Birkbeck
Department of Economics,
Mathematics and Statistics

Graduate Certificates and Diplomas
Economics, Finance, Financial Engineering;
BSc FE, ESP.

2009-2010

Applied Statistics and Econometrics

Notes and Exercises

Ron Smith

Email R.Smith@bbk.ac.uk

CONTENTS

1. PART I: Course Information

1.1. Aims

Economists have been described as people who are good with numbers but not creative enough to be accountants. This course is designed to ensure that you are good with numbers; that you can interpret and analyse economic and financial data and develop a critical awareness of some of the pitfalls in collecting, presenting and using data. Doing applied work involves a synthesis of various elements. You must be clear about why you are doing it: what the purpose of the exercise is (e.g. forecasting, policy making, choosing a portfolio of stocks, answering a particular question or testing a hypothesis). You must understand the characteristics of the data you are using and appreciate their weaknesses. You must use theory to provide a model of the process that may have generated the data. You must know the statistical methods, which rely on probability theory, to summarise the data, e.g. in estimates. You must be able to use the software, e.g. spreadsheets, that will produce the estimates. You must be able to interpret the statistics or estimates in terms of your original purpose and the theory. Thus during this course we will be moving backwards and forwards between these elements: purpose, data, theory and statistical methods. It may seem that we are jumping about, but you must learn to do all these different things together.

Part I of this booklet provides background information: reading lists; details of assessment (70% exam, 30% project) and instructions on how to do your project. Part II provides a set of notes. These include notes on the lectures, notes on economic and financial data, and applied exercises.

Not all the material in this booklet will be covered explicitly in lectures, particularly the sections on economic and financial data. But you should be familiar with that material. Lots of the worked examples are based on old exam questions. Sections labelled **background** contain material that will not be on the exam. If you have questions about these sections raise them in lectures or classes. If you find any mistakes in this booklet please tell me. Future cohorts of students will thank you.

1.2. Rough Lecture Outline

These topics roughly correspond to a lecture each, though in practice it may run a little faster or slower.

AUTUMN

1. Introduction
2. Descriptive Statistics
3. Index Numbers
4. Probability
5. Random Variables
SPRING
1. The normal and related distributions
2. Estimation
3. Confidence Intervals and Hypothesis Tests
4. Bivariate Least Squares Regression
5. Matrix Algebra & Multiple Regression
6. Properties of Least Squares estimates
7. Tests for regressions
8. Dynamics
9. Applications
10. Revison

Tutorial Classes run through the spring term, doing the exercises in section 2.

The sections in the notes on Economic and Financial Data and Applied Exercises, will be used for examples at various points in the lectures. You should work through them, where they come in the sequence in the notes. This material will be useful for class exercises, exam questions and your project.

1.3. Learning Outcomes

Students will be able to demonstrate that they can:

- Explain how measures of economic and financial variables such as GDP, unemployment and index numbers such as the RPI and FTSE are constructed, be aware of the limitations of the data and be able to calculate derived statistics from the data, e.g. ratios, growth rates, real interest rates etc.

- Use a spreadsheet to graph data and calculate summary statistics and be able to interpret the graphs and summary statistics.

- Use simple rules of probability involving joint, marginal and conditional probabilities, expected values and variances and use probabilities from the normal distribution.

- Explain the basic principles of estimation and hypothesis testing.
- Derive the least squares estimator and show its properties.
- Interpret simple regression output and conduct tests on coefficients.
- Read and understand articles using economic and financial data at the level of the FT or Economist.
- Conduct and report on a piece of empirical research that uses simple statistical techniques.

1.4. Your input

To achieve the learning outcomes (and pass the exams) requires a lot of independent work by you. We will assume that you know how to learn and that there are things that we do not have to tell you because you can work them out or look them up for yourself. The only way to learn these techniques is by using them.

- Read these notes.
- Get familiar with economic and financial data by reading newspapers (the FT is best, but Sunday Business sections are good), The Economist, etc. In looking at articles note how they present Tables and Graphs; what data they use; how they combine the data with the analysis; how they structure the article. You will need all these skills, so learn them by careful reading.
- Ensure that you can use a spreadsheet, such as Excel.
- Try to attend all lectures and classes, if you have to miss them make sure that you know what they covered and get copies of notes from other students.
- Do the exercises for the classes in the Spring term in advance. Continuously review the material in lectures, classes and these notes, working in groups if you can.
- Identify gaps in your knowledge and take action to fill them, by asking questions of lecturers or class teachers and by searching in text books. We are available to answer questions during office hours (posted on our doors) or by email.

- Do the applied exercise (section 8 of the notes) during the first term. We will assume that you have done it and base exam questions on it.
- Start work on your project early in the second term, advice on this is in section 4.

1.5. Reading

There are a large number of good text books on introductory statistics, but none that exactly match the structure of this course. This is because we cover in one year material that is usually spread over three years of an undergraduate degree: economic and financial data in the first year, statistics in the second year, and econometrics in the third year. Use the index in the text book to find the topics covered in this course.

These notes cross-reference introductory statistics to Barrow (2009) and the econometrics and more advanced statistics to Verbeek (2008). This is one of the books that is used on the MSc in Economics econometrics course. There are a large number of other similar books, such as Gujarati and Porter (2009) and Stock and Watson (2009).

There are a range of interesting background books on probability and statistics. The history of probability can be found in Bernstein (1996), which is an entertaining read as are other general books on probablility like Gigerenzer (2002), and Taleb (2004, 2007). A classic on presenting graphs is Tufte (1983).

Where economic or financial topics appear in these notes, they are explained. But it is useful to also do some general reading. On economics there are a range of paperbacks aimed at the general reader such as Kay (2004) and Smith (2003). Similarly, there are lots of paperbacks on finance aimed at the general reader. Mandelbrot and Hudson, (2005) is excellent. Mandelbrot a mathematician who invented fractals has done fundamental work on finance since the 1960s. Although he is highly critical of a lot of modern finance theory, he gives an excellent exposition of it. Das (2006) provides an excellent non-technical introduction to derivatives, as well as a lot of funny and often obscene descriptions of what life is actually like in financial markets. Although written before the credit crunch Taleb, Mandlrot and Das all pointed to the danger of such events.

References

Barrow, Michael,2009) *Statistics for Economics Accounting and Business Studies,* 5th edition, FT-Prentice Hall.

Bernstein, Peter L. (1996) *Against the Gods, the Remarkable Story of Risk*, Wiley.

Das, Satyajit (2006) *Traders Guns and Money,* Pearson

Gigerenzer, Gerd (2002) *Reckoning with Risk*, Penguin.

Gujarati D.N. and D.C. Porter, (2009) *Basic Econometrics*, 5th edition. McGraw Hill

Kay, John (2004) *The Truth about Markets*, Penguin

Mandelbrot, Benoit and Richard Hudson, (2005) *The (Mis) Behaviour of Markets* Profile Books

Smith, David (2003) *Free Lunch*, Profile Books

Stock, J.H. and M.W. Watson (2007) *Introduction to Econometrics*, 2nd edition, Pearson-Addison Wesley.

Taleb, Nassim Nicholas (2004) *Fooled by Randomness: the hidden role of chance in life and in the markets*, 2nd edition, Thomson

Taleb, Nassim Nicholas (2007) *The Black Swan: The impact of the highly improbable*, Penguin.

Tufte, Edward R (1983) *The Visual Display of Quantitative Information*, Graphics Press

Verbeek, Marno (2008) *A guide to modern econometrics*, 3rd edition, Wiley.

2. Class exercises Spring term (Many are past exam questions).

2.1. Week 1 Descriptive Statistics

(1) In a speech, *Why Banks failed the stress test*, February 2009, Andrew Haldane of the Bank of England provides the following summary statistics for the "golden era" 1998-2007 and for a long period. Growth is annual percent GDP growth, inflation is annual percent change in the RPI and for both the long period is 1857-2007. FTSE is the monthly percent change in the all share index and the long period is 1693-2007.

	Growth		Inflation		FTSE	
	98-07	long	98-07	long	98-07	long
Mean	2.9	2.0	2.8	3.1	0.2	0.2
SD	0.6	2.7	0.9	5.9	4.1	4.1
Skew	0.2	-0.8	0.0	1.2	-0.8	2.6
Kurtosis	-0.8	2.2	-0.3	3.0	3.8	62.3

(a) Explain how the mean; standard deviation, SD; coefficient of skewness and coefficient of kurtosis are calculated.

(b) What values for the coefficients of skewness and kurtosis would you expect from a normal distribution. Which of the series shows the least evidence of normality.

(c) Haldane says "these distributions suggest that the Golden Era" distributions have a much smaller variance and slimmer tails" and "many risk management models developed within the private sector during the golden decade were, in effect, pre-programmed to induce disaster miopia.". Explain what he means.

(2) The final grade that you get on this course (fail, pass, merit, distinction) is a summary statistic. 40-59 is a pass, 60-69 is a merit, 70 and over is a distinction. In the Grad Dips grade is based on marks (some of which are averages) in 5 elements. Merit or better is the criteria for entering the MSc.

Final overall grades are awarded as follows:

Distinction: Pass (or better) in all elements, with Distinction marks in three elements and a Merit (or better) mark in a fourth.

Merit: Pass (or better) in all elements, with Merit marks (or better) in four elements.

Pass: In order to obtain a Pass grade, a student should take all examinations and obtain Pass marks (or better) in at least four elements.

Notice that the grade is not based on averages. This is like the driving test. If you are good on average, excellent on steering and acceleration, terrible on braking, you fail; at least in the UK.

Consider the following four candidates.

	Mac	Mic	QT	AES	Opt
a	80	80	30	80	80
b	80	80	40	80	80
c	60	60	40	60	60
d	80	80	80	30	30

(a) What final grade would each get?

(b) How would rules grades based on mean, median or mode differ.

(c) What explanation do you think there is for the rules? Do you think that they are sensible?

2.2. Week 2 Probability

(1) Show that the variance equals the mean of the squares minus the square of the mean:

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

where $\overline{x}=\sum x_i/N$.

(2) Suppose you toss a fair coin three times in a row. What is the probability of:

(a) three heads in a row;

(b) a tail followed by two heads.

(c) at least one tail in the three throws.

Hint write out the 8 (2^3) possible outcomes and count how many are involved in each case.

(3) Students take two exams A and B. 60% pass A, 80% pass B, 50% pass both.

(a) Fill in the remaining five elements of the joint and marginal distributions below, where PA indicates pass A, FB fail B, etc.

(b) What is the probability of a student passing B given that they passed A?

(c) Are the two events passing A and passing B (i) mutually exclusive (ii) independent?

	PA	FA	B
PB	50		80
FB			
A	60		100

2.3. Week 3 Probability Continued

(1) Consider the following game. A fair coin is tossed until it comes up heads and you get paid £(2^n) if it comes up heads on the n-th throw. If it comes up heads the first time you get £2 and the game stops. If it comes up heads, for the first time on the second throw you get £4=$(2)^2$ and the game stops; and so on. What is the expected value of this game? How much would you personally pay to play it?

(2) Define $P(A)$ as the probability of event A happening; $P(B)$ the probability of event B happening; $P(A \cap B)$ the probability of both A and B happening; $P(A \cup B)$ the probability of either A or B happening; and $P(A \mid B)$ the probability of A happening conditional on B already having happened.

(a) What is $P(A \cap B)$ if A and B are mutually exclusive.

(b) What is $P(A \cap B)$ if A and B are independent?

(c) What is $P(A \cup B)$?

(d) Show that

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

(3) You are in a US quiz show. The host shows you three closed boxes in one of which there is a prize. The host knows which box the prize is in, you do not. You choose a box. The host then opens another box, not the one you chose, and shows that it is empty. He can always do this. You can either stick with the box you originally chose or change to the other unopened box. What should you do: stick or change? What is the probability that the prize is in the other unopened box?

(4) **(Optional)**. Calculate the probability that two people in a group of size N will have the same birthday. What size group do you need for there to be a 50% chance that two people will have the same birthday? Ignore leap years.

Use a spreadsheet for this and work it out in terms of the probability of not having the same birthday. In the first row we are going to put values for N (the number of people in the group), in the second row we are going to put the probability that no two people in a group of that size have the same birthday.

In A1 put 1, in B1 put =A1+1, copy this to the right to Z1.

In A2 put 1. Now in B2 we need to calculate the probability that two people will NOT share the same birthday. There are 364 possible days, i.e. any day but the first person's birthday, so the probability is 364/365. So put in B2 =A2*(365-A1)/365. Copy this right. Go to C2, the formula will give you $1 \times (364/365) \times (363/365)$. The third person, has to have birthdays that are different from the first and the second. Follow along until the probability of no two people having the same birthday falls below a half.

2.4. Week 4 Index numbers etc.

(1) UK GDP in current market prices in 1995 was £712,548m, while in 1997 it was £801,972m. GDP at constant 1995 market prices in 1997 was £756,144m.

(a) Construct index numbers, 1995=100 for: current price GDP; constant price GDP; and the GDP deflator in 1997.

(b) From these numbers calculate the average annual rate of inflation between 1995 and 1997.

(c) From these numbers calculate the average annual rate of growth between 1995 and 1997.

(d) If the interest rate on two year bonds in 1995 was 10% per annum what would the real per annum interest rate over this period be.

(e) Explain the difference between Gross Domestic Product and Gross National Product.

(f) Explain what Gross Domestic Product measures. What limitations does it have as a measure of the economic wellbeing of a nation.

(2) The Department buys bottles of red wine, white wine and orange juice for its parties. The table below gives prices per bottle and number of bottles for three years. Construct:

(a) an expenditure index using 1995=100;

(b) a party price index (i) using 1995 as a base (Laspeyres), (ii) using 1997 as a base (Paasche);

(c) a quantity index using 1995 as a base.

	1995		1996		1997	
	p	q	p	q	p	q
Red	3	20	4	15	5	10
White	4	20	4	25	4	30
Orange	1	10	2	10	3	10

.

2.5. Week 5 Properties of estimators and distributions.

(1) Suppose you have a sample of data, Y_i, $i = 1, 2, .., N$, where $Y \sim IN(\mu, \sigma^2)$.

(a) Explain what $Y \sim IN(\mu, \sigma^2)$ means.

(b)How would you obtain unbiased estimates of μ and σ^2? Explain what unbiased means.

(c)How would you estimate the standard error of your estimate of μ?

(d) Suppose that the distribution of your sample was not normal but highly skewed. Explain what this means and discuss what other measures of central tendency that you might use.

(2)Marks on an exam are normally distributed with expected value 50 and standard deviation 10.

(a) What proportion of the students get

(i) <30;

(ii) between 30 and 50;

(iii) over 50.

(b) What mark does a student need to get into the top 15%.

(c) In a class of 16, what is the probability that the average mark is greater than 53?

2.6. Week 6 Hypothesis testing and Regression

(1) For a sample of data $Y_1, Y_2, ..., Y_N$ on a random variable $Y \sim IN(\mu, \sigma^2)$.

(a)You want to test the hypothesis that μ equals a specified value μ_o. How would you test this hypothesis?

(b) Explain Type I and Type II errors. How did you deal with these two types of error in your answer to (a)?

(c) Suppose that you wanted to calculate the probability of observing a value of Y greater than a specified value Y_0. How would you do this?

(2) Consider the following bivariate regression model:

$$Y_i = \alpha + \beta X_i + u_i$$

estimated on a sample of data $i = 1, 2, ..., N$, where Y_i is an observed dependent variable, X_i is an observed exogenous regressor, u_i is an unobserved disturbance, and α and β are unknown parameters.

(a) Derive the least squares estimators for α and β.

(b) Under what assumptions about u_i will these least squares estimators be Best Linear Unbiased.

(c) Explain what Best Linear Unbiased means.

(d) Explain what exogenous means.

2.7. Week 7, Regression

It is believed that an energy demand equation takes the form:

$$q_t = \alpha + \beta y_t + \gamma p_t + \varepsilon_t,$$

where q_t is the logarithm of per capita energy demand in year t; p_t the logarithm of real energy prices; y_t the logarithm of per-capita real GDP; ε_t is a well behaved disturbance term. The following estimates (with standard errors in parentheses) were obtained using data for the period t =1974-1990.

	β	γ	R^2	SER
$India$	1.006	−0.068	0.38	0.027
	(0.102)	(0.080)		
$Indonesia$	1.564	−0.488	0.52	0.034
	(0.234)	(0.195)		
$Korea$	1.074	−0.136	0.54	0.031
	(0.125)	(0.189)		

SER is the standard error of the regression.

(a) How would you interpret β and γ?

(b) Explain what R^2 and SER are and what they tell you. How would you interpret the fact that while Korea has the largest R^2 it does not have the lowest SER?

(c) For Indonesia, test (i) the hypothesis $\beta = 1$ and (ii) the hypothesis $\gamma = 0$.

(d) Interpret the stochastic component of the model. How would you estimate it?

(e) Suppose that you believed that there was a trend increase in energy efficiency in these countries. How would you adjust the model to allow for this.

2.8. Week 8, Regression

Consider the linear regression model

$$y = X\beta + u$$

where y is a $T\times 1$ vector of observations on a dependent variable, X a full rank $T\times k$ matrix of observations on a set exogenous variables, β a $k \times 1$ vector of unknown coefficients, and u an unobserved disturbance with $E(u) = 0$ and $E(uu') = \sigma^2 I$.

(a) Derive the least squares estimator $\widehat{\beta}$.

(b) Derive the variance covariance matrix of $\widehat{\beta}$.

(c) Show that $\widehat{\beta}$ is unbiased.

2.9. Week 9, Regression

The original Phillips curve made the percentage rate of growth of money wages, Δw_t a function of the percentage rate of unemployment u_t. Subsequently, Friedman and others argued that it should be the rate of growth of expected real wages $\Delta w_t - \Delta p_t^e$ that was a function of the unemployment rate, where Δp_t^e is the expected rate of inflation. Suppose expected inflation equals actual inflation in the previous year, $\Delta p_t^e = \Delta p_{t-1}$ and that the rate of growth of wages is a reciprocal function of unempoyment. Then, both the Phillips and Friedman models can be regarded as special cases of:

$$\Delta w_t = \alpha + \beta(1/u_t) + \gamma \Delta p_{t-1} + \varepsilon_t,$$

where ε_t is assumed to be a well behaved error term. This equation was estimated using UK data for t =1967 to 1987. The results were:

$$\begin{array}{lcccccc} \Delta w_t & = 3.86 & + & 7.24 & (1/u_t) & + & 0.65 \quad \Delta p_{t-1} \\ & (2.17) & & (5.20) & & & (0.15) \end{array}$$

$R^2 = 0.534, SER = 3.20$. Standard errors are given in parentheses, SER is the standard error of regression.

(a) What is meant by a well behaved error term.

(b) What predictions for γ are implied by (i) the Phillips 'money wage' theory and (ii) the Friedman 'real wage' theory?

(c) Test the hypothesis that $\beta = 0$.

(d) Test the hypotheses (i) that $\gamma = 0$; (ii) that $\gamma = 1$.

(e) What would you conclude about the importance of unemployment and inflation expectations in this period.

2.10. Week 10, Regression.

Using US data on on company earnings, E_t, and the dividends paid out to shareholders, D_t, t =1872-1986 the following results were obtained (standard errors in parentheses):

$$\begin{array}{rccc} D_t = & 0.011+ & 0.088E_t+ & 0.863D_{t-1} \quad +\widehat{u}_{1t} \\ & (0.009) & (0.008) & (0.019) \end{array}$$

$R^2 = 0.998$, $SER = 0.074$.

$$\begin{array}{rccc} \ln D_t = & -0.136+ & 0.312 \ln E_t+ & 0.656 \ln D_{t-1} \quad +\widehat{u}_{2t} \\ & (0.015) & (0.025) & (0.029) \end{array}$$

$R^2 = 0.993$, $SER = 0.085$.

SER is the standard error of regression.

(a) Test whether the intercepts in each equation are significantly different from zero at the 5% level and interpret them. Do they have sensible values?

(b) It is suggested that the linear equation is a better equation than the logarithmic because it has a higher R^2. Do you agree with this?

Interpret the role of the lagged dependent variable and calculate the long-run effect of earnings on dividends in each case.

(c) A test for second order serial correlation had a p value of 0.008 in the linear model and 0.161 in the logarithmic model. Explain what second order serial correlation is and why it is a problem. Is it a problem in either of these models?

Extra questions

1. From observations taken over many years it is found that marks on a particular exam are normally distributed with an expected value of 50 and a variance of 100. For a standard normal distribution $\Pr(Z < z)$ =0.6915 for z=0.5; 0.8413 for z=1; 0.9332 for z=1.5; 0.9772 for z=2.

(a) What is the probability of a student getting below 40 marks on this exam?

(b) What is the probability of a student getting below 30 marks on this exam?

(c) Suppose that in a class of 16 students a new teaching method was used and the average mark in this class was 54. Is this statistically significant evidence, at the 5% level, that the new method is more effective? Suppose that the average of 54 had been obtained in a class of 36 students, would this have been statistically significant evidence? Assume that the new teaching method did not change the variance.

(d) Show that the arithmetic mean is an unbiased estimator of the expected value.

(e) Give an example of the type of distribution where the arithmetic mean would not be a good measure of the typical value of a random variable.

2. The following data is taken from Economic Trends Annual Supplement 1999.

	Y	C	RC	TBY
1995	494,574	454,171	454,171	6.31
1996	521,281	485,418	470,622	6.26
1997	554,641	517,032	488,936	7.13
1998	565,935	545,124	505,367	5.63

Y is gross households disposable income at current prices, C is households final consumption expenditure at current prices, RC is consumption expenditure at 1995 prices, TBY is the Treasury Bill Yield in percent.

(a) Calculate a price index for consumption 1995 to 1998, 1995=100.

(b) Calculate the rate of inflation for 1996, 1997 and 1998.

(c) Calculate the ex post real interest rate for 1995, 1996 and 1997.

(d) Calculate the savings rate for 1995 to 1997.

(e) Does there seem to be any relationship between the savings rate and the real interest rate? What relationship would you expect?

3. Data are available for the quantity of a good consumed Q_t, real income Y_t, the price of the good P_t, and the average of all other prices P_t^* for years $t = 1, 2, ..., T$. The demand function is assumed to take the form

$$Q_t = AY_t^{\alpha} P_t^{\beta_1} P^{*\beta_2} e^{u_t}$$

where u_t is a random error term.

(a) How would you estimate the parameters by least squares?

(b) How would you interpret the parameters and what signs would you expect for them?

(c) How would you test the hypothesis $\beta_1 + \beta_2 = 0$?

(d) How would you interpret this hypothesis?

2.11. Answers to selected exercises

2.11.1. Week 2, question 1.

Note

$$(x_i - \overline{x})^2 = x_i^2 + \overline{x}^2 + 2\overline{x}x_i$$

and that $\overline{x}$ is a constant, does not vary with i, so $\sum \overline{x}^2 = N\overline{x}^2$

$$\begin{aligned} N^{-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 &= N^{-1}\sum x_i^2 + N^{-1}N\overline{x}^2 - N^{-1}2\overline{x}\sum x_i \\ &= N^{-1}\sum x_i^2 + \overline{x}^2 - 2\overline{x}^2 \\ &= N^{-1}\sum x_i^2 - \overline{x}^2 \end{aligned}$$

2.11.2. Week 3, question 1.

This is the St Petersburg paradox. The expected value is infinite, but few would pay an infinite amount of money to play it. The usual explanation is in terms of the diminishing marginal utility of money, which makes the expected utility of the game less than infinity..

2.11.3. Week 3, question 2.

Suppose you chose A to start with. Consider the two strategies, stick S, or change C. If the prize is in A, the host can open either box B or C and show it is empty. You win with S, sticking with box A and lose with C, changing to the box the host left unopened. If the prize is in B, the host has to open box C. You lose with S, win with C, because you have to change to box B, box C is open. If the prize is in C, the host has to open box B. You lose with S, win with C because you change to box C, box B is open. Changing is the optimal strategy since you win 2 times out of three and the probability that the prize is in the other box is 2/3.

It can also be shown by Bayes theorem. Let W_A be the event that the prize is in box A, etc. Let H_A be the event that the host opens box A, etc. Suppose you choose box A. The probability that you win the prize if you switch is the probability that the prize is in B and the host opened C plus the probability that

the prize is in C and the host opened B

$$\begin{aligned} & P(W_B \cap H_C) + P(W_C \cap H_B) \\ = \; & P(W_B)P(H_C \mid W_B) + P(W_C)P(H_B \mid W_C) \\ = \; & \frac{1}{3} \times 1 + \frac{1}{3} \times 1 = \frac{2}{3} \end{aligned}$$

The second line folows from the definition of conditional probabilities. This formula might seem complicated, but it appears in a very popular work of teenage fiction: *The curious incident of the dog in the night-time*, by Mark Haddon, Random House, 2003.

2.11.4. Week 7.

(a) β is the income elasticity of demand for energy, γ is the price elasticity.

(b) $R^2 = 1 - \sum \widehat{\varepsilon}_t^2 / \sum (q_t - \overline{q})^2$; $SER = \sqrt{\sum \widehat{\varepsilon}_t^2 / T - k}$ where $k = 3$ here. R^2 measures the fit relative to the variance of the dependent variable, the SER just measures the fit. The rankings would only necessarily be the same if all the dependent variables had the same variance.

(e) Include a time trend

$$q_t = \alpha + \beta y_t + \gamma p_t + \delta t + \varepsilon_t,$$

2.11.5. Week 10

(a) Intercept of the linear equation has t ratio 0.011/0.009=1.22 not significantly different from zero. Intercept in log equation has t ratio -0.136/0.015=-9.06 significantly different from zero. The intercepts measure different things in the two equations. See (c) below.

(b) No, the R^2 cannot be compared because the dependent variables are different.

(c) The lagged dependent variable captures the effect that firms smooth dividends and only adjust them slowly in response to earnings changes. The long run relations are

$$\begin{aligned} D &= \frac{0.011}{1 - 0.863} + \frac{0.088}{1 - 0.863} E \\ &= 0.080 + 0.642E \end{aligned}$$

$$\begin{aligned}\ln D &= \frac{-0.136}{1-0.656} + \frac{0.312}{1-0.656}\ln E \\ &= -0.395 + 0.91 \ln E\end{aligned}$$

In the case of the linear model the long-run intercept should be zero, dividends should be zero when earnings are zero, in the logarithmic case it is a constant of proportionality $\exp(-0.395) = 0.67$ so the long-run is

$$D = 0.67E^{0.91}.$$

(d) Second order serial correlation is a relation between the residuals of the form

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + e_t$$

It is a problem because it indicates the model is likely to be misspecified. The linear model p value indicates that there is serial correlation p<0.05, the logarithmic model p value indicates that there is probably not serial correlation, p>0.05, at the 5% level.

2.11.6. Answers to Extra questions

1.a (40-50)/10=-1, P(Z<-1)=P(Z>1)=1-P(Z<1)=1-0.8413=0.1587

1.b (30-50)/10=-2, P(Z<-2)=1-P(Z<2)=1-0.9772=0.0228

1.c Test statistic for N=16, $\sqrt{N} = 4$

$$\frac{54-50}{10/4} = 1.6$$

This is not significantly different from zero at the 5% level. For N=36

$$\frac{54-50}{10/6} = 3.6$$

This is significantly different from zero at the 5% level.

1d Suppose $E(X) = \mu$ then $X_i = \mu + u_i$ with $E(u_i) = 0$

$$\overline{X} = N^{-1}\sum_{i=1}^{N} X_i = N^{-1}\sum_{i=1}^{N}(\mu + u_i) = \mu + N^{-1}\sum_{i=1}^{N} u_i$$

so

$$E(\overline{X}) = \mu + E(N^{-1}\sum_{i=1}^{N} u_i) = \mu.$$

1e In any very skewed distribution, such as income, the average can be very different from the typical, so the mode<median<mean.

2.a The price index for consumption is $P_t = 100 \times C_t/RC_t$, Inflation is $I_t = 100 \times (P_t - P_{t-1})/P_{t-1}$, the ex post real interest rate is $RIR = TBR_t - I_{t+1}$ the Savings Rate is $SR = 100(1 - C/Y)$. Thus

$Year$	P_t	I_t	RIR	SR
1995	100		3.21	8.16
1996	103.1	3.1	3.76	6.88
1997	105.7	2.5	5.03	6.78
1998	107.9	2.1		

The real interest rate is rising, while the savings rate is falling, the opposite of what one might expect.

3. (a) First take logarithms then estimate by LS.

(b) The income elasticity of demand $\alpha > 0$ (for a normal good); the own price elasticity $\beta_1 < 0$ (not a Giffen good); the cross price elasticity with all other goods $\beta_2 > 0$ (it cannot be a complement with all other goods).

(c) Reparameterise the estimated equation as

$$\ln Q_t = a + \alpha \ln Y_t + \beta_1(linP_t - \ln P_t^*) + (\beta_2 - \beta_1) \ln P_t^* + u_t$$

and conduct a t test on the hypothesis that the coefficient of $\ln P_t^*$ is zero.

(d) Only relative prices matter.

3. Assessment and specimen exam.

3.1. Assessment

Assessment is 70% on the exam, 30% on the empirical project submitted in mid May. You should read the advice on doing the project fairly early in the course to get the general idea of what we are looking for and then refer back to it regularly as you do your project.

The exam will have six questions, the first three questions are in Section A the last three questions are in Section B and are more applied. You must do three questions: at least one from each section and one other. The questions will be about:

1. Least Squares, e.g. deriving the least squares estimator and its variance covariance matrix, proving that it is unbiased Estimator, etc. This will be the only question that requires matrix algebra.

2. Interpretation of statistics associated with regression, e.g. R^2, standard error of regression, diagnostic tests like the Durbin-Watson statistics and the effects of the failures of the various assumptions.

3. Hypothesis testing, e.g. explain the basis of tests applied either to means or regression coefficients.

4. Economic and financial data, this will involve calculations, e.g. of index numbers, growth rates, ratios, derived measures and some interpretation.

5. Interpretation of regression results given to you in the question.

6. Probability and distributions, e.g. being able to use the basic rules of probability; given the mean and variance for a normal distribution, calculate the probability of various events happening, etc.

Before 2004 there were separate exams for Applied Finance and Statistics and Applied Economics and Statistics. Examples of question 1 can be found on AFS, though not AES papers; examples of question 4 can be found on AES but not AFS papers. Examples of the other questions appear in both.

The 2004 exam, with answers is given below.

3.2. Specimen Exam (2004)

Answer **THREE** Questions, at least one from each section and one other. All questions are weighted equally.

SECTION A

1. Consider the model:

$$y_t = \beta_1 + \beta_2 x_{2t} + u_t$$

$t = 1, 2, ..., T.$ where y_t are observations on a dependent variable and x_{2t} observations on a non-stochastic independent variable, u_t is an unobserved disturbance, with $E(u_t) = 0$; β_1 and β_2 are unknown parameters.

(a) Set out the system of T equations in matrix form $y = X\beta + u$ where y and u are $T \times 1$ vectors, X is a $T \times 2$ matrix, and β is a 2×1 vector.

(b) How would you express the sum of squared residuals $\sum_{t=1}^{T} u_t^2$ in matrix form. Show that this is a function of β.

(c) Show that the Least squares estimator of β is given by

$$\widehat{\beta} = (X'X)^{-1}X'y.$$

(d) Show that $\widehat{\beta}$ is unbiased, i.e. $E(\widehat{\beta}) = \beta$.

2. US data 1971-1999 were used to estimate a relationship between the rate of interest R_t and the rate of inflation π_t,

$$R_t = \alpha + \beta\pi_t + u_t.$$

The least squares estimates (with standard errors in parenthesese) were

$$\begin{array}{cccc} R_t = & 6.37 & +0.33\pi_t & +\widehat{u}_t \\ & (0.66) & (0.10) & \end{array}.$$

The standard error of regression, $s = 2.75$, the coefficient of determination $R^2 = 0.29$, and the Durbin Watson Statistic $DW = 0.63$.

(a) Explain what $\widehat{u}_t$ measures.

(b) Explain what R^2, Standard Error of the Regression and Durbin Watson statistic measure and what they tell you about this regression.

(c) Interpret the estimates of α and β. What do they tell you about the relationship between inflation and interest rates?

(d) What assumptions are required for least squares to give good estimates.

(3) Using information in question 2, and assuming that the 95% critical value for a t test is ± 2 :

(a) Test the hypotheses $\alpha = 0$ and $\beta = 1$ at the 5% level.

(b) Explain why the hypothesis $\beta = 1$ might be interesting to test.

(c) Explain what Type I and Type II errors are. What is the probability of Type I error in your test in part (a).

(d) Give a 95% confidence interval for $\widehat{\beta}$. Explain what a confidence interval is.

SECTION B

4. The following data were taken from Economic Trends, February 2004.

	NDY	RDY	RPI	CPI	HP	TBY
2000	654,649	654,649	170.3	105.6	87.7	5.69
2001	700,538	685,263	173.3	106.9	95.1	3.87
2002	721,044	696,224	176.2	108.3	111.2	3.92

NDY is nominal household disposable income in current prices; RDY is real household disposable income in constant 2000 prices; RPI is the retail price index, 1987=100; CPI is the consumer price index, 1996=100; HP is a house-price index; TBY is the Treasury Bill Yield a short term interest rate expressed as percent per annum.

(a) Calculate the price index (implicit deflator) for disposable income.

(b) Calculate the rate of inflation for 2001-2 for disposable income, RPI, CPI, and HP. Comment on the relation between them.

(c) Suppose that you owned a house, whose price increased at the average rate, and had a 100% mortgage paying the Treasury Bill Rate on the mortgage, what would be the real return on owning the house over 2001-2.

(d) The RPI includes housing prices in its measure of inflation, the CPI does not. Should house prices be included in inflation measures?

5. You have data on average earnings by age and education in the US in 1991 for 30 groups of men, $i = 1, 2, ..., N$. Define w_i the logarithm of earnings, A_i age in years (divided by 100), E_i years (divided by 100) of education. It is suggested that an appropriate model is:

$$w_i = \alpha + \beta_1 A_i + \beta_2 E_i + \gamma_1 A_i^2 + \gamma_2 A_i E_i + u_i.$$

The estimated coefficients, with their standard errors in parentheses are:

$$\begin{array}{rcccccc} w_i & = & 7.46 & +6.6A_i & +9.0E_i & -6.7A_i^2 & +4.1A_iE_i \\ & & (0.19) & (0.62) & (1.01) & (0.56) & (1.95) \end{array}$$

$R^2 = 0.99$, $SER = 0.05$.

(a) What is the stochastic component in this model? How would you interpret it?

(b) Explain the role of the A_i^2 term?

(c) Comment on these results. Are the signs what you would expect?

(d) Given these estimates of the parameters at what age do men with zero and twenty years of education earn the maximum amount? Are these sensible numbers?

6. Suppose that you have a machine for filling 1 kilogram bags of sugar. The machine is set so that the weight of a bag should be a normally distributed random variable with an expected value of $1005g$ and a standard deviation of $2g$. For a random variable Z with a standard normal, the cumulative probabilities are:

z	0	0.33	0.5	1	1.5	2	2.5
$P(Z < z)$	0.5	0.6293	0.6915	0.8413	0.9332	0.9772	0.9938

(a) Expain what $P(Z < z)$ tells you.

(b) What is the probability that a bag of sugar will weight (i) less than $1000g$ (ii) between $1004g$ and $1006g$?

(c) You take a sample of 16 bags of sugar, which have a mean of $1004g$. What is (i) the standard error of the mean and (ii) the probability of getting a sample mean of $1004g$ or less?

(d) On the basis of this estimate of the mean would you conclude that the machine was still working correctly?

(e) How would you investigate whether the variance of the machine might have increased?

3.3. Answers

Question 1

(a).

$$\begin{bmatrix} y_1 \\ .. \\ y_T \end{bmatrix} = \begin{bmatrix} 1 & x_{21} \\ .. & .. \\ 1 & x_{2T} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ .. \\ u_T \end{bmatrix}$$

(b). $u'u = (y - X\beta)'(y - X\beta) = y'y + \beta'X'X\beta - 2\beta'X'y$, which is clearly a function of β.

(c).

$$\begin{aligned} \frac{\partial u'u}{\partial \beta} &= 2X'X\beta - 2X'y = 0 \\ X'X\beta &= X'y \\ \widehat{\beta} &= (X'X)^{-1}X'y \end{aligned}$$

(d)

$$\begin{aligned} \widehat{\beta} &= (X'X)^{-1}X'(X\beta + u) \\ &= \beta + (X'X)^{-1}X'u \\ E\left(\widehat{\beta}\right) &= \beta + (X'X)^{-1}X'E(u) \\ E\left(\widehat{\beta}\right) &= \beta \end{aligned}$$

since $E(u) = 0$ and X is non-stochastic.

Question 2

(a) In $R_t = \widehat{\alpha} + \widehat{\beta}\pi_t + \widehat{u}_t$; $\widehat{u}_t$ is the estimated residual, the difference between the actual and predicted value of R_t.

(b) $R^2 = 1 - \sum \widehat{u}_t^2 / \sum (R_t - \overline{R})^2$ gives the proportion of the variation in the dependent variable explained by the regression, 29% in this case so quite low.

$s = \sqrt{\sum \widehat{u}_t^2 / T - 2}$ is a measure of the average error, 2.75 percentage points in this case, quite a large error in predicting the interest rate;

$DW = \sum \Delta \widehat{u}_t^2 / \sum \widehat{u}_t^2$ is a test for serial correlation, it should be close to two, so at 0.63 this regression suffers from severe positive serial correlation.

(c) α is the value the interest rate would take if inflation were zero, interest rates would be 6.37%; β is the effect of inflation on interest rates: a 1 percentage point increase in inflation raises interest rates by 0.33 percentage points.

(d) In the model;

the regressors should be exogenous, uncorrelated with the errors, $E(\pi_t u_t) = 0$,

the regressors should not be linearly dependent, the variance of π_t not equal zero, in the case of a single regressor.

The disturbances should have

expected value (mean) zero, $E(u_t) = 0$,

be serially uncorrelated, $E(u_t u_{t-s}) = 0, s \neq 0$

with constant variance, $E\left(u_t^2\right) = \sigma^2$.

Question 3

(a) $t(\alpha = 0) = 6.37/0.66 = 9.65$ reject the hypothesis that α equals zero; $t(\beta = 1) = (0.33 - 1)/0.10 = -6.7$ reject the hypothesis that β equals one.

(b) If the real interest rate is constant plus a random error (Fisher Hypothesis) $I_t = I + u_t$ then $R_t = I + \pi_t + u_t$ then in the regression $\alpha = I$ and $\beta = 1$.

(c) Type I error is rejecting the null when it is true, Type II error is accepting the null when it is false. The probability of type I error in (a) is 5%.

(d). The 95% confidence interval is $0.33 \pm 2 \times 0.10$ i.e. the range 0.13 to 0.53.

We are 95% confident that this range covers the true value of β.

Question 4

(a) Deflator is the ratio NDY/RDY

	NDY	RDY	$Ratio \times 100$
2000	654,649	654,649	100
2001	700,538	685,263	102.3
2002	721,044	696,224	103.6

(b) Inflation, 2001-2:

DY=1.27%, RPI=1.67%; CPI=1.31; HP=16.92.

Massive house price boom, RPI slightly higher than CPI or DY.

(c). Real return is capital gain on house prices, less interest cost, less rate of inflation. Using CPI (others are acceptable)= 16.92-3.92-1.31=11.69%.

(d) Yes even if you own your own home and have paid off your mortgage there is an implicit rental cost of home ownership and this will increase when house prices increase.

Question 5.

(a) The stochastic component is u_i it is the bit of log earnings not explained by the regressors, will reflect unmeasured ability etc.

(b) We would expect earnings to rise and then fall with age, the quadratic terms captures this feature.

(c) Yes. For earnings to rise and fall with age, we need $\beta_1 > 0$ and $\gamma_1 < 0$. You earn more with better education so $\beta_2 > 0$. The positive coefficient on the interaction term γ_2 makes peak earnings later for more highly educated men, which is likely.

(d) To get the maximum

$$\begin{aligned} \frac{\partial w}{\partial A} &= \beta_1 + 2\gamma_1 A + \gamma_2 E = 0 \\ A &= -(2\gamma_1)^{-1}(\beta_1 + \gamma_2 E) \\ &= 0.0746(6.6 + 4.1E) \\ &= 0.49 + 0.31E \end{aligned}$$

note A_i and E_i are divided by 100.

For $E_i = 0$ max w =100x0.49=49 years, .

For $E_i = 0.2$ w=100x(0.49+0.31x0.2)=55 years.

Although the idea of somebody with zero years of education is unlikely, these are plausible values for peak earning age.

Question 6.

(a) The probability that a random variable Z takes a value less than a particular value z.

(b)

(i) z=(1000-1005)/2=2.5. P(z<2.5)=1-P(z<2.5)=1-0.9938=0.0062=, 0.6%, roughly one chance in 200

(ii) z=(1004-1005)/2=0.5. P(-0.5<Z<0.5)=2x(0.6915-0.5)=0.383=38%.

(c) (i) standard error of the mean is $2/\sqrt{16} = 2/4 = 0.5$,

(ii) $P(Z < 1004)$: z=(1004-1005)/0.5=-2. P(z<-2)=1-P(z<2)=1-0.9772=0.0228, 2.28%.

(d) From c(ii) the probability of getting this value or less is 2.28%, which is a small number, so it is probably not working correctly. This is not needed for the answer but if you wanted to test at the 5% level the null hypothesis that it was working properly ($\mu = 1005$), you would need to be careful whether the alternative was $\mu \neq 1005$, in which case there would be 2.5% in each tail; or $\mu < 1005$, in

which case there would be 5% in the lower tail. Since the probability is less than 2.5%, you would reject either on a two tailed or one tail test.

(e) Estimate the sample variance from your sample of 16 $s^2 = (16-1)^{-1} \sum (x_i - \overline{x})^2$, to check whether the variance seemed to have increased from 4. This is not needed for the answer but you would use a variance ratio F test.

4. Doing Your Project

To do your project you need to choose a topic and collect some data; do some statistical analysis (e.g. graphs, summary statistics); draw some conclusions and write up your results clearly in a standard academic style in less than 3,000 words. The project will count for 30% of the total marks for the course and must be submitted in **mid May**. This is to test your ability to collect and interpret data, not a test of the material covered in this course, e.g. you do not need to do regressions or give text-book material on statistical procedures.

You must submit a hard-copy version of the project, with an electronic copy of the data (e.g. on CD). We will not return your project, which we keep on file for writing references, etc. Make a copy for your own use. We do not show projects to anyone but the examiners. This is to allow students to use confidential data from work.

Keep safe backup copies of your data and drafts of your text as you work (college computers are a safe place). We are very unsympathetic if you lose work because it was not backed up properly. If you lose work, it was not backed up properly.

4.1. BASIC RULES

Email a project project proposal to Ron Smith by **the end of the second week in March**. Your proposal should be less than 100 words. It should have (1) your name; (2) your course (3) an indicative title (4) a brief description of the data you are going to use and where you will get it and (5) a short indication of the questions you might investigate. You must check that the data are actually available, you cannot do your project without data. You can change your topic subsequently if you run into difficulties.

Submit your project to the Department Office by the deadline of **in the middle of the third term**. Refer to the College rules on late submission of work.

Length: 3,000 words maximum (excluding graphs, tables, appendices, title page and abstract, but including everything else). Do not exceed this upper limit: part of the exercise is to write up your results briefly.

Your project must be typed, securely stapled and have page numbers. Do not put it in a binder or folder. You must attach a disk, CD or USB stick with your data in an envelope labelled with your name and course, firmly attached to your report (eg by stapling to the last page). The data must be in a form that allows us to replicate what you have done, e.g. in a file from a standard program.

The first page should be a title page with the following information:

- The course title and year (eg GDE ASE Project 2010)
- Title of project
- Your name
- An abstract: maximum length 100 words
- The number of words in the project
- The programs you used

You must graph the data (line graphs, histograms or scatter diagrams)
All graphs and tables must have titles and be numbered
You must have a bibliography
You must detail the sources of your data and provide it.

The project must be your own work. You can discuss it with friends or colleagues and it is a good idea for students to read and comment on each others work but it must be your work which you submit. Plagiarism is a serious offence (see the section in the course handbook).

4.2. HOW WE WILL ASSESS YOUR WORK

The criteria are listed below. We do not assign fixed weights to them, you can trade them off, but be aware of diminishing returns.

Writing. You will not get any credit unless we can understand what you have done. We look for clear organisation of the piece as a whole; clear writing of individual paragraphs and sentences; logical arguments and careful use of evidence. Check your spelling and grammar. We have set a short word limit so make every word count.

Scholarly conventions. Are sources of ideas and quotations properly acknowledged. Is there a list of references? Are data sources properly documented? Is the project written in an academic (as opposed to, say, journalistic) style? If you do it on a topic related to your work, remember the style may have to be different from that appropriate at work. Copy the styles of articles in economics journals.

Originality/interest. Most topics can be made interesting if presented sufficiently well, but it is harder to find something interesting to say about a standard

topic, such as CAPM or the aggregate consumption function, than about a slightly more unusual topic. We get bored reading over 100 projects, try to make yours memorable.

Analysis. Does your work indicate a good understanding of the relevant context, e.g. economics, institutions? Have you brought appropriate concepts, e.g. economic or finance theory, to bear on your work? Can you develop a logical argument and use evidence effectively to support your argument? Did you answer the question you posed? Are you clear about the direction of causality?

Data collection/limitations. Have you collected appropriate data (given time limitations)? Have you taken reasonable care to check the raw data and derived variables? Do you understand what your data actually measure? Are you aware of the limitations of your data? You will receive some credit for any unusual amount of work you have put into collecting data. Unless you have experience in designing surveys, do not conduct a survey to collect data.

Data summary and presentation. Have you computed appropriate derived variables? Have you noticed apparently anomalous observations? Do you demonstrate the ability to summarize and present data in a clear and effective way?

Statistical Methods. Have you used appropriate statistical methods? Use the simplest technique that will answer your question. Have you qualified any conclusions that you have drawn? e.g. pointed out that the sample size is small or that you have been unable to control for certain factors, etc. Beware of using advanced statistical techniques that you do not understand; you will be penalised for any mistakes you make in their use.

Interpretation. How well have you interpreted your data? Have you borne its limitations in mind when interpreting it? Does your interpretation reveal understanding of the relevant concepts?

4.3. CHOOSING A TOPIC

You can do your project on anything that involves interpreting data, it does not have to be narrowly economic or financial. The topic may come from your work, issues that have come up in the course, items in the news, or anything that interests uou. Often choice of topic is prompted by the data available.

4.4. DATA

Barrow Ch. 9 discusses data. You must give us a copy of the data you have used. If you need to use confidential work-related data, we can provide a letter to your employer explaining that it will be kept confidential. You should choose a topic on which you can find data without too much effort. If you cannot make substantial progress in finding data in 2-3 hours systematic search, either in the library or over the internet, you should probably change your topic. There is a vast amount of statistics available on the Web from governments, central banks, international organisations (IMF, OECD or World Bank). Also check Birkbeck eLibrary, statistical databases; datastream is available in the library. The main UK source is the Office of National Statistics, US Data is available on the Federal Reserve Economic Database and the Bureau of Economic Analysis. Try Google or other search engines: just type the topic you are interested in and then data, e.g. "Road Traffic Deaths Data" got various sites with international data on road traffic deaths.

Check your data, no matter where it comes from. Errors (eg a decimal point in the wrong place) can cause havoc if you miss them. Check for units, discontinuities and changes in definitions of series (e.g. unification of Germany). Check derived variables as well as the raw data. Calculating the minimum, maximum and mean can help to spot errors. Carry out checks again if you move data from one type of file to another.

4.5. WHAT YOUR REPORT SHOULD LOOK LIKE

Your project report should tell a story, with a beginning, a middle and an end. It is a story about your investigation not part of your autobiography. The following structure is a suggestion, adapt it to suit your question. Look at the structure used in section 7.7, which describes UK growth and inflation.

4.5.1. ABSTRACT

Here you must summarize your project in 100 words or less. Many journals print abstracts at the start of each paper, copy their form

4.5.2. INTRODUCTION.

Explain what you are going to investigate, the question you are going to answer, and why it is interesting. Say briefly what sort of data you will be using (eg.

quarterly UK time-series 1956-2009 in section 7.7). Finish this section with a paragraph which explains the organization of the rest of your report.

4.5.3. BACKGROUND

This section provides context for the analysis to follow, discusses any relevant literature, theory or other background, e.g. explanation of specialist terms. Do not give standard textbook material; you have to tell us about what we do not know, not what we do know. On some topics there is a large literature on others there will be very little. The library catalogue, the EconLit database and the library staff can help you to find literature.

In many cases, this section will describe features of the market or industry you are analyzing. In particular, if you are writing about the industry in which you work, you should make sure you explain features of the industry, or technical terms used in it, which may be very well known to everyone in it, but not to outsiders.

4.5.4. DATA

Here you should aim to provide the reader with enough information to follow the rest of the report, without holding up the story line. Details can be provided in an appendix. You should discuss any peculiarities of the data, or measurement difficulties. You may need to discuss changes in the definition of a variable over time.

4.5.5. ANALYSIS

The background should guide you in suggesting features of the data to look at, hypotheses to test, questions to ask. You must have tables and graphs describing the broad features of the data. In the case of time series data these features might include trends, cycles, seasonal patterns and shifts in the mean or variance of the series. In the case of cross-section data they might include tables of means and standard deviations, histograms or cross-tabulations. In interpreting the data, be careful not to draw conclusions beyond those that are warranted by it. Often the conclusions you can draw will be more tentative than you would like; data limitations alone may ensure this. Do not allow your emotional or ethical responses to cloud your interpretation of what you find in the data.

If you run regressions, report: the names of variables (including the dependent variable); number of observations and definition of the sample; coefficents and either t-ratios, standard errors or p values; R-squared (or R-bar-squared); standard error of the regression; and any other appropriate test statistics such as Durbin-Watson for time-series.

4.5.6. SUMMARY AND CONCLUSIONS.

What are the main findings of your work, the answers to the questions you posed in the introduction? How must your findings be qualified because of the limitations of the data or the methods of analysis you have employed? Do they have policy implications (public or private)? Do you have suggestions for further work?

4.5.7. BIBLIOGRAPHY

You must give a bibliographic citation for any work referred to in the text, follow the Harvard system, used in section 1.5.

4.5.8. APPENDICES

You must have a data appendix, giving precise definitions of variables, and details of the sources. The guiding principle is that you should provide enough detail to enable the reader to reproduce your data. Give the data in electronic form attached to the project.

4.6. Good luck

You can learn a lot by doing your project. The skills you can develop in data analysis, interpretation and presentation are valuable in the labour market; and having a project to show a potential employer can be useful in getting a job. Doing your project can also be a very enjoyable experience. The more care you take with it, the more you will learn, and the more fun you will have.

5. PART II: NOTES

The word Statistics has at least three meanings. Firstly, it is the data themselves, e.g. the numbers that the Office of National Statistics collects. Secondly, it has a technical meaning as measures calculated from the data, e.g. an average. Thirdly, it is the academic subject which studies how we make inferences from the data.

Descriptive statistics provide informative summaries (e.g. averages) or presentations (e.g. graphs) of the data. We will consider this type of statistics first. Whether a particular summary of the data is useful or not depends on what you want it for. You will have to judge the quality of the summary in terms of the purpose for it is used, different summaries are useful for different purposes.

Statistical inference starts from an explicit probability model of how the data were generated. For instance, an empirical demand curve says quantity demanded depends on income, price and random factors, which we model using probability theory. The model often involves some unknown parameters, such as the price elasticity of demand for a product. We then ask how to get an estimate of this unknown parameter from a sample of observations on price charged and quantity sold of this product. There are usually lots of different ways to estimate the parameter and thus lots of different estimators: rules for calculating an estimate from the data. Some ways will tend to give good estimates some bad, so we need to study the properties of different estimators. Whether a particular estimator is good or bad depends on the purpose.

For instance, there are three common measures (estimators) of the typical value (central tendency) of a set of observations: the arithmetic *mean* or average; the *median,* the value for which half the observations lie above and half below; and the *mode,* the most commonly occurring value. These measure different aspects of the distribution and are useful for different purposes. For many economic measures, like income, these measures can be very different. Be careful with averages. If we have a group of 100 people, one of whom has had a leg amputated, the average number of legs is 1.99. Thus 99 out of 100 people have an above average number of legs. Notice, in this case the median and modal number of legs is two.

We often want to know how dispersed the data are, the extent to which it can differ from the typical value. A simple measure is the *range*, the difference between the maximum and minimum value, but this is very sensitive to extreme values and we will consider other measures below.

Sometimes we are interested in a single variable, e.g. height, and consider its average in a group and how it varies in the group? This is univariate statistics,

to do with one variable. Sometimes, we are interested in the association between variables: how does weight vary with height? or how does quantity vary with price? This is multivariate statistics, more than one variable is involved and the most common models of association between variables are correlation and regression, covered below.

A model is a simplified representation of reality. It may be a physical model, like a model airplane. In economics, a famous physical model is the Phillips Machine, now in the Science Museum, which represented the flow of national income by water going through transparent pipes. Most economic models are just sets of equations. There are lots of possible models and we use theory (interpreted widely to include institutional and historical information) and statistical methods to help us choose the best model of the available data for our particular purpose. The theory also helps us interpret the estimates or other summary statistics that we calculate.

Doing applied quantitative economics or finance, usually called econometrics, thus involves a synthesis of various elements. We must be clear about why we are doing it: the purpose of the exercise. We must understand the characteristics of the data and appreciate their weaknesses. We must use theory to provide a model of the process that may have generated the data. We must know the statistical methods which can be used to summarise the data, e.g. in estimates. We must be able to use the computer software that helps us calculate the summaries. We must be able to interpret the summaries in terms of our original purpose and the theory.

5.1. Example: the purpose of AA guns

The booklet contains a lot of examples, a number of which are not from economics or finance, because the issues are often simpler in other areas. This example is to illustrate the importance of interpreting statistical summaries in terms of purpose. At the beginning of World War II, Britain fitted some merchant ships with anti-aircraft (AA) guns. A subsequent statistical analysis showed that no German planes had ever been hit by merchant AA guns and it was decided to remove them. However, before this was done another statistical analysis showed that almost none of the AA equipped ships had been hit by bombs from German aircraft, whereas large numbers of those without AA had been hit. This was the relevant statistic and the AA guns were kept on merchant ships. Although the guns did not hit the bombers, but they kept the bombers further away from the

ships, reducing the probability of them damaging the ships. Other examples of this sort of use of statistics in World War II can be found in *The Pleasures of Counting, T.W.* Korner, Cambridge University Press, 1996.

5.2. Example: the Efficient Market model.

A simple and very powerful model in economics and finance is the random walk

$$y_t = y_{t-1} + \varepsilon_t.$$

This says that the value a variable takes today, time t, is the value that it had yesterday, time $t-1$, plus a random shock, ε_t. The shock can be positive or negative, averages zero and cannot be predicted in advance. Such shocks are often called 'White noise'. To a first approximation, this is a very good description of the logarithm of many asset prices such as stock market prices and foreign exchange rates, because markets are quite efficient: the change in log price (the growth rate) $\Delta y_t = y_t - y_{t-1} = \varepsilon_t$ is random, unpredictable. Suppose that people knew something that will raise the price of a stock tomorrow, they would buy today and that will raise the price of the stock today. Any information about the future that can be predicted will be reflected in the price of the stock now. So your best estimate of tomorrow's price is today's price. What will move the price of the stock will be new, unpredicted, information. The random shock or error ε_t represents that unpredictable information that changes prices. Most of our models will involve random shocks like ε_t. Sometimes a firm will report a large loss and its stock price will go up. This is because the market had been expecting even worse losses, which had been reflected in the price. When reported losses were not as bad as expected the price goes up. Whether the efficient market hypothesis is strictly true is a subject of controversy, but it is an illuminating first approximation.

If the variable has a trend, this can be allowed for in a random walk with drift

$$y_t = \alpha + y_{t-1} + \varepsilon_t.$$

Then the variable increases on average by α every period. If the variable is a logarithm, α is the average growth rate. This is a parameter of the model, which we we will want to estimate from data on y_t. Parameters like α and random errors or shocks like ε_t will play a big role in our analysis.

5.3. Notation

It is very convenient to express models in mathematical notation, but notation is not consistent between books and the same symbols means different things in different disciplines. For instance, Y often denotes the dependent variable but since it is the standard economic symbol for income, it often appears as an independent variable. It is common to use lower case letters to indicate deviations from the mean, but it is also common to use lower case letters to denote logarithms. Thus y_t could indicate $Y_t - \overline{Y}$ or it could indicate $\ln(Y_t)$. The logarithm may be written $\ln(Y_t)$ or $\log(Y_t)$, but in empirical work natural logarithms, to the base e, are almost always used. The number of observations in a sample is sometimes denoted T for time series and sometimes N or n for cross sections.

In statistics we often assume that there is some true unobserved parameter and wish to use data to obtain an estimate of it. Thus we need to distinguish the true parameter from the estimate. This is commonly done in two ways. The true parameter, say the standard deviation, is denoted by a Greek letter, say σ, and the estimate is denoted either by putting a hat over it, $\widehat{\sigma}$, said 'sigma hat' or by using the equivalent latin letter, s. In many cases we have more than one possible estimator (a formula for generating an estimate from the sample) and we have to distinguish them. This is the case with the standard deviation, there are two formulae for calulating it, denoted in these notes by $\widehat{\sigma}$ and s. However, books are not consistent about which symbol they use for which formula, so you have to be careful.

The Greek alphabet is used a lot. It is given below, with the upper case letter, lower case, name and example.

A α alpha; α often used for intercept in regression.

B β beta; β often used for regression coefficients and a measure of the risk of a stock in finance.

Γ γ gamma.

Δ δ delta; used for changes, $\Delta y_t = y_t - y_{t-1}$; δ often rate of depreciation.

E ϵ or ε epsilon; ε often error term.

Z ζ zeta.

H η eta; η often elasticity.

Θ θ theta; Θ sometimes parameter space; θ often a general parameter.

I ι iota.

K κ kappa.

Λ λ lambda; λ often a speed of adjustment.

M μ mu; μ often denotes expected value or mean.

N ν nu.

Ξ ξ xi.

O o omicron.

Π π pi; (ratio of circumference to diameter) often used for inflation. Π is the product symbol: $\prod y_i = y_1 \times y_2 \times ... \times y_n$.

P ρ rho; ρ often denotes autocorrelation coefficient.

Σ σ sigma; σ^2 usually a variance, σ a standard deviation, Σ is the summation operator, also sometimes used for a variance covariance matrix.

T τ tau.

Υ υ upsilon.

Φ ϕ φ phi; $\Phi(y)$ sometimes normal distribution function; $\phi(y)$ normal density function.

X χ chi; χ^2 distribution.

Ψ ψ psi.

Ω ω omega; Ω often a variance covariance matrix.

6. Descriptive statistics

Data tend to come in three main forms:

-time-series, e.g. observations on annual inflation in the UK over a number of years;

-cross-section, e.g. observations on annual inflation in different countries in a particular year; and

-panels e.g. observations on inflation in a number of countries in a number of years.

Time-series data have a natural order, 1998 comes after 1997; cross-section data have no natural order; the countries could be ordered alphabetically, by size or any other way.

The data are usually represented by subscripted letters. So the time-series data on inflation may be denoted y_t, $t = 1, 2, ..., T$. This indicates we have a sequence of observations on inflation running from $t = 1$ (say 1961) to $t = T$ (say 1997) so the number of observations $T = 37$. For a set of countries, we might denote this by y_i, $i = 1, 2, ..., N$. Where, if they were arranged alphabetically, $i = 1$ might correspond to Albania and $i = N$ to Zambia. Panel data would be denoted y_{it}, with a typical observation being on inflation in a particular country i, say the UK, in a particular year, t, say 1995, this gives TN observations in total. We will use both T and N to denote the number of observations in a sample.

Graphs are generally the best way to describe data. There are three types of graph economists commonly use. Firstly, for time-series data, we use a line graph, plotting the series against time. We can then look for trends (general tendency to go up or down); regular seasonal or cyclical patterns; outliers (unusual events like wars or crises). Secondly, we can plot a histogram, which gives the number (or proportion) of observations which fall in a particular range. Thirdly we can plot one variable against another to see if they are associated, this is a scatter diagram or X-Y Plot. Barrow Chapter 1 has lots of examples.

6.1. Summary Statistics

We will use algebra, particularly the summation operator, to describe operations on data. The formulae may look complicated, but they are just a set of instructions. Suppose we have a series of numbers: 2,4,6,8, which we denote, x_1, x_2, x_3, x_4;

x_i, $i = 1, 2, ..., N$, where $N = 4$. The sum of these is 20, which we denote

$$\sum_{i=1}^{N} x_i = 2 + 4 + 6 + 8 = 20$$

this simply says add together the N elements of x. If we multiply each number by a constant and add a constant to each number to create $y_i = a + bx_i$, then

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} (a + bx_i) = Na + b \sum_{i=1}^{N} x_i.$$

In the example above for $a = 1, b = 2$, then $y_i = 5, 9, 13, 17$, with sum 44, which is the same as $4 \times 1 + 2 \times 20$.

6.1.1. The mean, a measure of central tendency or typical value

The arithmetic mean (average) of x_i, usually denoted by a bar over the variable, said 'x bar', is defined as

$$\overline{x} = \sum_{i=1}^{N} x_i / N.$$

In this example, it is 20/4=5. The formula just says add up all the values and divide by the number of observations. There are other sorts of mean. For instance, the geometric mean is the Nth root of the product of the numbers

$$GM(x) = \sqrt[N]{x_1 \times x_2 \times ... \times x_N}$$

and can be calculated as the exponential (anti-log) of the arithmetic mean of the logarithms of the numbers, see Barrow P54.

6.1.2. The variance and standard deviation, measures of dispersion

The variance, often denoted σ^2, and standard deviation (square root of the variance, σ) measure how dispersed or spread out the observations are. The variance is more convenient for mathematical derivations, but in applied work always use the standard deviation. The standard deviation is in the same units as the original variable. If a variable is normally distributed two thirds of the observations will fall in the range of the mean plus and minus one standard deviation and 95%

of the observations will lie in the range the mean plus or minus two standard deviations.

One estimator of the variance of x_i, (sometimes called the population variance) is

$$\widehat{\sigma}^2 = \sum_{i=1}^{N} (x_i - \overline{x})^2 / N.$$

Notice here we distinguish between the true value σ^2 and our estimate of it $\widehat{\sigma}^2$. This formula gives a set of instructions. It says take each of the observations and subtract the mean, $(x_i - \overline{x})$; square them $(x_i - \overline{x})^2$; add them together $\sum_{i=1}^{N} (x_i - \overline{x})^2$ and divide them by the number of observations, 4 in this case: $\sum_{i=1}^{N} (x_i - \overline{x})^2 / N = 20/4 = 5$.

i	x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
1	2	-3	9
2	4	-1	1
3	6	1	1
4	8	3	9
sum	20	0	20

In this case both the Mean and the Variance are 5. The standard deviation, $SD(x) = \widehat{\sigma}$ is the square root of the variance: 2.24 in this case.

Another estimator of the variance of x_i, (sometimes called the sample variance) is

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

We discuss the difference between $\widehat{\sigma}^2$ and s^2 below.

6.1.3. Covariance and correlation, measures of association

The covariance, which is used to measure association between variables is

$$Cov(x, y) = \sum_{i=1}^{N} (x_i - \overline{x}) (y_i - \overline{y}) / N.$$

The Covariance will be positive if high values of x are associated with high values of y, negative if high values of x are associated with low values of y. It will be zero is there is no linear relationship between the variables. The covariance can be difficult to interpret, so it is often standardised to give the correlation coefficient,

by dividing the covariance by the product of the standard deviations of the two variables.

$$r = \frac{Cov(x, y)}{SD(x)SD(y)}$$

The correlation coefficient lies between plus and minus one, $-1 \leq r \leq 1$. A correlation coefficient of -1 means that there is an exact negative linear relation between the variables, $+1$ an exact positive linear relation, and 0 no linear relation. Correlation does not imply causation. Two variables may be correlated because they are both caused by a third variable.

6.1.4. Standardised data

Data are often standardised by subtracting the mean and dividing by the standard deviation (the square root of the sample variance),

$$z_i = \frac{x_i - \overline{x}}{\sigma}.$$

This new variable, z_i, has mean zero and variance (and standard deviation) of one. Notice the correlation coefficient is the covariance between the standardised measures of x and y.

6.1.5. Moments

A distribution is often described by:

- its moments, which are $\sum_{i=1}^{N} {x_i}^r/N$. The mean $\overline{x} = \sum x_i/N$ is the first moment, $r = 1$.
- its centred moments $\sum_{i=1}^{N}(x_i - \overline{x})^r/N$. The variance, $\sigma^2 = \sum_{i=1}^{N}(x_i - \overline{x})^2/N$ is the second centred moment, $r = 2$. The first centred moment, $\sum_{i=1}^{N}(x_i - \overline{x})/N = 0$.
- its standardised moments $\sum z_i^r/N$, where $z_i = (x_i - \overline{x})/s$. The third standardised moment, $r = 3$, is a measure of whether the distribution is symmetrical or skewed. The fourth standardised moment, $r = 4$, is a measure of kurtosis (how fat the tails of the distribution are). For a normal distribution, the coefficient of skewness $\sum z_i^3/N$ is zero, and the coefficient of kurtosis $\sum z_i^4/N$ is 3.

Some distributions do not have moments. The average (mean) time to get a PhD is not defined since some students never finish, though the median is defined, the time it takes 50% to finish.

6.2. Example, averages and reversals

Suppose a firm has two factories one in the low wage north, where it employs mainly male staff, and the other in the high wage south, where it employs mainly females. In both it pays males more than females. The Table below gives the number of male staff, NM, the male wage, WM, the number of females, NF and the female wage WF.

	NM	WM	NF	WF
N	200	350	50	300
S	50	500	200	450

The average male wage is $(200 \times 350 + 50 \times 500)/(200 + 50) = 380$. The average female wage is $(50 \times 300 + 200 \times 450)/(50 + 200) = 420$. Despite paying men more than women at both factories, the average female wage is higher than the average male wage, because it is employing more women in the high wage south. This reversal is know as Simpson's paradox, though he was not the first to note it.

6.3. Example, Do Economists agree?

It is said that when you get two economists together, you will get three opinions; because one will say 'on the one hand ..., on the other hand ...,'. In June 2004 the BBC sponsored a survey of the views of the Society of Business Economists (www.sbe.co.uk) on the trend in UK house prices over the next three years. 225 members replied. They had to choose from various ranges, so they could not say 'on the one hand they might go up, but on the other hand they might go down'.

The percentages $f(x_i)$ choosing each range, x_i, for expected growth in house prices were

x_i	$f(x_i)$
$> 15\%$	15%
$5 - 15\%$	25%
$0 - 5\%$	18%
$-10 - 0\%$	24%
$< -10\%$	18%

There are two modes: roughly a quarter think prices will go up by between 5 and 15% and a quarter think prices will fall by up to 10%. The median (half above

and half below) falls in the range 0 to 5%. The distribution is broadly symmetric. To calculate means and variances, we need to assign values to the ranges, which is inevitably arbitrary to some extent. We will use mid-points of closed ranges, e.g. the mid point of 5-15% is 10%; and for the open ranges treat more than 15% as 20%, and less than -10% as -15%. This gives the values X_i below. We also give p_i the proportions in each range, percentages divided by 100.

We cannot use our standard formula for the mean, so we need to adjust it. See Barrow p27. Call the total number of respondents, $N = 225$. Call the number who responded in each range N_i. So the number in the lowest range (who reponded that they would fall by more than 10%), $N_1 = 0.18 * 225 = 40.5$. The percentages are rounded, which is why it is not an integer. We could calculate the mean by multiplying the the value of each answer by the number who gave that value, adding those up over the 5 ranges and dividing by the total number, but it is easier to calculate, if we rewrite that formula in terms of the proportions.

$$\overline{X} = (\sum_{i=1}^{5} N_i X_i)/N = \sum_{i=1}^{5} \left(\frac{N_i}{N}\right) X_i = \sum_{i=1}^{5} p_i X_i.$$

We can do the same thing for calculating the variance, giving

$$\sigma^2 = \sum_{i=1}^{5} p_i (X_i - \overline{X})^2.$$

So in the table, we give the values X_i; the proportions p_i (note they sum to one); calculate the product $p_i X_i$, then sum these to get $\overline{X} = 2.05$. Then we calculate $(X_i - \overline{X})$ and $p_i(X_i - \overline{X})$ (note these sum to zero). Then multiply $p_i(X_i - \overline{X})$ by $(X_i - \overline{X})$ to get $p_i(X_i - \overline{X})^2$. We sum these to get the variance. The calculations are given below.

X_i	p_i	$p_i X_i$	$(X_i - \overline{X})$	$p_i(X_i - \overline{X})$	$p_i(X_i - \overline{X})^2$
20	0.15	3	17.95	2.6925	48.330375
10	0.25	2.5	7.95	1.9875	15.800625
2.5	0.18	0.45	0.45	0.081	0.03645
-5	0.24	-1.2	-7.05	-1.692	11.9286
-15	0.18	-2.7	-17.05	-3.069	52.32645
Sum	1	2.05	2.25	0	128.4225

Since the variance $\sigma^2 = 128.4225$ the standard deviation, its square root is $\sigma = 11.33$.

To summarise the mean forecast of SBE respondents was for house price growth of 2.05, with a standard deviation of 11.33= $\sqrt{128.4225}$, which indicates the large range of disagreement. The mean falls in the same range as the median, which also indicates that the distribution is fairly symmetric, but it is bimodal with about a quarter thinking prices will rise between 5 and 15%, and a quarter thinking that they would fall by up to 10%.

In June 2004 when the survey was conducted, the average house price in the UK according to the Nationwide Building Society was £151,254. In June 2007, three years later, it was £184,074; a rise of 22%, though prices subsequently fell and in June 2009 it was £156,442.

6.4. Background Example: risk and return

When judging whether a particular investment, e.g. buying the shares of a particular company) is worthwhile you need to try and judge the expected return (dividends plus capital gain for an equity stock) and the expected risk. Although the past is not always a reliable guide to the future, we use past data to try and get estimates of likely future risk and return. The variance (or standard deviation) of the returns on a stock are the standard way of measuring the risk or volatility of a share. You want to have a higher mean return but if you are risk averse you would like a lower variance: given the choice of two stocks with the same expected return, you would choose the one with the lower variance. Riskier stocks pay higher returns. One measure that reflects both is the Sharpe Ratio:

$$\frac{R_i - R}{\sigma_i}$$

the ratio of the difference between the mean return on the stock from the risk free rate of interest to the standard deviations of the stock's returns. There are a range of other financial performance measures that combine risk and return.

You can reduce risk by diversifying your portfolio, holding more than one share. Suppose there are two stocks, with mean returns μ_1 and μ_2, variances σ_1^2 and σ_2^2, and covariance between them of σ_{12}. Suppose you form a Portfolio with share w in stock one and $(1-w)$ in stock two. The mean return on the portfolio is

$$\mu_p = w\mu_1 + (1-w)\mu_2$$

the variance of the portfolio is

$$\sigma_p^2 = w^2\sigma_1^2 + (1-w)^2\sigma_2^2 + 2w(1-w)\sigma_{12}.$$

Therefore, if the covariance between the stocks is negative, the portfolio will have a smaller variance: when one stock is up, the other is down. Even if the covariance is zero there are gains from diversification. Suppose $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and $\sigma_{12} = 0$. Then

$$\begin{aligned} \sigma_p^2 &= w^2\sigma^2 + (1-w)^2\sigma^2 \\ &= (1 + 2(w^2 - w))\sigma^2 \end{aligned}$$

since $w < 1$, then $w^2 < w$ and the second term is negative, making $\sigma_p^2 < \sigma^2$.

7. Economic and Financial Data I: numbers and graphs

7.1. Tables and Calculations

Data will typically come in a Table, either electronic or hard-copy. **When constructing your own tables, make sure you put a Title, full definitions of the variables, the units of measurement and the source of the data.** Be clear on the units of measurement and get a feel for the orders of magnitudes: what are typical values, what is the range (the highest and lowest values). When comparing series or graphing them together make sure they are in comparable units.

Consider the following table.

Gross National Product (billions of constant 1995 US$), Population (millions), Military Expenditure (billions of constant 1995 US$) and the number in the armed forces (thousands) in Developed and Developing Countries, 1985 and 1995.

	1985	1995
GNP		
Developed	21190	23950
Developing	4184	7010
Population		
Developed	1215.7	1151.5
Developing	3620.8	4520.0
Military Expenditures		
Developed	1100.8	667.8
Developing	230.0	196.7
Number in Armed Forces		
Developed	11920	7667
Developing	16150	15120

Source World Military Expenditures and Arms Transfers, US Arms Control and Disarmament Agency. The developed group includes 33 high per-capita income countries, the rest of the world is developing.

Gross National Product, GNP, and Gross Domestic Product are two different measures of the output of a country. GNP includes production by the nationals of the country, GDP includes the production within the boundary of the country whether by nationals or foreigners. The difference is net property income from

abroad. The definition of Population is straightforward, though it can be difficult to count everybody. Number in the armed forces can raise problems of definition in countries with para-military units like the French Gendarmerie, are they armed forces or not? Defining what should be included in military expenditure also raises difficulties and many countries are secretive about what they spend on the military. There are quite large margins of error on all these numbers. Many economic measures are only ROMs (rough orders of magnitude) others are WAGs (wild arsed guesses).

From this table we can calculate derived measures like (a) per-capita income, GNP per head, by dividing GNP by population for the world as a whole and for the developed and developing countries (we would have to calculate the world totals) (b) the average annual growth rate of population or per-capita income (GNP per head) between 1985 and 1995; (c) the percentage share of military expenditure in GDP (d) the number of people in the armed forces per 1000 population for the world as a whole and for the developed and developing countries in 1985 and 1995.

When doing these calculations, it is crucial to be careful about units. These variables are all measured in different units and ratios will depend on the units of the numerator and denominator. Expressing the units as powers of 10 is often useful. $1 = 10^0$; $10 = 10^1$; $100 = 10^2$; $1,000,000 = 10^6$. The power gives you the number of zeros after the one.

GNP and Military Expenditure are measured in US billions, thousand millions, 10^9 1995 US$, so the military expenditure GDP ratio is a proportion. For developed countries in 1985 it is $1100.8/21190 = 0.0519$ to convert to percent multiply by 100, 5.19% of GNP was devoted to the military.

Population is measured in millions, 10^6; the number in the armed forces in thousands, 10^3. Divide the number in the armed forces by population for developed countries in 1985 it is $11920/1215.7 = 9.8$. This is in units of the numerator divided by the denominator: $10^3/10^6 = 10^{3-6} = 10^{-3}$; one per thousand. Thus there are roughly 10 members of the armed forces per thousand population in the developed countries in 1985; about 1% of the population.

GNP per capita is GNP (10^9) divided by population (10^6) so is (10^3) so the 1985 figure for developed countries $21190/1215 = 17.44$, is measured in thousands (10^3).Thus average income in the developed world in 1985 in 1995 dollars was about $\$17,500$ compared the figure for the developing world to $4184/3620.8 = 1.155$, \$1155.

The growth rate in GNP for the developed countries 1985 to 1995 is $(23950 -$

21190)/21190 = 23950/21190 − 1 = 0.13, 13% over ten years, roughly 1.3% per annum. Notice whether growth rates are expressed as proportions (0.13) or percentages (13%) and the period they are calculated over.

7.2. Graphs

Graphs are usually the most revealing way to present data. See examples in 7.7. For time-series data the starting point is a line graph, just a plot over time. If you want to plot more than one series on the graph to see if they move together, make sure that they are of the same scale. Do not put too many series on the same graph. When you look at a time-series graph look for trends (steady upward, or less often downward, movement in the series); seasonal patterns (if it is quarterly or monthly data); cycles, periods of boom or recession spreading over a number of years; outliers which look out of line with the other numbers. Outliers may be produced by unusual events (wars, financial crises, etc) or they may just be mistakes in the data.

Scatter diagrams plot one series against another and are typically used to investigate whether there is an association between two variables. Look to see how close the association is, whether it is positive or negative, whether there are outliers which do not fit into the standard pattern.

Histograms, or frequency distributions, are pictures which give the number (or proportion) of the observations that fall into particular ranges. We will use these extensively in lectures. Look to see whether the distribution is unimodal or bimodal; whether it is skewed; how dispersed it is and whether there are outliers.

7.3. Transformations

In many cases, we remove the effects of trends, changes in price levels etc. by working with either growth rates, or with ratios. In economics and finance certain ratios tend to be reasonably stable (i.e. not trended). An example is the Average Propensity to Consume (the ratio of Consumption to Income) or the Savings Ratio. Since income equals savings plus consumption $Y = S + C$, the average propensity to consume equals one minus the savings rate $APC = C/Y = 1 - S/Y$. Where Y is income, S savings, C consumption, SR savings ratio, APC average propensity to consume. SR and APC can be expressed either as proportions as here or multiplied by 100 to give percent. In finance, we work with ratios like the Price-Earnings Ratio or the Dividend Yield. Notice these ratios can be compared

across countries, because the units of currency in the numerator and denominator cancel.

Theory will often tell you what variables to construct, e.g.

-Real Interest Rates equal to the nominal ordinary interest rate minus the (expected) rate of inflation;

-real exchange rate, the nominal exchange rate times the ratio of foreign to domestic price indexes;

-the velocity of circulation, the ratio of nominal GDP to the money supply.

7.4. National Accounts

The output of an economy is usually measured by Gross Domestic Product, GDP, or sometimes Gross National Product, GNP. These measures are part of a system of National Accounts, which start from the identity that output equals expenditure equals income. Anything produced, output, is sold to somebody, and is then their expenditure, and the money generated by the sale is paid to somebody, it is their income. Expenditure is made up of consumption C_t, plus investment I_t, plus government spending on goods and services (i.e. excluding government expenditure on transfer payments) G_t, plus exports X_t, minus imports M_t. Income is made up of wages W_t plus profits P_t. In any period:

$$Y_t = C_t + I_t + G_t + X_t - M_t = W_t + P_t.$$

Output produced but not sold, left in the warehouse is treated as a change in inventories part of investment by firms. Investment includes Gross Domestic Fixed Capital Formation and acquisition of inventories. Although, in principle, output, expenditure and income are identically equal, in practice because of measurement errors, they are not and 'balancing items' are added to make them match. Coverage can be national (by the citizens of the country) or domestic (within the boundaries of the country). The difference between Gross National Income and Gross Domestic Product is net income (recipts less payments) from the rest of the world. The measures can be gross of the depreciation of capital stock or net of it. Be careful about Gross and Net measures, since it is often not clear from the name what is being netted out. The measure can be at the prices actually paid (market prices) or what the producers actually receive (factor cost) the difference is indirect taxes less subsidies. The UK national accounts conventions are based on the European System of Accounts 1995 and are published in the Blue Book.

GDP or other national income aggregates measure marketed outputs and there are lots of activities that are left out. These include domestic activities (household

production), environmental impacts, illegal activities, etc. If there is an increase in crime which leads to more security guards being hired and more locks fitted, this increases GDP. There are various attempts to adjust the totals for these effects although, so far, they have not been widely adopted. You should be aware of the limitations of GDP etc as measures. There is a good discussion of the issues and alternatives in the *Report by the Commission on the Measurement of Economic and Social Progress* (2009) available on www.stiglitz-sen-fitoussi.fr.

The accounts are divided by sector. The private sector covers firms (the corporate sector usually divided into financial and non-financial) and households; the public sector covers general government (which may be national or local) and sometimes state owned enterprises, though they may be included with the corporate sector; the overseas sector covers trade. Corresponding to the output, expenditure and income flows, there are financial flows between sectors. Define T_t as taxes less transfer payments. The total $Y_t - T_t$ factor income minus taxes plus transfer payments (e.g. state pensions or unemployment benefit) is known as disposable income. Subtract T_t from both sides of the income-expenditure identity

$$Y_t - T_t = C_t + I_t + G_t - T_t + X_t - M_t$$

note that savings $S_t = Y_t - T_t - C_t$. Move C_t and I_t to the left hand side to give:

$$(S_t - I_t) = (G_t - T_t) + (X_t - M_t)$$

$$(S_t - I_t) + (T_t - G_t) + (M_t - X_t) = 0$$

the three terms in brackets represent what each sector - private, public, overseas - needs to borrow or lend and total borrowing must equal total lending. If savings is greater than investment, the private sector has a surplus of cash which it can lend to the public or overseas sector. They sum to zero because for every borrower there must be a lender.

7.5. Unemployment

We often have a theoretical concept and need to provide an 'operational' definition, a precise set of procedures which can be used by statistical offices, to obtain measures. This raises questions like what is the best operational measure and how well does it correspond to the particular theoretical concept. Unemployment is a case in point. There are a number of different theoretical concepts of unemployment and a number of different ways of measuring it.

Do you think the following people are unemployed:

a student looking for a summer job who cannot find one;

a 70 year old man who would take a job if one was offered;

a mother looking after her children who would take a job if she could find good child care;

an actor 'resting' between engagements;

someone who has been made redundant and will only accept a job in the field they previously worked in for the same or better salary?

One method is the 'Claimant count', i.e. the number who are registered unemployed and recieving benefit. But this is obviously very sensitive to exact political and administrative decisions as to who is entitled to receive benefit.

An alternative is a survey, which asks people of working age such questions as

(i) are you currently employed; if not

(ii) are you waiting to start a job; if not

(iii) have you looked for work in the last four weeks.

Those in category (iii) will be counted as unemployed.

7.6. Example: real interest rates

World War I ended in November 1918. The table below gives data on the percentage unemployment rate U; the Retail Price Index 1963=100, RPI; the yield on 3 month Treasury Bills, R, (these were not issued during the war); GDP per capita in 1913 prices Y; and the dollar-sterling exchange rate, $\$/£$ (sterling was not convertible during the war); for 1918-1922.

$Year$	U	RPI	R	Y	$\$/£$
1918	0.8	42		54	
1919	2.1	56	3.5	48	4.42
1920	2.0	52	6.2	47	3.66
1921	12.9	47	4.6	42	3.85
1922	14.3	38	2.6	44	4.43

(a) Calculate the inflation rate and the growth rate of per capita income 1919-1922.

(b) Calculate the ex post real interest rate 1919-1922?

(c) Explain the difference between an ex ante and an ex post real interest rate.

(d) Use the data to evaluate the proposition: "Post war monetary policy was strongly contractionary in order to deflate the economy so that sterling could

return to gold at its pre-war parity of 4.87. The consequence was continued recession and mass unemployment."

Answer

$Year$	$Growth\%$	$INF\%$	$EARIR$	$EPRIR$	$\$/£$
1918					
1919	-11	33.3	-29.8	10.6	4.42
1920	-2	-7.1	13.3	15.8	3.66
1921	-11	-9.6	14.2	23.7	3.85
1922	5	-19.1	21.7		4.43

(a) Growth is $100(Y_t/Y_{t-1} - 1)$. For 1919 it is $100((48/54) - 1) \approx -11\%$, per capita income fell by over 10%, and continued falling till 1921. This fall produced rising unemployment. Inflation is calculated as the percentage change in the RPI. Prices were rising between 1918 and 1919, but then fell giving negative inflation, deflation. Between 1919 and 1922 prices fell by almost a third.

(b) If you lend £1 at 15% for a year you get £1.15 at the end of the year, but if prices rise at 10%, over the year what you can buy with your £1.15 has fallen, the real rate of return is only 5%=15%-10%. This is the expost (from afterwards) rate, using the actual inflation over the time you lend. So the ex post real interest rate for 1919 is EPRIR=$3.5 - (-7.1) = 10.6$.

(c) When you lend the money you do not know what the rate of inflation will be, the ex ante (from before) rate is the interest rate minus the expected rate of future inflation. In many cases the expected rate of inflation can be approximated by the current rate of inflation, which you know, so the ex ante real interest rate is the nominal interest rate minus the current rate of inflation. So the ex ante real interest rate is EARIR=$3.5 - 33.3 = -29.8$. At the beginning the ex ante real rate was negative because inflation was higher than the nominal interest rate, subsequently with quite high nominal rates and deflation (negative inflation) real rates became very high.

(d) The statement is true; the combination of sharply reduced military spending and high real interests rate caused deflation (falling prices), falling output, rising unemployment and after 1920 a strengthening of the exchange rate. The Chancellor of the Exchequer, Winston Churchill returned sterling to the gold standard at its pre-war parity in 1925. Keynes blamed this policy for the depression of the early 1920s.

7.7. Example: Were the nineties and noughties NICE?

7.7.1. Introduction

Mervyn King, the Governor of the Bank of England, described the UK economic environment at the end of the 20th century and the beginning of the 21st century as NICE: non-inflationary, consistently expanding. Subsequently it became VILE: volatile inflation, less expansion. This example uses descriptive statistics and graphs to compare UK growth and inflation over the period 1992-2007, with their earlier behaviour to see how nice this period was.

7.7.2. Data

The original series, from the Office of National Statistics, are for 1955Q1-2009Q2, Q_t = Gross Domestic Product, chained volume measure, constant 2003 prices, seasonally adjusted (ABMI) and 1955Q1-2009Q1 E_t = Gross Domestic Product at market prices: Current price: Seasonally adjusted (YBHA). The price index, the GDP deflator, is $P_t = E_t/Q_t$. Growth (the percentage change in output), g_t, and inflation (the percentage change in prices), π_t, are measured over the same quarter in the previous year as:

$$\begin{aligned} g_t &= 100 * (\ln(Q_t) - \ln(Q_{t-4})) \\ \pi_t &= 100 * (\ln(P_t) - \ln(P_{t-4})). \end{aligned}$$

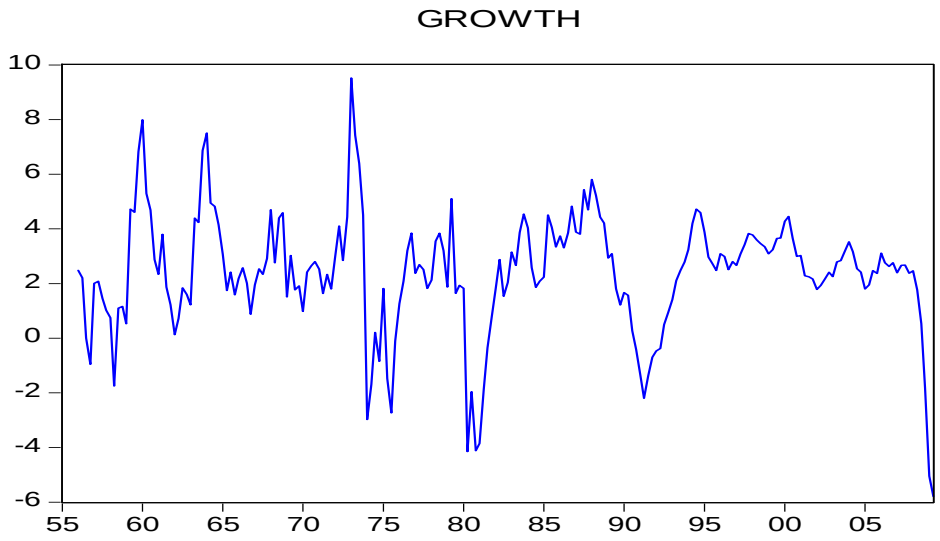
Such annual differences smooth the series and would remove seasonality if they were not already seasonally adjusted. Notice that by taking the four quarter change, whereas the data for ouput and prices starts in 1955Q1, the data for growth and inflation only starts in 1956Q1.

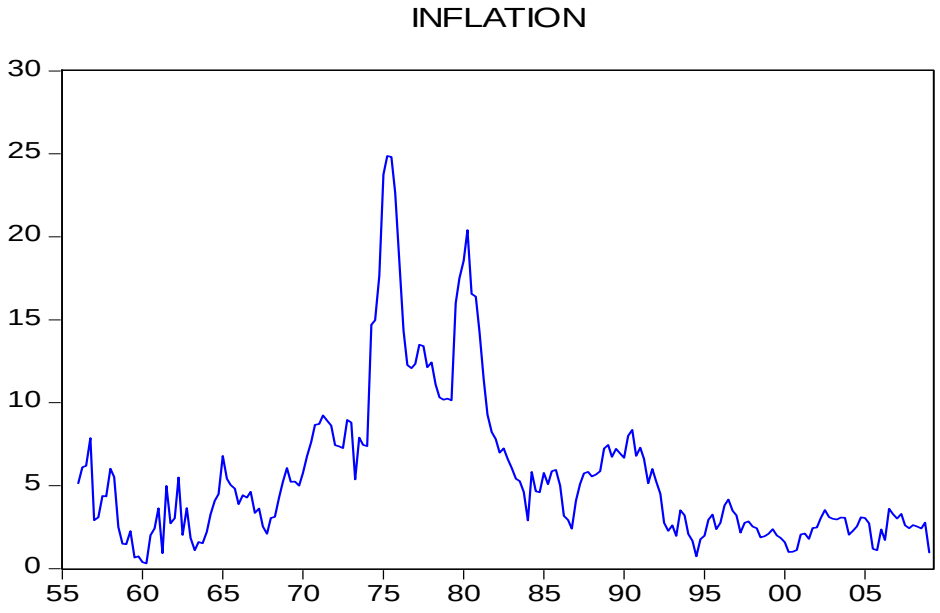
7.7.3. Line graphs

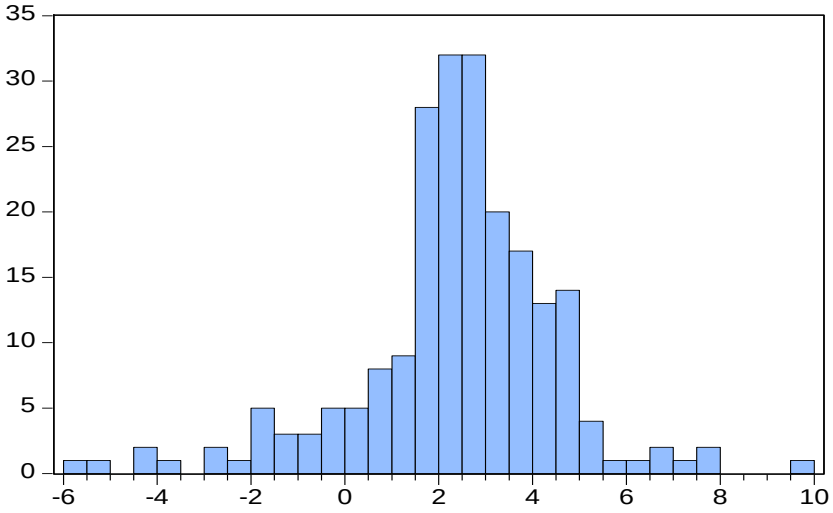
The broad pattern of UK economic events can be seen from the from the line graphs below for the time series of growth and inflation. UK economic policy since World War II can be described as a search for targets. Initially the target was the balance of payments to support a fixed exchange rate. Then with the end of the Bretton Woods system of fixed exchange rates in the early 1970s, there was a shift to monetary targets to control inflation, a shift which became more pronounced with the election of Mrs Thatcher in 1979. However, the monetary aggregates proved very unstable and there was a switch to exchange rate targets in

the middle 1980s, culminating in joining the European Exchange Rate Mechanism. With the ejection of sterling from the ERM in September 1992, inflation targets were adopted. The Bank of England was given independent responsibility for targetting inflation, when the Labour Government was elected in 1997. This history is reflected in the graphs for growth and inflation below. The "stop-go" pattern of the 1950s and 1960s is obvious, then there is a peak when growth reached almost 10% during the "Barber Boom".of the early 1970s following the collapse of the Bretton Woods system of fixed exchange rates. Anthony Barber was the conservative Chancellor at the time. The first oil price shock of 1973 following the Arab-Israeli war sent the economy into deep recession, with growth negative in most quarters between 1974Q1 and 1975Q4. During 1976 the UK had to borrow from the IMF. Growth recovered in the later 1970s, before a further recession in the early 1980s following the second oil price shock after the Iranian revolution and Mrs Thatcher's monetarist policies. Growth recovered in the later 1980s with the boom under Nigel Lawson, the Conservative Chancellor, then sank into recession again in the early 1990s, possibly worsened by the fixed exchange rate required by membership of the European Exchange Rate Mechanism. The UK left the ERM in September 1992 and adopted inflation targetting, with independence of the Bank of England in 1997. There was then a period of relative stability, before the effects of the 2007 Credit Crunch began to impact on the economy. Output fell by 5.8% in the year up to 2009Q2, the lowest observed in this sample; but the 2009 figures are likely to be revised as more data becomes available.

Inflation was fairly low, below 10%, though volatile during the 1960s and 1970s. In the mid 1970s it shot up to almost 25%, before falling back to almost 10%, then rising again over 20% following the election of Mrs Thatcher in 1979. It then came down below 5%, with a burst in the late 1980s and early 1990s, before stabilising at a low level subsequently. There are a number of different measures of inflation, CPI, RPI etc., and they show slightly different patterns from the GDP deflator used here.


GROWTH
10
8
6
4
2
0
-2
-4
-6
55
60
65
70
75
80
85
90
95
00
05

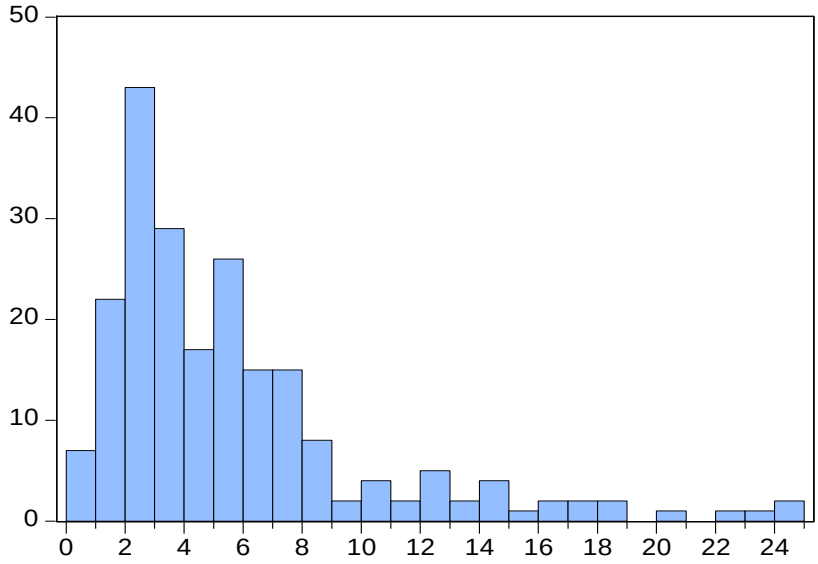

INFLATION
30
25
20
15
10
5
0
55
60
65
70
75
80
85
90
95
00
05



Series: GROWTH	
Sample 1955Q1 2009Q2	
Observations 214	
Mean	2.374050
Median	2.508702
Maximum	9.516651
Minimum	-5.795526
Std. Dev.	2.161040
Skewness	-0.678412
Kurtosis	5.299602
Jarque-Bera	63.56818
Probability	0.000000

7.7.4. Frequency distributions.

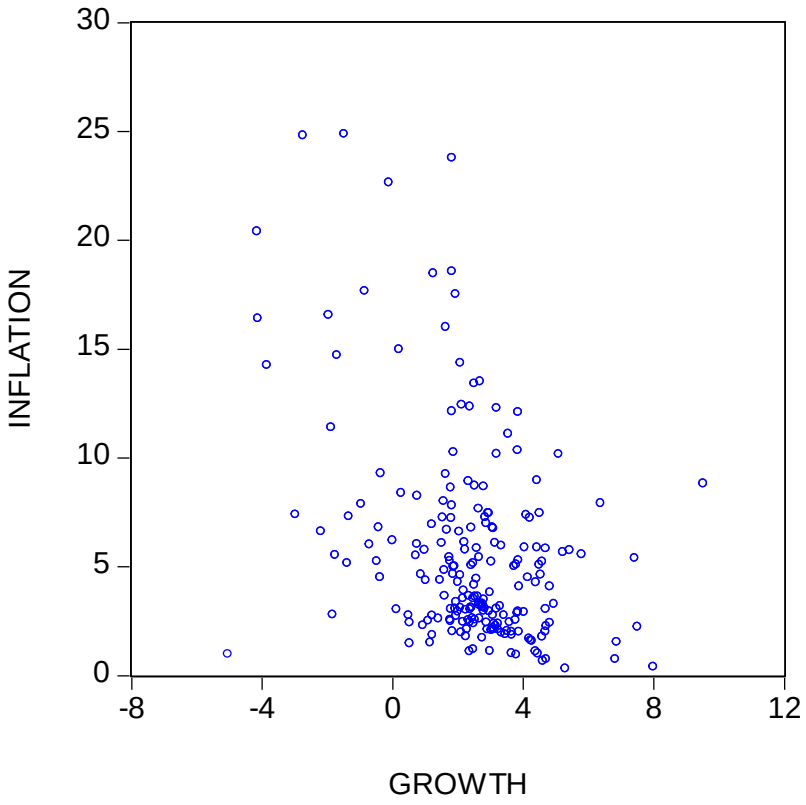
The histograms and summary statistics which describe the distribution of growth and inflation for the whole period are next shown. An average growth rate of about 2.5%, seems to be a long-term feature of the British economy, the mode is between 1.5-3%. The distribution is not quite normal, with a slight negative skew and slightly fatter tails than a normal distribution. If a random variable is normally distributed, the coefficient of skewness should be zero and the coefficient of kurtosis 3. The Jarque-Bera statistic , which is distributed as $\chi^2(2)$, tests whether the skewness and kurtosis are significantly different from these values. The p value, probability of the null hypothesis of normality being true, indicate that the hypothesis that growth is normal is rejected. If growth was normally distributed one would be very unlikely to observe a value of -5.8% as in 2009Q2, which is almost 4 standard deviations from the mean. Inflation is even less normal than growth as the larger value of the J-B statistic indicates and is strongly positively skewed as both the difference between the mean and the median and the coefficient of skewness show. The average of around 5% is not really typical of the whole period, the mode is between 2% and 3%.



Series: INFLATION	
Sample 1955Q1 2009Q2	
Observations 213	
Mean	5.656730
Median	4.360406
Maximum	24.87334
Minimum	0.310697
Std. Dev.	4.690943
Skewness	1.905004
Kurtosis	6.913777
Jarque-Bera	264.7751
Probability	0.000000

7.7.5. The relationship between growth and inflation

The scatter diagram below plots inflation against growth for the whole period. There is a negative relationship, the correlation is -0.4, but it is not close, and the association comes mainly from the periods of crisis, with high inflation and low growth in the top left hand quadrant of the scatter. Any relationship between output and inflation is complicated by slow adjustment and the influence of other factors. Notice that the patterns described by the graphs in this section are specific to this period and to the UK; other countries and other periods may show other features.


30
25
20
15
10
5
0
INFLATION
-8
-4
0
4
8
12
GROWTH

7.7.6. Differences between periods

We shall divide the sample into period A, the 37 years 1956-1992, and period B the 15 years 1993-2007. The table gives mean, median, standard deviation, coefficients of skewness and kurtosis and the number of observations for growth and inflation over these two periods Relative to the first period, inflation was much lower and much less volatile, while growth was slightly higher and much less volatile in the second period. The standard error of mean growth in period A is $2.3/\sqrt{148} \approx 0.18$ and in period B is $0.7/\sqrt{60} \approx 0.0.09$. So the 95% confidence interval for the period A mean growth is $2.3 \pm 2(0.18)$, that is 2.66 to1.94, while that for period B is $2.9 \pm 2(0.09)$ that is 3.08 to 2.72. They do not overlap.

Growth and inflation, A: 1956-1992, B: 1993-2007

	$Infl$		Gr	
periods	A	B	A	B
$Mean$	7.1	2.5	2.3	2.9
$Median$	5.8	2.5	2.2	2.9
$St.Dev$	5.0	0.7	2.3	0.7
$Skew$	1.5	-0.2	-0.2	0.4
$Kurt$	5.4	2.7	3.8	3.0
$NObs$	148	60	148	60

7.7.7. Conclusion

Compared to previous (and subsequent?) history, the period 1993-2007 was nice. From being high and volatile, inflation became low and stable. Growth was slightly higher and much less volatile. Although there was an economic cycle 1997-2007, it was less pronounced than in earlier years. Thus there is some basis for the claim by Gordon Brown, Chancellor then Prime Minister, to have abolished boom and bust. Whether this was a matter of luck or good policy remains a matter of debate, as does whether the easy-money policy of these years contributed to the subsequent financial crisis. This "Great Moderation" was not confined to the UK, but seems to have been a general feature of many advanced economies. There were global economics shocks over this period: the Asian crisis of 1997-8; the Russian default and the LTCM crisis of 1998; the dot.com boom and bust of 2001; the gyrations in the oil price, which went from around $10 in 1998 to $147 in 2008; and the 9/11 attacks and the wars in Iraq and Afghanistan. But despite these shocks, there was smooth non-inflationary growth in the UK, as in many economies. Whether the Great Moderation was merely a transitory interlude of

stability in a crisis-prone system remains to be seen, as the warning on financial products says: past performance is not necessarily a guide to the future.

8. Applied exercise I: Ratios and descriptive statistics

This exercise is to be done on your own. We will assume that you have done this exercise and are familiar with this material, so we can set exam questions using it. It will be started in the spreadsheet class. The data are available on the ASE web page. They were constructed by Robert J. Shiller, updating the series in Robert J Shiller, *Market Volatility*, MIT Press 1989. His later book, *Irrational Exhuberance*,Princeton University Press 2000 is a good description of the stock market bubble at the end of the twentieth century. At the end of this data, January 2000, the S&P was 1425, it peaked in March 2000 at 1527 He got timing of his publication right, since it then fell sharply and was around 1000 in late August 2003, with a 2003 low as of then of 800 (taken from FT World Markets at a glance table). Check what it has done since; bubbles (when prices shoot up to unsustainable levels) and financial crises (when they return to normal) are recurrent features of the system.

8.1. Data

This will be in an Excel file called Shiller.xls. Copy this file to your own directory and when you have finished the exercise copy the new file to your own disk using a new name.

In the file, rows 2 to 131, contain US data from 1871 to 2000; row 1 contains headings:

A-YEAR

B-NSP: Stock Price Index (Standard & Poors composite S&P) in current prices (Nominal terms), January figure.

C-ND: Dividends per share in current prices, average for the year.

D-NE: Earnings per share in current prices, average for the year.

E-R: Interest rate, average for the year.

F-PPI:The producer price index, January figure, 1967 average=100.

The letters A to F indicate the columns in the spreadsheet. Note 2000 data are missing for ND, NE, R.

8.2. Transformations: ratios, growth rates, correcting for inflation, etc

In column G construct the (backwards) Price Earnings Ratio. Type PER in cell G1 as a heading. Put **=B3/D2** in cell G3. Copy this down for the rest of the years ending in G131. The price earnings ratio is a measure of the underlying value of a share, how much you have to pay to buy a stream of earnings.

Highlight the data for PER over G3:G131. Use Chart wizard (on toolbar) and choose a line graph, top left sub-type, choose next and go to next. Comment on the picture. Can you identify the Stock Market Booms of the late 1920s and 1990s. What happened next in both cases? What other features of the data can you see?

In the same way, create the following new variables for the period 1872 to 1999, i.e. rows 3 to 130, with the variable name at the top:

H- Dividend Yield: DY (Type DY in cell H1 and type the formula **=C3/B3** in cell H3 and copy down)

I-Capital Gain: CG **(type CG in I1) type the formula =(B4-B3)/B3 in I3 and copy down**

J-Inflation:INF**=(F4-F3)/F3 type formula in J3**

K-Real Return on equities: RRE**=H3+I3-J3**

L- Real Interest Rate RIR=**(E3/100)-J3**.

Notice that these are proportions, e.g. numbers like 0.05. This corresponds to 5%. Why do we subtract the rate of inflation to get the real return, whereas we would divide the stock price index by the price level to get the real stock price index? Why do we divide the interest rate by 100? Plot inflation. Notice how we have matched the dates, e.g. in defining inflation, we have had to take account of the fact that the price index is a January figure.

8.3. Graphing Real Returns on Equities

We now want to analyse real return on equities (i.e. dividends plus capital gains less inflation) over the period 1872-1999 and compare it with the real interest rate. First we will do it graphically. Highlight K3:L130, click chart wizard, choose line graph and view real returns and the real interest rate. There is a lot of volatility from year to year in the RRE and the range is large: from about -0.5 to +0.5. Over this period, you could have made or lost almost 50% of your investment in exceptional years. The real interest rate is much less volatile. Go to cell J132 type in mean, then K132 and type in **=AVERAGE(K3:K130).** Go to J 133 type in SD, then K133 and type in **=STDEV(K3:K130).** Copy these two cells

to L132 and 133. You now have the mean (average) and the (unbiased estimate of) the standard deviation for the real return on equities and the real interest rate. The mean real return on equities is much higher than that from the interest rate. This is known as the equity premium. But the risk on equities, measured by the standard deviation, is also much higher. Calculate the average real return on equities, 1872-1990 and 1991-1999.

8.4. More Descriptive Statistics

We want to calculate descriptive statistics for the array K3:J130. Following the principles above construct the following table for RRE and RIR. Only the figures for RRE are given below.

line	J	K	RRE
132	**Mean**	**=AVERAGE(RR)**	0.088
133	**SD**	**=STDEV(RR)**	0.184
134	**MIN**	**=QUARTILE(k3:k130,0)**	-0.448
135	**25%**	**=QUARTILE(k3:k130,1)**	-0.021
136	**MEDIAN**	**=QUARTILE(k3:k130,2)**	0.091
137	**75%**	**=QUARTILE(k3:k130,3)**	0.219
138	**MAX**	**=QUARTILE(k3:k130,4)**	0.514
139	**SKEW**	**=SKEW(k3:k130)**	-0.389
140	**KURTOSIS**	**=KURT(k3:k130)**	0.296

You can also use the descriptive statistics command in data analysis to get most of these.

8.5. Interpretation of spreadsheet summary statistics

Call the series y_t, $t = 1, 2, ..., T$. Above we calculate the mean:

$$\overline{y} = \sum_{i=1}^{T} y_t / T$$

the unbiased 'sample' standard deviation, SD:

$$s = \sqrt{\sum_{i=1}^{T} (y_t - \overline{y})^2 / (T-1)}$$

The amount $T - 1$ is known as the degrees of freedom.

Then we calculate the minimum value; the value which 25% of the observations lie below; the median, the value which 50% of the observations lie below; the value which 75% of the observations lie below; and the maximum. These are known as Quartiles. Excel also allows you to calculate percentiles: the value for which $x\%$ lie below, for any x. Returns were negative in over 25% of the years. The median is very similar to the mean which suggests that the distribution is symmetric. The range of real returns is very large, between minus 50% and plus 50%.

The measure of skewness is roughly

$$\frac{1}{n}\sum_{i=1}^{T}\left(\frac{(y_t-\overline{y})}{s}\right)^3.$$

the standarised third moment. In fact Excel makes degrees of freedom adjustments, similar to the sample standard deviation above. If the distribution is symmetrical, the measure of skewness should be zero. In this case, it is pretty near zero.

The measure of (excess) kurtosis is roughly

$$\frac{1}{n}\sum_{i=1}^{T}\left(\frac{(y_t-\overline{y})}{s}\right)^4-3.$$

if the distribution is normal the expected value of the first term (the fourth standardised centred moment) is three, so values around zero indicate a roughly normal distribution. You can get exact definitions from HELP, STATISTICAL FUNCTIONS, SKEW & KURT.

9. Index Numbers

9.1. Introduction

Inflation, the growth rate of the price level, is measured by percentage change in a price index:

$$INF_t = 100 * (P_t - P_{t-1})/P_{t-1} = 100 * \{(P_t/P_{t-1}) - 1\} \approx 100 * (\ln P_t - \ln P_{t-1})$$

where P_t is a price index. There are a lot of different price indexes. In the past the Bank of England had a target for inflation in the Retail Price Index excluding mortgage interest payments (which go up when interest rates are raised) RPIX of 2.5%. In 2004 this was replaced by a 2% target for inflation in the Consumer Price Index, CPI. This was previously known as the Harmonised Index of Consumer Prices HICP, the type of index the European Central Bank uses. There are two main differences. One is in the method of construction. RPI uses arithmetic means, CPI uses geometric means. This difference makes the RPI run about 0.5% higher, hence the reduction in target from 2.5% to 2%. The other is that the CPI excludes housing. In August 2003 RPIX was 2.9%, CPI 1.3%, most of the difference accounted by the high rate of UK housing inflation, while in May 2009 the CPI was at +2.2% and the RPI at -1.1%, deflation, not inflation, because of falling house prices. The RPI and CPI measure consumer prices, the GDP deflator, another price index used in section 7.7, measures prices in the whole economy.

Distinguish between the price level and the rate of inflation. When the inflation rate falls, but is still positive, prices are still going up, just at a slower rate. If inflation is negative, prices are falling. Suppose the Price Index was 157 in 1995 and 163 in 1996, then the rate of inflation is 3.82%. Notice that this can also be expressed as a proportion, 0.0382. In many cases, we will calculate the growth rate by the change in the logarithm, which is very close to the proportionate change for small changes, e.g. < 0.1, i.e.10%. We usually work with natural logs to the base e, often denoted by LN rather than LOG, sometimes used just for base 10. Price indexes are arbitrarily set at 100, or 1, in some base year, so the indexes themselves cannot be compared across countries. The index can be used to compare growth relative to the base year if they all have the same base year, e.g. 1990=100 for all countries.

If the inflation rate rises from 3% to 6% it has risen by three percentage points. It has not risen by three percent, in fact it has risen by 100%. If something falls

by 50% and then rises by 50%, it does not get back to where it started. If you started at 100, it would fall to 50, then rise by 50% of 50, 25, to get to 75.

9.2. Prices and Quantities, Real and nominal variables

Suppose that a firm buys 2 million barrels of oil in 2003 at \$35 a barrel and one million in 2004 at \$40 a barrel, we can denote the price in 2003 as P_t, and the price in 2004 as P_{t+1} both measured in dollars. Similarly the quantities are Q_t and Q_{t+1}, both measured in million barrels. Total expenditure on oil in each year is $E_t = P_t Q_t$ and $E_{t+1} = P_{t+1}Q_{t+1}$, both measured in million dollars.

	P_t	Q_t	E_t
t	35	2	70
$t+1$	40	1	40

The change in expenditure from \$70m to \$40m, reflects both a 14.3% increase in price and a 50% fall in quantity. Notice that we need only two of the three pieces of information in the table for each year. Above knowing price and quantity we could work out expenditure. If we knew expenditure and quantity we could always work out price as $P_t = E_t/Q_t$. If we knew expenditure and price, we could work out quantity as $Q_t = E_t/P_t$.

Often we work with logarithms, where the proportionate change in expenditure can be decomposed into the sum of proportionate changes in price and quantity

$$\begin{aligned}
\Delta \ln E_t &= \ln E_t - \ln E_{t-1} \\
&= (\ln P_t + \ln Q_t) - (\ln P_{t-1} + \ln Q_{t-1}) \\
&= (\ln P_t - \ln P_{t-1}) + (\ln Q_t - \ln Q_{t-1}) \\
&= \Delta \ln P_t + \Delta \ln Q_t
\end{aligned}$$

Notice that the formulae would be more complicated if we worked with the original values

$$\begin{aligned}
\Delta E_t &= P_t Q_t - P_{t-1}Q_{t-1} \\
&= (P_{t-1} + \Delta P_t)(Q_{t-1} + \Delta Q_t) - P_{t-1}Q_{t-1} \\
&= P_{t-1}Q_{t-1} + P_{t-1}\Delta Q_{t-1} + Q_{t-1}\Delta P_{t-1} + \Delta P_{t-1}\Delta Q_{t-1} - P_{t-1}Q_{t-1} \\
&= P_{t-1}\Delta Q_{t-1} + Q_{t-1}\Delta P_{t-1} + \Delta P_{t-1}\Delta Q_{t-1}.
\end{aligned}$$

The change in quantity measured at last years prices, plus the change in prices measured at last years quantities plus an interaction term. The easiest way to

present this is on a graph with price and quantity on the two axes. Revenue is then the area of the rectangle, price times quantity. Draw the two rectangles for years t and $t-1$. The difference between their areas will be made up of the three components of the final equation.

Most of the time, we are not dealing with a single good, but with aggregates of goods, so that total expenditure is the sum of the prices times the quantities of the different goods, $i = 1, 2, ..., N$ whose prices and quantities change over time.

$$E_t = \sum_{i=1}^{n} p_{it} q_{it}.$$

This is like your supermarket receipt for one week, it lists how much of each item bought at each price and the total spent. To provide a measure of quantity, we hold prices constant at some base year, 0, say 2000 and then our quantity or constant price measure is

$$Q_t = \sum_{i=1}^{n} p_{i0} q_{it}.$$

Monetary series can be either in nominal terms (in the current prices of the time, like expenditures) or in real terms (in the constant prices of some base year to correct for inflation, to measure quantities). To convert a nominal series into a real series it is divided by a price index. So if we call nominal GDP E_t and real GDP Q_t, and the price index P_t then $E_t = P_t Q_t$. So given data on nominal (current price) GDP and a price index we can calculate real (constant price) GDP as $Q_t = Y_t/P_t$, where P_t is the value of a price index. Alternatively if we have data on current price (nominal) and constant price (real) GDP, we can calculate the price index (usually called the implicit deflator) as the ratio of the current to constant price series: $P_t = Y_t/Q_t$.

Most statistical sources only give two of the three of the possible series, nominal, real, price, assuming (somewhat implausibly) that users will know how to calculate the third from the other two.

9.3. Price Indexes

Suppose we wish to measure how the prices of a set of goods, $i = 1, 2, ..., N$ have moved over time, $t = 0, 1, 2, ..., T$, (e.g. 1990,1991,1992). We observe the prices, p_{it} and quantities, q_{it} of each good in each year. Total expenditure on all goods

in year t, e.g. current price GDP, is $E_t = \sum_{i=1}^{N} p_{it}q_{it}$. We could also express this as an index, relative to its value in some base year:

$$E_t^I = \frac{\sum_{i=1}^{N} p_{it}q_{it}}{\sum_{i=1}^{N} p_{i0}q_{i0}}$$

here the index would be 1 in the base year, usually they are all multiplied by 100 to make them 100 in the base year. If the base is 100, then $E_t^I - 100$ gives the percentage change between the base year and year t. Index numbers are 'unit free'. This is an expenditure index.

A constant price series would measure quantities all evaluated in the same base year prices. Suppose we used year zero, then the constant price measure of quantity would be

$$Q_t = \sum_{i=1}^{N} p_{i0}q_{it}.$$

Constant price GDP was a measure of this form, where the base year was changed every five years or so. Recently this fixed base approach has been relaced by a moving base called a chain-weighted measure.

We can construct a price index as the ratio of the expenditure series to the constant price series (in the case of GDP, this would be called the GDP deflator)

$$P_t^1 = \frac{E_t}{Q_t} = \frac{\sum_{i=1}^{N} p_{it}q_{it}}{\sum_{i=1}^{N} p_{i0}q_{it}}.$$

It measures prices in year t relative to prices in year zero, using quantities in year t as weights. Where $t = 0$, $P_t^1 = 1$. The index always equals 1 (or 100) in its base year. This is a price index.

We could also use quantities in year zero as weights, and this would give a different price index.

$$P_t^2 = \frac{\sum_{i=1}^{N} p_{it}q_{i0}}{\sum_{i=1}^{N} p_{i0}q_{i0}}.$$

Notice that these will give different measures of the price change over the period 0 to t. In particular, for goods that go up (down) in price, quantities in year t are likely to be lower (higher) than in year 0. Indexes that use beginning of the period values as weights are called Laspeyres indexes, those that use end of period values are called Paasche indexes. There are a range of other ways we could calculate price indexes; chain indexes use moving weights. Apart from the problem

of choosing an appropriate formula, there are also problems of measurement; in particular, measuring the quantities of services supplied, accounting for quality change and the introduction of new goods. Barrow Chapter 2 discusses index numbers.

You will often find that you have overlapping data. For instance, one edition of your source gives a current price series and a constant price series in 1980 prices for 1980 to 1990; the second gives you a current price series and a constant price series in 1985 prices for 1985 to 1995. This raises two problems. Firstly the current price series may have been revised. Use the later data where it is available and the earlier data where it is not. Secondly, you have to convert the data to a common price basis. To convert them, calculate the ratio in 1985 (the earliest year of the later source) of the 1985 constant price series to the 1980 constant price series; then multiply the earlier 1980 price series by this ratio to convert the 1980 constant price series to 1985 constant prices. If the two estimates of the current price series for 1985 were very different, you would also have to adjust for the ratio of the current price series.

9.3.1. Example Substitution

In 2000 a company bought 10 computers at £2000 each and 20 software licenses at £1000 each. In 2001 it bought 20 computers at £1000 each and 10 software licenses at £2000 each.

(a) What were its total computing costs in each year?

(b) What would have been its total computing costs in each year (i) if it had bought the 2000 quantities (ii) if it had bought the 2001 quantities?

(c) Use the estimates in (b) to calculate two measures of inflation (i) using 2000 quantities as weights and (ii) using 2001 quantities as weights.

(d) Comment on your results.

Answer

This example is a little extreme to indicate the effects substitution can have on the measurement of inflation.

(a) Total expenditure was £40,000 in both years: 2000=(10×2000+20×1000); 2001=($20 \times 1000 + 10 \times 2000$).

(b) (i) Using 2000 quantitities and 2000 prices, expenditure in 2000 would have been £40,000, which it was. Using 2000 quantities and 2001 prices expenditures in 2001 would have been $(10 \times 1000 + 20 \times 2000)$= 50,000 (ii) Similarly using 2001 quantities, 2000=50,000 and 2001=40,000

(c) Using 2000 quantities as weights inflation is 25%=100(50,000/40,000-1) , using 2001 quantities as weights inflation is –20% =100(40,000/50,000-1).

(d) Because of demand responses to price (the firm bought more hardware which had fallen in price and less software which had risen in price), base weighted measures tend to overestimate inflation (+25%) and terminal weighted measures tend to underestimate it (-20%). The truth lies somewhere in between.

9.3.2. Example House Prices

There are a number of different indexes of house-prices, which can show very different trends. A major difference is the stage of the transaction at which they measure the price. Some measure the asking price of houses put on the market. This provides early information, but they may sell for more or less than the initial asking price. The Building Society series are based on mortgage approvals, again this may not reflect the final price and about 25% of houses are paid for in cash and are not captured in these series since their purchase does not require a mortgage. Other series use surveys of completions (when the sale is completed). The Land Registry figure is based on when the transaction is registered, and covers the final price for all housing transactions. The gap between the house being put on the market and the transaction registered can be over six months. The indexes also differ in (a) whether they adjust for the mix of transactions; in unadjusted series average price will jump if there is the sale of a very expensive house (b) how often they are published and how long it takes to publish the data (c) whether they are seasonally adjusted. House prices are usually compared to average earnings, with a normal 20th century UK ratio about 3.5. In the 21st century the ratio rose higher than any previous peak, before dropping back.

9.3.3. Example Stock market indexes

See section 8 for the measurement of the returns on equities using the Standard & Poors (S&P) index. Here we will consider the measurement of average market return.

The most obvious definition of the value of the stock market is the total value of all the stocks quoted (the sum over firms of the price per share times the number of shares issued by the firm) $\sum_{i=1}^{N} V_{it}$, where there are N firms, with values V_{it}. This is scaled by the value in some base year to construct an index. This is how the S&P 500 is calculated. On this measure between $t = 0, (31/3/00)$ and

$t = T, (31/7/03)$ the S&P index fell 30.5%:

$$R_1 = \frac{\sum_{i=1}^{N} V_{iT} - \sum_{i=1}^{N} V_{i0}}{\sum_{i=1}^{N} V_{i0}} = -0.305.$$

An alternative measure is to take the average of the returns on the individual stocks:

$$R_2 = \sum_{i=1}^{N} (\frac{V_{iT} - V_{i0}}{V_{i0}})/N = 0.12.$$

Over the same period this measure shows an increase of 12%. A third measure, using a geometric average shows a fall of 18% (New York Times 31/8/03 B6).

R_2 is an equally weighted average of the individual stock's returns, R_1 is weighted by their size. To see this note that

$$1 + R_1 = \frac{\sum_{i=1}^{N} V_{iT}}{\sum_{i=1}^{N} V_{i0}} = \sum_{i=1}^{N} \left[\frac{V_{iO}}{\sum_{i=1}^{N} V_{i0}} \right] \frac{V_{iT}}{V_{i0}}$$

where the terms in [..] are the weights, the share of the market accounted for by firm i, in the base year. In $1 + R_2$ each of the weights is just $1/N$.

Most indexes are weighted by market capitalisation, but other forms of weighting are becoming more widespread, e.g. 'fundamental indices', which use measures like the firms revenues.

9.3.4. Example Index linked weapons contracts

Weapons procurement contracts often cover long periods, because it takes many years to develop and get them into production. Eurofighter/Typhoon, which only entered service in 2004, was started in the early 1980s. In such contracts it is common to link the agreed price to inflation, the issue then becomes which index to use. On the EH101 helicopter, the prime contractor, IBM at the time proposed a simple materials and fuels index (essentially an output price index). The UK Ministry of Defence insisted on the use of a combined costs index reflecting input costs including labour. Because of productivity growth output price indexes grow more slowly than input cost indexes. The National Audit Office, in its report *Accounting for Inflation in Defence Procurement*, para 2.25, December 1993, calculated that had the MOD used the index suggested by IBM, rather than the one it had insisted on, it could have saved itself £95 million or about 6%

of the contract price over the lifetime of the contract. This particular over-spend got very little publicity because most journalists and MPs tend to fall asleep once index numbers are mentioned.

To see the relation between prices and wages, write the total value of sales (price times quantity) as a markup on labour costs, wages times number employed

$$P_t Q_t = (1 + \mu) W_t E_t$$

prices are then a mark-up on unit labour costs

$$P_t = (1 + \mu) W_t E_t / Q_t$$

and noting that productivity is output per worker Q_t/E_t

$$\ln P_t = \ln(1 + \mu) + \ln W_t - \ln(Q_t/E_t)$$

so if mark-ups are constant, output price inflation is the rate of growth of wages minus the rate of growth of productivity:

$$\Delta \ln P_t = \Delta \ln W_t - \Delta \ln(Q_t/E_t)$$

10. Probability

10.1. Introduction

We need to analyse cases where we do not know what is going to happen: where there are risks, randomness, chances, hazards, gambles, etc. Probabilities provide a way of doing this. Some distinguish between (a) risk: the future is unknown but you can assign probabilities to the set of possible events that may happen; (b) uncertainty: you know the set of possible events but cannot assign probabilities to them; and (c) unawareness where you cannot even describe the set of possible events, what US Defense Secretary Donald Rumsfeld called the unknown unknowns, the things you do not even know that you do not know about. People seem to have difficulty with probabilities and it is a relatively recent branch of mathematics, Nobody seems to have regarded probabilities as things that could be calculated before about 1650 (after calculus) and the axiomatic foundations of probability theory were only provided in the 1930s by the Russian Mathematician Kolmogorov.

Probabilities are numbers between zero and one, which represent the chance of an event happening. Barrow chapter 2 discusses them. If an event is certain to happen, it has probability one; if an event is certain not to happen, it has probability zero. It is said that only death and taxes are certain, everything else is uncertain. Probabilities can either represent degrees of belief, or be based on relative frequency, the proportion of times an event happens. So if in past horse races the favourite (the horse with the highest probability, the shortest odds offered by bookmakers) won a quarter of the time, you might say the probability of the favourite winning was 0.25; this is a relative frequency estimate. Alternatively you could look at a particular future race, study the history (form) of the horses and guess the probability of the favourite in that race winning, this is a degree of belief estimate. You bet on the favourite if your estimate of the probability of the favourite winning is greater than the bookmakers estimate, expressed in the odds offered; the odds are the ratio of the probability to one minus the probability. There is a large literature on the economics and statistics of betting. Notice that although the probabilities of the possible events should add up to one (it is certain that some horse will win the race), the implied probabilities in the odds offered by bookmakers do not. That is how they make money on average. There are also systematic biases. For instance, the probability of the favourite winning is usually slightly better than the bookmaker's odds suggest and the the probability of an outsider slightly worse. This favourite-longshot bias has been noted for over 60

years in a variety of horse-races, but its explanation is still subject to dispute.

If you throw a dice (one dice is sometimes known as a die) there are six possible outcomes, 1 to 6, and if the die is fair each outcome has an equal chance; so the probability of any particular number is 1/6. On one throw you can only get one number, so the probability of getting both a 3 and a 4 on a single throw is zero, it cannot happen. Events which cannot both happen (where the probability of both happening is zero) are said to be mutually exclusive. For mutually exclusive events, the probability of one or the other happening is just the sum of their probabilities, so the probability of getting either a 3 or a 4 on one throw of a dice is 1/6+1/6=2/6=1/3.

Suppose two people, say A and B, each throw a dice the number B gets is independent of the number A gets. The result of A's throw does not influence B's throw. The probability of two independent events happening is the product of their probabilities. So the probability of both A and B getting a 3 is $1/6 \times 1/6 = 1/36$. There are 36 (6^2) possible outcomes and each are equally likely. The 36 outcomes are shown in the grid below, with the six cases where A and B get an equal score shown in bold. So there is a probability of $6/36 = 1/6$ of a draw. We can also use the grid to estimate the probability of A getting a higher score than B. These events correspond to the 15 events above the diagonal, so the probability of A winning is $15/36 = 5/12$; the probability of B winning is also 5/12 and the probability of them getting an equal score, is $1/6 = 2/12$. Notice the 3 events (A wins, B wins, a draw) are mutually exclusive and their probabilities sum to one, 12/12.

		A					
		1	2	3	4	5	6
	1	**x**	x	x	x	x	x
	2	x	**x**	x	x	x	x
B	3	x	x	**x**	x	x	x
	4	x	x	x	**x**	x	x
	5	x	x	x	x	**x**	x
	6	x	x	x	x	x	**x**

When events are not mutually exclusive, one has to allow for the probability of both events happening. This seems to have been first pointed out by Bernouilli in his *Ars cojectandi* in 1713, with a gruesome example. "If two persons sentenced to death are ordered to throw dice under the condition that the one who gets the smaller number of points will be executed, while he who gets the larger number will be spared, and both will be spared if the number of points are the same,

we find that the expectation of one of them is 7/12. It does not follow that the other has an expectation of 5/12, for clearly each of them has the same chance, so the second man has an expectation of 7/12, which would give the two of them of them an expectation of 7/6 of life, i.e. more than the whole life. The reason is that there is no outcome such that at least one of them is not spared, while there are several in which both are spared."[1] A will win 5/12 times, draw 2/12 times, so survives 7/12 times. Similarly for B. The probability of at least one surviving is the sum of the probability of each surviving minus the probability of both surviving: $5/6+5/6-1/6 = 1$. The probability of both has to be substracted to stop double counting. Check that the probability of getting either a 3 or a 4 on two throws of a dice is $1/3 + 1/3 - 1/9 = 20/36$. Notice this is different from the probability of getting either a 3 or a 4 on both throws of the dice, which is $(1/3)^2 = 1/9$. You must be careful about exactly how probability events are described.

10.2. Some rules

Denote the probability of event A happening as $P(A)$. Then the probability of event A not happening is $1 - P(A)$. This is called the complement of A, sometime written $\overline{A}$. If event A is certain to happen $P(A) = 1$. If event A cannot happen $P(A) = 0$. Denote both events A **AND** B happening as $P(A \cap B)$ (the intersection); this is often known as the joint probability. If the events are mutually exclusive they cannot happen together so $P(A \cap B) = 0$. Denote the probability of event A **OR** event B happening as $P(A \cup B)$ (the union). Then as we saw above with the dice:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

The probability of A or B equals the probability of A plus the probability of B minus the probability of both A and B happening. Notice that if the events are mutually exclusive, $P(A \cap B) = 0$ so $P(A \cup B) = P(A) + P(B)$. If the events are independent the joint probability, i.e. the probability of both happening; is

$$P(A \cap B) = P(A) \times P(B).$$

The probability of A happening given that event B has happened is called a conditional probability and is given by:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{10.1}$$

[1] Quoted by Ian Hacking, *The emergence of probability*, p144.

Below we will calculate the probability of winning the jackpot in the lottery. Strictly this is a conditional probability: the probability of an event A (winning the jackpot), given event B (buying a lottery ticket). Winning the jackpot and not buying a ticket are mutually exclusive events. Conditional probabilities play a very important role in decision making. They tell you how the information that B happened changes your estimate of the probability of A happening. If A and B are independent $P(A \mid B) = P(A)$, knowing that B happened does not change the probability of A happening. Similarly, the probability of B happening given that A happens is:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}. \tag{10.2}$$

Multiply both sides of (10.1) by $P(B)$ and both sides of (10.2) by $P(A)$, and rearrange to give

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

the joint probability is the product of the conditional probability and the marginal probability in each case. Using the two right hand side relations gives Bayes Theorem:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

This formula is widely used to update probabilities of an event A, in the light of new information, B. In this context $P(A)$ is called the prior probability of A, $P(B \mid A)$ is called the likelihood, and $P(A \mid B)$ is called the posterior probability.

10.2.1. Example: Screening

There is considerable controversy about the value of screening for diseases like breast cancer, prostate cancer or HIV. These are cases where the disease is not apparent but may be revealed by a test, which may not always give the correct answer. Call D having the disease, N not having the disease. Call TP the test showing positive (suggesting that you have the disease), TN the test showing negative (suggesting that you do not have the disease). In medical terminology the probability of the test showing positive given that you have the disease $P(TP \mid D)$ is called the sensitivity of the test. The probability of the test showing negative given that you do not have the disease $P(TN \mid N)$ is called the specificity of the test. Both are about 0.9 for a mamogram for breast cancer, and the higher they are the better. The probability of testing positive when you do not have the

disease $P(TP \mid N) = 1 - P(TN \mid N)$ is called the probability of a false positive. The probability of testing negative when you do have the disease, $P(TN \mid D) = 1 - P(TP \mid D)$ is the probability of a false negative. The fact that a decision can lead to two sorts of error, false positives and false negatives in this case, appears under a variety of different names in many areas.

Question

Suppose there is a disease which 1% of the population suffer from $P(D) = 0.01$. There is a test which is 99% accurate, i.e. 99% of those with the disease test positive and 99% of those without the disease test negative: $P(TP \mid D) = P(TN \mid N) = 0.99$. Suppose you test positive, what is the probability that you actually have the disease?

Answer

It is often simpler and clearer to work with numbers rather than probabilities and present the results as numbers. This is also often more useful for non-specialist audiences. Imagine a population of one hundred thousand: 100,000. Then a thousand ($1000 = 0.01 \times 100,000$) have the disease and 99,000 are healthy. Of those with the disease, 990 (0.99×1000) test positive, 10 test negative. Of those without the disease, 990 ($0.01 \times 99,000$) also test positive, 98010 test negative. Of the $2 \times 990 = 1980$ people who test positive, half have the disease, so the probability of having the disease given that you tested positive is 50%. Thus you should not worry too much about a positive result. A negative result is reassuring since only 10 out of 98020, who test negative have the disease. Positive results are usually followed up with other tests, biopsies, etc.

We could represent the joint and marginal frequencies as a table.

	D	N	
TP	990	990	1,980
TN	10	98,010	98020
	1,000	99,000	100,000

We could also calculate the conditional probability directly using Bayes Theorem

$$P(D \mid TP) = \frac{P(TP \mid D)P(D)}{P(TP)} = \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.01 \times 0.99} = 0.5$$

In practice screening is confined to groups where $P(D)$ is high to avoid this problem. The decision to establish a screening program depends on a judgement of the balance between (a) the benefits of detecting the disease, e.g. whether early treatment saves lives, (b) the costs of false positives: inappropriate treatment, worry

etc. and (c) the cost of testing, e.g. time off work to take the test. Since people disagree about these costs and benefits, screening is controversial. For instance, there is a test for an indicator of prostate cancer, PSA. The British Medical Association say "two-thirds of men with high PSA do not have prostate cancer, some men with prostate cancer do not have high PSA and no evidence exists to show whether treating localised prostate cancer does more harm than good". There is a profitable private industry in screening tests.

10.2.2. Example Innovation

Suppose that there is a survey of $1,000$ firms. Of these 500 report introducing a new product in the previous year, 400 report introducing a new production process and 350 report having introduced both a new product and a new process.

(a) What is the probability that a firm has done no innovation: neither introduced a new product nor a new process?

(b) Are the probabilities of introducing a new product and a new process independent?

(c) What is the conditional probability of introducing a new product given that the firm introduced a new process?

(d) What is the conditional probability of introducing a new process given that the firm introduced a new product?

Answers

(a) There were 550 innovators: 350 did both, 150 just product, 50 just process. Thus 450 did not innovate so the probability of not innovating was 0.45. Formally if event A, is make a product innovation, $P(A) = 0.5$; event B, make a process innovation, $P(B) = 0.4$. and the probability of doing both, $P(A \cap B) = 0.35$. For the event not making an innovation, $P(N)$

$$\begin{aligned} P(N) &= 1 - P(A \cup B) \\ &= 1 - (P(A) + P(B) - P(A \cap B)) \\ 0.45 &= 1 - (0.5 + 0.4 - 0.35). \end{aligned}$$

Notice the categories innovator, 550, and non-innovator 450 are mutually exclusive, the probability of being both an innovator and a non-innovator is zero by definition.

(b) If they were independent the product of the probability of product innovation times the probability of process innovation would give the probability of doing both: $P(A)P(B) = P(A \cap B)$. In this case $0.5 \times 0.4 = 0.2$ which is much

less than 0.35, so they are not independent. You are more likely to do a second type of innovation if you have already done one type.

(c) The probabilty of doing product innovation conditional on process innovation is the probability of doing both divided by the probability of doing process

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.35}{0.4} = 0.875$$

87.5% of process innovators also introduce a new product.

(d) The probability of doing process conditional on doing product is the probability of doing both divided by the probability of doing product:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{0.35}{0.5} = 0.7$$

70% of product innovators also introduce a new process. Notice that the answers to (c) and (d) are different.

10.2.3. Background Example: Hit and Run

Two cab companies operate in a city, 85% are green, 15% are blue. A cab hit a pedestrian at night and drove away. The person who had been hit said they thought the cab was blue. Subsequent tests showed that the person could correctly identify the color of a cab at night 80% of the time. What is the probability that the person was hit by a blue cab?

Answer.

We know the proportion of blue B and green G cabs are $P(B) = 0.15$, $P(G) = 0.85$. We know that the probability of the person reporting that it is blue RB given that it is blue is $P(RB \mid B) = 0.8$ and from this the probability of wrongly reporting that it is blue $P(RB \mid G) = 0.2$. What we need to know is the probability that it was blue given that they report it is blue $P(B \mid RB)$. The probability of the person reporting a blue cab is the probability of them seeing a blue cab times the probability of reporting it as blue plus the probability of seeing a green cab times the probability of wrongly reporting the cab as blue:

$P(RB) = P(B)\,P(RB \mid B) + P(G)\,P(RB \mid G) = 0.15 \times 0.8 + 0.85 \times 0.2 = 0.29$.

We have all the terms needed to apply Bayes Theorem

$$P(B \mid RB) = \frac{P(RB \mid B) \times P(B)}{P(RB)} = \frac{0.8 \times 0.15}{0.29} = 0.41$$

The report that the cab was blue increases the probability that the cab was blue from the unconditional prior probability of 0.15 to the conditional posterior probability of 0.41, but it is still a lot less than 0.8.

In this case we knew the prior probabilities, the proportion of blue and green cabs, that we used to adjust the report. In other cases where people report events we do not know the prior probabilities, e.g. when 15% of people in California report having being abducted by aliens.

10.2.4. Background Example: Lost H bomb

In 1966 a B52 crashed with an air-tanker while refueling at 30,000 feet off the coast of Palomares Spain, losing its four H bombs. Three were recovered quickly, the fourth was somewhere on the seabed. The US Navy constructed a map of the seabed, then got a group of various experts to bet on different scenarios that might have happened, (e.g. the bomb had two parachutes, the scenarios might be (i) both opened, (ii) one opened, (iii) none opened). Each scenario left the weapon in a different location. They then used Bayes theorem to combine the experts different subjective estimates of the probability (derived from the bets) to work out the (posterior) probability of the bomb being at each location. The highest probability location was far from where the other three bombs or the wreckage of the B52 were found. Fortunately the bomb was there. The details are in *Blind Man's Buff* S. Sontag and C. Drew, Harper Paperbacks, 1999, which gives various other examples of the use of Bayes theorem in submarine warfare.

11. Discrete Random Variables

Above we dealt with events where the outcomes are uncertain, now we want to consider how we apply probability to variables where we are uncertain what values they will take. These are called random variables. Forecasting involves estimating future values of random variables and should provide not only an estimate "our central forecast of CPI inflation in two years is 2.0%", but also an indication of the likely uncertainty "and we are 90% certain that it will lie between 1.0% and 3.0%". Inflation is a continuous random variable, it can take any value. We will begin with discrete random variables. Barrow discusses Random variables at the beginning of chapter 3.

A discrete random variable, X can take a number of distinct possible values, say $x_1, x_2, .., x_N$. with probabilities $p_1, p_2, ..., p_N$. The observed values are called

the realisations of the random variable. For instance, X the total obtained from throwing two dice is a discrete random variable. It can take the values 2 to 12. After you throw the dice, you observe the outcome, the realisation, a particular number, x_i. Associated with the random variable is a probability distribution, $p_i = f(x_i)$, which gives the probability of obtaining each of the possible outcomes the random variable can take. The cumulative probability distribution,

$$F(x_j) = \sum_{i=1}^{j} f(x_i) = P(X \leq x_j)$$

gives the probability of getting a value less than or equal to x_j. So in the dice case:

x_i	$f(x_i)$	$F(x_j)$
1	0	0
2	1/36	1/36
3	2/36	3/36
4	3/36	6/36
5	4/36	10/36
6	5/36	15/36
7	6/36	21/36
8	5/36	26/36
9	4/36	30/36
10	3/36	33/36
11	2/36	35/36
12	1/36	36/36

Make sure that you can calculate all the probabilities, use the 6x6 grid in section 10.1 if necessary. Notice $f(1) = 0$, it is impossible to get 1, and $F(12) = 1$, you are certain to get a value less than or equal to 12. $f(7) = 6/36$, because there are six different ways of getting a 7: (1,6), (6,1), (2,5), (5,2), (3,4), (4,3). These are the diagonal elements (running from bottom left to top right) in the grid above in section 10.1. $\sum f(x_i) = 1$. This is always true for a probability distribution. This probability distribution is symmetric with mean=median=mode=7.

The mathematical expectation or expected value of a random variable (often denoted by the Greek letter mu) is the sum of each value it can take, x_i, multiplied by the probability of it taking that value $p_i = f(x_i)$:

$$E(X) = \sum_{i=1}^{N} f(x_i)x_i = \mu. \tag{11.1}$$

The expected value of the score from two throws of a dice is seven; calculated as

$$7 = 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + 4 \times \frac{3}{36} ..... + 12 \times \frac{1}{36}.$$

If all the values are equally likely, $f(x_i) = 1/N$, so the expected value is the arithmetic mean.

The variance of a random variable is defined as

$$V(X) = E(X - E(X))^2 = \sum_{i=1}^{N} f(x_i)(x_i - \mu)^2 = \sigma^2. \tag{11.2}$$

If $f(x_i) = 1/N$ this is just the same as the population variance we encountered in descriptive statistics, section 6.1.2. This is the same formula that we used in section 6.2 with $f(x_i) = p_i$. In the dice example, the variance is 5.8 and the standard deviation 2.4.

Suppose that there are two random variables X and Y with individual (marginal) probabilities of $f(x_i)$ and $f(y_i)$ and joint probabilities $f(x_i, y_i)$. The joint probability indicates the probability of both X taking a particular value, x_i, and Y taking a particular value, y_i, and corresponds to $P(A \cap B)$ above. So if X is the number on the first dice and Y is the number on the second dice

$$f(6, 6) = P(X = 6 \cap Y = 6) = 1/36$$

If the random variables are independent, then the joint probability is just the product of the individual probabilities as we saw above

$$f(x_i, y_i) = f(x_i)f(y_i)$$

and if they are independent, the expected value of the product is the product of the expected values

$$E(XY) = E(X)E(Y).$$

Expected values behave like $N^{-1} \sum$. So if a is a constant $E(a) = a$. If a and b are constants $E(a + bx_i) = a + bE(x_i)$.

The Covariance between two random variables is

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))].$$

If $f(x_i) = 1/N$ this is

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y}).$$

as we saw in section 6.1.3. If the random variables are independent the covariance is zero. However, a covariance of zero does not imply that they are independent, independence is a stronger property.

11.0.5. Background Example The Lottery

In the lottery six numbers are randomly chosen from 49 possible numbers. Over the long-run the expected value of playing the lottery is $-55p$: you pay them £1 and they pay out $45p$ in prizes for every pound they take in. The $55p$ you lose, on average, goes on tax, good causes and their costs and profits. You win the jackpot if you match all six numbers, though not necessarily in the order drawn. Order the numbers from the smallest to the largest. For the first number you chose there are six chances of getting it (six draws). So the probability of your first number coming up is 6/49. To get your second number, you only have five chances (your first number has been drawn leaving 48 remaining numbers), so it is 5/48. Similarly the third is 4/47, fourth is 3/46, fifth is 2/45, sixth is 1/44. The probability of getting all 6 is the product of these probabilities

$$\left(\frac{6}{49}\right)\left(\frac{5}{48}\right)\left(\frac{4}{47}\right)\left(\frac{3}{46}\right)\left(\frac{2}{45}\right)\left(\frac{1}{44}\right) = \frac{720}{10068347520}$$

this is a 1 in $13,983,816$ chance, 1 in 14 million. Notice that low probability events are not necessarily rare, it depends on the population exposed to them. Winning the jackpot is a low probability event for any particular person, but it happens to someone almost every week. Always check the time horizon that the probability applies to. Someone shouting "we are all going to die" is not very worrying, since that is certainly true eventually, though if they mean in the next five minutes, it may be more worrying.

The usual formula for calculating the lottery is the number of ways in which a group of r objects (in this case 6) can be selected from a larger group of n objects (in this case 49) where the order of selection is not important. It is just the inverse of the formula above.

$${}^{n}C_{r} = \frac{n!}{r!(n-r)!} = \frac{49 \times 48 \times 47 \times 46 \times 45 \times 44}{6 \times 5 \times 4 \times 3 \times 2 \times 1}$$

The expected value of any particular game depends on whether the jackpot has been increased by being rolled over from previous games where it was not won. Even if the jackpot is over £14m, the expected value may not be positive,

because you may have to share the jackpot with other winners who chose the same number, (unless you are a member of a gang that bought all the available tickets and made sure nobody else could buy any tickets). Choosing an unpopular number, that others would not choose, will not change the probability of winning but may increase the probability of not having to share the jackpot. For instance, people sometimes use birthdays to select numbers, so do not choose numbers over 31. You can choose to buy random numbers to avoid this problem. Optimal design of lotteries raises interesting economic questions.

11.0.6. Background Example: churn

On average a customer is worth £20 a month to a mobile phone company. But the churn rate (the probability of a customer ceasing to subscribe) is 5% a month. The company's discount rate is 2% a month. The company wants to know how much it can subsidise handsets, while keeping the present value of a customer positive. What is a customer worth to the company?

Answer

Call the value per month V, the churn rate p, and the discount rate r. Every month a proportion of $(1-p)$ continue to the next month. The present value is

$$\begin{aligned} PV &= \sum_{t=0}^{\infty} \frac{V(1-p)^t}{(1+r)^t} = V \sum_{t=0}^{\infty} \left(\frac{(1-p)}{(1+r)} \right)^t \\ &= V \left(1 - \frac{(1-p)}{(1+r)} \right)^{-1} \\ &= V \left(\frac{1+r-(1-p)}{1+r} \right)^{-1} \\ &= \frac{(1+r)V}{r+p} \end{aligned}$$

since this is just the sum of a geometric progression in $(1-p)/(1+r)$. On these numbers

$$PV = \frac{1.02 \times 20}{0.07} = 291$$

a customer is worth £291.

12. Continuous random variables

Whereas a discrete random variable can only take specified values, continuous random variables (e.g. inflation) can take an infinite number of values. Corresponding to the probabilities $f(x_i)$ for discrete random variables there is a probability density function, *pdf*, also denoted $f(x_i)$ for continuous random variables and a distribution function $F(x_i) = \Pr(X \leq x_i)$ which gives the probability that the random variable will take a value less than or equal to a specified value x_i. The Bank of England publishes its estimate of the probability density function for inflation as a fan chart. Since there are an infinite number of points on the real line, the probability of any one of those points is zero, although the *pdf* will be defined for it. But we can always calculate the probability of falling into a particular interval, e.g. that inflation will fall into the range 1.5% to 2.5%. In the definitions of expected value and variance for a continuous random variable we replace the summation signs in (11.1) and (11.2) for the discrete case by integrals so

$$\begin{aligned} E(X) &= \int xf(x)dx = \mu \\ V(X) &= \int (x-\mu)^2 f(x)dx = \sigma^2. \end{aligned}$$

12.1. Uniform Distribution

The simplest continuous distribution is the uniform The probability density function takes equal values over some range (support) a to b. It is zero, $f(x_i) = 0$ outside the range and $f(x_i) = 1/(b-a)$ within the range. The mean of a uniform random variable, $E(x) = (a+b)/2$ and its variance is $Var(x) = (b-a)^2/12$. Thus if the range was $a = 0$ to $b = 1$ $E(x) = 0.5, Var(x) = 1/12$, and the standard deviation of x is 0.29. Notice in this case $f(x) = 1$ over the range of x, but the probabilities sum to unity $\int f(x)dx = 1$, since the graph of $f(x)$. has height 1, and length 1, so area 1. The uniform distribution is used in section 14.

12.2. The normal distribution

The most common distribution assumed for continuous random variables is the normal or Gaussian distribution. This has a bell shape. One source of normality comes from the central limit theorem. This says that the distribution of the

sample mean will be approximately normal, whatever the distribution of the original variable and that this approximation to normality will get better the larger the sample size. The normal distribution is completely defined by a mean (first moment) and variance (second moment), it has a coefficient of skewness (third moment) of zero and a coefficient of kurtosis (fourth moment) of three. The standard deviation is the square root of the variance. For a normal distribution roughly two thirds of the observations lie within one standard deviation of the mean and 95% lie within two standard deviations of the means.

Many economic variables, e.g. income or firm size, are not normally distributed but are very skewed and not symmetrical. However, the logarithm of the variable is often roughly normal. This is another reason we often work with logarithms of variables in economics.

Suppose that we have a random variable Y which is normally distributed with expected value $E(Y) = \alpha$ and variance

$$V(Y) = E(Y - E(Y))^2 = E(Y - \alpha)^2 = \sigma^2$$

We write this $Y \sim N(\alpha, \sigma^2)$. If we have an independent sample from this distribution, Y_i; $i = 1, 2, ..., N$ we write this $Y_i \sim IN(\alpha, \sigma^2)$. This is said Y_i is independent normal with expected value α and variance σ^2. Although the expected value of a normal distribution is often denoted μ, we use α to establish a link with regression below.

If one variable Y is normally distributed with mean α, and variance σ^2 . Then any linear function of Y is also normally distributed.

$$\begin{aligned} Y &\sim N\left(\alpha, \sigma^2\right) \\ X &= a + bY \sim N\left(a + b\alpha, b^2\sigma^2\right) \end{aligned}$$

It is b^2 because the variance is a squared concept, X has standard deviation $b\sigma$. So if temperature over a year measured in centigrade is normally distributed, temperature in Farenheit (which is a linear transformation) is also normal.

Using this we can write

$$Y_i = \alpha + u_i$$

where $u_i = Y_i - \alpha$, and from the rules $u_i \sim N\left(0, \sigma^2\right)$. Decomposing an observed random variable into its expected value and an error, u_i, is very convenient for many purposes.

An important linear function of Y is

$$z_i = \frac{Y_i - \alpha}{\sigma} \sim N(0, 1)$$

This is called the standard normal, has expected value zero and variance (and standard deviation) of one (like any standardised variable) and is tabulated in most statistics and econometrics books. Barrow Table A2 gives the table of $1 - F(z) = P(Z > z)$ for values of $z > 0$. So from the table in Barrow $P(Z > 0.44) = 0.33$. Read down the first column till 0.4 and then go across the row to the 0.04 column. Since the normal distribution is symmetric $P(Z > z) = P(Z < -z)$. So $P(Z < -0.44) = 0.33$ also.

The standard normal is useful because we can always convert from Y to z using the formula above and convert from z back to Y using

$$Y_i = \sigma z_i + \alpha.$$

The distribution has a bell shape and is symmetric with mean=median=mode. The formula for the normal distribution is

$$\begin{aligned} f(y_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2}(\frac{y_i - \alpha}{\sigma})^2\} \\ f(z_i) &= (2\pi)^{-1/2} \exp -\frac{z_i^2}{2} \end{aligned}$$

where $z_i = (y_i - \alpha)/\sigma$ is $N(0, 1)$ standard normal. The normal distribution is the exponential of a quadratic. The $(2\pi\sigma^2)^{-1/2}$ makes it integrate (add up) to unity, which all probability distributions should.

12.2.1. Areas under a normal distribution

For a standard normal distribution, with mean zero and standard deviation one, the probability that the random variable Z is less than a specific value z is given below for various values of z. Note we give $P(Z < z)$ for values of $z > 0$ whereas Barrow Table A2 gives $P(Z > z)$ for values of $z > 0$. There is no standard way to present areas under the normal distribution, so check how the table you are using presents it.

z	$P(Z < z)$
0	0.5
0.5	0.6915
1	0.8413
1.5	0.9332
2	0.9772

Since the normal distribution is symmetric, the probability of being less than the mean, (corresponding to $z = 0$) is 0.5, the same as the probability of being

greater than the mean. There is an 84% chance, of getting a value less than the mean plus one standard deviation, $z = 1$. The chance of being within one standard deviation of the mean is $P(-1 < Z < +1) = 0.6826$=0.8413-(1-0.8413). There is a 16% (1-0.84) chance of being less than one standard deviation below the mean, and a 16% chance or more than one standard deviation above the mean. The chance of being more than two standard deviations from the mean is 0.0456=2(1-0.9772), roughly 5%. Strictly 95% of the normal distribution lies within 1.96 standard deviations from the mean, but 2 is close enough for most practical purposes.

12.2.2. Example: equities

Suppose the average real return on equities in a particular year is,

$$Y_t \sim N(10\%, (15\%)^2),$$

normal with an expected value of $\alpha = 10\%$ and a standard deviation of $\sigma = 15\%$. We know that there is a 50% chance of getting 10% or better, but what return can we expect that there is a 33% chance of getting that return or better? From the normal tables we can find that $P(Z > 0.44) = 0.33$. So for a standard normal we know that we have a third chance of getting $z = 0.44$ or better. We now want to convert this to the distribution of returns using $Y = \sigma z + \alpha$. This is $Y = 15 \times 0.44 + 10 = 16.6\%$. So there is a one third chance of getting a return of 16.6% or better. Since the distribution is symmetric, we can also use -0.44 in the formula. This indicates that there is also a one third chance of getting a return of $Y = 15 \times (-0.44) + 10 = 3.4\%$ or less. For these values of mean and standard deviation, a third of the time you would be between 3.4% and 16.6%, a third of the time below 3.4% and a third of the time above 16.6%. This calculation assumes that the future distribution will be the same as that we observed in the past and this may not be true. It also assumes normality and equity returns can show much more extreme events than would be expected from a normal distribution: they are said to be fat tailed.

12.2.3. Example; test scores

Suppose that a maths class is made up of an equal number of Blue and Green Students. Within each group marks are distributed normally, but blues are better at maths with a mean of 60 compared to a mean of 55 for green students. Blue

students are also less erratic with a standard deviation of 5 compared to a standard deviation of 10 for green students.

(a)What proportion of blue students get more than 70?

(b) What proportion of green students get more than 70?

(c) Of those who get over 70 what proportion are green and what proportion are blue?

Answer

We have $B \sim N(60, 5^2)$, $G \sim N(55, 10^2)$

(a) We want to find the probability that the mark is over 70. For Blue students

$$z = \frac{70 - 60}{5} = 2$$

so the probability of a mark over 70 is $P(Z > 2) = 1 - P(Z < 2) = 1 - 0.9772 = 0.0228$ or 2.28%. The 0.9772 came from the table of areas under a normal distribution.

(b) For Green Students

$$z = \frac{70 - 55}{10} = 1.5$$

so the probability of a mark over 70 is $P(Z > 1.5) = 1 - P(Z < 1.5) = 1 - 0.9332 = 0.0668$ or 6.68%. The 0.9332 came from the table of areas under a normal distribution.

(c) In a large class with equal number of blue and green students, 4.48% of all students, (2.28+6.68)/2, would get over 70. The proportion of those that are blue is 25% (=2.28/(2.28+6.68)), the proportion that are green is 75% (=6.68/(2.28+6.68)).

Even though the question says blues are better at maths and it is true that their average is higher, three quarters of the top group in maths are green (as are three quarters of the bottom group). The lesson is to think about the whole distribution, not just the averages or parts of the distribution (e.g. the top of the class), and try not to be influenced by value-laden descriptions: 'better' or 'less erratic'.

12.3. Distributions related to the normal

We also use distributions which are functions of normally distributed variables.

12.3.1. Chi-squared

Suppose z_i is $IN(0,1)$, independently distributed, standard normal and we form

$$A = \sum_{i=1}^{n} z_i^2 \sim \chi^2(n)$$

Then A is said to have a Chi squared distribution with n degrees of freedom. Notice the Chi squared is only defined over positive values. As well as being the number of observations less the number of parameters estimated, degrees of freedom are a parameter of the distribution. The normal distribution has two parameters, which determine its shape, the mean and the variance. The mean determined its centre and the variance determined its spread. The Chi-squared distribution has one parameter, its degrees of freedom, that determines its shape. Its expected value equals its degrees of freedom; its variance equals twice its degrees of freedom. For small degrees of freedom the Chi squared distribution is skewed, for large degrees of freedom it approaches the normal. It arises naturally because we calculate estimates of the variance and $(n-1)s^2/\sigma^2$ has a χ^2 distribution with $n-1$ degrees of freedom, where

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

12.3.2. t distribution

A standard normal divided by the square root of a Chi-squared distribution, divided by its degrees of freedom, say n, is called the t distribution with n degrees of freedom

$$t(n) = z/\sqrt{\frac{\chi^2(n)}{n}}$$

We often divide an estimate of the mean or a regression coefficient (which are normally distributed) by their standard errors (which are the square root of a χ^2 divided by its degrees of freedom, and this is the formula for doing this. The t distribution has fatter tails than the normal, but as the sample size gets larger, about 30 is big enough, the uncertainty due to estimating the standard error becomes small and the distribution is indistinguishable from a normal. It is sometimes called the Student's t distribution. W.S. Gosset, who discovered it, worked for Guiness and because of a company regulation had to publish it under a pseudonym, and he chose Student.

12.3.3. F distribution

Fisher's F distribution is the ratio of two independent Chi-squared divided by their degrees of freedom.

$$F(n_1, n_2) = \frac{\chi^2(n_1)/n_1}{\chi^2(n_2)/n_2}.$$

It arises as the ratio of two variances.

Barrow chapter 6 discusses the χ^2 t and F distributions.

13. Economic and Financial Data II: interest rates, growth rates, exchange rates, etc

13.1. Interest rates

There are a set of basic rules that apply to a series of different rates: the rate of interest, the rate of return, the rate of inflation, the rate of growth of GDP or other variables. We will use rates of interest or return as an example, but the same rules apply to the others. Suppose we invest £100, in 2000, the value of the asset rises to £110 in 2001, £121 in 2001, 133.1 in 2002, etc. We can write this

$$\begin{aligned} V_0 &= 100 \\ V_1 &= 110 \\ V_2 &= 121 \\ V_3 &= 133.1 \end{aligned}$$

For other examples, V might be GDP (growth rates), a price index (inflation), etc. The gross rate of return in the first year is $(1 + r_1) = V_1/V_0 = 1.1$. The (net) rate of return in the first year is $r_1 = (V_1 - V_0)/V_0 = V_0/V_1 - 1 = 0.1$, the percentage rate of return is $100r_1 = 10\%$. Be aware of the difference between proportionate, 0.1 and percentage, 10%, rates of interest and return. Interest rates are also often expressed in basis points, 100 basis points is one percentage point.

From the definition of the gross return, $(1 + r_1) = V_1/V_0$, we can write $V_1 = (1 + r)V_0$. The rate of return in this example is constant at 10%, $r_i = r = 0.1$. Check this by calculating r_2 and r_3. The value of the investment in year 2 is

$$V_2 = (1 + r)V_1 = (1 + r)^2 V_0$$

and for year t

$$V_t = (1 + r)^t V_0. \tag{13.1}$$

Notice how interest compounds, you get interest paid on your interest. Interest rates are often expressed at annual rates even when they are for shorter or longer periods. If interest was paid out quarterly during the year, you would get 2.5% a quarter, not 10% a quarter. However it would be paid out four times as often so the formula would be

$$V_t = (1 + r/4)^{4t} V_0$$

or if it is paid out n times a year

$$V_t = (1 + r/n)^{nt} V_0.$$

As $n \to \infty$, continuous compounding, this converges to

$$V_t = e^{rt} V_0. \tag{13.2}$$

The irrational number $e \approx 2.718$ seems to have been discovered by Italian bankers doing compound interest in the late middle ages. Since $\ln V_t = rt + \ln V_0$ the continuously compounded return is

$$\frac{d \ln V}{dt} = \frac{1}{V} \frac{dV}{dt} = r.$$

For discrete data this can be calculated as

$$r = (\ln V_t - \ln V_0) / t.$$

The return in any period is often calculated as $r_t = \ln V_t - \ln V_{t-1}$.
Notice that the discrete version (13.1) is strictly

$$r = \exp(\{\ln V_t - \ln V_0\}/t) - 1$$

In addition

$$\ln V_t - \ln V_{t-1} = \ln \left(\frac{V_t}{V_{t-1}} \right) = \ln(1 + r_t) \approx r_t$$

if r is small, e.g. 0.1 another justification for using the difference of the logarithms. Growth rates and inflation rates are also calculated as differences in the logarithms. Multiply them by 100 if you want percentage rates.

So far we have assumed the rate of return is constant. Suppose instead of investing in a safe asset with a fixed return we had invested in a speculative equity and the values of our asset in 2000, 2001, and 2002 were

$$\begin{aligned} V_0 &= 100 \\ V_1 &= 1000 \\ V_2 &= 100 \end{aligned}$$

so $r_1 = 1000/100 - 1 = 9 = 900\%$, $r_2 = 100/1000 - 1 = -0.9 = -90\%$. The average (arithmetic mean) return is $(900-90)/2 = 405\%$. This example brings out two issues. Firstly percentages are not symmetric. Our investment got back to where it started after a 900% increase and only a 90% fall. Secondly, the arithmetic mean does not seem a very good indicator of the average return (except perhaps to the salesman who sold us the stock). We get an average return of 405% and have exactly the same amount as when we started. The geometric mean return in this case is zero

$$\begin{aligned} GM &= \sqrt{(1+r_1)(1+r_2)} - 1 \\ &= \sqrt{(10)(0.1)} - 1 \\ &= \sqrt{1} - 1 = 0 \end{aligned}$$

which seems more sensible.

There are also interest rates at various maturities, depending on how long the money is being borrowed or lent. The pattern of interest rates with respect to maturity is called the term structure of interest rates or yield curve. Typically the term structure slopes upwards. Long-rates, interest rates on money borrowed for a long period of time, such as 10 year government bonds, are higher than short rates, money borrowed for a short period of time, such as 3 month Treasury Bills. Interest rates are usually expressed at annual rates, whatever the length of the investment. When monetary policy is tight, the term structure may slope downwards, the yield curve is inverted: short-rates are higher than long-rates. This is often interpreted as a predictor of a forthcoming recession. Monetary policy is operated by the Central Bank through the control of a short overnight interest rate called the policy rate, Repo rate, Bank Rate or in the US Federal Funds Rate. Usually other short rates such as LIBOR (London Inter-Bank Offer Rate, the rate at which banks lend to each other) are very close to the policy rate. However, during the credit runch starting in August 2007 they diverged: Banks required a risk premium to lend to other banks.

13.2. Exchange Rates

13.2.1. Spot and forward

The spot exchange rate is the rate for delivery now: the exchange takes place immediately. The spot exchange rate is usually quoted as domestic currency per unit of foreign currency, with the dollar being treated as the foreign currency: Swiss Francs per Dollar for instance. A rise indicates a depreciation in the Swiss Franc: more Swiss Francs are needed to buy a dollar. Some are quoted as foreign currency per unit domestic, in particular Sterling, which is quoted Dollars per Pound. In this case a rise indicates an appreciation of the Pound, a pound buys more dollars. Forward rates are for delivery at some time in the future. The one year forward rate is for delivery in a years time when the exchange takes place at a rate quoted and agreed upon today.

13.2.2. Covered interest parity

Suppose that you have a $1 million and know that you are going to need Sterling in a years time. You can either convert the Dollars into Sterling now giving you £1/S m, where S is the spot rate, (£0.548 m at the spot rate on 27/7/4 of 1.8245$/£). You can invest this for a year at UK interest rates R£, 5.3375% (one year £ Libor (London Inter Bank Offer Rate) on the same day). Thus in a years time you will have £(1+R)/S m (£0.577m). Alternatively you could invest the Dollars in the US getting rate R$, 2.43% ($ Libor) so you will have $(1+R$) m ($1.0243m) in a years time. You can sell these dollars for delivery in a year at forward rate F, (1.7752). This means that in a years time you will have £(1+R$)/F m (£0.577m). These are both risk free (assuming the bank repays your deposit, i.e. no default risk) so the returns must be equal, as they are. You get £577,000 by either route. These rates are given every day in the FT. Expressing it as a formula

$$\begin{aligned} \frac{1+R\pounds}{S} &= \frac{1+R\$}{F} \\ \frac{F}{S} &= \frac{1+R\$}{1+R\pounds} \end{aligned}$$

Or approximately

$$\frac{F-S}{S} \approx R\$ - R\pounds$$

where the term on the left hand side is called the forward premium, when positive or discount when negative and the term on the right hand side the interest differ-

ential. This relation is called covered interest parity and follows from arbitrage. If it did not hold banks could make a riskless return by exploiting the difference. Notice that the forward rate on 27/7/4 indicated that the market expected sterling to fall in value over the next year. In fact because it is determined by the interest rate differential, the forward rate tends to be a poor predictor of the future spot rate. In this case the rate on 29/7/5 was \$1.759, so the forward rate was quite a good predictor. Verbeek section 4.11 gives some empirical examples of the use of these relationships.

13.2.3. Real and effective rates

Suppose a Big Mac costs \$X in the US and £Y in the UK and the spot rate is S, e.g. 1.6\$/£. Then the relative price of Big Macs in the UK and US is dollars

$$Q^M = (£Y \times S)/\$X$$

this is the Big Mac real exchange rate, which the Economist publishes. Similar calculations are done for the whole economy using price indexes, to calculate the real exchange rate for the country: $Q = P^*S/P$, where P^* is a foreign price index and P a domestic price index. Purchasing Power Parity (PPP) says that trade will equalise prices between countries and the real exchange rate will be constant, in the long run. As Keynes said 'In the long run, we are all dead' and deviations from PPP can be very persistent.

Any currency has a large number of exchange rates with all the other currencies, so 'trade weighted averages' or effective exchange rates are often calculated, which give an average exchange rate with all the other currencies, the weights reflecting their relative importance in trade. Suppose that we denote the base year as year zero, e.g. 2000, then the index in year t is

$$I_t = \sum w_i(\frac{S_{it}}{S_{i0}})$$

where the w_i are the percentage shares of trade with country i so that $\sum w_i = 100$, S_{io} is the exchange rate with country i in the base year and S_{it} is the exchange rate in year t. The value of the index in year zero, $I_0 = 100$.

14. Applied Exercise II: Sampling distributions

The remainder of the notes are concerned with estimation and hypothesis testing. These rest on the idea of a sampling distribution, the particular sample we take

is just one of a large number of possible samples that we might have taken. This exercise is designed to illustrate the idea of a sampling distribution. You do not need any data for this exercise, you create it yourself.

Go into Excel and in cell A1 type =RAND( ). This generates a random number uniformly distributed over the interval zero-one. Each number over that interval has an equal chance of being drawn. Copy this cell right to all the cells till O1. In P1 type =AVERAGE(A1:O1). In Q1 type =STDEVP(A1:O1). In R1 type =STDEV(A1:O1). Copy this row down to line 100.

You now have 100 samples of 15 observations from a uniform distribution and 100 estimates of the mean and 100 estimates for each of two estimators of the standard deviation. An estimator is a formula which tells you how to calculate an estimate from a particular sample; an estimate is the number that results from applying the estimator to a sample.

We can then look at the sampling distribution of the mean and standard deviation. Calculate and draw the histogram for the 100 estimates of the mean. Do the same for the two estimators of the standard deviation. What do you think of their shapes? Are they close to a normal distribution? Go to P101, type in =AVERAGE(P1:P100). Go to P102 type in =STDEV(P1:P100). Go to P103 type in =SKEW(P1:P100). Go to P104 type in KURT(p1:p100). Copy these to the Q and R columns to give the descriptive statistics for the two estimates of the standard deviation.

If x is uniformly distributed over $[a, b]$ then $E(x) = (a + b)/2$ and $Var(x) = (b - a)^2/12$. In this case $a = 0, b = 1$, so the theoretical mean should be 0.5 (compare this with the number in cell P101) and the theoretical variance 1/12, with standard deviation $\sqrt{1/12} = 0.288675$ (compare this with the number in Q101, which should be biased downwards and with the number in R101, which should be closer). The standard deviation of the mean from a sample of size N (in this case 15) is

$$SD(\overline{x}) = \sqrt{Var(x)/N}$$

so we would expect the standard deviation of our distribution of means to be 0.07453 (compare this with the number in P102). As N becomes large (the number of observations in each sample, 15 in this case which is not very large), the distribution of the mean tends to normality (the central limit theorem). Do the measures of skewness and excess kurtosis given in P103 and P104 suggest that these means are normal? The values should be close to zero for normality. Is the mean more normally distributed than the standard deviation?

What we have done here is called a 'Monte Carlo' simulation. We have ex-

amined the properties of the estimators by generating lots of data randomly and looking at the distributions of the estimates from lots of samples. In practice, 100 replications of the sample is rather small, in Monte Carlo studies many thousands of replications are typically used. Because 100 is quite small, you will get rather different answers (e.g. for the overall mean) from me. However, by making the number of replications sufficiently large, we could make the difference between you and me as small as we like (law of large numbers).

In this case, we did not need to do a Monte Carlo because we can derive the properties of the estimators theoretically. But for more complicated problems this is not the case and we must do it numerically like here. However, doing the Monte Carlo gives you a feel for what we mean when we discuss the distribution of an estimator.

15. Estimation

15.1. Introduction

In the first part of the course we looked at methods of describing data, e.g. using measures like the mean (average) to summarise the typical values the variable took. In the second part of the course, we learned how to make probability statements. Now we want to put the two together and use probability theory to judge how much confidence we have in our summary statistics. The framework that we will use to do this is mathematical, we will make some assumptions and derive some results by deduction. Chapter 4 and 5 of Barrow covers these issues. There are a number of steps.

1. We start with a model of the process that generates the data. For instance, the efficient market theory says that the return on a stock in any period t, $Y_t = \alpha + u_t$ Where Y_t is the return, which we can observe from historical data, α is the expected return, an unknown parameter, and u_t is an unpredictable random error that reflects all the new information in period t. We make assumptions about the properties of the errors u_t. We say that the error u_t is 'well behaved' when it averages zero, $E(u_t) = 0$; is uncorrelated through time $E(u_t u_{t-1}) = 0$; and has constant variance, $E(u_t^2) = \sigma^2$.

2. We then ask how we can obtain an estimator of the unknown parameter α. An estimator is a formula for calculating an estimate from any particular sample. We will use two procedures to choose an estimator $\widehat{\alpha}$, (said alpha

hat) of α, that gives $Y_t = \widehat{\alpha} + \widehat{u}_t$: (1) method of moments, which chooses the estimator that makes our population assumptions, e.g. $E(u_t) = 0$, hold in the sample so $N^{-1} \sum \widehat{u}_t = 0$ (2) least squares, which chooses the estimator that has the smallest variance and minimises $\sum \widehat{u}_t^2$. In the cases we look at these two procedures give the same estimator, but this is not generally true.

3. We then ask how good the estimator is. To do this we need to determine what the expected value of the estimator is and the variance of the estimator, or its square root: the standard error. We then need to estimate this standard error. Given our assumptions, we can derive all these things mathematically and they allow us to determine how confident we are in our estimates. Notice the square root of the variance of a variable is called its standard deviation, the square root of the variance of an estimator is called its standard error.

4. We then often want to test hypotheses. For instance, from Applied Exercise I, we found that the mean real return on equities over the period 1872-1999 was 0.088 (8.8%) with a standard deviation of 0.184; but the mean 1872-1990 was only 0.080, while the return during 1991-1999 was 0.18 (18%) with a standard deviation of 0.15. In the year 2000, you might have wanted to ask whether there really was a New Economy, with significantly higher returns (over twice the historical average) and lower risk (a lower standard deviation); or whether you might just get the numbers observed during the 1990s, purely by chance.

5. Since our mathematical derivations depend on our assumptions, we need to check whether our assumptions are true. Once we have estimated $\widehat{\alpha}$ we can estimate $\widehat{u}_t = Y_t - \widehat{\alpha}$. Then we can ask whether our estimates of the errors are uncorrelated and have constant variance.

We will go through this procedure twice, first for estimating the sample mean or expected value and testing hypotheses about it, and then follow the same procedure for regression models, where the expected value is not a constant, but depends on other variables.

15.1.1. A warning

The procedures we are going to cover are called classical statistical inference and the Neyman-Pearson approach to testing. When first encountered they may seem

counter-intuitive, complicated and dependent on a lot of conventions. But once you get used to them they are quite easy to use. The motivation for learning these procedures is that they provide the standard approach to dealing with quantitative evidence in science and other areas of life, where they have been found useful. However, because they are counter-intuitive and complicated it is easy to make mistakes. It is claimed that quite a large proportion of scientific articles using statistics contain mistakes of calculation or interpretation. A common mistake is to confuse statistical significance with substantive importance. Significance just measures whether a difference could have arisen by chance it does not measure whether the size of the difference is important. There is another approach to statistics based on Bayes Theorem. In many ways Bayesian statistics is more intuitive, since it does not involve imagining lots of hypothetical samples as classical statistics does. It is conceptually more coherent, since it just involves using your new data to update your prior probabilities in the way we did in section 10.2.2. However, it is often mathematically more complex, since it usually involves integrals. Modern computers are making this integration easier. Gary Koop, *Bayesian Econometrics*, Wiley 2003 provides a good introduction.

It is important to distinguish two different things that we are doing. First, in theoretical statistics we are making mathematical deductions: e.g. proving that an estimator has minimum variance in the class of linear unbiased estimators. Second, in applied statistics, we making inductions, drawing general conclusions from a particular set of observations. Induction is fraught with philosophical difficulties. Even if every swan we see is white, we are not entitled to claim 'all swans are white', we have not seen all swans. But seeing one black swan does prove that the claim 'all swans are white' is false. Given this, it is not surprising that there are heated methodological debates about the right way to do applied statistics and no 'correct' rules. What is sensible depends on the purpose of the exercise. Kennedy, *A Guide to Econometrics*, fifth edition chapter 21 discusses these issues.

15.2. Estimating the expected value.

Suppose we have an independent sample of data over time $Y_1, Y_2, ..., Y_T$,

$$Y_t = \alpha + u_t$$

where u_t is a random variable with mean zero and variance σ^2 and the observations are uncorrelated or independent through time, i.e. $E(u_t) = 0$, $E(u_t^2) = \sigma^2$,

$E(u_t u_{t-i}) = 0$. Notice the number of observations here is T, earlier we used N or n for the number of observations. We wish to choose a procedure for estimating the unknown parameter α from this sample. We will call the estimator $\widehat{\alpha}$ (said alpha hat). We get an estimate by putting in the values for a particular sample into the formula. We derive the estimator $\widehat{\alpha}$ in two ways: method of moments which matches the sample data to our population assumptions and least squares which mimimises the variance.

15.2.1. Method of moments.

We assumed that $E(Y_t) = \alpha$, which implies $E(u_t) = E(Y_t - \alpha) = 0$. Let us choose an estimator $\widehat{\alpha}$ such that the sample equivalent of the expected value, (the mean) of $(Y_t - \widehat{\alpha})$ also equals zero. That is we replace $E(Y_t - \alpha) = 0$ with $T^{-1}\sum_{t=1}^{T}(Y_t - \widehat{\alpha}) = 0$. This imples $T^{-1}\left\{\sum_{t=1}^{T} Y_t - T\widehat{\alpha}\right\} = 0$ or $T^{-1}\sum_{t=1}^{T} Y_t = \widehat{\alpha}$. So the estimator which makes the sample equivalent of $E(Y_t - \alpha) = 0$ hold is the mean, so $\widehat{\alpha} = \overline{Y}$. Notice this derivation also implies that $\sum_{t=1}^{T}(Y_t - \widehat{\alpha}) = 0$, the sum of deviations from the mean are always zero

15.2.2. Least squares

Alternatively suppose, we choose the estimator, $\widehat{\alpha}$ that makes the sum of squared deviations, $S = \sum(Y_t - \widehat{\alpha})^2$ as small as possible. This will also minimise the estimated variance, $\widehat{\sigma}^2 = \sum(Y_t - \widehat{\alpha})^2/T$.

$$\begin{aligned} S &= \sum_{t=1}^{T}(Y_t - \widehat{\alpha})^2 \\ &= \sum(Y_t^2 + \widehat{\alpha}^2 - 2\widehat{\alpha}Y_t) \\ &= \sum Y_t^2 + T\widehat{\alpha}^2 - 2\widehat{\alpha}\sum Y_t \end{aligned}$$

To find the $\widehat{\alpha}$ that minimises this, we take the first derivative of S with respect to $\widehat{\alpha}$ and set it equal to zero:

$$\frac{\partial S}{\partial \widehat{\alpha}} = 2T\widehat{\alpha} - 2\sum Y_t = 0.$$

Divide through by 2, move the $-\sum_{t=1}^{T} Y_t$ to the other side of the equality,

gives $T\widehat{\alpha} = \sum Y_t$ or

$$\widehat{\alpha} = \sum_{t=1}^{T} Y_t/T.$$

so again $\widehat{\alpha} = \overline{Y}$.

15.3. Properties of the estimator

We distinguish, between the true (or population) parameter α and the estimator $\widehat{\alpha} = \sum Y_i/n$, the formula telling you how to calculate an estimate from a particular sample. A different sample would give a different estimate, so $\widehat{\alpha}$ is a random variable. When different estimators are available, in this case the median might be an alternative estimator, we need criteria to choose between different estimators. One criterion is that the estimator is unbiased, on average (over lots of hypothetical samples) it is equal to the true value. The expected value of the estimator is equal to the true value of the parameter. Another property that is often desirable is that the estimates tends to be close to the true value; for unbiased estimators this implies that the estimator has a small variance.

15.3.1. The expected value of $\widehat{\alpha}$

To find out whether the mean is unbiased we need to calculate the expected value of $\widehat{\alpha}$. This is

$$\begin{aligned} \widehat{\alpha} &= \sum Y_t/T = \sum(\alpha + u_t)/T = T\alpha/T + (\sum u_t/T) \\ E(\widehat{\alpha}) &= \alpha + E(\sum u_t/T) = \alpha \end{aligned}$$

Since $E(u_t) = 0$. So $\widehat{\alpha}$ is unbiased under our assumptions. From this derivation we see that:

$$\widehat{\alpha} - \alpha = \sum u_t/T$$

while on average over lots of hypothetical samples, $\sum u_t/T$ may be zero, it will not be zero in any particular sample, so our estimate will differ from the true value. Now let us calculate how large the difference is likely to be.

15.3.2. The variance and standard error of the mean $\widehat{\alpha}$

The variance of $\widehat{\alpha}$, say $V(\widehat{\alpha}) = E(\widehat{\alpha} - E(\widehat{\alpha}))^2 = E(\widehat{\alpha} - \alpha)^2$ since $E(\widehat{\alpha}) = \alpha$. Since $\widehat{\alpha} - \alpha = \sum u_t/T$

$$E(\widehat{\alpha} - \alpha)^2 = E(\sum u_t/T)^2$$

The right hand side can be written

$$= E\left(\frac{u_1}{T} + \frac{u_2}{T} + ... + \frac{u_T}{T}\right)\left(\frac{u_1}{T} + \frac{u_2}{T} + ... + \frac{u_T}{T}\right)$$

This product will have T^2 terms. There are T terms with squares like u_1^2, and $T^2 - T$ terms with cross-products like $u_1 u_2$. The expectation of the squares are $E(u_t^2)/T^2 = \sigma^2/T^2$, since the variance of the u_t, $E(u_t^2) = \sigma^2$, is assumed constant for all t. There are T terms like this, so the sum is $T(\sigma^2/T^2) = \sigma^2/T$. The expectations of the cross products are of the form $E(u_t u_{t-j})/T^2$. But since the errors are assumed independent $E(u_t u_{t-i}) = 0$, for $i \neq 0$, so the expectation of all the cross-product terms equals zero. Thus we have derived the variance of the mean, which is:

$$V(\widehat{\alpha}) = E(\widehat{\alpha} - E(\widehat{\alpha}))^2 = \frac{\sigma^2}{T}$$

where T is the number of observations.

The square root of the variance $\sigma/\sqrt{T}$ is called the standard error of the mean. It is used to provide an indication of how accurate our estimate is. Notice when we take the square root of the variance of a **variable** we call it a **standard deviation**; when we take the square root of a variance of an **estimator**, we call it a **standard error**. They are both just square roots of variances.

15.3.3. Estimating the variance

There are two common estimators of the variance of Y:

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{\sum(Y_t - \widehat{\alpha})^2}{T} \\ s^2 &= \frac{\sum(Y_t - \widehat{\alpha})^2}{T-1} \end{aligned}$$

the first estimator, $\widehat{\sigma}^2$, sometimes called the population variance, which divides by T is a biased estimator of σ^2,

$$E(\widehat{\sigma}^2) = \frac{T-1}{T}\sigma^2 < \sigma^2.$$

The proof of this is simple, but long, so we do not give it. The second estimator, s^2, sometimes called the sample variance, is an unbiased estimator. The bias arises because we use an estimate of the mean and the dispersion around the estimate is going to be smaller than the dispersion around the true value because the estimated mean is designed to make the dispersion as small as possible. If we used the true value of α there would be no bias. The correction $T-1$ is called the degrees of freedom: the number of observations minus the number of parameters estimated, one in this case, $\widehat{\alpha}$. We estimate the standard error of the mean by

$$SE(\widehat{\alpha}) = \frac{s}{\sqrt{T}}$$

On the assumptions that we have made it can be shown that the mean is the minimum variance estimator of the expected value among all estimators which are linear functions of the Y_i and are unbiased. This is described as the mean being the Best (minimum variance) Linear Unbiased Estimator (BLUE) of the expected value of Y. This is proved later in a more general context, but it is a natural result because we chose this estimator to minimise the variance.

15.3.4. Asymptotic properties

In many cases we are interested in what happens to the properties of the estimator when T gets large. So although $\widehat{\sigma}^2$ which divides by T is a biased estimator of σ^2,

$$E(\widehat{\sigma}^2) = \frac{T-1}{T}\sigma^2$$

as $T \to \infty$; $(T-1)/T \to 1$ and the bias goes away. In addition as $T \to \infty$ the standard error of the mean $\sigma/\sqrt{T} \to 0$, the distribution of estimates get closer and closer to the true value so with $T = \infty$ there is no dispersion at all, the estimator converges to its true value, we can estimate it exactly. Estimators which have this property are said to be consistent. Verbeek section 2.6 discusses asymptotic properties.

15.3.5. Summary

So far we have (1) found out how to estimate the expected value of Y $\alpha = E(Y)$ by the mean; (2) shown that if the expected value of the errors is zero the mean is an unbiased estimator, (3) shown that if the errors also have constant variance σ^2 and are independent, the variance of the mean is σ^2/T where T is the number

of observations (4) shown that the standard error of the mean can be estimated by $s/\sqrt{T}$, where s^2 is the unbiased estimator of the variance and claimed (5) that the mean had the minimum variance possible among linear unbiased estimators (the Gauss-Markov theorem) and (6) that for large T the distribution of $\widehat{\alpha}$, will be normal whatever the distribution of Y, (the central limit theorem).

15.3.6. Background Example: proportions

We also often use sample proportions to estimate probabilities. Barrow Chapter 4 covers proportions. For instance in 10.2.2 we found from a sample of n =1000 firms that 450 reported doing neither product and process innovations, so we estimated $P(N) = p = 0.45$. Had we sampled different firms we would have got a different estimate and only if we had sampled all firms in the economy, the population, would we be sure that we got the true proportion of non-innovators. We want to know how accurate our estimate is, what is its standard error? In the case of a proportion the standard error is

$$SE(p) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.45 \times 0.55}{1000}} = 0.0157$$

So with a sample of 1000 our standard error on the estimated proportion of 45% is 1.57%. Often the formula is given using $q = 1 - p$.

In designing surveys it is important to check that: your sample is representative, random samples are best; not biased by non-response (innovators may be more likely to fill out the form); the questions are clear, in this case firms may differ on what they think an innovation is; and the sample is large enough. Barrow chapter 9 discusses these issues. Suppose that you thought that a standard error of 1.5% was too large and wanted to reduce it to 1%, how large would the survey need to be. Call our desired $SE(p)$ x. Then we need

$$\begin{aligned} x &= \sqrt{\frac{p(1-p)}{n}} \\ x^2 &= p(1-p)/n \\ n &= p(1-p)/x^2 \end{aligned}$$

to get $x = 1\%$ we need $n = (0.45 \times 0.55)/(0.01)^2 = 2475$. You would need to more than double the sample.

16. Confidence intervals and Hypothesis Tests

Earlier we noted that if a variable was normally distributed, the mean plus or minus two standard deviations would be expected to cover about 95% of the observations. This range, plus or minus two standard deviations is called a 95% confidence interval. In addition to constructing confidence intervals for a variable we also construct confidence intervals for our estimate of the mean, where we use the standard error of the mean instead of the standard deviation. Barow Chapter 4 discusses these issues.

16.1. Example: was the New Economy different?

We often want to compare our estimate,$\widehat{\alpha}$ with some hypothesised value α_0. To illustrate the procedures, suppose the average real return on equities 1991-1999 was 18%, and the average 1872-1990 was 8%. Check the actual numbers from Applied Exercise I. At the end of the 1990s many people believed that there had been a structural change and the arrival of the "New Economy" produced much higher expected returns than the historical average. In this case $\alpha_0 = 8\%$, our hypothesis is that the true expected return during the 1990s is the long-run historical average. Our estimate is the mean during the 1990s, $\widehat{\alpha} = 18\%$, estimated on $T = 9$ observations. As our estimate of the standard deviation (square root of variance) of returns, we will use the estimate over the whole period, $\sigma = 18.4\%$, .assuming that the variance has not changed and that we know it exactly. This simplifies the calculation a lot, if we used the estimated variance 1991-99 we would have a larger sampling error and have to use the t distribution rather than the normal. Using the formula above, the standard error of our mean for the 1990s is thus

$$se(\widehat{\alpha}) = \sigma/\sqrt{T} = 18.4/\sqrt{9} = 6.13.$$

Notice that we are using the mean calculated over 9 years, 1991-99, so we divide by $\sqrt{9}$, not by the number of years that we used to calculated the standard deviation. If our estimate of the mean is normally distributed, with a true expected value $\alpha_0 = 8\%$ we would expect there to be a 95% chance of the estimate falling into the range

$$\alpha \pm 1.96 se(\widehat{\alpha}) = 8 \pm 1.96\,(6.13) = 8 \pm 12.$$

Thus we would be 95% confident that the range -4 to$+20$ would cover the true value. There is a 2.5% chance that a 9 year average would be above 20% and a

2.5% chance it would be below -4%. The historical estimate falls in this range so even if the true expected return is equal to its historical value, 8%; we would expect to see 9 year average returns of 18% just by chance, more than 5% of the time. Suppose the true expected value is 8%, what is the probability of observing 18% or more? We can form

$$z = \frac{\widehat{\alpha} - \alpha}{se(\widehat{\alpha})} = \frac{18 - 8}{6.13} = \frac{10}{6.13} = 1.63$$

Using the tables of the normal distribution we find $P(Z \geq 1.63) = 0.0516$. Just over 5% of the time, we will observe periods of 9 years with mean returns of 18% or greater, if expected return is constant at 8%.

This assumes everything is normally distributed, if the distribution had heavier tails than a normal, these probabilities would be a little larger. We could also centre our confidence interval on $\widehat{\alpha}$ rather than α_0 and report $\widehat{\alpha} \pm 1.96 se(\widehat{\alpha})$. This would be 18 ± 12 the range, 6 to 30. Notice this confidence interval covers the historical mean.

We would conclude that, assuming normality, at the 95% level, the 1990s return is not statistically significantly different from the historical mean.

16.2. Confidence intervals

Suppose $Y_t = \alpha + u_t$. We can get an estimate, $\widehat{\alpha}$, of α from our sample of T observations . If u_t is normally distributed, $\widehat{\alpha}$ will also be normally distributed, because $\widehat{\alpha}$ is just a linear function of u. If the sample is large enough $\widehat{\alpha}$ will also be normally distributed, even if u_t is not normal. We saw that

$$\widehat{\alpha} = \alpha + \sum_t u_t / T$$

So

$$\widehat{\alpha} \sim N(\alpha, \sigma^2/T)$$

$s(\widehat{\alpha}) = \sigma/\sqrt{T}$ is the standard error of $\widehat{\alpha}$ and can be estimated by, $\widehat{s(\alpha)} = s/\sqrt{T}$, where

$$s = \sqrt{\frac{\sum (Y_t - \widehat{\alpha})^2}{T - 1}}$$

From tables of the normal distributions, we know that it is 95% certain that $\widehat{\alpha}$ will be within 1.96 standard errors of its true value α. The range $\widehat{\alpha} \pm 1.96 s(\widehat{\alpha})$ is

called the 95% confidence interval. The 68% confidence interval is $\widehat{\alpha} \pm s(\widehat{\alpha})$: the range covered by the estimate plus and minus one standard error will cover the true value just over two thirds of the time. If that confidence interval covers some hypothesised value α_0, then we might be confident that the true value could be α_0. If $\widehat{\alpha}$ is more than about 2 standard errors from the hypothesised value, α_0, we think it unlikely that the difference could have occurred by chance (there is less than a 5% chance) and we say the difference is statistically significant. That is we calculate the test statistic

$$\tau = \frac{\widehat{\alpha} - \alpha_0}{\widehat{s(\alpha)}}$$

and reject the null hypothesis that $\alpha = \alpha_0$ at the 5% level if the absolute value of the test statistic is greater than 1.96. We can also calculate the 'p value', the probability of getting the observed value of τ if the hypothesis was true and reject the hypothesis if the p value is less than 0.05.

16.3. Small samples

Suppose that we have a sample of T observations. If we knew the true standard deviations, the standard error would be $s(\widehat{\alpha}) = \sigma/\sqrt{T}$ then $(\widehat{\alpha} - \alpha_0)/s(\widehat{\alpha})$ would have a normal distribution. But when we estimate the variance by s^2 our estimated standard error is $\widehat{s(\alpha)} = s/\sqrt{T}$. This adds extra uncertainty, from not knowing σ, and $(\widehat{\alpha}-\alpha_0)/\widehat{s(\alpha)}$ follows another distribution called the t distribution, introduced above, which is more spread out. How much more spread out depends on the degrees of freedom: $T - 1$. As the number of observations, T, becomes large the effect of estimating the variance becomes smaller and the t distribution becomes closer to a normal distribution. For a normal distribution 95% of the distribution lies within the range ± 1.96 standard errors. For a t distribution with 3 degrees of freedom ($n = 4$) 95% lies within ± 3.182 standard errors, with 10 degrees of freedom it is ± 2.228,with 30 degrees of freedom it is ± 2.042. Tables of the t distribution are given at the back of statistics textbooks, e.g. Barrow Table A3. The practice in economics is to use 2 as the critical value and not use very small samples. Most computer programs give p values, the probability of the null hypothesis being true, and they may be more convenient to use. At the 5% level you reject the null hypothesis if the p value is < 0.05. You have to know what the null hypothesis is.

In testing the new economy above we used the estimate of the standard deviation from the whole sample, so the number of observations used in estimating

the variance was large and we could use the normal distribution. If we had used the estimate of the standard deviation from the 1990s the number of observations would have been small, 9, and we would have had to use the t distribution.

16.4. Testing

In testing we start with what is called the null hypothesis, $H_0 : \alpha = \alpha_0$. It is called null because in many cases our hypothesised value is zero, i.e. $\alpha_0 = 0$. We reject it in favour of what is called the alternative hypothesis, $H_1 : \alpha \neq \alpha_0$; if there is very strong evidence against the null hypothesis. This is a two sided alternative, we reject if our estimate is significantly bigger or smaller. We could also have one sided alternatives $\alpha < \alpha_0$ or $\alpha > \alpha_0$. The convention in economics is to use two sided alternatives.

The problem is how do we decide whether to reject the null hypothesis or not to reject the null hypothesis. In criminal trials, the null hypothesis is that the defendant is innocent. The jury can only reject this null hypothesis if the evidence indicates guilt "beyond reasonable doubt". Even if you think the defendant is probably guilty (better than 50% chance) you have to acquit, this is not enough. In civil trials juries decide "on the balance of the evidence", there is no reason to favour one decision rather than another. So when OJ Simpson was tried for murder, a criminal charge, the jury decided that the evidence was not beyond reasonable doubt and he was acquitted. But when the victims family brought a civil case against him, to claim compensation for the death, the jury decided on the balance of the evidence that he did it. This difference reflects the fact that losing a criminal case and losing a civil case have quite different consequences.

Essentially the same issues are involved in hypothesis testing. We have a null hypothesis, defendant is innocent. We have an alternative hypothesis that the defendant is guilty. We can never know which is true. There are two possible decisions. Either accept the null hypothesis (acquit the defendant) or reject the Null hypothesis (find the defendant guilty). In Scotland the jury has a third possible verdict: not proven. Call the null hypothesis, H_0, this could be defendant innocent or $\alpha = \alpha_0$. Then the possibilities are

	H_0 true	H_0 false
Accept H_0	Correct	Type II error
Reject H_0	Type I error	Correct

In the criminal trial Type I error is convicting an innocent person. Type II error is acquitting a guilty person. Of course, we can avoid Type I error completely:

always accept the null hypothesis: acquit everybody. But we would make a lot of type II errors, letting guilty people go. Alternatively we could make type II errors zero, convict everybody. Since we do not know whether the null hypothesis is true (whether OJ is really innocent), we have to trade off the two risks. Accepting the null hypothesis can only be tentative, this evidence may not reject it, but future evidence may.

Statistical tests design the test procedure so that there is a fixed risk of Type I error: rejecting the null hypothesis when it is true. This probability is usually fixed at 5%, though this is just a convention.

So the procedure in testing is

1. Specify the null hypothesis, $\alpha = \alpha_0$.

2. Specify the alternative hypothesis $\alpha \neq \alpha_0$.

3. Design a test statistic, which is only a function of the observed data and the null hyposthesis, not a function of unknown parameters

$$\tau = \frac{\widehat{\alpha} - \alpha_0}{\widehat{s(\alpha)}}$$

4. Find the distribution of the the test statistic if the null hypothesis is true. In this case the test statistic, τ, has a t distribution in small samples (less than about 30), a normal distribution in large samples.

5. Use the distribution to specify the critical values, so that the probability of $\widehat{\alpha}$ being outside the critical values is small, typically 5%.

6. Reject the null if it is outside the critical values, (in this case outside the range ± 2); do not reject the null otherwise.

7. Consider the power of the test. The power is the probability of rejecting the null hypothesis when it is false $(1 - P(typeIIerror))$, which depends on the true value of the parameters.

In the medical example, of screening for a disease, that we used in section 11, we also had two types of errors (false positives and false negatives), and we had to balance the two types of error in a similar way. There we did it on the basis of costs and benefits. When the costs and benefits can be calculated that is the best way to do it. In cases where the costs and benefits are not known we use significance tests.

Statistical significance and substantive significance can be very different. An effect may be very small of no importance, but statistically very significant, because we have a very large sample and a small standard error. Alternatively, an effect may be large, but not statistically significant because we have a small

sample and it is imprecisely estimated. Statistical significance asks: 'could the difference have arisen by chance in a sample of this size?' not 'is the difference important?'

When we discussed confidence intervals we said that the 68% confidence interval is $\widehat{\alpha} \pm s(\widehat{\alpha})$: the range covered by the estimate plus and minus one standard error will cover the true value, α, just over two thirds of the time. There is a strong temptation to say that the probability that α lies within this range is two thirds. Strictly this is wrong, α is fixed not a random variable, so there are no probabilities attached to α. The probabilities are attached to the random variable $\widehat{\alpha}$, which differ in different samples. Bayesian statistics does treat the parameters as random variables, with some prior probability distribution; uses the data to update the probabilities; and does not use the Neyman-Pearson approach to testing set out above.

16.4.1. Example equities

Suppose the average real return on equities over $T = 100$ years was $\widehat{\alpha} = 10\%$; the standard deviation of real returns $s = 20\%$ and they appeared normally distributed (in reality equity returns are not quite normally distributed). For a Random Variable Z following a standard normal distribution

z	0.5	1	1.5	2	2.5	3
$P(Z < z)$	0.6915	0.8413	0.9332	0.9772	0.9938	0.9987

(a) Explain what $P(Z < z)$ means. What is $P(Z < 0)$?
(b) What is the standard error of the mean?
(c) Is the mean return significantly different from zero?
(d) What is the probability of a return less than –50%?
(e) What is the probability of a positive return.

Answer.

(a) $P(Z < z)$ is the probability that the random variable Z takes a value less than a specified value, z. $P(Z < 0) = 0.5$ since the standard normal distribution is symmetric around zero, there is 50% below zero and 50% above.

(b) Standard error of mean is $s/\sqrt{T} = 20/\sqrt{100} = 2$.

(c) To test the hypothesis, we use the formula $\tau = (\widehat{\alpha} - \alpha_0)/s(\widehat{\alpha}), \widehat{\alpha} = 10, \alpha_0 = 0, s(\widehat{\alpha}) = 2.$: $\tau = (10 - 0)/2 = 5$ is greater than 2. So the mean return is significantly different from zero. We reject the null hypothesis that the expected return is zero at (better than) the 5% level.

(d) Probability of a return less than -50%? z=(–50-10)/20=-3. Distribution is symmetrical so $P(Z < -3) = P(Z > 3) = 1 - P(Z < 3)$: Prob=1-0.9987=0.0013 or 0.13%

(e) Probability of a positive return:

z=(0-10)/20=-0.5; P(Z>-0.5)=P(Z <0.5)=0.6915 or 69%.

Notice the importance of whether we are using the standard deviation of returns σ or the standard error of the mean $\sigma/\sqrt{T}$.

16.4.2. Background Example: clinical trials

Clinical trials tend to be done in three phases. Phase I is a small trial to determine toxicity and effective dosage. Phase II is a larger trial to determine effectiveness. Phase III is an even larger trial to compare effectiveness with alternative treatments, if any. If there is no alternative treatment, patients are randomly assigned to a treatment group who are given the drug and to a control group who are given a placebo, made to look as much like the drug as possible. The placebo effect is the fact that any treatment, however ineffective, tends to make patients get better, if they believe in it. Randomisation is important because otherwise the two groups of patients may differ in ways that influence the effect of the drug. The trials are double blind in that neither the patient not the physician knows whether the patient is getting the drug or the placebo. This is to stop the physician selecting those treated, e.g. giving it to the ones who were more ill, which would bias the result of the trial. Giving some people an ineffective placebo raises ethical issues, but so does giving the others an untried and potentially dangerous drug. Again we are trying to balance two sorts of errors.

Suppose we have 100 patients, 50 in the treatment group, 50 in the control group; 18 of the treatment group die within a time-period, 22 of the control group die; is this difference significant?

Answer

As we saw above, the standard error for an estimate of a proportion is $se\,(p) = \sqrt{pq/n}$ where n is the number of observations on which it is based, and $q = 1 - p$. We estimate $\widehat{p} = N/n$, where N is the number who die. The number of observations in the treatment group $n_1 = 50$, as is the number in the control group, n_2. The estimated proportions who die are $\widehat{p}_1 = 0.36$, $\widehat{p}_2 = 0.44$. If the number of observations n_1 and n_2 are sufficiently large, the difference of the sample proportions $\widehat{p}_1$ and $\widehat{p}_1$ will be approximately normal with mean $p_1 - p_2$ and variance

$$V(p_1 - p_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$$

where $q_i = 1 - p_i$. Our null hypothesis is $p_1 - p_2 = 0$. If the null hypothesis is true there is no difference, then our best estimate of $p = p_1 = p_2$ is (18+22)/100=0.4. and the standard error is

$$se(\widehat{p}) = \sqrt{\frac{0.4 \times 0.6}{50} + \frac{0.4 \times 0.6}{50}} \approx 0.1$$

our test statistic is then

$$\tau = \frac{\widehat{p}_1 - \widehat{p}_2}{se(\widehat{p})} = \frac{0.36 - 0.44}{0.1} = -0.8$$

This is less than two in absolute value, so we would not reject the null hypothesis that the proportion who died was the same in the treatment and control group. The differences could have easily arisen by chance. To check this we would need to do a larger trial. Barrow chapter 7 discusses these issues.

It should not make a difference, but in practice how you frame the probabilities, e.g. in terms of proportion who die or proportion who survive, can influence how people respond.

17. Bivariate Regression

A large part of the use of statistics in economics and finance (econometrics) involves measuring the effect of one variable (e.g. price) on another variable (e.g. quantity demanded). Regression is the statistical tool used to measure the effects. In this case price would be the independent variable or regressor and quantity demanded the dependent variable. Barrow Chapters 7 and 8 discusses this material.

17.1. Example: CAPM.

Suppose the risk free interest rate over a period is R_t, the return on a particular stock is R_t^i and the return on the stock market (e.g. the FTSE or S&P index) was R_t^m. These returns would usually be measured as the changes in the logarithms of the stock prices. The Capital Asset Pricing Model (CAPM) can be written as a regression

$$\left(R_t^i - R_t\right) = \alpha + \beta \left(R_t^m - R_t\right) + u_t$$

the excess return on stock i is equal to a constant α (which should be zero) plus a coefficient β times the excess return on the market, plus a random error or

disturbance, which reflects the factors that shift the return on stock i other than movements of the whole market. The riskiness of a stock is measured by β, if $\beta = 1$ it is as volatile as the market; if $\beta > 1$, it is more volatile than the market; if $\beta < 1$, it is less volatile than the market. The riskier the stock, the higher the return required relative to the market return. Given data on R_t^i, R_t and R_t^m, for time periods $t = 1, 2, ..., T$ we want to estimate α and β for the stock and determine how much of the variation of the stock's returns can be explained by variation in the market. Verbeek, section 2.7 discusses this example in more detail.

17.2. Example: Fisher Relation

The real interest rate is the nominal interest rate R_t less the rate of inflation π_t

$$r_t = R_t - \pi_t$$

suppose the real interest rate is roughly constant, equal to a constant plus a random error

$$r_t = r + u_t$$

then we can write

$$R_t = r_t + \pi_t = r + \pi_t + u_t.$$

Then if we ran a regression

$$R_t = \alpha + \beta\pi_t + u_t$$

the theory says $\alpha = r$ and $\beta = 1$, the hypothesis $\beta = 1$ can be tested. The interpretation of α is that it is the rate of interest that would be expected on average when the rate of inflation is zero. β tells you how much the interest rate rises in response to a rise in inflation by a percentage point.

17.3. Deriving the Least Squares Estimator

In both the examples above, there is data, $t = 1, 2, ..., T$, and a model of the form

$$Y_t = \alpha + \beta X_t + u_t,$$

and we will continue to assume that u_t is a random variable with expected value zero and variance σ^2 and the observations are uncorrelated or independent through time, i.e. $E(u_t) = 0$, $E(u_t^2) = \sigma^2$, $E(u_t u_{t-i}) = 0$. We will further assume that the

independent variable varies, $Var(X_t) \neq 0$, and is independent of the error so that the covariance between them is zero $E\{(X_t - E(X_t))u_t\} = 0$. If we can estimate α and β, by $\widehat{\alpha}$ and $\widehat{\beta}$, then we can predict Y_t for any particular value of X:

$$\widehat{Y}_t = \widehat{\alpha} + \widehat{\beta}X_t$$

these are called the fitted or predicted values of the dependent variable. We can also estimate the error:

$$\widehat{u}_t = Y_t - \widehat{Y}_t = Y_t - (\widehat{\alpha} + \widehat{\beta}X_t)$$

these are called the residuals. Notice we distinguish between the true unobserved errors, u_t, and the residuals, $\widehat{u}_t$ the estimates of the errors.

As with the expected value above there are two procedures that we will use to derive the estimates, method of moments and least squares.

17.3.1. Method of Moments

Our two population assumptions (moment-conditions) are that the expected values of the errors are zero, $E(u_t) = 0$ and the covariance of the independent variables and the errors are zero: $E\{(X_t - E(X_t)u_t\} = 0$. We choose our estimates so that the sample equivalents of these equations are true. The sample equivalent of $E(u_t) = 0$ is that the mean of the estimated errors is zero

$$\begin{aligned} T^{-1}\{\sum_t \widehat{u}_t\} &= T^{-1}\left\{\sum_t (Y_t - \widehat{\alpha} - \widehat{\beta}X_t)\right\} = 0 \\ T^{-1}\left\{\sum_t Y_t - T\widehat{\alpha} - \widehat{\beta}\sum_t X_t\right\} &= T^{-1}\sum_t Y_t - \widehat{\alpha} - \widehat{\beta}(T^{-1}\sum_t X_t) = 0 \\ \widehat{\alpha} &= \overline{Y} - \widehat{\beta}\overline{X} \end{aligned}$$

Our first moment-condition implies that the estimate of $\widehat{\alpha}$ is the mean of Y minus $\widehat{\beta}$ times the mean of X. We do not yet know what $\widehat{\beta}$ is, but as long as we define $\widehat{\alpha}$ this way, the errors will sum to zero, whatever the value of $\widehat{\beta}$. We can substitute this estimate of $\widehat{\alpha}$ into the estimated equation

$$Y_t = \widehat{\alpha} + \widehat{\beta}X_t + \widehat{u}_t = \left(\overline{Y} - \widehat{\beta}\overline{X}\right) + \widehat{\beta}X_t + \widehat{u}_t$$

which we can write

$$Y_t - \overline{Y} = \widehat{\beta}(X_t - \overline{X}) + \widehat{u}_t$$

using lower case letters to denote deviations from the mean, this is.

$$y_t = \widehat{\beta} x_t + \widehat{u}_t \tag{17.1}$$

We use this to find $\widehat{\beta}$. The sample equivalent of our second moment-condition: $E\{(X_t - E(X_t)u_t\} = 0$ is

$$\begin{aligned} T^{-1}\sum_t x_t \widehat{u}_t &= 0 \\ T^{-1}\sum_t x_t(y_t - \widehat{\beta} x_t) &= T^{-1}\sum_t x_t y_t - \widehat{\beta}\left\{T^{-1}\sum_t x_t^2\right\} = 0 \\ \widehat{\beta} &= \left\{T^{-1}\sum_t x_t y_t\right\} / \left\{T^{-1}\sum_t x_t^2\right\} = \left\{\sum_t x_t y_t\right\} / \left\{\sum_t x_t^2\right\}. \end{aligned}$$

This says that our estimate of $\widehat{\beta}$ is the ratio of the (population) covariance of X_t and Y_t to the variance of X_t, (remember lower case letters denote deviations from the means).

17.3.2. Least squares

As with the expected value we can also find the $\widehat{\beta}$ that minimises $\sum \widehat{u}_t^2$, where from (17.1)

$$\sum_t \widehat{u}_t^2 = \sum_t \left(y_t - \widehat{\beta} x_t\right)^2 = \sum_t y_t^2 + \widehat{\beta}^2 \sum_t x_t^2 - 2\widehat{\beta}\sum_t x_t y_t$$

the derivative of $\sum \widehat{u}_t^2$ with respect to $\widehat{\beta}$ is

$$\frac{\partial \sum \widehat{u}_t^2}{\partial \widehat{\beta}} = 2\widehat{\beta}\sum_t x_t^2 - 2\sum_t x_t y_t = 0 \tag{17.2}$$

Writing this $2\widehat{\beta}\sum_t x_t^2 = 2\sum_t x_t y_t$ and dividing both sides by $2\sum_t x_t^2$; gives $\widehat{\beta} = \sum_t x_t y_t / \sum_t x_t^2$, as before.

The second order condition is

$$\frac{\partial^2 \sum \widehat{u}_t^2}{\partial \widehat{\beta}^2} = 2\sum x_t^2 > 0 \tag{17.3}$$

since squares are positive, so this is a minimum.

Our estimates

$$\begin{aligned}\widehat{\alpha} &= \overline{Y} - \widehat{\beta}\overline{X} \\ \widehat{\beta} &= \frac{\sum(X_t - \overline{X})(Y_t - \overline{Y})}{\sum(X_t - \overline{X})^2}\end{aligned}$$

(i) make the sum of the estimated residuals zero and the estimated residuals uncorrelated with the explanatory variable and (ii) minimise the sum of squared residuals.

17.4. Properties of the estimates

Since

$$\widehat{\beta} = \frac{\sum x_t y_t}{\sum x_t^2} = \frac{\sum x_t(\beta x_t + u_t)}{\sum x_t^2} = \beta + \frac{\sum x_t u_t}{\sum x_t^2}$$

then $E(\widehat{\beta}) = \beta$, and it is unbiased; since because of independence

$$E\left\{\frac{\sum x_t u_t}{\sum x_t^2}\right\} = E\left\{\frac{\sum x_t}{\sum x_t^2}\right\} E(u_t)$$

and $E(u_t) = 0$. To derive the variance of $\widehat{\beta}$, note since $\widehat{\beta}$ is unbiased

$$V(\widehat{\beta}) = E(\widehat{\beta} - \beta)^2 = E(\frac{\sum x_t u_t}{\sum x_t^2})^2 = \frac{\sigma^2}{\sum x_t^2} \tag{17.4}$$

using the same sort of argument as in deriving the standard error of the mean above.

17.5. Measuring how well the regression fits

Note if we define the covariance between X and Y : $S_{xy} = \sum(X_t - \overline{X})(Y_t - \overline{Y})/T$, and the variance $S_{xx} = \sum(X_t - \overline{X})^2/T$, then $\widehat{\beta} = S_{xy}/S_{xx}$ as we saw above and the correlation coefficient is:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

as we saw in section 6. The correlation coefficient lies between -1 (perfect negative relation) and +1 (perfect positive relation) with 0 indicating no linear relationship.

Notice that covariance and correlation only measure linear relationships. You could have an exact non-linear relationship (e.g. a circle) and the correlation would be zero. In regression we use the square of the correlation coefficient, r^2, usually written R^2 and called the coefficient of determination. This gives you the proportion of the variation in Y that has been explained by the regression.

We measure the dispersion around the line in exactly the same way that we measured the dispersion around the mean, either using the biased estimator

$$\widehat{\sigma}^2 = \sum \widehat{u}_t^2/T$$

or the unbiased estimator

$$s^2 = \sum \widehat{u}_t^2/(T-2)$$

where $\widehat{u}_t = Y_t - \widehat{\alpha} - \widehat{\beta}X_t$. Now there are $T-2$ degrees of freedom because we estimated two parameters $\widehat{\alpha}$ and $\widehat{\beta}$. The estimate s^2 is called the variance of the regression and its square root s, the Standard Error of The Regression (SER), which gives you an idea of the average size of the errors. The SER is measured in the same units as the dependent variable. If the dependent variable is a logarithm, the SER can be multiplied by 100 and interpreted as a percentage error. We can then measure the standard error of $\widehat{\beta}$ as $se(\widehat{\beta}) = s/\sqrt{\sum x_t^2}$ by putting the estimate in (17.4) and taking square roots.

The coefficient of determination, or R squared, the proportion of the variation of Y that has been explained, or 1 minus the proportion of the variation in Y that has not been explained. is defined as

$$R^2 = 1 - \frac{\sum \widehat{u}_t^2}{\sum (Y_t - \overline{Y})^2}$$

Show this is the same as r^2 defined above. If $\widehat{\beta} = 0$, then nothing has been explained: $\widehat{u}_t = Y_t - \widehat{\alpha}$, where $\widehat{\alpha}$ is just the mean and $R^2 = 0$.

Computer packages also often calculate adjusted R^2 or $\overline{R}^2$ (R bar squared). This corrects the numerator and denominator for degrees of freedom:

$$\overline{R}^2 = 1 - \frac{\sum \widehat{u}_t^2/(T-k)}{\sum (Y_t - \overline{Y})^2/(T-1)}$$

where k is the number of explanatory variables, including the intercept. This can be negative.

From our estimates we can also calculate the predicted value of Y for a particular X:

$$\widehat{Y}_t = \widehat{\alpha} + \widehat{\beta} X_t$$

and the residual or unexplained part of Y for a particular observation is:

$$\widehat{u}_t = Y_t - \widehat{Y}_t = Y_t - (\widehat{\alpha} + \widehat{\beta} X_t)$$

17.6. Assumptions for Least Squares to give good estimates

We are assuming:

That we have the correct model for the process, e.g. that it is linear and we have not left any relevant variables out.

That the expected values of the errors is zero: $E(u_t) = 0$, on average the true errors are zero. Notice that the average of our residuals is always zero, ($T^{-1} \sum \widehat{u}_t = 0$ by construction), as long as we have an intercept in the equation or work with deviations from the mean.

That $E(u_t^2) = \sigma^2$, errors have constant variance. This assumption is sometimes expressed the errors are homoskedastic (their variances are the same), its failure is that the errors are heteroskedastic (their variances are different).

$E(u_t u_{t-i}) = 0$, for $i \neq 0$. Different errors are independent. This assumption is also called: no serial correlation or no autocorrelation.

That $\sum (X_t - \overline{X})^2 \neq 0$,the X's must vary over the sample. We cannot calculate $\widehat{\beta}$ if this is not the case. This would also fail if there were not enough observations to estimate the parameters. We need $T > 2$. If $T = 2$, we can fit the observations exactly by a line through the two points.

That $E\{(X_t - E(X_t)u_t\} = 0$. This assumption is usually described as the X's being exogenous: not related to the errors. Notice that for our estimates $T^{-1} \sum_t (X_t - \overline{X})\widehat{u}_t = 0$ by construction. For exogeneity, the X's may either be non stochastic, fixed numbers, though this is rare in economics where our X variables are usually random or random variables distributed independently of the errors. Independence implies that X and the errors are not correlated, but is a stronger assumption than being uncorrelated.

Properties of estimators were discussed in section 15.3. With all these assumptions we can show that among all estimators of α and β that are linear functions of Y and are unbiased; the least squares estimator has the smallest variance. This is the Gauss-Markov theorem: under these assumptions the least squares estimator is the Best (minimum variance) Linear Unbiased Estimator, it is BLUE. If in

addition we add another assumption to the model, that the errors are normally distributed, then our estimates will also be normally distributed and we can use this to construct test statistics to test hypotheses about the regression coefficients. Even if the errors are not normally distributed, by the central limit theorem our estimates will be normally distributed in large samples; in the same way that the mean is normally distributed whatever the distribution of the variable in large samples.

We need the assumption that X is exogenous, to make causal statements about the effect of X on Y. When we are only interested in predicting Y, as in the height-weight example, we do not need the exogeneity assumption and have the result that the least squares prediction $\widehat{Y}_t$ is the Best (minimum variance) Linear Unbiased Predictor of Y_t.

17.7. Predicted values and residuals.

We assume that there is a true model or "data generating process" as its sometimes called: $Y_t = \alpha + \beta X_t + u_t$. We estimate $Y_t = \widehat{\alpha} + \widehat{\beta} X_t + \widehat{u}_t$ or $Y_t = \widehat{Y}_t + \widehat{u}_t$. Thus the least squares procedure splits Y_t into two bits, the explained bit, the expected or predicted value, and the unexplained bit, the residual, $\widehat{u}_t$. Its called the residual because its the bit left over after we have explained all we can. The predicted value is an estimate of the conditional expectation for Y conditional on X: $\widehat{Y}_t = E(Y_t \mid X_t) = \widehat{\alpha} + \widehat{\beta} X_t$.

Notice that the predicted values and the residuals are uncorrelated, their covariance is exactly zero:

$$\sum_{t=1}^{T} \widehat{Y}_t \widehat{u}_t = \sum (\widehat{\alpha} + \widehat{\beta} X_t) \widehat{u}_t = \widehat{\alpha} \sum \widehat{u}_t + \widehat{\beta} \sum X_t \widehat{u}_t$$

But our moment-condition was that $\sum \widehat{u}_t = 0$ and the least squares estimate of α is chosen to make this true, so the first term is zero, while our second moment-condition was $\sum X_t \widehat{u}_t = 0$, so the second term is also zero. So the predicted or fitted values are uncorrelated with the estimated residuals. One of the main uses of the predicted values is in forecasting, we make an assumption about how X will change in the future, use the equation to forecast Y and calculate a standard error for the forecast.

By construction the residuals have mean zero (if there is an intercept, i.e. α, in the equation, and you should always include an intercept) and they are uncorrelated with the explanatory variables. But we can check whether other

assumptions we made about the errors hold for the estimated residuals. We can test whether our assumption that $E(u_t^2) = \sigma^2$ a constant, holds in the data. We can also test whether $E(u_t u_{t-i}) = 0$, for i not equal 0. This is the assumption of independence, or no serial correlation or no autocorrelation. We might also have assumed that u_t is normally distributed and we can test whether the skewness and kurtosis of the residuals are those of a normally distributed variable. We will discuss how we test these assumptions later, but in many cases the best way to check them is to look at the pictures of the actual and predicted values and the residuals. The residuals should look random, with no obvious pattern in them and the histogram should look roughly normal.

What do you do if you have unhealthy residuals, that show serial correlation or heteroskedasticity? The text books tend to suggest that you model the disturbances, typically by a procedure called Generalised Least Squares. However, in most cases the problem is not that the true disturbances are heteroskedastic or serially correlated. The problem is that you have got the wrong model, and the error in the way you specified the model shows up in the estimated residuals. Modelling the disturbances often is just treating the symptoms, the solution is to cure the disease: get the model specified correctly.

18. Multiple Regression.

18.1. Example and logarithmic models

Most of the time we have more than one right hand side variable, so our regression may be a demand function like

$$\ln Q_t = \beta_1 + \beta_2 \ln Y_t + \beta_3 \ln P_t + \beta_4 \ln P_t^* + u_t \tag{18.1}$$

where Q_t is quantity demanded, Y_t real income and P_t the price of the good, P_t^* a measure of the price of all other goods, and ln denotes natural logarithms. Given the log equation then β_2 is the income elasticity of demand (the percentage change in demand in response to a one percent change in income), which we would expect to be positive, and β_3 the own price elasticity, which we expect to be negative and β_4 is the cross-price elasticity, which for all other goods should be positive. It is standard to use logarithms of economic variables since (a) prices and quantities are non-negative so the logs are defined (b) the coefficients can be interpreted as elasticities, so the units of measurement of the variables do not matter (c) in many cases errors are proportional to the variable, so the variance is more likely

to be constant in logs, (d) the logarithms of economic variables are often closer to being normally distributed (e) the change in the logarithm is approximately equal to the growth rate and (f) lots of interesting hypotheses can be tested in logarithmic models. For instance in this case if $\beta_3 = -\beta_4$ (homogeneity of degree zero)only relative prices matter. Notice the original model is non-linear

$$Q_t = BY_t^{\beta_2} P_t^{\beta_3} P_t^{*\beta_4} \exp(u_t)$$

where $B = \exp(\beta_1)$, but can be made linear by taking logs.

Another common logarithmic model is the Cobb-Douglas production function explaining output at time t Q_t, by capital K_t and labour L_t and an error

$$Q_t = AK_t^b L_t^c e^{dt+u_t}$$

Notice output will be zero if either capital or labour are zero. We can make this linear by taking logarithms

$$\ln Q_t = \ln A + b \ln K_t + c \ln L_t + dt + u_t. \tag{18.2}$$

The rate of growth of technical progress is measured by d, it is the amount log output changes between periods if all inputs are constant. The residual u_t is often treated as a measure of efficiency, how much higher or lower output is than you would expect.

If $b + c = 1$ there is constant returns to scale: if both inputs go up by 10%, output goes up by 10%. We can test this by rewriting the equation

$$\ln Q_t - \ln L_t = \ln A + b \left[\ln K_t - \ln L_t\right] + (b + c - 1) \ln L_t + dt + u_t \tag{18.3}$$

and do a t test on the coefficient of $\ln L_t$, which should be not significantly different from zero if there is constant returns to scale. Notice (18.2) and (18.3) are identical statistical equations, e.g. the estimates of the residuals would be identical.

18.2. Matrix form of the Linear Regression Model

We could write (18.1)

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t$$

where the dependent variable Y_t corresponds to $\ln Q_t$, X_{1t} is a variable that always takes the value one so we do not need to write it in, X_{2t} corresponds to $\ln Y_t$, and

X_{3t} to $\ln P_t$, X_{4t} to $\ln P_t^*$. The problem is the same as before. We want to find the estimates of β_i, $i = 1, 2, 3, 4$ that minimise the sum of squared residuals, $\sum \widehat{u}_t^2$.

$$\sum \widehat{u}_t^2 = \sum (y_t - \widehat{\beta}_1 - \widehat{\beta}_2 X_{2t} - \widehat{\beta}_3 X_{3t} - \widehat{\beta}_4 X_{4t})^2$$

we have to multiply out the terms in the brackets and take the summation inside and derive the first order conditions, the derivatives with respect to the four parameters. These say that the residuals should sum to zero and be uncorrelated all four X_{it} . The formulae, expressed as summations are complicated. It is much easier to express them in matrix form. Verbeek Appendix A reviews matrix algebra.

We can write this is vector form

$$Y_t = \beta' X_t + u_i$$

where β and X_i are 4×1 vectors, so the product $\beta' X_i$ is $(1 \times 4) \times (4 \times 1) = 1 \times 1$ a scalar, just like Y_i.

We can also write this in matrix form in terms of y a $T \times 1$ vector and X a $T \times 4$ matrix

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ Y_T \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & X_{41} \\ 1 & X_{22} & X_{32} & X_{42} \\ . & . & . & . \\ 1 & X_{2T} & X_{3T} & X_{4T} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_3 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ . \\ u_T \end{bmatrix}$$

$$\underset{(T \times 1)}{y} = \underset{(T \times 4)}{X} \underset{(4 \times 1)}{\beta} + \underset{(T \times 1)}{u} .$$

This gives us a set of T equations. Notice, in writing X_{it}, we have departed from the usual matrix algebra convention of having the subscripts go row column. This generalises to the case where X is a $T \times k$ matrix and β a $k \times 1$ vector, whatever k. Notice that for matrix products, the inside numbers have to match for them to be conformable and the dimension of the product is given by the outside numbers.

18.3. Assumptions

We now want to express our assumptions about the errors in matrix form. The assumptions were: (a) that $E(u_t) = 0$, on average the true errors are zero; (b) that $E(u_t^2) = \sigma^2$, errors have constant variance; and (c) $E(u_t u_{t-i}) = 0$, for $i \neq 0$,

different errors are independent. The first is just that the expected value of the random $T \times 1$ vector u is zero $E(u) = 0$. To capture the second and third assumptions, we need to specify the variance covariance matrix of the errors, $E(uu')$ a $T \times T$ matrix. u' is the transpose of u, a $1 \times T$ vector. The transpose operation turns columns into rows and vice versa. Note $u'u$ is a scalar, 1×1 the sum of squared errors. Writing out $E(uu')$and putting our assumptions in:

$$\begin{aligned} E(uu') &= \begin{bmatrix} E(u_1^2) & E(u_1u_2) & .. & E(u_1u_T) \\ E(u_1u_2) & E(u_2^2) & .. & E(u_2u_T) \\ .. & .. & .. & .. \\ E(u_1u_T) & E(u_2u_T) & .. & E(u_T^2) \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2 & 0 & .. & 0 \\ 0 & \sigma^2 & .. & 0 \\ .. & .. & .. & .. \\ 0 & 0 & .. & \sigma^2 \end{bmatrix} \end{aligned}$$

So our assumptions say that $E(uu') = \sigma^2 I_T$. Where I_T is a $T \times T$ identity matrix with ones on the diagonal and zeros on the off diagonal.

The assumption that X is exogenous, distributed independently of the errors, u, implies $E(X'u) = 0$, which corresponds to our earlier assumption $E(X_t - \overline{X})u_t = 0$. We also assume that X has full rank k. This implies that the different regressors vary independently, are not perfectly correlated, and corresponds to our earlier assumption that X_t.varies.

18.4. Estimating $\widehat{\beta}$

As before we will consider two methods for deriving the estimators method of moments and least squares. This time for the model $y = X\beta + u$, where y and u are $Tx1$ vectors, β is a $k \times 1$ vector and X a $T \times k$ matrix.

18.4.1. Method of moments

Our exogeneity assumption is $E(X'u) = 0$, the sample equivalent is $X'\widehat{u} = 0$, a $k \times 1$ set of equations, which for the case $k = 4$ above, gives

$$\begin{aligned} \sum \widehat{u}_t &= 0; \\ \sum X_{2t}\widehat{u}_t &= 0; \\ \sum X_{3t}\widehat{u}_t &= 0; \\ \sum X_{4t}\widehat{u}_t &= 0. \end{aligned}$$

So

$$X'\widehat{u} = X'(y - X\widehat{\beta}) = X'y - X'X\widehat{\beta} = 0.$$

Since X is of rank k, $(X'X)^{-1}$ exists ($X'X$ is non-singular, its determinant is non-zero) so

$$\widehat{\beta} = (X'X)^{-1}X'y.$$

18.4.2. Least Squares

The sum of squared residuals is

$$\begin{aligned} \widehat{u}'\widehat{u} &= (y - X\widehat{\beta})'(y - X\widehat{\beta}) \\ &= y'y + \widehat{\beta}'X'X\widehat{\beta} - 2\widehat{\beta}'X'y \end{aligned}$$

To derive the least square estimator, we take derivatives, and set them equal to zero. If β is a $k \times 1$ vector we get k derivatives, the first order conditions are the $k \times 1$ set of equations,

$$\frac{\partial \widehat{u}'\widehat{u}}{\partial \widehat{\beta}} = 2X'X\widehat{\beta} - 2X'y = 0$$

Compare this to (17.2). So the least squares estimator is $\widehat{\beta} = (X'X)^{-1}X'y$ as before. Again our assumptions ensures that $(X'X)^{-1}$ exists The second order condition is

$$\frac{\partial^2 \widehat{u}'\widehat{u}}{\partial \widehat{\beta}\partial \widehat{\beta}'} = 2X'X$$

which is a positive definite matrix, ensuring a minimum. Compare this to (17.3).

19. Properties of Least Squares

We can derive the expected value of $\widehat{\beta}$.

$$\begin{aligned}\widehat{\beta} &= (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + u) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u = \beta + (X'X)^{-1}X'u\end{aligned}$$

So

$$\widehat{\beta} = \beta + (X'X)^{-1}X'u \tag{19.1}$$

$$E(\widehat{\beta}) = \beta + E((X'X)^{-1}X'u)$$

since β is not a random variable, and if X and u are independent $E((X'X)^{-1}X'u) = E((X'X)^{-1}X')E(u) = 0$ since $E(u) = 0$. Thus $E(\widehat{\beta}) = \beta$ and $\widehat{\beta}$ is an ubiased estimator of β.

From (19.1) we have

$$\widehat{\beta} - \beta = (X'X)^{-1}X'u$$

The variance-covariance matrix of $\widehat{\beta}$ is

$$E(\widehat{\beta} - E(\widehat{\beta}))(\widehat{\beta} - E(\widehat{\beta}))' = E(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)'$$

since $\widehat{\beta}$ is unbiased. But from (19.1) we have

$$\widehat{\beta} - \beta = (X'X)^{-1}X'u$$

so

$$\begin{aligned}E(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)' &= E((X'X)^{-1}X'u)((X'X)^{-1}X'u)' \\ &= E((X'X)^{-1}X'u\ u'X(X'X)^{-1} \\ &= (X'X)^{-1}X'E(u\ u')X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}\end{aligned}$$

since $E(uu') = \sigma^2 I$, σ^2 is a scalar, and $(X'X)^{-1}X'X = I$. Compare this to (17.4) above.

If we also assume normality then we can write that $u \sim N(0, \sigma^2 I)$. Linear functions of normally distributed variables are also normal. We saw that if one

variable Y is normally distributed with mean α, and variance σ^2 . Then any linear function of Y is also normally distributed.

$$\begin{aligned} Y &\sim N\left(\alpha, \sigma^2\right) \\ X &= a + bY \sim N\left(a + b\alpha, b^2\sigma^2\right) \end{aligned}$$

where a and b are scalars. The matrix equivalent of this is that, for $k \times 1$ vectors, Y and M and $k \times k$ matrix Σ (not the summation sign) $Y \sim N\left(M, \Sigma\right)$, then for $h \times 1$ vectors X and A, and $h \times k$ matrix B

$$X = A + BY \sim N\left(A + BM, B\Sigma B'\right)$$

Notice the variance covariance matrix of X say $V(X) = B\Sigma B'$ is $(h \times k) \times (k \times k) \times (k \times h) = h \times h$.

Since y is a linear function of u it follows that y is also normally distributed: $y \sim N(X\beta, \sigma^2 I)$. In this case B is the identity matrix. Since $\widehat{\beta} = (X'X)^{-1}X'y$ is a linear function of y, it is also normally distributed.

$$\begin{aligned} \widehat{\beta} &\sim N((X'X)^{-1}X'X\beta, (X'X)^{-1}X'\left(\sigma^2 I\right)X(X'X)^{-1} \\ &\sim N(\beta, \sigma^2(X'X)^{-1}) \end{aligned}$$

in this case A is zero, $B = (X'X)^{-1}X'$. $(X'X)^{-1}$ is equal to its transpose because it is a symmetric matrix. $(X'X)^{-1}X'X = I$.

This says that $\widehat{\beta}$ is normally distributed with expected value β (and is therefore unbiased) and variance covariance matrix $\sigma^2(X'X)^{-1}$.

The variance covariance matrix is a $k \times k$ matrix and we estimate it by

$$V(\widehat{\beta}) = s^2(X'X)^{-1}$$

where $s^2 = \widehat{u}'\widehat{u}/(T-k)$. The square roots of the diagonal elements of $s^2(X'X)^{-1}$ give the standard errors of the estimates of the individual elements of β, which are reported by computer programs.

19.1. Ommitted variables

When you add another variable, the estimate of the coefficient of X will generally change.

Express the data in deviations from mean, to get rid of all the constant terms. Suppose we first just include x_t, leaving out z_t and get estimates of:

$$y_t = bx_t + v_t. \tag{19.2}$$

Then we run another regression which includes z_t:

$$y_t = \beta x_t + \gamma z_t + u_t \tag{19.3}$$

We get two estimates of the coefficient on x_t : b and β. What is the relation between them? To understand this we need to look at the relationship between x_t and z_t. We can summarise the relationship between x_t and z_t by another regression equation:

$$z_t = dx_t + w_t \tag{19.4}$$

w_t is just the bit of z_t that is not correlated with x_t. d may be zero, if there is no relationship. Put (19.4) into (19.3)and we get (19.2):

$$y_t = \beta x_t + \gamma(dx_t + w_t) + u_t$$

$$y_t = (\beta + \gamma d)x_t + (\gamma w_t + u_t)$$

So $b = (\beta + \gamma d)$, the coeficient of x_t picks up the bit of z_t that is correlated with x_t. Bits of z_t that are not correlated with x_t end up in the error term $(\gamma w_t + u_t)$. This is why looking for patterns in the error term is important, it may suggest what variable you have left out. If you add a variable that is not correlated with x_t, the coefficient of x_t will not change. If you add a variable that is highly correlated with x_t, the coefficient of x_t will change a lot.

20. Tests in regression

20.1. Tests for a single hypothesis on individual coefficients

Suppose we have the model

$$Y_t = \alpha + \beta X_t + \gamma Z_t + u_t$$

we can test the significance of the individual coefficients using t ratios exactly as we did for the mean

$$t(\beta = 0) = \frac{\widehat{\beta}}{se\left(\widehat{\beta}\right)}$$

where $se\left(\widehat{\beta}\right)$ is the estimated standard error of $\widehat{\beta}$. This tests the null hypothesis $H_0 : \beta = 0$. If this t ratio is greater than two in absolute value we conclude that

$\widehat{\beta}$ is significant: significantly different from zero at the 5% level. Computers often print out this t ratio automatically. They usually give the coefficient, the standard error, the t ratio and the p value. The p value gives you the probability that the null hypothesis is true. If it was less than 0.05, we would reject the hypothesis at the 5% level.

We could test against other values than zero. Suppose economic theory suggested that $\gamma = 1$ the t statistic for testing this would be

$$t(\gamma = 1) = \frac{\widehat{\gamma} - 1}{se\left(\widehat{\gamma}\right)}$$

and if this t statistic is greater than two in absolute value we conclude that $\widehat{\gamma}$ is significantly different from unity at the 5% level.

20.2. Tests on joint hypotheses

Suppose that we wanted to test the hypothesis that none of the independent variables had any effect on the dependent variable in

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t.$$

The hypothesis is that $\beta_2 = \beta_3 = \beta_4 = 0$. The test statistics used to test joint hypotheses follow a different distribution called the F distribution, introduced in section 12.2. Just as the t distribution is described by its degrees of freedom, the F distribution is described by its degrees of freedom, though it has two of them. The first is the number of hypotheses being tested, three in this case, and the second is the degrees of freedom, $T - 4$ in this case. This would be written $F(3, T-4)$ and critical values are given in statistics books. The F statistic for this hypothesis (that all the slope coefficients are equal to zero) is often printed out by computers, they also usually give a p value. With the p value you do not have to look up tables, just reject the null hypothesis if $p < 0.05$. Notice that the joint hypothesis that both coefficients are equal to zero can give different conclusions from a sequence of individual hypotheses that each are equal to zero, they are testing different hypotheses.

20.3. Diagnostic Tests for our assumptions about the errors.

If our assumptions about the errors are valid, the estimated residuals should be normally distributed and random: without any pattern in them, so our null hypothesis is that the model is well specified and there is no pattern in the residuals,

e.g. other variables should not be able to explain them. Our alternative hypothesis is that the model is misspecified in a particular way, and since there are lots of ways that the model could be misspecified (the errrors could be serially correlated, heteroskedastic, non-normal or the model could be non-linear) there are lots of these tests, each testing the same null, the model is well specified, against a particular alternative that the misspecification takes a particular form. This is like the fact that there are lots of different diagnostic tests that doctors use. There are lots of different ways that a person, or a regression, can be sick.

The Durbin-Watson test for serial correlation is a diagnostic test for serial correlation. It is given by

$$DW = \frac{\sum_{t=2}^{T}(\widehat{u}_t - \widehat{u}_{t-1})^2}{\sum_{t=1}^{T}\widehat{u}_t^2}$$

it should be around 2, say 1.5 to 2.5. If it is below 1.5 there is positive serial correlation, residuals are positively correlated with their previous (lagged) values, above 2.5 negative serial correlation. It is only appropriate if (a) you are interested in first order serial correlation; (b) there is an intercept in the equation, so the residuals sum to zero and (c) there is no lagged dependent variable in the equation. First order (one lag) serial correlation assumes that errors are related to their values in the previous period

$$u_t = \rho u_{t-1} + \varepsilon_t$$

but there may be higher order serial correlation. For instance,. in quarterly data, the errors may be related to errors up to a year ago: the size of the error in the alcohol equation at Christmas ($Q4$) is related not just to the previous quarters error but to the size of the error last Christmas:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + \rho_4 u_{t-4} + \varepsilon_t$$

this is fourth order (four lags) serial correlation. Suppose you ran a regression

$$y_t = \beta' x_t + u_t$$

The test involves running a regression of the residuals on the variables included in the original regression and the lagged residuals

$$\widehat{u}_t = b'x + \rho_1 \widehat{u}_{t-1} + \rho_2 \widehat{u}_{t-2} + \rho_3 \widehat{u}_{t-3} + \rho_4 \widehat{u}_{t-4} + \varepsilon_t$$

then testing the joint hypothesis $\rho_1 = \rho_2 = \rho_3 = \rho_4$.

There are many diagnostic tests which involve regressing the estimated residuals or powers of the residuals on particular variables. Technically, most of these tests are known as Lagrange Multiplier Tests. It is important that you check your equation for various diseases before you regard it as healthy enough to be used. Statistical packages like EViews (section 22.2), Microfit, Stata, etc. have built in tests of the assumptions that are required for Least Squares estimates to be reliable. If the assumptions do not hold the estimated standard errors are likely to be wrong and corrected standard errors that are 'robust' to the failure of the assumptions are available.

21. Economic and Financial Data III: Variables and Relationships

Linear regression is very flexible because we can redefine the variables by some transformation, which allows the underlying relationship to be non-linear, but the relationship we estimate to be linear, as in 18.1. We can also construct variables, like the trend in 18.1.

21.1. Dummy variables.

Suppose that we had UK annual data on consumption and income for 1930 to 1960 and wanted to estimate a consumption function. This period includes the second world war, 1939-1945, when there was rationing and consumption was restricted. This would shift the consumption function and could be allowed for by estimating an equation

$$C_t = \alpha + \beta Y_t + \gamma D_t + u_t$$

where D_t is a "Dummy" variable which takes the value one 1939-45 and zero in other years. The intercept during the War is then $\alpha + \gamma$ and we would expect $\gamma < 0$. We could also write this

$$C_t = \delta_1 D_t + \delta_2(1 - D_t) + \beta Y_t + u_t \tag{21.1}$$

we would get an identical estimate $\widehat{\beta}$, in both equations, $\widehat{\delta}_1$ is the estimated intercept 1939-45, $\widehat{\delta}_2 = \widehat{\alpha}$ is the intercept for other years and $\widehat{\gamma} = \widehat{\delta}_1 - \widehat{\delta}_2$, the difference in the intercept between the two periods. Notice that had you included

a constant in (21.1) the computer would have refused to estimate it and told you that the data matrix was singular. This is known as 'the dummy variable trap'.

A similar technique allows for seasonal effects. Suppose that we had quarterly data on consumption and income, and wanted to allow for consumption to differ by quarters (e.g. spending more at Christmas). Define $Q1_t$ as a dummy variable that is one in quarter one and zero otherwise; $Q2_t$ is one in quarter two zero otherwise, etc. Then estimate

$$C_t = \alpha_1 Q1_t + \alpha_2 Q2_t + \alpha_3 Q3_t + \alpha_4 Q4_t + \beta Y_t + u_t$$

then the intercept in $Q1$ is α_1 in $Q2$ is α_2, etc.

21.2. Non-linearities

21.2.1. Powers

We can easily allow for non-linearities by transformations of the data as we saw with logarithms above. As another example imagine y (say earnings) first rose with x (say age) then fell. We could model this by

$$y_i = a + bx_i + cx_i^2 + u_i$$

where we would expect $b > 0$, $c < 0$. Although the relationship between y and x is non-linear, the model is linear in parameters, so ordinary least squares can be used, we just include another variable which is the square of the first. Notice that the effect of x on y is given by

$$\frac{\partial y}{\partial x} = b + 2cx_i$$

thus is different at different values of x_i, and has a maximum (or minimum) which can be calculated as the value of x that makes the first derivative zero. Earnings rise with age to a maximum then fall, self reported happiness tends to fall with age to a minimum then rise. The middle-aged are wealthy but miserable. We can extend this approach by including the product of two variables as an additional regressor. There is an example of this in the specimen exam, section 3.2 question 5. Verbeek section 3.5 has an extensive discussion.

21.2.2. Background: Regressions using Proportions

Suppose our dependent variable is a proportion, $p_t = N_t/K$, whereN_t is a number affected and K is the population, or a maximum number or saturation level. Then p_t lies between zero and one and the logistic transformation $(\ln(p_t/(1-p_t))$ is often used to ensure this. If the proportion is a function of time this gives,

$$\ln\left(\frac{p_t}{1-p_t}\right) = a + bt + u_t \tag{21.2}$$

which is an S shaped curve for p_t over time. This often gives a good description of the spread of a new good (e.g. the proportion of the population that have a mobile phone) and can be estimated by least squares. Although this is a non linear relationship in the variable p_t it is linear in parameters when transformed so can be estimated by least squares. The form of the non-linear relationship is

$$p_t = \frac{N_t}{K} = \frac{1}{1+\exp -(a+bt)} \tag{21.3}$$

We could estimate this directly, treating K as an unknown parameters in a programs like EViews which does non-linear least squares. So if N_t is the number of mobile phone owners we would enter this in Eviews as

$$N = C(1)/(1+\exp(C(2)+C(3)*@trend)). \tag{21.4}$$

@trend in EViews provides a trend, t. C(1) would be an estimate of K, C(2) of a and C(3) of b. In practice, unless the market is very close to saturation it is difficult to estimate K precisely. Notice that (21.2) and (21.4) imply different assumptions about how the error term enters (21.3) so are not equivalent.

22. Applied Exercise III: Running regressions

You can run regressions in Excel but in most cases it is easier to use a specialised package. There are many of them and if you are familiar with a particular package you can use that. EViews is a very easy package to use and is installed on our machines. Microfit is another econometrics package that is easy to use. This example uses the Shiller.xls file that was used in Applied Exercise I and is on the ASE home page. It has data 1871-2000 (130 observations) on 5 variables NSP (nominal stock prices), ND (nominal dividends) NE (nominal earnings, profits)

and R (interest rates) and PPI (producer price index). Note that there is no 2000 data on three of the variables. Figures are given on NSP and PPI which are January figures, the other three are averages for the year. Even if you are not using EViews read the explanation and carry out the exercise on the software you are using. The example below regresses dividends on earnings for the period 1871-1986. First we describe some of the output the computer produces.

22.1. Regression Output

Computer programs will print out a range of information, which may include

- the estimates of the regression coefficients $\widehat{\beta}_i$, $i = 1, ..., k$ including the constant
- the standard error of each coefficient $SE(\widehat{\beta}_i)$ which measures how precisely it is estimated,
- the t ratio $t(\beta_i = 0) = \widehat{\beta}_i / SE(\widehat{\beta}_i)$ which tests the null hypothesis that that particular coefficient is really zero (the variable should not appear in the regression). If the t ratio is greater than 2 in absolute value, we can reject the null hypothesis that $\beta_i = 0$ at about the 5% level. In this case the coefficient is said to be significantly different from zero or significant.
- the p value for the hypothesis that $\beta_i = 0$. This gives the probability that the null hypothesis is true. If this is less than 0.05 again we can reject the hypothesis that $\beta_i = 0$.
- The Sum of Squared residuals $\sum \widehat{u}_t^2$. This is what least squares minimises.
- The standard error of regression

  $$s = \sqrt{\sum \widehat{u}_t^2 / (T - k)}$$

  where k is the number of regression coefficients estimated and T the number of observations. This is an estimate of the square root of the error variance σ^2 and gives you an idea of the average size of the errors. If the dependent variable is a logarithm, multiply s by 100 and interpret it as the average percent error.

- R squared, which tells you the proportion of the variation in the dependent variable that the equation explains

$$R^2 = 1 - \frac{\sum \widehat{u}_t^2}{\sum (Y_t - \overline{Y})^2} = \frac{\sum (\widehat{Y}_t - \overline{Y})^2}{\sum (Y_t - \overline{Y})^2}$$

- R bar squared, which corrects R squared for degrees of freedom

$$\overline{R}^2 = 1 - \frac{\sum \widehat{u}_t^2/(T-k)}{\sum (Y_t - \overline{Y})^2/(T-1)}$$

  where k is the number of regression coefficients estimated and T is the number of observations. Whereas R^2 is always positive and increases when you add variables, $\overline{R}^2$ can be negative and only increases if the added variables have t ratios greater than unity.

- Durbin Watson Statistic is a measure of serial correlation of the residuals

$$DW = \frac{\sum_{t=2}^{T} (\widehat{u}_t - \widehat{u}_{t-1})^2}{\sum_{t=1}^{T} \widehat{u}_t^2}$$

  it measures whether the residuals are correlated. It should be around 2, say 1.5-2.5. It tests the hypothesis $\rho = 0$ in the autoregression $u_t = \rho u_{t-1} + \varepsilon_t$. Roughly $DW = 2(1-\rho)$. This statistic depends on the ordering of the data, since it calculates $u_t - u_{t-1}$. In time-series there is a natural ordering of the data, in cross-section there is not. So in cross-section the DW should be interpreted with caution.

- An F statistic which tests the hypothesis that none of the slope variables (i.e. the right hand side variables other than the constant $\widehat{\alpha}$) is significant. Notice that in the case of a single slope variable, this will be the square of its t statistic. Usually it also gives the probability of getting that value of the F-statistic if the slope variables all had no effect.

22.2. Excel

Go into Excel, Load the Shiller.xls file. Click Tools; Data Analysis; Choose Regression from the list of techniques. You may have to add-in the data-analysis module. Where it asks you Y range enter C2:C117. Where it asks you X range

enter D2:D117. Click Output Range and enter in the output range box G1. Alternatively you can leave it at the default putting the results in a separate sheet. Click OK. It gives you in the first box, Multiple R, which you can ignore, R squared and Adjusted R Squared and Standard Error of the Regression. Then it gives you an ANOVA box which you can ignore. Then it gives you estimates of the coefficients (intercept, X Variable 1, etc), their standard errors, t statistics, and P values, etc. shown in the summary output.

In this case we have run a regression of dividends on earnings and the results for the sample 1871 1986 are:

$$\begin{array}{rccc} ND_t = & 0.169 + & 0.456 NE_t & + \widehat{u}_t \\ & (0.036) & (0.007) & \\ & [4.67] & [61.65] & \\ & \{0.000\} & \{0.000\} & \end{array}$$

$R^2 = 0.971$, $s = 0.31$.

Standard errors of coefficients are given in parentheses, t statistics in brackets, and p values in braces. You would normally report only one of the three, usually just standard errors. The interpretation is that if earnings go up by \$10, then dividends will go up by \$4.56. If earnings were zero, dividends would be 16.9 cents. Earnings explain 97% of the variation in dividends over this period and the average error in predicting dividends is 0.31. We would expect our predictions to be within two standard errors of the true value 95% of the time. Both the intercept and the coefficient of earnings are significantly different from zero at the 5% level: their t statistics are greater than 2 in absolute values and their p values are less than 0.05.

In Excel, if you click the residuals box it will also give you the predicted values and residuals for every observation. If you use Excel you must graph these residuals to judge how well the least squares assumptions hold. You can have more right hand side, X, variables but they must be contiguous in the spreadsheet, side by side. So for instance we could have estimated

$$ND_t = \alpha + \beta NE_t + \gamma NSP_t + u_t$$

by giving the X range as D2:E117.

22.3. EViews

22.3.1. Entering Data

Open the EViews program, different versions may differ slightly. Click on File, New, Workfile, accept the default annual data and enter the length of the time series 1871 2000 in the box. OK. You will now get a box telling you that you have a file with two variables C (which takes the value unity for each observation) and RESID which is the variable where estimates of the residuals will be stored.

Click on File, Import, Read Text-Lotus-Excel, then click on the Shiller file. It will open a box. Tell it, in the relevant box, that there are 5 variables. Note the other options, but use the defaults, note that B2 is the right place to start reading this data file. Click Read. You should now also have PPI, ND, NE and NSP R in your workfile. Double click on NSP and you will see the data and have various other options including graph.

Highlight NE and ND. Click on Quick, then Graph, OK line graph, and you will see the graphs of these two series. Close the graph. **Always graph your data.**

Use Save As command to save the Workfile under a new name and keep saving it when you add new data, transformations etc.

22.3.2. Estimating a Regression

Click on Quick, Estimate Equation and you will get a box. Enter ND C NE; set the sample as 1871 1986. OK and you will get a box with equation estimates. Notice that you have menu buttons both on the equation box and the main window. The estimates are the same as given above for Excel.

Dependent Variable: ND
Method: Least Squares
Date: 07/29/04 Time: 14:00
Sample: 1871 1986
Included observations: 116

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.168391	0.035869	4.694672	0.0000
NE	0.456537	0.007328	62.2974	1 0.0000

R-squared 0.971464	Mean dependent var 1.445948
Adjusted R-squared 0.971214	S.D. dependent var 1.868085
S.E. of regression 0.316949	Akaike info criterion 0.556937

Sum squared resid 11.45203	Schwarz criterion 0.604413
Log likelihood -30.30234	F-statistic 3880.968
Durbin-Watson stat 0.659872	Prob(F-statistic) 0.000000

Suppose the equation is

$$Y_t = \alpha + \beta X_t + u_t,\ t = 1, 2, ..., T.$$

The program gives you the dependent variable Y_t, in this case ND; the method of estimation, in this case least squares; the date and time; the sample you used, in this case 1871 1986; and the number of observations, 116. Then for each right hand side variable, it gives you the estimate of the coefficient, $\widehat{\alpha}$ and $\widehat{\beta}$; its standard error, its t statistic and p value (Prob).

Reading down the first column it gives: R-squared; Adjusted R-squared, usually called R bar squared; the Standard Error of the Regression, which measures the average size of the error in predicting Y_t; Sum of Squared Residuals $\sum_{t=1}^{T} \widehat{u}_t^2$; Ignore Log-likelihood; Durbin Watson Statistic In this regression, the DW of 0.662 shows that there is something badly wrong. The second column gives the mean and standard deviation of the dependent variable, two criteria which you can ignore for now and the F-statistic and p value for the hypothesis that all of the slope coefficients are zero.

In Eviews when you have the equation box on the screen, click View on the box toolbar and you will see a range of options. Actual Fitted Residuals allows you to graph the actual and fitted (predicted values) for the dependent variable and the residuals. **Always look at the graph of predicted values and residuals.** Under Residual you can test for normality, serial correlation, heteroskedasticity and under stability you can test for non-linearity RESET or structural change in the parameters at some point. Use Chow Break point if you know when the relationship shifted, or Cusum graphs if you do not. If the graphs go outside the confidence bands there is a problem. In each case the null hypothesis is that the model is well specified (does not have the problem) so small p values ($p<0.05$) lead you to reject the hypothesis that this is a healthy equation. Verbeek section 3.3 and 4.4 and 4.7 discusses many of these tests.

22.3.3. A different specification.

Close the equation box. Click on Quick, Generate, enter in the box the equation

$$LRD = \log(ND/PPI)$$

and click OK. You will see that a new variable, log of real dividends, has been added to the workfile.

Do the same to generate log of real earnings: LRE=log(NE/PPI). Graph LRD and LRE. Click on Quick, Estimate equation, and enter LRD C LRE LRD(-1) in the box. This estimates an equation of the form

$$Y_t = \alpha + \beta X_t + \gamma Y_{t-1} + u_t$$

where the dependent variable is influenced by its value in the previous period. The estimates are:

$$\begin{array}{lllll} LRD_t = - & 0.517 + & 0.248 LRE_t + & 0.657 LRD_{t-1} & R^2 = 0.94 \\ & (0.107) & (0.034) & (0.046) & s = 0.108 \end{array}$$

Although this has a R^2 of 0.94, it does not mean that it is worse than the previous equation, which had an R^2 of 0.97, because the two equations have different dependent variables. Above the dependent variable was nominal dividends, here it is log real dividends. The Durbin Watson statistic for this equation is 1.72 which is much better. This is a dynamic equation since it includes the lagged dependent variable. The long-run elasticity of dividends to earnings is $0.248/(1 - 0.657) = 0.72$. A 1% increase in earnings is associated with a 0.72% increase in dividends in the long-run.

23. Dynamics

With cross-section data a major issue tends to be getting the functional form correct; with time-series data a major issues tends to be getting the dependence over time, the dynamics, correct.

23.1. Autoregressions and distributed lags

We have already come across autoregressions (AR), regressions of a variable on lagged values of itself when discussing serial correlation in error terms. They can also used for variables just as above we ran a regression of dividends on a constant, earnings and the lagged value of dividends. A first order autoregression would take the form

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + u_t$$

the parameters can be estimated by least squares. The random walk with drift, our first example of a model in section 5.2, is the special case where $\alpha_1 = 1$. If

$-1 < \alpha_1 < 1$ the process is stable, it will converge back to a long-run equilibrium after shocks. The long run equilibrium can be got from assuming $y_t = y_{t-1} = y$ (as would be true in equilibrium with no shocks) so

$$\begin{aligned} y &= \alpha_0 + \alpha_1 y \\ y^* &= \alpha_0/(1-\alpha_1). \end{aligned}$$

Using the star to indicate the long-run equilibrium value. A random walk does not have a long-run equilibrium it can wander anywhere.

A second order (two lags) autoregression takes the form

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + u_t.$$

This is stable if $-1 < \alpha_1 + \alpha_2 < 1$, in which case its long run expected value is

$$y^* = \frac{\alpha_0}{1-\alpha_1-\alpha_2}$$

We may also get slow responses from the effects of the independent variables, these are called distributed lags (DL). A first order distributed lag takes the form

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + u_t$$

and we could have higher order versions.

We can put the first order AR1 and DL1 together to get an ARDL(1,1)

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + u_t$$

Again it can be estimated by least squares. There are often strong theoretical reasons for such forms.

In many cases adjustment towards equilibrium is slow. This can be dealt with by assuming a long-run equilibrium relationship, e.g. between consumption and income

$$C_t^* = \theta_0 + \theta_1 Y_t \tag{23.1}$$

and a partial adjustment model (PAM)

$$\Delta C_t = \lambda(C_t^* - C_{t-1}) + u_t. \tag{23.2}$$

$\Delta C_t = C_t - C_{t-1}$, is the change in consumption. People adjust their consumption to remove part of the difference between the equilibrium consumption and consumption in the previous period. The coefficient λ is an adjustment coefficient, it

measures the proportion of the deviation from equilibrium made up in a period and we would expect $0 < \lambda \leq 1$, with $\lambda = 1$ indicating instantaneous adjustment and $\lambda = 0$ no adjustment. We can write this

$$\Delta C_t = \lambda\theta_0 + \lambda\theta_1 Y_t - \lambda C_{t-1} + u_t$$

or

$$C_t = \lambda\theta_0 + \lambda\theta_1 Y_t + (1-\lambda)C_{t-1} + u_t$$

we would just run a regression of consumption on a constant, income and lagged consumption,

$$C_t = \alpha_0 + \beta Y_t + \alpha_1 C_{t-1} + u_t$$

We can recover the theoretical parameters λ, θ_0, θ_1 from the estimated parameters given by the computer α_0, α_1, β. So we estimate the speed of adjustment as $\widehat{\lambda} = (1 - \widehat{\alpha}_1)$; and the long run effect as $\widehat{\theta}_1 = \widehat{\beta}_1/\widehat{\lambda}$. This is a dynamic equation it includes lagged values of the dependent variable. Whether we estimate it using the first difference ΔC_t or level C_t as the dependent variable does not matter, we would get identical estimates of the intercept and the coefficient of income. The coefficient of lagged consumption in the levels equation will be exactly equal to the coefficient in the first difference equation plus one. Sums of squared residuals will be identical, though R^2 will not be, because the dependent variable is different. This is one reason R^2 is not a good measure of fit.

For more complex adjustment processes, we can keep the long-run relationship given by (23.1) and replace the partial adjutment model (23.2) by the error correction model (ECM), which assumes that people respond to both the change in the target and the lagged error

$$\begin{aligned}
\Delta C_t &= \lambda_1 \Delta C_t^* + \lambda_2 (C_{t-1}^* - C_{t-1}) + u_t \\
\Delta C_t &= \lambda_1\theta_1 \Delta Y_t + \lambda_2(\theta_0 + \theta_1 Y_{t-1} - C_{t-1}) + u_t \\
\Delta C_t &= a_0 + b_0 \Delta Y_t + b_1 Y_{t-1} + a_1 C_{t-1} + u_t.
\end{aligned}$$

and we can estimate the last version which gives us the estimated parameters, which are functions of the theoretical parameters, ($a_0 = \lambda_2\theta_0$, $b_0 = \lambda_1\theta_1$, $b_1 = \lambda_2\theta_1$, $a_1 = -\lambda_2$), so we can solve for the theoretical parameters from our estimates, e.g. the long run effect is, $\widehat{\theta}_1 = -\widehat{b}_1/\widehat{a}_1$.

We can also rearrange the ECM to give a (reparameterised) estimating equation of the ARDL(1,1) form

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + u_t.$$

Where $\alpha_1 = 1 + a_1$, $\beta_1 = b_0 + b_1$.We can find the equilibrium solution if the model is stable, $-1 < a_1 < 1$ by setting $y_t = y_{t-1} = y$; $x_t = x_{t-1} = x$ so that in long-run equilibrium

$$\begin{aligned} y &= \alpha_0 + \alpha_1 y + \beta_0 x + \beta_1 x \\ y &= \frac{\alpha_0}{1-\alpha_1} + \frac{\beta_0 + \beta_1}{1-\alpha_1} x \\ y^* &= \theta_0 + \theta_1 x \end{aligned}$$

These long-run estimates, θ_0 and θ_1 can be calculated from the short-run estimates of α_0, α_1, β_0, β_1 which the computer reports. Economic theory usually makes predictions about the long-run relations rather than the short-run relations. Notice our estimate of θ_1 will be identical whether we get it from the ECM or ARDL equation or from estimating a non-linear version.

23.2. Background example: ARCH

Asset prices tend to show volatility clustering, periods of high volatility followed by periods of low volatility. This is often captured by assuming that the variance of the asset price is positively serially correlated, so a high variance in one period makes it more likely that there will be a high variance in the next period. Suppose the logarithm of the asset price is a random walk

$$\begin{aligned} p_t &= p_{t-1} + \varepsilon_t \\ \Delta p_t &= \varepsilon_t \end{aligned}$$

Usually we assume that $E(\varepsilon_t) = 0$ and $E(\varepsilon_t^2) = \sigma^2$, it has constant variance, is homoskedastic. Here we shall assume that the variance changes through time so $E(\varepsilon_t^2) = \sigma_t^2$, it is heteroskedastic, and that the variance follows a first order autoregression:

$$\sigma_t^2 = \alpha + \rho \sigma_{t-1}^2 + v_t.$$

This is Auto-Regressive Conditional Heteroskedasticity, ARCH. If we can estimate this equation, we can use it to predict the variance in the future. This is straightforward, our best estimate of $\sigma_t^2 = \varepsilon_t^2$ and since $\varepsilon_t^2 = (\Delta p_t)^2$ we can just run a regression of

$$(\Delta p_t)^2 = \alpha + \rho (\Delta p_{t-1})^2 + v_t.$$

The unconditional variance of returns is $\alpha/(1-\rho)$ assuming the process is stable $-1 < \rho < 1$.

Above we could estimate the variance directly, usually we assume that the error from an estimated regression equation exhibits ARCH. Suppose that we estimate

$$y_t = \beta' x_t + \varepsilon_t$$

where $E(\varepsilon_t) = 0$ and $E(\varepsilon_t^2) = \sigma_t^2$. The GARCH(1,1) first order Generalised ARCH model is then

$$\sigma_t^2 = a_0 + a_1 \varepsilon_{t-i}^2 + b_2 \sigma_{t-i}^2$$

more lags could be added. Eviews can estimate GARCH models of various forms.

23.3. Final thought

Modern statistics and econometrics programs are very powerful, you can easily estimate almost anything you want. The difficult task is interpreting what the program output is telling you. You have a better chance of interpreting the output if you have: (a) graphed the data and know what it really measures; (b) investigated the historical or institutional context of the data; (c) thought about the economics or other theoretical context; (d) understood the statistical technique being used and (e) know the purpose of doing the statistical analysis. Try to let the data speak for themselves rather than beating a confession out of them.

24. Additional matrix results

24.1. The bivariate case in matrix algebra

Consider our earlier model

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

for $t = 1, 2, ..., T$, that is a set of T equations. The errors, are $u_t = Y_t - \beta_1 - \beta_2 X_t$. Least squares chooses β_i to minimise $\sum u_t^2$.

In matrix form the model is

$$y = X\beta + u$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ .. \\ Y_T \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ .. & .. \\ 1 & X_T \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ .. \\ u_T \end{bmatrix}$$

and we will use $X'X$ where X' is the $2 \times T$ transpose of X, so $X'X$ is 2×2

$$(X'X) = \begin{bmatrix} T & \sum x_t \\ \sum x_t & \sum x_t^2 \end{bmatrix}$$

$X'X$ is a symmetric matrix so $(X'X)' = (X'X)$. We will also use the inverse of $X'X$

$$(X'X)^{-1} = \frac{1}{T\sum X_t^2 - (\sum X_t)^2} \begin{bmatrix} \sum X_t^2 & -\sum X_t \\ -\sum X_t & T \end{bmatrix}$$

For this to exist, the variance of $X_t = T\sum X_t^2 - (\sum X_t)^2$ must not equal zero. Use class exercise (1) week 2 to show this is the variance of X_t. It is also the determinant of $X'X$. For the inverse to exist the matrix must be non-singular, with a non-zero determinant.

The sum of squared residuals is a scalar (a 1×1 matrix), $\widehat{u}'\widehat{u}$ the product of a $(1 \times T)$ vector $\widehat{u}'$ and a $(T \times 1)$ vector $\widehat{u}$. If A is a $n \times m$ matrix, and B is an $m \times k$ matrix, the product AB is an $n \times k$ matrix. The transpose of the product $(AB)'$ is $B'A'$ the product of a $k \times m$ matrix B' with a $m \times n$ matrix A'. $A'B'$ is not conformable, you cannot multiply a $m \times n$ matrix by a $k \times m$ matrix. Below we set $y'X\widehat{\beta} = \widehat{\beta}'X'y$. In general a matrix is not equal to its transpose, but these are both scalars, so are equal to their transposes. Check this. We expand the sum of squared residuals as we did above:

$$\begin{aligned} \widehat{u}'\widehat{u} &= (y - X\widehat{\beta})'(y - X\widehat{\beta}) \\ &= y'y + \widehat{\beta}'X'X\widehat{\beta} - 2\widehat{\beta}'X'y \\ \sum_{t=1}^{T} \widehat{u}_t^2 &= \sum Y_t^2 + (\widehat{\beta}_1^2 T + \widehat{\beta}_2^2 \sum X_t^2 + 2\widehat{\beta}_1\widehat{\beta}_2 \sum X_t) \\ &\quad -2(\widehat{\beta}_1 \sum X_t + \widehat{\beta}_2 \sum X_t Y_t) \end{aligned}$$

The scalar $\widehat{\beta}'X'X\widehat{\beta}$ is a quadratic form, i.e. of the form $x'Ax$ and the $\widehat{\beta}_i^2$ appear in it. Quadratic forms play a big role in econometrics. A matrix, A, is positive definite if for any a, $a'Aa > 0$. Matrices with the structure $X'X$ are always positive definite, since they can be written as a sum of squares. Define $z = Xa$, then $z'z = a'X'Xa$ is the sum of the squared elements of z.

24.2. Differentiation with vectors and matrices

Consider the linear relation:

$$P = \underset{1\text{x}n}{x'} \; \underset{n\text{x}1}{a}$$

Then the differential of P with respect to x or x' is defined as :

$$\frac{dP}{dx} = a \text{ and } \frac{dP}{dx'} = a'$$

In the case n= 2, we can write:

$$\begin{aligned} P &= [x_1, x_2] \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \\ &= x_1 a_1 + x_2 a_2 \end{aligned}$$

Then

$$\frac{dP}{dx_1} = a_1 \text{ and } \frac{dP}{dx_2} = a_2$$

So

$$\frac{dP}{dx} = \begin{bmatrix} \frac{dP}{dx_1} \\ \frac{dP}{dx_2} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = a$$

and

$$\frac{dP}{dx'} = \left[\frac{dP}{dx_1}, \frac{dP}{dx_2}\right] = [a_1, a_2] = a'$$

Consider the quadratic form:

$$Q = \underset{1\text{x}n}{x'} \; \underset{(n\text{x}n)}{A} \; \underset{(n\text{x}1)}{x}$$

Then the derivative of Q with respect to x or x' is defined as :

$$\frac{dQ}{dx} = 2Ax \text{ and } \frac{dQ}{dx'} = 2x'A$$

In the case n= 2, we can write:

$$Q = [x_1, x_2] \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

where for simplicity A is assumed to be symmetric. Expanding this gives:

$$\begin{aligned} Q &= [x_1, x_2] \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{12}x_1 + a_{22}x_2 \end{bmatrix} \\ &= a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 \end{aligned}$$

So:

$$\frac{dQ}{dx_1} = 2a_{11}x_1 + 2a_{12}x_2 \text{ and } \frac{dQ}{dx_2} = 2a_{12}x_1 + 2a_{22}x_2$$

Then

$$\begin{aligned}\frac{dQ}{dx} &= \begin{bmatrix} \frac{dQ}{dx_1} \\ \frac{dQ}{dx_2} \end{bmatrix} = \begin{bmatrix} 2a_{11}x_1 + 2a_{12}x_2 \\ 2a_{12}x_1 + 2a_{22}x_2 \end{bmatrix} = 2 \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= \underset{(2\times2)}{2\,A}\,\underset{(2\times1)}{x}\end{aligned}$$

and

$$\begin{aligned}\frac{dQ}{dx'} &= \left[\frac{dQ}{dx_1}, \frac{dQ}{dx_2}\right] = [2a_{11}x_1 + 2a_{12}x_2, 2a_{12}x_1 + 2a_{22}x_2] \\ &= 2\,[x_1, x_2] \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \\ &= \underset{1\times2}{2\,x'}\,\underset{2\times2}{A}\end{aligned}$$

24.3. Gauss-Markov Theorem

We can derive many results without assuming normality. We have often claimed that the mean or regression coefficients are the minimum variance estimators in the class of linear unbiased estimators. We now prove it for the general case. Verbeek section 2.3 covers this. It applies to the mean when β contains a single element and X is just a column of ones.

Consider another linear estimator of the form $\widetilde{\beta} = Cy$ and we assume that X and C are fixed (non-stochastic) matrices

$$\begin{aligned}\widetilde{\beta} &= Cy = C(X\beta + u) = CX\beta + Cu \\ E(\widetilde{\beta}) &= CX\beta + CE(u)\end{aligned}$$

so $\widetilde{\beta}$ will be unbiased as long as $CX = I$. Write $\widetilde{\beta} = Cy = ((X'X)^{-1}X' + W)y$, that is $W = C - (X'X)^{-1}X$. Then $CX = I$ implies $((X'X)^{-1}X' + W)X = I$ or $(X'X)^{-1}X'X + WX = I$ or $I + WX = I$. This can only be true if $WX = 0$. This also implies that $X'W' = 0$. Assume that this is the case to ensure that $\widetilde{\beta}$ is unbiased. Let us look at the variance covariance matrix of $\widetilde{\beta}$.This is

$$E(\widetilde{\beta} - (\widetilde{\beta}))(\widetilde{\beta} - (\widetilde{\beta}))' = E(\widetilde{\beta} - \beta)(\widetilde{\beta} - \beta)'$$

since $\widetilde{\beta}$ is unbiased by assumption. From above

$$\begin{aligned}\widetilde{\beta} &= \beta + Cu = \beta + ((X'X)^{-1}X' + W)u \\ \widetilde{\beta} - \beta &= (X'X)^{-1}X'u + Wu\end{aligned}$$

$$E(\widetilde{\beta} - \beta)(\widetilde{\beta} - \beta)' = E((X'X)^{-1}X'u + Wu)((X'X)^{-1}X'u + Wu)'$$

When we multiply out the brackets we have four terms, the first is

$$E((X'X)^{-1}X'uu'X(X'X)^{-1}) = \sigma^2(X'X)^{-1}$$

the variance-covariance matrix of the least squares estimator. The second is

$$E(Wu\ u'W') = \sigma^2 WW'$$

the third is

$$E((X'X)^{-1}X'uu'W') = \sigma^2(X'X)^{-1}X'W' = 0$$

since $X'W' = 0$. Similarly

$$E(Wuu'X(X'X)^{-1}) = \sigma^2 WX(X'X)^{-1} = 0$$

since $XW = 0$. So the Variance of any other linear unbiased estimator is

$$\begin{aligned}V(\widetilde{\beta}) &= E(\widetilde{\beta} - (\widetilde{\beta}))(\widetilde{\beta} - (\widetilde{\beta}))' = \sigma^2[(X'X)^{-1} + WW'] \\ &= V(\widehat{\beta}) + \sigma^2 WW'\end{aligned}$$

since WW' is a positive definite matrix for $W \neq 0$, we have shown that in the class of linear unbiased estimators the OLS estimator has the smallest variance. An $n \times n$ matrix, A is positive definite if the scalar quadratic form $b'Ab > 0$ for any $T \times 1$ vector b. In this case we requires $b'WW'b > 0$. Define $z = W'b$ an $n \times 1$ vector then $b'WW'b = z'z = \sum z_t^2$ a sum of squares which must be positive.

25. Index