eman ta zabal zazu

Universidad del País Vasco | Euskal Herriko Unibertsitatea

CAMPUS OF INTERNATIONAL EXCELLENCE

Lesson 4

The Multiple Regression Model

Specification

Pilar González and Susan Orbe

Dpt. Applied Econometrics III (Econometrics and Statistics)

Learning objectives

- To analyse the functional form in Gretl.
- To identify the main elements of an econometric model.
- To understand the basic assumptions of the linear regression model.
- To interpret the coefficients of the regression model.
- To include qualitative explanatory variables in the model by means of dummy variables.

Contents

Contents

Econometric approach to model specification.

Econometric model.

$$
\begin{aligned}
\text{Dependent variable} &= \text{Systematic Part} + \text{Random Part} \\
& \text{Systematic Part} = f(\text{explanatory variables}) \\
& \text{Random Part} = \text{error term}
\end{aligned}
$$

$$
\begin{aligned}
Y &= f(\text{explanatory variables}) + \text{error term} \\
Y &= f(X_2, X_3, \ldots, X_k) + u
\end{aligned}
$$

Econometric approach to model specification.

It is necessary to choose:

- Explanatory variables: quantitative and/or qualitative.
- Functional form $f(.)$: linear, quadratic, logarithmic, ...
- The error term (or disturbance) distribution: mean, variance, ...

Contents

Graphic analysis and functional form.

Example: household consumption function.

Economic model:

$$C = f(I) \qquad \frac{dC}{dI} > 0 \quad \text{(marginal propensity to consume)}$$

Econometric model:

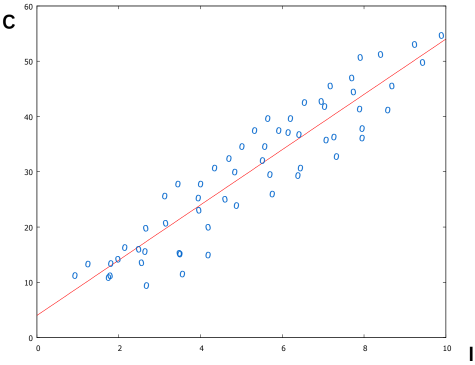
$$C = f(I) + u$$

Income is a quantitative explanatory variable.

What about the functional form? How is the relationship between consumption and income?

Tool: plot of consumption against income

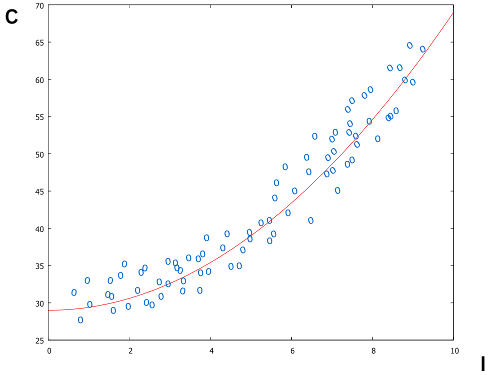
Graphic analysis and functional form.



Functional form: linear.

$$C_i = \beta_1 + \beta_2 I_i + u_i \quad i = 1, 2, ..., N$$

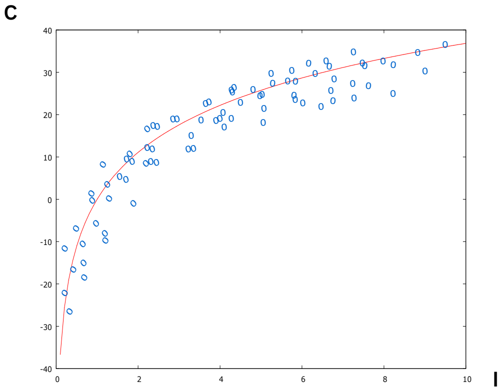
Graphic analysis and functional form.



Functional form: quadratic.

$$C_i \,=\, \beta_1 \,+\, \beta_2\, I_i \,+\, \beta_3\, I_i^2 \,+\, u_i \quad i = 1, 2, ..., N$$

Graphic analysis and functional form.



Functional form: double logarithmic.

$$\ln C_i = \beta_1 + \beta_2 \ln I_i + u_i \quad i = 1, 2, ..., N$$

Graphic analysis and functional form.

$$
\begin{aligned}
\text{Linear}: \quad & C_i && = \beta_1 + \beta_2 I_i + u_i \\
\text{Quadratic}: \quad & C_i && = \beta_1 + \beta_2 I_i + \beta_3 I_i^2 + u_i \\
\text{Log-Log}: \quad & \ln C_i && = \beta_1 + \beta_2 \ln I_i + u_i \\
\text{Log-Linear}: \quad & \ln C_i && = \beta_1 + \beta_2 I_i + u_i \\
\text{Linear-Log}: \quad & C_i && = \beta_1 + \beta_2 \ln I_i + u_i
\end{aligned}
$$

Is valid any functional form within the framework of the general linear regression model?

Assumption. Linearity

The linear regression model must be linear in the coefficients.

Graphic analysis and functional form.

Example 4.1. Graphic analysis and functional form.

1. Plotting functions using Gretl.

   Example 4.1.1

2. Graphic analysis of data and functional form.

   Example 4.1.2

Contents

Qualitative explanatory variables.

Some examples:

- Gender, race, level of studies, position, location,...
- Seasonality, structural breaks (war/peace, crisis/no crisis, ...)
- Quantitative variables measured by intervals: income, age, ...

Qualitative explanatory variables.

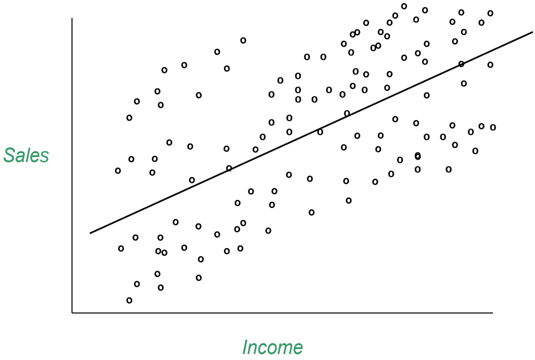
Example: Sales of a chain of shops.

A chain of shops with establishments in France, Spain and Italy wants to analyse the factors that determine sales. They collect data from 350 shops on sales and average income in the neighborhood where the shops are located.

Qualitative explanatory variables.

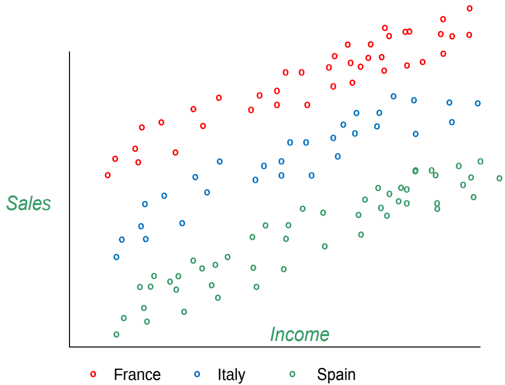
Consider the simple regression model:

$$S = f(I) \quad \Rightarrow \quad S_i = \beta_1 + \beta_2 I_i + u_i \qquad i = 1, 2, \ldots, 350$$



Qualitative explanatory variables.

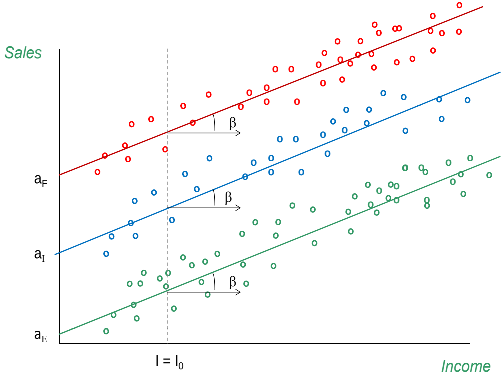
Now, have a look at this graph that includes information about the location of the shops (Spain, France and Italy).



$\Rightarrow$ Sales seem to be different in the three countries considered.

$$S = f(Income, Country)$$

Qualitative explanatory variables.



The “country” effect should be included in the econometric model as follows: the marginal effect of income on sales is the same for Spain, Italy and France, but given an income, the average level of sales is different in each country.

Qualitative explanatory variables.

Tool to include qualitative explanatory variables in the model:

DUMMY VARIABLES

Dummy variables.

A dummy variable is a binary variable:

$$D_i = \begin{cases} 1 & \text{if the characteristic is present in observation i} \\ 0 & \text{otherwise} \end{cases}$$

Qualitative explanatory variables.

In principle, we need as many dummy variables as categories of the qualitative variable.

Example.

Number of categories: 3 $\Rightarrow$ 3 Dummy variables

$$It_i = \begin{cases} 1 & i \in \text{ Italy} \\ 0 & \text{otherwise} \end{cases}$$

$$F_i = \begin{cases} 1 & i \in \text{ France} \\ 0 & \text{otherwise} \end{cases}$$

$$E_i = \begin{cases} 1 & i \in \text{ Spain} \\ 0 & \text{otherwise} \end{cases}$$

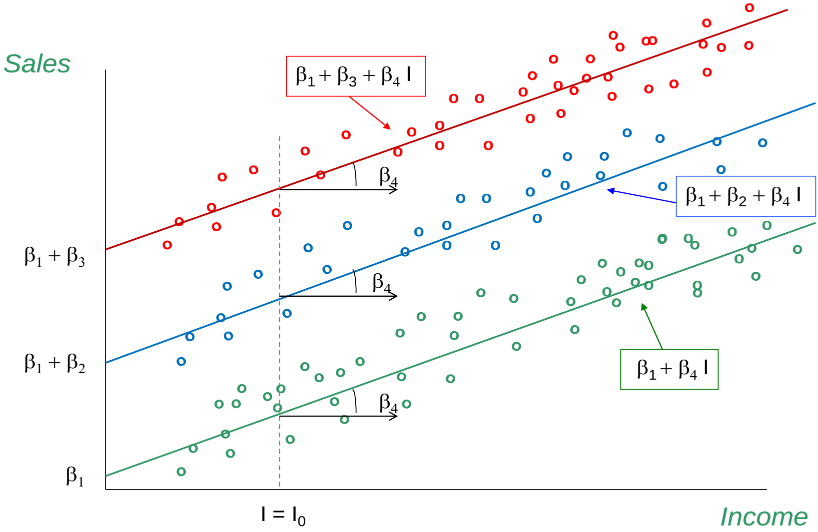
Specification

We include in the model the intercept and as many dummy variables as categories of the qualitative variable minus 1.

$$S_i = \beta_1 + \beta_2\, It_i + \beta_3\, F_i + \beta_4\, I_i + u_i \quad i = 1, 2, ..., 350$$

$$\begin{aligned} S_i \,|\, i \,\epsilon \text{ Italy} &= \beta_1 + \beta_2 + \beta_4\, I_i + u_i \\ S_i \,|\, i \,\epsilon \text{ France} &= \beta_1 + \beta_3 + \beta_4\, I_i + u_i \\ S_i \,|\, i \,\epsilon \text{ Spain} &= \beta_1 + \beta_4\, I_i + u_i \end{aligned}$$

Graphic analysis.



Qualitative explanatory variables.

Why not include all the dummy variables and the intercept in the model?

$$S_i = \beta_1 + \beta_2\, It_i + \beta_3\, F_i + \beta_4\, E_i + \beta_5\, I_i + u_i \quad i = 1, 2, ..., 350$$

$$\begin{aligned} S_i \,|\, i \,\epsilon \text{ Italy} &= \beta_1 + \beta_2 + \beta_5 I_i + u_i \\ S_i \,|\, i \,\epsilon \text{ France} &= \beta_1 + \beta_3 + \beta_5\, I_i + u_i \\ S_i \,|\, i \,\epsilon \text{ Spain} &= \beta_1 + \beta_4 + \beta_5\, I_i + u_i \end{aligned}$$

The coefficients $\beta_1, \beta_2, \beta_3, \beta_4$ **are NOT identified** in this model because $F_i + It_i + E_i = 1, \forall i$.

Assumption. NO perfect collinearity

There are not linear combinations among the regressors of the model.

Generating dummy variables in Gretl.

Example 4.2. Design dummy variables in Gretl.

1. Enter dummy variables manually.
   See **Example 4.2.1** for applications.

2. Generate dummy variables for discrete variables.
   See **Example 4.2.2** for applications.

3. Define a dummy variable for a range of observations.
   See **Example 4.2.3** for applications.

4. Dummy variables generated by Gretl.
   See **Example 4.2.4** for applications.

5. Time trend variable.
   See **Example 4.2.5** for applications.

Contents

Multiple Regression Model. Specification and assumptions.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i \quad i = 1, 2, \ldots, N$$

- Y: dependent variable, endogeneous variable, regressand.
- $X_j \;\; j = 1, \ldots, k$: explanatory variables, independent variables, regressors.
- $\beta_j \;\; j = 1, \ldots, k$: unknown coefficients.
- u: error term or disturbance (non observable).
- N: sample size.

Multiple Regression Model. Specification and assumptions.

Systematic part: $\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki}$

- Include all the relevant factors to determine sales and all the factors included are relevant.
- Represents the expected behaviour of the dependent variable Y conditional on the sample values of X:

$$E(Y_i|X) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} \quad \Rightarrow \quad E(u_i|X) = 0$$

Random part:

u is a non observable random variable that includes:

- factors other than income that affect sales not explicitly included in the model,
- uncertainty of economic relationships,
- small data discrepancies or measurement errors.

Basic assumptions.

Assumption A1.

The model in the population can be written as:

$$Y_i = \beta_1 + \beta_2\, X_{2i} + \beta_3\, X_{3i} + \ldots + \beta_k\, X_{ki} + u_i \qquad i = 1, 2, , \ldots, N$$

where:

- $X_2, X_3, \ldots, X_k$: regressors.
- $\beta_1, \beta_2, \beta_3, \ldots, \beta_k$: unknown parameters of interest (constant).
- The model is linear in the coefficients.
- The model is well specified, that is, all the relevant factors are included in the model and all the factors included in the model are relevant.

Basic assumptions.

Assumption A.2. No perfect collinearity

In the sample, none of the explanatory variables is constant and there are no exact linear relationship among the explanatory variables.

Example:

$$Sales_i = \beta_1 + \beta_2\, I_i + \beta_3\, Italy_i + \beta_4\, France_i + \beta_5\, Spain_i + u_i \quad i = 1, 2, \ldots, N$$

Does this assumption hold if all the shops are located in Italy, France or Spain?

Example:

$$Sales_i = \beta_1 + \beta_2\, Price_i + u_i \qquad i = 1, 2, , \ldots, N$$

Does this assumption hold if all the firms set the same price?

Basic assumptions.

Assumptions on the error term

A.3 Zero conditional mean:

$$E(u_i|X_2, X_3, \ldots, X_k) = E(u) = 0 \quad \forall i = 1, 2, , \ldots, N$$

A.4 Homoskedasticity: $V(u_i|X_2, X_3, \ldots, X_k) = V(u) = \sigma_u^2 \quad \forall i = 1, 2, , \ldots, N$

A.5 No serial correlation: $cov(u_i u_j|X_2, X_3, \ldots, X_k) = 0 \quad \forall i \neq j$

A.6 Normality: The errors u_i are independent of X and identically normally distributed.

$$\boxed{u_i \sim NID(0, \sigma_u^2) \qquad \forall i = 1, 2, , \ldots, N}$$

Interpretation of the coefficients.

Given the assumptions of the GLRM,

$$\begin{aligned} E(Y_i|X) &= E[(\beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i)|X] \\ &= \beta_1 + \beta_2\, X_{2i} + \ldots + \beta_k\, X_{ki} + \underbrace{E(u_i|X)}_{=0} \end{aligned}$$

$$E(Y_i|X) = \beta_1 + \beta_2\, X_{2i} + \ldots + \beta_k\, X_{ki}$$

$E(Y_i|X)$ is the *Population Regression Function* (PRF):

the expected value of Y_i given the values of the explanatory variables.

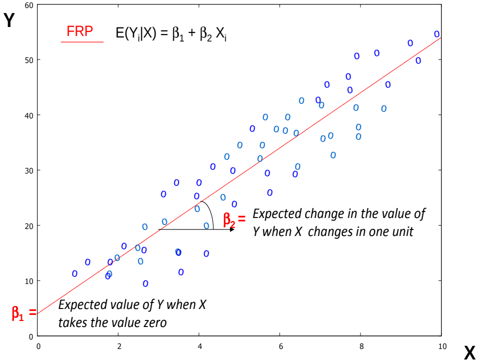
The Simple Regression Model.

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \text{PRF: } E(Y_i|X) = \beta_1 + \beta_2 X_i$$

where:

β_1 : intercept of the population regression function.

β_2 : slope of the population regression function.



Interpretation of the coefficients.

The Multiple Regression Model.

The Population Regression Function:

$$E(Y_i|X) = \beta_1 + \beta_2 \, X_{2i} + \ldots + \beta_k \, X_{ki}$$

Let's assume that all the regressors are quantitative explanatory variables.

- $\beta_1 = E[Y_i | X_{2i} = 0, \ldots, X_{ki} = 0]$

  β_1 (intercept) is the expected value of Y when all the explanatory variables equal 0.

- $\beta_j, j = 2, 3, \ldots, k$ (slopes): the expected change (increase or decrease) in Y resulting from changing X_j by one unit, holding the rest of the explanatory variables constant.

Interpretation of the coefficients.

1. The interpretation of the coefficients depends on the type of relationship between the dependent variable and the regressor.

Functional form	Partial effect	Elasticity
Linear $Y_i = \beta_1 + \beta_2 X_i + u_i$	β_2	$\beta_2 \dfrac{X}{Y}$
Quadratic $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i$	$\beta_2 + 2\beta_3 X$	$(\beta_2 + 2\beta_3 X)\dfrac{X}{Y}$
Log-Log $\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i$	$\beta_2 \dfrac{Y}{X}$	β_2
Log-Linear $\ln Y_i = \beta_1 + \beta_2 X_i + u_i$	$\beta_2 Y$	$\beta_2 X$
Linear-Log $Y_i = \beta_1 + \beta_2 \ln X_i + u_i$	$\beta_2 \dfrac{1}{X}$	$\beta_2 \dfrac{1}{Y}$

Interpretation of the coefficients.

2. Qualitative explanatory variables.

   Consider the example about the chain of shops:

$$S_i = \beta_1 + \beta_2\, It_i + \beta_3\, F_i + \beta_4\, I_i + u_i \quad i = 1, 2, ..., 350$$

Population Regression Function.

$$E(S_i|X) = E(\beta_1 + \beta_2\, It_i + \beta_3\, F_i + \beta_4\, I_i + u_i|X) = \beta_1 + \beta_2\, It_i + \beta_3\, F_i + \beta_4\, I_i$$

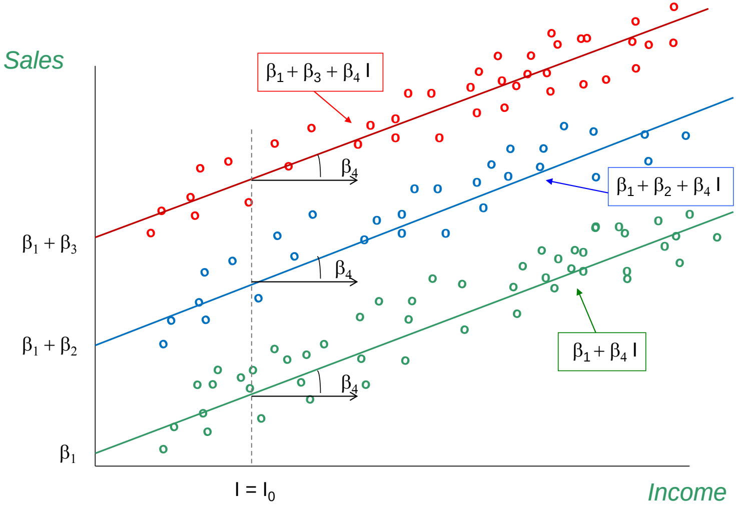
Spain: $E(S_i|I_i, It_i = 0, F_i = 0) = \beta_1 + \beta_4\, I_i.$

Italy: $E(V_i|I_i, It_i = 1, F_i = 0) = (\beta_1 + \beta_2) + \beta_4\, I_i$

France: $E(V_i|I_i, It_i = 0, F_i = 1) = (\beta_1 + \beta_3) + \beta_4\, I_i$

Interpretation of the coefficients.

Graphic analysis.



Interpretation of the coefficients.

Coefficient of variable income.

▷ $\beta_4 =$ The expected change in the value of sales when income changes in one unit holding the variable country fixed.

Intercept.

▷ $\beta_1 = E(S_i | I_i = 0, It_i = 0, F_i = 0)$

Expected value of sales in the Spanish shops when the value of income is zero.

Coefficients related to the dummy variables.

▷ β_2, β_3 **do not have a slope interpretation** because the dummy variables are not continuous; they are binary variables.

Spain: $E(S_i|I_i, It_i = 0, F_i = 0) = \beta_1 + \beta_4\, I_i$

Italy: $E(S_i|I_i, It_i = 1, F_i = 0) = (\beta_1 + \beta_2) + \beta_4\, I_i$

$$\Longrightarrow \quad \beta_2 = E(S_i|I_i, It_i = 1, F_i = 0) - E(S_i|I_i, It_i = 0, F_i = 0)$$

$\Longrightarrow \quad \beta_2 =$ Expected difference in the value of sales between Italy and Spain, holding variable income constant.

Interpretation of the coefficients.

Spain:

$$E(S_i|I_i, F_i = 0, It_i = 0) = \beta_1 + \beta_4 \, I_i$$

France:

$$E(S_i|I_i, F_i = 1, It_i = 0) = (\beta_1 + \beta_3) + \beta_4 \, I_i$$

$$\Longrightarrow \quad \beta_3 = E(S_i|I_i, F_i = 1, It_i = 0,) - E(S_i|R_i, F_i = 0, It_i = 0)$$

$\Longrightarrow \quad \beta_3 =$ Expected difference in the value of sales between France and Spain, holding variable income constant.

Contents