CHAPTER 1

Introduction

There are many types of professionals working in the world today who are interested in financial data. Academics in universities often derive and test theoretical models of behavior of financial assets. Civil servants and central bankers often study the merits of policies under consideration by government. Such policies often depend on what is happening with stock markets, interest rates, house prices and exchange rates. In the private sector, practitioners often seek to predict stock market movements or the performance of particular companies.

For all of these people, the ability to work with data is an important skill. To decide between competing theories, to predict the effect of policy changes, or to forecast what may happen in the future, it is necessary to appeal to facts. In finance, we are fortunate in having at our disposal an enormous amount of facts (in the form of "data") that we can analyze in various ways to shed light on many economic issues.

The purpose of this book is to present the basics of data analysis in a simple, non-mathematical way, emphasizing graphical and verbal intuition. It focuses on the tools used by financial practitioners (primarily regression and the extensions necessary for time series data) and develops computer skills that are necessary in virtually any career path that the student of finance may choose to follow.

To explain further what this book does, it is perhaps useful to begin by discussing what it does **not** do. **Financial econometrics** is the name given to the study of quantitative tools for analyzing financial data. The field of econometrics is based on probability and statistical theory; it is a fairly mathematical field. This book does not attempt to teach much probability and statistical theory. Neither does it contain much mathematical content. In both these respects, it represents a clear departure from traditional financial econometrics textbooks. Yet, it aims to teach most of the tools used by practitioners today.

Books that merely teach the student which buttons to press on a computer without providing an understanding of what the computer is doing, are commonly referred to as "cookbooks". The present book is **not** a cookbook. Some econometricians may interject at this point: "But how can a book teach the student to use the tools of financial econometrics, without teaching the basics of probability and statistics?" My answer is that much of what the financial econometrician does in practice can be understood intuitively, without resorting to probability and statistical theory. Indeed, it is a contention of this book that most of the tools econometricians use can be mastered simply through a thorough understanding of the concept of correlation, and its generalization, regression (including specialized variants of regression for time series models). If a student understands correlation and regression well, then he/she can understand most of what econometricians do. In the vast majority of cases, it can be argued that regression will reveal most of the information in a data set. Furthermore, correlation and regression are fairly simple concepts that can be understood through verbal intuition or graphical methods. They provide the basis of explanation for more difficult concepts, and can be used to analyze many types of financial data.

This book focuses on the **analysis** of financial data. That is, **it is not a book about collecting financial data**. With some exceptions, it treats the data as given, and does not explain how the data is collected or constructed. For instance, it does not explain how company accounts are created. It simply teaches the reader to make sense out of the data that has been gathered.

Statistical theory usually proceeds from the formal definition of general concepts, followed by a discussion of how these concepts are relevant to particular examples. The present book attempts to do the opposite. That is, **it attempts to motivate general concepts through particular examples**. In some cases formal definitions are not even provided. For instance, P-values and confidence intervals are important statistical concepts, providing measures relating to the accuracy of a fitted regression line (see Chapter 5). The chapter uses examples, graphs and verbal intuition to demonstrate how they might be used in practice. But no formal definition of a P-value nor derivation of a confidence interval is ever given. This would require the introduction of probability and statistical theory, which is not necessary for using these techniques sensibly in practice. For the reader wishing to learn more about the statistical theory underlying the techniques, many books are available; for instance *Introductory Statistics for Business and Economics* by Thomas Wonnacott and Ronald Wonnacott (Fourth edition, John Wiley & Sons, 1990). For those interested in how statistical theory is applied in financial econometrics, *The Econometrics of Financial Markets* by John Campbell, Andrew Lo and Craig MacKinlay (Princeton University Press, 1997) and *The Econometric Modelling of Financial Time Series* by Terrence Mills (Second edition, Cambridge University Press, 1999) are two excellent references.

This book reflects my belief that the use of concrete examples is the best way to teach data analysis. Appropriately, each chapter presents several examples as a means of illustrating key concepts. One risk with such a strategy is that some students might interpret the presence of so many examples to mean that a myriad of concepts must

be mastered before they can ever hope to become adept at the practice of econometrics. This is not the case. At the heart of this book are only a few basic concepts, and they appear repeatedly in a variety of different problems and data sets. The best approach for teaching introductory financial econometrics, in other words, is to illustrate its specific concepts over and over again in a variety of contexts.

Organization of the book

In organizing the book, I have attempted to adhere to the general philosophy outlined above. Each chapter covers a topic and includes a general discussion. However, most of the chapter is devoted to empirical examples that illustrate and, in some cases, introduce important concepts. Exercises, which further illustrate these concepts, are included in the text. Data required to work through the empirical examples and exercises can be found in the website which accompanies this book **http://www.wiley.com/go/koopafd**. By including many data sets, it is hoped that students will not only replicate the examples, but will feel comfortable extending and/or experimenting with the data in a variety of ways. Exposure to real-world data sets is essential if students are to master the conceptual material and apply the techniques covered in this book.

Most of the empirical examples in this book are designed for use in conjunction with the computer package Excel. However, for the more advanced time series methods used in the latter chapters of the book, Excel is not appropriate. The computer package Stata has been used to do the empirical examples presented in these latter chapters. However, there is a wide range of other computer packages that can be used (e.g. E-views, MicroFit, Gauss, Matlab, R, etc.).

The website associated with this book contains all the data used in this book in Excel format. Excel is a simple and common software package and most other common packages (e.g. Stata) can work with Excel files. So it should be easy for the student to work with the data used in this book, even if he/she does not have Excel or Stata. Appendix B at the end of the book provides more detail about the data.

Throughout this book, mathematical material has been kept to a minimum. In some cases, a little bit of mathematics will provide additional intuition. For students familiar with mathematical techniques, appendices have been included at the end of some chapters.

The content of the book breaks logically into two parts. Chapters 1–7 cover all the basic material relating to graphing, correlation and regression. A very short course would cover only this material. Chapters 8–12 emphasize time series topics and analyze some of the more sophisticated financial econometric models in use today. The focus on the underlying intuition behind regression means that this material should be easily accessible to students. Nevertheless, students will likely find that these latter chapters are more difficult than Chapters 1–7.

Useful background

As mentioned, this book assumes very little mathematical background beyond the pre-university level. Of particular relevance are:

- Knowledge of simple equations. For instance, the equation of a straight line is used repeatedly in this book.
- Knowledge of simple graphical techniques. For instance, this book is full of graphs that plot one variable against another (i.e. standard *XY*-graphs).
- Familiarity with the summation operator is useful occasionally.
- In a few cases, logarithms are used.

For the reader unfamiliar with these topics, the appendix at the end of this chapter provides a short introduction. In addition, these topics are discussed elsewhere, in many introductory textbooks.

This book also has a large computer component, and much of the computer material is explained in the text. There are a myriad of computer packages that could be used to implement the procedures described in this book. In the places where I talk directly about computer programs, I will use the language of the spreadsheet and, particularly, that most common of spreadsheets, Excel. I do this largely because the average student is more likely to have knowledge of and access to a spreadsheet rather than a specialized statistics or econometrics package such as E-views, Stata or MicroFit.

I assume that the student knows the basics of Excel (or whatever computer software package he/she is using). In other words, students should understand the basics of spreadsheet terminology, be able to open data sets, cut, copy and paste data, etc. If this material is unfamiliar to the student, simple instructions can be found in Excel's on-line documentation. For computer novices (and those who simply want to learn more about the computing side of data analysis) *Computing Skills for Economists* by Guy Judge (John Wiley & Sons, 2000) is an excellent place to start.

Appendix 1.1: Concepts in mathematics used in this book

This book uses very little mathematics, relying instead on intuition and graphs to develop an understanding of key concepts (including understanding how to interpret the numbers produced by computer programs such as Excel). For most students, previous study of mathematics at the pre-university level should give you all the background knowledge you need. However, here is a list of the concepts used in this book along with a brief description of each.

The equation of a straight line

Financial analysts are often interested in the relationship between two (or more) variables. Examples of variables include stock prices, interest rates, etc. In our context a variable is something the researcher is interested in and can collect data on. I use capital letters (e.g. Y or X) to denote variables. A very general way of denoting a relationship is through the concept of a function. A common mathematical notation for a function of X is f(X). So, for instance, if the researcher is interested in the factors which explain why some stocks are worth more than others, she may think that the price of a share in a company depends on the earnings of that company. In mathematical terms, she would then let Y denote the variable "price of a share" and X denote the variable "earnings" and the fact the Y depends on X is written using the notation:

$$Y = \mathrm{f}(X)$$

This notation should be read "Y is a function of X" and captures the idea that the value for Y depends on the value of X. There are many functions that one could use, but in this book I will usually focus on linear functions. Hence, I will not use this general "f(X)" notation in this book.

The equation of a straight line (what was called a "linear function" above) is used throughout this book. Any straight line can be written in terms of an equation:

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

where α and β are *coefficients* which determine a particular line. So, for instance, setting $\alpha = 1$ and $\beta = 2$ defines one particular line while $\alpha = 4$ and $\beta = -5$ defines a different line.

It is probably easiest to understand straight lines by using a graph (and it might be worthwhile for you to sketch one at this stage). In terms of an XY graph (i.e. one which measures Y on the vertical axis and X on the horizontal axis) any line can be defined by its intercept and slope. In terms of the equation of a straight line α is the intercept and β the slope. The intercept is the value of Y when $X = 0$ (i.e. point at which the line cuts the Y-axis). The slope is a measure of how much Y changes when X is changed. Formally, it is the amount Y changes when X changes by one unit. For the student with a knowledge of calculus, the slope is the first derivative, $\frac{dY}{dX}$.

Summation notation

At several points in this book, subscripts are used to denote different observations of a variable. For instance, a researcher in corporate finance might be interested in the earnings of every one of 100 companies in a certain industry. If the researcher

uses Y to denote this variable, then she will have a value of Y for the first company, a value of Y for the second company, etc. A compact notation for this is to use subscripts so that Y_1 is the earnings of the first company, Y_2 the earnings of the second company, etc. In some contexts, it is useful to speak of a generic company and refer to this company as the i-th. We can then write, Y_i for $i = 1, \ldots, 100$ to denote the earning of all companies.

With the subscript notation established, summation notation can now be introduced. In many cases we want to add up observations (e.g. when calculating an average you add up all the observations and divide by the number of observations). The Greek symbol, Σ (pronounced "sigma"), is the summation (or "adding up") operator and superscripts and subscripts on Σ indicate the observations that are being added up. So, for instance,

$$\sum\nolimits_{i=1}^{100} Y_i = Y_1 + Y_2 + \ldots + Y_{100}$$

adds up the earnings for all of the 100 companies. As other examples,

$$\sum\nolimits_{i=1}^{3} Y_i$$

adds up the earnings for the first 3 companies and

$$\sum\nolimits_{i=47}^{48} Y_i$$

adds up the earnings for the 47th and 48th companies.

Sometimes, where it is obvious from the context (usually when summing over all companies), the subscript and superscript will be dropped and I will simply write:

$$\sum Y_i.$$

Logarithms

For various reasons (which are explained later on), in some cases the researcher does not work directly with a variable but with a transformed version of this variable. Many such transformations are straightforward. For instance, in comparing different companies financial economists sometimes use the price-to-earnings ratio. This is a transformed version of the stock price and earnings variables where the former is divided by the latter.

One particularly common transformation is the logarithmic one. The logarithm (to the base B) of a number, A, is the power to which B must be raised to give A. The notation for this is: $\log_B(A)$. So, for instance, if $B = 10$ and $A = 100$ then the logarithm is 2 and we write $\log_{10}(100) = 2$. This follows since $10^2 = 100$. In finance, it is common to work with the so-called natural logarithm which has $B = e$ where $e \approx 2.71828$. We will not explain where e comes from or why this rather unusual-

looking base is chosen. The natural logarithm operator is denoted by ln; i.e. $\ln(A) = \log_e(A)$.

In this book, you do not really have to understand the material in the previous paragraph. The key thing to note is that the natural logarithmic operator is a common one (for reasons explained later on) and it is denoted by $\ln(A)$. In practice, it can be easily calculated in a spreadsheet such as Excel (or on a calculator).