Econometric Principles and Data Analysis

Unit 1 An Introduction to Econometrics and Regression Analysis

Contents

Unit Overview

This unit provides an introduction to econometrics and regression analysis. It outlines the differences between financial and economic theory and econometrics. The unit explains how stochastic relations between variables are different to mathematical relations between variables. It explains how uncertainty may be modelled using a disturbance term. The unit introduces the steps involved in an econometric investigation.

Learning outcomes

After studying this unit, the readings, and the exercises, you will be able to discuss and apply the following:

- the population regression function
- the sample regression function
- the disturbance (or error) term
- the residual term
- how to read data from pre-existing text files
- how to create and interpret a scatter plot
- how to obtain summary statistics
- how to create transformations of variables.

🕮 Reading for Unit 1

Jeffrey M Wooldridge (2020) Chapter 1 'The nature of econometrics and economic data'. In: *Introductory Econometrics*. 7th Edition. Boston MA, Cengage. pp. 1–17.

1.1 What is Econometrics?

Welcome to this module. The aim of the module is to give you an introduction to econometric methods or, more specifically, to linear regression, which is the major statistical foundation of econometric work. This module requires that you work with data; we hope you will find this interesting and useful, and that you enjoy the module.

A principal concern of financial and economic theory is relations between variables. In finance, you may have already studied many of these including the capital asset pricing model; arbitrage pricing theory; efficient markets hypothesis; optimal hedging ratios; bid-ask spreads. If you have studied economics, you may be familiar with consumption, investment, and demand for money functions; labour supply and labour demand functions; the expectations-augmented Phillips curve; and many others. You could, in fact, view the whole of economic and finance theory as a set of relations among variables.

What is econometrics? Econometrics is concerned with quantifying financial and economic relations. Econometrics is of use in providing numerical estimates of the parameters involved and for testing hypotheses embodied in the theoretical relationships. Broadly defined, econometrics is:

> the application of statistical and mathematical methods to the analysis of economic data, with the purpose of giving empirical content to economic theories and verifying them or refuting them.
>
> Source: Maddala (1992) p. 1.

This definition is not the only possible one; in fact, in your key text you will come across a number of definitions, which each puts the emphasis slightly differently. Common to all definitions, however, is the stress on the empirical nature of econometric work: the subject matter of econometrics concerns the interaction of, and confrontation between, theory and data in quantifying economic and financial relationships.

Hence, econometrics is not purely a branch of mathematical economics or mathematical finance. Indeed, mathematical finance or economics need not have any empirical content at all. Econometrics makes use of mathematical methods, but its emphasis is on empirical analysis. However, econometrics is not just a ‘box of tools’ to work with data. It requires, undoubtedly, a good training in statistical techniques, but these techniques need to be situated in an interactive process between theory and the data.

To give empirical content to financial and economic theories and to verify them or refute them, the econometrician is confronted with three types of problems, which are of lesser or no concern to the theorist.

First, in economic or financial theory we develop models out of *a priori* reasoning based on relatively simple assumptions. To do this, we abstract from secondary complications by assuming that ‘other things remain equal’ while we investigate the relations between a few key economic or financial variables. In effect, this method reduces to ‘intellectual experimentation’ with causal relations postulated by theory. For example, in demand theory we say that the quantity demanded of a commodity (which is not an inferior good) will fall if its price rises, other things being equal.

This method is fruitful in theory but, unfortunately, in empirical economics and finance the scope for experimentation is severely limited. A researcher cannot alter a commodity’s price (or an asset’s price), holding other things constant, in order to see what happens to demand. In general, financial and economic data are not the outcome of experiments, but rather the product of observational programmes of data gathering and collection in a world where other things are never equal. In econometrics, therefore, we can only resort to careful observation; the basic art of econometric work is more like unravelling a complex puzzle than setting up an experiment in a laboratory.

Second, we need to address the difference between deterministic and stochastic relationships. This issue arises in a different way in economics and in finance. To make the point, we will explain the distinction between deterministic and stochastic relationships with an example from economics, and then address it from a financial perspective.

In most economic theory we work with *deterministic relationships* between economic variables. Take a simple example: the Keynesian consumption function. In economic theory we assume that if we know the level of aggregate real income, consumption will be uniquely determined. That is, for each value of aggregate real income there corresponds a given level of aggregate consumption. This is a convenient device to enable us to work out exact solutions for the interplay between variables within the confines of the assumptions of an economic model.

In reality, however, we do not expect this relationship to be exact: it may be stable perhaps, but it is surely imperfect. Hence, in econometric work we deal with imperfect relationships between variables. It follows that our models cannot be deterministic in nature. We investigate functions between variables that we believe to be reasonably stable, on average, but there will always be a degree of uncertainty about outcomes and conclusions derived from such a model. Econometric modelling requires that we make explicit assumptions about the character of these imperfections, or *disturbances* as they are more commonly labelled. That is, we work with stochastic variables and we need to model their stochastic nature. This is what makes us enter the areas of probability theory and statistical inference and estimation.

How does the distinction between deterministic and stochastic relationships arise in finance? Uncertainty is a fundamental element of risk, and the measurement and management of risk are central aspects of finance. To demonstrate this, consider the single-index model (which you will examine and estimate in Unit 2). In the single-index market model the return on a company stock is considered to be a function of three elements. There is a fixed element which is specific to the company. There is also a deterministic relationship between the return on the company stock and the return on a relevant market index: For each value of the return on the market index there corresponds a given value for the return on the company stock. (This part of the model captures the concept of market-determined risk.) In addition, the return on the company stock is explained by a company-specific disturbance or error. (The company-specific error captures the concept of company-specific risk.) The single-index model includes the company-specific disturbance not just to make the model more realistic; it is included because we specifically want to understand the stochastic nature of the return on the company stock, and thus get a better understanding of the risk associated with the stock.

Third, in financial and economic models we work with *theoretical* variables. Econometrics, in contrast, deals with *observed* data. Obviously, there is a certain correspondence between them; data collection is inspired by theoretical frameworks. For example, national income account data were constructed after the ascendancy of Keynesian economics, which concerns the analysis of theoretical aggregates such as output, demand, employment and the price level. However, observed variables do not fully correspond to their theoretical equivalents because of errors in measurement, conceptualisation and coverage. This is usually less of a problem for econometrics applied in a financial context than it is for economics. Financial data on asset prices, for example, is more closely related to the actual transactions taking place, so measurement error is less likely. However, we should be aware that movements in financial data may be the result of the particular operating or reporting features of a market, say, in addition to the desired trading activities of the participants that our theories suggest. In econometrics we need to be aware of the nature of the observed data and its implications for investigating theoretical propositions.

These three elements:

- the fact that we cannot hold other things constant in empirical analysis
- the imperfect nature of relationships between variables and
- the discrepancies between theoretical variables and observed data

give econometrics its distinctive flavour. We cannot move straight from a financial or an economic model (as formulated by theory) to the data before we come to terms with these issues. Econometric methods, therefore, aim to address these issues so as to enable us to engage in meaningful investigation of economic and financial theories.

Note that we talk about methods and, hence, emphasise the need for methodological groundwork to approach these types of problems. There are no hard and fast rules to deal with them. There is not a box of magic tricks, which always work and give us straight answers. Rather, we are left with the task of studying methodological approaches to issues, which are complex, varied, but challenging.

This module, *Econometric Principles and Data Analysis*, deals with regression analysis. Why this focus? We have seen that, in empirical analysis, our data never behave exactly as our theoretical models would lead us to believe. Theoretical models are useful abstractions, which provide the applied researcher with analytical handles to make sense of an often bewildering economic and financial reality. Good theory allows us to search for patterns within the data and to give meaning to such patterns. But we need to disentangle these patterns in the middle of a great deal of chance variations and uncertainties of outcomes, which our theories could not possibly aim to explain. Regression analysis provides us with an analytical framework to handle relations between variables, especially between variables whose relation is imperfect.

Indeed, regression analysis seeks to establish *statistical regularities* among observed variables. To do this, we need to come to terms with the uncertainty inherent in the behaviour of our data. For this, we need to equip ourselves with statistical theory which allows us to model uncertainty as part of relations between variables. This is the purpose of this module, *Econometric Principles and Data Analysis*, of which this is the first unit.

The following are the main points to remember.

- In econometrics we pose the question how to confront theory with data so as to quantify our financial and economic relationships, to verify them or to refute them.
- In practice, we deal with *imperfect* relationships between variables which we can only *observe* (with errors and, often, through proxies) in a context which we do *not control* (we cannot experiment).
- It follows that we can only resort to careful observation of complex phenomena in order to check our theories against the empirical evidence. This raises questions about econometric methods: methodological issues about gathering and evaluating such empirical evidence. Whatever conclusions we draw in such a context will always involve a considerable degree of uncertainty, even if our models are correctly specified. For this reason, we resort to probability theory and statistical inference to deal with uncertainty in assessing outcomes and conclusions of empirical analysis.
- Since our concern is primarily with investigating relations between variables, regression analysis constitutes the major tool of statistical analysis in econometrics.

1.2 How to Use the Module Units

It is quite possible that you are worried about studying econometrics. After all, it involves working with mathematics and statistics, and you may feel that this is not one of your greatest strengths. Alternatively, you may be one of those who welcome this greater emphasis on mathematics and statistics. Whichever view you hold, it is useful to be aware of a particular problem that invariably arises when studying econometrics.

Teaching and learning econometrics almost inevitably involves a preoccupation with technical details: definitions of technical terms, mathematical derivations, step by step descriptions of statistical procedures *etc*, all phrased in technical notation. This is normal and, indeed, necessary. But with this preoccupation on technical detail there is a possibility that students lose a perspective on 'What is it all about?' or 'Why are we doing this?' Therefore, there is a need to keep a focus on the kind of basic questions, uncluttered by notation and technical detail, which give substance to the subsequent technical exercises. We need to get an overview of a problem before we explore it aided by our technical skills. We need to know the simple questions and intuitive insights which often prompted elaborate technical enquiries.

For this reason, the *module units* will always start with a section on *ideas* or *issues*.

The purpose of this is to explain in simple words, with the minimum of technical notation, the basic substance of the unit. The aim is to give you an intuitive feel for the subject matter before going into technical detail. If you feel that mathematics and statistics are not your strongest subjects, this regular section will give you a few 'analytical handles' to hold on to when studying relevant techniques.

But, alternatively, if you are confident with mathematics and statistics, it is important *not* to skip this section. Technical expertise is not just a question of one's ability to work out the steps in a technical procedure or to understand a mathematical derivation. It also involves understanding the type of questions a technique tries to address as well as the assumptions on which it is based. Good technical expertise is more than understanding a set of technical skills (narrowly defined); it also involves analytical insights and judgement of the appropriateness of particular technical procedures in specific conditions.

The section on ideas or issues will be self-contained; no references will be made to reading parts of the assigned key text. Take your time to read it carefully, and to reflect whether you understand the type of questions which will be addressed subsequently in technical detail: 'get familiar with the forest before you start looking at the trees'. In other words, use this section to provide you with the 'analytical handles' to facilitate the study of the relevant techniques.

Next, the module units will have a reading section, or *Study Guide*, which guides your study of the key text, *Introductory Econometrics* by Jeffrey M Wooldridge. The purpose of these sections is to structure your reading of the key text as well as to provide brief comments, elaborations and cross-references to exercises and examples, and to suggest short cuts in coping with the material.

The section after that will normally contain one example. This section has two purposes. Firstly, the example highlights a specific aspect of the topic under study in a particular unit of the module. Secondly, the example also tries to give you a bit of the flavour of econometrics in action. Generally, you will be asked to participate in the analysis of the example. The examples aim to highlight the links between economic theory and empirical investigation, and try to illustrate the problems that can arise when we work with real data.

The next section will provide a brief summary of the main issues raised in the unit. This will be followed by a section of *exercises*. It is most important that you work through all of these exercises. The exercises have three purposes:

- to check your understanding of basic concepts and ideas
- to verify your ability to use technical procedures in practice and
- to develop your skills in interpreting the results of empirical analysis.

The final section of the units will include brief answers to these exercises, which you should not look at until after you've worked out the answers for yourself!

You will be using R to do the econometric exercises. Instructions to use R will accompany the exercises.

This basic structure of the module units will be maintained throughout your study of this module. The section on *ideas* or *issues* gives you an overview of the topic of the unit, using non-technical language. The core of the unit is the *study guide*. This guides you through your reading of the key text and refers you to the exercises whenever appropriate. The *example* in each unit demonstrates a problem dealt with in the module material using real data. By using examples drawn from areas of finance, using real data, this section also aims to provide cross-references to the theory modules.

The *summary* draws your attention to the main points made in the unit. The *exercises* are important and you should always work through them. The exercises will help you to understand the module material. In addition, the knowledge and experience you gain from doing the exercises will help you to write assignments and answer examination questions.

1.3 Ideas – The Concept of Regression

The remainder of this unit will deal with the introduction to regression analysis. As you will see, it is structured along the pattern outlined above.

1.3.1 What is regression?

Regression is the main statistical tool of econometrics. What is regression? Broadly speaking,

> regression methods bring out relations between variables, especially between variables whose relations are imperfect in that we do not have one Y for each X.
>
> Source: Mosteller & Tukey (1977) p. 262.

But what do we mean by imperfect relations?

An example may help. Consider the relation between corporate bond spreads (this is the Y-variable) and the earnings before interest of companies (this is the X-variable). The spread for a corporate bond is the difference between the interest rate on the corporate bond and the interest rate on government bonds of equivalent maturity. Interest rates on corporate bonds are higher than those on government bonds to reflect expected default loss, different tax treatments and the riskier return associated with corporate bonds. We would expect that a company with higher earnings before interest would be less likely to default, and hence the bond spread for that company would be lower.

Hence, we expect that, on average, the corporate bond spread is inversely related to earnings before interest. But we do not expect this relation to be perfect. That is, if we were to sample 10 companies with identical earnings before interest (ie equal X-values), we would not expect to get 10 identical corporate spreads (the Y-values). Differences between the markets in which the firms operate, in management and in other financial variables (eg coupon rates, coverage ratios) will account for differences in bond spreads. But, importantly, it is still valid to say that, on average, the bond spread declines as the level of earnings before interest increases. That is what Mosteller and Tukey (quoted above), mean when they say that a relation exists between two variables but that it is imperfect in that we do not have one Y for each X.

This leads us to the discussion of the concept of regression. Regression methods aim to bring out this *average* relation between a dependent variable on the one hand and one or more independent variables on the other. In our example the average inverse relation between the bond spread and the level of earnings before interest is the *regression* of the former variable on the latter. But, obviously, there will be variation in how markets view the bonds of individual companies that have broadly the same earnings.

In fact, anyone familiar with data analysis knows very well that we can always take an average of one or another aspect of a number of individuals, but we rarely meet the 'average individual'. So, it is also with regression as an average relation: individual observations will rarely conform to the average relationship between Y and X. Hence, in regression analysis we seek to establish statistical regularities in the middle of a great deal of chance variation and uncertainty in outcomes. For this reason, regression methods involve statistical *modelling* of the chance variation in the data as well as of the average relationship.

In summary, we hope that our model captures the basic structure of interaction between economic and financial variables, and we expect that the behavioural relations are reasonably stable, but imperfect. At most, we expect these relations to hold 'on average'. In other words, we seek to discover structure and regularity within data in the middle of a great deal of uncertainty in outcomes. It is similar to separating *sound* from *noise* when trying to listen to a badly tuned radio.

Therefore, a regression model embraces two components:

- a regression line (which defines the basic structure) and
- disturbances.

Firstly, the regression line models the *average* relation between the dependent variable and its explanatory variable(s). To do this we make an explicit assumption about the shape of the regression curve: linear, quadratic, exponential, *etc*.

Secondly, we recognise the existence of chance fluctuations due to a multitude of factors beyond our control. We model this element of uncertainty (the noise) in the form of a disturbance term, which constitutes an integral part of our model. This disturbance term is a 'catch all for all the variables considered as irrelevant for the purpose of the model as well as all unforeseen events' (Maddala, 1992: p. 3). It is a random variable which we cannot observe or measure in practice.

Sometimes we are not interested in the disturbance term as a variable in its own right, but we are interested in understanding how the disturbance term affects our attempts to investigate the behavioural relations in the model. In other circumstances we might be particularly interested in the properties of the disturbance term, if it reflects an element of uncertainty and risk that we are trying to understand.

In both cases, we need to model the probabilistic nature of the disturbance term. In other words, we try to model the character of the uncertainty inherent in the data. This is no easy task, and we always need to think carefully whether the assumptions we make about the nature of these chance variations are indeed appropriate for the type of issue under study. Not surprisingly, a great deal of econometric theory and practice is concerned with these assumptions.

It is useful to express these important ideas a little more formally. We start with the population regression function. This is a theoretical construct, which contains a hypothesis about how the data are generated. For the simple, two-variable linear regression model we have

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{1.1}$$

in which Y is the dependent variable, X is the explanatory variable – sometimes called the regressor, u is the disturbance term, and the subscript i indicates the ith observation. β_0 and β_1 are the regression parameters; β_0 is the intercept, or constant, and β_1 is the slope coefficient. Typically, the variables Y and X are observable, the disturbance is not observable, and the parameters β_0 and β_1 are unknown. The presence of the random disturbance means that Y is stochastic; for each value of the explanatory variable, X, there is a distribution of Y-values.

In this explanation of regression we will continue to use the i subscript to indicate the ith observation. In many financial applications we will examine series that vary over time, and it will be more meaningful to use a t subscript to indicate that the observation refers to period t. This will allow us to use $t - 1$ to refer to the previous period, *etc.*

The population regression function may be viewed as comprising two components: a systematic element represented by a straight line which shows the statistical dependence of Y on X; and a random, or stochastic, element represented by the disturbance (error) term u. The systematic element can be expressed as

$$E\left[Y_i \mid X_i\right] = \beta_0 + \beta_1 X_i \tag{1.2}$$

that is, the average (or expected), value of Y conditional on a given value of X is a linear function of X – or, more concisely, the average value of Y for each value of X. That is, the population regression function joins the conditional means of Y. The disturbance term, u, is the focus of much attention. It accounts for the variation in Y around the population regression line. In Unit 2 you will learn about the important assumptions made about u.

A prime objective of econometrics is to quantify the unknown parameters β_0 and β_1. Using a sample of data on Y and X, we obtain estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, of the unknown population parameters.[1]

We have the sample regression function

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i \tag{1.3}$$

in which $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables (the particular estimates obtained depend on the particular sample of data on Y and X used) that differ from the population parameters β_0 and β_1. Consequently, the sample residuals, e_i, differ from the unknown population disturbances, u_i. Whereas the *disturbance term* accounts for the variation in Y around the *population regression line*, the *residuals* give us the vertical deviations of the observed Y-values from the *estimated regression line* derived from sample data. The residuals, therefore, are not identical with the disturbances, but clearly they do tell a story, which may enable us to assess whether or not our assumptions about the behaviour of the disturbances seem reasonable. How to analyse the story or stories told by residuals is a matter we address in the second half of the module.

The predicted value of the dependent variable is given by the sample regression line

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{1.4}$$

in which $\hat{Y}_i$ is the fitted value of the dependent variable, the estimator of $E\left[Y_i \mid X_i\right]$, that is the estimator of the population conditional mean. The sample regression line is an estimator of the population regression line.

Notice that we focus on the *linear* regression model. That is, we are concerned with a model that is linear in the parameters to be estimated. The model

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{1.5}$$

is linear in β_0 and β_1. With the sample regression line

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{1.6}$$

$\hat{\beta}_0$ is the predicted value of Y (in units of Y) if $X = 0$. Also

$$\hat{\beta}_1 = \frac{d\hat{Y}}{dX}.$$

This implies that a 1 unit increase in X (measured in units of X) results in a $\hat{\beta}_1$ unit increase in $\hat{Y}$ (measured in units of Y).

Now consider the model (in which e stands for exponential, not the residual)

$$Y_i = \alpha X_i^{\beta_1} e^{u_i} \tag{1.7}$$

[1] ^ is read as 'hat', hence $\hat{\beta}_1$ is 'beta one hat'.

which, after taking natural logarithms of both sides of the relation, can be written as

$$\ln Y_i = \ln \alpha + \beta_1 \ln X_i + u_i$$

or

$$\ln Y_i = \beta_0 + \beta_1 \ln X_i + u_i \tag{1.8}$$

where $\beta_0 = \ln \alpha$.

This model is also linear in the parameters to be estimated, β_0 and β_1. We may view the model as

$$Y_i^* = \beta_0 + \beta_1 X_i^* + u_i \tag{1.9}$$

where $Y_i^* = \ln Y_i$ and $X_i^* = \ln X_i$. This model is known by a number of different names – logarithmic, double log, log-log, log linear, and constant elasticity – and is frequently used in applied work when it characterises the form of the functional relationship between the variables. It has the useful property that the slope coefficient measures the *elasticity* of Y with respect to X because

$$\beta_1 = \frac{d \ln Y_i}{d \ln X_i} = \frac{dY_i}{Y} \Bigg/ \frac{dX_i}{X_i} \tag{1.10}$$

With this logarithmic model, a 1 *per cent* increase in X results in a β_1% increase in Y. Note that here we mean a 1% proportionate increase in X, not that X increases by 100 basis points (1 basis point equals 0.01%).

Although regression analysis is related to correlation analysis, conceptually these two types of analysis are very different. The main aim of correlation analysis is to measure the degree of linear association between two variables, and this is summarised by a sample statistic, the correlation coefficient. The two variables are treated symmetrically. Both are considered random; there is no distinction between dependent and explanatory variables, and no implication of causality in a particular direction from one variable to the other.

Regression analysis, however, can incorporate relationships between two or more variables and the variables are not treated symmetrically. The dependent and explanatory variables are carefully distinguished. The former is random and the latter is often assumed to take the same values in different samples – often referred to as 'fixed in repeated samples'. The underlying economic or financial theory implies that X, an explanatory variable, causes Y, the dependent variable. Moreover, with more than one explanatory variable, regression analysis quantifies the influence of each explanatory variable on the dependent variable.

1.3.2 Data and regression

Regression methods allow us to investigate associations between variables, but the inspiration as to which relations to investigate obviously comes from theory. We are not interested in detecting spurious (false or bogus) associations between variables. Indeed, relations have to be meaningful – and whether they are, or not, depends on theoretical argument.

This does not mean, however, that data play only a passive role in economic and financial analysis. The role of data is not just to provide numerical support to theoretical arguments. Empirical investigation is an active part of theoretical analysis in as much as it is concerned with testing theoretical hypotheses against the data as well as, in many instances, providing clues and hints towards new avenues of theoretical enquiry. This requires that we translate our theoretical insights into empirically testable hypotheses, which we can investigate with observed data. Hence, the process between theory and the data is interactive: we must continuously investigate the empirical content of our theoretical propositions in order to test our theories, and pick up signals from the data that enable us to improve our theoretical insights.

Most of the data we use in applied economic analysis are not obtained through experimentation but are the result of observational programmes. National income accounts, agricultural and industrial surveys, financial accounts, employment surveys, population census data, household budget surveys and price and income data, among others, are collected by various statistical offices. They are partial records of what happens; they are not the outcome of experiments. As we have noted, finance data more closely relate to actual transactions, but, like economic data, they are not the outcome of experiment.

The character of this economic and financial data makes the work of an econometrician quite different from that of a psychologist or an agricultural scientist. In the latter cases, experiments play a prominent role in analysis, and much of the emphasis in research work is put on the careful design of experiments in order to be able to isolate effect and response between two variables while controlling for the influence of other variables (that is, by holding them constant). In economics and finance, the scope for experimentation is very limited.

We cannot change the price of a stock, holding all other prices constant, merely to see what would happen in its demand. In theory, we do just that by assuming that 'other things are equal' and postulating cause and effect between the remaining variables. In empirical analysis, however, other things are never equal, and we can only carefully observe the behaviour of economic agents from survey data. As you will see in subsequent units, multiple regression techniques allow us to 'account' for the influence of other variables while investigating the interaction between two key variables, but this is not the same as 'holding other variables constant'.

The econometrician, therefore, needs to be, above all, a careful observer. Empirical analysis in economics and in finance allows us to search for patterns in our data through careful observation backed by theoretical understanding; but experimentation is not really an option we have available, because we do not have control over the overall context that determines the movement of our variables.

In analysing data, we should follow the advice ascribed to Darwin. It is obviously pleasing if the empirical evidence seems to support our theoretical hypotheses, but – more importantly – we should take special note of any signs given by the data that go against our arguments. That is, we should not approach our data merely to confirm answers to well-defined questions derived from theoretical argument, but we should also look out for hints from the data about what we do not know – that is, about questions that we have not confronted yet. A careful observer uses data not just to confirm their theories, but also to get clues from empirical analysis to advance one's theoretical grasp of a problem. It is primarily this aspect that enables data to be used to play an active part in the process of analysis.

1.3.3 Rates of return

Much analysis in financial econometrics is concerned with rates of return, including returns on shares, stock indices, commodities and exchange rates. Therefore, at this point in Unit 1 it might be useful, briefly, to refresh your understanding of returns. In your study of finance or risk management, or in your work, you may already be familiar with arithmetic and logarithmic rates of return. For example, logarithmic returns are used especially in the Black–Scholes–Merton model of options pricing.

First consider arithmetic returns. Suppose we have a stock that is worth $1,000 at the start of the year and $1,050 at the end of the year. Ignoring any dividends, we say that the arithmetic or simple or proportionate rate of return is

$$r = \frac{\left(1,050 - 1,000\right)}{1,000} = 0.05 \text{ or } 5\%.$$

It is the increase (or decrease) in value, divided by the original value.

Put another way, if the stock, initially valued at $1,000, benefits from a 5% return over the year, then the value at the end of the year will be

$$1,000\left(1 + 0.05\right) = 1,050$$

In general terms, if the price at the start of the year is P_0, and the stock experiences a return of *r*, the price at the end of the year will be

$$P_1 = P_0\left(1 + r\right) \tag{1.11}$$

and the rate of return is

$$r = \frac{\left(P_1 - P_0\right)}{P_0}. \tag{1.12}$$

To understand logarithmic returns and continuous compounding, it may help to conduct a short thought exercise. In the previous example, we can think of the return, *r*, being applied to the asset once a year (if it makes more sense to you, think of *r* as the interest paid on a sum of money in a bank account, paid annually). Now suppose that this growth rate is applied at more times through the year, but the rate of return at each point of the year is adjusted to take account of the increased number of times the return is experienced. Continuing the 5% example, if the return is applied twice in a year, the stock will benefit from a return of 2.5% in the first six months, and another 2.5% in the second six months. After six months the asset price will be

$$1{,}000\left(1+\frac{0.05}{2}\right)=1{,}025$$

And after one year the asset price will be

$$P_1=1{,}000\left(1+\frac{0.05}{2}\right)^2=1{,}050.625$$

The growth of 0.025 or 2.5% in the first six months also benefits from growth of 0.025 or 2.5% in the second six months. This is known as *compounding*, and it explains why the value of the stock at the end of the year is more than 1,050. In general, if the return is applied *m* times in a year, the asset price at the end of the year will be

$$P_1=P_0\left(1+\frac{r}{m}\right)^m \tag{1.13}$$

We could increase *m* to 12 or 365, to see what the price of the stock would be if the return were applied (or compounded) every month or every day. We could also ask what continuous compounding would look like. Continuous compounding or continuous growth is when the return is experienced an infinite number of times in the year, but the return at each point of the year is infinitesimally small. That is, what happens if *m* approaches infinity? You can see that r/m will approach zero, but the expression in brackets in equation (1.13) will be raised to the power infinity. The limit of this expression when *m* approaches infinity is e^r, where *e* is equal to 2.718 (to three decimal places). The value *e* is known as the base of natural logarithms.

Going back to our example, if the stock is initially valued at $1,000, and experiences continuous growth at an annual rate of 5% (or 0.05), it will be valued at the end of the year at

$$1{,}000e^{0.05}=1{,}051.27$$

and in general terms

$$P_1=P_0e^r \tag{1.14}$$

We can calculate the logarithmic rate of return (also known as the continuously compounded return) as

$$r=\ln\frac{P_1}{P_0}=\ln P_1-\ln P_0 \tag{1.15}$$

where ln represents the natural logarithm, or the logarithm to base *e*. To see this, take natural logarithms of the end-of-year continuously compounded stock price

$$\ln P_1 = \ln\left(P_0 e^r\right) = \ln P_0 + \ln\left(e^r\right) = \ln P_0 + r \ln e = \ln P_0 + r$$

since the natural log of *e* is 1.

In one of the exercises at the end of the unit you will show that arithmetic returns are not symmetric: if a stock valued at $1,000 experiences first a return of minus 10% and then a return of 10%, it will not be equal to $1,000 at the end. On the other hand, you will find out that logarithmic returns are symmetric. You will also use R to calculate arithmetic and log returns.

Note that in this module, returns will always be calculated as decimals, so a return of 5%, for example, will be shown as 0.05. It will not be shown as 5.00. A consistent approach is necessary, and the decimal representation makes calculations a little bit simpler.

1.4 Study Guide

First, let us consider notation. In econometrics, population parameters and their estimators are normally denoted by Greek letters; the module units and your key text follow this standard practice. Your key text uses lowercase for the names of variables, and the units use uppercase. Table 1.1 summarises the principal differences and similarities in notation.

Table 1.1 Notation

	Module units	Key text
Population parameters	β_0, β_1	β_0, β_1
Their estimators	$\hat{\beta}_0$, $\hat{\beta}_1$	$\hat{\beta}_0$, $\hat{\beta}_1$
Variable names	Y_i, X_i	y_i, x_i
Disturbances	u_i	u_i
Residuals	e_i	$\hat{u}_i$
Number of observations	N	n

For this unit you are requested to study Chapter 1 of the module key text, *Introductory Econometrics* by Wooldridge. This chapter has four main sections, the first one of these addresses the question: What is econometrics? This section is straightforward, and you can read it relatively quickly.

🕮 Reading 1.1

Please now read Section 1-1, pages 1–2, of your key text by Wooldridge.

✍ Make notes of the important points.

Wooldridge (2020) Section 1-1 of Chapter 1 'The nature of econometrics and economic data' in *Introductory Econometrics*.

The next section of the key text is particularly important. It sets out a methodology of econometrics; that is, it explains how you might proceed in a typical econometric study. Wooldridge considers the steps associated with the typical econometric investigation. An empirical analysis begins by stating the question or questions you are interested in. We can then specify a formal economic model that describes the relationships between the variables with which we are concerned. The economic model is influenced by economic theory, but the extent of this influence varies depending on the context. Some economic models are derived from complex formal theoretical analysis, while in other models the relations between the variables can be suggested by economic theory without more formal model building.

The statement of the econometric model is a particularly important step. The economic model might not specify a precise form for the relations between variables; but the econometric model requires more precision, and a statement of functional form. The economic model might include theoretical variables which cannot be observed; the econometric model must include variables we can observe and get data on. The econometric model also needs to take account of the fact that we cannot include every influence in the model, and that relations between variables are not exact; so the econometric model includes the error or disturbance term.

The empirical analysis continues with data collection, estimation of the econometric model, and testing of hypotheses.

You should be able to see that the steps described are relevant to econometric investigation in any discipline, including finance. Notice also the central role of estimating the parameters of the model and so obtaining the estimated *regression* line.

The notation in the key text differs slightly from the notation in these units. Suppose we are interested in a simplified version of the wage equation in equation [1.4] in the reading, and the simplified equation involves only wages and education. Wooldridge defines *wage* as the hourly wage and *educ* as years of formal education, and we could write the bivariate population regression function using the Wooldridge variables as

$$wage = \beta_0 + \beta_1 educ + u \qquad (1.16)$$

which is comparable to our population regression function

$$Y_i = \beta_0 + \beta_1 X_i + u_i \,. \qquad (1.1)$$

Reading 1.2

Please read *carefully* Section 1-2, pages 2–5, of the key text

Wooldridge (2020) Section 1-2 of Chapter 1 'The nature of econometrics and economic data' in *Introductory Econometrics*.

The next section in Wooldridge concerns the structure of data, including cross-sectional data, time series data, pooled cross-sections and panel data. In brief, cross-sectional data involves observations on individuals, households, firms, or other economic units, at a particular point in time. Time series data involves observations on one or more variables over time. Pooled cross-sectional data involves cross-sections observed at two or more points in time, and where the units (the individuals, households, firms, *etc*) may not be the same in the different cross-sections. Panel data involves time series observations for each member of a cross-section.

In this module you will study how to use econometric methods to examine the relationships between financial variables. The examples and exercises will mainly use time series data. However, you should be aware of the different data structures, how the data is collected, and the differences between them.

Wooldridge makes two very important points in this section.

The first point is that the essential econometric procedures that you will study in this module can be used with cross-sectional *and* time series data. These fundamental procedures are covered in Part 1 of your key text, titled 'Regression analysis with cross-sectional data'. Although we will mainly be concerned with time series data in the examples, we can make use of these essential econometric methods, and most of the readings for the module will be from Part 1 of your key text.

The second point is that we cannot view observations in a time series data set as being from a random sample. If you think about the share price of a company, for example, the price in one period is very likely to be influenced by the price in the previous period. Some financial and economic time series exhibit what are called trends over time, in which previous values influence current and future values. This can present a particular challenge for estimation and interpretation of econometric models. The methods that have been developed to deal with these challenges are considered in Part 2 of your key text, 'Regression analysis with time series data' and Part 3 'Advanced topics'. These methods are studied in the module *Econometric Analysis and Applications*.

To summarise, for this module you will study the essential econometric methods that can be used for both time series and cross-sectional data. You will examine how these methods can be applied with time series data, and you will study some of the particular features of time series data and time series econometric models.

Reading 1.3

Please read Section 1-3, pages 5–10, of your key text. You should read all of this section, but please give particular attention to Section 1-3b Time series data, and Section 1-3e A comment on data structures. Note the importance of time and data frequency in time series data.

As Wooldridge notes, we can use many of the same tools to analyse time series data and cross-sectional data, and in this module you will study these essential techniques. However, some aspects of time series data require more complex econometric techniques, compared to cross-sectional data (covered in Parts 2 and 3 of the key text). These more advanced methods are studied in the module *Econometric Analysis and Applications*.

Wooldridge (2020) Section 1-3 of Chapter 1 'The nature of econometrics and economic data' in *Introductory Econometrics*.

Earlier in the unit, in Section 1.3.2, we noted the difficulty of 'holding other things equal' when trying to identify the effect of changes in one variable on another variable. For example, how can we identify a causal effect of changes in the price of an asset on the demand for the asset, if we are unable to keep all other prices unchanged? This idea of *ceteris paribus* or 'other things being equal', and the way we isolate causal effects, are explored in more detail in the final reading in this unit.

Reading 1.4

Please read Section 1-4 and the Summary to Chapter 1, pages 10–14, of the key text. The problem of identifying causality in time series models is examined in Example 1.6, and in Example 1.7 on the expectations hypothesis.

Wooldridge (2020) Section 1.4 and Summary in Chapter 1 'The nature of econometrics and economic data' in *Introductory Econometrics*.

1.5 An Example – Efficiency in the Foreign Exchange Market

This example examines the hypothesis of efficiency in the foreign exchange market, and provides a demonstration of the eight steps that are typical of any econometric investigation.

Statement of the theory

Efficiency in markets is a central assumption of many theories in finance and economics. The efficient markets hypothesis states that current prices will reflect all available information. Applied in the exchange rate market, the hypothesis suggests that the forward exchange rate is the market's expectation of the spot rate that will exist in the future. Any difference between the forward rate formed in the previous period and the spot rate in the current period should be entirely random and unpredictable. In addition, there should be a close relation between the forward rate from the previous period and the spot rate in the current period.

Collection of data

The data to be used are monthly time series data for the spot exchange rate between UK sterling and the US dollar, measured in dollars per pound, and the one-month-ahead forward exchange rate, also measured in dollars per pound. The data cover the period January 1982 to January 2012. The source of the data is www.bankofengland.co.uk. (Bank of England, nd accessed June 2019)

Figure 1.1 Scatter plot of S (current spot rate) on lag(F, −1) (previous forward rate), 1982–2012

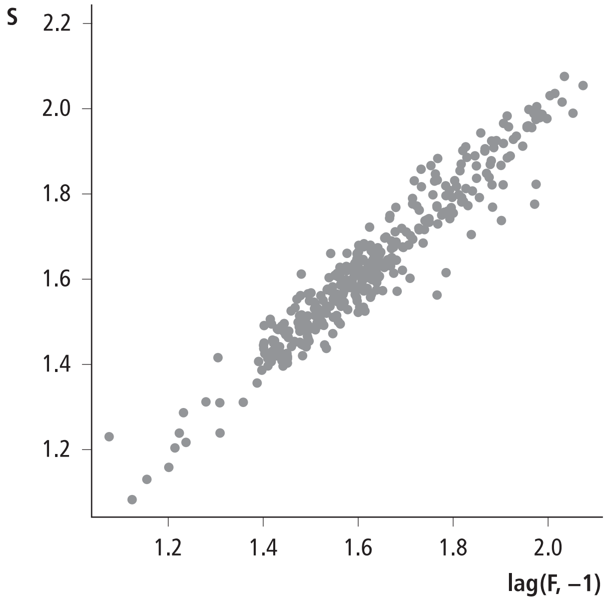


Figure 1.1 shows a scatter plot of the current spot rate, S, against the forward rate available in the previous month, lag (F, −1). The figure suggests that the relationship is upward sloping and it seems to be reasonably linear.

Mathematical model of the theory

The relation between the current spot rate and the forward rate in the previous month in its simplest form can be presented as a linear relationship

$$S_t = \beta_0 + \beta_1 F_{t-1} \tag{1.17}$$

where S_t is the spot rate in period *t*; F_{t-1} is the one-month ahead forward rate available in the previous period, $t-1$; β_0 is a constant (or intercept) and β_1 is the slope of the function. For the efficient markets hypothesis to hold we would expect $\beta_0 = 0$ and $\beta_1 = 1$.

Econometric model of the theory

The econometric model is stochastic. It includes a random error, u_t, which captures the influence of all the other variables that may influence the spot exchange rate.

$$S_t = \beta_0 + \beta_1 F_{t-1} + u_t \tag{1.18}$$

The disturbance term u_t is crucial to the distinction between a mathematical model and an econometric model. In the mathematical model we have a function – there is a unique value of the spot rate for each value of the previous forward rate. With the econometric model, we have a relation in which there is no longer a unique value of the spot rate for each value of the previous forward rate. In the context of the efficient markets hypothesis the disturbance term has additional interpretation: according to the hypothesis, any difference between the previous forward rate and the current spot rate should be random and unpredictable.

Parameter estimation

Using these data and R, it is possible to obtain estimates of the parameters β_0 and β_1 to obtain the average relationship between S_t and F_{t-1}. The problem of estimating the coefficients of the population regression function will be discussed in Unit 2. The function estimated with our data is

$$\hat{S}_t = 0.064 + 0.962F_{t-1} \tag{1.19}$$

and this represents the average relationship between the spot exchange rate and the previous forward exchange rate. The estimated value of β_0, $\hat{\beta}_0$, is 0.064 and the estimated value of β_1, $\hat{\beta}_1$ is 0.962. Consequently if the forward exchange rate increases by 0.01, the spot rate in the next period increases on average by 0.00962. The interpretation of the intercept is not as straightforward. Mechanical interpretation of the estimate tells us that the spot exchange rate is $0.064 per pound if the forward exchange rate in the previous period is zero. On its own, this statement is without meaning. However, in the context of the efficient markets hypothesis, we may ask if the estimated constant indicates there is a systematic and predictable difference between the average spot rate in a period, and the spot rate expected by the markets in the previous period (as measured by the forward rate), and whether this difference could be exploited by traders.

Checking for model adequacy

How appropriate is the model? Should some other variable(s) be included, and is the functional form correct? For example, research on the efficient markets hypothesis in exchange markets has used the natural logarithms of the spot and forward exchange rates. Alternatively, researchers have focussed on the rates of return on the spot and forward exchange rates, and not the levels. Researchers have also examined if the difference between the spot rate and the previous forward rate ($S_t - F_{t-1}$) can be explained by the difference that was observed in earlier periods. With the relevant data, we could estimate various specifications of the relation between spot and forward exchange rates. How do we choose the best model? This is discussed in Unit 8.

Tests of the hypothesis

Do the results conform to the theory of the efficient markets hypothesis? With our theory we expect $\beta_0 = 0$ and $\beta_1 = 1$. Is each of these hypotheses supported by the results? Our estimates would appear to be consistent with what we expected to obtain, but we should conduct formal tests to check that this is actually the case. Formal tests of hypotheses will be discussed in Unit 4.

Prediction

How might the estimated model be used for prediction? We could use it to predict what the spot exchange rate would be if the forward rate in the previous period was a particular amount. Suppose the forward exchange rate in the previous month was \$1.50 per £1.00. The predicted level of the spot rate is

$$\hat{S}_t = 0.064 + 0.962 \times 1.50 . \tag{1.20}$$

Therefore

$$\hat{S}_t = 1.507 .$$

That is, the spot rate is predicted to be \$1.507 per £1.00 if the forward rate in the previous period is \$1.50 per £1.00.

1.6 Conclusion

In this unit we introduced some basic ideas on econometrics and regression analysis. The most important points to remember are the following:

- Econometrics is the application of statistical and mathematical methods to the analysis of data, with a purpose of giving empirical content to economic and financial theories and verifying them or refuting them.

Three elements account for the difference in the work of an econometrician in relation to an economic or finance theorist:

1. the fact that we cannot 'hold other things constant' in empirical analysis
2. the imperfect nature of relations between variables which makes the conclusions and outcomes of empirical analysis always contain a considerable element of uncertainty, and
3. the discrepancy between theoretical variables and observed data in terms of coverage and precision of measurement.

Regression analysis constitutes the statistical foundation of econometric theory and practice. Its aim is to bring out relations between variables, especially between variables whose relation is subject to chance variation and to the influence of unforeseen events.

Regression involves finding an average line, which summarises the relation of Y on X among considerable chance variation and uncertainty of outcome.

The uncertainty inherent in conclusions and outcomes based on regression analysis is formally modelled through the introduction of a disturbance term in our behavioural equations. This is a stochastic variable, which we cannot observe in practice. However, the residuals of a sample regression function may provide us with an indication as to the behaviour of these unknown disturbances.

Regression allows us to investigate the association between variables, but this does not imply any causality between them. To establish causality we need to use economic and finance theory.

In empirical work in economics and finance we cannot use experimentation. Econometric analysis, therefore, is based on careful observation of data drawn from a context that we do not control.

In terms of practical skills, this unit requires that:

- you are familiar with the scatter plot as a practical tool of empirical analysis
- you know how to enter data in R by opening a pre-existing text file
- you know the R commands or operations to obtain a summary of descriptive statistics of a variable, make a scatter plot, create logarithms of variables, and create rates of return.

1.7 Working with R

If you have not done so already, now would be a good time to install R.

In this unit you will also be using the zoo package, so please also install that now. *Installation* of a package is a one-off procedure.

But remember that to use a function from a package in a particular R session, the package must be *loaded* before you try to use the function (loading is a very quick procedure). So, for example, to use the zoo() function, the zoo package must be loaded.

The data files for the exercises are available on the VLE in the module area for this study session.

The commands to perform the operations required in the exercises are provided in the unit, following each exercise.

The commands to read the data files are written on the basis that the working directory for the R session includes the data file you are working with (there is no file path in the command to read the data file). So please make sure the working directory includes the data file.

(Our preference is to organise data files in folders using a system that makes sense from an organisational point of view, and then change the working directory for the R session so it is the required folder. Alternatively, you could save the data file in the default R working directory.)

The command to find the current working directory is

```
getwd()
```

and you can change the working directory with File | Change dir…

Please note that R is case sensitive. Also note that <- has the same effect as =. Spaces can be used to separate elements in the commands, to make them more legible, but spaces are not necessary for the command to operate.

R is a very powerful and flexible programming environment. R and the packages written for R have many features that you will not use in this module, so don't worry if you see methods or notation in the Help files that are not covered in this module. Everything you need to understand is described in the module units, readings, and exercises.

Lastly, answers to the exercises are provided for you to check you have understood and done the exercises correctly. If you do the exercises yourself, you will develop a good understanding of the module materials, and the models and methods described in the units; you will also become more confident using these methods and using R.

Do not go straight to the answers!

1.8 Exercises

Question 1

What is the critical distinction between econometrics and (i) economic or finance theory and (ii) mathematical finance and economics?

Question 2

The file M430_U1_Q2.txt contains the data used in the example in the unit. It is monthly time series data on the exchange rate between the US dollar and UK sterling, measured in dollars per pound. The current spot exchange rate is denoted *S*, and the one-month ahead forward rate is denoted *F*. The data relate to the period January 1982 to January 2012, and the source of the data is www.bankofengland.co.uk (Bank of England, nd accessed June 2019).

a) Produce a plot of the spot rate, *S*, over time. Comment on the plot. Are there any noteworthy episodes?

b) Produce a scatter plot of the current spot rate, S_t, on the vertical axis and the forward rate available in the previous period, F_{t-1}, on the horizontal axis. Comment on the scatter plot; would a linear regression seem appropriate?

c) Produce a plot over time of the difference between the current spot rate and the forward rate available in the previous period, $S_t - F_{t-1}$. Comment on the plot; are there periods when the current spot rate differs noticeably from what is predicted by the previous forward rate?

d) Produce a scatter plot with the difference between the current spot rate and the forward rate available in the previous period, $S_t - F_{t-1}$, on the vertical axis, and this difference one month ago, $S_{t-1} - F_{t-2}$, on the horizontal axis. Comment on the scatter plot; does there appear to be a relationship between the two transformed series?

Data files

The file M430_U1_Q2.txt is a tab separated text file. Text files are very basic, they are readable by many applications, (you could open them in a spreadsheet, for example, or a text editor, or in R), and they are robust to upgrades in software. For these reasons, the data files for the module are all provided in the simplest (and most accessible) format, text files.

The first line of M430_U1_Q2.txt contains the labels for the three columns: Date, S and F. The next row contains the data for the first observation: 31/01/1982, 1.8835 and 1.8837, separated by tabs. Row 3 is 28/02/1982, 1.8225 and 1.8237, and so on. The final row contains 31/01/2012, 1.578 and 1.5777. A useful tip when working with data is to note the first and last observations for your variables, so that you can check files have been opened successfully (and completely).

Read data from text file and create zoo objects

The following commands first read the data from the text file M430_U1_Q2.txt into the data frame M430_U1_Q2.

A zoo object containing data on S and F is created from the data frame; within this zoo object the data are indexed by the dates contained in the text file.

For easier manipulation, two separate zoo series called S and F are also created (these will be indexed by the correct dates).

If you copy the commands to R one-by-one, you will need to press Enter to execute each command in R. If you copy the four commands together, the first three will be executed immediately and in sequence, and you will need to press Enter to run the fourth command.

The first line assumes the file M430_U1_Q2.txt is located in the working directory accessed by R. You can change the working directory using File | Change dir… to select the folder where you have saved M430_U1_Q2.txt.

Before you execute these commands, make sure you have installed the zoo package on your device, and that it is currently loaded.

```
M430_U1_Q2 <- read.table("M430_U1_Q2.txt", sep = "\t",
header = TRUE)

M430_U1_Q2_zoo <- read.zoo(M430_U1_Q2, format =
"%d/%m/%Y")

S = zoo(M430_U1_Q2_zoo$S)

F = zoo(M430_U1_Q2_zoo$F)
```

In the read.table command, sep = "\t" indicates the data are separated by tabs. Within the read.zoo command, the format for the dates as they appear in the data file is specified with format = "%d/%m/%Y".

Note the use of the quotation symbol " in the R commands. The symbols “ and ” will not be recognised in R. After copying the commands to R you may see a plus sign at the start of a line of code; this indicates the R command continues over two lines.

To see the data for *S*, say, type S (followed by Enter).

To see which objects have been created, use

```
        ls()
```

Saving a Workspace

To save a Workspace, use File | Save Workspace… (or use Control and s). Provide a name for the file, which will be saved as an R image with the extension .Rdata. The file will be saved in the current working directory (or you can browse to another folder).

To load a Workspace that you have saved previously, use File | Load Workspace…, and select the file to load. You can browse to another folder if the .Rdata file you want to load is not in the current working directory.

Deleting an object

To delete an object, use rm(). For example, to remove the object S you would use rm(S).

To remove all objects (for example, if you want to start with a new data set but you would like to keep working with the same loaded packages), go to Misc | Remove all objects. The same thing can be achieved with

```
        rm(list = ls(all = TRUE))
```

Producing a graph

Analysing graphs of your data is a very useful method for identifying general patterns, relations between series, or noteworthy changes in the data. To demonstrate this, Q2a examines a plot of a series over time; Q2b examines a scatter plot where one series is plotted against another; Q2c requires a plot of transformed series over time; and Q2d considers a scatter plot of two transformed series.

To produce a plot of the current spot rate, *S*, use

```
        plot(S, lwd = 2, xaxs = "i")
```

The text lwd = 2 determines the width of the plot line. xaxs = "i" extends the plot to the left and right edges of the plot area. The label on the horizontal axis will be Index. This label can be suppressed by including xlab = "" in the plot command

```
        plot(S, lwd = 2, xlab = "", xaxs = "i")
```

or you could add a Date label to the horizontal axis by including xlab = "Date" in the plot command.

To save the graph, make sure the graph window is highlighted, then go to File | Save as, then select the type of file you wish to save, and provide a name for the file. The file will be saved in the current working directory, or you can browse to another folder.

To copy the file to another application, use File | Copy to the clipboard, and copy as a bitmap or a metafile.

(These procedures can also be accessed by right-clicking over the graph.)

To continue inputting commands to R, make sure the R Console is selected. If the open graph is selected, R will not respond to typing or pasting.

Producing a scatter plot

Next, produce a scatter plot with S_t on the vertical axis and F_{t-1} on the horizontal axis. This can be done with

```
plot(lag(F,-1), S, pch = 19)
```

The first series in this plot command is measured on the horizontal axis. The second series in this plot command is measured on the vertical axis. pch = 19 specifies the symbol of character to use in the plot.

Let us explain how the lag( ) operator works in R. The first three observations for S and F are shown in Table 1.2.

Table 1.2 S and F, first three observations

Date	S	F
1982-01-31	1.8835	1.8837
1982-02-28	1.8225	1.8237
1982-03-31	1.7833	1.7866
…	…	…

We need to consider the series F_{t-1}, which is the one-month ahead forward exchange rate that was observed in the previous month. To get this series we need to use the term lag(F, -1) in the plot command in R.

If you type

```
lag(F,-1)
```

you will see that the first two values of lag(F,-1) are as shown in the next table.

Table 1.3 lag(F,-1) first two observations

Date	lag(F,-1)
-	-
1982-02-28	1.8837
1982-03-31	1.8237
…	…

There is no observation for lag(F,-1) for the date 1982-01-31. The data set does not include an observation for the one-month ahead forward exchange rate for December 1981, so we cannot create the lagged value of F for the observation 1982-01-31. By using the first lag of F we lose the first observation from the sample period.

In this module we will use this lag feature a lot. The unit will always provide the required R script to create the lags, but it is important you understand what is being done.

Why do we include -1 in the specification lag(F,-1)? In R the lag operator works by shifting a series backwards (and the default is to shift the series back by one period). So, in R, lag(F) is the one-month ahead forward exchange rate that will be observed in the following month, which is F_{t+1}.

You can check this by typing

```
lag(F)
```

and you will see the observations for lag(F) shown in Table 1.4. The value for lag(F) for 1982-01-31 is the value of F on 1982-02-28.

Therefore, to obtain F_{t-1} we use the specification lag(F,-1).

Table 1.4 lag(F) first three observations

Date	lag(F)
1982-01-31	1.8237
1982-02-28	1.7866
1982-03-31	1.7999
...	...

Question 2c requires a plot of $S_t - F_{t-1}$ over time.

The commands to obtain this are

```
plot(S-lag(F,-1), lwd = 2, xlab = "", xaxs = "i")
abline(h = 0)
```

The command abline(h = 0) draws a horizontal straight line through the origin on the vertical axis.

Question 2d requires a scatter plot with $S_t - F_{t-1}$ on the vertical axis, and $S_{t-1} - F_{t-2}$ on the horizontal axis. Can you think what two expressions are required for this graph (to go in the plot command)? And what order should they be in to produce this scatter plot?

The scatter plot of $S_t - F_{t-1}$ against $S_{t-1} - F_{t-2}$ is obtained with

```
plot(lag(S,-1)-lag(F,-2), S-lag(F,-1), pch = 19)
abline(h = 0)
abline(v = 0)
```

The command abline(v = 0) draws a vertical straight line through the origin on the horizontal axis.

Notice how we obtain the one-month ahead forward exchange rate that was observed two months ago with lag(F, -2).

Generating new series

As you can see, it is possible to type expressions for series (transformations of series) directly into the plot commands. Sometimes it will be more convenient, or you may prefer it, to create new series that incorporate the transformations, and then work with the new series. So, in Q2c you could create a new series, call it *Z*, equal to the difference between the current spot rate and the forward rate from the previous period, and then plot *Z*.

To generate $Z = S_t - F_{t-1}$ you would use

```
        Z = S - lag(F,-1)
```

The first few observations for *Z* are shown in Table 1.5.

Table 1.5 S - lag(F,-1), first two observations

Date	Z = S - lag(F,-1)
-	-
1982-02-28	–0.0612
1982-03-31	–0.0404
…	…

You can confirm that the value of Z on 1982-02-28 is the value of S on 1982-02-28 minus the value of F on 1982-01-31

$$1.8825 - 1.8837 = -0.0612$$

Question 3

A share is valued at $1,000 at the start of year 1. In year 1 it experiences a return of −20%, and in year 2 it experiences a return of +20%. Calculate the value of the share at the end of year 1 and the end of the year 2 using

a) arithmetic returns, and

b) logarithmic returns.

Comment on the values you have obtained for the share price at the end of year 2.

Question 4

The tab delimited text file M430_U1_Q4.txt contains the share price of Delta Airlines Inc. (DAL) and the New York Stock Exchange Composite Index (NYA). The data are daily, for the period 1 March 2010 to 1 March 2012, and both series are measured in US dollars (Yahoo!, nd accessed June 2019). The text file also includes a column of dates.

a) Plot the series DAL and NYA over time. Comment on the plots.

b) Plot the daily logged return of Delta Airline shares over time, and comment on the plot.

c) Produce a scatter plot with the daily logged return of Delta shares on the vertical axis, and the daily logged return on the NYSE composite index on the horizontal axis. Comment on the scatter plot.

d) Calculate the means, standard deviations, and minimum and maximum values for the Delta Airline daily logged return and daily arithmetic return. Comment on the values you have obtained.

The file contains three columns. The first column contains the date; the second column contains the share prices for Delta Airlines Inc. (DAL); and the third column contains the value for the NYSE Composite Index (NYA). For reference, on 1/3/2010 DAL is 13.17 and NYA is 7100.75; on 2/3/2010 DAL is 12.78 and NYA is 7135.97; and on 1/3/2012 DAL is 9.64 and NYA is 8175.11.

Read the data from the text file and create the zoo objects with the following four commands. Recall that the zoo functions are available in the zoo package.

```
M430_U1_Q4 <- read.table("M430_U1_Q4.txt", sep = "\t",
header = TRUE)

M430_U1_Q4_zoo <- read.zoo(M430_U1_Q4, format =
"%d/%m/%Y")

DAL = zoo(M430_U1_Q4_zoo$DAL)

NYA = zoo(M430_U1_Q4_zoo$NYA)
```

To produce a multi-panel plot of DAL and NYA, use

```
plot(M430_U1_Q4_zoo, lwd = 2, main = "", xlab = "", xaxs =
"i")
```

For Q4b you need to plot the daily logged return for the shares of Delta Airlines. Recall from the unit that the daily logged return is equal to

$$R_t = \ln P_t - \ln P_{t-1} \tag{1.21}$$

that is, the natural logarithm of the share price minus the natural logarithm of the share price from the day before.

To plot the daily logged return of DAL

```
plot(diff(log(DAL)), xlab = "", lwd = 2, xaxs = "i")

abline(h = 0)
```

The term diff(log(DAL)) provides the first difference of the log of the Delta share price, that is the current logged value minus the logged value in the previous period. Note that log in R is the natural logarithm by default.

Equivalently, you could create a new series for the logged return with

```
r = log(DAL) - lag(log(DAL),-1)
```

and then obtain a plot of r.

Q4c requires a scatterplot of the daily logged return on Delta shares against the daily logged return on the market index. This is achieved with

```
plot(diff(log(NYA)), diff(log(DAL)), pch = 19)
abline(h = 0)
abline(v = 0)
```

Sample statistics

You can produce the mean, standard deviation, minimum value and maximum value for a series with mean( ), sd( ), min( ) and max( ). Remember that DAL has been created as a zoo object. To use these functions on transformations of DAL, please make sure you have loaded the zoo package before you try to execute the functions.

The mean of the daily logged return for Delta Airlines shares is computed with

```
mean(diff(log(DAL)))
```

The standard deviation of the logged return is computed with

```
sd(diff(log(DAL)))
```

The maximum value of the logged return is found with

```
max(diff(log(DAL)))
```

The minimum value of the logged return is found with

```
min(diff(log(DAL)))
```

The arithmetic return is

$$RA_t = \frac{\left(P_t - P_{t-1}\right)}{P_{t-1}} \tag{1.22}$$

that is, the current share price minus the share price in the previous period, all divided by the share price in the previous period.

The mean for the arithmetic return can be computed with

```
mean((diff(DAL))/lag(DAL,-1))
```

The standard deviation for the arithmetic return can be computed with

```
sd((diff(DAL))/lag(DAL,-1))
```

The maximum value of the arithmetic return is found with

```
max((diff(DAL))/lag(DAL,-1))
```

And the minimum value of the arithmetic return is found with

```
min((diff(DAL))/lag(DAL,-1))
```

1.9 Answers to Exercises

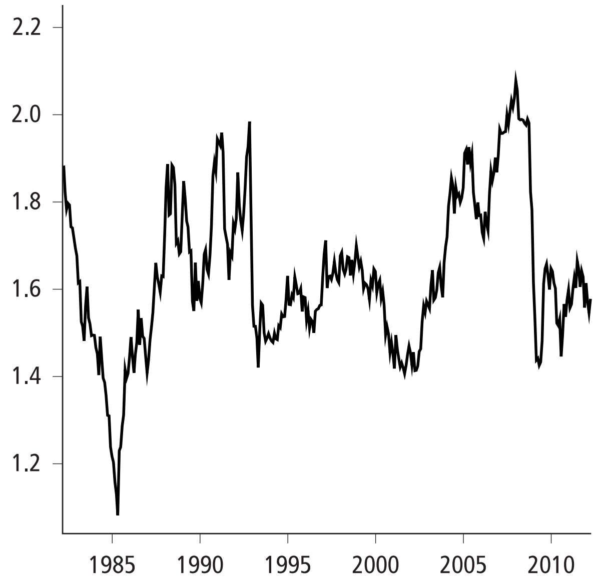
Question 1

Economic and finance theory can be viewed as a set of qualitative relations among variables. Such theory can frequently be written in the form of a mathematical model. An econometric model may be obtained from an appropriate mathematical model with the addition of a random error term. By using data to estimate the econometric model we can in effect quantify financial and economic relations.

Question 2

a) The plot of S over time is shown in Figure 1.2. The plot of the current spot rate, measured as US dollars per pound, reveals a number of notable episodes. For example, there is a sharp depreciation of sterling in 1992, when sterling left the European Exchange Rate Mechanism. (A lower value for S means that one pound will buy fewer dollars, or equivalently, it takes fewer dollars to buy one pound). There is another sharp depreciation of sterling against the dollar (and other currencies) after the 2008 financial crisis.

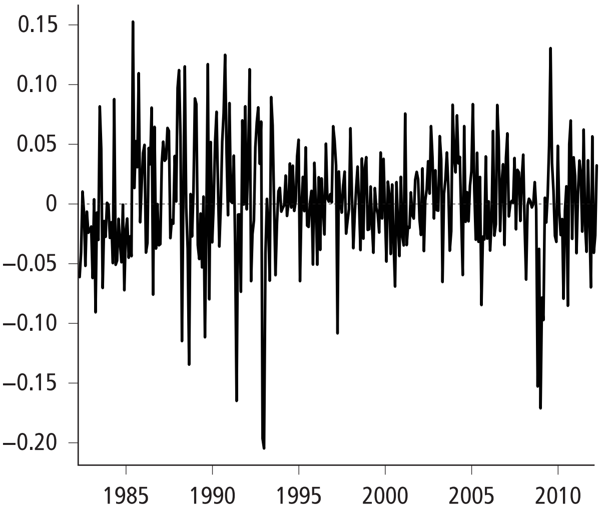
Figure 1.2 Plot of S 1982–2012



b) The scatter plot of S_t against F_{t-1} is shown in Figure 1.1 in the unit. The scatter plot shows that S_t and F_{t-1} have the expected positive relationship. The relationship seems to be approximately linear and seems to be relatively strong, in that the observations appear close to a regression line drawn in the scatter plot.

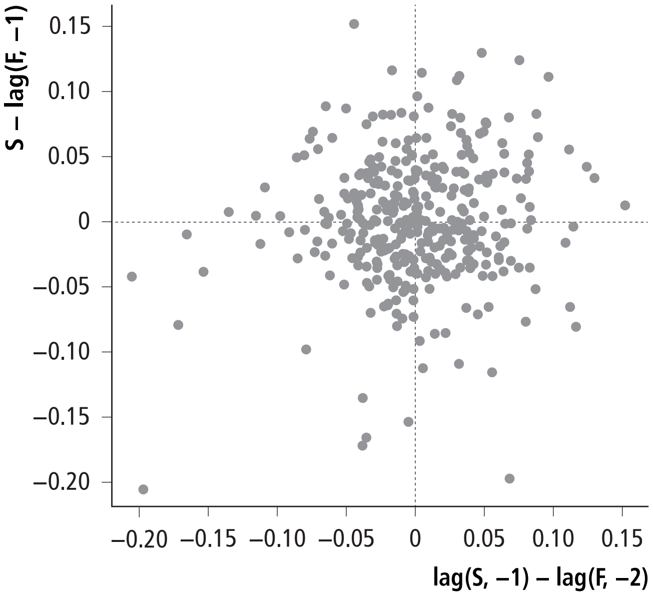
c) Figure 1.3 shows the plot of $S_t - F_{t-1}$ over time. This is the difference between the current spot rate and the one-month ahead forward rate that was available one month previously. According to the efficient markets hypothesis, the forward rate should be a good predictor of the spot rate, so that any differences between S_t and F_{t-1} should be random. Any differences also reflect information that has become available between the time the forward exchange rate was formed, and the current spot rate was formed. In the first few years of the sample there are months when $S_t - F_{t-1}$ is consistently negative. If F_{t-1} is consistently greater than S_t it suggests that the forward market is consistently under-predicting the value of the spot rate. Looking back at Figure 1.2, sterling was steadily depreciating in this period. This means the forward market is consistently underpredicting the extent of the depreciation in the spot rate. Again in 2008, there are relatively large negative values for $S_t - F_{t-1}$ for a few months, and the same interpretation might be applied: the forward market is not adequately predicting the depreciation in sterling.

Figure 1.3 Plot of S – lag(F,–1) 1982–2012



d) Figure 1.4 shows the scatter plot of $S_t - F_{t-1}$ against $S_{t-1} - F_{t-2}$.

Figure 1.4 Scatterplot of S – lag(F,–1) against lag(S, –1) – lag(F,–2) 1982–2012



That is, the difference between the current spot rate and the forward rate one month ago, plotted against the difference in the previous period.

The scatter plot allows us to examine whether the forecasting error between S_t and F_{t-1} can be explained by the forecasting error in the earlier period, $S_{t-1} - F_{t-2}$. Figure 1.4 suggests there is no obvious relationship, positive or negative, between the forecasting error in one period and the forecasting error in the period that follows.

Question 3

The value of the share at the start of year 1 is \$1,000, and in year 1 it experiences a return of −20% or −0.20. In year 2 the return is +20% or +0.20.

a) Using arithmetic returns, the share price at the end of year 1 is

$$P_1 = P_0(1+r) = 1,000(1-0.20) = \$800$$

The share price at the end of year 2 is

$$P_2 = P_1(1+r) = 800(1+0.20) = \$960$$

b) Using logarithmic returns, the share price at the end of year 1 is

$$P_1 = P_0 e^r = 1,000e^{-0.20} = 1,000 \times 0.8187307 = \$818.73$$

The share price at the end of year 2 is

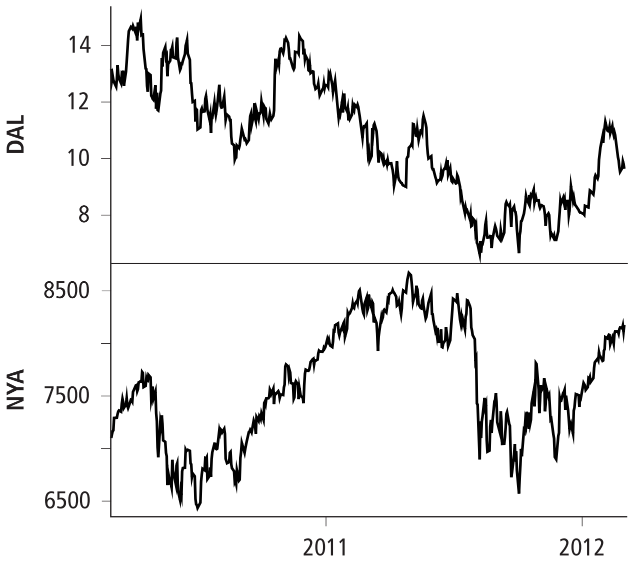
$$P_2 = P_1 e^r = 818.73e^{0.20} = 999.999 \text{ or } \$1,000$$

Arithmetic returns are not symmetric: a negative return, followed by a positive return of the same magnitude, does not restore the share to the original price. However, logarithmic returns are symmetric: a negative return followed by a positive return of the same magnitude does restore the share to the original value.

Question 4

a) The plot of the Delta Airlines Inc. share price and the NYSE Composite Index is shown in Figure 1.5.

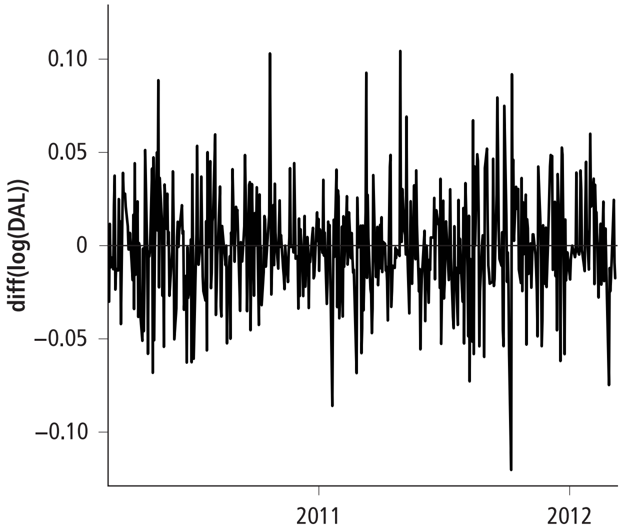
Figure 1.5 Plot of DAL and NYA March 2010 to March 2012



The Delta Airlines share price is displayed in the top panel and the NYSE Composite Index is displayed in the bottom panel. You can see there are periods when both series generally move together, and there are other times when one series exhibits sharp movements that are not shown in the other series.

b) Figure 1.6 shows the plot of the daily logged return of the Delta Airlines share price. The daily logged return crosses the zero line frequently. Occasionally there are large positive and negative daily returns, of around +0.12 (12%) and −0.12 (minus 12%).

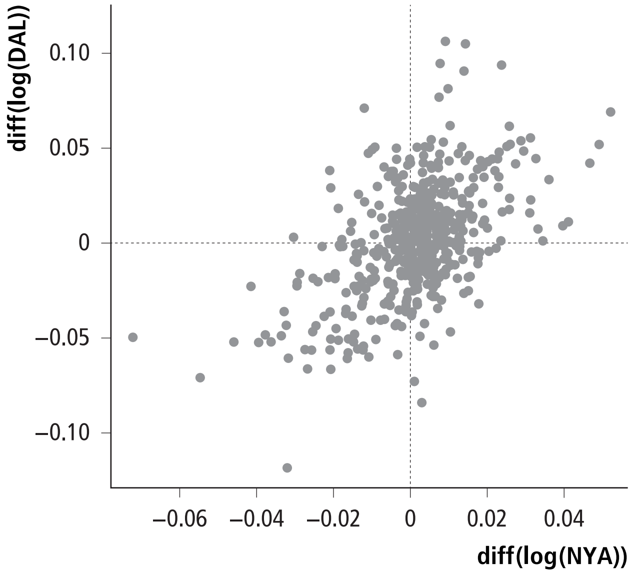
Figure 1.6 Plot of DAL daily logged return March 2010 to March 2012



c) Figure 1.7 shows the scatter plot of the daily logged return on Delta shares (on the vertical axis) against the daily logged return on the

NYSE Composite Index (on the horizontal axis). There would seem to be a positive, linear relationship between the two series.

Figure 1.7 Scatter plot of DAL daily logged return and NYA daily logged return



d) The descriptive statistics for the daily logged return and daily arithmetic return for the Delta share price are shown in Table 1.6.

Table 1.6 Descriptive statistics for logged and arithmetic returns, Delta Airlines

	diff(log(DAL)	(diff(DAL))/lag(DAL, −1)
Mean	−0.000617	−0.000185
Maximum	0.104360	0.110000
Minimum	−0.120286	−0.113333
Std. Dev.	0.029401	0.029448

For small changes, the logged return and arithmetic return are approximately equal. However, for larger changes this approximation is not so close. You can see this in the maximum and minimum values for the two series in Table 1.6.

References

Bank of England (nd) *The Bank of England*. [Online]. Available from: https://www.bankofengland.co.uk [Accessed 26 June 2019]

Maddala GS (1992) *Introduction to Econometrics*. New York: Macmillan.

Mosteller F and JW Tukey (1977) *Data Analysis and Regression: A Second Course in Statistics*. Boston MA: Addison-Wesley.

R Core Team (2019) *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Available from: https://www.r-project.org/ [Accessed 27 November 2019]

Yahoo! (nd) *Finance.* [Online]. Available from: http://finance.yahoo.com [Accessed 26 June 2019]

Zeileis A and G Grothendieck (2005) 'zoo: S3 Infrastructure for Regular and Irregular Time Series'. *Journal of Statistical Software*, 14 (6), 1–27. doi:10.18637/jss.v014.i06