Applied Statistics and Econometrics, Notes

Ron Smith: r.smith@bbk.ac.uk
Pedro Gomes: p.gomes@bbk.ac.uk

2020-21

Department of Economics Mathematics and Statistics
Birkbeck, University of London

Economics: Graduate Diplomas and BScs

Material labelled "background" in the notes will not be used directly in exam questions.

Lectures

The section in the notes corresponding to a lecture is denoted by #

Autumn term: Statistics: Pedro Gomes
Week 1: Introduction, #1
Week 2: Descriptive Statistics, #2
Week 3: Index Numbers #3 #4 #5
Week 4: Probability, #6
Week 5: Discrete Random Variables, #7
Week 6: Reading Week
Week 7: Continuous Random Variables: Normal and related distributions #8
Week 8: Estimation of the Expected Value, #9
Week 9: Confidence Intervals and Hypothesis Testing, #10
Week 10: Overview
Week 11: Project

Spring term: Econometrics, Ron Smith

Week 1: Introduction,
Week 2: Bivariate Regression, #11
Week 3: Example: Life Expectancy, #12
Week 4: Multiple linear regression and properties of least squares, #13
Week 5: Matrix form of the Linear Regression Model, #14
Week 6: ReadingWeek: Computer Classes
Week 7: Matrix form of the properties of Least Squares #15
Week 8:Testing #16
Week 9 Regression in Practice, #17
Week 10 Econometric Relationships, #18
Week 11 Time-series and Dynamics , #19

Contents

1. Introduction

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, such as. 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. British annual net household income in 2014 had a mean of £29,198 and a median of £23,574. 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 methods of measuring 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.

1.1. 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, they kept them 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.

1.2. 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, y_t, takes today, at 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 random 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 represented by ε_t. 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 the random walk model 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 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.

1.3. 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. So y_t could indicate $Y_t - \overline{Y}$, where $\overline{Y}$, said Y bar, is the mean. But it is also common to use lower case letters to denote logarithms, so y_t 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.

We use the Greek alphabet a lot, so is given below.

1.4. The Greek alphabet.

A α alpha; α often used for intercept in regression and a measure of performance in finance.

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.

2. Descriptive statistics

Data tend to come in three main forms:

-time-series, such as. observations on annual inflation in the UK over a number of years 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 $T = 37$.

-cross-section, e.g. observations on annual inflation in different countries in a particular year, denoted by y_i, $i = 1, 2, ..., N$. Where, if they were arranged alphabetically, $i = 1$ might correspond to Albania and $i = N$ to Zambia. Whereas time-series data has a natural ordering, 1996 comes after 1995, cross-section does not.

-panels (also called longitudinal data) e.g. observations on inflation in a number of countries, $i = 1, 2, ..., N$, in a number of years, $t = 1, 2, ..., T$, denoted y_{it},

The number of observations in a sample will be denoted T for time-series and N for cross-sections.

Graphs are generally the best way to describe data. There are three types of graph economists commonly use, each are illustrated in section 4.

(1) 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, crises or pandemics).

(2) We can plot a histogram, which gives the number (or proportion) of observations which fall in a particular range.

(3) 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.

2.1. 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; or 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

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} (a + bx_i) = Na + b\sum_{i=1}^{N} x_i. \tag{2.1}$$

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

2.1.1. 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.

2.1.2. 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 we use the standard deviation because it is in the same units as the original variable. 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 $\widehat{\sigma}^2$, said sigma hat squared. This formula gives a set of instructions. It says take each of the observations and subtract the mean, $(x_i - \overline{x})$; square these deviations from the mean $(x_i - \overline{x})^2$; add the squared deviations together $\sum_{i=1}^{N}(x_i - \overline{x})^2$ and divide the sum of squared deviations 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, by coincidence, both the Mean and the Variance happen to be 5. The standard deviation, $SD(x) = \widehat{\sigma}$ is the square root of the variance: 2.24 in this case. Notice that the sum of the deviations from the mean is always zero:

$$\sum_{i=1}^{N}(x_i - \overline{x}) = \sum_{i=1}^{N} x_i - N\overline{x} = \sum_{i=1}^{N} x_i - N\frac{\sum_{i=1}^{N} x_i}{N} = 0. \tag{2.2}$$

Notice we use (2.1) in deriving the second term.

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.

If x has standard deviation σ, then for constants a and b, the new variable $y = a + bx$ has standard deviation $b\sigma$ and thus variance $b^2\sigma^2$.

2.1.3. 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. To prove this note that if $y = x$, $Cov(x, y) = Var(x)$, and $SD(x)SD(y) = Var(x)$, so $r = 1$.

Correlation measures linear relationships, there may be a strong non-linear relationship and zero correlation. Correlation does not imply causation. Two variables may be correlated because they are both caused by a third variable.

2.1.4. Standardised data

To remove the effect of units of measurement, data are often standardised by subtracting the mean and dividing by the standard deviation (the square root of

the sample variance),

$$z_i = \frac{x_i - \overline{x}}{SD(x)}.$$

This new variable, z_i, has mean zero, from (2.2), and variance (and standard deviation) of one. If x has standard deviation $SD(x) = \sigma$, $z = \sigma^{-1}x - \sigma^{-1}\overline{x}$ has standard deviation $\sigma^{-1}\sigma = 1$. This follows from our rule above that if x has standard deviation σ, then $y = a + bx$ has standard deviation $b\sigma$. In this case $a = -\sigma^{-1}\overline{x}$ and $b = \sigma^{-1}$. The correlation coefficient is the covariance between the standardised measures of x and y.

$$r = \frac{\sum_{i=1}^{N}(x_i - \overline{x})\,(y_i - \overline{y})\,/N}{SD(x)SD(y)} = \sum \left(\frac{(x_i - \overline{x})}{SD(x)}\right)\left(\frac{(y_i - \overline{y})}{SD(y)}\right)/N$$

2.1.5. Moments

A distribution is often described by:

- its moments, m, which are $\sum_{i=1}^{N} {x_i}^k/N$, $k = 1, 2, ...$ The mean $\overline{x} = \sum x_i/N$ is the first moment, $k = 1$.
- its centred moments, cm, $\sum_{i=1}^{N}(x_i - \overline{x})^k/N$. The variance, $\sigma^2 = \sum_{i=1}^{N}(x_i - \overline{x})^2/N$ is the second centred moment, $k = 2$. The first centred moment, $\sum_{i=1}^{N}(x_i - \overline{x})/N = 0$.
- its standardised moments, sm, $\sum z_i^k/N$, where $z_i = (x_i - \overline{x})/\sigma$. The third standardised moment, $k = 3$, is a measure of whether the distribution is symmetrical or skewed. The fourth standardised moment, $k = 4$, is a measure of kurtosis (how fat the tails of the distribution are). A distribution we will use a lot is called the normal or Gaussian distribution. It is symmetrical and bell shaped. Its coefficient of skewness $\sum z_i^3/N$ is zero, and the coefficient of kurtosis $\sum z_i^4/N$ is 3.

Moments

k	m	cm	sm
1	$\overline{x}$	0	0
2	$\sum_{i=1}^{N} {x_i}^2/N$	σ^2	1
3	$\sum_{i=1}^{N} {x_i}^3/N$	$\sum_{i=1}^{N}(x_i - \overline{x})^3/N$	$\sum z_i^3/N$
4	$\sum_{i=1}^{N} {x_i}^4/N$	$\sum_{i=1}^{N}(x_i - \overline{x})^4/N$	$\sum z_i^4/N$

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.

2.2. 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.

3. Economic and Financial Data

3.1. Tables and Calculations

Data will typically come in a Table, either electronic or hard-copy. **When constructing your own tables, make sure the table has: 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 are the maximum and minimum values? When comparing series or graphing them together make sure they are in comparable units.

Be careful about units. 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, as in done in scientific notation. $1 = 10^0$; $10 = 10^1$; $100 = 10^2$; $1,000,000 = 10^6$. The power gives you the number of zeros after the one. Millions are 10^6, billions 10^9, trillions 10^{12}. UK GDP in 2009 was about £1.4 trillion (10^{12}) and its population was about 61 million (10^6) so its per-capita GDP was about £$(1.4 \times 10^{12}/61 \times 10^6) = 0.0229 \times 10^6 =$ £22,900

Be careful about coverage. In the UK statistics are sometimes for England and Wales, sometimes Great Britain (England, Wales & Scotland) and sometimes UK (England, Wales, Scotland and Northern Ireland). Team GB at the Olympics should have been Team UK.

3.2. 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 (not trended). An example is the Average Propensity to Consume (the ratio of Consumption to Income) or the Savings Ratio. In finance, ratios like the Price-Earnings Ratio or the Dividend Yield are used. Notice these ratios can be compared across countries, because the units of currency in the numerator and denominator cancel. These can be expressed either as proportions or multiplied by 100 to give percent.

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.

In economics and finance it is common to work with the logarithms of the variables, for reasons discussed in detail in section 12.1.

3.3. National Accounts

The output of an economy is usually measured by Gross Domestic Product, GDP. This measure is part of a system of National Accounts, which start from the identity that

- output = expenditure = 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 Product 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.

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.

3.3.1. A good measure?

"Material wellbeing, and measures of it—GDP, personal income, and consumption—have recently received a bad press. Spending more, we are often told, does not bring us better lives, and religious authorities regularly warn against materialism. Even among those of us who endorse economic growth, there are many critics of GDP as it is currently defined and measured. GDP excludes important activities, such as services by homemakers; it takes no account of leisure; and it often does a poor job of measuring those things that are included. It also includes things that arguably should be excluded, like the cost of cleaning up pollution or building prisons or commuting. These "defensive" expenditures are not good in and of themselves but are regrettably necessary to enable things that are good. If crime goes up, and we spend more on prisons, GDP will be higher. If we neglect climate change, and spend more and more on cleaning up and repairing after storms, GDP will go up, not down; we count the repairs but ignore the destruction."

Angus Deaton

3.4. 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.

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. The unemployment rate is the number unemployed divided by the number in the labour force (em-

ployed plus unemployed). Youth unemployment rates are high partly because the denominator, the labour force, is reduced by the number of people studying.

3.5. 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

$V_0 = 100$	$V_1 = 110$	$V_2 = 121$	$V_3 = 133.1$

.

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 $r_1 = (V_1 - V_0)/V_0 = V_1/V_0 - 1 = 0.1$, 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

$$V_t = (1 + r)^t V_0. \tag{3.1}$$

Notice how interest compounds, you get interest paid on your interest. Interest rates are often expressed at annual rates, per annum, p.a., even when they are for shorter or longer periods. If interest at 10% p.a. 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

$$V_t = e^{rt} V_0. \tag{3.2}$$

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 (3.1) 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, 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.

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 crunch starting in August 2007 policy rates and LIBOR diverged: Banks required a risk premium to lend to other banks and LIBOR was described as the rate at which banks were unwilling to lend to each other. Evidence that LIBOR was manipulated in the interests of the banks led to reform of the process of setting it. From the crisis till summer 2015 interest rates in the US, UK and eurozone have been close to zero and not changed and other sorts of monetary policy called quantitative easing used.

3.6. Exchange Rates

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.

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, $I_0 = 100$.

4. Example: Were the nineties and noughties NICE?

4.1. Introduction

To give an example of how you use descriptive statistics and as an example of what a project might look like we will look at UK economic history. Mervyn King, the previous 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 and later behaviour to see how nice this period was.

4.2. Data

The original series, from the Office of National Statistics, are for 1955Q1-2016Q2, Q_t = Gross Domestic Product, chained volume measure, constant 2013 prices, seasonally adjusted (ABMI) E_t = Gross Domestic Product at current prices: Seasonally adjusted (YBHA). U_t is the ILO unemployment rate: UK: All: Aged 16-64: %, seasonally adjusted: (LF2Q), which is only available from 1971. 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), inf_t, are measured over the same quarter in the previous year as $g_t = 100 * (\ln(Q_t) - \ln(Q_{t-4}))$ and $\text{inf}_t = 100 * (\ln(P_t) - \ln(P_{t-4}))$, which are very close to the percentage changes. Such annual differences smooth the series and would remove seasonality if they were not already seasonally adjusted. Whereas the data for ouput and prices starts in 1955Q1, the data for growth and inflation only starts in 1956Q1, because of taking the 4 quarter change. The most recent figures are very likely to get revised and many have questioned the accuracy of the figures.

4.3. Line graphs

The broad pattern of UK economic events can be seen from the line graphs for growth, inflation, and unemployment. 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, ERM. With the ejection of sterling from the ERM in September 1992, inflation targets were adopted. In 1997, the newly elected Labour government gave the Bank of England independent responsibility for monetary policy, setting interest rates to target inflation. This history is reflected in the graphs for growth and inflation below.

The "stop-go" pattern of growth in 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.

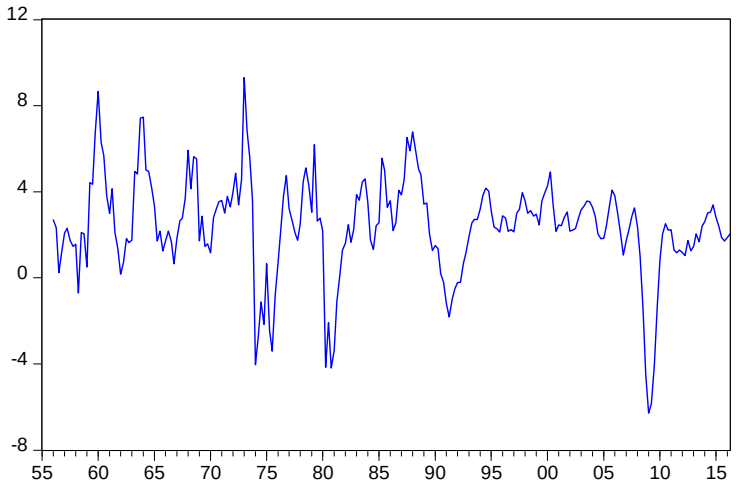
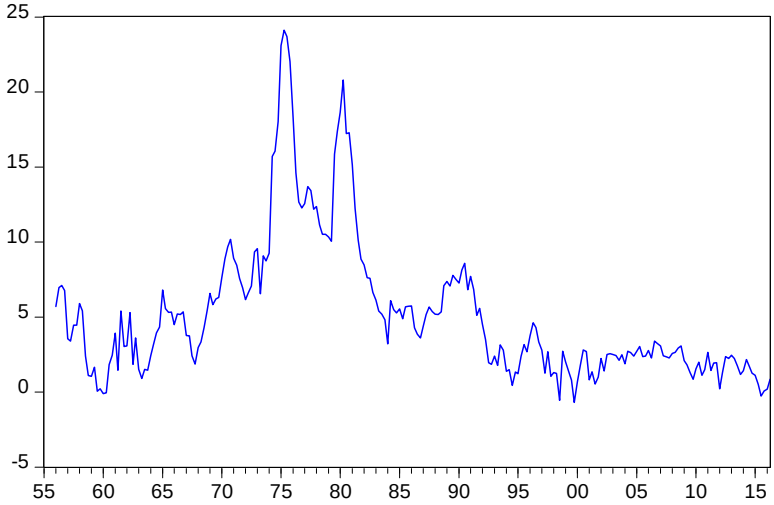
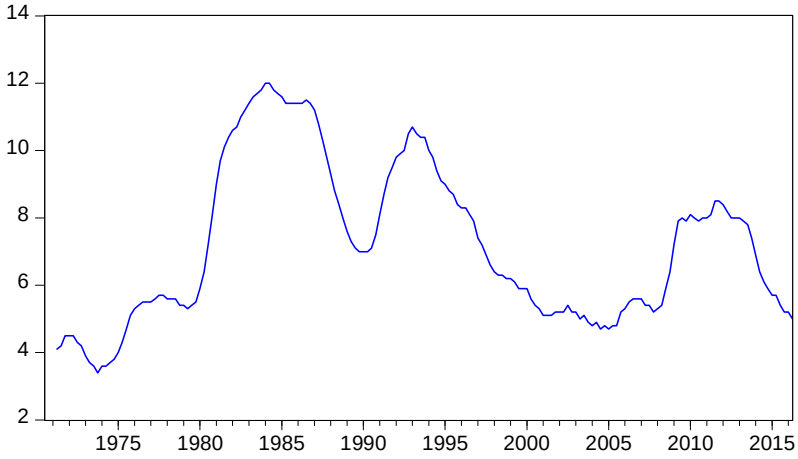


Figure 4.1: Growth 1956-2016



Inflation 1956-2016



Unemployment 1972-2016

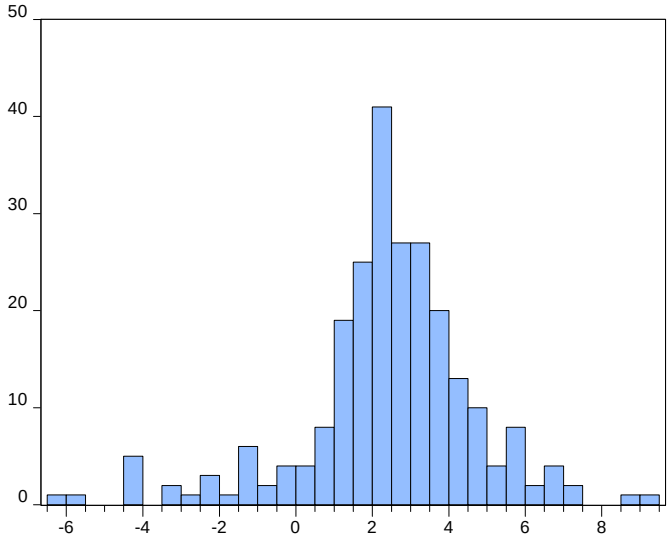
Growth recovered in the later 1980s with the boom under Nigel Lawson, the Conservative Chancellor, then sank into recession again in the early 1990s, possibly worsened by the fixed exchange rate required by membership of the ERM. The UK left the ERM in September 1992 and there followed a long period of relative stability, before the effects of the 2007 financial crisis began to impact on the economy. Output fell by about 6% in the year up to 2009Q1, the lowest observed in this sample and while the economy bounced back, subsequent growth was slow.

Inflation is the percentage change in prices. Negative inflation, prices faling, is called deflation. Inflation was 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. The unemployment rate, which is only available on a consistent basis since 1971, shows the effects of the recessions of the mid 70s, early 80s, early 90s and late 00s. The combination of high unemployment and high inflation is sometimes called stagflation.

4.4. Frequency distributions.

The histograms describes the distribution of growth are shown below. The average growth rate over the 242 observations on overlapping four quarter changes is

2.40 (standard deviation 2.29), the median is slightly higher at 2.46. An average growth rate of just over 2%, has seemed to be a long-term feature of the British economy, the mode is between 2 and 2.5%. The distribution is not normal, with a slight negative skew (-0.75) and excess kurtosis (5.28) 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 $\chi^2(2)$, and tests whether the skewness and kurtosis are significantly different from these values has a value of 75 and a p value of zero. The maximum value is 9.3 and the minimum value is -6.3. One would be very unlikely to observe such values if the distribution was normal.

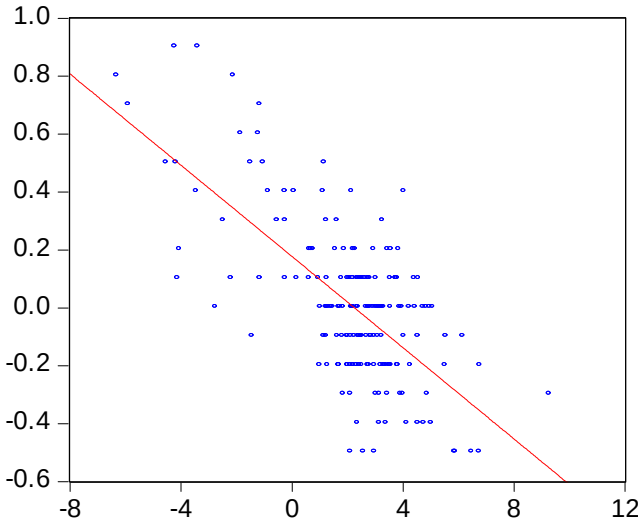


Histogram for growth

4.5. The relationship between growth and the change in unemployment

Growth is a major factor determining the change in unemployment: if growth is high unemployment falls, if growth is low, unemployment rises. The relationship between the change in unemployment and growth is known as Okun's Law, after the US economist Arthur Okun, who also introduced the misery index defined as the sum of the unemployment rate and the inflation rate. The scatter diagram plots the change in the unemployment rate over a quarter, on the Y axis, against the growth rate over the year, on the X axis, for the period 1971-2015. The relationship is clearly negative, the correlation is -0.67. One can fit a line to the

data which takes the form: $D1U_t = 0.18 - 0.079G_t + \varepsilon_t$, with a standard error of regression of 0.20. Okun estimated the relationship the other way round with growth a function of the change in unemployment. Unemployment increased less in the recent recession than one would expect from the historical pattern, though there is dispute about the reason for this. Notice the pattern of horizontal lines is because of rounding in the change in unemployment.



Scatter growth and change in unemployment

4.6. Conclusion

Compared to previous (and subsequent) history, the period 1993-2007 was nice. Inflation became low and stable rather than high and volatile, as it was before. Growth was slightly higher and much less volatile. Although there was an economic cycle 1997-2007, it was less pronounced than in earlier years, hence Gordon Brown's claim to have abolished boom and bust. Whether this was good 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 and Great Recession. 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; 9/11 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.

5. Index Numbers

5.1. Introduction

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

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

where P_t is a price index. Distinguish between the price level, P_t and the rate of inflation, π_t. 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%. The approximation works because $\ln(1 + \pi_t) = \pi_t$ if π_t is small.

Be careful with percentages. 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.

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.

5.2. Example: deflation after WWI

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. From these we can calculate: the growth rate of per capita income, $Growth_t = 100(Y_t/Y_{t-1} - 1)$; the inflation rate, $INF_t = 100(RPI_t/RPI_{t-1} - 1)$; and and the real interest rate, $RIR_t = R_t - INF_t$, for 1919-1922.

$Year$	U	RPI	R	Y	$\$/£$	$Growth$	INF	RIR
1918	0.8	42		54				
1919	2.1	56	3.5	48	4.42	−11	33.3	−29.8
1920	2.0	52	6.2	47	3.66	−2	−7.1	13.3
1921	12.9	47	4.6	42	3.85	−11	−9.6	14.2
1922	14.3	38	2.6	44	4.43	5	−19.1	21.7

In 1919 per capita income fell by over 10%, and continued falling till 1921. This fall produced rising unemployment. Inflation was very high during 1919, but then fell giving negative inflation, deflation. Between 1919 and 1922 prices fell by almost a third.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%. When you lend the money you do not know what the rate of inflation will be. In many cases the expected rate of inflation can be approximated by the current rate of inflation, which you know, so the real interest rate is often measured as the nominal interest rate minus the current rate of inflation. In 1919 the real interest rate was negative because inflation was higher than the nominal interest rate, subsequently with quite high nominal rates and deflation (negative inflation) real interest rates became very high.

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.

5.3. Prices and Quantities

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	Q	E
2003	35	2	70
2004	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. Knowing price and quantity we can calculate expenditure, $E_t = PQ$. Knowing expenditure and quantity we can calculate price as $P_t = E_t/Q_t$. Knowing expenditure and price, we can calculate 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

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

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.

5.4. Real and nominal variables

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 $Q_t = E_t/P_t$, 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: $P_t = E_t/Q_t$.

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

5.5. 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. The index always equals 1 (or 100) in its base year, $P_0^1 = 1$. This is a price index.

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

$$P_t^2 = \frac{\sum_{i=1}^{N} p_{it} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}}. \tag{5.1}$$

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 10 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.

5.5.1. 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=($10\times2000+20\times1000$); 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.

5.5.2. CPI and RPI

There are a lot of different price indexes. Currently the Bank of England has a target for annual CPI inflation of 2% and the governor has to write a letter to the Chancellor of the Exchequer if inflation falls below 1% or goes above 3%. Before 2004 the target was for inflation in the Retail Price Index excluding mortgage interest payments (which go up when interest rates are raised) RPIX of 2.5%. The CPI is. the Harmonised Index of Consumer Prices HICP, the standard European index. 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 measures prices in the whole economy.

There are three main differences between the CPI and RPI: CPI excludes owner occupied housing, RPI includes it; they use different formulae and their target populations are slightly different, the RPI excludes the very rich and very poor. The formula effect was thought to contribute about 0.5% to the difference (hence the replacement of a 2.5% RPI target with a 2% CPI target) though this has increased to almost 1% recently. Consider the Laspeyres price index, P_t^2 above (5.1)

$$P_t = \frac{\sum_{i=1}^N p_{it}q_{i0}}{\sum_{i=1}^N p_{i0}q_{i0}} = \sum_{i=1}^N \frac{p_{it}q_{i0}}{\sum_{i=1}^N p_{i0}q_{i0}} = \sum_{i=1}^N \left(\frac{p_{it}}{p_{i0}}\right)\left(\frac{p_{i0}q_{i0}}{\sum_{i=1}^N p_{i0}q_{i0}}\right) = \sum_{i=1}^N \left(\frac{p_{it}}{p_{i0}}\right) w_{i0}.$$

The term p_{it}/p_{i0} is called a price relative and w_{i0} is the share of good i in total expenditure in year zero. So the price index is a weighted average of price relatives. Here p_{it} would be a category like clothing, w_{i0} the share of expenditure on clothing. But one needs to average the prices of all the different items of clothing, (jeans, shirts, ..) where you do not have expenditure weights, to get what are called elementary aggregates, EAs. There are 3 widely used methods of averaging named after the people who developed them:

(a) arithmetic mean of price relatives (Carli) used for 35% of EAs in the RPI:

$$I^C = \sum_{j=1}^n \left(\frac{p_{jt}}{p_{j0}}\right)/n$$

(b) ratio of arithmetic means of prices (Dutot) used for 55% of EAs in the RPI

and 30% of EAs in the CPI:

$$I^D = \frac{\sum p_{jt}/n}{\sum p_{j0}/n}$$

(c) the geometric mean of price relatives (Jevons) which is identical to the ratio of geometric mean prices, used for 70% of the EAs in the CPI:

$$I^J = \sqrt[n]{\prod_{i=1}^{n}\left(\frac{p_{jt}}{p_{j0}}\right)} = \frac{\sqrt[n]{\prod p_{jt}}}{\sqrt[n]{\prod p_{0t}}}$$

Methods (b) and (c) tend to give quite similar results, method (a) can give very different ones particularly when the variance of prices is large. Method (a) is prohibited by the EU regulations that govern construction of the CPI and most of the formula difference between RPI and CPI comes from the use of method (a) in the RPI. Method (c) tends to deal with substitution, of the sort seen in the previous example rather better.

The government is trying to shift indexing to the CPI from the RPI, though Index linked government debt is linked to the RPI, which raises contractual issues. The ONS is introducing a variant of the RPI which uses the Jevons rather than Carli procedure to construct elementary aggregates.

5.5.3. 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.

5.5.4. 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, like ASE students, 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:

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

6. Probability

6.1. 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[1] and it is a relatively recent branch of mathematics, nobody seems to have regarded probabilities as things that could be calculated before about 1650 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

[1] Daniel Kahneman, *Thinking, fast & slow*, Penguin 2011, discusses these difficulties and David Hand, *The improbability principle: why coincidences, miracles and rare events happen every day,* Farrar Straus & Giroux 2015, contains lots of examples.

usually slightly better than the bookmaker's odds suggest and 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 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."[2] A will win 5/12 times, draw 2/12 times, so survives 7/12 times. Similarly for B. The probability of A or B surviving is the sum of the probability of each surviving minus the probability of both surviving: $7/12 + 7/12 - 2/12 = 1$. 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.

6.2. 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, sometimes 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 or conjunction); this is often known as the joint probability. If the events are mutually exclusive they cannot happen together so $P(A \cap B) = 0$. For isntance, one cannot have A and not A (getting a six and not getting a six) happening together so $P(A \cap \overline{A}) = 0$.

Denote the probability of event A **OR** event B happening as $P(A \cup B)$ (the union or disjunction). 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

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

[2] Quoted by Ian Hacking, *The emergence of probability*, p144, CUP 1975.

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

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{6.1}$$

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:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}. \tag{6.2}$$

Multiply both sides of (6.1) by $P(B)$ and both sides of (6.2) 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.

Statistics tends to divide into the approach this course emphasises, called classical or frequentist, and Bayesian statistics. A very readable account of the development of Bayesian statistics is Sharon Bertsch McGrayne, *The theory that would not die: how Bayes rule cracked the enigma code, hunted down Russian submarines & emerged triumphant from two centuries of controversy*, Yale University Press 2011.

As an example of its application, 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. This account comes from *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, though some of the details are disputed by McGrayne.

6.2.1. 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. For covid-19 the accuracy of antigen tests, for currently having it, and antibody tests, for having had it, was a major issue during the pandemic.

Call D having the disease, N not having the disease, with probabilities $P(D)$ and $P(N)$ where $P(D)+P(N) = 1$. 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) with $P(TP) + P(TN) = 1$.

The medical literature on testing reports the sensitivity and specificity of the test. Sensitivity is probability of a positive test given that you have the disease $P(TP \mid D)$. Specificity is the probability of a negative test given that you do not have the disease $P(TN \mid N)$. Both are about 0.9 for a mamogram for breast cancer, and the higher they are the better. However, they are not what you want to know. For diagnosis you want to know the positive and negative predictive values, the probability of having the disease given that you test positive, $P(D \mid TP)$ and the probability of not having the disease given you test negative, $P(N \mid TN)$.

However, specificity and sensitivity are used because they are much easier to measure. You can start with a sample that you are sure have the disease and see what proportion test positive and similarly with a sample that you are sure do not have the disease. You can relate specificity and sensitivity to predictive value

if you know the prevalence of the disease, here the proportion of the population with it, $P(D)$. But with covid this was difficult to estimate since many who had it did not show symptoms.

False positives are when you test positive and do not have the disease, with probability $P(TP \mid N) = 1 - P(TN \mid N)$. False negatives are testing negative when you do have the disease, with probability $P(TN \mid D) = 1 - P(TP \mid D)$. These two sorts of errors are given different name in different contexts.

For instance, the PSA test detects prostate specific antigen. The NHS advice says:

"About 75 out of every 100 men who have an abnormal PSA test result do not have prostate cancer. This is called a false positive result. About 15 out of every 100 men who have a normal PSA test result do have prostate cancer. This is called a false negative result."

Consider a much more accurate hypothetical test.

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 do 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	98,020
	1,000	99,000	100,000

We could also calculate the conditional probability directly using Bayes Theorem and noting that $P(TP) = P(TP \mid D)P(D) + P(TP \mid N)P(N)$,

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

Comment

Diagnostic screening is usually confined to groups where $P(D)$ is high to avoid the problem of false positives. The decision to establish a screening program depends on a judgement of the balance between (a) the benefits of detecting the disease, for instance if early treatment saves lives, (b) the costs of false positives such as inappropriate treatment, worry etc. and (c) the cost of testing, including time off work to take the test. Suppose the benefit of a true positive is B, and the cost of a false positive is C, the expected value is $EV = P(D \mid TP) \times B - P(ND \mid TP) \times C$. If the rule is to test if the present value is positive you test if

$$\frac{P(D \mid TP)}{P(ND \mid TP)} > \frac{C}{B}$$

Since people disagree about these costs and benefits, diagnostic screening is controversial.

When the screening is for an infectious disease to trace the infected persons contacts the costs of a false negative, the disease spreads, are different to the costs of a false negative from a diagnostic test when the person is no worse than they were before. Purpose matters.

6.2.2. 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 $0.45 = 1 - (0.5 + 0.4 - 0.35)$. 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, $P(N)$

$$P(N) = 1 - P(A \cup B) = 1 - (P(A) + P(B) - P(A \cap B))$$

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 probability 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.

6.3. Background Examples

These examples below illustrate some of the issues in probability, but are more complex than the exam questions that would be typically asked.

6.3.1. Birthdays

What is the probability that two people in a class of N people have the same birthday? It is easier to calculate the probability that they do not have the same birthday. In a spreadsheet put $N = 1, 2, 3, ..$ in the first row. Put 1 in cell A1. In A2 put =A1+1, copy this to the right. In Cell B1 put 1. In cell B2 put =A2*((365-B1)/365). Copy this to the right. For N=2, this is 364/365=0.9945, because one day is occupied; for N=3 it is (364/365)(363/365).because for the third person two days are occupied. See for what N it equals 50%.

6.3.2. The Lottery

In the UK lottery, before it was revised, six numbers were 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. In the new game of 6 out of 59 the probability of winning the jackpot in roughly 1 in 45 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.

6.3.3. 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.

7. 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 6.1 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 section6.1 . $\sum f(x_i) = 1$. 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 $p_i = f(x_i)$:

$$E(X) = \sum_{i=1}^{N} f(x_i)x_i = \mu. \tag{7.1}$$

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 as in section 2.

The variance of a random variable is defined as

$$V(X) = E(X - E(X))^2 = \sum_{i=1}^{N} f(x_i)(x_i - \mu)^2 = \sigma^2. \tag{7.2}$$

If $f(x_i) = 1/N$ this is just the same as the population variance we encountered in descriptive statistics. 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 in descriptive statistics. If the random variables are independent the covariance is zero. However, a covariance (or correlation) of zero does not imply that they are independent, unless the variables are normally distributed. Independence is a stronger property than uncorrelated.

8. Continuous random variables

Whereas a discrete random variable can only take specified values, x_i, $i = 1, 2, ..., N$; continuous random variables x (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*, denoted $f(x)$ for continuous random variables. Corresponding to the discrete cumulative distribution function, *cdf*, $F(x_i) = \Pr(X \leq x_i)$, there is a continuous *cdf*, $F(x) = \Pr(X \leq x)$ which gives the probability that the random variable will take a value less than or equal to a specified value x. 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 $P(X = x)$ is zero, although the *pdf* will be defined for it, $f(x) \neq 0$. 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 (7.1) and (7.2) for the discrete case by integrals so

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

8.1. 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, $f(x) = 0$ outside the range and $f(x) = 1/(b-a)$ within the range. By doing the integration you can show 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.

8.2. 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 whatever the distribution of the original variable, e.g. uniform, the distribution of the sample mean will be approximately normal 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 centred moment), it has a coefficient of skewness (third standardised moment) of zero and a coefficient of kurtosis (fourth standardised moment) of three..See section 2.1.5. 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, such as 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 $Y \sim N(\alpha, \sigma^2)$ which is said "Y normally distributed with expected value α and variance σ^2, where

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

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

If one variable $Y \sim N(\alpha, \sigma^2)$ then any linear function of Y is also normally distributed.

$$X = a + bY \sim N\left(a + b\alpha, b^2\sigma^2\right) \tag{8.1}$$

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 $u_i = Y_i - E(Y_i) = Y_i - \alpha$, and from the rules $u_i \sim N(0, \sigma^2)$. 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.

8.2.1. 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	0	0.5	1	1.5	2	2.5	3
$P(Z < z)$	0.5	0.6915	0.8413	0.9332	0.9772	0.9938	0.9987

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 of 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. The probability of being more than 3 standard deviations from the mean is 0.0013, once in a thousand observations, if the distribution is normal.

8.2.2. 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'.

8.3. Distributions related to the normal

Barrow chapter 6 discusses the χ^2 t and F distributions which are widely used functions of normally distributed variables. Suppose we have a sample $i = 1, 2, ..., n$ of standardised data $z_i \tilde{} \ IN(0, 1)$, independently distributed, standard normal.

8.3.1. Chi-squared

Suppose we form the sum of squares

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

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 estimate the variance as $s^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2/(n-1)$, so if x_i is normal $(n-1)s^2/\sigma^2$ has a χ^2 distribution with $n-1$ degrees of freedom.

8.3.2. t distribution

A standard normal divided by the square root of a Chi-squared distribution, divided by its degrees of freedom, 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 from the central limit theorem) 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.

8.3.3. 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.

9. Estimation

9.1. 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 $\sum \widehat{u}_t^2$. We will use these two procedures in three cases: (1) expected values (2) bivariate regression (3) multiple regression. In these three cases the two procedures happen to give identifical estimators, 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 whether a measured effect could really be zero or whether the number observed only differed from zero 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 three times, first for estimating the sample mean or expected value and testing hypotheses about it; then for the bivariate

regression model, where the expected value is not a constant, but depends on another variable, then for multiple regression, where it depends on many variables.

9.1.1. 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. In response to this the American Statistical Association issued in 2016 a "Statement on Statistical Significance and P-values".

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, discussed above in sections 6.2 and 6.3.3. 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 6.3.3. However, it can be mathematically more complex, usually involving integrals, though computers now make 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 are 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*, chapter 21 discusses these issues.

9.2. Estimating the expected value of Y_t

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.

9.2.1. Method of moments 1

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

9.2.2. Least squares 1

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, $\widehat{\sigma}^2 = \sum(Y_t - \widehat{\alpha})^2/T$.

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

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

9.3. 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.

9.3.1. 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.

9.3.2. 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

$$\begin{aligned} &= E\left(\frac{u_1}{T} + \frac{u_2}{T} + ... + \frac{u_T}{T}\right)\left(\frac{u_1}{T} + \frac{u_2}{T} + ... + \frac{u_T}{T}\right) \\ &= E\left(\frac{u_1^2}{T^2} + \frac{u_2^2}{T^2} + ... + \frac{u_T^2}{T^2} + \frac{u_1}{T}\frac{u_2}{T}...\right) \\ &= \frac{\sigma^2}{T^2} + \frac{\sigma^2}{T^2} + ... + \frac{\sigma^2}{T^2} + 0 + ... \end{aligned}$$

This product has 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:

$$V(\widehat{\alpha}) = E(\widehat{\alpha} - E(\widehat{\alpha}))^2 = \frac{\sigma^2}{T} \tag{9.1}$$

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.

9.3.3. 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 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.

In many cases we are interested in what happens to the properties of the estimator when T gets large: asymptotic properties. So although $\widehat{\sigma}^2$ which divides by T is a biased estimator of σ^2, as $T \to \infty$; $(T-1)/T \to 1$ and the bias goes away asymptotically. 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.

9.3.4. Background: Bias in estimating the variance

Let $x_t = \alpha + u_t$, with $E(x_t) = \alpha$; $Var(x_t) = E(x_t - E(x_t))^2 = E(u_t^2) = \sigma^2$; $E(u_k u_h) = 0$ for $k \neq h$ and $\bar{x} = \sum_{t=1}^{T} x_t/T$.

Consider

$$\begin{aligned} T^{-1}\sum_{t=1}^{T}(x_t-\bar{x})^2 &= T^{-1}\sum_{t=1}^{T}(x_t^2+\bar{x}^2-2\bar{x}x_t) = T^{-1}\sum_{t=1}^{T}x_t^2+\bar{x}^2-T^{-1}2\bar{x}\sum_{t=1}^{T}x_t \\ &= T^{-1}\sum_{t=1}^{T}x_t^2-\bar{x}^2. \end{aligned}$$

Consider the two terms separately

$$A \;:\; T^{-1}\sum_{t=1}^{T}x_t^2 = T^{-1}\sum_{t=1}^{T}(\alpha+u_t)^2 = \alpha^2+T^{-1}\sum_{t=1}^{T}u_t^2+2\alpha T^{-1}\sum_{t=1}^{T}u_t;$$

$$B \;:\; \bar{x}^2 = \left(T^{-1}\sum_{t=1}^{T}(\alpha+u_t)\right)^2 = \left(\alpha+T^{-1}\sum_{t=1}^{T}u_t\right)^2 = \alpha^2+T^{-2}\left(\sum_{t=1}^{T}u_t\right)^2+2\alpha T^{-1}\sum_{t=1}^{T}u_t.$$

Then take expected values of each term using $E(u_t)=0$, and $E(u_t^2)=\sigma^2$

$$A \;:\; E(\alpha^2+T^{-1}\sum_{t=1}^{T}u_t^2+2\alpha T^{-1}\sum_{t=1}^{T}u_t) = \alpha^2+T^{-1}(T\sigma^2) = \alpha^2+\sigma^2$$

$$B \;:\; E\left(\alpha^2+T^{-2}\left(\sum_{t=1}^{T}u_t\right)^2+2\alpha T^{-1}\sum_{t=1}^{T}u_t\right) = \alpha^2+T^{-2}(T\sigma^2) = \alpha^2+T^{-1}\sigma^2$$

Since $E\left(\sum_{t=1}^{T}u_t\right)^2 = T\sigma^2$ because $E(u_t^2)=\sigma^2$, and $E(u_k u_h)=0$.

So taking the difference of the expected values of $(A-B)$

$$E(T^{-1}\sum_{t=1}^{T}x_t^2-\bar{x}^2) = \left(\alpha^2+\sigma^2\right)-\left(\alpha^2+T^{-1}\sigma^2\right) = \frac{T-1}{T}\sigma^2.$$

9.3.5. Background Example: proportions

We also often use sample proportions to estimate probabilities. Barrow Chapter 4 covers proportions. For instance in 6.2.2 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.

9.4. Summary

So far we have (1) found out how to estimate the expected value of Y, $E(Y) = \alpha$ 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).

10. Confidence intervals and Hypothesis Tests

Earlier we noted that if a variable was normally distributed, the mean plus or minus 1.96 standard deviations would be expected to cover 95% of the observations. This range, plus or minus 1,96 (often approximatd to 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.

10.1. Confidence intervals

When we make an estimate or forecast we also need to indicate the degree of uncertatinty as well as the expected value. In answer to the question how long does it take to get to the airport, the answer "Between one and three hours, depending on traffic" is more useful than "Two hours, on average".

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, by the central limit theorem, even if u_t is not normal. We saw that

$$\widehat{\alpha} = \alpha + \sum_t u_t/T$$

So

$$\widehat{\alpha} \sim N(\alpha, \sigma^2/T)$$

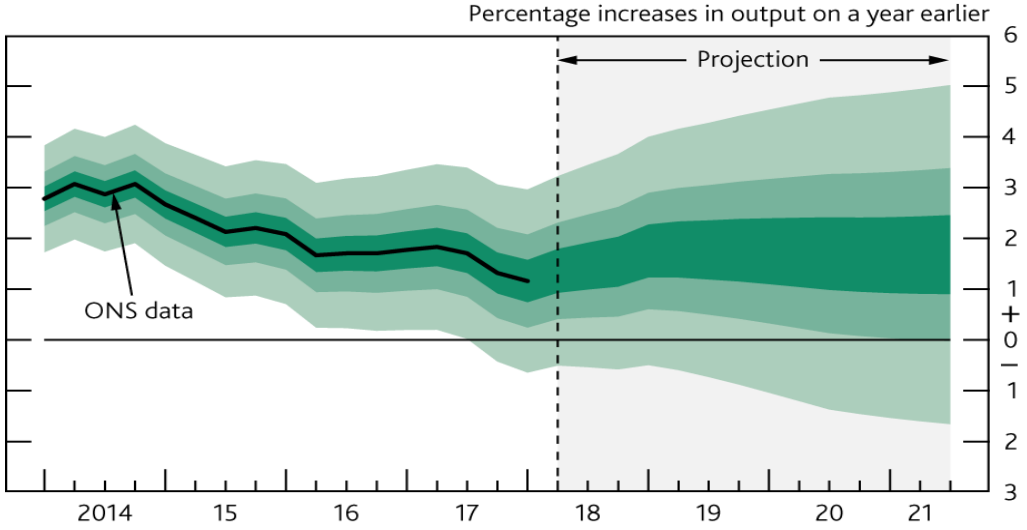
$s(\widehat{\alpha}) = \sigma/\sqrt{T}$ is the standard error of $\widehat{\alpha}$. Be clear when we are taling about the variance (standard deviation) of the variable or the variance (standard error) of the mean.

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 $\widehat{\alpha} \pm s(\widehat{\alpha})$: in repeated samples 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. We discuss this idea more formally below under testing.

10.1.1. Confidence Intervals for GDP growth

In its forecasts of GDP growth, the Bank of England wants to provide various confidence intervals, 30%, 60% and 90%. It wants to do this for every quarter. The way it presents this is as a fan chart that depicts the probability of various outcomes for GDP growth. The one for August 2018 is shown below. The central band gives the 30% confidence interval, the next band the 60% confidence interval and the outside band the 90% conficence interval. The 90% confidence interval for the Banks forecast of growth in mid 2021 is in the range +5% to -1.5%. In any particular quarter of the forecast period, GDP growth is therefore expected to lie somewhere within the fan on 90 out of 100 occasions. The forecast is based on various assumptions including that interest rates will follow market expectations. The dark line is the Office of National Statistics estimate of the growth rate at the time of the forecast. Since GDP data are often revised, the Bank put a confidence interval over the past data as well. In August 2018 ONS thought growth in 2014 was around 3%, but the 90% confidence interval stretches from 2% to 4%. We do not know where we have been, let alone where we are going.



Bank of England Fan Chart for GDP growth, August 2018.

10.2. Small samples

Above we assumed that we knew the true standard deviation, σ, so the standard error is $s(\widehat{\alpha}) = \sigma/\sqrt{T}$ then $(\widehat{\alpha} - \alpha_0)/s(\widehat{\alpha})$ has a normal distribution. But when we estimate the standard deviation σ by

$$s = \sqrt{\frac{\sum(Y_t - \widehat{\alpha})^2}{T-1}}$$

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 in 8.3, 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.

10.3. 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, something has no effect, i.e. $\alpha_0 = 0$. We reject it in favour of what is called the alternative hypothesis, say $H_1 : \alpha \neq \alpha_0$; if there is very strong evidence against the null hypothesis.

Above in discussing confidence intervals, we said that 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.

The test statistic τ measures the difference between the data $\widehat{\alpha}$ and the hypothesised value α_0 given by the model. We can also calculate the 'p value', the probability that the test statistic τ would have been at least as large as its observed value if all the assumptions of the model including the hypothesis being tested were correct. The assumptions of the model include such things as the normal distribution and how the data were collected. It can be regarded as a measure of the fit of the model, including the hypothesis being tested, to the data: the smaller the value the worse the fit.[3] Most computer programs give p values and they are more convenient to use for distributions other than the normal, since you do not need to look up critical values for t, F or χ^2 distributions. 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. The p value is the probability that the test test statistic τ would have been at least as large as its observed value if all the assumptions of the model including the hypothesis being tested were correct.

With the null hypothesis, $H_0 : \alpha = \alpha_0$ the alternative hypothesis, $H_1 : \alpha \neq \alpha_0$; 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. 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, H_0 : defendant is innocent. We have an alternative hypothesis, H_1 : 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

[3] P values are often misinterpreted and the American Statistical Association has issued "The ASA's statement on p-values: context, process and purpose" R.L. Wasserstein & N. A. Lazar, The American Statistician, 2016.

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 villains 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 costs or losses that result from the two types of error. The 18th century English lawyer, William Blackstone, said "it is better that 10 guilty escape than one innocent suffers."

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 discussing probability, 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.

10.3.1. 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 $\sigma/\sqrt{T}$.

10.3.2. 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.

Question

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 $\widehat{p}_2$ 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 people can be influenced by how you frame the probabilities, e.g. in terms of proportion who die or proportion who survive.

11. 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.

11.1. Examples

11.1.1. 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.

11.1.2. Fisher Relation

The real interest rate, RR_t is the nominal interest rate R_t less the rate of inflation π_t

$$RR_t = R_t - \pi_t$$

suppose the real interest rate is roughly constant, equal to a constant plus a random error

$$RR_t = r + u_t$$

then we can write

$$R_t = RR_t + \pi_t = r + \pi_t + u_t.$$

Then if we ran a regression

$$R_t = \alpha + \beta\pi_t + u_t \tag{11.1}$$

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 of a percentage point. Using the data in the Shiller file for short term interest rates, R_t and CPI inflation, π_t, INF, for $t = 1960 - 2011$ and the program EViews gives:

Dependent Variable: R
Method: Least Squares
Date: 10/05/16 Time: 18:12
Sample (adjusted): 1960 2011
Included observations: 52 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.580207	0.551243	4.680706	0.0000
INF	0.901890	0.115322	7.820640	0.0000

R-squared	0.550208	Mean dependent var	6.096442
Adjusted R-squared	0.541212	S.D. dependent var	3.395464
S.E. of regression	2.299880	Akaike info criterion	4.541294
Sum squared resid	264.4725	Schwarz criterion	4.616342
Log likelihood	-116.0736	Hannan-Quinn criter.	4.570065
F-statistic	61.16241	Durbin-Watson stat	0.828518
Prob(F-statistic)	0.000000		

We can write this as

$$\begin{array}{rcccccc} R_t = & 2.58 & +0.90 & \pi_t & +\widehat{u}_t & R^2 = 0.55 \\ & (0.55) & (0.12) & & & SER = 2.3 \end{array}$$

where standard errors of the coefficients are given in parentheses. $\widehat{\alpha} = 2.58$ is the value that would be predicted for interest rates if inflation was zero. The equation says that if inflation increases by one percentage point, interest rates increase by 0.90 percentage points. This is not significantly different from unity since $t(\beta = 1) = (0.90 - 1)/0.12 = -0.83$. Thus the hypothesis that $\beta = 1$ is not rejected. The estimated residuals are

$$\widehat{u}_t = R_t - \widehat{R}_t = R_t - (\widehat{\alpha} - \widehat{\beta}\pi_t)$$

the difference between the actual value of the interest rate and its predicted value. R^2 tells you the proportion of the variation in the interest rate over the period

that has been predicted, 55%. The Durbin Watson statistic, DW, should be 2 and at 0.83 it tells us that one of our assumptions, no serial correlation, $E(u_t u_{t-1}) = 0$ fails to hold.

11.2. Deriving the Least Squares Estimator

In the examples above, there is data, $t = 1, 2, ..., T$, and a model, which explains the dependent variable Y_t by an independent variable X_t of the form

$$Y_t = \alpha + \beta X_t + u_t.$$

This is a set of T equations. As above we assume that u_t is a random error with expected value zero, $E(u_t) = 0$ and constant variance $E(u_t^2) = \sigma^2$, where the errors are uncorrelated or independent through time, $E(u_t u_{t-i}) = 0$. We will further assume that X_t varies, $Var(X_t) \neq 0$, and is exogenous, independent of the error so that the covariance between X_t and u_t 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 estimated residuals, $\widehat{u}_t$.

As with the expected value above there are two procedures that we will use to derive the estimates, method of moments and least squares.

11.2.1. Method of Moments 2

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: $E\{(X_t - E(X_t))u_t\} = 0$. The sample equivalent of the expected value is the mean. The method of moment procedure says 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 residuals, $\widehat{u}_t$, 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

$$\begin{aligned} Y_t &= \widehat{\alpha} + \widehat{\beta}X_t + \widehat{u}_t = \left(\overline{Y} - \widehat{\beta}\overline{X}\right) + \widehat{\beta}X_t + \widehat{u}_t \\ Y_t - \overline{Y} &= \widehat{\beta}(X_t - \overline{X}) + \widehat{u}_t \\ y_t &= \widehat{\beta}x_t + \widehat{u}_t. \end{aligned} \tag{11.2}$$

Where we use lower case letters to denote deviations from the mean, $y_t = Y_t - \overline{Y}$, $x_t = X_t - \overline{X}$.

In terms of deviations from the means the sample equivalent of our second moment-condition: $E\{(X_t - E(X_t))u_t\} = 0$ is that the mean of $(X_t - \overline{X})\widehat{u}_t$ is zero:

$$\begin{aligned} T^{-1}\sum_t x_t\widehat{u}_t &= 0 \\ T^{-1}\sum_t x_t(y_t - \widehat{\beta}x_t) &= T^{-1}\sum_t x_t y_t - \widehat{\beta}\left\{T^{-1}\sum_t x_t^2\right\} = 0 \\ \widehat{\beta} &= \left\{T^{-1}\sum_t x_t y_t\right\} / \left\{T^{-1}\sum_t x_t^2\right\} = \left\{\sum_t x_t y_t\right\} / \left\{\sum_t x_t^2\right\}. \end{aligned}$$

This says that our estimate of $\widehat{\beta}$ is the ratio of the covariance of X_t and Y_t to the variance of X_t, (remember lower case letters denote deviations from the means and we can divide both the denominator and the numerator by either T or $T-1$).

11.2.2. Least squares 2

As with the expected value we can also find the $\widehat{\beta}$ that minimises $\sum \widehat{u}_t^2$, where from (11.2)

$$\sum_t \widehat{u}_t^2 = \sum_t \left(y_t - \widehat{\beta} x_t\right)^2 = \sum_t y_t^2 + \widehat{\beta}^2 \sum_t x_t^2 - 2\widehat{\beta} \sum_t x_t y_t$$

the derivative of $\sum \widehat{u}_t^2$ with respect to $\widehat{\beta}$ is

$$\frac{\partial \sum \widehat{u}_t^2}{\partial \widehat{\beta}} = 2\widehat{\beta} \sum_t x_t^2 - 2 \sum_t x_t y_t = 0 \tag{11.3}$$

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

$$\frac{\partial^2 \sum \widehat{u}_t^2}{\partial \widehat{\beta}^2} = 2 \sum x_t^2 > 0 \tag{11.4}$$

since squares are positive, so this is a minimum.

Our estimates

$$\begin{aligned} \widehat{\alpha} &= \overline{Y} - \widehat{\beta} \overline{X} \\ \widehat{\beta} &= \frac{\sum_t (X_t - \overline{X})(Y_t - \overline{Y})}{\sum_t (X_t - \overline{X})^2} = \frac{\sum_t x_t y_t}{\sum_t x_t^2} \end{aligned}$$

(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.

11.3. Properties of the estimates

If the expected value of the random variable $\widehat{\beta}$ (it is different in every sample) equals its true value

$$E(\widehat{\beta}) = \beta$$

then $\widehat{\beta}$ is said to be unbiased.

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 and treating x_t as fixed

$$\begin{aligned} V(\widehat{\beta}) &= E(\widehat{\beta} - \beta)^2 = E(\frac{\sum x_t u_t}{\sum x_t^2})^2 \\ &= \frac{1}{(\sum x_t^2)^2} E\left(\sum x_t u_t\right)^2 \end{aligned} \tag{11.5}$$

We can write $E\left(\sum x_t u_t\right)^2$ as

$$\begin{aligned} &E(x_1 u_1 + x_2 u_2 ... + x_T u_T)(x_1 u_1 + x_2 u_2 ... + x_T u_T) \\ &E(x_1^2 u_1^2 + x_2^2 u_2^2 + ... + x_T^2 u_T^2 + 2 x_1 u_1 x_2 u_2 + ...) \\ &x_1^2 \sigma^2 + x_2^2 \sigma^2 + ... + x_T^2 \sigma^2 + 0 + ... \\ &\sigma^2 \sum x_t^2 \end{aligned}$$

since for $t = 1, 2, ..., T$, $E(u_t^2) = \sigma^2$ and $E(u_t u_{t-i}) = 0. So$

$$\begin{aligned} V(\widehat{\beta}) &= \frac{1}{(\sum x_t^2)^2} \left(\sigma^2 \sum x_t^2\right) \\ &= \frac{\sigma^2}{\sum x_t^2} \end{aligned}$$

Note we are using the same sort of argument as in deriving the standard error of the mean (9.1) above. Note that $\sum x_t^2$ rises and the variance falls with T, the sample size as in the case of a mean.

The residuals are

$$\widehat{u}_t = y_t - \widehat{\beta} x_t = Y_t - \widehat{\alpha} - \widehat{\beta} X_t$$

and the unbiased estimator of σ^2 is

$$s^2 = \sum \widehat{u}_t^2 / (T - 2)$$

because we estimate two parameters $\widehat{\alpha}$ and $\widehat{\beta}$. Our estimator for the standard error of $\widehat{\beta}$ is the square root of $V(\widehat{\beta})$ with σ replaced by s :

$$se(\widehat{\beta}) = s / \sqrt{\sum x_t^2}.$$

11.4. Predicted values and residuals are uncorrelated

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 or fitted value, and the unexplained bit, the residual, $\widehat{u}_t$: the part of Y_t 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-conditions made $\sum \widehat{u}_t = 0$ and $\sum X_t \widehat{u}_t = 0$, so both terms are zero. 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.

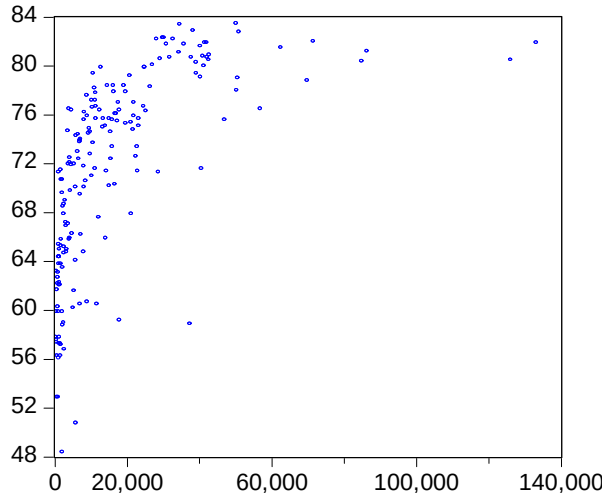
By construction the residuals have mean zero (if there is an intercept, i.e. α, in the equation, and you should always include an intercept) and they are uncorrelated with the explanatory variables. But we can test whether other assumptions we made about the errors hold for the estimated residuals, see section 16.5. In many cases the best way to check the assumptions is to look at graphs of residuals (which should look random, with no obvious pattern in them). and of the histogram (which should look roughly normal).

12. Example: Life expectancy

We will examine a simple linear regression using cross-country data for life expectancy taken from Gapminder. .

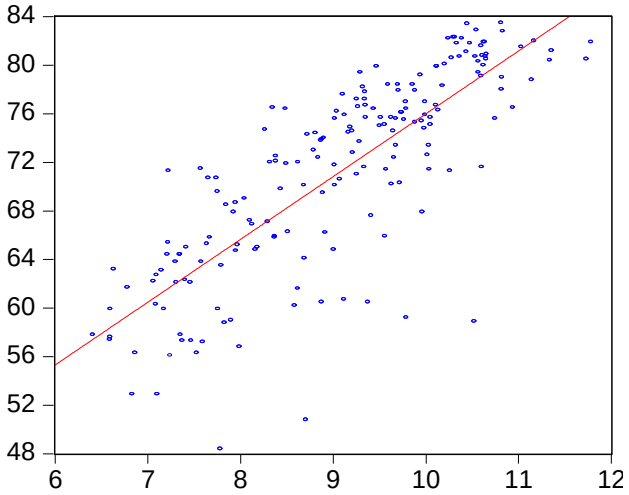
Linear regression is very flexible because we can redefine the variables by some transformation, which allows the underlying relationship to be non-linear, but the relationship we estimate to be linear. The most common transformation is to logarithms We can also construct artificial variables, like trends and dummy variables. We look at relationships involving these below.

Below is a scatter diagram of life expectancy on per-capita income in 2013, from Gapminder. This data is on Moodle



Scatter LE with per capita income

However if we use log per-capita income we get a relationship that is much closer to being linear. What counts is not the dollar increase in income but the percentage increase in income.



Scatter LE with log per capita income.

12.1. Logarithms

We often use logarithms of economic variables since

1. prices and quantities are non-negative so the logs are defined
2. the coefficients can be interpreted as elasticities, % change in the dependent variable in response to a 1% change in the independent variable, so the units of measurement of the variables do not matter
3. in many cases errors are proportional to the variable, so the variance is more likely to be constant in logs,
4. the logarithms of economic variables are often closer to being normally distributed
5. the change in the logarithm is approximately equal to the growth rate and
6. lots of interesting hypotheses can be tested in logarithmic models.
7. often effects are proportional, which is captured by logarithmic models.

Normally we use natural logarithms to the base e.

12.2. Regression output

Below is the output that gives the regression line shown on the scatter diagram above between LE and log per capita income.

Dependent Variable: LE
Method: Least Squares
Date: 08/18/18 Time: 16:13
Sample: 1 189
Included observations: 189

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	24.27829	2.568390	9.452729	0.0000
LPCI	5.172650	0.279985	18.47472	0.0000

R-squared	0.646045	Mean dependent var	71.28771
Adjusted R-squared	0.644152	S.D. dependent var	8.050099
S.E. of regression	4.802125	Akaike info criterion	5.986520
Sum squared resid	4312.296	Schwarz criterion	6.020824
Log likelihood	-563.7261	Hannan-Quinn criter.	6.000417
F-statistic	341.3155	Durbin-Watson stat	2.176566
Prob(F-statistic)	0.000000		

We can write this as for countries $i = 1, 2, ..., 189$

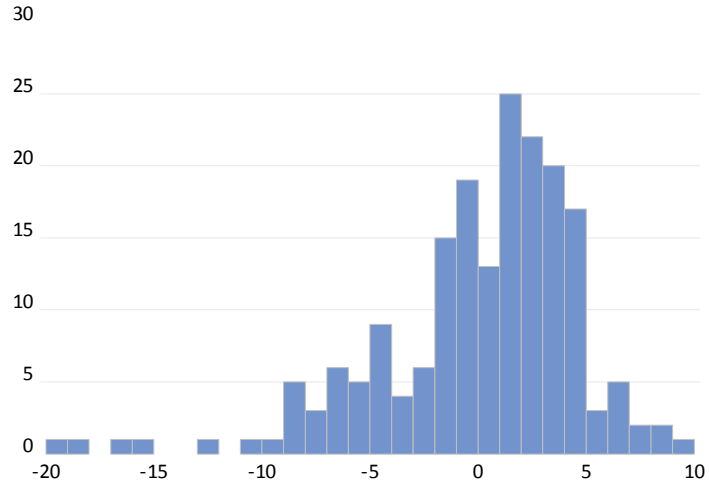
$$\begin{aligned} LE_i &= \alpha + \beta LPCI_i + u_i \\ LE_i &= 24.27 + 5.17 LPCI + \widehat{u}_i \end{aligned}$$

Consider a country with annual per-capita income of \$22,026, which is a bit above the mean. The natural log of 22,026. The predicted life expectancy for this country, rounding the numbers, is

$$\widehat{LE} = 24 + 5.2 * 10 = 76.$$

The S.E. of regression of 4.8 gives you the average error, which is about 5 years. If the errors were normally distributed, which they are not, the 95% confidence interval would be $76 \pm 1.96 * 4.8$, roughly 67 to 85.

There are a number of countries which have large negative residuals, a lot below the regression line. In the hisogram, you can see these on the left, which casuses the distribution to be negatively skewed, (skewness =-1.3, rather than zero) and to have fat tails (kurtosis=5.76 rather than 3).



Histogram of residuals.

We can remove the effect of these 4 "outliers" by using dummy variables discussed below.

13. Multiple Regression and properties of least squares

Most of the time we have more than one right hand side variable. For instance, the price of house $i = 1, 2, ..., N$ is determined by lots of variables: number of rooms, floor area, location, amenities, schools etc. which we can write:

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i.$$

$X_{1i} = 1$ all i so we do not need to write it in. Suppose X_{2i} is the number of rooms and X_{3i} is the floor area. Then β_2 is the effect of a change in the number of rooms on price holding all the other included variables constant. This may not always have what you think is the expected sign. Generally houses with more rooms cost more. But incresing the number of rooms, while holding floor area constant, means that on average rooms are smaller. This may not be desirable, so β_2 may be negative.

Similarly for time series variables. Suppose we have a constant, $X_{1i} = 1$, and three variables which we can write

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t \tag{13.1}$$

for $t = 1, 2, ..., T$. Again 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 with all 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.

13.1. Assumptions for Least Squares to give good estimates

We are assuming a linear regression model with k parameters:

$$Y_t = \beta_1 + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t \tag{13.2}$$

- 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 (which will pick up the average of the error) 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 (same variances), its failure is that the errors are heteroskedastic (different variances).

- That $E(u_t u_{t-i}) = 0$, for $i \neq 0$. Errors are independent, with no serial correlation or no autocorrelation.

- In a bivariate regression 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. In a multiple regression this corresponds to there being no exact linear relationship between any of the variables (no exact multicollinearity). Again we would not be able to calculate the regression coefficients if this assumption failed.

- That $E\{(X_{it} - E(X_{it}))u_t\} = 0$. This assumption is usually described as each of the X_i's, X_{2t}, X_{3t}, X_{4t} being exogenous: not related to the errors. Notice that for our estimates $T^{-1}\sum_t (X_{it} - \overline{X}_i)\widehat{u}_t = 0$ by construction. For exogeneity, the X's may either be non stochastic, fixed numbers, (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. If the explanatory variable is only uncorrelated, typically the case for a lagged dependent variable included in the equation, it is said to be predetermined or weakly exogenous.

Implicitly 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. Having the wrong model would generally cause failure of some of the assumptions above.

13.2. Properties of Least Squares

With these assumptions we can show that the least squares estimators are unbiased and among all estimators of 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, 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. We can predict weight from height, or height from weight without assuming one causes the other.

13.3. Consequences of failure of the assumptions

- If the X are not exogenous, independent of the errors, the estimates are biased. If the X are not exogenous but a weaker assumption holds, the

X are predetermined (or weakly exogenous), in that they are uncorrelated with the errors, the estimates are biased but consistent, the bias goes away as the sample size gets large. This is typically the case for lagged dependent variables in time series. If the X are correlated with the errors the estimates are inconsistent.

- If either (a) $E(u_t^2) \neq \sigma^2$, heteroskedasticity, the variance is not constant, or (b) $E(u_t u_{t-i}) \neq 0$, serial correlation the errors are correlated; then $\widehat{\beta}_i$ remains unbiased, but is no longer minimum variance and the standard errors are wrong. Robust standard errors are available. Serial correlation or heteroskedasticity of the estimated residuals may be caused by misspecification rather than serial correlation or heteroskedasticity of the errors.

- If a variable does not vary of there is an exact linear relationship between some of the independent variables one cannot obtain estimates $\widehat{\beta}$. EViews will not give you any estimates and the message "Near singular matrix error. The regressors may be perfectly collinear." Stata will drop a variable to break the exact linear relationship.

14. Matrix form of the Linear Regression Model

We can write (13.1) in vector form

$$Y_t = \beta' X_t + u_t$$

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 (13.1) in matrix form in terms of y a $T \times 1$ vector and X a $T \times 4$ matrix

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ Y_T \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & X_{41} \\ 1 & X_{22} & X_{32} & X_{42} \\ . & . & . & . \\ 1 & X_{2T} & X_{3T} & X_{4T} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ . \\ u_T \end{bmatrix}$$

$$\underset{(T \times 1)}{y} = \underset{(T \times 4)}{X} \underset{(4 \times 1)}{\beta} + \underset{(T \times 1)}{u}.$$

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.

14.1. 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:

$$E(uu') = \begin{bmatrix} E(u_1^2) & E(u_1u_2) & .. & E(u_1u_T) \\ E(u_1u_2) & E(u_2^2) & .. & E(u_2u_T) \\ .. & .. & .. & .. \\ E(u_1u_T) & E(u_2u_T) & .. & E(u_T^2) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & .. & 0 \\ 0 & \sigma^2 & .. & 0 \\ .. & .. & .. & .. \\ 0 & 0 & .. & \sigma^2 \end{bmatrix}$$

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.

14.2. Differentiation with vectors and matrices

Consider the linear relation:

$$P = \underset{1 \times n}{x'} \; \underset{n \times 1}{a}$$

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:

$$P = [x_1, x_2] \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = x_1 a_1 + x_2 a_2$$

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:

$$Q = \underset{1\times n}{x'} \underset{(n\times n)}{A} \underset{(n\times 1)}{x}$$

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:

$$Q = [x_1, x_2] \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{12}x_1 + a_{22}x_2 \end{bmatrix} = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2$$

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

$$\frac{dQ}{dx} = \begin{bmatrix} \frac{dQ}{dx_1} \\ \frac{dQ}{dx_2} \end{bmatrix} = \begin{bmatrix} 2a_{11}x_1 + 2a_{12}x_2 \\ 2a_{12}x_1 + 2a_{22}x_2 \end{bmatrix} = 2 \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2 \underset{(2\times 2)}{A} \underset{(2\times 1)}{x}$$

and

$$\begin{aligned}\frac{dQ}{dx'} &= \left[\frac{dQ}{dx_1}, \frac{dQ}{dx_2}\right] = [2a_{11}x_1 + 2a_{12}x_2, 2a_{12}x_1 + 2a_{22}x_2] \\ &= 2\,[x_1, x_2]\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} = 2\underset{1\times 2}{x'}\underset{2\times 2}{A}\end{aligned}$$

14.3. 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 $T \times 1$ vectors, β is a $k \times 1$ vector and X a $T \times k$ matrix.

14.3.1. Method of moments 3

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

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

14.3.2. Least Squares 3

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 $\widehat{u}$

$$\begin{aligned}\widehat{u}'\widehat{u} &= (y - X\widehat{\beta})'(y - X\widehat{\beta}) \\ &= y'y + \widehat{\beta}'X'X\widehat{\beta} - 2\widehat{\beta}'X'y\end{aligned} \tag{14.1}$$

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. Note $\widehat{\beta}' X'y = y'X\widehat{\beta}$ because they are both scalars, 1×1 matrices. In general we cannot set matrices equal to their transposes unless they are symmetric, but we can with scalars. $X'X$ and $(X'X)^{-1}$ are symmetric matrices equal to their transposes, see (14.2) and (14.3) . 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.

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, like (11.3) , are the $k \times 1$ set of equations,

$$\frac{\partial \widehat{u}'\widehat{u}}{\partial \widehat{\beta}} = 2X'X\widehat{\beta} - 2X'y = 0$$

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, like (11.4) , , 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. 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.

14.4. The bivariate case in matrix algebra

Consider the bivariate case

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

for $t = 1, 2, ..., T$. The residuals are $\widehat{u}_t = Y_t - \widehat{\beta}_1 - \widehat{\beta}_2 X_t$. Least squares chooses $\widehat{\beta}_i$ to minimise $\sum \widehat{u}_t^2$.

The model, $y = X\beta + u$, is

$$\begin{bmatrix} Y_1 \\ Y_2 \\ .. \\ Y_T \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ .. & .. \\ 1 & X_T \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ .. \\ u_T \end{bmatrix}$$

$$(X'X) = \begin{bmatrix} T & \sum X_t \\ \sum X_t & \sum X_t^2 \end{bmatrix} \tag{14.2}$$

$$(X'X)^{-1} = \frac{1}{T\sum X_t^2 - (\sum X_t)^2} \begin{bmatrix} \sum X_t^2 & -\sum X_t \\ -\sum X_t & T \end{bmatrix} \tag{14.3}$$

For the inverse to exist the matrix must be non-singular, with a non-zero determinant. The determinant of $X'X$ is $T\sum X_t^2 - (\sum X_t)^2$. Note $\sum X_t^2/T - (\sum X_t/T)^2$ is the variance of X_t, so we need a non-zero variance as before.

$$X'y = \begin{bmatrix} \sum Y_t \\ \sum X_t Y_t \end{bmatrix}.$$

In the bivariate case (14.1) is

$$\begin{aligned} \sum_{t=1}^{T} \widehat{u}_t^2 &= \sum Y_t^2 + (\widehat{\beta}_1^2 T + \widehat{\beta}_2^2 \sum X_t^2 + 2\widehat{\beta}_1\widehat{\beta}_2 \sum X_t) \\ &\quad -2(\widehat{\beta}_1 \sum Y_t + \widehat{\beta}_2 \sum X_t Y_t) \end{aligned}$$

15. Properties of Least Squares

15.1. Expected Value of $\widehat{\beta}$

$$\begin{aligned} \widehat{\beta} &= (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + u) \\ \widehat{\beta} &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u \\ &= \beta + (X'X)^{-1}X'u \end{aligned} \tag{15.1}$$

$$E(\widehat{\beta}) = \beta + E((X'X)^{-1}X'u)$$

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.

15.2. Variance Covariance matrix of $\widehat{\beta}$

From (15.1) we have

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

The variance-covariance matrix of $\widehat{\beta}$ is a $k \times k$ matrix

$$V(\widehat{\beta}) = E(\widehat{\beta} - E(\widehat{\beta}))(\widehat{\beta} - E(\widehat{\beta}))' = E(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)'$$

since $\widehat{\beta}$ is unbiased. But from (15.1) 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 (11.5) above. We derive the variance covariance matrix conditional on the observed sample, which is why we can take the expected value inside in line 3.

We estimate $V(\widehat{\beta})$ by

$$\widehat{V(\beta)} = s^2(X'X)^{-1}$$

where $s^2 = \widehat{u}'\widehat{u}/(T-k)$. The square roots of the ith diagonal element of $s^2(X'X)^{-1}$ gives the standard errors of $\widehat{\beta}_i$ the ith elements of $\widehat{\beta}$, which is reported by computer programs.

15.3. Predicted Values and residuals

The predicted values are $\widehat{y} = X\widehat{\beta}$; the residuals

$$\widehat{u} = y - \widehat{y} = y - X\widehat{\beta} = y - X(X'X)^{-1}X'y = (I - X(X'X)^{-1}X')y = My;$$

where $M = I - P_x$, and $P_x = X(X'X)^{-1}X'$ where P_x is called a projection matrix. Both M and P_x are idempotent, equal to their product $P_xP_x = P_x$ and $MP_x = 0$. So

$$\widehat{u} = My = M(X\beta + u) = MX\beta + Mu = Mu,$$

since $MX\beta = (I - P_x)X\beta = X\beta - X(X'X)^{-1}X'X\beta = X\beta - X\beta = 0$.

15.4. Gauss-Markov Theorem

We have often claimed that the mean or regression coefficients are the minimum variance estimators in the class of linear unbiased estimators, they are BLUE, best linear unbiased estimators. We now prove it. Note this does not require assuming normality. 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 any other linear estimator $\widetilde{\beta} = Cy$ where 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 this is the case to ensure that $\widetilde{\beta}$ is unbiased.The variance covariance matrix of $\widetilde{\beta}$ is:

$$E(\widetilde{\beta} - E(\widetilde{\beta}))(\widetilde{\beta} - E(\widetilde{\beta}))' = E(\widetilde{\beta} - \beta)(\widetilde{\beta} - \beta)'$$

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:

$$\begin{aligned} E((X'X)^{-1}X'uu'X(X'X)^{-1}) &= \sigma^2(X'X)^{-1} \\ E(Wu\ u'W') &= \sigma^2 WW' \\ E((X'X)^{-1}X'uu'W') &= \sigma^2(X'X)^{-1}X'W' = 0 \\ E(Wuu'X(X'X)^{-1}) &= \sigma^2 WX(X'X)^{-1} = 0 \end{aligned}$$

The last two terms are zero since $WX = X'W' = 0$. 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.

15.5. Consequences of failure of the assumptions

$E(X'u) = 0$. If this does not hold, because of failure of the exogeneity assumption, estimates are biased and inconsistent. Exogeneity fails because of simultaneity (the dependebnt variable also determines the independent variable), measurement error, or ommitted variables.

$E(uu') = \sigma^2 I_T$. If this does not hold, because of serial correlation or heteroskedasticity of the errors, $\widehat{\beta}$ remains unbiased, but is no longer minimum variance and the standard errors are wrong: $V(\widehat{\beta}) \neq s^2(X'X)^{-1}$. Robust standard errors are available. Serial correlation or heteroskedasticity of the estimated residuals may be caused by misspecification rather than serial correlation or heteroskedasticity of the errors.

$(X'X)$ non-singular, rank of X equals k. If this does not hold, because of exact linear dependence in the data, there are no estimate.

Notice that a normal distribution is NOT one of the assumptions that we require for the Gauss-Markov theorem. So if the other assumptions hold OLS is BLUE even if the residuals are non-normal.

15.6. Ommitted variables

If you have unhealthy residuals, that show serial correlation or heteroskedasticity, some 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 may just treat the symptoms, it may be better to cure the disease: specify the model correctly.

Suppose the model is

$$y_t = \beta x_t + \gamma z_t + u_t \tag{15.2}$$

expressed in deviations from mean, to remove the constant. But we leave out z_t

and estimate:

$$y_t = bx_t + v_t. \tag{15.3}$$

There are two estimates of the coefficient on x_t : b and β. To understand the relation between thems we need to look at the relationship between x_t and z_t, summarised by the regression equation:

$$z_t = dx_t + w_t \tag{15.4}$$

w_t is just the part of z_t that is not correlated with x_t. d may be zero, if there is no relationship. Put (15.4) into (15.2) and we get (15.3):

$$y_t = \beta x_t + \gamma(dx_t + w_t) + u_t$$

$$y_t = (\beta + \gamma d)x_t + (\gamma w_t + u_t)$$

So $b = (\beta + \gamma d)$. When z_t is left out the coeficient of x_t picks up the part of z_t that is correlated with x_t. Parts of z_t that are not correlated with x_t end up in the error term $(\gamma w_t + u_t)$. This causes correlation between x_t and the error term, a failure of exogeneity. Looking for patterns in the error term is important, it may suggest a 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.

16. Testing

16.1. Tests for a single hypothesis on individual coefficients

Suppose we have the model

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t \tag{16.1}$$

we can test the significance of the individual coefficients, say β_2 using t ratios exactly as we did for the mean, where the null hypothesis is $H_0 : \beta_2 = 0$, and the test statistic is

$$t(\beta_2 = 0) = \frac{\widehat{\beta}_2}{se\left(\widehat{\beta}_2\right)}$$

where $se\left(\widehat{\beta}_2\right)$ is the estimated standard error of $\widehat{\beta}_2$. If this t ratio is greater than two in absolute value we conclude that $\widehat{\beta}_2$ 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 (roughly) the probability that the null hypothesis is true in this data. If the p value is less than 0.05, we reject the hypothesis at the 5% level.

We can test against other values. Suppose economic theory suggested that $\beta_3 = 1$ the t statistic for testing this would be

$$t(\beta_3 = 1) = \frac{\widehat{\beta}_3 - 1}{se\left(\widehat{\beta}_3\right)}$$

and if this t statistic is greater than two in absolute value we conclude that $\widehat{\beta}_3$ is significantly different from unity (one) at the 5% level. Notice that it might be significantly different from both zero and one, if the confidence interval did not cover either value.

16.1.1. Normality

If, in addition to our Gauss-Markove assumptions, 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, see (8.1). The matrix equivalent is that, for $k \times 1$ vectors, Y and M and $k \times k$ matrix Σ (not the summation sign) $Y \sim N(M, \Sigma)$, then for $h \times 1$ vectors X and A, and $h \times k$ matrix B

$$X = A + BY \sim N(A + BM, B\Sigma B') \tag{16.2}$$

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: $y \sim N(X\beta, \sigma^2 I)$. Since $\widehat{\beta} = (X'X)^{-1}X'y$ is a linear function of y, applying (16.2) it is normally distributed $\widehat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$. This says that $\widehat{\beta}$ is normally distributed with expected value β (and is therefore unbiased) and variance covariance matrix $\sigma^2(X'X)^{-1}$.

While we require normality for our distributional results and tests in small samples, in practice we rely on the Central Limit Theorem which says that in large samples $\widehat{\beta}$ is normally distributed even if u is not normally distributed.

16.2. Tests on joint hypotheses

Suppose that in (16.1) we wanted to test both the hypotheses, $\beta_2 = 0$ and $\beta_3 = 1$, at the same time, rather than separately as above. To do this we would use an F test. The test statistics used to test joint hypotheses follows a F distribution, introduced in section 8.3. Just as the t distribution is described by its single degree of freedom, the F distribution is described by its two degrees of freedom: the number of hypotheses being tested, two in this case, and the number of observations less the number of parameters estimated: $T - 4$ in this case. This would be written $F(2, T-4)$ and critical values are given in statistics books. To do the test, we would first estimate the unrestricted model (16.1) and get the unrestricted residual sum of squares: $URSS = \sum \widehat{u}_t^2$. Then we would estimate the restricted model which imposes the restrictions, (dropping X_{2t} since $\beta_2 = 0$ and bringing X_{3t} to the left hand side, since we do not need to estimate its coefficient)

$$Y_t - X_{3t} = \beta_1 + \beta_4 X_{4t} + r_t$$

from this we get the restricted residual sum of squares $RRSS = \sum \widehat{r}_t^2$. For this case with T observations, 2 restrictions and 4 parameters in the unresricted model, the test statistic is

$$\frac{\left(\sum \widehat{r}_t^2 - \sum \widehat{u}_t^2\right)/2}{\sum \widehat{u}_t^2/(T-4)} \tilde{} F(2, T-4)$$

For the general case with m restrictions and k parameters in the unrestricted model it is

$$\frac{(RRSS - URSS)/m}{URSS/(T-k)} \tilde{} F(m, T-k).$$

16.3. Tests for the insignificance of regressors

A common hypothesis that most regression programs test is that none of the independent variables had any effect on the dependent variable. For (16.1) the hypothesis is $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$. 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 at the 5% level if $p < 0.05$. Notice that the joint hypothesis that all coefficients are equal to zero can give different conclusions from a sequence of individual hypotheses that each are equal to zero. These are different hypotheses. If $\beta_2 = \beta_3 = \beta_4 = 0$, then nothing has been explained:

$\widehat{u}_t = Y_t - \widehat{\beta}_1$, where $\widehat{\beta}_1$is just the mean and $R^2 = 0$. The $RRSS = \sum(Y_t - \overline{Y})^2$ in this case and this F statistic can be related to R^2 by

$$F(k-1, T-k) = \frac{R^2/(k-1)}{(1-R^2)/T-k)}.$$

If there is a single hypothesis, $m = 1$, the F and t tests are related by the fact that: $F(1, T-k) = t^2(T-k)$. This follows from the definitions in 8.3.

16.4. Tests for structural stability

Another common hypothesis tested with F statistics is structural stability between two periods, 1 and 2, $H_0 : \beta_{1i} = \beta_{2i};, i = 1, 2, ..., k$. In the unrestricted model we allow the k parameters to change and estimate the k parameters: $\widehat{\beta}_{i1}$ over the period one $t = 1, 2, ..., T_1$, with $RSS_1 = \sum \widehat{u}_{1t}^2$; and $\widehat{\beta}_2$ over period two $t = T_1 + 1, T_1 + 2, ..., T_1 + T_2$ with $RSS_2 = \sum \widehat{u}_{2t}^2$ so $URSS = RSS_1 + RSS_2$.In the restricted model the parameters do not change and we estimate $\widehat{\beta}_0$ over the whole period $t = 1, 2, ..., T$ with $T = T_1 + T_2$ and $RRSS = \sum \widehat{u}_{0t}^2$. Then the Chow, breakpoint test for parameter constancy $H_0 : \beta_1 = \beta_2$ conditional on the variances in the two periods being equal is

$$\frac{(\sum \widehat{u}_{0t}^2 - (\sum \widehat{u}_{1t}^2 + \sum \widehat{u}_{2t}^2))/k}{(\sum \widehat{u}_{1t}^2 + \sum \widehat{u}_{2t}^2)/(T-2k)} \sim F(k, T-2k).$$

Note the restricted model estimates k parameters $\widehat{\beta}_{0i}$, the unrestricted $2k : \widehat{\beta}_{1i}$ and $\widehat{\beta}_{2i}$, so $m = k$ restrictions.

The Chow predictive failure test sets $\widehat{u}_2'\widehat{u}_2 = 0$ and is

$$\frac{(\sum \widehat{u}_{0t}^2 - \sum \widehat{u}_{1t}^2)/T_2}{\sum \widehat{u}_{1t}^2/(T_1 - k)} \sim F(T_2, T_1 - k).$$

This tests whether the period 1 model predicts period 2.

16.5. Diagnostic Tests

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 diagnostic tests, each with the same null hypothesis, the model is well specified, but with different alternatives: the specific form of misspecification. Just as doctors use lots of different diagnostic tests since there are many different ways a person can be sick, there a many different ways a regression can be sick. Often the best diagnostic indicator to graph the residuals and look for patterns.

16.5.1. Tests for serial correlation and heteroskedasticity

The Durbin-Watson test for serial correlation 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. 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 by an autoregressive, AR1, model

$$u_t = \rho u_{t-1} + \varepsilon_t$$

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 and AR4 model with fourth order (four lags) serial correlation. Suppose you ran a regression

$$y_t = \alpha + \beta x_t + u_t$$

The test involves running a regression of the residuals on the variables included in the original regression and the lagged residuals

$$\widehat{u}_t = a + bx_t + \rho_1 \widehat{u}_{t-1} + \rho_2 \widehat{u}_{t-2} + \rho_3 \widehat{u}_{t-3} + \rho_4 \widehat{u}_{t-4} + \varepsilon_t$$

then testing the joint hypothesis $H_0 : \rho_1 = \rho_2 = \rho_3 = \rho_4$. You reject the null, no serial correlation up to fourth order if the test statistic (Chi-squared or F) is above the critical value for the required size, say 5%, or the p value is below 0.05.

There are many diagnostic tests which involve regressing the estimated residuals or powers of the residuals on particular variables. One test for heteroskedasticity regresses the squared residuals on the regressors:

$$\widehat{u}_t^2 = a + bx_t + v_t$$

to check whether the variance of the residuals is correlatd with the regressors. The null hypothesis is $b = 0$. Another test for autoregressive conditional heteroskedasticity (ARCH) regresses the squared residuals on the lagged squared residuals;

$$\widehat{u}_t^2 = a + b\widehat{u}_{t-1}^2 + v_t$$

and again tests $b = 0$.

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, 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.

17. Regression in practice

In practice, you will estimate regressions using a computer. Below we set out the sort of things that computer programs tell you and summarise how to estimate models. There will be classes to gain practice in estimating models.

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 use that. EViews is an easy package to use and is installed on our machines. gretl is a free open-source program similar to EViews. Microfit is another free econometrics package that is easy to use. EViews, Gretl and Microfit are menu driven programs, other programs like Stata are command driven (which ensures an audit trail of what you have done). R in a popular open source statistics program. For your projects you can use any program you wish.

There are applied exercises using the Shiller16 data for EViews, Stata and gretl, which provide more detail on using the programs. There is more background in

the EViews version, so even if you are not using EViews read the explanation and carry out the exercise on the software you are using.

The Shiller16 file is on Moodle. It has data 1871-2016 on NSP nominal stock prices, ND, nominal dividends, NE, nominal earnings, R, short interest rates, RL, long interest rates, CPI, consumer price index, RR real interest rates, RC real consumption Note that there is no data for some eyars on some. The example below regresses dividends on earnings for the period 1871-1986.

17.1. 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 (very loosely) gives the probability that this data support the null hypothesis. 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, which measures, roughly the average size of the error

  $$s = \sqrt{\sum \widehat{u}_t^2 / (T - k)}$$

  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 (11.5) and taking square roots.

- R squared, which tells you the proportion of the variation in the dependent variable that the equation explains,

$$R^2 = 1 - \frac{\sum \widehat{u}_t^2}{\sum (Y_t - \overline{Y})^2}$$

you can show that in a bivariate regression this is the square of the correlation coefficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}.$$

$S_{xy} = \sum (X_t - \overline{X})(Y_t - \overline{Y})/T$, is the covarianceand the variance is $S_{xx} = \sum (X_t - \overline{X})^2/T$,

- R bar squared, which corrects the numerator and denominator of R squared for degrees of freedom.

$$\overline{R}^2 = 1 - \frac{\sum \widehat{u}_t^2/(T-k)}{\sum (Y_t - \overline{Y})^2/(T-1)}$$

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.

17.2. Excel

Go into Excel, Load the Shiller16.xls file and find the Data Analysis module, you may have to add it in. Choose Regression from the list of techniques. Where it asks you Y range enter C2:C145. Where it asks you X range enter D2:D145.. Click in the residuals box. Click OK and you get output below, with the residuals listed below that, which you can graph.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.952245							
R Square	0.906771							
Adjusted R Square	0.906115							
Standard Error	2.444335							
Observations	144							
ANOVA								
	df	*SS*	*MS*	*F*	*gnificance F*			
Regression	1	8251.99	8251.99	1381.139	4.83E-75			
Residual	142	848.4176	5.974772					
Total	143	9100.408						
	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	0.839403	0.230853	3.636098	0.000386	0.383051	1.295755	0.383051	1.295755
X Variable 1	0.354001	0.009525	37.16368	4.83E-75	0.335171	0.372832	0.335171	0.372832

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 (rounded) results for the sample 1871 2014 are:

$$\begin{array}{rccc} ND_t = & 0.84+ & 0.35NE_t & +\widehat{u}_t \\ & (0.23) & (0.009) & \\ & [3.63] & [37.16] & \\ & \{0.000\} & \{0.000\} & \end{array}$$

$R^2 = 0.907$, $s = 2.44$.

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 \$3.54. If earnings were zero, dividends would be 84 cents. Earnings explain 90% of the variation in dividends over this period and the average error in predicting dividends is 2.44. 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

$$ND_t = \alpha + \beta NE_t + \gamma R_t + u_t$$

by giving the X range as D2:E130.

17.3. EViews

There is more explanation in the applied exercise. There is an example of EViews output in 11.1.2.

17.3.1. 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 2016 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, Import from file, then click on the Shiller file. It will guide you through loading the data.

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.**

That graph was not very revealing dominated by trends. Create the payout ratio. Click on Quick, generate series, and type into box PO=ND/NE. you will see it has been added to the variable list. Click on it and choose graph from view in the top left corner of the box that appears.

Use Save As command to save the Workfile under a new name and keep saving it when you add new data, transformations etc.

17.3.2. Estimating a Regression

Click on Quick, Estimate Equation and you will get a box. Enter ND C NE; OK and you will get a box with equation estimates, the same as for Excel above, but with slightly different extra information. Notice that you have menu buttons both on the equation box and the main window. 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.

18. Econometric Relationships

18.1. Logarithmic models

Consider a logarithmic demand function like

$$\ln Q_t = \beta_1 + \beta_2 \ln I_t + \beta_3 \ln P_t + \beta_4 \ln P_t^* + u_t \tag{18.1}$$

where Q_t is quantity demanded, I_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. The hypothesis $H_0 : \beta_3 = -\beta_4$ (homogeneity of degree zero) means only relative prices matter.

Notice the original model from which (18.1) is derived is non-linear

$$Q_t = BI_t^{\beta_2} P_t^{\beta_3} P_t^{*\beta_4} \exp(u_t)$$

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, CRS, if both inputs go up by 10%, output goes up by 10%. We can test this by rewriting (reparameterising) the equation as

$$\ln Q_t - \ln L_t = \ln A + b \left[\ln K_t - \ln L_t\right] + (b + c - 1) \ln L_t + dt + u_t \tag{18.3}$$

and do a t test on the coefficient of $\ln L_t$, which should be not significantly different from zero if there is CRS. Notice (18.2) and (18.3) are identical statistical equations, e.g. the estimates of the residuals would be identical. The restricted version, which imposes CRS so reduces the number of parameters estimated by one is

$$\ln Q_t - \ln L_t = \ln A' + b' \left[\ln K_t - \ln L_t\right] + d't + u'_t \tag{18.4}$$

where the primes indicate that the estimates will be different. Another way to test CRS is to use an $F(1, T - 4)$ test, to compare the fit of (18.4) with (18.3) .

18.2. 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

$$C_t = \delta_1 D_t + \delta_2(1 - D_t) + \beta Y_t + u_t \tag{18.5}$$

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 (18.5) the computer would have refused to estimate it and told you that the data matrix $(X'X)$ 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. We could also drop one dummy and include a constant.

18.3. Powers

We can easily allow for non-linearities by transformations of the data as we saw with logarithms above. As another example imagine Y_i (say earnings for individual i) first rose with X_i (say their age) and Z_i, (say education) then fell. We could model this by

$$Y_i = a + b_1 X_i + b_2 X_i^2 + c_1 Z_i + c_2 Z_i^2 + d Z_i X_i + u_i \tag{18.6}$$

where we would expect $b_1, c_1 > 0$, $b_2, c_2 < 0$. Although the relationship is non-linear, the model is linear in parameters, so ordinary least squares can be used, we just include other variables which are the squares of the original variables and an interaction term which is the product of the variables. Notice that the effect of X on Y is given by

$$\frac{\partial Y}{\partial X} = b_1 + 2b_2 X_i + dZ_i \tag{18.7}$$

thus is different at different values of X_i and Z_i, and has a maximum (or minimum) X_i^* which can be calculated as the value of X_i that makes the first derivative zero:

$$X_i^* = -(b_1 + dZ_i)/2b_2$$

On average earnings rise with age to a maximum then fall, and X^* is higher for more educated people. Verbeek section 3.5 has an extensive discussion. Self reported happiness tends to fall with age to a minimum then rise: the middle-aged are wealthy but miserable.

18.4. Proportions

Suppose our dependent variable is a proportion, $p_t = N_t/K$, where N_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 that the predicted value lies between zero and one. If the proportion is a function of time this gives,

$$\ln\left(\frac{p_t}{1-p_t}\right) = a + bt + u_t \tag{18.8}$$

which is an S shaped curve for p_t over time. This often gives a good description of the spread of a disease, e.g. the proportion of the population that have a mobile phone. 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

$$p_t = \frac{N_t}{K} = \frac{1}{1 + \exp(-(a + bt))} \tag{18.9}$$

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

$$N = C(1)/(1 + \exp(C(2) + C(3) * @trend)). \tag{18.10}$$

@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 there is data very close to saturation it is difficult to estimate K precisely. Notice that (18.8) and (18.10) imply different assumptions about how the error term enters (18.9) so are not equivalent.

18.5. Interpreting regression coefficients with transformed data

We distinguish (1) equations which are non-linear in variables because of transformations, like (18.1) or (18.6), but which can be estimated by a linear regression on the transformed data and (2) equations which are non-linear in parameters, like (18.10) above, where we need a non-linear estimation routine. We will consider models that are linear in parameters.

Typically we interpet regression coefficients as derivatives or elasticities. Often the derivative or elasticity is not constant, but a function of the variables, as in (18.7).above. We consider some other cases below.

18.5.1. Linear in variables

In the standard linear regression bteween continuous untransformed variables

$$Y_t = \alpha + \beta X_t + u_t,$$

β measures the change in Y_t that result from a one unit change in $X_t : \Delta X_t = 1$. It corresponds to the derivative

$$\frac{\partial Y_t}{\partial X_t} = \beta.$$

β depends on the units that the variables are measured in. Suppose, X_t is per-capita GDP measured in dollars and Y_t is life expectancy in years, then β is the number of extra years of life bought by an extra dollar. The standard error of regression measures the size of a typical error and is in the same units as the dependent variable, here years.

The elasticity is the percentage (proportionate) change in Y_t that results from a one percent change in X_t. It is invariant to units, but does depend on where we measure it.

$$\eta = \frac{\partial Y_t / Y_t}{\partial X_t / X_t} = \frac{\partial \log Y_t}{\partial \log X_t} = \frac{\beta X_t}{Y_t}$$

For a linear relationship, the elasticity is different at every point on the line. A convenient place to measure it is at the typical values, the means of X_t and Y_t.

18.5.2. Log-log regression

In a logarithmic regression

$$\log Y_t = \alpha + \beta \log X_t + u_t$$

then β is the elasticity. The standard error of regression measures a typical proportional error (multiply by 100 to get percentage error). To provide a rough comparison with the fit of a linear model, divide the standard error of the linear model by the mean of the dependent variable (assuming the mean is positive and non-zero) which will also give a proportionate error.

Suppose we have a dummy variable in the equation

$$\log Y_t = \alpha + \beta \log X_t + \gamma D_t + u_t$$

where $D_t = 0$ or $D_t = 1$. The effect on Y_t of the dummy variable going from zero to one is $\exp(\beta) - 1$.

18.5.3. Percentages

Suppose Y_t and X_t are both percentages. For instance, Y_t is the inflation rate, measured in percent, and X is the unemployment rate, also measured in percent. Then β measures the percentage **point** change in Y_t in response to a one percentage **point** change in X_t. If unemployment rises from 1% to 2%, it increases by one percentage point and 100%. The Phillips Curve relationship between inflation and unemployment may be non-linear: the effect on inflation of a one percentage point change in unemployment is much greater when unemployment is 1% than when it is 9%. We can represent this by using the reciprocal of unemployment

$$\begin{aligned} Y_t &= \alpha + \beta X_t^{-1} + u_t, \\ \frac{\partial Y_t}{\partial X_t} &= -\beta X_t^{-2}. \end{aligned}$$

So at 1% unemployment the effect is just $-\beta$, but at 9% unemployment it is $-\beta/9^2$, very small.

18.5.4. Semi log

In the life expectancy example we found that a better fit was given by using log per-capita income, i.e. an equation, sometimes called linear-log, of the form

$$Y_t = \alpha + \beta \log X_t + u_t$$

If X_t changes by 1%, then $\Delta X_t/X_t = 0.01$, so Y_t changes by 0.01β.

One can have it the other way round a log-linear model

$$\log Y_t = \alpha + \beta X_t + u_t.$$

Here a unit change in X_t, $\Delta X_t = 1$, causes a $100\beta\%$ change in Y_t.

19. Time-series and 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.

19.1. Autoregressions

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 1.2 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

$$y^* = \frac{\alpha_0}{1-\alpha_1-\alpha_2}$$

19.2. ARDL and ECM

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)

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + u_t \tag{19.1}$$

Again it can be estimated by least squares. There are often strong theoretical reasons for such forms. We can reparameterise this to an error correction model, ECM form,

$$\begin{aligned} y_t - y_{t-1} &= \alpha_0 + (\alpha_1 - 1)y_{t-1} + \beta_0 x_t - \beta_0 x_{t-1} + \beta_0 x_{t-1} + \beta_1 x_{t-1} + u_t \\ \Delta y_t &= \alpha_0 + (\alpha_1 - 1)y_{t-1} + \beta_0 \Delta x_t + (\beta_0 + \beta_1)x_{t-1} + u_t \end{aligned}$$

$$\Delta y_t = a_0 + a_1 y_{t-1} + b_0 \Delta x_t + b_1 x_{t-1} + u_t \tag{19.2}$$

These dynamic forms can be interpreted in terms of a long-run equilibrium relationship,

$$y_t^* = \theta_0 + \theta_1 x_t \tag{19.3}$$

and slow adjustment. The simplest form of adjustment is the partial adjustment model (PAM)

$$\Delta y_t = \lambda(y_t^* - y_{t-1}) + u_t. \tag{19.4}$$

$\Delta y_t = y_t - y_{t-1}$, is the change. For instance, people change their consumption, y_t, to remove part of the difference between the equilibrium consumption and consumption in the previous period. The coefficient λ is an adjustment coefficient, or speed of adjustment, it measures the proportion of the deviation from the target or equilibrium level of consumption (which iss a function of income, x_t), which is made up in a period. We would expect $0 < \lambda \leq 1$, with $\lambda = 1$ indicating instantaneous adjustment and $\lambda = 0$ no adjustment.

We can write the PAM as

$$\Delta y_t = \lambda\theta_0 + \lambda\theta_1 x_t - \lambda y_{t-1} + u_t$$

or

$$y_t = \lambda\theta_0 + \lambda\theta_1 x_t + (1 - \lambda)y_{t-1} + u_t$$

we would just run a regression of y_t on a constant, x_t and lagged y_t, i.e. y_{t-1}consumption,

$$y_t = \alpha_0 + \beta x_t + \alpha_1 y_{t-1} + u_t$$

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 $\widehat{\theta}_1 = \widehat{\beta}/\widehat{\lambda} = \widehat{\beta}/(1 - \widehat{\alpha}_1)$. This is a dynamic equation it includes lagged values of the dependent variable. Whether we estimate it using the first difference Δy_t or level y_t as the dependent variable does not matter, we would get identical estimates of the intercept and the coefficient of income, x_t. 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.

The error correction model (ECM) keeps the long-run relationship given by (19.3) , but unlike the partial adjutment model (19.4) assumes agents respond to both the change in the target and the lagged error

$$\begin{aligned}
\Delta y_t &= \lambda_1 \Delta y_t^* + \lambda_2 (y_{t-1}^* - y_{t-1}) + u_t \\
\Delta y_t &= \lambda_1 \theta_1 \Delta x_t + \lambda_2 (\theta_0 + \theta_1 x_{t-1} - y_{t-1}) + u_t \\
\Delta y_t &= a_0 + b_0 \Delta x_t + b_1 x_{t-1} + a_1 y_{t-1} + u_t.
\end{aligned}$$

This is the same as (19.2) above and the estimated parameters 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, $\widehat{\theta}_1 = -\widehat{b}_1/\widehat{a}_1$. Notice that λ_2 is also a speed of adjustment, it measures the proportion of the deviation from target that is made up in a period.

If the model is stable, $-1 < a_1 < 1$, the equilibrium solution for (19.1) is got by setting $y_t = y_{t-1} = y$; $x_t = x_{t-1} = x$ so that in the long-run

$$\begin{aligned}
y &= \alpha_0 + \alpha_1 y + \beta_0 x + \beta_1 x \\
y &= \frac{\alpha_0}{1 - \alpha_1} + \frac{\beta_0 + \beta_1}{1 - \alpha_1} x \\
y^* &= \theta_0 + \theta_1 x
\end{aligned}$$

Economic theory usually makes predictions about long-run rather than the short-run relations. Notice our estimate of θ_1 will be identical whether we get it from the ECM (19.2) or ARDL (19.1) or from estimating a non-linear version.

A common restriction on dynamic models in log is that the long-run coefficient is unity. This can be tested by estimating

$$\Delta y_t = a_0 + b_0 \Delta x_t + b_1(x_{t-1} - y_{t-1}) + (a_1 + b_1) y_{t-1} + u_t.$$

If $\widehat{\theta}_1 = -b_1/a_1 = 1$; then $(a_1 + b_1) = 0$. This "reparameterisation" is not unique, there are other ways that we could write the equation to get a coefficient equal to $(a_1 + b_1)$.

19.3. Vector autoregressions

Suppose that we have two variables y_{1t} and y_{2t} which interact, but neither are exogenous, as we assumed x_t above was. We can let them each depend on the lags of themselves and the other variable:

$$\begin{aligned} y_{1t} &= a_{10} + a_{11} y_{1,t-1} + a_{12} y_{2,t-1} + u_{1t}, \\ y_{2t} &= a_{20} + a_{21} y_{1,t-1} + a_{22} y_{2,t-1} + u_{2t}. \end{aligned}$$

If past y_{1t} predicts current y_{2t}, i.e. $a_{21} \neq 0$, then y_{1t} is said to be "Granger-Causal" for y_{2t}. Granger causality can go in bothe directions and does not correspond to the usual economic definition of causality. Weather forecasts are Granger causal for the weather. Below log earnings, LE seem granger causal for log dividends, LD, but at the 5% level LD does not seem Granger causal for LE.

Vector Autoregression Estimates
Date: 10/06/16 Time: 15:28
Sample (adjusted): 1872 2014
Included observations: 143 after adjustments
Standard errors in ( ) & t-statistics in [ ]

	LE	LD
LE(-1)	0.697047	0.255720
	(0.08956)	(0.03245)
	[7.78291]	[7.87996]
LD(-1)	0.215712	0.672732
	(0.11252)	(0.04077)
	[1.91702]	[16.4996]
C	-0.132133	-0.220331
	(0.13875)	(0.05028)
	[-0.95229]	[-4.38243]
@TREND	0.005032	0.001968
	(0.00205)	(0.00074)
	[2.45017]	[2.64455]
R-squared	0.976061	0.995984
Adj. R-squared	0.975545	0.995898
Sum sq. resids	10.81936	1.420510
S.E. equation	0.278993	0.101092
F-statistic	1889.180	11492.09
Log likelihood	-18.33042	126.8376
Akaike AIC	0.312314	-1.718008
Schwarz SC	0.395190	-1.635131
Mean dependent	0.905255	0.364613
S.D. dependent	1.784054	1.578354
Determinant resid covariance (dof adj.)		0.000565

19.4. 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, including allowing for a t distribution for the errors rather than a normal distribution.

19.5. 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.