AN INTRODUCTION TO APPLIED ECONOMETRICS

(Lecture notes)

Jean-Pierre Laffargue
University of Paris 1, PSE and CEPREMAP

CONTENTS

INTRODUCTION

These notes are intended for students having no knowledge in econometrics and little knowledge in statistics and in probability.

For a long time, the tradition in France was to teach econometrics the hard way. In a first stage, students had to learn plenty of mathematical results on various classes of estimators and tests. During this time they had to believe that their arid investment will be profitable in the future and will allow them to deal with economic data and to answer economic problems. Later on they could turn to applications.

A problem with this method is that many students became discouraged during the first step of the process. Another problem was that many students who had reached the second step had a tendency to turn to very sophisticated and fragile methods when they faced simple practical problems. Sometimes, the results they reached were crazy when they mixed very complicated methods, with very elementary mistakes contradicting basic common sense. The most serious mistakes in econometrics, which can even be found in articles published by good journals, come from not spending time enough looking at the data, in a pragmatic way, without a priori and without the strong desire to apply complicated methods that are totally inappropriate for these data.

These notes follow a completely different principle. I will introduce econometrics through a series of simple applications. I will use little mathematics, and I will be little rigorous. I will appeal to the common sense and the intuition of the reader to introduce the basic concepts, methods and traps of econometrics. I will try to show that econometrics is simple, and thinking in an econometric way is the same as thinking in an economic way. Sometimes, the developments will be a bit tricky, and I hope as funny as the kind of riddles and puzzles you can find in newspapers and magazines. The book by Berndt (quoted among the references) is entertaining and pleasant to read (with much gossip on the profession, so you can discover that econometricians are also human beings). Finally, econometric methods give answers to economic questions, and these answers must be understandable and look convincing to people who are experts of these questions and not econometricians. So, introducing students to econometrics through applications is sensible.

There is a limit to the approach followed in these notes, and students are expected to feel it more and more when they progress in this course. Examples and intuition quickly meet their limits and to go further we must use logical and rigorous methods. So mathematics is unavoidable, and, after having read these notes students must learn a book of econometrics, which includes the mathematical foundations of this field. However, doing that in a second stage of learning, after having gone through these notes, will be a task much easier than starting directly with the mathematics of econometrics.

There are many user-friendly econometric software. I will advise you to use E-Views or Stata. Both are known and used in the whole world and it is much wiser to learn and use a software that is a world standard. Loosely speaking E-Views is well adapted to macroeconomics (time series-data) and Stata to microeconomics (individual data). However, the availability of such user-friendly software may encourage laziness and the absence of reflection. This is a pity because these software include powerful graphic capacities, and plenty of descriptive statistics, which are extremely precious to look at the data and to learn much of them[1].

Alexander Pope wrote: “Little learning is a bad thing”. Henri de Monfreid tells a nice story. When he was smuggling weapons on the Red Sea, he had a good friend, who was a policeman and the only French official on a small island near Djibouti. He told of his friend that he had the wisdom of men of the people who had not been spoiled yet by compulsory public education. Both comments are a bit arrogant and even reactionary, but they are basically true. Many applied econometricians use sophisticated methods, which were developed by experts in the field, and they apply them to their problems and data without further reflection. This imitation process has become very easy with the existing software and the availability of many programs on the websites of Eviews and Stata. In general these methods were correct for the problem they were designed for, and their developers did not make mistakes in their applications. But the adaptation of these methods to problems they were not designed for can be awfully wrong. So, these applied econometricians make logical mistakes and draw silly conclusions. In my lifetime I read many strange papers of applied econometrics with results, which were meaningless, incomprehensible and unbelievable. These papers generally were in development economics and macroeconomics, but this can result from the fact that most of my readings are in these fields. These applied econometricians had a superficial knowledge of theoretical econometrics and tried to substitute recipes to logic. So, they had “little learning”[2]. However, they could have avoided their mistakes if they had not lost their common sense, the wisdom of the ignorant. These notes are quite insufficient to help you to solve the first problem. However, they will give you advice to help you not to progressively lose your common sense when you become more and more learned. These advices can be summed up in tow sentences. First, do not forget that you are an economist and that your econometric results must be explainable in plain French (or English) and without cheating, to a non-econometrician economist. Secondly, look carefully at the data and do not apply a method based on assumptions, which are contradicted by these data. In summary, econometrics must not make you lose your common sense.

Econometrics is a set of quantitative tools for analysing economic data. Economists need to use economic data for three reasons: 1) to decide between competing theories; 2) to predict the effect of policy changes; 3) to forecast what may happen in the future. Three examples: Have PC increased the productivity of clerks and secretaries; How to evaluate and compare the efficiency of various policies against AID in Africa; How to forecast the demand for public transportation in a big city?

Economists deal with different kinds of data:

1. Time series data. For instance GDP data are collected every quarter. Macroeconomics and finance use such data. In macroeconomics frequencies are

[1] If you want to write sophisticated programs, to solve computational problems, you’d better use Gauss or Matlab.

[2] This insufficiency can be combined with straight dishonesty. A complicated scientific method can be manipulated to defend conclusions, which agree with your ideology or your interest. This unfortunate situation especially happens in fields where there are hard political and ideological debates and conflicts.

annual, quarterly or monthly. Frequencies are much shorter in finance. I will use the following mathematical notation for a variable or series: Y_t, with: $t+1,2,..,T$. T represents the number of observations.

2. Cross-sectional data. For instance in a labour survey you interview 1000 workers of the chemistry industry on their wages, their labour conditions, etc. All these interviews take place at about the same date. Each question gives you as many answers as interviewed workers. Let us take for instance wages. I will use the following mathematical notation: Y_i, with: $i=1,2,..,I$. Y_i represents the wages of worker i, and I is the number of surveyed workers. Cross-sectional data are mainly met in microeconomics (observations can bear on workers, households or firms). But, macroeconomics can use such data when it compares different countries (for instance their GDP per head). Time series data and cross-sectional data differ on a very important point. Time is oriented. The past comes before the future. You can use the past to forecast the future, but you cannot use the future to forecast the past. Of course, the past and the present depend on the expectations of the future by economic agents. But, the expected future is based on the experienced past, and not on the true future which is unknown. On the other hand, there is no natural way for orienting cross-sectional data. Because of its specificity the econometrics of time series data is a bit special (and in my opinion it is the most difficult part of econometrics).

3. Panel data. In the above example you can track the same workers for several years and interview them periodically over this period of time. For instance, you can interview the same people on their wages every year for five years. I will use the following mathematical notation for the wages of worker i for year t: Y_{it}, with: $i=1,2,..,I$ and $t=1,2,...,T$. There is a special field of econometrics to deal with this kind of data. In most case, you will have a large number of individual units, and a small number of time periods (5 for example). This is the case in my example, and in most of microeconomics. Some macroeconomists use to work on a panel of countries or regional areas: for instance yearly data of 30 OECD countries on the 1970-2000 period. Their problem is a bit different: a smaller number of individuals (30 instead of 1000) and a larger number of periods (31 instead of 5). It is a controversial question if the traditional econometrics of panel data is appropriate to the kind problems macroeconomists face. The traditional econometric of panels can be a bit tricky and cumbersome, but it is quite consistent with intuition and common sense, and the mathematics it uses is elementary.

4. Quantitative data and qualitative data. GDP is a quantitative data: it takes real values, for instance 900 thousand million dollars. But some other data can take only two values, which in general are not numbers. Examples of such data are genders (male or female), if a worker kept or lost his job for a given year, if a household owns or does not own a car. A difference between these two kinds of data is that quantitative data have a natural order: having a large GDP is better than having a small GDP. But the qualitative data above have no natural order. A more complicated case is when you consider qualitative data, which can take more than two values. For instance a household can own no car, 1 car, more than 1 car. In this case a natural order appears among the three values (but I could build examples when such an order would not appear, like spending your vacation in the countryside, in the

mountains or on the seaside). Sometimes, a qualitative order appears, but you cannot make any quantitative comparisons between the possible values taken by the variable. For instance, you can survey African newspapers to discover if, after an IMF intervention, you will have: nothing, a big strike, big riots, or a revolution. A revolution is worse than a riot, but you cannot tell if it is three times or five times worse. On the other hand you can tell that French people are 8 times wealthier than Tunisian people. However, sometimes quantitative comparisons can be made for qualitative data, for instance if they represent the number of patents applied and received by firms. The econometrics of qualitative data, also called limited dependent variable, is a popular field of econometrics. The basis is simple to understand and to apply. But this simplicity disappears quickly when you try to go a little further than the most elementary cases.

When your analyse data, you quickly discover that it is sometimes better to transform them. For instance, if you want to compare the quality of life between several countries, you should divide your data by the populations of the associated countries. Thus, you will compare the numbers of medical doctors per 1000 people. A useful, but tricky, transformation is to go from a variable in level (GDP for instance) to the growth rate of this variable. The growth rate of variable Y_t will be denoted: $\dot{Y}_t = (Y_t - Y_{t-1}) / Y_{t-1}$. Sometimes, economists prefer the approximation: $\dot{Y}_t = \ln(Y_t) - \ln(Y_{t-1})$. You can multiply these formulae by 100 if you like having your growth rate in percentage points. You can notice that GDP and the growth rate of GDP do not have the same units. GDP is in dollars or euros. Its growth rate is a pure number, for instance 0.03 or 3%. Moreover, GDP has a trend, but its growth rate has no trend. You must not forget these differences when you write an economic equation. For instance it is sensible (and Keynesian) to write that consumption increases with income. But it is queer to write that the growth rate of consumption increases with income (and Keynes never claimed such a thing).

A last transformation is to substitute a variable in level by its natural logarithm. This has two advantages. First, the values taken by the log are much smaller and vary on a much smaller range. Secondly, when you draw a graph of the evolution on the variable relatively to time, if it grows at a constant rate, you get an exponential for the variable in level but a straight line for the variable in logarithm

References

- Marno Verbeek: Modern Econometrics, 2nd edition, John Wiley, 2004.

This is the book I advise students to buy and to read as soon as they have been through my course. In 400 pages it covers the whole field of econometrics at the introductory and medium-advanced levels (so the book can be used as a tutorial and a reference). It mixes econometric theory and applied econometrics, with plenty of examples on real data and interesting problems. The data can be unloaded from the website of the author. The theoretical elements are well explained, without excessive use of abstract mathematics, but with much precision and without mistakes. A difficulty with basic econometrics is that many elementary results are unimportant for applications and should be skipped in a first course. However, sophisticated and recent results are sometimes essential for applications and must be introduced in a first course, but in a simplified way. The author of this book succeeds well in doing that. Some questions, which I think important for practitioners, robust econometrics and the foundations of the concepts of exogeneity, are absent of this book. However, other concepts, which are also very important for practitioners, do not appear in my notes, but are well developed in the book. The author of this book takes strong positions on the various methods he presents, and gives advises on the mistakes many practitioners do when they apply these methods. The first half the book is about general econometrics. The second half presents the econometrics of specific fields: limited dependant variables, time series, and panel data). So, I advise students to buy this book (orders through amazon.co.uk are executed promptly, do not try cheaper but unreliable booksellers on line).

- Peter Kennedy: A Guide to Econometrics, 4th edition, The MIT Press, 1998.

This book is for practitioners and is centred on the mistakes to avoid when you do applied econometrics. Its chapters are divided into three parts. The first, written with large letters, presents very elementary things. The second, written with medium-sized letters, is more advanced. The third, written with small letters, is very advanced. There are no mathematical developments, but there are many theoretical results, some formulae, and mathematics are never very far in the background. Thus, this book is more advanced than my notes (and much more rigorous). There have been (at least) four editions, each much richer than the previous one. Thus, you can see that this book met a tremendous success in the whole world.

- Gary Koop: Analysis of Economic Data John Wiley and Sons, Chichester, 2000.

This is a book written for students in business, who hate mathematics. Many ideas of my notes were taken from this volume

- Chandan Mukherjee, Howard White and Marc Wuyts: Econometrics and data analysis for developing countries, Routledge, London and New York, 1998.

The first version of my course was based on this book. The parts of my notes on robust regressions are mostly taken from it.

- Ernst R. Berndt: The Practice of Econometrics, Addison-Wesley, Reading, 1991.

Each chapter presents the history of an economic problem (the capital asset pricing model, hedonic prices, etc.) and of the solutions applied economists gave to this problem. Each chapter ends with problems using the data used by these economists and asking to reproduce their results or variants of their results. Moreover, each chapter deals with a specific econometric problem (bivariate analysis, multivariate analysis, etc.). The book is extremely pleasant to read and very lively. One or two chapters (those connected to time series econometrics and macroeconomics) have become a little old fashioned, but the others are still fully actual.

- The tutorial manual of E-Views is an excellent book of applied econometrics. You will find it in the help of the softwarefor versions 4 and 5. You will find an excellent course teaching how to use Stata on South African survey data on the website http://saproject.psc.isr.umich.edu/ .

Finally, if you want to go further than these notes, I advise you to attend one or several courses in theoretical econometrics. In this field self learning is difficult, even with the excellent books I quoted above.

CHAPTER 1. DESCRIPTIVE STATISTICS

Graphs

When you have data, the first thing to do is to look at them by drawing well-chosen graphs. Excel and E-Views are complementary instruments to do that. We will go through a few examples.

First, Koop\exruk.xls, gives monthly time series data from January 1947 through October 1996 of the UK pound / US dollar exchange rate. Draw the time series graph. Comment

Second, Koop\gdppc.xls, contains cross-sectional data on real GDP per capita in 1992 for 90 countries in US dollars using PPP exchange rates. Draw the histogram. Draw the kernel density. Notice the bimodal distribution of GDP per head.

Third, Koop\forest.xls, contains data on deforestation, and on population density for 70 tropical countries. Deforestation is the average annual forest loss over the period 1981-1990 expressed as a percentage of the total forested area. Population density is the number of people per thousands of hectares. Draw the scatter diagram between these two variables. Notice the positive relationship, and the outliers.

Mean and other numerical summaries

Sometimes, you would like to sum up the distribution of a variable, for instance the above histogram, by a few numbers. There are two traditional summaries. The mean indicates the value around which all the values taken by the variable are equally distributed. The formula is $\bar{Y} = (\sum_{i=1}^{N} Y_i)/N$, N is the number of observations. The standard deviation is defined by the formula: $s = \left[\left(\sum_{i=1}^{N} (Y_i - \bar{Y})^2 \right) /(N-1) \right]^{1/2}$. It is a measure of dispersion. Compute these two indicators on Koop\gdpp.xls.

Now, we should think a little harder about the previous concepts. A histogram represents the distribution of a sample of observations of a random variable. The histogram is a reflection of the true distribution function of the random variable. The first feature of this function is the level around which it is located, for instance 300 dollars or 30000 dollars. The concept of mean is a choice (among others) of a measure of this location. Sometimes this choice is good. Sometimes it is bad. I will later introduce the equation given by a linear regression. Such an equation will determine the mean (or expected value) of the explained or dependent variable, conditional on the knowledge of the explanatory variables. Thus, the concept of mean is central to econometrics, and we must understand its limits.

There exist three measures of the location of the distribution of a variable: the mode, which is the value the most often observed, the median, which is the value such that the larger values are as many as the smaller values (the centre of probability) and the mean (the centre of gravity).

Some distributions are bimodal (for instance GDP per capita in the above example). In this case the mean is a bad summary. Very often a bimodal distribution suggests that the sample includes two kinds of individuals which should be considered separately, at least at the beginning of the analysis. For instance you can have such a bimodal distribution for the wages of a sample of workers. This may result from the fact that women are uniformly less paid than men.

Some distributions are much skewed to the right, for instance the payment of overhead hours to a sample of workers over different weeks. Then, the mode is smaller than the median, which is smaller than the mean.

When the distribution has the shape of a bell (so is unimodal) and is symmetric, the mode, the median and the mean are equal. In this case, it is justified to compute the mean. Then, the arithmetic mean computed on a sample of observations is a BLUE estimator of the expected value of the underlying distribution. Moreover, if the distribution is normal, the arithmetic mean is the maximum likelihood estimator of the expected value of the underlying distribution.

As the median and the mean are equal, we could think about estimating their common value, not by the arithmetic mean of the sample, but by its empirical median. This would be a bad idea. In the case of a normal distribution and for a large sample, the standard deviation of the estimator median is 1.25 higher than the standard deviation of the estimator arithmetic mean. This means that the empirical median is a less precise estimator than the empirical mean. But for skewed distributions for which the mean has little meaning (for instance for wages), then the empirical median deserves to be computed and looked at.

The concept of robust summary

I will consider the GDP per capita of a sample of sub-Saharan African countries. The first sample includes 7 countries and concludes that in 1990 GDP per capita had a mean of \$354, a median of \$370 and a standard deviation of \$196. The mean appears as an interesting summary of this distribution. Now, in this sample, I will substitute Botswana to Lesotho. Then, the mean becomes equal to \$570, the median remains equal to \$370 and the standard deviation becomes equal to \$673. Thus, by changing only one individual into the sample, I did not change the median, but I changed the mean a lot. So, the mean is not a robust indicator of GDP per capita in African countries. But the median is a robust indicator. There is an economic and a political dimension in this discussion. If we fix to 500 dollars the level under which a country is considered to be poor, the mean indicates that African countries are above this poverty level. The median gives the opposite conclusion.

The problem with the mean is that it minimises the mean square error (explain). So, an outlier has an excessive weigh in its determination.

I consider the distribution of a random variable with mean μ and variance (the square of the standard deviation) σ^2. These indexes are estimated by the empirical mean and variance $\bar{Y}$ and s^2. However, besides μ and σ^2 there exists two other important indices summarising the distribution of the random variable, and which are computed by using the operator expected value, denoted E:

The skewness is: $\alpha_3 = \frac{E(Y-\mu)^3}{\sigma^3}$

The kurtosis is: $\alpha_4 = \frac{E(Y-\mu)^4}{\sigma^4}$

Skewness is zero for a symmetric distribution (hence for a normal distribution). Kurtosis is equal to 3 for a normal distribution. When the kurtosis is larger than 3 we will say that the distribution has a fat tail. If you remember the formula of the normal distribution you can notice that the value of this distribution decreases as the square of the inverse of an exponential when we go further and further from the mean. This means that this decrease is very fast, and, for a sample with a reasonable size the probability of having one observation or more farther than 4 standard deviations from the mean, is practically zero.

Variance, skewness and kurtosis are an arithmetic average, that is they take into account the value taken by the variable for each individual of the sample. Hence, an outlier, that is an observation with a value far from the mean, will strongly affect these three indices. If I want to remove the perverse influence of outliers on the numerical indices which sum up the distribution of the investigated variable, I can compute other indices, which do not take into account the values taken by this variable for each observation, but only the rank of this value in the sample. To do that, I will rank the observations in increasing order. The median is the observation, which divides the sample into two parts of equal size. The lower quartile Q_L is the median of the part of the sample, which is located below the true median. The upper quartile Q_U is the median of the part of the sample, which is located above the true median.

These definitions must be refined to take into account that, for instance if the number of observations is even, there are two observations which apply to be the median. In this case we will take their mean as the median. The same problem occurs for quartiles.

The range is the difference between the highest value and the lowest value observed in the sample. It is very sensitive to outliers. However, the inter-quartile range IQR is the difference between the upper quartile and the lower quartile $Q_U - Q_L$. It is not sensitive to outliers.

An outlier is a point that is located very far outside the IQR. For instance Y_O can be considered as an outlier if: $Y_O < Q_L - 1.5IQR$, or $Y_O > Q_U + 1.5IQR$
We can decrease or increase coefficient 1.5 to define near-outliers and far-outliers.

How can we test if a variable is distributed according to a normal law? The first thing to test is the symmetry of the distribution that is the absence of skewness. The most natural test is to compare if the difference between the mean and the median is large. The importance of this difference is evaluated relatively to the IQR.

This difference may be large, not because the distribution is asymmetric, but because of a small number of outliers. So, it is interesting to measure if the value of the median is near the mean of the two quartiles. If this is the case, the central values of the distribution are symmetrically distributed, and we can expect to have a symmetric distribution after having eliminated outliers. A robust index of skewness is the coefficient of Bewley:

$b_s = (Q_U + Q_L - 2Md) / IQR$

The second thing to test is kurtosis. For a normal distribution we have $o = IQR / 1.35$. So, we can evaluate in a way robust to outliers, if the tail of the distribution is fat or thin by computing the difference between the empirical standard deviation and IQR divided by 1.35. When the tail of the distribution is fat, the empirical median becomes a better estimator of the mean than the empirical mean, which becomes very sensitive to outliers. We can build a 95% confidence interval of the mean as going from rank $\text{int}\,eger((n+1)/2 - \sqrt{n})$ to rank $\text{int}\,eger((n+1)/2 + \sqrt{n})$. The results of these formulae must be rounded up to the nearer, lower integer and to the nearer, larger integer, respectively.

The Jarque and Bera test is more sophisticated but cannot discriminate between the presence of outliers and a true asymmetry or true fat tails. I must compute $Z_3 = a_3\sqrt{n} / \sqrt{6}$ and $Z_4 = (a_4 - 3)\sqrt{n} / \sqrt{24}$ where a_i is the estimator of de α_i. These two expressions follow standard normal distributions. Moreover, for big samples (1000), these two expressions become independent. Thus the sum of their squares follows a χ^2. This is the Jarque and Bera test.

To remove true skewness from a series we can transform it:
Y^3 reduces extreme negative skewness
Y^2 reduces negatives skewness.
$\log(Y)$ reduces positive skewness.
$-1/Y$ reduces extreme positive skewness.

If the transformation is well made the mean and the median of the transformed series will approximately be equal. As this transformation preserves the order of the observations, the inverse transformation of the mean-median will give the median of the original series, but not its mean. In the same way the application of the inverse transformation to the confidence interval will give the confidence interval of the median of the original series. This does not matter very much, because in case of skewness, the mean does not have great economic meaning for the original variable.

CHAPTER 2. BIVARIATE ANALYSIS

Correlation

Let us come back to the file Mukherjee\forest.xls. The scatter diagram between deforestation and population density shows that these variables are related, but that the relationship is imperfect. If we draw a line in the middle of the scatter diagram, we can see that the points of the diagram are distributed around this line, but they are not on the line. The correlation between the two variables is a number, included between –1 and 1, which measures the intensity of their relationship. This intensity is very strong if the correlation is near 1 or –1. It is very weak if this intensity is near 0. More precisely the square of the correlation measures the proportion of the cross-country variability in deforestation that matches up with the variance in population density. In our forest example, the correlation coefficient is equal to 0.66. As $0.66^2 = 0.44$, we can say that 44% of the cross-country variance in deforestation can be explained by the cross-country variance in population density.

I will spend the rest of the paragraph thinking about what "explaining" means. Sometimes I will use the verb to cause instead of to explain. Both words will have the same meaning. This meaning will be the general meaning that people give to these words. Econometricians defined Granger-causality. This is quite a specific and technical concept, and I will not use it in this chapter.

My first example will be about an exercise, which is given in Koop\hprice.xls. This file contains data relating to 546 houses in Windsor (Canada) in the summer of 1987. It contains the selling price (in Canadian dollars) along with many characteristics for each house. First, I can find a correlation of 0.54 between the price and the size of its lot. Thus, we can think that the size of the lot causes the price of the house, which is located on it. If, I have a house and if I buy some land besides it, its price will increase.

Second, I find a correlation of 0.37 between the price of a house and the number of its bedrooms. Thus, I can think that a house with 4 bedrooms will have a higher price than a house with 3 bedrooms. Third, I find a correlation of 0.15 between the size of the lot and the number of bedrooms of a house. This number is surprisingly low. I would have thought that big houses often go with large lots. This relationship exists, but it is pretty weak.

Now, let us think about the strong relationship between the price of a house and the size of its lot. One reason for this relationship is that a house with a big garden will have a higher price than a house without a garden. A second reason is that for a house, large lots are (weakly) connected with large numbers of bedrooms, and buyers are ready to pay for a large number of bedrooms. Thus, there is something spurious about the strong correlation of 0.54 found between price and size of the lot: if I buy some land besides my house, without adding one more bedroom, its price might increase by less than expected. This is the essence of multivariate regressions, which will be considered in next chapter. However, this is also the essence of the main difficulty, which is met by applied econometricians.

I will develop this last idea, which may be called direct against indirect causality, on another example. We can find a strong correlation between holding a university degree and pay. However, does that mean that education increases productivity and earning as the theory of the human capital assumes? The strong correlation could be because people with a university degree are intelligent, and that firms are ready to pay high wages to intelligent people. In microeconomics this is called the theory of screening, which is part of the theory of signalling. The theory of the human capital considers that there is a direct causality from education to pay. The theory of screening considers that there is an indirect causality from intelligence to pay, which passes through education. However, education by itself would be useless. Econometricians developed plenty of tricky methods to determine which, of these two theories is right. You can imagine that using intelligence tests to discriminate between these theories is controversial: such tests are sensitive to social or ethnic backgrounds and are not a wholly convincing measure of intelligence. However, you can build your econometrics, for instance, by comparing twins with different levels of education (one of them did not go to university because of a car accident…).

The last example. If you take a sample of people, you will find a strong correlation between the number of cigarettes each person smokes per week, and on whether they have lung cancer. This result is normal, because smoking causes cancer. You will also find strong correlation between the number of cigarettes smoked every week and the amount of alcohol drunk in a typical week. This result is also normal and may be related to social attitude: there exist people who do not care much about nutrition and who like spending their evenings in pubs. These people drink, smoke and eat much fat. Of course, you will also find a strong correlation between drinking alcohol and having a lung cancer. However, this correlation is spurious: drinking does not cause lung cancer. Only, people who drink a lot use to smoke a lot, and smoking causes cancer.

My conclusion is that correlation is a very helpful tool when you want to analyse a problem. However, correlation (and econometrics) is quite insufficient by itself. You still have to make a clever analysis of the problem, using common sense and a few well-thought tricks. Computers are not substitutes for human intelligence.

Compute the correlation matrix of koop\cormat.xls.

The correlation between variables X and Y is given by: $r = \frac{\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)}{\sqrt{\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2}\sqrt{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}}$. X and Y appear in a symmetric way in the formula: the correlation between X and Y is the same as the correlation between Y and X. This is another way to consider the ambiguous meaning of correlation: if it is high, does X cause Y or Y cause X?

An introduction to simple regression

In the example with hprice.xls, we drew a scatter diagram of the price of a house relatively to the size of the lot where it is located. We found a positive relationship between these two variables. Let us call X the size of the lot and Y the price of the house. This relationship can be written:

$$Y = \alpha + \beta X$$

This is the equation of the regression line of Y relatively to X. X is the explanatory variable. Y is the explained or dependent variable. α and β are the coefficients or the parameters of the regression line. You remember the line the computer drew in the middle of the scatter diagram.

Now, this relationship is only approximately true: the observations for each house are distributed around this line, but there are not on the line. Thus, the true equation is:

$$Y = \alpha + \beta X + \varepsilon$$

ε is called the error term. For some houses it is positive. For others it is negative. Sometimes it is large. Sometimes it is small. Thus, it represents the fact that the regression line is only an approximation of truth. The error term sums up all the omitted variables: the number of rooms, the quality of the district, the ability of the seller to get a good price, etc.

Now, the econometrician does not know the true values of parameters α and β. He will use statistics to infer estimates of these values based on the observation of the data he has for 546 houses. Of course, these estimates will be a little wrong: they will differ from the true values α and β. We will denote them: $\hat{\alpha}$ and $\hat{\beta}$. The true equation can be substituted by the estimated equation:

$$Y = \hat{\alpha} + \hat{\beta} X + u$$

There are two differences between the estimated and the true equations. First, the true values of parameters are substituted by their estimated values. Second, the error term ε is substituted by the residual u. The residual cumulates all the approximations in the true equation, which are included in the error term ε, plus the error resulting from the approximation of the true values of the parameters by their estimates.

How do econometricians estimate parameters? They want the estimated regression line well in the middle of the scatter diagram. Or, to be more precise, they want to minimise the values of the residuals of the equation. If, I call u_i, the residual for house i, a way to measure the importance of the residuals is to compute the sum of the squared residuals: $SSR = \sum_{i=1}^{N} u_i^2$.

The most popular way to compute estimates $\hat{\alpha}$ and $\hat{\beta}$ is to look for the values, which minimise the SSR. This method of estimation is called ordinary least square (OLS). It is easy to compute the formulae of $\hat{\alpha}$ and $\hat{\beta}$. The result is $\hat{\beta} = \dfrac{\sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)}{\sum_{i=1}^{N} \left(X_i - \vec{X}\right)^2}$,

$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$.

OLS estimation, and the minimisation of the SSR, is a very popular method in econometrics. However, it has a few disadvantages, which will be considered later on.

In the example of hprice.xls, I found: $\hat{\alpha} = 34136$, and: $\hat{\beta} = 6.60$. $\hat{\alpha}$ has no simple interpretation. The result for $\hat{\beta}$ means that if you increase the size of your lot by 1 square foot, the price of your house will increase by C$ 6.60. However, you must remember two things. First, the equation is an approximation. Secondly, the relationship between the size and price may be partly spurious and reflect, for instance, that large lots are often associated with a large number of rooms.

Let us consider house i. The estimated equation explains its price by: $Y_i = \hat{\alpha} + \hat{\beta} X_i + u_i$. Y_i is the true price of house i. If this price was on the estimated regression line it would be: $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$. $\hat{Y}_i$ represents the fitted or predicted value of the true price of house i. Actually, this price can be decomposed between the fit and the residual of the equation: $Y_i = \hat{Y}_i + u_i$. E-Views computes the fit and the residual for each house.

The total sum of the squares of the prices of houses is given by $TSS = \sum_{i=1}^{N} \left(Y_i - \overline{Y} \right)^2$. Actually, divided by $N-1$, it represents the variance of this price. The regression sum of squares is defined by $RSS = \sum_{i=1}^{N} \left(\hat{Y}_i - \overline{Y} \right)^2$. Divided by $N-1$ it represents the variance of the fit of the price. We can prove that: $TSS = RSS + SSR$. This means that the dispersion of prices is the sum of the dispersion of the fitted values of the equation and of the dispersion of the residuals. The precision of the equation or the quality of its fit are better if SSR is relatively small and RSS is relatively high. Thus, it is natural to measure the quality of the fit by the correlation coefficient $R^2 = RSS / TSS$. This coefficient is included between 0 and 1. In the bivariate case investigated in this chapter, the correlation coefficient is equal to the correlation coefficient between the two variables.

Until now I have considered a linear relationship between Y and X: $Y = \alpha + \beta X$. There are many cases when Y and X are strongly related, but in a non-linear way. For instance $Y = \alpha + \beta X^2$. In this case, we substitute X^2 to X and we proceed exactly as before. Thus, we just transform variables in an adequate way to get a linear regression model with the transformed variables.

The most popular transformation is the natural logarithm transformation: $\ln(Y) = \alpha + \beta \ln(X)$. To understand its meanings consider the relationship between households' consumption C and households' income Y. According to Keynes we should have: $C = a + cY$. Keynes called c the marginal propensity to consume and considered it to be positive but smaller than 1. However, we can get a better fit with the regression equation: $\ln(C) = b + d \ln(Y)$. d is the elasticity of consumption relatively to income: if income increases by 1%, consumption will increase by d%. Econometricians generally find d to be near to 1. This is a homogeneity constraint which must be satisfied if you do not want to find queer differences between countries of different sizes.

Most macroeconomic series have a geometric trend if they are not transformed. Their logarithm exhibits a linear trend (which is easier to analyse on a graph). The first difference of a series in logarithm is the growth rate of the original series. Thus, macroeconomists love to work with series in logarithms and with log-linear functions. However, a few series should not be transformed in this way, for instance, those which fluctuate around 0 (the logarithm of 0 is minus infinity). Examples of such series are interest rates, inflation rates and the rate of unemployment.

We can think a little bit more on non-linearity and common sense with the data of mukherjee\maputo. Each morning the authorities of the harbour of Maputo declare that they have the number DEMD of positions of dockers to fill for the day. Of course, this number varies with the numbers of ships arrivals to the harbour. The number of dockers hired for the day is RECD. Of course, this number cannot be larger than DEMD, and depends on the number of dockers who are ready to work this day. We have data on DEMD and RECD for 400 successive days. First, draw the histograms of the two series. Then, regress the first series on the second and a constant term. The result looks apparently econometrically good. Is not it economically silly? Draw the scatter diagram of both variables and think a little bit. Of course $RECD \leq DEMD$ and the difference between both variables increases with DEMD.

Statistical aspects of regressions

I will consider again the relationship between the price of a house and the size of the lot where it is located. The true relationship is: $Y = \alpha + \beta X + \varepsilon$
The values of parameters α and β are unknown by the econometrician. However, the econometrician has observations on a sample of 546 houses, and he can estimate by OLS equation:

$$Y = \hat{\alpha} + \hat{\beta} X + u$$

$\hat{\alpha}$ and $\hat{\beta}$ are called the estimates of α and β. They depend on the precise set of 546 observed houses. With another sample of 546 houses I would have got different estimates. As the sample of 546 houses is random, you could consider that the estimates are also random. However, most estimates computed on different samples of houses will be near the true values of the parameters. It is a good idea to consider an estimate as the realisation of a random variable, called an estimator. An estimator is function which links a random sample of houses to two values for parameters α and β. An estimate is the value taken by the estimator for a specific sample of houses. In the rest of this paragraph I will only consider parameter β. It is the most interesting parameter of the equation because it gives the sensitivity of the price of a house to the size of its lot. However, the main reason for doing this choice is that I do not want to complicate my notations and explanations by dealing with two parameters instead of one.

When you have computed an estimate, you know that it will probably differ from the true value of the parameter. Thus, it is clever to choose a significance level, for instance 95%. Then, statistics allows us to compute an interval centred on $\hat{\beta}$ such that the true value β has a probability of 95% to belong to this interval. This interval is called a confidence interval. The wider it is, the more imprecise the estimation is. For instance, newspapers publish such intervals with their political polls surveys: the conservative party will get between 32% and 36% votes in next election.

Koop makes Monte Carlo simulations pp. 58-61. He takes the true equation and he fixes the values of the parameters to $\alpha = 0$, $\beta = 1$. Thus, the true equation he will simulate is $Y = X + \varepsilon$. Then he chooses a sample of values for X, he makes random drawings for the error term ε (in E-Views you have a command which generates random numbers), he computes Y and he draws the scatter diagram of X and Y.
An econometrician would observe the scatter diagram, but would not know the true values of the parameters. Instead he would try to infer these values, that is to compute estimates, from the scatter diagram. Koop draws 4 scatter diagrams and we can notice on them:

1. More observations will increase the accuracy of the estimation.
2. Smaller errors (i.e. a smaller variance of ε) will increase the accuracy of the estimation.
3. A larger spread of values of X (that is a larger variance of the explanatory variable) will increase the accuracy of the estimation. This is normal: if all the lots had a size between 5000 square feet and 6000 square feet, estimating the effect of size on price would be difficult.

I advise you to look at the 4 scatter diagrams drawn by Koop.

Econometric theory shows that the confidence interval of β is $(\hat{\beta} - t_\alpha s_b, \hat{\beta} + t_\alpha s_b)$. s_b is the standard deviation of the estimator of β. We saw that this estimator is a random variable (it depends on randomly chosen houses), and its standard deviation measures the accuracy of the estimation. Econometric theory gives the formula: $s_b = \sqrt{\dfrac{SSR}{(N-2)\sum_{i=1}^{N}(X_i - \bar{X})^2}}$. To interpret it I will write it a bit differently: $s_b = \sqrt{\dfrac{SSR/N}{(N-2)[(\sum_{i=1}^{N}(X_i - \bar{X})^2)/N]}}$. s_b decreases, that is the accuracy of the estimation increases, when: 1) $N-2$, that is the number of observation increases ; 2) SSR/N , that is the variability of the error terms decreases ; 3) $[(\sum_{i=1}^{N}(X_i - \bar{X})^2)/N]$, that is the variability of the explanatory variable increases. Thus, this formula is consistent with the four scatter diagrams drawn by Koop.

α is the significance level, for instance 95%. If the error term is normally distributed, t_α is a value, which can be found in a Student statistical table. It depends on α, of course, but also on the number of observations N. However except for very low values of N, t_α does not change much with N, and its value can be found in a normal distribution statistical table. Nowadays, nobody looks at statistical tables: the computer looks at them instead. However, 20 years ago students had to learn how to use these tables, which was a bit cumbersome.

What happens if the error term is not normally distributed, that is if its distribution includes a fat tail or if there are outliers. Econometric theory proves that the previous results are still true if the number of observation is large. however, large has ambiguous meaning: does large mean 100 observations, 500 or 5000? The answer to this question will make a big difference for the practitioner.
Fat tails are bad[3], but outliers are very bad. This explains why I spent some time on the concept of robustness in chapter 1. I will come back to this concept a bit later.

In this paragraph, I have dealt until now with the estimations of parameters. Now, I will consider testing. Does the price of a house really depend on the size of the lot where it is located? If you find the question silly, does deforestation is really sensitive to population density? After all it could be only sensitive to the greed of foreign capitalists and the locals could cherish their forests. Or in a more general way is Y sensitive to movements of X, or does β differ or not from 0.

I will call the hypothesis: $\beta = 0$ the null hypothesis. If it is true, the explanatory variable has no effect on the explained variable. $\beta \neq 0$ is the alternative hypothesis. If it is true, changes in the explanatory variable affect the dependent variable.
I will test the null hypothesis against the alternative hypothesis. Statisticians are pessimistic people. Either, the evidence given by the observed data contradicts the null hypothesis, and the null hypothesis is rejected. Or, the evidence given by the observed data does not contradict the null hypothesis, and the null hypothesis is not rejected. This conclusion is different from "the null hypothesis is accepted". The null hypothesis could be completely wrong, but the observations could be insufficiently informative to discover that.

To process a test you need to compute an appropriate test statistic. For our problem, this statistics is called a t-statistics, or t-ratio, and is defined as $t = \frac{\hat{\beta}}{s_b}$. If t is small, I will not reject the null hypothesis, if it is large, I will reject the null hypothesis. Now, what do large and small mean?

Econometric theory proves that if the null hypothesis is true (and if the error term is normally distributed or if the number of observations is large), t is distributed as a Student distribution (or as a normal distribution if the number of observations is large). Student and normal distributions are implemented in econometric software. Thus, assume that the value of the statistics is 2.36. If the null hypothesis is true that is if: $\beta = 0$, the computer can tell us that the probability for having a statistic equal or larger than 2.36 is 1.1% (to be fairly honest I did not check this last number because I have no statistical table with me). 1.1% is called the P-value of the test. 1.1% is a low probability. If the null hypothesis was true, it would have been very unlikely to get a statistic as high as 2.36. Thus, I will reject the null hypothesis.

A rule of thumb is to reject the null hypothesis when the P-value is smaller than 5%, and not to reject it when it is larger than 5%. 5% is called the significance level of the test. For some problems you could prefer choosing another significance level, for

[3] Actually, if the tails of the distribution of the error term are a little too fat, its standard deviation, and even its mean are not defined, and usual econometrics becomes invalid.

instance 10% or 1%.

We can be a little more precise. We saw above that a larger number of observations will increase the accuracy of the estimation. Thus, if the null hypothesis is a little wrong, that is if the explanatory variable has a weak effect on the explained variable, the P-value of the test will be high if the size of the sample is small, but very low if this size is large. Thus, it will be wise to take a significance level, which is high for a small sample size (10%) and low for a large sample size (1%)[4].

A rule of thumb is to reject the null hypothesis when the P-value is smaller than 5%, and not to reject it when it is larger than 5%. 5% is called the significance level of the test. For some problems you could prefer choosing another significance level, for instance 10% or 1%.

If the number of observations is large enough, 1.96 is associated with a P-value of 5%. Then, you will reject the null hypothesis $\beta = 0$ if the t statistic is larger than 1.96. This value increases when the number of observation decreases. For instance, with 20 observations, 2.09 must be substituted for 1.96. The difference between these two numbers is tiny. However, you must remember that the Student test rests upon the assumption that the error term follows a normal law (or otherwise that the number of observations is large).

If you are a bit imaginative, you have already noticed, that you would have the same test if you computed the confidence interval associated with a significance level of 95% (which is 100%-5%), and checked if 0 belonged to this interval. If 0 belongs to this interval, you will not reject the null hypothesis. Otherwise you will reject it.

If you are a little more than a bit imaginative, you will have noticed that testing $\beta = 0$, is testing if the explanatory variable has an influence on the dependent variable, i.e. if the square of the correlation coefficient of the regression R^2 is equal or not to 0.

You could like testing not the null hypothesis: $\beta = 0$, but the null hypothesis: $\beta = c$, where c is a nonzero number which you selected. The test of this new null hypothesis proceeds as before, except that you must use the new statistics: $t = \frac{\hat{\beta} - c}{s_b}$, instead of: $t = \frac{\hat{\beta}}{s_b}$. This is a form of the Frisch-Waugh theorem. This theorem proves that if you estimate by OLS equation: $Y = \alpha + \beta X + \varepsilon$, and equation; $Y - cX = \alpha + (\beta - c)X + \varepsilon$, if you denote by: $\hat{\alpha}$ and $\hat{\beta}$ the estimates of the first equation, the estimates of the second equation will be: $\hat{\alpha}$ and $\hat{\beta} - c$. Thus, testing if the coefficient of the explanatory variable is equal to c in the first equation is equivalent to testing that the coefficient of the explanatory variable is equal to 0 in the second equation. Then, we can apply the test developed before to the second equation, and we get the above formula for the test statistics.

[4] To make this sentence clearer we would need to introduce the concepts of type II error and of power of a test.

Robustness of regressions

I will consider the true model: $Y = \alpha + \beta X + \varepsilon$
and its estimation by OLS: $Y = \hat{\alpha} + \hat{\beta} X + u$

For observation i, residual u_i is an estimate of the error term ε_i. A basic assumption for OLS is that the standard deviation of the error term is the same for all observations. Its estimate is $s = \sqrt{\frac{SSR}{N-2}}$. However, the standard deviation of residuals changes with observations. For observation i it is $se(u_i) = s\sqrt{1-h_i}$, with: $h_i = \frac{1}{N} + \frac{(X_i - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}$.
h_i is called the hat statistic. The more distant X_i is from its mean, the higher h_i is.

In a regression, an *outlier* is an observation with a large residual, compared to the residual of most other observations. An observation with high leverage is an observation with an explanatory variable, which takes a value very different from its mean, that is an observation with a high hat statistic. An *influential observation* is such that if you remove it, the estimate of β will change a lot. These concepts are related to one another, but they are different. Loosely speaking, an observation with high leverage has the potential of being influential. If this observation is an outlier this potential is realised and the observation is influential. So, to be influential an observation must be simultaneously with high leverage and an outlier. When an observation is influential the results of the estimation of the regression depend at a strong extent on this observation. So, they become fragile: you would like the results of your estimation to be almost insensitive to the the deletion of any arbitrary set of a small number of observations

To identify an *outlier* we must compute the *studentized residuals*. To do that, we will divide each residual by its standard deviation. However, as the standard deviation of the error term is sensitive to outliers, in the formula giving the standard deviation of residual i, we will use an estimator of the standard deviation of the error term, which does not use this residual, let be s_i instead of s. Then, we get the expression: $t_i = \frac{u_i}{s_i\sqrt{1-h_i}}$
There exists a table for this statistics, so we can test if u_i significantly differs from 0. The critical values of this table are higher than for a Student table. It is easy to understand why. Let us assume that I am interested by the different problem: is observation i an aberration in the sample? Then, I will estimate the equation after having added a new explanatory variable, which is a dummy variable with a value equal to 0 except for observation i where it is equal to 1. Then, I will compute the Student-t statistics of the coefficient of the dummy variable and I can notice that it is exactly equal to the studentized residual for this observation: t_i. I can compare this statistic to the critical value of a Student table, which is of the order of 1.96. The null hypothesis is that observation i is not an outlier. However, in this paragraph the

problem is different. I wonder if among the N observations there exist one or several outliers. My null hypothesis is that there exist no outlier. To test this hypothesis, I will not select an observation i and check if t_i is above some critical value. I must check if $\max(t_i)$ is above some critical value, or if all t_i are under this critical value. To compare N t_i to a common value is a much stricter requirement than to compare only one of them to this value. Then, a statistical analysis of the set of all the studentized residuals must use a critical value higher than 1.96.

Actually, it is not very important to bother with an exotic table. I can simply compute the statistics for all the residuals and investigate if it does not sometimes take excessive values. I also can, like in last chapter, compare the median, the mean and the IQR, and look for outliers in the distribution of the t_i

The *leverage* of an observation is measured by its hat statistics. If $X_i = \overline{X}$, i.e. if the observation is located in the middle of the sample, then it has no leverage and $h_i = 1/N$. If the observation is far from its mean, h_i increases and tends to 1. It would be a good idea to compute the maximum of the hat statistics. If it is less than 0.2 we have no reason to worry. If it is larger than 0.5, the leverage is too high.

An *influential observation* is identified by its DFBETA statistics. This statistics is equal to $DFBETA_i = \dfrac{\hat{\beta} - \hat{\beta}_i}{se(\hat{\beta})_i}$, where $\hat{\beta}_i$ and $se(\hat{\beta})_i$ are the estimates of β and the standard deviation of this estimate after having removed observation i. Usually, this statistics is compared to $2/\sqrt{N}$, or to $3/\sqrt{N}$. Observations with DFBETA higher than these limits are worrisome. I also can use the robust methods of last chapter to look for outliers in the series of the DFBETA.

It is possible that no observation is influential, but that a grape of 2 or 3 successive observations is influential. We can check that by applying the same formula to a grape of such observations.

Tutorial: the capital asset pricing model (CAPM) (Berndt, chapter 2)

I will consider a stock denoted j. Its price at current time 0 is p_{j0}. Its expected price at future time 1 will be p_{j1}. As p_{j1} is an expectation, I can consider it a random variable. For instance you will expect that the price of the stock will belong to interval ($29, $29.10) with probability 0.07, to interval ($29.10, $29.20) with probability 0.15, etc. d_j represents the dividends which will be paid at time 1. Then I will define the expected return of the stock over the period as $r_j = \dfrac{p_{j1} - p_{j0} + d_j}{p_{j0}}$. r_j is a random variable.

If expectations are given, having available a model which can explain the values taken by r_j, is equivalent to having a model which can explain the current value of the price of a stock p_{j0}. Finance theory prefers working with rates of return than with prices. I will follow this practice.

I will denote by r_f the risk free interest rate (for instance the rate of return of 30-days Treasury bills) and by r_m the market rate of return. r_m represents the rate of return of all the risky assets taken together which are available in the economy. It is usually computed as a financial market index. The CAPM will be a better representation of reality if this index is very wide. This index is called the market rate of return.

The CAPM proves the following equation:

$$Er_j - r_f = \beta_j (Er_m - r_f)$$

At date 0, r_j and r_m are random variables. Er_j and Er_m represent their expected or mean values (that is the expected or mean value of the price of the stock and the financial index in one period time). $Er_j - r_f$ can be interpreted as a risk premium: the expected rate of return of your stock is, for example, 6 percentage points above the risk free interest rate. $Er_m - r_f$ can be interpreted as the risk premium on the market portfolio. If we take this portfolio as a reference, associated to a risk equal to 1, $Er_m - r_f$ can be interpreted as the price of risk. β_j is the beta of stock j, or its risk. Finance theory proves that it is equal to:

$\beta_j = \sigma_{jm} / \sigma_m^2$ where σ_{jm} is the covariance between r_j and r_m, and σ_m^2 is the variance of r_m.

Now, let us consider the regression model:

$$r_j - r_f = \alpha_j + \beta_j (r_m - r_f) + \varepsilon_j,$$

If we take the expectation of this equation we get the CAPM equation (notice that in a regression $E\varepsilon_j = 0$, which means that as many errors are by excess as by default). According to the CAPM: $\alpha_j = 0$. Moreover, we can see that:

$$Var(r_j) = \beta_j^2 Var(r_m) + Var(\varepsilon_j)$$

This formula uses a property of the regression model, which is that the error term and the explanatory variable are not correlated. The CAPM equation shows that the price of stock j (proportional to the inverse of r_j) only prices the beta of this stock. Thus, in the above formula, we say that $Var(r_j)$ is the total risk of the stock. It is split between a systematic risk, which cannot be removed by clever diversification and which has to be priced $\beta_j^2 Var(r_m)$, and a unsystematic risk, which disappears in a well diversified portfolio and which must have a zero price $Var(\varepsilon_j)$.

Now, we can estimate the regression equation

$$r_j - r_f = \hat{\alpha}_j + \hat{\beta}_j (r_m - r_f) + u_j$$

This estimation can be made, for instance, by using monthly data over a 5 years period. The formulae given above show that $\hat{\beta}_j = \hat{\sigma}_{jm} / \hat{\sigma}_m^2$. The hat on the left-hand side represents estimates of the covariance and the variance.

It is easy to test if $\alpha_j = 0$ or if $\hat{\beta}_j$ stays constant over time (this by using a Chow test which was not presented in this lecture, but which is implemented in E-Views).

The square of the correlation coefficient R^2 can be interpreted as an estimate of the proportion of the total risk of the stock, which is systematic. $1 - R^2$ is the proportion of the total risk, which is unsystematic.

The arbitrage-pricing model (APM) of Stephen Ross was built on foundations very different from those of the CAPM. However, its final equation is the same as the CAPM equation, except that, besides of the measure of the price of risk, other explanatory variables appear in the final equation. Thus it is easy to test the CAPM against the APM, by testing if the coefficients of these new variables differ significantly or not from 0. If there is only one more variable, we can use the Student test introduced before. If there are more than 1 supplementary variable, we must test for the simultaneous nullity of 2 or more coefficients. We use a generalisation of the Student test, which is called a Fisher test and which is implemented in E-Views (under the more general name of Wald test).

The exercises are given pp. 41 to 54. Exercises 1, 2, 3 and 4 are directly related to this chapter. Exercises 7 and 9 are related to next chapter. Exercises 6, 8 and 10 use notions which were not introduced in this chapter. Exercise 5 is outside our subject.

CHAPTER 3. MULTIVARIATE ANALYSIS

Multiple regressions

Multiple regressions extend simple regressions to the case when there are many explanatory variables. Most of the intuition and statistical techniques of multiple regressions are very similar to those of simple regressions.

X_1: The lot size of the property (in square feet);
X_2: The number of bedrooms;
X_3: The number of bathrooms;
X_4: The number of storeys (excluding the basement).

The true model is: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$

The econometrician does not know the values of the parameters in the true model. He will try to estimate it by using the 546 observations, and he will get the estimated model:

$$Y = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + u$$

He will compute the estimates of the parameters by minimising the sum of squared residuals:

$$SSR = \sum_{i=1}^{N} u_i^2 = \sum_{i=1}^{N} \left(Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{13} - \hat{\beta}_4 X_{4i} \right)^2$$

The formulae giving the estimated values of the parameters, $\hat{\alpha}$, $\hat{\beta}_1$, etc. are a bit cumbersome. However, the computer can easily compute them. The correlation coefficient of the regression, R^2, is still a measure of the fit. It must be interpreted as a measure of the explanatory power of the explanatory variables together. Confidence intervals for each parameter can be computed as in chapter 2. Testing if any parameter is equal to 0 or to a value c can be done as in last chapter.

In chapter 2 only the first explanatory variable, the size of the lot, was taken into account. Its estimated coefficient $\hat{\beta}$ was equal to 6.6. In this chapter, three more explanatory variables have been added. The estimated coefficient of the size of the lot $\hat{\beta}_1$ is now equal to 5.4, which is a lower value.

This last estimation means that if you own a house, if you keep unchanged the number of bedrooms, the number of bathrooms and the number of storeys, and if you buy one square foot of garden to your neighbour, the price of your house will increase by Ca $5.4.

In chapter 2, I concluded that this price should increase by Ca $6.6 and I was wrong. The problem is that big lots, many bedrooms, many bathrooms and many storeys go together. Or, in technical terms, the 4 explanatory variables are positively correlated to one another. Thus, in chapter 2 where 3 important variables were omitted from the regression, the size of the lot was also an index of the number of bedrooms, the number of bathrooms and the number of storeys of the houses of the sample. The coefficient 6.6 represented for a part equal to 5.4, the fact that a big garden increases the price of a house. However, for a part equal to: 6.6-5.4=1.2, big gardens are indices for large numbers of bedrooms, bathrooms and storeys, variables which were omitted from the equation. In econometric language 6.6 is a biased estimation of the effect of the size of the lot on the price of a house. 5.4 is an unbiased estimation. Of course, both estimations differ from the unknown true value. However, there is something like a systematic error in the estimated value 6.6. This systematic error is missing in the estimate 5.4, which is much more reliable.

Thus, if you consider buying a piece of land from your neighbour, you should consider that the price of your house will increase by $5.4 per square foot bought, and not by $6.6. Bad econometrics can have serious practical consequences.

Some important variables are still missing from the equation of this chapter. If they are positively correlated with the size of the lot, their absence increases the coefficient of this variable (the "true value" of it should be less than 5.4). If they are negatively correlated with the size of the lot (for instance the proximity of the centre of the city is negatively correlated with the size of the lot), their absence should decrease the value of the estimated coefficient. We cannot include all the required explanatory variables for two reasons. First, because many of them are difficult to measure, or are simply omitted in surveys. Second, because your estimation becomes less precise (that is the standard deviations of the estimated parameters becomes higher) when the number of explanatory variables increases.

The previous problem is the most serious problem econometricians have to deal with. There is no magical recipe. A good econometrician must, not only know how to use the computer, but he still must be a good analyst (of the housing market for instance), able to mix his technical ability with common sense and the experience of experts (for instance sales agents).

Another problem arises if some or all the explanatory variables are highly correlated with one another. Let us imagine that we work on a survey of households. It is reasonable to consider that consumption is related to income and wealth. However, income and wealth are strongly correlated to each other. Thus, when you run your regression, you will be unable to identify the precise effect of income, compared to wealth, on consumption. Technically, the manifestation of this impossibility will be very large confidence intervals for the estimated coefficients of both variables and very high P-values for tests of nullity of these coefficients. These two problems will disappear if you omit one of the two explanatory variables, for instance wealth. The confidence interval of the coefficient of income will be small, and the P-value for the test of the nullity of this coefficient will be very small. However, the estimation of this coefficient will be biased, because it will cumulate the direct effect of income on consumption, and the indirect effect of wealth for which income acts as an index.

Tutorial: Costs, learning curves and scales economies (Berndt, chapter 3)

Cost function

The econometrics of production very much uses the concept of the cost function. Consider a firm, or a plant, with a production function: $y = f(x_1,...,x_n;A)$. y

represents the output of the firm, $x_1,..,x_n$ are the inputs and A represents the state of technical knowledge available at the firm.

Let us double all the quantities of inputs used by the firm: $x_i \to 2x_i$, with $i = 1,..n$. If: $y \to 1.8y$, I will say that the firm has *decreasing returns to scale*. If: $y \to 2y$ I will say that the firm has *constant returns to scales*. If: $y \to 2.2y$ I will say that the firm has *increasing returns to scales*. 1.8-2=-0.2 or 2.2-2=0.2 represent the economies of scale.

I will denote the price of input i as p_i and I will consider it to be fixed and exogenous (the assumption of perfect competition on the inputs markets). The total production cost of the firm is $C = \sum_{i=1}^{n} p_i x_i$. For any given level of output y (determined outside the models considered in this tutorial) the firm determines the quantities of the various inputs it will use to minimise its total production under the constraint of its production function. Thus, I define the Lagrangian of this optimisation problem: $\sum_{i=1}^{n} p_i x_i + \lambda[y - f(x_1,..,x_n;A)]$, where λ is the Lagrange multiplier. Then, I compute the partial derivatives of this expression, relatively to all the x_i, and I put them equal to 0. I get:

$\frac{p_i}{p_n} = \frac{f_i(x_1,..,x_n;A)}{f_n(x_1,..,x_n;A)}$, for: $i = 1,..n-1$.

I can use these n-1 first order conditions with the production function to compute the optimal quantities of each input in function of the quantity of output and of the n prices of inputs. Then, I can substitute these expressions in the definition of the total cost and I get: $C = g(p_1,.., p_n, y; A)$

This expression is called the *cost function* of the firm. *Unit cost or average cost* is defined as $c = C / y$.

Frequently, c is a decreasing function of y when y is small (increasing returns to scale) and an increasing function of y when y is large (decreasing returns to scale).

Learning curve

The average production cost of an output decreases with the cumulated past production of this output. This empirical result is caused by the fact that the more a firm has produced, the better it knows how to produce in an efficient way. This result can be formalised by equation:

$c_t = c_1 n_t^{\alpha} \exp(\varepsilon_t)$, $\alpha < 0$

c_t represents the average production cost in period t, n_t represents past cumulated production (not including period t), ε_t is an error term.

This equation can be given the following interpretation. Each time n_t doubles c_t will decline to proportion d of its previous level, with: $d = 2^{\alpha}$. d is called the learning curve slope. Berndt indicates page 80 that in empirical research many products have a learning curve slope of 0.80 to 0.84.

Cobb-Douglas production function

I will try to be more specific. I will start by the case when there is no learning. A Cobb-Douglas production function can be written:

$$y = Ax_1^{\alpha_1} x_2^{\alpha_2} x_3^{\alpha_3}$$

The returns to scale are: $r = \alpha_1 + \alpha_2 + \alpha_3$

If we apply the procedure of computation presented above, we can prove that the cost function is:

$$\log(C) = \log(k) + (1/r)\log(y) + (\alpha_1/r)\log(p_1) + (\alpha_2/r)\log(p_2) + (\alpha_3/r)\log(p_3),$$

with: $k = r[A\alpha_1^{\alpha_1}\alpha_2^{\alpha_2}\alpha_3^{\alpha_3}]^{-1/r}$

When we estimate this cost function we must not forget that its parameters are related by constraint: $r = \alpha_1 + \alpha_2 + \alpha_3$. Thus, to make the estimation easier I will use this equation to eliminate parameter α_3 from the cost function. I will get:

$$\log(C) - \log(p_3) = \log(k) + (1/r)\log(y) + (\alpha_1/r)[\log(p_1) - \log(p_3)] + (\alpha_2/r)[\log(p_2) - \log(p_3)]$$

To estimate this equation I will define the new variables:

$\log(C^*) = \log(C) - \log(p_3)$, $\log(p_1^*) = \log(p_1) - \log(p_3)$, $\log(p_2^*) = \log(p_2) - \log(p_3)$,

and the new parameters:

$\beta_0 = \log(k)$, $\beta_y = 1/r$, $\beta_1 = \alpha_1/r$, $\beta_2 = \alpha_2/r$

Thus, the cost function can be re-written as:

$$\log(C^*) = \beta_0 + \beta_y \log(y) + \beta_1 \log(p_1^*) + \beta_2 \log(p_2^*)$$

This equation can easily be estimated by OLS. Then we can compute the estimates of the old parameters from the estimates of the new parameters by using formulae:

$\alpha_1 = \beta_1/\beta_y$, $\alpha_2 = \beta_2/\beta_y$, $\alpha_3 = (1 - \beta_1 - \beta_2)/\beta_y$

To introduce learning in the cost function I will simply assume that learning increases the state of knowledge available to the firm:

$$A_t = n_t^{-\alpha}$$

I will define the new parameter: $k' = r[\alpha_1^{\alpha_1}\alpha_2^{\alpha_2}\alpha_3^{\alpha_3}]^{-1/r}$. The cost function becomes:

$$\log(C_t) = \log(k')$$
$$+ (\alpha/r)\log(n_t) + (1/r)\log(y_t) + (\alpha_1/r)\log(p_{1t}) + (\alpha_2/r)\log(p_{2t}) + (\alpha_3/r)\log(p_{3t})$$

A practical difficulty with this equation is that the cost of the firm is measured in current dollars, and that the econometrician must know the prices of inputs. Very often these prices are unknown. Berndt suggests a solution to this problem (a not very good solution in my opinion). Let us denote the GNP price deflator by $GNPD_t$. This price is an average of all the prices prevailing in the economy. If we are lucky we will have approximately the relationship:

$$\log(GNPD_t) = (\alpha_1 / r)\log(p_{1t}) + (\alpha_2 / r)\log(p_{2t}) + (\alpha_3 / r)\log(p_{3t})$$

If we subtract this equation from the previous cost function we get:

$$\log(C_t^{'}) = \log(C_t) - \log(GNPD_t) = \log(k') + (\alpha / r)\log(n_t) + (1/r)\log(y_t)$$

$C_t^{'} = C_t / GNPD_t$ represents production costs in constant dollars. If I denote unit costs in constant dollars as: $c_t = C_t^{'} / y_t$, the previous equation can be re-written:

$$\log(c_t) = \log(k') + (\alpha / r)\log(n_t) + [(1-r)/r]\log(y_t)$$

This equation can easily be estimated by OLS. Then, we can immediately deduce the learning curve slope and the returns to scale. We can also test if these returns are constant or not (null hypothesis r=1).

In pages 76-78 of his book Berndt gives a very clear discussion of the omitted variable bias. In page 83 you can read: "The econometric literature on estimated returns to scale in the electrical industry in the United States appears to suggest that substantial economies of scale have been available, that such economies may have been largely exploited by the early 1970s, and that today the bulk of electricity generation comes from firms generating electricity at the bottom of their average cost curves".

The exercises are given pages 83 to 95. Exercises 1, 2, 4, 5, 6 are excellent. Exercise 3 is very interesting but academic. Exercise 7, 8, 9 and 10 are more advanced than this course.

Partial correlations

Let us start with the data in Mukhrjee \Khrishnaj.wk1. We want to investigate if the demand of industrialised consumption goods ln(M) decreases in India when the price of cereals increases, because of an income effect, which dominates substitution effect. The explanatory variables are the logarithm of real income ln(R) and the relative price of cereal ln(p). We can regress the explained variable relatively to the two explanatory variables and a constant term:

$$\ln(M) = \hat{\alpha} + \hat{\beta}_1 \ln(R) + \hat{\beta}_2 \ln(p) + u$$

The *Frisch-Waugh theorem* establishes the following result.

First, let us regress ln(M) on a constant term and on ln(R). We get:

$$\ln(M) = \hat{\gamma}_0 + \hat{\gamma}_1 \ln(R) + \nu$$

Second, let us regress ln(p) on a constant term and on ln(R). We get:

$$\ln(p) = \hat{\delta}_0 + \overleftarrow{\delta}_1 \ln(R) + \eta$$

ν and η represent respectively the parts of ln(M) and ln(p), which are left unexplained after having taking care of the information brought on these variables by $\ln(R)$.

Now, if we run a regression of ν on η (without a constant term), we will get:

$$v = \hat{\beta}_2 \eta + u$$

Thus, there are two ways to compute the estimate $\hat{\beta}_2$ and the residual u:
Directly, by running the first multivariate regression.
Indirectly, by running the three next univariate regressions.

An advantage of the second method is that we can apply to the three bivariate equations it includes the methods of analysis of chapter 2. These methods can be helpful, for instance to look for outliers, non-linearities, etc. Scatter plots can be very instructive.

The correlation coefficient of the last equation is called the partial correlation coefficient between the logarithm of manufactured consumption and the logarithm of price. It is a good measure of the supplementary information brought by the logarithm of price to the logarithm of manufactured consumption, after having taken care of the information brought by the logarithm of income. Thus, if you consider adding an explanatory variable to an equation, partial correlation is a good measure of the new information brought by this variable. The significance of the new explanatory variable can be tested by a Student test on its coefficient in the expanded equation.

Fragility analysis

We will use the data in Mukherjee\fertility.wk1 . We want to explain the variability of the fertility rate of women over a large number of countries by 4 explanatory variables. No fully convincing theory exists which could help us to choose a specification of the equation. Thus, we run 15 regressions: all the possible combinations of explanatory variables, 4 regressions, with only one explanatory variable, 6 regressions with 2 explanatory variables, 4 regressions with 3 explanatory variables, 1 regression with the 4 explanatory variables.

In all the regressions where it appears, the coefficient of the variable family planning takes about the same value, which significantly differs from 0. This robustness of the result concerning this variable suggests that we must put it in the equation. Thus, we discard the 7 equations where it does not appear.

We use the same method on another variable for the 8 equations left. Progressively, we discard equations.

Finally, we are left with two equations between which we cannot choose on econometric grounds. These equations differ by the fact that infant mortality appears in the first and not in the second, and the alphabetisation rate of women appears in the second and not in the first. However, common sense suggests that, when women become more educated, their fertility rate decreases (for instance because they can get a job, or because they become more independent from their husband). Moreover, their children have a higher probability to survive (because educated mothers are abler to follow basic rules of hygiene). Thus, we could choose to keep the second equation. Another argument would be that when a child dies, his parents would want to have another child to take the place he left. Thus, we could choose to keep the first equation. Thus, we face an economic problem, which is well discussed in the book by Mukherjee and *alii.*

Regressions with dummy variables

Some of the explanatory variables can be qualitative variables, of the kind we met in chapter 1. This does not change anything for the theory of multivariate regressions. However, the economic interpretation of the results can be a bit tricky. I will explain how on the example given in Koop\hprice.xls. The explained variable is still the price of a house Y. I have data on 546 houses. I will use two former explained variables: X_1 the size of the lot, and X_2 the number of bedrooms. I will not use the 2 other explanatory variables introduced before, just not to make my presentation too cumbersome. I will add two qualitative variables: having or not having a driveway having or not having a recreation road.

I will denote the first variable as D_1. This variable can take values 1 (if the house has a driveway) or 0 (if it does not). I will denote the second variable as D_2. This variable can also take values 1 (if the house has a recreation room) or 0 (if it does not). Such variables, which can take on only two values, 0 or 1, are called dummy variables.
Now, I will run the regression of Y on the four explanatory variables and a constant. I will get:

$$\hat{Y} = -2736 + 12598 D_1 + 10969 D_2 + 5.197 X_1 + 10.562 X_2$$

That means that if the house has no driveway ($D_1 = 0$) and no recreation room ($D_2 = 0$), its value will be $\hat{Y} = -2736 + 5.197 X_1 + 10.562 X_2$.
If the house has a driveway, its value will be, holding all other variables constant (or ceteris paribus), \$12598 more. If the house has a recreation room, its value will be, ceteris paribus, \$10969 more. If the house has a driveway and a recreation room, its value will be, ceteris paribus, 12598+10969 = \$23567 more.

What is peculiar in the above result is that the marginal effect of the size of the lot, or the number of bedrooms, is the same for the four kinds of houses. Increasing the size of the garden by 1 square foot will increase the price of the house by \$5.197 whether the house has or not a driveway or a recreation room. This assumption may be justified or not. To understand this point I will give another example.

I am still interested in explaining the price of a house Y. But now, I will use only two explanatory variables: D = 1 if the house has air conditioning, 0 if not; X = lot size. However, I will consider that increasing the size of the lot can affect the price of the house, which will differ if the house has or does not have air conditioning. For that I will introduce a third variable $Z = DX$. I run the regression and I get

$$\hat{Y} = 35684 + 7613 D + 5.02 X + 2.25 DX$$

If the size of the lot is increased by 1 squared foot, the price of the house will increase by \$5.02 if the house has no air conditioning, and by 5.02+2.25=\$7.27 if the house has air conditioning. We can check, with a Student test that the P-value of the coefficient of DX is 0.02, so that this coefficient is significantly different from 0. Thus, the difference in the marginal effect of the size of the lot on price for the two kinds of houses is significant.

How more will be priced a house with air conditioning? In the case of interacting dummy and non-dummy variables, we do not get as simple results as before. A house with air conditioning will be priced $ 7613+5.02X. The premium of having air conditioning is no more constant; it increases with the size of the lot.

Thus, you can see that dummy *explanatory variables* do not change the econometrics. In this paragraph we have just introduced a few ideas on the best way to give good economic interpretation for the result. However, dealing with an *explained or dependent dummy variable* is a much more difficult problem. In this case the econometrics of OLS does not work any more and we must use another econometrics, called probit and logit models. Actually, the basis of this new econometrics is not very difficult. However, it is different and will not be investigated in these lectures. E-Views has these two new methods implemented.

Tutorial: Analysing determinants of wages and measuring wage discrimination (Berndt, chapter 5)

Theory.

Education has a cost. This cost takes the form of foregone earnings, as well as direct expenses such as tuition. However, education will increase productivity in the workplace, and thus will be gratified by higher earnings over work life. Thus, education includes all the ingredients of an investment.
The length of time over which you will benefit from the return of education will be higher if you study when you are young than if you study when you are old. Moreover, as earnings increase with experience, foregone earnings will be higher for an old student than for a young student. These stylised facts can explain why most people attend university when they are young.

Learning is a tough task. However, if you are clever and smart, this task will be easier for you[5]. Thus, there should be a high positive correlation between education and abilities. When you find in your regressions that a higher number of years of schooling goes with higher wages, you will face the difficulty of discovering if firms pay for education or for abilities.

On-the-job training is also related to wages. Here things become a bit subtle. General training increases the productivity of the worker at any task (inside or outside the

[5] The theory of *screening*, which was presented before, and which is part of the microeconomic theory under asymmetric information, uses the assumption that learning is *much easier* for smart people than for stupid people. Thus, if people with many years of schooling are paid *a little more* than people without any schooling, smart people will be induced to go to university, but not stupid people. Thus, when a firm will hire somebody with a university degree, it will know that this person is clever. Of course, if the differential of wages between educated and non-educated people was too large, everybody, silly people included, would go to university, and having a degree would not bring any information to firms. In this approach, education is only a *signal* allowing firms to identify among people who apply for a job those who are clever and those who are stupid. Education itself is useless and does not increase productivity. If this theory were fully true you could look for the best field of studies signalling a given ability. For instance, his employer could give somebody able to learn Greek philology a very boring task like classifying all the administrative files of the firm since its birth.

firm). Thus, a firm will be hesitant to provide this kind of training (for example learning a foreign language) because it knows that the worker can threaten to leave the firm afterward and ask for higher wages. Thus, firms will be ready to give this training if the worker pays at least for part of it, for instance by accepting lower wages over the period of training.

Specific training increases the productivity of the worker at specific tasks inside the firm, but cannot be used in other firms (for example learning how the computerised information system of the firm works). Thus, the firm can be ready to fully pay for this training, knowing that it will not have any reason to increase the wages of its trained worker. But, is it truly so? The firm and the worker with its new specific knowledge are in the situation of bilateral monopolists, which have to share a rent. The specific training of the worker has increased his productivity. How to share this increase between more profit for the firm and higher wages for the worker? If the worker separates from the firm the rent will be lost for both of them. Game theory has investigated this question for a long time.

Like education, training is an investment. Thus, it should decrease when the worker becomes older, so closer to the age of retirement.

Econometrics.
Measuring earnings and wages is difficult. When you survey workers you do not get the same answers as when you survey firms. Non wage benefits are badly known, and they do not benefit to all workers in equal amounts. The number of hours worked are also badly known (and the declarations by firms and workers differ).

The distribution of wages and earnings across a sample of randomly selected workers is log normal (or if you prefer the logarithm of earnings and wages follow a normal distribution). Thus, for reason discussed above, we will take the logarithm of earnings and wages as the explained variable of our regressions.

The usual statistical earnings function is:

$$\log(W_i) = f(s_i, X_i, z_i) + \varepsilon_i\,,\ i = 1,..,N$$

W_i represents earnings or wages for the *ith* individual, s_i is a measure of schooling or educational attainment, X_i indexes the human capital stock of experience, z_i are other factors affecting earnings such as the race, gender, or/and geographical region, ε_i is a random disturbance term reflecting unobserved ability characteristics and the inherent randomness of earnings statistics. It is usually assumed that ε_i is normally distributed with mean zero and constant variance.

To run a regression we must be more specific with the earnings function. I will assume that it is:

$$\log(W_i) = \log(W_0) + \beta_1 s_i + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 s_i X_i + \varepsilon_i$$

The fourth term of the right-hand side introduces the fact that returns on experience is increasing when people are at the beginning of their professional career, and decreasing when they have become old. Thus, we expect: $\beta_2 > 0$ and $\beta_3 < 0$, which implies that experience has the highest return when $X = -\beta_2 / (2\beta_3)$. The fifth term means that return on experience is higher for educated people than for people without education. Thus, we expect $\beta_4 > 0$.

We can add dummy variables to this equation. For instance, if we think that women are less paid than men, *ceteris paribus* we can add the gender dummy variable DG_i. It is equal to 1 if individual i is female and to 0 if it is male. If the coefficient of this new variable is denoted as β_5, we will expect it to be negative. The average earnings of a woman will be equal to the proportion $\exp(\beta_5)$ from the earnings of an average man, or approximately will be lower than these earnings by proportion β_5. We can also investigate if the return to education is the same for a woman as for a man. To do that we will add to the right-hand side of the equation the new variable $DG_i * X_i$.

We saw before that education can also be an index of abilities. In this case, the estimate of the coefficient of education will be upward biased. Griliches and Mason tried to deal with this problem by adding as explanatory variable the result of intelligence tests given by the army to its new conscripts. The authors found that the correlation between schooling and the results of these tests was very low. They concluded that omitting any measure of abilities in the equation does not induce serious bias in the estimate of the return to education.

Taubman worked with a sample of 1000 twins. Twins were assumed to have benefited of the same social background and to have the same abilities. Taubman found that when differences in earnings between pairs of twins were related to differences in schooling, the estimated rate of return to schooling was only about 3%, much less than the typical findings of 8% found elsewhere in the literature.

However, the majority of the economists believe that differences in abilities do not account for a sizeable proportion of earnings differentials among individuals who have different amounts of schooling. Thus they side with Griliches and Masson and with the human capital theory, against Taubman and the screening theory.

Pages 167-179 are a good survey of empirical results on the earnings function. Approximately, one year more schooling increases earnings by 8%. There are results showing that this rate can decrease with the number of years of schooling. Change in this rate of return over time is an important problem for labour economics. For instance, in the eighties and the beginning of the nineties this rate increased in the US, but decreased in France.

Discrimination.

Women (or blacks, or minorities) are often discriminated against on the labour market. The earnings statistical function is a good way to investigate the problem of discrimination and to introduce the concept of *statistical discrimination*. I presented above the following earnings function:

$$\log(W_i) = \log(W_0) + \beta_1 s_i + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 s_i X_i + \varepsilon_i$$

If I define β as the line vector of coefficients (the first element is $\log(W_0)$), and Z_i as the column vector of variables (the first element is a column of 1 and its coefficient is the constant term of the regression), this equation can be re-written in matrix form:
$\log(W_i) = \beta Z_i + \varepsilon_i$.

The estimate of this equation will be:
$\log(W_i) = \hat{\beta} Z_i + u_i$

A result of the theory of regression is that the mean of residuals over the total number of observations is zero. I will denote as $\overline{\log(W)}$ the mean of the explained variable[6], and as $\overline{Z}$ the mean of the vector of the explanatory variables. Then, I have:
$\overline{\log(W)} = \hat{\beta}\overline{Z}$.

Now, I will consider that my sample includes male and female workers. I will divide this sample into two sub samples. The first only includes female workers. I will run an OLS estimation of the earnings function on this subsumable, then I will compute the average of this relationship over the subsumable, and I will get:
$\overline{\log(W^F)} = \hat{\beta}^F \overline{Z}^F$

The right-member of this equation represents the logarithm of the average wages of female workers. I do the same estimation on the second subsumable, which only includes male workers and I get:
$\overline{\log(W^M)} = \hat{\beta}^M \overline{Z}^M$

In general, the average wages of the female worker are smaller than the average wages of male workers. The statistical results I have just obtained will bring some light on the reasons for this difference. I subtract the two last equations and I get:

$$\overline{\log(W^F)} - \overline{\log(W^M)} = \hat{\beta}^M (\overline{Z}^F - \overline{Z}^M) + (\hat{\beta}^F - \hat{\beta}^M)\overline{Z}^F$$

Blinder and Oaxaca commented on this equation. They stated that the mean difference in log earnings between male and female workers can be decomposed into the effects of differences in their average endowments (the first term on the right-hand side of the equation), and the effects of discrimination, as revealed by differences in estimated coefficients (the second term). Note that average endowment differences are weighted by male workers' estimated coefficients while differences in the estimated coefficients are weighted by average characteristics of female workers.

This decomposition means that, if women as a group are less paid than men, this can result from the fact that women spent fewer years at school or at university than men (as an average), or this can be because a university degree is less paid when it is held by a woman than by a man. Blinder and Oaxaca consider only the second reason to be discrimination.

[6] The exponential of this mean is equal to the geometric average of all the wages of the sample of observations.

This decomposition is extremely useful. However, the concepts we introduced and the definition of discrimination are purely statistical. For example, we would like to measure discrimination on the labour market, or better discrimination in a given firm. If in a firm, women are less paid than men (at an average level), this can result from less seniority for women related to their strong turnover. This can be unrelated to the firm policy. However, it can also result from an unfriendly policy of the firm with its female workers (no flexibility for maternity leaves or for working hours). Inversely, if a firm pays less a female engineer than a male engineer, this can be due not to a sexist attitude of the manager of the firm, but to a sexist attitude of the male workers whom engineers have to supervise. Not to go into trouble the manager of the firm may be reticent to hire a female engineer, except if she does not cost much. Thus, here again, we can see that econometrics is not a substitute for economic analysis or even common sense.

We can make a symmetric decomposition:

$$\overline{\log(W^F)} - \overline{\log(W^M)} = \hat{\beta}^F(\overline{Z}^F - \overline{Z}^M) + (\hat{\beta}^F - \hat{\beta}^M)\overline{Z}^M$$

Now, average endowment differences are weighted by female workers' estimated coefficients while differences in the estimated coefficients are weighted by the average characteristics of male workers. Applied econometricians used both decompositions. I do not know of any argument to prefer one to the other. I think (but I am not sure) that the choice between the two decompositions has no serious practical implications[7].

Trade unionism.

I will use the same technique as for discrimination. Only the economic interpretation will change. I split my sample of workers between two sub samples: unionised workers, identified by superscript U, and non-unionised workers, identified by superscript N. I run the same regressions as in the case for discrimination. Simply I substitute unionised for female workers and non-unionised for male workers. I get the two equations:

$$\overline{\log(W^U)} - \overline{\log(W^N)} = \hat{\beta}^N(\overline{Z}^U - \overline{Z}^N) + (\hat{\beta}^U - \hat{\beta}^N)\overline{Z}^U$$

$$\overline{\log(W^U)} - \overline{\log(W^N)} = \hat{\beta}^U(\overline{Z}^U - \overline{Z}^N) + (\hat{\beta}^U - \hat{\beta}^N)\overline{Z}^N$$

Let us consider the first decomposition, for example. The first term of the right-hand side represents differences in wages, which are related to differences between endowments of the two sub samples of workers. The second term represents the effect of unionisation on wages. Belonging to a union increases wages *ceteris paribus* (the constant term of the earnings equation), but it also changes the effects of all the variables on wages (for instance formal education has less effect and seniority has more effect in unionised firms). The same technique could be applied to answer plenty of other questions such as: do small firms pay lower wages than big firms?

[7] An alternative method to investigate the question of discrimination per gender would be to assume that the earning functions of men and women differ only by their constant term. This assumption is probably wrong and prevents some interesting investigation such as is the return on education the same for men and women. However, when the size of the available sample is too small the econometrician can be forced to make such an extreme assumption.

Exercise 1 is interesting, but bears on descriptive statistics and is more related to chapter 1. Exercises 2 to 7 are very good. Exercise 8 is more advanced than the course.

A last example as a conclusion

I will use the data of mukherjee\birth.wk1. I want to explain the variability of the birth rate between several countries, by the GDP per capita of these countries and by their child death rates. I have data on 109 countries in 1985. The birth rate is the number of births per 1000 inhabitants in 1985 in the country. It is a coarse index, which does not take into account the proportion of women who have an age where they can bear a child. The death rate is the proportion of born children dead before the age of 1. Economic theory teaches us that the birth rate should decrease with income per capita. The intuition for this result is that the opportunity cost of time spent bringing up children increases with income. We can also think that if a family wishes to have a certain number of children reaching an adult age, birth rate must increase when the proportion of children dying young increases. The regression over 109 countries is (the Student-t statistics appear under the coefficients)

$$\begin{array}{l} Birth_i = 18.8 - 0.00039Y_i + 0.22IMR_t..R^2 = 0.79 \\ ....(11.65)(-2.23).......(14.04) \end{array} :$$

The results look good. 79% of the total variance of the birth rate is explained, which is good for a cross-section regression, the signs of the coefficients are right, the Student-t are significant. The apparent weakness of the coefficient of GDP per capita only results from the choice of units. So, we are tempted to stop our inquiry here.

However, it is wise to check if the residuals of the equation satisfy the assumptions required for OLS to give good estimates. To check the normality of the error terms we use a Jarque and Bera test. The P-value of this test is 5.74%, which is more than 5%, so we do not reject the null hypothesis of normality. We test homoscedasticity by checking if the variance of the residuals does not increase with any of the explanatory variable. The Goldfeld and Quandt test does not reject homoscedasticity. So, we are tempted to stop our inquiry here.

Let us try a graphical analysis and look at the histograms of each of the three variables. We can see that the histogram of the birth rate is approximately rectangular. The histogram of income per capita is strongly decreasing, and the histogram of the child death rate is slowly decreasing. How, can we explain a feature as essential as a rectangular histogram with two decreasing histograms?

Now, let us look at the scatter diagrams of the variables taken by pairs. We can notice that child mortality is a decreasing, strongly convex, function of GDP per capita. We observe the same thing for the rate of birth. However, the birth rate is an increasing function, slightly concave, of the child mortality rate.

Thus, it seems that we should transform the variables, first to get the symmetry of the histograms of the death rate and of GNP per capita, then to get linear relationships in the scatter diagrams. We can identify the best transformations by a succession of trials and errors. Or we can use more precise and sophisticated methods (not considered in these lectures). Finally, we substitute GNP per capita by its logarithm and the child death rate by its squared root. After these transformations both variables have approximately rectangular histograms, and the scatter diagrams exhibit linear relationships. However, some scatter diagrams present outliers. For instance, China and Sri Lanka have a birth rate lower than what is consistent with their low levels of income on the scatter diagram. They also have child death rates too low, for the levels associated with their low GNP on the scatter diagram. Oil producer countries are in the opposite situation. However, the scatter diagram between birth and death rates does not exhibit any outlier.

If we estimate the equation with the transformed variables, we get:

$$Birth_i = -2.59 - 0.63\log(Y_i) + 4.06(IMR_t)^{0.5} .. R^2 = 0.85$$

$$.............(-0.38)(0.925).......(13.78)$$

Now, 85%, instead of 79%, of the variance of the birth-rate is explained by the regression (note that this comparison is possible because we did not transform the dependent variable). The coefficient of GNP per capita has the wrong sign and is not significant. Thus it is wiser to remove this variable. We get:

$$Birth_i = 3.61 + 3.83(IMR_t)^{0.5} .. R^2 = 0.85$$

$$.............(2.75)(24.17)$$

This regression looks satisfactory. It is associated with a scatter diagram without outliers and to two rectangular histograms. The Jarque and Bera test has a P-value of 16.5% that is better than for the previous equation.

Thus, our result is that GNP per capita has no direct effect on the birth rate. Of course, it has an indirect effect: a higher GNP per capita is associated with a lower child death rate. However, this last relationship knows some exceptions, which were very helpful to reach our conclusion. China and Sri Lanka have low child mortality rates in comparison to their income, oil producer countries are in the opposite situation. We can also notice that social development is not always determined by income. For instance, if you consider the various states of India, you will find states with a high degree of social development and a low income per capita (Kerala) and states in the opposite situation.

CHAPTER 4. THE ECONOMETRICS OF TIME SERIES

Regression with time lags: distributed lag models

I will consider example Koop\safety.xls. A company bears due to industrial accidents. I will denote as Y_t the losses (in pounds) in month t. The same company can provide safety training to its workers. I will denote as X_t the number of hours of training provided to each worker in month t. I have data available for T=60 successive months.

An increase in the safety training of each worker should decrease accidents. Thus, I expect a negative influence of X on Y. I could regress Y on a constant term and X as in previous chapters. However, things are a little more complicated with time series. The safety training of a month will probably reduce accidents in the same month. However, it will be more effective the month after when every worker has better understood how to use its new knowledge in his workplace. After a few more months, the training will begin to be forgotten, workers will come back to their old working habits, and the number of accidents will increase again.

This suggest to run a regression of Y_t on a constant term and X_t, but also on lagged values of X_t, denoted X_{t-1}, X_{t-2}, X_{t-3} and X_{t-4}. For instance, if t represents May 1999, X_t will represent the number of hours of training in May 1999, X_{t-1} the number of hours of training in April 1999, X_{t-2} the number of hours of training in March 1999, etc. Thus, I will estimate equation:

$$Y_t = \alpha + \beta_1 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + \beta_4 X_{t-4} + \varepsilon_t$$

and I will get the estimate:

$$Y_t = 92001.51 - 145 X_t - 462.14 X_{t-1} - 424.47 X_{t-2} - 199.55 X_{t-3} - 36.90 X_{t-4} + e_t$$

This equation can be given the following interpretation. If the safety training of each worker is increased by 1 hour in April 1999 (only in April 1999, neither before nor after), losses due to industrial accidents will decrease by pounds 145 in the same month, by pounds 462.14 pounds in May, by pounds 424.74 in June, by pounds 199.55 in July and by pounds 36.90 in August. The policy of training will have no effect in the following months.

The above equation is called a distributed lag model. It includes lagged explanatory variables with lags equal to 1, 2, 3 and 4. 4, the highest lag in the equation, is called the lag order or the lag length of the equation. If I have 60 observations on the contemporary variables, I only have 59 observations on the variable with lag 1, 58 observations on the variable with lag 2, ..., 56 observations on the variable with lag 4. Actually, I can run the regression only for t going from 5 to 60, that is over 56 observations. This problem is well-taken care of by E-Views. However, you must be careful, check the sample used for the regression and compare it to the total available sample.

How to select the lag order of a distributed lag model? Why did I take an order lag of 4 in previous equation, instead of 3, or 5? There are several methods to make this choice. I will explain one of them.

First, I select a high lag order q. Then, I run the regression on a constant term and on X_t, X_{t-1}, .., X_{t-q}. The coefficient of the last variable is β_q. I use a Student test to check if the estimated coefficient $\hat{\beta}_q$ significantly differs from 0. If it does, I stop here and I keep the estimate of the regression. If it does not significantly differ from 0, I discard variable X_{t-q} from the equation. Then, I run the regression on a constant term, X_t, X_{t-1}, .., X_{t-q+1}. Now this regression has one less explanatory variable than before. The coefficient of the last variable is β_{q-1}. I use a Student test to check if the estimated coefficient $\hat{\beta}_{q-1}$ significantly differs or not from 0. If it does, I stop here and keep the estimate of the regression. If it does not significantly differ from 0, I discard variable X_{t-q+1} from the equation. Then, I run the regression on a constant term, X_t, X_{t-1}, .., X_{t-q+2}. Now this regression has one less explanatory variable than before. Etc.

For instance, in the above example, the P-value of the Student test on $\hat{\beta}_4$ is 44%. Thus, I can discard this lag and estimate the regression:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + \varepsilon_t$$

I get:

$$Y_t = 90402.22 - 125.90 X_t - 443.49 X_{t-1} - 417.61 X_{t-2} - 179.90 X_{t-3} + e_t$$

For this new regression, the P-value of the Student test on $\hat{\beta}_3$ is 0.003%. Thus, $\hat{\beta}_3$ significantly differs from 0, I stop here and keep the last estimated equation.

Thus to compute the lag order of the regression, I process recursively, starting with a large order, estimating the equation, computing a t-test on the last explanatory variable (the variable with the highest lag), removing this variable if the null hypothesis of the test is not rejected, estimating the new equation, etc.

This should remind you of what I told about partial correlation in chapter 3. Actually, the above procedure could equivalently be expressed in terms of partial correlation.

Now, I will introduce more advanced topics. I am interested in estimating the following equation:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + \varepsilon_t$$

A problem with this estimation is that, very often, a variable is strongly correlated to its lagged values (auto correlated in statistical language). Thus, very often, the four explanatory variables of this regression will be correlated to one another, and this multi-collinearity will result in very imprecise estimations of the parameters, with very wide confidence intervals. There are two solutions to this problem.

1) I can orthogonalise the explanatory variables by using transformations of these variables, which are little correlated to one another. The previous equation can be re-written:

$$Y_t = \alpha + (\beta_0 + \beta_1 + \beta_2 + \beta_3) X_t$$
$$-(\beta_1 + \beta_2 + \beta_3)(X_t - X_{t-1}) - (\beta_2 + \beta_3)(X_{t-1} - X_{t-2}) - \beta_3(X_{t-2} - X_{t-3}) + \varepsilon_t$$

I will define the first difference of explanatory variable X_t as $\Delta X_t = X_t - X_{t-1}$. Of course, I will have the lagged first difference variable defined as: $\Delta X_{t-1} = X_{t-1} - X_{t-2}$, etc.
I will introduce the new parameters: $\gamma_0 = \beta_0 + \beta_1 + \beta_2 + \beta_3$, $\gamma_1 = \beta_1 + \beta_2 + \beta_3$, $\gamma_2 = \beta_2 + \beta_3$, $\gamma_3 = \beta_3$. These 4 formulae determine the four new parameters γ_i in function of the 4 old parameters β_i, with i = 0, 1, 2, 3. It would be very easy to invert these 4 equations and get the expressions of the 4 old parameters in function of the 4 new parameters. Thus, the correspondence between the β and the γ is one to one. The regression can be re-written:

$$Y_t = \alpha + \gamma_0 X_t - \gamma_1 \Delta X_t - \gamma_2 \Delta X_{t-1} - \gamma_3 \Delta X_{t-2} + \varepsilon_t$$

The advantage of this new expression is that very often, X_t, ΔX_t, ΔX_{t-1} and ΔX_{t-2} are weakly correlated to one another (this can easily be checked under E-Views). So I am rid of the problem of multi-collinearity. The Frisch-Waugh theorem establishes that: 1) if I estimate the new parameters γ_i and compute the associated values of the old parameters β_i, or 2) if I directly estimate the old parameters β_i, I will get the same estimates. However, the new parameters will be estimated with good precision (small confidence intervals) and the old parameters will be estimated with large imprecision (large confidence intervals).

The above transformation is extremely used by practitioners, not only for the explanatory variable but for the lagged values of the explained variable also, as we will see in the next paragraph.

If, for instance, X represents the logarithm of GDP, ΔX will represent the growth rate of GDP. Then, the variables of the transformed equation have an economic meaning, and the equation itself can be given an economic interpretation. This makes this kind of transformation attractive.

2) I can put some clever constraints on the parameters β_i. We saw in last example that the β_i were hump-shaped: safety training had an immediate weak effect, it had a stronger effect in the short run, a weak effect in the medium run and no effect in the long run. Parabolas, that are the graphs of quadratic functions, can have the same shape. This suggests to constraint the four β_i to be located on a parabola, which will depend on three parameters, which will have to be estimated. These constraints can be mathematically expressed by $\beta_i = \gamma_0 + \gamma_1 i + \gamma_2 i^2$, $i = 0,1,2,3$.

We can easily prove that, with these new parameters, the regression can be re-written:

$$Y_t = \alpha + \gamma_0 V_t + \gamma_1 W_t + \gamma_2 Z_t + \varepsilon_t ,$$

with the new transformed variables:
$V_t = X_t + X_{t-!} + X_{t-2} + X_{t-3}$, $W_t = X_{t-1} + 2X_{t-2} + 3X_{t-3}$, $Z_t = X_{t-1} + 4X_{t-2} + 9X_{t-3}$.

It is easy to estimate the new equation by OLS, then to compute the values of the original parameters with the formulae relating the parameters β to the parameters γ. The three new variables will in general be less correlated to one another than the four old variables. Moreover, now we have just three parameters to estimate instead of four. We saw before that this will increase the precision of the estimation.

However, everything comes with a price. Here, we run the risk to have imposed wrong constraints on the parameters. Thus, the results of the regression can appear good. But a test of the constraints on the β parameters can reject these constraints with very low P-value.

This second method is known as polynomial distributed lags or Almon lags. It is implemented in E-Views (chapter 13 of the User's guide). It was extremely popular in the seventies and the beginning of the eighties. Then, it was killed by the error correction model, which will be presented later.

We can notice that now the transformed variables have no economic meaning, and the transformed equation cannot be given an economic interpretation.

Univariate time series analysis: the autoregressive model of order 1 (the AR(1) model)

In the previous paragraph I introduced a model explaining a variable by contemporary and lagged values of another variable. Could I explain a variable by lagged values of this variable itself? The answer to this question very often is yes.

The simplest equation, based on the above principle, explains variable Y_t by a constant term and its own value lagged once;
$$Y_t = \alpha + \phi Y_{t-1} + \varepsilon_t$$
It is called an AR(1) model.

To get a good understanding of such a model, I will make the very simple transformation of variable: $Y_t^{'} = Y_t - \alpha / (1 - \phi)$. So, I only subtract a fixed amount from Y_t. The equation becomes:
$$Y_t^{'} = \phi Y_{t-1}^{'} + \varepsilon_t$$

ε_t results from a series of independent random drawings. Usually, ε_t is assumed to follow a normal law. Here, for simplicity sake, I will assume that it can take value 1, with probability 0.5 and value -1, with probability 0.5.

If $\phi = 0$, at each period t, $Y_t^{'}$ will take value 1 with probability 0.5 and value -1 with probability 0.5. There will not be any time dependence in the values taken by $Y_t^{'}$. We will say that $Y_t^{'}$ follows a white noise.

If $\phi = 0.5$ and if initially $Y_0^{'} = 0$, $Y_1^{'}$ will take value 1, with probability 0.5 and value -1, with probability 0.5. If $Y_1^{'}$ takes value 1, $Y_2^{'}$ will take value 1.5, with probability 0.5 and value -0.5, with probability 0.5. If $Y_2^{'}$ takes value 1.5, $Y_3^{'}$ will take value 1.75, with probability 0.5 and value -0.25, with probability 0.5, etc. Thus, a high value of $Y_{t-1}^{'}$ will be followed by a high value of $Y_t^{'}$, equal to $0.5Y_{t-1}^{'} \pm 1$. However, as soon as $Y_{t-1} > 2$, $Y_t^{'}$ will be smaller than $Y_{t-1}^{'}$. Thus, even if $Y_t^{'}$ can from time to time go a little far from 0 and stay at some distance from this value, finally it will be drawn toward 0, and over a long period of time it will fluctuate around 0. Statisticians use the expression *mean-reverting process* that I find very expressive. Thus, we have a short run time dependence: $Y_t^{'}$ strongly depends on $Y_{t-1}^{'}$, less strongly on $Y_{t-2}^{'}$, and is almost independent of its far past values. This result is sensible to understand the real world. The present depends much on the near past, but little on the far past. A series of random variables Y_t which may depend on their near past, but which are almost independent of their far past will be called a stationary stochastic process[8].

To have a better understanding of this result, we can notice that $Y_t^{'} = 0.5Y_{t-1}^{'} + \varepsilon_t$, implies that $Y_{t-1}^{'} = 0.5Y_{t-2}^{'} + \varepsilon_{t-1}$. If I use the second equation to substitute $Y_{t-1}^{'}$ in the second equation, I will get $Y_t^{'} = 0.25Y_{t-2}^{'} + \varepsilon_t + 0.5\varepsilon_{t-1}$. If I continue recursively to substitute for the lagged variable backward, I will finally get:

$$Y_t^{'} = 0.5^t Y_0^{'} + \varepsilon_t + 0.5\varepsilon_{t-1} + 0.5^2 \varepsilon_{t-2} + .. + 0.5^{t-1} \varepsilon_1$$

ε_1 can be interpreted as a random shock which hit the variable of interest $Y_1^{'}$ at time 1. At time t this shocks has still an effect on $Y_t^{'}$, equal to $0.5^{t-1}\varepsilon_1$. But the size of this effect has become very small. Actually, it decreases geometrically at rate 0.5. Then, the recent past affects the present, but this effect decreases geometrically over time.

If I assume that t is high, that ε_t has a mean equal to 0, and a constant variance equal to σ^2, at time 0, the expected value of Y_t is 0, and its variance is $\sigma^2(1 + 0.5^2 + 0.5^4 + 0.5^6 + ..) = \sigma^2 / (1 - 0.5^2) = \sigma^2 / (1 - \phi^2)$. This term is finite and constant.

[8] Actually, this definition is incorrect. However, being more rigorous would require much more sophisticated mathematics than the ones used here. The stationarity of a random process is the constancy of its mean, its variance and its autocovariances over time. A change in regime in an economic time series, for instance a period of time with a high inflation rate followed by a period of stable prices, is an example of non-stationarity. On the other hand, there is a stochastic trend in a series if it is non stationary, but if its first difference is stationary. Thus, stochastic trends are very specific cases of non-stationarity. This kind of non stationarity has destructive implications for econometrics. The effects of the other kinds of non-stationarity are less clear. However, we can hope that if we follow the advice of the previous chapters we will avoid the worst mistakes.

Time series theory proves that the correlation coefficient between Y_t and Y_{t-q} is equal to 0.5^q, or to ϕ^q in the general case when $0 \leq \phi < 1$. Thus, it decreases with q at geometrical rate ϕ. The correlation between two random variables measures the strength of the dependence between these variables. Thus, the previous result means that the present depends very much on the near past, and very little on the far past. The correlation coefficient between Y_t and Y_{t-q} will be called r_q, the autocorrelation of order q, and we can draw a graph, with r_q on the Y-axis and q on the X-axis. This graph is called the autocorrelation function. Koop, page 129, gives two examples of such graphs.

What will happen if $\phi = 1$? First, I cannot make the previous transformation of variable, which would include a division by 0. The original equation is

$$Y_t = \alpha + Y_{t-1} + \varepsilon_t$$

The transformation of variable, which I will use now, is $Y_t^{'} = Y_t - \alpha t$. Thus, instead of subtracting a constant term from the original variable, I will subtract a linear deterministic trend αt. This operation is called detrending the series.
The equation with the transformed variable is:

$$Y_t^{'} = Y_{t-1}^{'} + \varepsilon_t .$$

This is referred to as the random walk model. If initially $Y_0^{'} = Y_0 = 0$, $Y_1^{'}$ will take value 1, with probability 0.5 and -1, with probability 0.5. If $Y_1^{'}$ takes value 1, $Y_2^{'}$ will take value 2, with probability 0.5 and value 0, with probability 0.5. If $Y_2^{'}$ takes value 2, $Y_3^{'}$ will take value 3, with probability 0.5 and value 1, with probability 0.5. Etc. Thus, a high value for $Y_{t-1}^{'}$ will be followed by a high value for $Y_t^{'}$, equal to $Y_{t-1} \pm 1$. But now, there is no more strength pulling $Y_t^{'}$ with 0.

Actually, we have $Y_t^{'} = \varepsilon_t + \varepsilon_{t-1} + ... + \varepsilon_2 + \varepsilon_1$. The past eternally keeps its influence on the present, instead of having an effect, which decreases at a geometrical rate. Nothing is ever forgotten. If I was lucky when I was in my twenties, I will still benefit from the gains of my past luck in my sixties. In this case, when the coefficient of the lagged variable is equal to 1, I will say that the stochastic process of the $Y_t^{'}$ is non stationary with a unit root. We can also say that the series is integrated of order one (or has a stochastic trend), which means that its first difference is stationary.

If I assume that ε_t has a mean equal to 0, and a constant variance equal to σ^2, at time 0, before the firs drawing of an error term, the expected value of Y_t is 0, and its variance is $\sigma^2 t$. Thus, the variance of the variable of interest increases linearly with time. If, at the same time I compute the square of the autocorrelation between Y_t and Y_{t-q}, with $t - q \geq 1$, I get $1 - q / t$. Thus, for given t, this autocorrelation decreases linearly with q. In the stationary case it decreases geometrically. Thus, we have identified a series of differences between the stationary case and the non stationary unit root case.

Unit roots are very frequent in macroeconomics. For instance, example koop\income.xls, considers the logarithm of the personal income in the US from 1954Q1 to 1994Q4. The estimation of the AR(1) equation on this series gives $Y_t = 0.039 + 0.996Y_{t-1} + u_t$. We are very near the unit root equation. In this case $Y_t - Y_{t-1}$ is the growth rate of personal income. If I estimate this equation under the constraint that the coefficient of Y_{t-1} is equal to zero, I will get $Y_t - Y_{t-1} = 0.008 + u_t$. That means that the growth rate of real personal income in the US can be split between two components. The first is deterministic and its estimate is equal to 0.8% per quarter. It is called a deterministic trend. The second is random and equal to ε_t. Its estimates on the past are the u_t, the residuals of the equation. Sometimes u_t is positive, sometimes it is negative. It is called a *stochastic trend*.

When you look at an economic time series over, for instance, 80 quarters, it generally has an apparent positive trend. However, it is very difficult to separate by mere sight the part of this trend which is deterministic from the part which is stochastic. If you simulate the equation: $Y_t^{'} = Y_{t-1}^{'} + \varepsilon_t$ over 80 periods, you might get an increasing path, or a decreasing path, or a path having the shape of a U, or of an inverted U, etc. Identifying that this path is a pure random realisation of the previous model instead of a deterministic movement with important economic meaning cannot be decided by simply looking at the graph. You must use sophisticated statistical tests.

In conclusion, the equation $Y_t = \alpha + \phi Y_{t-1} + \varepsilon_t$, generates very different paths for variable Y, when $0 \le \phi < 1$ and when $\phi = 1$. If we remember that Δ is the first difference operator, the previous equation can be re-written
$\Delta Y_t = \alpha + \rho Y_{t-1} + \varepsilon_t$, with $\rho = \phi - 1$. Thus, the non-stationary case is associated to: $\rho = 0$, and the stationary case to: $-1 \le \rho < 0$.

In this paragraph, I have considered only the cases when $0 \le \phi < 1$, and $\phi = 1$. The cases when ϕ is negative, or is bigger than 1, are not very interesting because they are associated with realisation paths of variable Y, which are unlikely to exist for economic time series.

Univariate time series analysis: the autoregressive model of order p (AR(p) model

Let us consider for example the case when p=4. An AR(4) is defined by:
$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \phi_4 Y_{t-4} + \delta t + \varepsilon_t$
I have introduced a deterministic trend δt in the right-hand side of the equation just to be more general. A yearly time series can sometimes be represented with a good fit by an AR(1). A quarterly or a monthly time series will need an AR with a higher order. This order can be computed as in the first paragraph by a series of Student tests on the coefficient of the variable with the highest lag. A quarterly or a monthly time series will probably exhibit some seasonality. There are several ways to deal with this feature, but they will be better presented in a course on time series analysis. In his book, Koop presents the simplest method, which is to introduce seasonal dummies variables in the equation. The only question I will investigate about this equation will be: is the series (i.e. the stochastic process) of the Y_t stationary or not? To answer

this question, I will re-write the previous equation by using the operator first difference Δ :

$$Y_t - Y_{t-1} = \alpha - (1 - \phi_1 - \phi_2 - \phi_3 - \phi_4)Y_{t-1} - (\phi_2 + \phi_3 + \phi_4)(Y_{t-1} - Y_{t-2})$$
$$-(\phi_3 + \phi_4)(Y_{t-2} - Y_{t-3}) - \phi_4(Y_{t-3} - Y_{t-4}) + \delta t + \varepsilon_t$$

or with an evident transformation of parameters:

$$\Delta Y_t = \alpha + \rho Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \gamma_3 \Delta Y_{t-3} + \delta t + \varepsilon_t$$

As in the previous paragraph, the non-stationary case with a unit root is associated to: $\rho = 0$, and the stationary case to: $-1 \leq \rho < 1$. I want to test the presence of a unit root against the alternative of stationarity. The problem is that under the hypothesis of non-stationarity the Student statistics of ρ will not follow a Student distribution. Thus, I cannot simply compare this statistic to 1.96. Moreover, in the case of a unit root, the Student statistics of the other coefficients of the regression do not follow Student distributions (I think that page 140 of the book by Koop is wrong). Thus testing $\rho = 0$ is more complicated than we could have expected. The tests I will use are called augmented Dickey-Fuller (ADF) tests. The term augmented is relative to the lags on ΔY_t which appear in the equation in order to improve its fit. I will present a strategy of nested ADF tests. I will apply it to three series of Mukherjee\tanmon.wk1. These series concern Tanzania and are log(M2), the logarithm of the quantity of money, log(CPI), the logarithm of the consumer price index and XGROW, the growth rate of exports[9]. The statistical tables of Dickey and Fuller are given page 480 of the book. You will find the same tables in the book by Walter Enders, *Applied Econometric Time Series*, John Wiley and Sons, pages 419-421. The ADF test is explained in this book on pages 221-227.

First step.

I run the regression $\Delta Y_t = \alpha + \rho Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \gamma_3 \Delta Y_{t-3} + \delta t + \varepsilon_t$

The null hypothesis is $\alpha = \rho = \delta = 0$. The model can be re-written under this hypothesis:

$$\Delta Y_t = \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \gamma_3 \Delta Y_{t-3} + \varepsilon_t$$

The lag order of the autoregression can be computed on this last equation by a succession of Student tests, as explained before. The alternative hypothesis is $\alpha * \rho * \delta \neq 0$.

I will compute the Fisher statistics associated with the three constraints defining the null hypothesis (this statistic is computed by E-Views under the title Wald test). But, I will not use the P-value given by E-Views, nor a Fisher table.

I will use the critical value given for statistics Φ_2 of page 480 of the book, which at the 5% significance level for a sample of size 25 is 5.68 (4.67 at 10%). This means that if the null hypothesis is true, there is a probability of 5% to get a statistic larger than 5.68. I get for this statistic a value of 3.30 for log(CPI), of 6.83 for log(M2) and of 5.60

[9] These series are yearly. Thus, I will limit myself to only one lag for the first difference term (instead of three lags)

for XGROW. Thus, I will accept the null hypothesis for log(CPI), and conclude that this series presents a stochastic trend without a deterministic trend.

If I want to be fully assured of the validity of this conclusion, I can re-estimate the equation for log(CPI) after having put $\delta = 0$, that is the equation:

$$\Delta Y_t = \alpha + \rho Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \gamma_3 \Delta Y_{t-3} + \varepsilon_t$$

Then, I test the null hypothesis: $\alpha = \rho = 0$. The alternative hypothesis is: $\alpha * \rho \neq 0$. Both hypothesis do not assume anything about δ.

I compute the Fisher statistics of this test and I compare it to the critical value of statistics $\Phi_1$of page 480 of the book, which at the 5% significance level for a sample of size 25 is 5.18 (4.12 à 10%). I get a statistic much lower than these values and I keep the previous conclusion.

Second step

I have not reached any conclusion yet for series log(M2) and XGROW. I will test for these series the null hypothesis: $\delta = \rho = 0$. The model can be re-written under this hypothesis:

$$\Delta Y_t = \alpha + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \gamma_3 \Delta Y_{t-3} + \varepsilon_t$$

The lag order of the autoregression can be computed on this equation by a succession of Student tests, as explained before. The alternative hypothesis is $\delta * \rho \neq 0$. Both hypothesis assume nothing on α .

I compute the Fisher statistics associated to these constraints (with E-Views). I compare them to the critical value of statistics $\Phi_3$of page 480 of the book, which at the 5% significance level for a sample of size 25 is 7.24 (5.91 at 10%). I do not reject the null hypothesis for log(M2) (its statistic is equal to 3.61) but I reject it for XGROW (its statistics equal to 8.39). Thus, the logarithm of the quantity of money presents a stochastic trend and a deterministic trend.

If I want to feel secure, I can test only $\rho = 0$. The null hypothesis is $\rho \neq 0$. Under both hypothesis i do not make any assumption on $\alpha$and δ. I will compute the Student statistic which is equal to –0.849. I will compare it to the critical values given for τ_3 page 480. At 5% significance level, the critical value is –3.60 (-3.24 at 10%). This means that if the null hypothesis is true, there is a probability of 5% of having a statistic smaller than –3.60. Thus, I will accept that log(M2) has a stochastic trend.

If I did the same test for the equation without a deterministic trend (estimated after having imposed $\delta = 0$, as in the end of the first step) I would get a positive statistics, which I should compare to the critical values given for τ_2 page 480 (-3 and –3.62). Thus, I will still not reject the null hypothesis of a stochastic trend.

Thus, I find that in Tanzania, the logarithm of the quantity of money exhibits a deterministic trend and a stochastic trend, and the logarithm of price only includes a stochastic trend. Does this difference between money and price have an economic meaning? Over the observation period the inflation rate was always positive. But, it also increased, and we can notice a break in the graph of log(CPI). If this variable had followed a stochastic trend without a deterministic trend, we should have observed some negative inflation rates, which is not the case. I will regress $\Delta \log(CPI)$ on its lagged value and on a constant term. The constant term has a Student statistic slightly greater than 2. Residuals are not distributed uniformly around 0 (there is a difference between the beginning and the end of the sample. If I run the regression over various subsamples, I get very different estimates of the coefficient of the lagged variable. Thus, I will add to the equation a dummy variable equal to 0 before 1980 and to 1 in 1980 and afterward. Its Student statistics and the Student statistics of the constant term are larger than 3. Thus, it seems that the logarithm of price has a deterministic trend with a break. The automatic process of successive steps, which I used, was insufficient to detect this feature in the series. Thus, I will conclude that the logarithms of the quantity of money and of the consumption price present both a deterministic trend and a stochastic trend. But the deterministic trend of the second series has a break, which could be identified only after some reflection.

Step 3

If I have not yet reached a conclusion, which is the case for XGROW, I will test in the original equation $\rho = 0$, and I will compare the Student statistics of this test to the critical value associated with a significance level of 5% for a standard normal distribution. This critical value is -1.96. The value if the statistics of the test is -4.1, which is much lower So, I reject the null hypothesis $\rho = 0$.

Step 4

If I have reached this step, which is the case for XGROW, I will conclude that the series has no stochastic trend. Then, I can test if it has a deterministic trend by computing the Student statistics of δ and by comparing it to the critical values of a Student table (by simply running a Wald test under E-Views). In the present case, I do not reject the null hypothesis and I conclude that XGROW has no deterministic trend.

There is a problem with ADF tests. They test the null hypothesis of non-stationarity with a unit root against the alternative hypothesis of stationarity. If we use these tests on a series which is non stationary, but which does not have a unit root, for instance a series going through several regimes, like the inflation rate or the interest rate which were very high until the mid eighties and low afterward, the tests will probably conclude that the series has a unit root. It is of course a bit silly to conclude that the interest rate follows a random walk when there exist a powerful strength pulling it toward levels around 5% per year. In statistical language I will say that ADF tests lack power against states of the world, which differ from the null and the alternative hypothesis[10]. Moreover, these tests require samples large enough (around 100 at least). They become doubtful if they are used on small samples.

[10] Actually, we cannot discriminate between a series of 80 points generated by a random walk and a series generated by a stationary autoregressive process with a few well-chosen structural breaks. The robust methods presented in chapters 1 and 2 can be of some help, and should be used besides of ADF tests.

Regression with time series variables : the case with stationary variables

If I merger the specifications of equations introduced in the first and the third paragraph of this chapter, I will get the autoregressive distributive lag (ADL) model:

$$Y_t = \alpha + \delta t + \phi_1 Y_{t-1} +..+ \phi_p Y_{t-p} + \beta_0 X_t + \beta_1 X_{t-1} +..+ \beta_q X_{t-q} + \varepsilon_t$$

We saw how to re-write this equation into the equivalent form:

$$\Delta Y_t = \alpha + \delta t + \rho Y_{t-1} + \theta X_{t-1} + \gamma_1 \Delta Y_{t-1} .. + \gamma_p \Delta Y_{t-p} + \theta X_t + \omega_0 \Delta X_t + .. + \omega_{q-1} \Delta X_{t-q+1} + \varepsilon_t$$

When variables X and Y are both stationary, we face an easy problem. Estimation and tests can proceed as in Chapter 3. Actually, successive Student tests can be used to determine both lag orders p and q.

The second equation can be used to introduce the concept of the long run multiplier (LRM). If I increase X_t by an amount of 1, for every value of t larger than q, in the long run Y_t will increase by a constant amount equal to LRM, and ΔX_t and ΔY_t will not move. Thus, we must have $LRM = -\theta / \rho$. In the language of system analysis I will tell that a permanent increase in the input by 1, will induce in the long run a permanent increase in the output by $LRM = -\theta / \rho$. This must be interpreted all other things being kept equal, or *ceteris paribus*. You remember that in the example on industrial accidents and training, I considered transitory changes in the input (limited to 1 month). Here, these changes are made permanent.
Koop gives a nice example of an ADL model in pages 150 and 151.

The ADL model can be written

$$\Delta Y_t = \rho[Y_{t-1} + \alpha / \rho + (\delta / \rho)t + (\theta / \rho) X_{t-1}] + \gamma_1 \Delta Y_{t-1} + .. + \gamma_p \Delta Y_{t-p} + \omega_0 \Delta X_t + .. + \omega_{q-1} \Delta X_{t-q+1} + \varepsilon_t$$

The first term on the right-hand side represents the difference between the level on the explained variable and the level of the explanatory variables in the previous period. If this difference is zero, we can interpret it as a long-run equilibrium. If it is nonzero, we can interpret this value as disequilibrium or an “error”. The proportion ρ of this error is corrected by a movement in the explained variable in the current period. However, this movement also depends on the error term, lagged value of this movement and current and lagged values of the variation of the exogenous variable, which compose the short run dynamics of the model. This model is called an error component model. IT is much more used when variables are non-stationary, but it has already a meaning with stationary variables.

Regression with time series variables: the case with non-stationary variables: spurious regressions

Consider the very simple time series regression:

$$Y_t = \alpha + \beta X_t + \varepsilon_t$$

It can be considered as a simplified version of the ADL model presented above. When X and Y are non stationary with a unit root, this regression meets plenty of problems. For instance, if the true value of β is zero, its estimate could differ a lot from 0, a Student test could conclude that it significantly differs from 0 and the R^2 of the regression could be quite high. Thus, the regression can be wholly spurious. We can easily understand that with the following exercise with E-Views.

Go into E-Views. Create a work file with yearly data over the period 1960-2000 (41 observations).

Write on the command line the following instructions. At the end of each line push the Enter key. Nrnd is a command, which produces a number which can be considered as resulting from the drawing of a random variable with a standard normal distribution.

```
Smpl1960 1960
Series x=1
Series y=1
Series z=1
Series t=1
Smpl 1961 2000
Series x=x(-1)+nrnd
Series y=y(-1)+nrnd
Series z=0.5*z(-1)+nrnd
 Series t=t(-1)+1
Smpl 1960 2000
```

Look at series t.

Look at series x. Does it have a trend? Regress x on t. Does the coefficient of t significantly differ from 0? Look at the R2 of the regression. Do the residuals look homoscedastic? Process a White test. In theory the variance should be a linear function of time. Look at the Durbin and Watson statistics.

Do the same things with y and t.

Do the same things with y and x. In 20% cases the Student statistic is smaller than 1.67, in 5% cases it is between 1.67 and 2, in 9% cases between 2 and 2.67, in 66% it is larger than 2.67.

We get what is called a *spurious regression*. Understanding such a regression, and realising that many regressions run in macroeconomics are actually spurious, might be the most important result of the econometrics of variables with stochastic trends. In a spurious regression, the R2 is often high, but the DW is low. Experience showed that when R2>DW, then the regression is probably spurious. Moreover, in such a regression, the Student's test rejects more and more often the right hypothesis that $\beta = 0$ when the size of the sample increases.

Look at the series z. Compare its mean-reverting feature to the random walk feature of the previous series. For this last kind of series history has eternal consequences (hysteresis).

Exercise. Use the data in Mukherjee\Tanmon. Look at variables CPI and log(CPI) for Tanzania over 1966-1992. Look at M2 and log(M2). Look at the trends and the dispersions of the differences from the trend.

Regress log(CPI) on a constant term and on log(M2). Will this regression turn you into a monetarist?

Regression with time series variables: the case with non-stationary variables: Cointegration

We saw that running the regression; $Y_t = \alpha + \beta X_t + \varepsilon_t$ can be very dangerous when variables X and Y are non-stationary with a unit root. However, there is a very important case where this regression has a meaning. Let us write it a bit differently: $Y_t - \alpha - \beta X_t = \varepsilon_t$. X_t and Y_t both have stochastic trends. That means that the paths they follow go in any direction, without any mechanisms pulling them toward equilibrium values. It is this feature which was given the appellation of a random walk. Very often, the paths of the two variables are unconnected or loosely connected. That means that ε_t will also follow a random walk that is that no strength will force the two variables to follow their erratic paths together (as two drunken friends leaving the last pub on Saturday night and mutually holding their shoulders).

However, sometimes, for a precise value of α, ε_t will follow a stationary stochastic process without any stochastic trend. That means that variables X_t and Y_t will wander randomly, but together (like the two drunken friends). In this case, X_t and Y_t are said to be cointegrated[11].

This concept has a very intuitive economic meaning. Economic theory investigates behavioural relationships, like consumption being a function of present and past incomes and wealth. Or it builds equilibrium equations, for instance equalising to zero the excess demand for a commodity, which is a function of the price of this commodity and close substitutes to it. The economic variables, which appear in these functions and equations, most often, have stochastic trends. In the real world, you do not expect these functions and equations to be perfectly valid. Thus, you will introduce in them a random error term. However, for economic functions and equations to have meaning, this error term must fluctuate around zero. It cannot have a stochastic trend. If your equilibrium equation is wrong by 2% some day, you will expect it to be wrong by -2% some other day: your error term must follow a mean-reverting process pulling it toward 0. Otherwise, you could hardly consider your equilibrium relationship to be true equilibrium. Thus, macroeconomists love finding co-integration relationships, because most often they can be given an economic interpretation.

[11] Another denomination would be to say that they have a common (stochastic) trend.

Koop's book gives (pages 154, 155, 157, 158, 161 and 162) an illuminating example of two co-integrated variables, which are the price of regular oranges and the price of organic oranges.

How can we test if two variables X_t and Y_t are co-integrated? Engle and Granger devised a very simple test. First run an OLS regression of Y_t on X_t and a constant term and save the residuals. Then, carry out a unit root test on the residuals. You will have to run the same equation as in page 39, but without including in it a constant term nor a deterministic trend (so, after having put $\alpha = \delta = 0$). In this test the null hypothesis is that the residuals include a unit root, which implies that X_t and Y_t are non-cointegrated. If you reject the null hypothesis, you will accept that the residuals are stationary, that is that X_t and Y_t are cointegrated. The test uses the t Student statistics on ρ. A natural idea would be to compare this statistics to the values given for the ADF test τ_1, page 480 in the book by Mukherjee and alii (or in the book by Enders). However, there is a small complication taking care of. Using an ADF table would be right if we observed the values of the error terms ε_t. But we only observe the values of the residuals of the equation u_t, which are estimates of the error terms. Because of that we must use a different statistical table, built by Engle and Granger. However, the ADF table and the Engle and Granger table are rather similar.

If we conclude that X_t and Y_t are cointegrated, the OLS regression, which was run in the first step of the test, is no more spurious, and estimate $\hat{\beta}$ is a good estimate of the long run multiplier of X on Y. Actually, the theory of cointegration proves that $\hat{\beta}$ is still a better estimator than in the classic situation where all the variables are stationary (this property is called *super-convergence* in statistical language). However, although I think that this result is theoretically interesting, I am less convinced by its practical implications. In the regression of Y_t on X_t we are not allowed to make Student or Fisher tests on the estimates of the parameters. These estimates follow non-classical and complex distributions.

.

If X_t and Y_t are cointegrated, $Y_t = \alpha + \beta X_t + \varepsilon_t$ represents the long run relationship existing between X and Y, and β is the long run multiplier, which measures what effect in the long run a permanent increase in X will have on Y. The short run and medium run dynamics of Y are still to be determined. Engle and Granger noticed that u_t, the residual of the cointegration process, is stationary. Moreover, the first differences of the two variables, ΔX_t and ΔY_t, are also stationary. Thus it seems natural to look for and ADL equation relating these variables:

$$\Delta Y_t = \alpha' + \lambda u_{t-1} + \gamma_1 \Delta Y_{t-1} .. + \gamma_{p-1} \Delta Y_{t-p+1} + \omega_1 \Delta X_t + .. + \omega_q \Delta X_{t-q+1} + \nu_t$$

In this equation I have denoted the error term ν_t, to avoid confusion with the error term of the cointegration relationship which was denoted ε_t. This equation is called an error correction model (ECM). To understand this appellation you must remember that when u_{t-1} is positive, that means that Y_{t-1} is above is normal or equilibrium value. For a well-behaved ECM we must have $-1 < \lambda < 0$. That means that, all other things being constant, we will have: $-u_t < \Delta Y_t < 0$, or $Y_{t-1} - u_t < Y_t < Y_{t-1}$. Thus, the desequilibrium or the error on Y, observed at time t, will be partly corrected at time $t-1$.

An interesting property of an ECM equation is that all the variables of the equations are stationary. Thus, we can estimate it as an ordinary ADL equation, and test the significance of its parameters with classical Student and Fisher tests.

Engle and Granger proposed to estimate an ECM in two steps. First, test for cointegration. If this hypothesis is accepted, keep the residuals of the cointegration equation, put them in the ECM, and estimate it with OLS. The estimation of the cointegration equation will give the long run multiplier of the relationship between X and Y. The estimation of the ECM will give the short and medium run dynamics.

Exercise 1. Use the data in Mukherjee\pakep.wk1. LX represents the logarithm of the volume of exports by Pakistan. RER is the logarithm of the real exchange rate of Pakistan. If RER increases, the Pakistanese currency depreciates in real terms. A strong devaluation occurred in 1980. Then, its effects were cancelled by a strong inflation. Then, a structural adjustment plan has induced a continuous real depreciation of the Pakistanese currency since 1985. The graphs of the two variables look similar. Draw a scatter plot. ADF tests conclude that both variables are non-stationary with unit roots.
Run the regression of LX on a constant term and RER. Compute the ADF test on the residuals. The Student statistics on ρ takes value –1.88. It is smaller than the critical value given by Engle and Granger (-3.37). Thus, we will reject the hypothesis that the two series are cointegrated. However, after having looked at the graphs of the two series, are you fully convinced by this conclusion?

Exercise 2. Use the data in Mukherjee\crcon.wk1. Look at the series log(C) and log(Y), which are the logarithms of real consumption and GDP per capita in Costa Rica. Both variables are non-stationary with unit roots. Look at the graphs of both variables. Run the regression of log(C) on a constant term and log(Y). Think a little about the strange elasticity you get. Does the Engle and Granger test reject cointegration?
Mukherjee and *alii* wrote that they estimated such a consumption function for 8 countries (industrialised and developing). Each time they rejected the hypothesis of cointegration between consumption and income. This conclusion looks a bit strange and raises some suspicion on the Engle and Granger test.

In the above examples the relationship between the two series which were investigated, was each time strong but imperfect. The robust methods and fragility analysis presented in previous chapters, can precise the nature of these imperfections. We should consider that the rejection of cointegration might come from these imperfections and not from a stochastic trend in the error term. Anyway, a convincing identification of such a trend would require a lengthier period of observation and smoother economic conditions, consistent with the stationarity of the first differences of economic variables in Pakistan and Costa Rica. In Costa Rica, something special happened at the end of the seventies, which increased the share of consumption in GDP. Thus, something important is lacking in the consumption equation, which is related to economics and not to the theory of cointegration. A last remark: remember what we told about the fragility of the ADF test when the true state of the world strongly differs from the null hypothesis and the alternative hypothesis. And the Engle and Granger test has some kind of ADF test embodies in it.

My personal experience is that Engle and Granger test almost always rejects the hypothesis of cointegration. This result can be proved mathematically and interpreted as a weak power of the test. There is better way testing for cointegration and estimating an ECM, which does not suffer from this weakness in power. Remember that an ECM can be written:

$$\Delta Y_t = \alpha' + \lambda(Y_{t-1} - \alpha - \beta X_{t-1}) + \gamma_1 \Delta Y_{t-1} .. + \gamma_{p-1}\Delta Y_{t-p+1} + \omega_1 \Delta X_t + .. + \omega_q \Delta X_{t-q+1} + \nu_t$$

or

$$\Delta Y_t = \alpha'' + \lambda Y_{t-1} - \beta' X_{t-1} + \gamma_1 \Delta Y_{t-1} .. + \gamma_{p-1}\Delta Y_{t-p+1} + \omega_1 \Delta X_t + .. + \omega_q \Delta X_{t-q+1} + \nu_t$$

with $\alpha'' = \alpha' - \lambda\alpha \;\; \beta' = \lambda\beta$

Variables X and Y appear with one lag. This equation can be estimated by OLS. λ must significantly differ from 0 for X and Y to be cointegrated. To test this assumption we shall compute the Student statistics of λ. This statistics does not follow a Student distribution[12]. However, Bosjwyk computed tables of critical values for this statistic. These tables differ when there is (or not) a constant term or a trend in the ECM. The opposite results often given by the Engle and Granger test and the Bosjwyk test raise some suspicion on cointegration tests. Ericsson and MacKinnon added improvements to Bosjwyk' test.

My personal experience is that the two steps Engle and Granger method is inadequate in macroeconomics: the available series are too short and they cannot be made stationary by simply taking their first differences. The direct estimation of the ECM and the test of the significance of the coefficient of the lagged explained variable is better. However, we must not forget to use the robust methods and exploratory analysis presented in previous chapters.

In the example on spurious regressions, we introduced two independent series, and noticed that they were not cointegrated. However, two series can depend on each other and be non-cointegrated. I will give an example. Let us denote by ε_t and ν_t two independent white noises with standard normal distributions. I will define two random walks by:

$$x_t = x_{t-1} + \varepsilon_t \quad x_0 = 0$$
$$z_t = z_{t-1} + \nu_t \quad z_0 = 0$$

I can easily show that at time 0, I have:

$$Ex_t = Ez_t = 0, \; Ex_t^2 = Ez_t^2 = t$$

[12] Under the assumption of cointegration, the Student statistics of the coefficients of the non-stationary variables of the ECM do not follow a Student distribution. However, the Student statistics of the coefficients of all the stationary variables of the equation and of the constant term follow a Student distribution.

I will build the non-stationary variable: $y_t = 10x_t + z_t$

y_t is a random walk: $y_t - y_{t-1} = 10\varepsilon_t + \nu_t$, with $10\varepsilon_t + \nu_t$ a Gaussian white noise of mean 0, standard deviation 11 and with: $y_0 = 0$. At time 0 I have: $Ey_t = 0$, $\mathrm{E}\, y_t^2 = 11t$. y_t and x_t are non-cointegrated : the difference $y_t - 10x_t$ follows a random walk, and has a variance equal to t that is increasing indefinitely over time. This does not prevent that the knowledge of x_t brings an important information on y_t. If we are at time 0, and if we try to forecast the value which will betaken by y at time t, if we do not know x_t we will forecast that $y_t = 0$, and the variance of the forecast error will be $11t$. If we know x_t, we will forecast $10x_t$ and the variance of the forecast error will be t. Of course, this variance increases over time. But it is 11 times smaller than when we do not know the value taken by x_t.

Thus, knowing the dependence relationship between two non-stationary variables with a unit root y_t et x_t is useful, even when these variables are non-cointegrated. For instance, knowing that: $y_t = 6x_t + z_t$, with z_t non-stationary and independent of x_t, is a precious information if we want to forecast y conditionally to x, or if we want to understand economic behaviour. Unfortunately, time series econometrics does not know how to estimate equations relating non-stationary and non-cointegrated variables. But panel data econometrics recently developed methods to estimate such equations.

The literature on the econometric theory of non-stationary variables and of cointegration is huge and sophisticated. However, its applications to applied macroeconomics are sometimes unconvincing. For instance, integration and cointegration tests often assume null and alternative hypotheses, which are both clearly wrong. What does the conclusion of these tests mean under these circumstances? You remember that I have advocated using robust methods and fragility analysis to explore these situations. An automatic application of the theory of integration and cointegration to economic problems can easily give incredible and queer results because the methods developed by this theory are fragile. Thus, when you apply these methods you must not forget the rest of econometrics nor economics, if you want to reach serious results.

A different problem is that the tests and methods that I have presented in this chapter are the most ancient and the best known ones. They are neither the most efficient nor the most robust ones. There exist plenty of more recent methods and tests. However, even if they look better, their precise qualities in realistic situations are not well understood. This is why most of them are not implemented in econometric software like E-Views. This is also why econometric books are not unanimous in the strategies they advise dealing with non-stationary variables. There is still much controversy among the best econometricians about the best ways to apply their theoretical results in applied econometrics.

Two last remarks. First, it is unreasonable to use the methods of this chapter on series which do not have at least 80 observations and which present (themselves or their first-differences) a regular pattern (without breaks in levels or trends). Second, the negative results on spurious regressions are very important: plenty of macroeconometric relationships estimated in the 60s and the 70s, before econometricians were aware of this problem, were simply fully wrong.

CHAPTER 5. EXOGENEITY

The main difficulty with applied econometrics is that notions, which are very advanced from a theoretical point of view, can be extremely useful for the most elementary applications. However, these notions are taught in advanced level courses of econometrics, are tricky (even when they are not truly difficult) and sometimes are not fully understood, even by theoreticians. In the previous chapter I gave the example of spurious regressions. When this problem was discovered in the seventies, many of the well-established results in applied macroeconometics were found to be spurious, so without any meaning.

The correct understanding of the concept of exogeneity is still more recent: the seminal paper dates back of 1983. However, there are still applied works in econometrics that are based on a misunderstanding of this concept, or of theoretical developments connected to it like *Sim's critique*. So, these works are simply totally wrong.

Actually there are three concepts of exogeneity: *weak exogeneity*, which is useful for estimation, *strong exogeneity*, which is useful for forecasting (in a time series context), *super exogeneity*, which is useful for the framing of economic policies and is connected to the *Lucas' critique.* I will limit myself in this course to the first concept.

An example

The model represents a closed economy. It includes two equations. The first one is a traditional Phillips curve:

(1) $y_t = \alpha_0 + \alpha_1 \pi_t + \alpha_2 \pi_{t-1} + v_t$

y_t is the GDP and π_t is the inflation rate, both computed on period starting at time t and ending at time $t+1$. α_0, α_1 and α_2 are parameters. The level of activity is positively related to inflation. Thus, we assume: $\alpha_1 + \alpha_2 > 0$, even if economists have given good reasons for α_2 being negative.

The second equation represents the policy of the Central Bank, which sets inflation according to the monetary rule:

(2) $\pi_t = \bar{\pi} + \rho_1(\pi_{t-1} - \bar{\pi}) + \rho_2(y_{t-1} - \bar{y}) + \eta_t$

$\bar{\pi}$ is the inflation target rate chosen by the Central Bank, $\bar{y}$ is potential output[13], ρ_1 and ρ_2 are parameters (the first included between 0 and 1 and representing the adjustment speed of the Central Bank to its target).

v_t and η_t are the error terms of the equations. Each error term is assumed to be identically distributed, independent of information available at time $t-1$ (which includes past values of GDP and inflation), with respective variance σ_v^2 and σ_η^2. Moreover, I will assume that the two error terms are normally distributed. Thus independence and non correlation will be equivalent. I am interested by estimating

[13] To simplify the example I will assume that potential output is constant and non observable.

the Phillips curve, that is the parameters α_0, α_1, α_2 and σ_ν, which are called the *parameters of interest.*

In period t, the model determines the current values of GDP and of the inflation rate as functions of their past values and of the current values of the error terms. I will look at the probability distribution of variables y_t and π_t conditionally on their past values. I will give asymmetric roles to these two variables. I will start by the inflation rate, which is determined by the monetary reaction function (2). Its distribution, conditional on past values of both variables is normally distributed, with an expected value of $\bar{\pi} + \rho_1(\pi_{t-1} - \vec{\pi}) + \rho_2(y_{t-1} - \bar{y})$ and a variance of σ_η^2.

I will turn to the probability density distribution of y_t, conditionally on the past values of both variables and the current value of the inflation rate π_t. In general, the error term of the Phillips curve ν_t is correlated to the error term of the monetary reaction function η_t. I can represent this relation by the equation:

(3) $\nu_t = \gamma\eta_t + \nu_t^{'}$,

The error term $\nu_t^{'}$ of this equation has an expected value of zero, is uncorrelated with η_t and has a variance equal to: $\sigma_\nu^2 - \gamma^2\sigma_\eta^2$. If I substitute equation (3) in equation (1) I will get:

$$(4)\quad \begin{array}{l} y_t = \alpha_0 + \alpha_1\pi_t + \alpha_2\pi_{t-1} + \gamma\eta_t + \nu_t^{'} = \\ \alpha_0 + \alpha_1\pi_t + \alpha_2\pi_{t-1} + \gamma[(\pi_t - \bar{\pi} - \rho_1(\pi_{t-1} - \bar{\pi}) - \rho_2(y_{t-1} - \bar{y})] + \nu_t^{'} \end{array}$$

Thus, the conditional probability distribution of y_t is normally distributed, with expected value: $\alpha_0 + \alpha_1\pi_t + \alpha_2\pi_{t-1} + \gamma[(\pi_t - \bar{\pi} - \rho_1(\pi_{t-1} - \bar{\pi}) - \rho_2(y_{t-1} - \bar{y})]$, and variance: $\sigma_\nu^2 - \gamma^2\sigma_\eta^2$. I can estimate equation (4), by an OLS regression of y_t on a constant term and on variables π_t and π_{t-1}. However, I will not get the estimates of the parameters of interest of the Phillips curve α_0, α_1, α_2 and σ_ν^2, but instead of parameters: $\alpha_0 - \gamma(1-\rho_1)\bar{\pi} + \gamma\rho_2\bar{y}$, $\alpha_1 + \gamma$, $\alpha_2 - \gamma\rho_1$ and $\sigma_\nu^2 - \gamma^2\sigma_\eta^2$. Thus, I will get a mix of estimates of the parameters of the Phillips curve and of the monetary reaction function.

The concept of weak exogeneity

Now, I will turn to a general approach. I define a model by the probability density of a set of variables conditional on the past values of these variables. This model depends on a set of parameters. In the example, the model is defined by equations (1) and (2). The set of parameters is denoted by vector: $\theta = (\alpha_0, \alpha_1, \alpha_2, \sigma_\nu, \gamma, \sigma_\eta, \rho_1, \rho_2, \bar{\pi}, \bar{y})$, with: $\theta \in \Theta$. I chose among the parameters of the model the parameters of interests. In the example, they are the parameters of the Phillips curve and we denote them by vector: $\psi = (\alpha_0, \alpha_1, \alpha_2, \sigma_\nu)$.

I split the set of variables between a set of explanatory variables (the inflation rate here) and a set of explained variables (the GDP here). The probability density of the total set of variables is the product of the density of the explained variables conditional on the explanatory variables (here the conditional density of GDP) and of the marginal density of the explanatory variables (here the marginal density of the inflation rate). I call ϕ_1 and ϕ_2 the parameters, which respectively appear in the conditional and the marginal densities. Of course, these parameters can be defined in many different ways by a simple reparametrisation. I denote: $\phi = (\phi_1, \phi_2)$ the vector of both sets of new parameters, with: $\phi \in \Phi$. In a parsimonious parametrisation of both densities, the dimension of vector ϕ_i, with: $i = 1,2$, will not be larger than the dimension of vector θ.

I will say that the explanatory variables are *weakly exogenous* for the parameters of interest if there exists a parametrisation of the conditional and the marginal probability densities such that ϕ_1 and ϕ_2 are variation free. That means that the sets they must belong to, Φ_1 and Φ_2, are independent of each other, that is: $\Phi = \Phi 1 \times \Phi_2$[14].

The parameters of interest depend on the parameters, which appear in the conditional sub model for the explained variables, but not on the parameters, which appear in the sub model for the explanatory variables: $\psi = g(\phi_1)$ alone.

In this case, I can estimate the parameters of interest by only using the conditional sub model of the endogenous variables, without knowing the marginal sub model of the explanatory variables and without loss of information[15].

In the example, the inflation rate is weakly exogenous for the parameters of the Phillips curve (1) only if: $\gamma = 0$, which is the absence of correlation between the error terms of the Phillips curve (the equation we want to estimate) and of the monetary reaction function.

Comments

In the example I have introduced a complete structural economics model of the economy. Then I have isolated an equation and I have expressed my wish to estimate only the parameters of this equation. So, the concept of *weak exogeneity* is an economic concept connected to a *limited information method of estimation* (like double least squares, instrumental variables or GMM). A purely statistical approach, like the direct estimation by OLS of equation (4) is perfectly correct at a statistical

[14] A trick is that if both ϕ_1 and ϕ_2 include a common parameter of θ, this condition will not be satisfied, even if this parameter can belong to the whole real line.

[15] This last qualification has some importance. Weak exogeneity is necessary to get an efficient estimation of the parameters of interest. However, it is not a necessary condition for consistent but inefficient estimations.

level. However, it does not answer the question of estimating the structural economic parameters of interest of the Phillips curve. In the paragraph dealing with the general problem I have only extended the previous ideas.

So, when an applied econometrician wants to estimate an equation he must be very clear if he wants to estimate a statistical equation or an economic equation. If he wants to estimate an economic equation the theoretical meaning of this equation must be perfectly clear. For instance the equation must represent the behaviour of an agent or a class of agents, or an equilibrium relationship on a well-defined market. If the econometrician does not succeed in giving a rigorous economic meaning to his equation he must accept that he is estimating a statistical relation (like a VAR model) and he must use the adequate statistical tools (without mixing them with economics).

If the equation has a clear economic meaning, some of its explanatory variables will appear as being determined simultaneously with the explained variable, as a result of the current economic equilibrium. So, the applied econometrician must close his model and write all the equations determining the values of the explanatory variables of his equation of interest. However, as the estimation method is under limited information, these supplementary equations will not be estimated and have to be written only to check the consistency of the estimation problem. What will matter very much is the list of explanatory variables which will appear in these complementary equations and, which do not appear in the equation of interest. The *exclusion* of these variables from the equation of interest is an economic assumption. We will see in the next paragraph that this exclusion is essential for the estimation of the parameters of interest and that it can be only partly tested.

We can give another example of the difference between a statistical model and a structural econometric model. We saw that the level of education of a worker has a strong effect on his wages. We also saw that the ability of the worker (is he clever or stupid) is difficult to measure. Because of that this variable does not explicitly appear in the equation, and is implicitly included in the error term. We also saw, that the ability of the worker might have a positive influence on the level of education. Thus, the error term of the wages equation could be correlated with the level of education. In this case, the estimated coefficient of education in the equation is a statistical relation, but not an economic one. For instance, if we find that one year more of schooling increases wages by 7%, we have a statistical regularity, which can be used for instance for forecasting, but not a measure of the return of education. We gave arguments explaining, that probably the correlation between ability and education was low. However, more recent research tried to estimate the economic effect of education on wages, by using instrumental variables. These variables must be correlated with the level of education, but uncorrelated with the error term of the wages equation (that is to the ability of the worker). Finding instrumental variables with this property is a difficult problem, but it is essentially an economic problem. These instrumental variables must influence the level of education, but must not influence wages in a direct way (only through their effect on the level of education). Economists took as instrumental variables family background variables, for instance parents' education. However, this choice is an economic hypothesis, which cannot be tested, and the meaning of the results of the estimation will be conditional on this assumption. All economists will not agree with this assumption. Other instruments can be imagined, such as the proximity of a college. Marno Verbeek gives a fascinating discussion of this problem (pages 137-141). The surprising result is that the return of education increases when we estimate it with instrumental variables (we expected that it would decrease).

Identification and estimation

The database Mukherjee\malta.wk1 includes yearly data on the foreign trade of Malta over the period 1963-1989. There are 8 series: Year (the year), X (the exports in US dollars), Y (an index of world demand), p (the exports price), pw (the world price), e (the exchange rate), cpi (an index of the consumption price), I (investment). The model is:

Log(Xd)=b1+b2*log(p)+b3*log(pw)+b4*log(Y) +error term
Log(Xs)=c1+c2*log(p)+c3*log(e)+c4*log(cpi)+c5*log(I)+error term
Xd=Xs=X

This model simultaneously determines the volume and the price of exports, Y and p. All the other variables are *assumed* to be weakly exogenous for the parameters of these two equations. This assumption is a bit dangerous. However, it could be criticised and tested only relatively to a more complete structural model of the economy of Malta, with equations explaining the determination of these other variables. Now, the two equations of the model have (more or less) a clear economic meaning. The first one explains that the demand for exports by foreign countries depends on the price of these exports compared to the world price and on the level of world demand. The second equation explains that the supply of these exports depends on their price (in US dollars), on the exchange rate, on the consumption price in Malta (which is an index of the labour cost) and of investment (which is an index of the imports demand by Malta that will be satisfied if exports are high enough). Personally, I find that the precise meaning of the equations (especially of the second equation) presents some ambiguity, but I will not develop my point.

The simplest method to estimate these two equations is by two stages least squares (TSLS). The logarithm of the exports price log(p) is projected on the linear variety defined by the weakly exogenous variables of the model and the constant term, which will be called *instrumental variables.* This projection $\log(\hat{p})$ will be substituted to log(p) in both equations and they will be estimated by OLS. Of course, this method will be valid only if the instrumental variables are truly weakly exogenous, and if their correlation with log(p) is high enough. Now, even if this is true, this method suffers from some difficulties.

Let us start with the simpler model:
Log(Xd)=b1+b2*log(p)+error term
Log(Xs)=c1+c2*log(p)+ error term
Xd=Xs=X
When you regress (by TSLS) the logarithm of exports on the logarithm of their price, you do not know if you estimate the demand function for exports, their supply function or a linear combination of both. So, none of these two equations is *identifiable*.

Let us continue with the model of intermediary complexity:
Log(Xd)=b1+b2*log(p)+error term

Log(Xs)=c1+c2*log(p)+c3*log(e)+c4*log(cpi)+c5*log(I)+error term
Xd=Xs=X

All linear combinations of the demand equation with the supply equation includes the same variables as the supply equation. So, when you try to estimate this last equation by TSLS, you do not know if you are estimating the supply equation or a linear combination of the supply and demand equations. Thus, the supply equation is *non identifiable* (but if you run a TSLS regression you will probably get a numerical result, only it will have no economic meaning). On the other hand all the linear combinations of the supply and the demand equations include explanatory variables besides of the price of exports. So, when you regress (by TSLS) the logarithm of exports on the logarithm of their price and a constant term, you will get an estimate of the demand for exports. This equation is *identifiable.*

You have noticed that the *identification* of the demand for exports results of the *exclusion* from this equation of explanatory variables that are present in the supply equation. This exclusion is based on economic theory and the economic meaning of the demand equation. We will see at the end of this section that the exclusion of some variables from an equation can be tested only if the number of excluded variable is high enough, higher than the number just necessary to make the equation identifiable.

Sims, wrote a devastating paper on this question in 1980, which presented what has become called the *Sims' critique*. Sims explained that economic theory can exclude explanatory variables from a decision equation only for some precise horizon (in general the very short run or the very long run). Available data (with a quarterly or yearly periodicity) do not conform to these horizons. So, the estimated equation is basically mis-specified, and the omitted variables are correlated to observed explanatory variables, which should not appear in the equation according to economic theory. Sims showed that tests of the exclusion conditions generally reject them. A more ordinary reason for this rejection is that the theoretical status of econometric equations is often fuzzy and the basis for excluding such and such an explanatory variable is often weak. When Sims published his paper, huge macroeconometric models with thousand's equations were popular. Many of these equations were purely ad hoc, with a very imprecise and unsatisfying theoretical foundation and quite arbitrary exclusion assumptions. So, the article hit a very sensitive weakness of the economics of its time.

In the original model of Maltese exports, no linear combination[16] of the equations of demand and supply includes only the variables present either in the supply or in the demand equation. So, both equations are identifiable and can be estimated by TSLS.

There is a necessary condition of identification, which is very easy to check when there do no exist constraints between parameters appearing in different equations nor between the error terms of different equations. This condition is sufficient in cases that are not too queer. Let M be the number of endogenous variables. Here it is equal to 4: Xd, Xc, X, p. Let K be the number of exogenous variables (the constant term included) of the system of equations. Here it is equal to 6: c, log(pw), log(Y), log(E), log(CPI), log(I). Let m be the number of endogenous variables in the investigated equation. It is equal to 2 in the demand equation and in the supply

[16] With non zero weights for both equations, of course.

equation (Xd and p, Xs and p respectively). Let k be the number of exogenous variables (the constant term included) present in the investigated equation. It is equal to 3 in the demand equation and to 4 in the supply equation. A necessary condition for the identification of the investigated equation is: $K-k \geq m-1$. For the supply and demand equations respectively, we get:
6-3>2-1
6-4>2-1
So, both equations are identified.

We can separate equations that are just identified (those for which the above identification condition is just satisfied) from the equations that are overidentified (those for which the identification condition is more than just satisfied). In this last case the difference between both hands of the inequation is the *degree of overidentification* (2 for the demand equation and 1 for the supply equation). An overidentified equation uses more instrumental variables than the minimum required number. This excess of instrumental variables allows for a more precise estimation[17]. Overidentification also allows some useful tests. These tests can reject a list of instrumental variables, but they do not say which variables have the responsibility of the rejection (which can also result from a wrong specification of the equation). These test do not teach anything when the equation is just identified. So, they are not substitute for more direct evaluation of the property of weak exogeneity of the instrumental variables[18].

To conclude, there exists a general necessary and sufficient condition of identification of an equation. It is more complicated than the one we have given her, but it is implemented in most of econometric software.

The explanatory endogenous variable is substituted by a linear combination of instrumental variables, which are not correlated with the error term. This procedure will give a precise estimation of the coefficient of the explanatory endogenous variables, if we do not lose too much information in this substitution that is if there is a strong correlation between the endogenous variable and the combination of the instruments. This can be measured with the F statistics of the regression of the explanatory endogenous variable relatively to the instruments. We can check if the whole set of these instruments is significant, if it is high (a value higher than10 is an usual benchmark), and we also can check on the information brought by each instrument.

How to test the weak exogeneity of a variable: the Haussman's test

Testing for the weak exogeneity can be done with the Haussman's test. I will present this test on two successive examples.

[17] However, as the size of the sample of observations is finite, the projection of log(p) on these variables could not be computed if they are too many. If this number is high, but not high enough to prevent this estimation, the lack of degree of freedom would make the estimation of log(p) very imprecise. The choice of the number of instrumental variables is a difficult question: it must be high enough but not too high.

[18] The results of overidentification tests are sensitive to the number of instruments: the power of these tests is weak for too few or too many instruments.

1. Ouvrir la base Mukherjee\indona.wk1 (Indonésie, annuel, 1968-1992, 6 variables). La deuxième variable est le PIB y, la troisième variable est la consommation C (comme C est un mot réservé pour la constante, on lit en fait ser03). On peut régresser le log de la consommation sur un constante et le log du PIB. La régression n'a pas une mauvaise allure : le coefficient de log(y) est 1.006288 avec un écart type de 0.024299. Maintenant, les économistes keynésiens disent que le niveau de la consommation et du PIB sont simultanément déterminés par le modèle d'équilibre (déséquilibre ?) auquel ils croient. Cela veut dire que dans la régression précédente le PIB n'est pas exogène, c'est-à-dire qu'il est corrélé avec le terme d'erreur. Dans ce cas on sait que les MCO donnent des estimateurs (ici de l'élasticité de la consommation au PIB) qui ne sont pas convergents. On peut remédier à ce problème en recherchant des variables instrumentales. Ces variables doivent être corrélées avec la variable explicative (le log du PIB) mais pas avec le terme d'erreur de l'équation de consommation. La tradition keynésienne conseille de prendre l'investissement I (qu'elle considère comme exogène). On peut aussi prendre l'excédent commercial (X-M). L'idée est que les exportations de matières premières sont largement indépendantes de la conjoncture indonésienne et déterminent largement les importations qui sont contraintes par les devises disponibles. Dans ce cas on trouve comme coefficient du log du PIB 1.008644, avec un écart type de 0.024885. Le résultat des MCO et celui des IV sont très voisins pour notre paramètre d'intérêt. L'idée du test d'Haussman est de comparer ces deux résultats.

2. Si le log du PIB est exogène, alors les MCO sont BLUE, et les IV sont convergents, mais moins précis. On remarque ci-dessus que l'estimateur IV a un écart type plus élevé que l'estimateur MCO. Si le log du PIB n'est pas exogène, alors les MCO ne sont pas convergents, mais l'estimateur IV l'est. Ainsi, si l'estimateur IV et l'estimateur des MCO sont très différents, le premier est correct, le second ne l'est pas et on rejette l'exogénéité du log du PIB. Si les deux estimateurs sont voisins, alors les deux sont corrects, mais celui des MCO est plus précis et on retient l'exogénéité du log du PIB. Pour faire le test on appelle u la différence entre les deux estimateurs de l'élasticité de la consommation au PIB. Ici on a : u=0,002356. Le rapport de u au carré à sa variance, noté m, suit un chi2 à 1 degré de liberté. Haussman démontre que la variance de u est égale à la différence des variances des estimateurs de l'élasticité par les IV et par les MCO. Finalement on obtient : m=0.19258795. Cette statistique est négligeable à côté du seuil à 5% du test du chi2, qui est 3.84. Aussi, on retient l'estimateur des MCO et on ne rejette pas l'exogénéité du log du PIB.

3. Le test d'Haussman s'étend naturellement au cas multivarié, quand on veut tester l'exogénéité d'une ou plusieurs variables explicatives, mais que le total de variables explicatives est strictement supérieur à 1, à une méchanceté pratique près. Dans ce cas, au lieu de diviser par la variance de l'écart des deux estimateurs, on multiplie par l'inverse de la matrice de variance-covariance des deux vecteurs d'estimateurs. Hélas, cette matrice est souvent singulière, et pour s'en sortir il faut recourir à des inverses généralisés, ce qui est embêtant quand on n'est pas très savant. Si on est astucieux, on peut cependant s'en sortir. Comme nous allons le voir.

4. Revenons au problème des exportations maltaises présenté dans la section précédente et supposons que l'on souhaite estimer la fonction de demande d'exportations. A priori la variable explicative prix des exportations est endogène, et on ne peut pas utiliser les MCO. Vérifions cela par un test d'Haussman. On commence par estimer cette équation par les MCO. Le coefficient de log(p) est –2,88 avec un t de Student de –4,44. On estime la même équation par les IV. Les instruments sont : c, log(pw), log(Y), log(E), log(CPI), log(I). Le coefficient de log(p) devient –5.07 avec un t de Student de –4,81. La grosse variation de l'estimation de ce coefficient suggère que p n'est pas exogène.

5. La mise en œuvre pratique du test d'Haussman (il n'est pas évident de démontrer que cette mise en œuvre pratique est correcte) consiste à régresser d'abord log(p) sur toutes les exogènes : c, log(pw), log(Y) log(p), log(E), log(CPI), log(I) On déduit alors le fit de log(p). Puis on réestime l'équation de demande d'exportations par les MCO en rajoutant la variable fit de log(p) dans la liste des explicatives. Le coefficient de log(p) devient négligeable, mais celui de fit est –5,05 avec un t de Student de –6,37. Le test d'Haussman consiste à tester la significativité de cette nouvelle variable par un test de Student. Elle est visiblement significative, et on rejette donc l'exogénéité du log du prix des exportations dans la fonction de demande d'exportations[19].

[19] To implement the Haussman's test on the Phillips curve, which was presented at the beginning of this chapter, we will add to the right side of equation (1) the fit of the inflation rate given by equation (2) and we will check if the coefficient of this new variable significantly differs from zero. The tutorial of Eviews gives a small program for this very practical implementation of the Haussman's test.

CONCLUSION

L'économétrie a deux aspects, complémentaires mais différents.

La théorie économétrique établit les propriétés mathématiques de méthodes d'estimation et de tests. Par exemple on établit que sous certaines hypothèses l'estimateur des MCO est BLUE.

L'économétrie appliquée utilise les résultats de l'économétrie théoriques pour établir à partir d'un ensemble de données un modèle susceptible de rendre compte de certaines caractéristiques importantes de ces données. La stratégie de construction d'un modèle économétrique a fait l'objet de débats intenses au cours de ces 50 dernières années. S'ils ont abouti à certains accords, les désaccords restent nombreux, et on peut identifier jusqu'à des clivages idéologiques sur ce sujet.

En tous les cas la modélisation économétrique sans stratégie rigoureuse, en appliquant rapidement des résultats théoriques récents et sophistiqués, aboutit à des résultats peu convaincants. Je pense que cette faiblesse est plus répandue en macroéconomie qu'en microéconomie. Peut-être parce que le développement théorique de l'économétrie des variables avec tendances stochastiques a progressé plus vite que la réflexion sur la façon dont il convenait d'utiliser ses résultats. Peut-être parce que la spécialisation entre économètres théoriciens et économètres appliqués est moins poussée en microéconomie qu'en macroéconomie[20].

Une idée majeure de ce cours est qu'en économétrie appliquée il convient de passer du temps à analyser les données. La facilité avec laquelle les logiciels récents effectuent des opérations de statistique descriptive et tracent des graphiques, rend cette exigence facile à mettre en œuvre.

Spécification, estimation et tests : l'approche de la Cowle Commission

On peut commencer par quelques définitions. Le *data generating process* (DGP) représente le vrai processus de générations des données. Un *modèle* est une classe de processus de générations de données, dépendant d'un certain nombre de paramètres. Il est possible que le DGP appartienne au modèle. Mais le contraire est également possible. L'économètre dispose d'un échantillon d'observations, et va l'utiliser pour estimer le modèle, mais aussi pour tester si le DGP appartient au modèle. Le test se fait contre un modèle plus général auquel on espère que le DGP appartient. Ce souhait n'est pas toujours réalisé.

L'approche orthodoxe de la *Cowles Commission* est très exigeante. Elle note que la théorie économétrique d'estimation et de tests suppose toujours que le modèle est donné initialement. Pour ce modèle elle propose alors d'utiliser l'échantillon d'observations pour estimer les paramètres et tester soit des propriétés de ceux-ci, soit si les hypothèses faites pour le modèle sont vérifiées. Dans le cas contraire on

[20] Ou peut-être qu'étant macroéconomiste je tends à surestimer la qualité des travaux d'un domaine que je connais moins bien.

rejette le modèle. Un nouvel essai avec un nouveau modèle pourra se fonder sur l'expérience acquise lors de l'échec précédent. Mais pour pouvoir utiliser validement la théorie économétrique il faudra effectuer la nouvelle estimation et les nouveaux tests sur un échantillon différent.

Evidemment cette exigence n'est pas réaliste. Aussi les praticiens partaient de leur modèle théorique préféré. Si des tests le rejetaient, ils en analysaient la raison. Puis ils faisaient des séries de corrections *ad hoc*, jusqu'à ce que plus aucun test ne rejette leur modèle amendé. Cette procédure est appelée du *spécifique au général*. Elle permet de ne rejeter jamais aucune théorie, à condition de passer suffisamment de temps sur les corrections *ad hoc* et à cause du caractère fini de l'information contenue dans les données. Bien sûr les tests, par exemple de Student, faits sur le modèle final, sont malhonnêtement utilisés, puisque ce modèle n'est pas indépendant des données.

Exemple : fonction de consommation française dont la réestimation tous les deux ans donne des résultats complètement différents (*data mining* et critique de Lucas).
Exemple : expliquer le PIB français de 1980 à 2000 par la situation en Thaïlande de 1960 à 1980 (exemple de *data mining*).
Le *data mining* est le défaut ultime que combattait justement la Cowles Commission.

Une première philosophie alternative : du général au spécifique (David Hendry)

Le modélisateur commence avec un modèle très général. Ce modèle englobe, comme cas particuliers, plusieurs théories concurrentes. Son caractère éclectique ne le rend pas théoriquement satisfaisant. La première chose à vérifier est qu'il est suffisamment général pour que ses hypothèses ne soient pas rejetées par les données (qu'il est congruent aux données). Par exemple il faut le plus souvent que ses résidus soient homoscédastiques, indépendants entre eux et gaussiens.

Une fois cela vérifié, l'économètre procède par une succession de simplifications. Chaque simplifications consiste le plus souvent à tester la nullité de certains paramètres. Pour chaque modèle simplifié on vérifie que les hypothèses requises (par exemple l'homoscédasticité des résidus) sont vérifiées. A la fin, toute simplification supplémentaire est rejetée par les données, et l'économètre retient le modèle simplifié au delà duquel il ne peut plus continuer.

Cette démarche est devenue une norme de nos jours, alors que la démarche du spécifique au général est très mal vue. Elle n'est pas cependant sans défaut. D'abord, deux économètres différents peuvent effectuer leurs simplifications successives dans des directions différentes et aboutir à deux modèle simplifiés différents. Les seuils de significativité des tests devraient se cumuler. Ainsi le modèle final n'est pas rejeté contre le modèle initial à un seuil bien supérieur aux 5% habituels ce qui rend cette méthode fragile. Ensuite, pour certaines questions, la théorie économique est assez précise pour permettre une définition du modèle général initial qui fasse l'unanimité. Dans d'autres cas cette théorie est trop vague et on ne sait pas trop comment choisir ce modèle général.

Une deuxième philosophie alternative : l'analyse exploratoire des données (EDA)

L'idée est d'examiner les données sous différents angles, et de cet examen, complété par une réflexion théorique, de déduire progressivement un bon modèle. D'abord un modèle très simple est ajusté sur les données. Puis les résidus de ce modèle sont examinés. Les caractéristiques de ces résidus qui contredisent les hypothèses de base des MCO (par exemple une autocorrélation ou la présence d'*outliers*) suggère alors des améliorations du modèle. On regarde notament les histogrammes, les *scatter diagrams* et les *outliers*. Evidemment, si l'échantillon d'observation est de petite taille, l'information qu'il contient risque d'avoir été entièrement utilisée pour spécifier un modèle qui en rende compte, comme dans le *data mining*. Les tests n'ont plus alors aucun sens. L'idéal serait de partager les données en deux, une partie pour l'EDA, une partie pour les tests et l'estimation. En général on ne fait pas cela et on se borne à ne pas pousser l'EDA trop loin.

Une troisième philosophie : l'analyse de fragilité ou de sensibilité (Leamer)

Supposons que nous nous intéressions à la théorie du rattrapage : selon cette théorie le taux de croissance d'un pays sur la période 1965-1995 est d'autant plus fort que son PIB par tête en 1965 était bas. Il est facile de construire un échantillon portant sur 100 pays, et de régresser sur 100 points le taux de croissance par rapport à une constante et au PIB initial, puis de tester par un t de Student si le coefficient du PIB initial est significativement négatif.

Maintenant les différences de performance de croissance entre pays sur ces 30 années ne dépendent pas que du PIB initial. Elles dépendent aussi du niveau d'éducation initial, de l'investissement en éducation, de l'expansion démographique, de la politique économique, etc. En rajoutant certaines de ces variables (on ne peut bien sûr pas toutes les mettre) on change l'estimation du coefficient du PIB initial. Ici c'est ce coefficient qui nous intéresse (on l'appelle un *paramètre d'intérêt*) alors que les valeurs des autres coefficients ne nous intéresse pas (on les appelle les *paramètres de nuisance*).

L'analyse de fragilité examine si le coefficient du PIB initial reste significativement négatif quand on change les autres variables du modèle. Elle essaie de voir aussi le champ de variation de l'estimation du paramètre d'intérêt.

Ces trois philosophies ne sont pas exclusives, et dans un problème concret, certaines s'avèrent plus adéquates que d'autres.

Conclusion

Aux recommandations précédentes on peut en ajouter d'autres plus basiques, mais importantes. D'abord l'équation estimée ne doit pas paraître étrange à un économiste, c’est-à-dire ne doit pas inclure des incohérences logiques.

La première distinction est entre variables de stock, définies à un instant précis (par exemple le capital, la richesse, les prix, le taux d'intérêt, le taux de change), et variables de flux définies sur une période (la consommation, le revenu, l'investissement). Le flux d'une période peut dépendre d'un stock de début de période, pas de fin de période. Souvent un stock est le cumul de flux passés (par exemple le capital est le cumul de l'investissement). Mais le flux le plus récemment observé est un mauvais indicateur du stock (cette évidence a été oubliée dans des articles célèbres).

Egalement si la théorie économique établit que la bonne variable est le taux d'intérêt, on ne peut pas remplacer cette variable par sa différence première, sauf à introduire une dynamique bizarre. Il ne faut pas mettre non plus des délais d'ajustement étranges, du type ma consommation dépend de mon seul revenu d'il y a trois ans : ce résultat économétrique risque de provenir de *data mining* dans un échantillon de petite taille (penser à mon exemple France-Thaïlande).