1. Introduction

1.1 What is econometrics?

Econometrics deals with the quantitative study of economic relations. It is a tool to interpret reality in the light of economic theory, using statistical techniques.
The starting point is always a question that requires a quantitative answer. This can be for example the determination of which fraction of disposable income is consumed and which spared, of the effect of the increase in the price of a good or asset on the quantity demanded, of what happens to a stock index if the Central Bank raises the interest rate, of the effect of an increase in advertising expenditure on the sales of a product, of the extent to which public investment can increase economic growth, of the amount by which real wages vary when productivity increases, of the effect of a reduction in the rate of pollution on health spending, of the increase in aggregate investments if business taxes are reduced, of the interest rate on deposits or loans offered by a bank to different types of customers, of the effect of increased research and development spending on the number of registered patents, of the changes in the likelihood of losing the job determined by the level of education, of the effect of terrorist attacks on the growth rate of an economy, of the impact of a change in oil prices on the price of gasoline, of the extent of the reduction in exports when the exchange rate appreciates, …
Virtually every economic question that requires a quantitative response may be the subject of an econometric study.

1.2 Elements of an econometric study

Once you have identified the problem you want to tackle, you must assess whether and at what level of accuracy economic theory provides guidance for its solution.

For example, if you want to determine which fraction of disposable income is consumed and which spared, you can resort to the microeconomic theory of optimizing consumers or to Keynesian macroeconomic theory, depending on whether the interest focuses on a specific consumer or the entire community. Assuming you want to study the entire collectivity, Keynesian theory links the consumption level to that of disposable income:

$$C = C_0 + c * Y_d \,, \tag{1.2.1}$$

where C indicates aggregate consumption, C_0 autonomous consumption, Y_d disposable income and c the marginal propensity to consume. Moreover, it must be $C_0 > 0$ and $0 \leq c \leq 1$. The proposed relationship between consumption and income is therefore very precise but to calculate savings as

$$S = Y_d - C = (1-c) * Y_d - C_0 \,, \tag{1.2.2}$$

we need to know the precise values of c and C_0.
Also, the answer to the original question is based on the assumption hat the Keynesian theory is correct, but there is no consensus about this at the theoretical level. For example, the alternative life cycle theory suggests that consumption depends not only on income but also on wealth.
Therefore, we need to provide a value for the parameters c and C_0, but also to check whether the Keynesian theory is correct or not. Providing a value for c and C_0, econometrics allows *a more accurate description of the economic reality*. Veryfing whether consumption also depends on wealth, econometrics allows you to *test hypotheses about the validity or otherwise of an economic theory*.
For other questions of interest – for example, to evaluate the effects of a reduction in the rate of pollution on health spending –, there is no specific economic theory. In these cases, the role of econometrics is, therefore, all the more important because it allows you to *derive empirical regularities* from the analysis of the economic reality, which can then *provide an opportunity to develop an appropriate economic theory*.
Having established an economic theory of reference, or lack thereof, the next step is to develop an *econometric model*. This typically requires specifying a relationship between the expected value of the variable of interest and potential explanatory variables, as suggested by economic theory or empirical observation, with additional assumptions about the difference between the actual and the expected value of the variable.
Continuing with the example of consumption, and taking the Keynesian theory as valid, an econometric model is expressible as:

$$E(C) = C_0 + c * Y_d \,, \tag{1.2.3}$$

$$C - E(C) \sim N(0, \sigma_C^2) \,, \tag{1.2.4}$$

where $E(C)$ indicates the expected value of consumption and N the Normal (Gaussian) density. Combining (1.2.3) and (1.2.4), we get:

$$C \sim N(C_0 + c * Y_d, \sigma_C^2) \tag{1.2.5}$$

In the example of pollution, although economic theory does not help us, we can assume that the expected value of health spending (*SS*) is still linearly related to the pollution level (*LI*),

$$E(SS) = a + b * LI, \tag{1.2.6}$$

and that deviations of SS from its expected value satisfy:

$$SS - E(SS) \sim N(0, \sigma_{SS}^2). \tag{1.2.7}$$

Hence, the econometric model becomes:

$$SS \sim N(a + b * LI, \sigma_{LI}^2) \,. \tag{1.2.8}$$

The problem is now better defined: the parameters that we want to estimate, or on whose size we want to conduct tests, are those of the expected value of a variable that has a certain Normal distribution, in the case of the examples. Statistical theory for parameter estimation and hypothesis testing are then helpful, as we shall see in detail in the following chapters.

In order to apply statistical theory to the econometric model, it is necessary to collect a sample of data that provides relevant information about the parameters of the model. Continuing the previous example, to determine which fraction of disposable income is consumed and which spared — to estimate the parameters C_0 and c in (1.2.5), we need data on aggregate consumption, on disposable income and, possibly, on wealth. Similarly, in order to assess the effect of a reduction in the rate of pollution on health spending, which is to estimate the parameters a and b in (1.2.8), we need measurements of the rate of pollution and data on health care expenditures. We will see in the next section that the data can be of many different types, and we will therefore also need to select the most appropriate one for the application of interest.

Assuming at this point that we have estimated the model parameters, it is necessary to interpret them correctly. With reference, for example, to equation (1.2.3), the parameter C_0 indicates that if the disposable income is 0, then the expected consumption will be equal to C_0. If instead there is a marginal change in disposable income, then there will be a corresponding variation in the expected value of consumption given by $c * \Delta Y_d$. We will return in more detail to the interpretation of the results in the next chapters, but it is good to clarify here three concepts.

First, we do not know the model parameters, rather we estimate them using statistical procedures and so there is a *more or less broad uncertainty around their values* that must be borne in mind when interpreting the results. For example, if the estimated value for the marginal propensity to consume, c, is 0.8, with a confidence interval at 90% for c of [0.4, 1.2], then we need to keep in mind that the effects of a change in disposable income on consumption are very uncertain. This is particularly important when the results of the econometric study are used to support economic policy decisions. For example, on the basis of the estimated coefficient b in (1.2.6) you may decide to allocate more funds to reduce pollution to obtain savings on health spending. However, if there is substantial uncertainty regarding the value of b, then the positive effects on health spending are very uncertain, while the costs to reduce pollution are certain.

Second, you might think to use the model to assess the effects on the variable of interest of changes of any size in the explanatory variables, and not just their marginal variations as previously indicated. For example, you may want to estimate the effects of a 10% increase in disposable income on aggregate consumption to gauge whether a substantial tax reduction will help to revive the economy. The problem is that when the explanatory variable changes substantially even the model parameters could change. For example, if the massive increase in disposable income is associated with a tax reduction, consumers might decide to reduce the marginal propensity to consume and increase the savings rate to cope with possible future tax increases needed to rebalance the fiscal budget. This interpretation problem was noted by Bob Lucas in the 1970s and is known as the *Lucas Critique*.

Third, the causal interpretation of the results of an econometric model is sustainable only when the model is based on economic theory. For example, although there is disagreement on other details, all theories of consumption agree that an increase in income leads to an increase in consumption. Therefore, an econometric model in which the estimated marginal propensity to consume is positive and of meaningful size can be used to support the statement that income causes consumption.

If instead the econometric model is not based on a valid and shared economic theory, the fact that a variable x has an estimated coefficient significantly

different from 0 to explain a variable y does not imply that x causes y. This is because the estimators of the parameters of the model, as we shall see in detail in the next chapter, rely fundamentally on the statistical notion of covariance, which cannot be used to substantiate causal relationships. For example, if y causes x, but the econometric model assumes by mistake that y depends on x, then x will typically have a significant estimated coefficient, but from this we cannot infer that x causes y. Similarly, when the relationship between x and y is not based on economic theory, the link between x and y might be spurious and due for example to a third variable z is not included in the model.

1.3 Data

In order to apply statistical techniques to the econometric model, it is necessary to have a sample of data. In experimental sciences, such as physics or biology, data are the result of experiments and are therefore available in virtually unlimited quantities (apart from a cost factor). Instead, in economics the data are usually not the result of experiments but choices and outcomes of real actions of economic actors. For example, we cannot ask ourselves what Mr. Smith's consumption would be if he had an income of EUR 1,000,000, we can only observe what Mr. Smith's consumption is given his actual level of income.

The problem of not being able to repeat the experiment with Mr. Smith is partly compensated by having data on many individuals (or units, more generally). There are for example polls where a vast group of people are asked to indicate their income, consumption and a host of other potentially interesting characteristics for the econometric analysis. Data with only one observation for a large number of units are referred to as longitudinal or *cross-section*.

Another partial solution to the non-availability of experimental data in econometrics is the possibility to observe the features and choices of the same individual or unit over an extended period of time. For example, if we are interested in studying the relationship between aggregate consumption and disposable income, we have a very long sample of observations on these two variables. Data consisting of a number of observations over time for the same unit are defined *time series*.

As a third possibility, we could have a sample with both a longitudinal and a temporal dimension. Such data are defined *panels*. For example, we could observe the consumption and income of several individuals for several months, or aggregate income and consumption of different countries for several years.

Note that when we use cross-section, time series or panel data we make an implicit assumption of homogeneity of the econometric model across units and/or over time. For example, it is assumed that several individuals have the same marginal propensity to consume, or that this remains unchanged over time for the same individual, or that both of these propositions are valid in the case of panel data. This assumption is required to have a sufficiently large sample of observations to ensure that statistical procedures (e.g., parameter estimation or hypothesis testing) give reliable results.
Finally, it is worth noting that the examples considered so far include *continuous variables*, such as consumption, income, the rate of pollution or health spending. There are also cases where the variable of interest or those that are used to explain it are instead of *discrete type*. For example, we might be interested in assessing what determines whether an economy is in recession or expansion, by associating the value 1 to the phases of expansion and the value 0 to those of recession, so that the variable of interest is binary or dichotomous. Or we might want to explain what determines the choice of a group of individuals to enroll in the University or not, to grant a mortgage or not, or to buy a certain product or not. Even in these cases the choices we want to explain can be represented by a binary variable. We will see in Chapter 9 that when the variable of interest is binary (or, more generally, discrete), it is necessary to adopt different econometric models from those used in the case of continuous dependent variables.

1.4 The descriptive analysis

A very important component of an econometric study is the descriptive analysis, which precedes and often simplifies the specification of a formal model.
This requires you to produce and analyze simple descriptive statistics for the variables under consideration, such as the mean, the variance or the correlation, possibly for the whole sample and subsamples. A graph of the behavior of the variables is also always recommended.
These simple steps can already provide useful guidance on which variables are potentially more important to explain that of interest, on the stability of the relation in the sample and/or the presence of abnormal observations, different from most others, that could distort the results of the analysis.
It is also important to consider whether and to what extent the available data represent a good approximation for the theoretical variables to which they relate. While there are typically no problems for variables such as interest rates or the number of employees, for other notions such as the potential output or

the natural rate of unemployment or the expectation of a future variable, the matching with the available data is much less clear. We will see that these cases require special treatment in order to avoid possible distortions in the results.
Finally, it is also useful to include in the descriptive analysis a discussion of the institutional context, although this cannot be fully formalized in the model. For example, the presence of exceptional events such as wars or substantial increases in the price of raw materials, but also temporary changes in legislation on value added tax, can have a significant impact on the relationship between income and consumption. We will see how to, at least partially, take into account these events in the formulation of the econometric model.

1.5 Some examples

Various computer softwares are available to conduct econometric analysis. Among them we chose EViews for its popularity, ease of use, flexibility and the availability of an online clear and complete user manual (examples and exercises have been carried out with version 7 but run with version 8 as well).
EViews workfiles use the extension ".wf1" and all those related to the examples and exercises in this book can be downloaded from the webpage www.igier.unibocconi.it/marcellino.
Each workfile has a predetermined data type (undated to be used for cross-sectional data, dated for time series data, or panel) and sample size. The data type and sample size should be indicated when creating a new file or automatically loaded when opening an existing file. Each workfile can then contain different types of objects, such as the "series" objects that contain data on the variables to be considered in the study.
By opening as an example the workfile "example_cons_chap1.wf1", we see from Figure 1.1 that it contains the following items.

- Two so-called default objects, used by the program for internal processing: a vector of coefficients c, denoted by the β icon, and the series "resid", with the  icon typically associated with data series.
- The series labeled CONS, YD, WEALTH and DEFLATOR, containing, respectively, data on nominal private consumption, disposable income and wealth, all at the aggregate level, and the GDP deflator for Italy. Data are sampled at a quarterly frequency and expressed in millions of euros, with the first observation relating to the first quarter of 1990 and the last one to the first quarter of 2012.
- Other series and objects that we will define shortly.

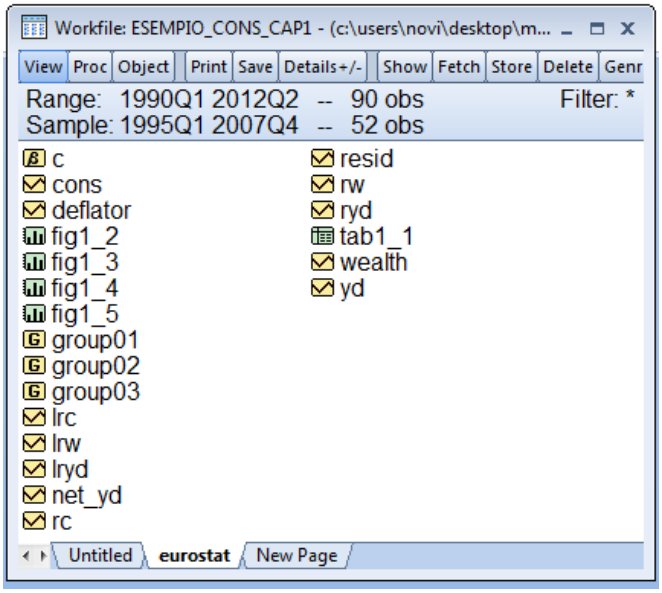

Figure 1.1: An example of an EViews workfile

The four variables CONS, YD, WEALTH and DEFLATOR, reporting national accounts data, can be used for a basic econometric study of the aggregate consumption function. For example, we could compare the simple Keynesian theory, where consumption depends only on current disposable income, with the life cycle theory of consumption, where wealth also plays an important role.

In this first example it is convenient to start the sample in 1995 and end it in 2007, excluding from the analysis the complex years of the financial crisis. To work with a subsample, we simply write in the command line of EViews:

```
smpl 1995 2007
```

indicating, after the command "*smpl*" the start date of the subsample (1995) and then the end date (2007). It is convenient to also introduce the commands "@first", to indicate the first observation in the sample (e.g., smpl @first 2000), "@last" for the last observation in the sample (e.g., smpl 1995 @last) and @all (e.g., smpl @all is equivalent to smpl 1990 2012).

It is always appropriate to start an econometric analysis with a graph of the temporal (and/or cross-sectional) evolution of the variables under consideration. To do this, you can highlight the four series by clicking with the left mouse button on their name (and holding down the Ctrl key), then click with the right mouse button and select the "open group" option from the scroll down menu, click on "View", and select the options "graph" and then "line & symbol", leaving all the other options at their default values. The result is shown in Figure 1.2:

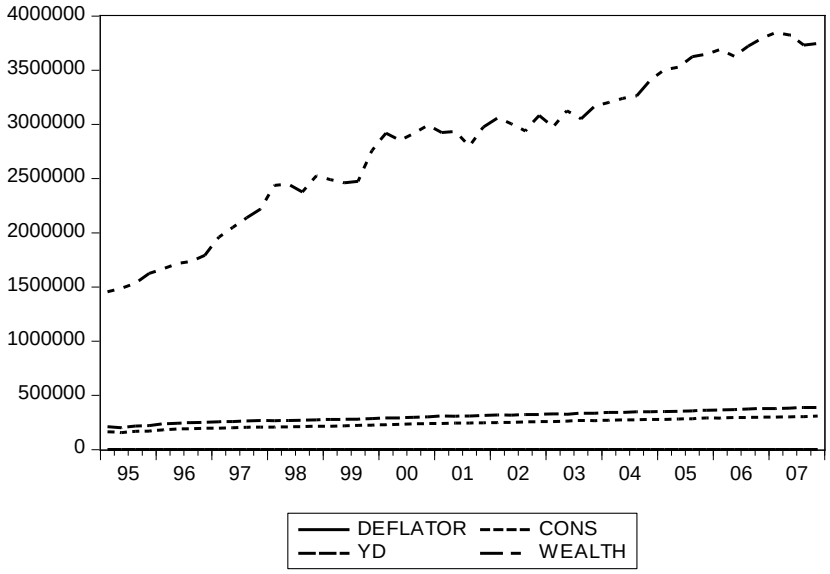

Figure 1.2: An example of a graph of the variables

Many other types of charts are available in the menu and the characteristics of the figure can be changed by clicking on it with the right mouse button and selecting "options". For example, in Figure 1.2 the GDP deflator is hardly distinguishable due to the different measuring scale of the variables. From the menu "Axes and scaling", we can click on the "+" sign and get some submenus such as "scaling". Choosing "normalized data" and the box "left axis scaling method", EViews subtracts to each variable its sample mean and divides the resulting variable by its standard deviation. The result is shown in Figure 1.3.

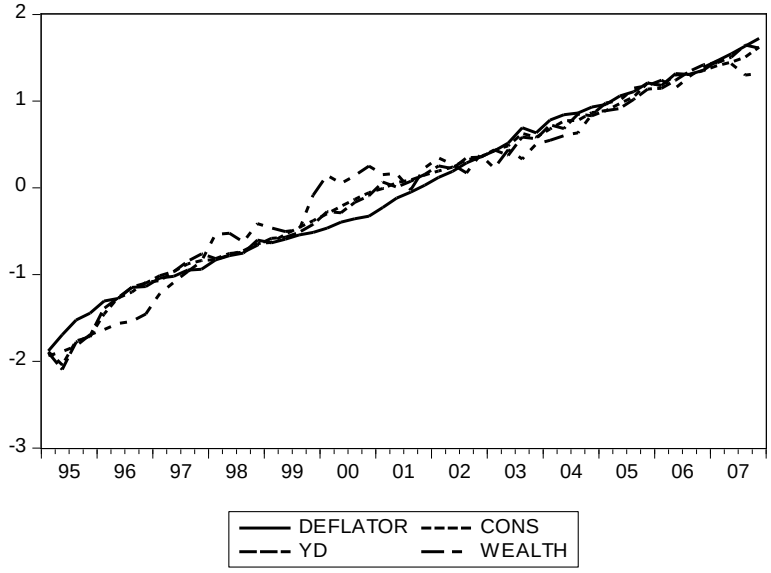

Figure 1.3: An example of a graph of standardized variables

After being "frozen" (by clicking the button "freeze"), each figure can be saved in the EViews workfile by assigning it a name (from the menu "name") or as a separate file (by clicking with the right mouse button and choosing "save graph to disk"), to be then imported directly into other computer software.

The variables that we have considered so far are expressed in nominal terms while economic theory typically refers to real variables. We can create real variables by dividing the nominal variables by the GDP deflator. We use the commands:

```
series rc = cons/deflator
series rw = wealth/deflator
series ryd = yd/deflator
```

and then present a graph of the real variables in Figure 1.4.

Comparing Figures 1.2 and 1.4, we note that the pattern of the three real variables is quite similar to that of their nominal counterparts, with an increasing trend for real wealth at least throughout the first half of the sample, and a fairly constant relationship between real consumption and real disposable income.

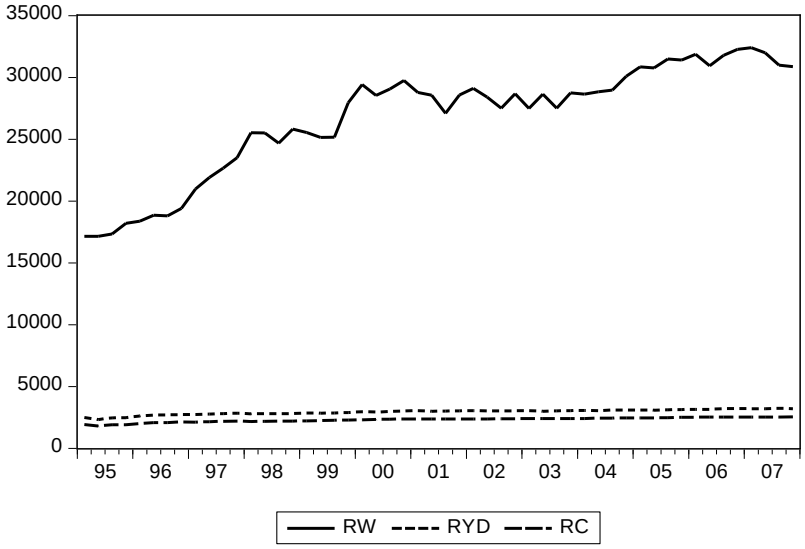

Figure 1.4: An example of a graph of real variables

The logarithmic (log) transformation is also common in *econometrics*, when the variables take non-negative values, as in the case of income, consumption and wealth. The range of data values can be so reduced and made more consistent across variables, allowing also a decrease in the volatility of the variables and making more apparent linear relations across them. The log transformation is also monotonic and strictly increasing, so that it leaves unaffected both variable trends and the position of the minimum and maximum values, if any. We will see though that, being the logarithmic transformation nonlinear, we should pay attention to the interpretation of the parameters in the model for the transformed variables.

To run a logarithmic transformation with EViews, we can use the following commands:

```
series lrc = log (rc)
series lrw = log (rw)
series lryd = log (ryd)
```

and the resulting variables are graphed in Figure 1.5.

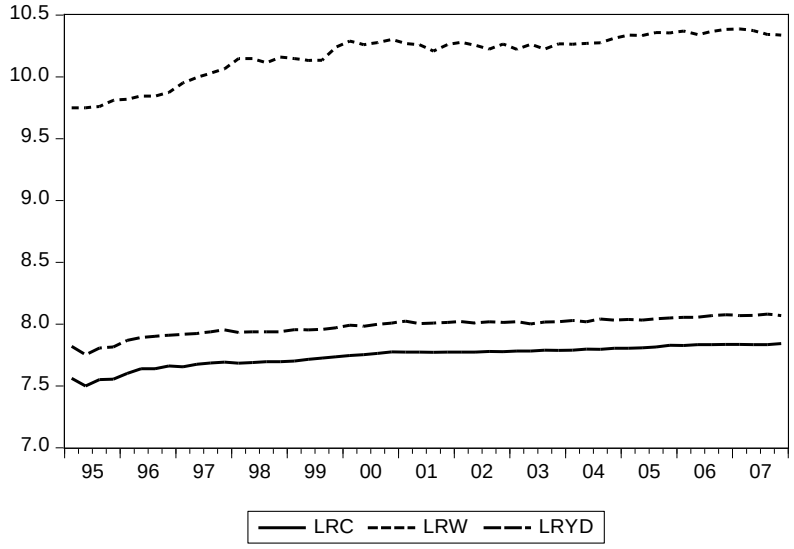

Figure 1.5: An example of a graph of log transformed variables

In Figure 1.5, there seems to be a high and positive correlation between the logs of real consumption and real income. Therefore, there is also a high and positive correlation between the two un-transformed variables (since the log transformation is monotonic). The growing trend of wealth could also generate a positive correlation with consumption. To calculate the three correlations, we select the three variables LRYD, LRC and LRW and from the Group object we click on "View", "Covariance analysis", and check the box "Correlations". The result, shown in Table 1.1, confirms our expectations. Interestingly, the positive correlation between consumption and wealth is in line with the theory of the life cycle, although the latter refers to the consumption-wealth link when also conditioning on disposable income, while the correlations in Table 1.1 are unconditional.

	LRC	LRW	LRYD
LRC	1.000000	0.962472	0.994394
LRW	0.962472	1.000000	0.951921
LRYD	0.994394	0.951921	1.000000

Table 1.1: Correlations between the logs of real variables

As a second example, the workfile named "example_regional_chap1.wf1" contains data on total real wages (W), employment (E) and per worker productivity (PR) in all the 20 Italian regions, for the period 1995-2009. It is therefore a *panel* type dataset, where the longitudinal dimension is represented by the regions. If we focus on data for a single year for all the 20 regions we obtain a *cross-sectional* dataset, whereas if we consider data from 1995 to 2009 for a given region we have a *time series*, as in the first example. In both cases, however, the overall sample size would be somewhat reduced – hence the usefulness of taking advantage of both the temporal and the longitudinal dimensions.
We may construct graphs of the variables even for panels, but the large number of series to be considered suggests using other tools. For example, we can build a table reporting the temporal average of the three series for each region. This can be done manually, selecting for each variable the option "Descriptive Statistics" in the menu "View", or by writing a simple EViews program, which is a series of commands that tell EViews to compute the quantities of interest.
Writing programs may seem more time consuming than using the EViews menus but it can actually save a lot of time when the same or similar actions or analyses have to be repeated several times, for example for different countries or when periodically updating the data.
After clicking on "File", "New", "Program", the list of commands required by EViews to generate the table of interest is reported below, where comment lines are preceded by the symbol " **'** " and are not read by EViews when running the program.

```
smpl 2007 95 ' specifies the sample amplitude (in this example too, we leave aside the years of crisis)
! q = 1 ' defines a counter variable whose value is equal to 1
table (20.4) medie ' defines a table object with 20 rows and 4 columns, called "medie"
for %s LOM PIEM VAL FRI FRI LIG EMIL TOS UMB MAR LAZ ABR MOL PUG BAS CAL CAM SIC SAR TREN ' defines a "loop" on all regions
series {%s}ow={%s}w/{%s}e ' generates wages per employee
!q=!q+1  ' updates the counter
medie (!q,1) =%s ' replaces the !q,1 element of the table with the region's name %s
medie(!q,2)=@mean({%s}e) ' replaces the !q,2 element of the table with the average of E for region %s
medie(!q,3)=@mean({%s}pr) ' replaces the !q,3 element of the table with the average of PR for region %s
medie(!q,4)=@mean({%s}ow) ' replaces the !q,4 element of the table with the average of OW for region %s
next ' indicates the end of the loop (i.e., the code switches to the next region)
!t = 1 ' sets another counter
for %u EMPL PROD WAGEPE ' defines another loop
!t = !t+1 ' updates the counter
medie (1,!t) =%u ' defines the header of each column in the table
```

```
next ' indicates the end of the loop
smpl @all ' reconsiders the whole sample
```

The result is a new object in the workfile, a table named "medie", shown in Table 1.2. The table points out sizable differences across regions in the total number of employees, caused by different surfaces and population, but also by differences in the average productivity of labor, partly reflected in the wages per worker.

	EMPL	PROD	WAGEPE
LOM	4313.023	59.54618	23.51026
PIEM	1894.546	53.07935	21.49163
VAL	56.66923	56.96148	21.49799
VEN	2132.685	52.59079	20.96652
FRI	549.7846	50.18644	22.18595
LIG	638.2231	53.55893	21.20674
EMIL	1980.431	53.29711	21.29569
TOS	1579.077	51.70392	20.35534
UMB	355.4769	47.61522	19.30201
MAR	668.8538	46.68140	19.01356
LAZ	2250.577	59.01183	24.65786
ABR	487.8462	46.63591	19.21301
MOL	114.8846	42.99700	17.88512
CAM	1756.277	44.90474	19.18502
PUG	1268.785	43.84201	19.19500
BAS	204.0077	42.47083	18.87000
CAL	611.9462	43.35827	18.61562
SIC	1443.938	47.27515	20.19538
SAR	579.2923	45.83074	19.32625
TREN	454.2385	54.04006	22.26839

Table 1.2: Average values of the variables for each region

So far in our examples we focused on sample averages and correlations. In the online guide of EViews, available under the "Help" button, "EViews Help Topics" under the heading "Descriptive Statistics" you can find a list of available commands to compute other descriptive statistics, with explanations and examples.

The data of the two previous examples were already in the form of an EViews workfile. However, you will typically download data from large databases, such as

those of Eurostat or the OECD, as text files or Excel spreadsheets. Assuming that you have data in Excel, for example in the file "example_regional_chap1.xls", you can import them into an EViews workfile by using the following procedure (in a similar way, you can import data into other formats).

- Create the workfile by clicking on the main window of EViews on "File", "New", "Workfile". You get the window reported in Figure 1.6.
- From the "Frequency" menu, for this example you should select "Annual"; in the "Start date" field, type 1995, the earliest date for which observations are available in the Excel file; in the "End date" field, type 2009. By clicking OK you get the EViews workfile (containing the default objects only).
- Click on the main window of EViews on "File", "Import", "Import from file". From the resulting menu, browse your folders until you find the "example_regional_chap1.xls" file. The Excel file should be closed and the data to be imported must be in the first worksheet.
- Click on the file, getting the window in Figure 1.7.

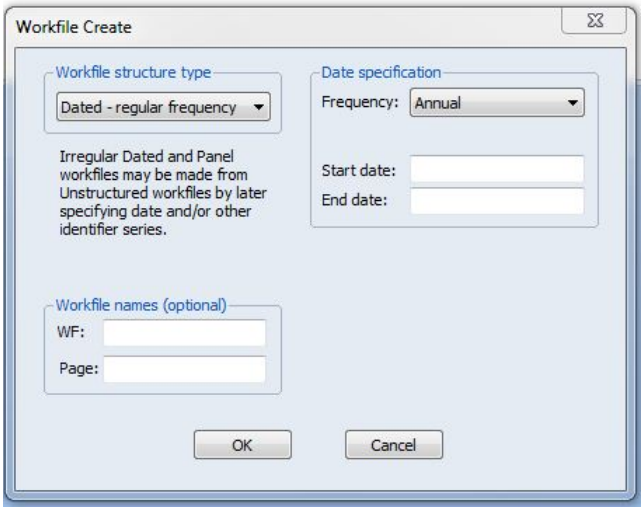

Figure 1.6: Creating a new Workfile

By clicking on "Next", you reach a similar screen where you can tell the program how many rows match the header file. In our case, it is a single line, so we just write "1" in the box "Header lines" and click the "Finish" button. At this point, the data are imported into the workfile, and series are created with the name specified in the Excel sheet.

You can follow a similar procedure to export data from Excel to EViews. In addition, EViews interacts with external programs to also allow the export of output. For example, by clicking with the right mouse button on a graph of EViews and choosing "Copy" and the option "EMF – enhanced metafile", you copy the chart to the Clipboard – ready to be pasted, for example, in MS Word.

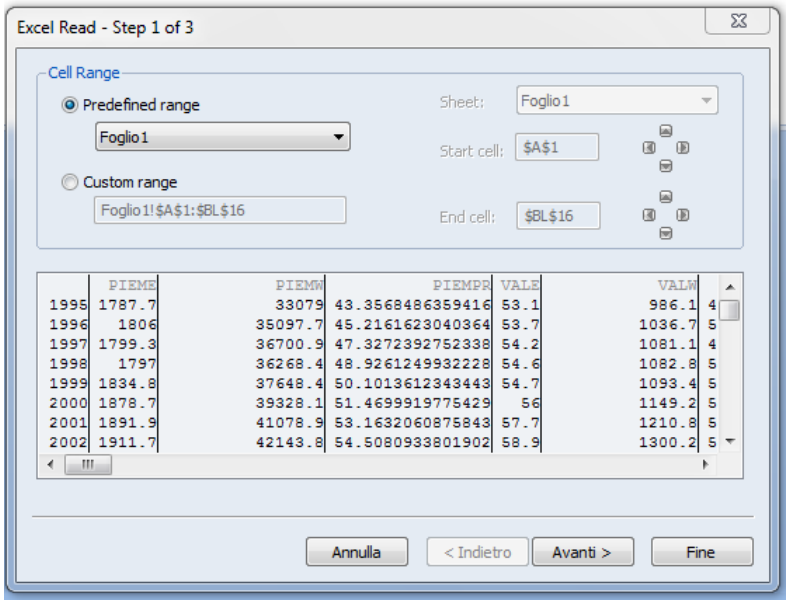

Figure 1.7: EViews window to import data

The main concepts of this chapter

In this chapter we saw that econometrics deals with the quantitative study of economic relations. econometrics males it possible to more accurately describe the economic reality, allowing you to test hypotheses about the validity or otherwise of an economic theory. When there is no specific economic theory, the role of econometrics is even more important, because it allows you to derive empirical regularities from the analysis of economic data, which can then provide an opportunity to develop an appropriate economic theory.
The results of econometric analysis should be interpreted with caution. We do not know the parameters of the model, we estimate them using statistical procedures, and so there is a more or less broad uncertainty around their values, which must be borne in mind when interpreting the results. The

problem is that when the explanatory variable changes substantially, the model parameters could change too. This interpretation problem was spotted by Bob Lucas in the 1970s and is known as the Lucas Critique.

In addition, if the econometric model is not based on a valid and broadly accepted economic theory, the fact that a variable x has an estimated coefficient significantly different from 0 to explain a variable y does not imply that x causes y.

We have then considered different types of data sets, which require different techniques of analysis. Data sets composed of a single data point for a number of units are referred to as *cross-sections*. Data sets consisting of many temporal observations for the same variable are known as *time series*. Data sets with both a longitudinal and temporal dimension are defined as *panels*. In addition, variables can be continuous, as consumption, income, the rate of pollution or health spending, or discrete, as in the case of binary variables.