\nu+1}{2}\right)\left(1 + \dfrac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}}{\sqrt{\nu\pi}\cdot\Gamma\left(\dfrac{\nu}{2}\right)} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}, \qquad (4.1.2)$$

4.2 Wald Test

The Wald test is named after the statistician Abraham Wald. It can be used in a number of different contexts to estimate the distance between the estimate $\hat{\theta}$ of a parameter (that is, its MLE) and the proposed value of the same parameter θ_0. Substantially, it is a 'significance test', meaning that its principle is to fit the regression without any restriction, and then assess whether the results seem to agree with the hypothesis.

We begin from a very simple case, referring to the above example on art market. Suppose that we want to test a null hypothesis H_0: $\beta_2 = \beta_2^0$, where β_2^0 is the assumed value (in this case, zero) of the regression coefficient. We aim to evaluate the **Wald distance** W_j of a coefficient estimate from its hypothesized value:

$$W_j = \frac{b_j - \beta_j^0}{\sqrt{\sigma^2 S_{jj}}}, \tag{4.2.1}$$

where s^2 is given by (2.4.1) and S_{jj} is the j-th diagonal element of the matrix $(X^T X)^{-1}$. If we assume that $\mathbb{E}[\beta_j] = \beta_j^0$, W_j is normally distributed. We can call $W_j = t_j$ because it has a distribution with $M - N$ degrees of freedom.

We first identify a confidence interval with which we would like to verify our model, for example the standard value of 95%. So, we can state that it is unlikely that a single value of t_j falls outside the interval:

$$(-t^*_{(1-\alpha/2),[M-N]},\ t^*_{(1-\alpha/2),[M-N]}).$$

The null hypothesis H_0 should be rejected if W_j is sufficiently large.

In Greene's own words, since 'it is so unlikely that we would conclude that it could not happen if the hypothesis were correct, so the hypothesis must be incorrect'.

Back to the results in the previous Chapter, if we compute W_k using the sample estimate of σ^2, i.e. s^2, we have:

$$t_j = \frac{b_j - \beta_j^0}{\sqrt{s^2 S_{jj}}}. \tag{4.2.2}$$

The variable t_j in the form (4.2.2) has a t distribution with $M - N$ degrees of freedom. The t **ratio** is the ratio between the estimator b_j and its standard error, so:

$$t_j = \frac{b_j}{\sqrt{s^2 S_{jj}}}$$

can be used for tests. If it is larger than 1.96, this means that the coefficient is significantly different from 0 at 95% confidence level, the null hypothesis should be rejected, so the related coefficient can be considered **statistically significant**.

In the next Example we are going to deal with an econometric model derived from a study by Mroz [2] (published in 1987 on *Econometrica*), corresponding to Example 5.2 ([1], pages $156 - 157$). It will be very useful to outline how to read and understand the regression results in a Table.

Example 7. *Consider the following regression equation which aims to investigate the relation between married women's earnings and other relevant data such as their age, education and children:*

$$\ln(Earnings) = \beta_1 + \beta_2 \cdot Age + \beta_3 \cdot Age^2 + \beta_4 \cdot Education + \beta_5 \cdot Kids + \epsilon. \quad (4.2.3)$$

Note the presence of the same covariate in 2 *different positions: Age is considered both in the linear and in the quadratic forms. This structure violates the assumption of independence among covariates, but it justified by the well-known effect of age on income, which has a parabolic concave behaviour over time. This scenario is easily explained by the relation between wages and pensions, for example. For this reason, we expect that the coefficient* β_2 *is positive and that* β_3 *is negative.*

The number of observations is 428*, corresponding to* 428 *white married women whose age was between* 30 *and* 60 *in* 1975*, and consequently the number of degrees of freedom of the model is* $428 - 5 = 423$*. The following Table presents all the results, including the t ratio:*

Variable	*Coefficient*	*Standard error*	*t ratio*
Constant	3.24009	1.7674	1.833
Age	0.20056	0.08386	2.392
Age^2	−0.0023147	0.00098688	−2.345
Education	0.067472	0.025248	2.672
Kids	−0.35119	0.14753	−2.38

To augment the above Table, we also know that the sum of squared residuals SSE *is* 599.4582*, that the standard error of the regression s is* 1.19044*, and that* $R^2 = 0.040995$*. In short, we can summarize the following:*

- *The t ratio shows that at* 95% *confidence level, all coefficients are statistically significant except the intercept, which is smaller then* 1.96*.*

- *The signs of all coefficients are consistent with our initial expectations: education affects earnings positively, the presence of children affects earnings negatively. We can estimate that an additional year of schooling yields* 6.7% *increase in earnings.*

- *The age acts as an inverted U on earnings, i.e. β_2 is positive and β_3 is negative. Specifically, the form of the age profile suggests that the peak of earnings can be approximately found at 43 years of age.*

4.3 The F statistic

The **F test**, or **F statistic**, is a way to test a hypothesis against another one, for example the null hypothesis against the alternative hypothesis. We are going to treat this fundamental test as easily as possible (for further reading and technical details, see [1], $157 - 161$).

First, we should rigorously define the Fisher - Snedecor distribution: consider 2 random variables X and $\widetilde{X}$, which are respectively distributed according to the chi-squared distributions $\chi^2(k)$ and $\chi^2(l)$, having k and l degrees of freedom. The **Fisher - Snedecor distribution** $\mathcal{F}(k,l)$ is the distribution of the random variable

$$F = \frac{X/k}{\widetilde{X}/l}.$$

Its p.d.f. is given by

$$f(x;k,l) = \frac{1}{xB(k/2,l/2)}\sqrt{\frac{k^k l^l x^k}{(kx+l)^{k+l}}},$$

where $B(\cdot)$ is **Euler's Beta function**, i.e.

$$B\left(\frac{k}{2},\frac{l}{2}\right) = \int_0^1 t^{\frac{k}{2}-1}(1-t)^{\frac{l}{2}-1}dt,$$

which is connected to Gamma by the following identity:

$$B\left(\frac{k}{2},\frac{l}{2}\right) = \frac{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{l}{2}\right)}{\Gamma\left(\frac{k+l}{2}\right)}.$$

The mean value of such a random variable is $\frac{l}{l-2}$ for $l > 2$, and its variance is $\frac{2l^2(k+l-2)}{k(l-2)^2(l-4)}$ for $l > 4$.

To carry out the F test, we are going to assume that the 2 random variables under consideration are normally distributed with variances σ^2_X and $\sigma^2_{\widetilde{X}}$ and observed standard errors s^2_X and $s^2_{\widetilde{X}}$. Now, since the random variables

$$\frac{(k-1)s^2_X}{\sigma^2_X} \qquad \text{and} \qquad \frac{(l-1)s^2_{\widetilde{X}}}{\sigma^2_{\widetilde{X}}}$$

are respectively distributed according to $\chi^2(k-1)$ and $\chi^2(l-1)$, then the random variable

$$F = \frac{\sigma^2_{\widetilde{X}}}{\sigma^2_X} \frac{s^2_X}{s^2_{\widetilde{X}}}$$

follows $\mathcal{F}(k-1, l-1)$.

The easiest way to use the F test in Econometrics can be described as follows. First, we should also note that the F ratio quantifies the relationship between the relative increase in the SSR and the relative increase in degrees of freedom between 2 models. Call SSR_1 and SSR_2 the sums of squares of residuals of the 2 models, which respectively have p_1 and p_2 degrees of freedom. Model 1 is the 'simple' model, whereas model 2 is 'complicated'. Clearly, we can take one of the 2 models based on the null hypothesis and the remaining one based on the alternative hypothesis, to test them against one another. We can also write:

$$F = \frac{SSR_1 - SSR_2}{SSR_2} \frac{p_2}{p_1 - p_2}, \tag{4.3.1}$$

under the assumption that model 1 is simpler then model 2, then it has a larger SSR. What we expect is that if the more complicated model (2) is correct, the following inequality holds:

$$\frac{SSR_1 - SSR_2}{SSR_2} > \frac{p_1 - p_2}{p_2},$$

which is equivalent to saying that if the F ratio (4.3.1) is smaller than 1, the simpler model is the correct one. On the other hand, if (4.3.1) is larger than 1, we can have 2 occurrences:

- either the more complicated model is the correct one;
- or the simpler model is the correct one, but the impression of a better fit achieved by the more complicated model is caused by the random scatter.

In order to try to answer this question, we can employ the **P value**, which provides an assessment of the probability of this last case. Basically, this verification works as follows:

- **if the** P**-value is low**, we can conclude that **model** 2 **is significantly better than model** 1;

- **if the P-value is high**, no evidence exists which supports model 2, so **we accept model** 1.

To conclude with, a few explanatory words about the **P-value**, also known as the **asymptotic significance**, which is the probability that, when the null hypothesis is assumed to be true, a result is obtained which is equal or more extreme than the one which is actually observed. Hence:

the smaller the P-value $\Longrightarrow$ the higher the significance $\Longrightarrow$

$\Longrightarrow$ the higher the probability that the null hypothesis does not appropriately explain the scenario.

In other words, given a significance level α_l selected by the investigator, if the P-value is smaller than α_l, the data are inconsistent with the null hypothesis, so it must be rejected.

To conclude with, here is a very simple example to show how to use the P-value.

Example 8. *Suppose to flip a coin* 7 *times in a row. If the coin is fair, at every flip we have the following trivial probabilities:*

$$Prob\{Outcome\ is\ Head\} = Prob\{Outcome\ is\ Tail\} = \frac{1}{2}.$$

Assume the fairness of the coin as the null hypothesis, i.e.

Null hypothesis: the coin is fair.

Alternative hypothesis: the coin is unfair, or fixed.

Suppose that the P-value is calculated based on the total number of Heads obtained, and that the confidence cutoff is 0.05*.*

If the researcher gets 'Head' 7 *times, the probability of such an event, provided each flip of the coin is independent of the remaining flips, is*

$$(1/2)^7 = 0.0078 < 0.05,$$

that is the result is significant at this confidence level.

Therefore, ***the null hypothesis should be rejected.*** *We conclude that a very high probability exists that* ***the coin is fixed.***

On the other hand, if the researcher gets 'Head' 4 *times and 'Tail'* 3 *times, the probability of such an event is*

$$\frac{7!}{4!3!}\frac{1}{2^7} = 0.2734375 > 0.05,$$

so this result is not significant. In this case, ***the null hypothesis can be accepted.***

Chapter 5

Dummy variables

The **dummy variables** (sometimes referred to as **binary variables**) are variables which can only be equal to 0 or to 1. They are typically employed when a certain effect or situation occurs under some circumstances or in some periods but not in other ones. They can be either summed in a regression equation or multiplied by the explanatory variables, depending on the context at hand.

We have already encountered a dummy variable (i.e., *kids*) in the previous Example 7, where the dummy intended to highlight the effect of possible presence of children in the involved female population.

On the other hand, the following worked example is borrowed from basic Microeconomics, and it can be useful for comprehension.

Example 9. *Suppose that we are constructing the regression line to estimate the quantity of ice creams consumed by the population in the* 4 *seasons. We consider the following variables:*

- Q*: demanded quantity of ice creams;*
- P*: price of ice creams;*
- E*: total expenditure of consumers.*

We can construct the linear relations with the help of a dummy variable in 2 *ways: either additive or multiplicative.*

In the additive case, the linear relation to be analyzed is:

$$Q = \beta_1 + \alpha_1 D + \beta_2 E + \beta_3 P + \epsilon, \tag{5.0.1}$$

where the regression parameters are $\beta_1, \beta_2, \beta_3$*, as usual. More than that, we have a further dummy variable* D*, which is equal to* 1 *during summertime, when ice creams are typically sold, and equal to* 0 *in the remaining* 3 *seasons. The dummy* *variable D is multiplied by a further regression parameter which is indicated by α_1 to highlight its difference with respect to the other ones. Finally, ϵ is the usual disturbance. Passing to the expected values, we have 2 possible regression equations:*

$$\mathbb{E}[Q] = \begin{cases} \beta_1 + \alpha_1 + \beta_2 E + \beta_3 P & \textit{during the summer} \\ \beta_1 + \beta_2 E + \beta_3 P & \textit{in the remaining seasons} \end{cases} .$$

Clearly, the estimation of α_1 can be carried out only in the first case.

A dummy variable can also be used as a multiplicative variable, by modifying the linear equation (5.0.1) as follows, for example:

$$Q = \beta_1 + \beta_2 E + \alpha_1 ED + \beta_3 P + \epsilon. \tag{5.0.2}$$

In this case, in the period in which $D = 1$, its effect is not separated from the other variables, because it 'reinforces' the expenditure variable E. When $D = 0$, the equation coincides with the one in the additive dummy case. The 2 regression equations read as

$$\mathbb{E}[Q] = \begin{cases} \beta_1 + (\beta_2 + \alpha_1) E + \beta_3 P & \textit{during the summer} \\ \beta_1 + \beta_2 E + \beta_3 P & \textit{in the remaining seasons} \end{cases} .$$

Further advanced details go beyond the scope of the present lecture notes. As usual, for further explanation and technical details, I encourage students to read [1], Chapter 6.

Bibliography

[1] Greene, W. H. (1997). *Econometric analysis 3rd edition.* New Jersey: Prentice-Hall International.

[2] Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions, *Econometrica: Journal of the Econometric Society* **55(4)**: 765-799.

[3] Stock, J. H., Watson, M. W. (2007). *Introduction to Econometrics.* Pearson Education Inc. New York.

[4] Thomas, R. L. (1997). *Modern econometrics: an introduction.* Addison-Wesley Longman.

[5] Wooldridge, J. M. (2013). *Introduction to Econometrics.* Cengage Learning.

[6] Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach.* Nelson Education.