1

Introduction

COPYRIGHTED MATERIAL

As mentioned in the preface, this book is a graduate text and a reference book for those who are interested in statistical theories and methods with economic and business applications. The role of econometrics has changed significantly during the last decade. Businesses and governments are now made accountable for making knowledge-based decisions. This requirement, coupled with the development in information and communication technology, has generated an enormous amount of data as a major source of information. This voluminous data is also coupled with some subjective but still useful knowledge in the hands of the decision makers. All of this information needs to be converted into meaningful and useful knowledge. The field of statistical knowledge itself, which came into existence barely a century ago, has expanded during the last few decades. Thus any new book such as this in econometrics must address itself as to how to handle large amounts of data and how to use cutting-edge statistical tools in order to discover the patterns in the data.

Gathering the right data, deciding which data are useful and which are not, data cleaning, data editing, combining quantitative data with other pieces of information and discovering patterns in all that information, etc are the building blocks of useful knowledge. It is that knowledge which is needed for making better business and economic decisions.

Fortunately there has also been a remarkable degree of acceptance in recent years of quantitative analysis in business and economics. The fear of mathematics and statistics, that was a characteristic feature of top management in business and government in the past, has now given way to an appreciation of their usefulness in making knowledge-based decisions. This is due mainly to the developments in computing software with graphics that have made mathematics and statistics a part of a black box. Their importance, however, is demonstrated by innovative graphics in terms of the end results of productivity gains, revenues, profits, reduced risk, etc that such methods can generate. This last part, an effective communication system between the quantitative analyst and the decision makers, is still in its infancy, and needs a great deal more development. We hope that the illustrative examples we give in this book, and the graphics that are built into our software, will go a long way in this direction. There is nevertheless a great danger of excessive use of such software without a proper understanding of

Developing Econometrics, First Edition. Hengqing Tong, T. Krishna Kumar and Yangxin Huang.
© 2011 John Wiley & Sons, Ltd. Published 2011 by John Wiley & Sons, Ltd.

the underlying statistical procedures. A misuse by incompetent people of the analytic tools, which can be easily implemented through the click of a mouse, using freely available open source software, might bring more discredit to analytics than credit. It is the main aim of this book to provide that link between business analytics, analytics software, and the required statistical knowledge. From that perspective this book differs in its scope from several other econometrics books, in the sense that it is aimed at the practitioner of business analytics or an applied econometrician. By providing a new orientation it also helps an academically oriented scholar to pursue academic interests in econometrics with a practical orientation.

We assume that the reader has had an introductory course on probability distributions and statistics, and also on the basic principles of statistical inference. This chapter introduces the types of economic problems and data that require quantitative analysis for business and public policy decisions. The competitive business environment requires that the analysis be done using the best possible statistical tools. Extensive treatment of these statistical methods will engage us in the subsequent chapters of this book. This chapter emphasizes the need to understand clearly the domain of application; as such knowledge is vital to understanding the data generating process or mechanism. Such an understanding is necessary for obtaining the best possible model.

1.1 Nature and scope of econometrics

1.1.1 What is econometrics and why study econometrics?

This book deals with the application of mathematical models, statistical theories and methods, to economic problems. This is an area of both economics and statistics which is called econometrics. When the International Econometric Society was founded in 1930 it defined econometrics as a science devoted to the advancement of economic theory in relation to mathematics and statistics. It was meant to be an interaction between economic theory or mathematical economics and measurement of economic variables, with the theory guiding the attempts at measurement, and measurement in turn modifying theory. Statistical methods become relevant in economics in two different ways. First, there is some variation in the observed economic data that needs to be understood through **exploratory data analysis**. Second, by making certain assumptions regarding the stochastic mechanism that generated the economic data one can attempt to specify and estimate an underlying statistical pattern in the data of the sample to make statements about the **data generating process**. It is this which constitutes the quantitative knowledge regarding the domain of application.

The development of econometrics, however, became unbalanced and leaned more towards mathematical theories and models that were divorced from reality. Fortunately econometrics today has another entirely different meaning and purpose that will amend this lack of relevance. Econometrics today is known simply as knowledge based on quantitative economic data and its analysis. It is this knowledge that businesses are seeking to exploit so as to gain an edge over their competitors. In the new age of digital on-line information firms gain competitive advantage by leveraging knowledge gained from data, such as optically scanned bar code data at the point of sale, or information on peoples' socioeconomic background and their preferences that can be mined from social network data. According to Davenport, the **business analytics** Guru, business analytics (econometrics), is about using '…sophisticated data-collection technology and analysis to wring every last drop of value from all your business processes' (Davenport, 2006: p. 1).

There is a value chain from information to knowledge, and knowledge to decision making based on that knowledge. Econometrics deals with processing information to filter out noise and redundant information and to discover patterns in the information so gathered. It is this **pattern recognition** that constitutes knowledge in business analytics. This book is about: (i) exploratory data analysis (including processing of raw data, data reduction, and data classification), and pattern recognition that separates signals from noise (**signal extraction**)[1]; (ii) model building and choice between alternate models; and (iii) prediction or forecasts based on the selected and estimated or calibrated model, along with probabilistic statements on the credibility of those predictions and forecasts. All of this is achieved through mathematical models, statistical theories, and numerical computations. These are the constituent molecules this book is made of.

It is estimated that in the year 2007 the average daily volume of traditional market transactions (spot, forward, and swaps) in foreign exchange markets globally was $3.21 trillion. For the same year the foreign exchange derivatives markets recorded an average daily turnover of $2.1 trillion. It is estimated that the ten most active traders account for almost 73% of the trading volume. One can imagine the importance of the econometric modeling of the foreign exchange markets for these ten traders, as well as for the other traders who wish to encroach on the privileged territory of these ten by leveraging the knowledge gained from econometric models. While the economic theory of **efficient capital markets** postulates that stock prices follow a **random walk**[2] it is also clear that there are asymmetries in information and knowledge available to the investors and asset management companies. Such asymmetries can lead to value addition for those who have better knowledge of how the market behaves under such asymmetries. Various asset management companies manage several thousands of individual accounts of corporations and families, each account maintaining a longitudinal database pertaining to the account holders' characteristics, their preferences, and how their portfolios performed in the market over time. This extensive database is not fully exploited for the knowledge inherent in that information. To quote T.S. Elliott, 'Where is the knowledge lost in the information?' Asset management companies can do better by using a better knowledge extraction from such information. Retail sales data, collected all over the world using bar codes and optical scanners and computers, includes data on consumer preferences.

Consumer feedback received by customer service departments has some textual information. This information can be exploited through data mining and **text mining** to gather knowledge on product quality, individual preferences, and individual willingness to pay so as to improve product design and advertising strategies. Click stream data stored on servers offer excellent information on peoples' preferences regarding the products and services available on the worldwide web. Using knowledge from that data one can do target marketing to improve sales, as is being done by Amazon.com.[3] These

[1] All these constitute what is now called data mining. Data mining is a specialized subject and is not discussed in detail here. Our discussion here is limited to some preliminary methods of pattern recognition through correlations and scatter plots. However, in Chapter 9 we discuss the methods of factor analysis and principal component analysis which form a part of data mining. For a more detailed description of data mining one may see Section 1.1 of the electronics references for Chapter 1 and Han and Kamber (2006).

[2] A random walk is achieved by taking successive random steps.

[3] Michael Lynch, a Bayesian with a Ph.D. degree in Engineering from Cambridge University, exploited his computational and statistical skills to discover patterns in a variety of digitized data such as text, visual, and numerical data employing Bayesian principles of searching for useful information. He is regarded as Bill Gates of Europe and started Autonomy Inc, a company specializing in meaning-based computing in 1996.

are just a few examples of the application of econometrics in business. More examples mentioned subsequently in this chapter will substantiate this point.

1.1.2 Econometrics and scientific credibility of business and economic decisions

There is an increasing tendency among business firms to base their decisions on credible knowledge. Knowledge derived from an arbitrarily chosen model, however scientific the subsequent statistical analysis might be, suffers from a lack of credibility to the extent that the basis for choosing the model is not made explicit and defended. Credibility of knowledge is being judged by scientific approaches used in generating knowledge, including evaluating the performance of alternate models and testing the chosen model. Scientific credibility is to be achieved through objectivity, reproducibility, testability or falsifiability, efficiency in use of information, closeness to reality of the assumptions made and results obtained.[4] There are two kinds of information, beneficial information or signalling information, and non-beneficial information or noise. The first question is: 'Is all beneficial information used?' The second question is: 'Is all available information classified into beneficial and non-beneficial information?' The third question is:'What is the knowledge gathered or pattern discovered from the beneficial information?'

Not all information is in the form of quantitative information of comparable quality. When information comes from sources of different quality or reliability, scientific credibility calls for the best way of pooling such information. When information needed for analysis is not available one may have to collect it, if the resources permit it, or obtain it using some proxy variable. Alternately, one may obtain the value of such a crucial unavailable variable by eliciting its likely value from experts. Not using such a relevant variable in a model, as data on it were not available, is equivalent to ignoring that variable altogether! The question then arises as to what is the most credible way to combine such subjectively ascertained information with objectively collected information. **Bayesian analysis** deals with this method of credible ways of combining the subjective non-sample information with sample information. Bayesian analysis is explained in some detail in Chapter 10. Ultimately we must try to extract the maximum possible credible knowledge from *all* the available and *useful* information.

Most situations in economics call for using data generated by either a designed **random experiment** or a **sample survey** or a naturally occurring economic process that is viewed as a random data generation process. Scientific credibility in the former two cases can be established through design of economic experiments using the statistical theory of **design of experiments**, and design of sample surveys. In the third type of non-experimental situation credibility of econometric knowledge depends on how convincing the model is in reflecting the truth of the underlying data generating process. Where the sample data used is not from a random experiment or a random sample we need a much greater degree of effort to establish statistical credibility for modeling. To summarize: achieving credibility through pattern recognition is the essence of this book. We might quote the famous Indian poet and Nobel Laureate in literature, Rabindranath Tagore, who wrote this to inaugurate the launching of *Sankhya*, the Indian journal of statistics:

[4] Carl Sagan (1996) calls the scientific method the most effective 'baloney detector' ever invented.

> The enchantment of rhythm is obviously felt in music, the rhythm which is inherent in the notes and their groupings. It is the magic of mathematics, this rhythm, which is in the heart of all creation, which moves in the atom and in its different measures fashions gold and lead, the rose and the thorn, the sun and the planets, the variety and vicissitudes of man's history. These are the dance steps of numbers in the arena of time and space, which weave the maya of appearance, the incessant flow of changes that ever is and is not. *What we know as intellectual truth, is that also not a perfect rhythm of the relationship of facts that produce a sense of convincingness to a person who somehow feels that he knows the truth?* We believe any fact to be true because of harmony, a rhythm in reason, the process of which is analysed by the logic of mathematics. (*Sankhya*, Vol. 2, Part 1, Page 1, 1935. Emphasis in italics is by the authors).

1.2 Types of economic problems, types of data, and types of models[5]

1.2.1 Experimental data from a marketing experiment

Practical situations often arise where the questions that are of interest to us are such that there are no data that are actually available to answer the questions. We may have to generate the required data. We give a simple example. A coffee powder manufacturer would like to design a packaging and pricing strategy for the product that maximizes his revenue. He knows that using a plastic bag with color has a positive effect on the consumer's choice, while a colored plastic bag is more costly than a plain plastic cover. He needs to estimate the net benefit he would have in introducing a colored plastic bag. He also knows that consumers prefer to have fresh coffee powder and thus depending on the weekly rate of consumption they choose the size of the packet. The larger the size of the packet that a household wants the lower is its willingness to pay, but smaller packets will increase the cost of packaging. He would like to know what would be the net benefits to the firm of different sizes of the packets at different levels of prices he could fix for them given different types of demand.

To introduce more realism and more complexity let us assume that there is a cost saving coffee substitute, called chicory, that when mixed with coffee brings thickness and bitterness to coffee that some people may like. But too much chicory is not liked by many consumers. As a result the manufacturer expects that the greater the content of chicory the lower the price the customer is willing to pay. Are consumers willing to trade a part of their preference for colored plastic bag for the optimal size of the packet? Historically collected data on coffee sales may be of no use to answer these questions as colored plastic bags were not used in the past. The manufacturer cannot go ahead and introduce the new colored package incurring higher cost. The coffee manufacturer wishes to conduct a small-scale pilot **marketing experiment** to estimate the effects on net revenue of different types of packaging, different levels of chicory and different sizes of the packets. How should one conduct the experiment? How should one analyze the data collected through such an experiment? Designing economic

[5] The authors have some managerial and statistical consultancy experience. They are also fortunate to have students with industry experience who bring practical problems to the classroom for discussion. The examples given below are drawn from such teaching and consultancy experience.

Table 1.1 The kind of data from marketing experiment.

Design Point	Chicory Content	Size of Packet	Colour of Packet	Mean Price (Customer's Willingness to Pay for 100 grams)
1	L(−)	L(−)	L(−)	15
2	H(+)	L(−)	L(−)	14
3	L(−)	H(+)	L(−)	12
4	H(+)	H(+)	L(−)	10
5	L(−)	L(−)	H(+)	17
6	H(+)	L(−)	H(+)	15
7	L(−)	H(+)	H(+)	16
8	H(+)	H(+)	H(+)	15

experiments and their analysis has become a new econometric tool widely used in recent years. Data in Table 1.1 summarizes the kind of data obtained for one such marketing experiment when each of the factors is set at two levels labeled Low (L) and High (H) for chicory content of 10%, size of packet 100 gms and 200 gms, plain cover and colored cover.

The questions of interest are: 1. How to choose the factors and assign them to the experimental subjects of the pilot experiment? 2. How do the changes in the three factors affect people's willingness to pay for 100 gms of coffee powder? 3. Is the relation between these factors and willingness to pay linear or nonlinear? 4. How can we estimate the effects? These questions can be answered using the statistical theory of design of experiments and the statistical method of **analysis of variance**. The first question is discussed in specialized texts on the design of experiments (see Anderson and Whitcomb (2000) for details on how to design **factorial experiments**).[6] The rest of the topics on the statistical analysis of experimental data are discussed in greater detail in Chapter 3 and Chapter 9.

1.2.2 Cross-section data: national sample survey data on consumer expenditure

The National Sample Survey Organization of India conducts nation-wide sample surveys of households to record their consumption expenditure pattern. This is a very rich database that was initiated to aid Indian planners to plan economic development. It is now an excellent data base for understanding consumer behavior in India in order to develop retail marketing strategies. The data is now made available at a reasonable cost and is at the household level, by means of a fine grid of geographic strata both in the rural and urban regions of India. One can delineate market areas and for each such a market estimate the consumer demand patterns. Typical information available from the NSSO database is presented in Table 1.2. Data presented in the table are only representative of the original data.

[6] In the field of marketing research there are very sophisticated designs of experiment. These are intended to meet the needs of a marketing researcher so as to perform a choice based conjoint analysis. For example, see Raghavarao, Wiley, and Chitturi (2011).

Table 1.2 National sample survey data on consumer expenditure (representative).

Sl. No	State	Region	Sub Sample	House hold Number	Multiplier	House hold Size	Cereal Consumption		Total Expenditure
							Quantity	Value	
1	2	1	1	1	13291	1	14.5	184	1023
2	2	1	1	3	13291	4	56	672	1744
3	2	1	1	5	13291	3	42	500	1476
4	2	1	1	7	13291	5	70.5	846	2630
5	2	1	1	1	2492	4	54	800	3760
6	2	1	1	3	2492	1	14.5	184	1028
7	2	1	1	1	35165	5	45.75	553	2145
8	2	1	1	3	35165	3	40	480	1068
9	2	1	1	5	35165	4	40.5	406	1536
10	2	1	1	7	35165	4	43	559	2860
11	2	1	1	9	35165	6	73	1062	4608
12	2	1	1	11	35165	5	60	550	1875
13	2	1	1	1	4786	2	14	192	2474
……	2	1	1	3	4786	2	27.5	405	10008
N	2	1	1	1	10608	3	35	362	1695

Source: Unit level data from National Sample Survey Organization, Government of India, used in a research study on consumption deprivation reported in Kumar, Mallick, and Holla (2009).

Sample surveys such as this are usually multi-stage stratified samples, giving different weights to different strata. Unit level data such as these cannot all be regarded as equivalent, ignoring the different over- and under-sampling of strata. The column labeled multiplier gives the weight one must attach to each observation to convert it into what would have been the case, if the sample was a simple random sample that gives an equal chance for every sampled unit to be included in the sample. These multipliers are derived from the sample design chosen. Sampling is a specialized topic and one may see Thompson (2002) for the details. These multipliers must be used as weights for the recorded observations before any modeling is attempted. Given this sample information one might want to know (i) if there is any pattern implied by the theory of consumer behavior that relates expenditure on cereals to household size and total expenditure; (ii) if such a relation is linear or nonlinear; (iii) how to estimate alternate specifications; and (iv) how to choose between alternate specifications. This type of data is called cross-section multivariate data. The data analyses of such cross-sectional data will be discussed in detail in Chapters 2–5, 9, and 10.

Another recently popular way of generating data for analysis is through web surveys. Before using such data for analyzing the underlying pattern one must make sure whether the analysis pertains to only that sample of respondents and their behavior or refers to a wider population of which the web survey is only a sample. If the latter is the case one must determine the probability that a unit is selected for web survey, and the probability that a selected unit responds. Based on these two probabilities one must make a sample selection correction. In order to have a credible model this kind of data adjustment must be made before modeling.

1.2.3 Non-experimental data taken from secondary sources: the case of pharmaceutical industry in India

An advertising company noted that the pharmaceutical industry is poised for rapid growth in India owing to several factors such as switching over to a new product patenting regime, economic reforms that permitted foreign direct investment, low cost of doing research and development work in India, and the large pool of scientific and technical manpower that exists in India. It wanted to make a pitch for new customer accounts from some of the major pharmaceutical companies. It examined the data on sales and advertisement expenditure and wished to demonstrate that advertisement expenditure pays rich dividends in terms of generating a substantial increase in sales. The data the agency collected from an industry database, such as PROWESS from the Centre for Monitoring the Indian Economy, is presented below in Table 1.3. The figures quoted in the table are in Rs Crore (Rs. 10 million) per year. The advertising agency found a simple relationship between advertising expenditure and sales and argued in favour of spending on advertising. The marketing and **supply-chain** manager of the company argued that the results demonstrated by the advertisement agency referred to all the pharmaceutical companies in India, while they themselves were different from the typical average pharmaceutical company. He also said that sales were also affected by marketing effort and through supply-chain management, of which the distribution expenses were a proxy. He thus said that he was not convinced that the effect of advertising on sales *in his company* was what was suggested by the advertising company. The issues to be examined are: 1. Is the effect of advertising on sales the same for all companies in the database? 2. Do all companies in the database have the same structural pattern so as to be treated as one sample? 3. What are the various drivers of sales? 4. What is the most plausible

Table 1.3 The data collected from an industry database (PROWESS).

Sl. No	Company Name	Sales	Advertising	Marketing	Distribution
1	A C E Laboratories Ltd.	133.02	0.28	1.04	0.71
2	Aarey Drugs & Pharmaceuti	–	–	–	–
3	Aarti Drugs Ltd.	230.79	0.46	2	6.86
4	Abbott India Ltd.	274.17	4.29	39.28	6.65
5	Add-Life Pharma Ltd.	0.46	0	0	0
6	Adinath Bio-Labs Ltd.	12.02	0.01	0.06	0
7	Aditya Medisales Ltd.	0	0	45.38	0
8	Advik Laboratories Ltd.	15.25	0	4.68	0.48
9	Aesculapius Remedies Ltd.	–	–	–	–
10	Ahlcon Parenterals (India	21.13	0.15	0.05	0.4
11	Ajanta Pharma Ltd.	101.62	1.6	10.08	4.81
12	Albert David Ltd.	124.41	0	30.32	0
13	Alembic Ltd.	608.66	0	57.39	10.87
14	Alpha Drug India Ltd.	22.7	0	0.23	0.4
15	Alta Laboratories Ltd.	12.28	0	0.01	0.01
16	Ambalal Sarabhai Enterpri	120.44	0	3.4	0.46
17	American Remedies Ltd. [M	94.3	6.84	1.18	1.52
18	Amit Alcohol & Carbon Dio	20.25	0	0.32	0
19	Amol Drug Pharma Ltd.	2.49	0	0.04	0

Source: Company level data extracted from PROWESS: A company level data base of the Indian economy from the Centre for Monitoring the Indian Economy (CMIE).

functional form for the multivariate relation between sales and these drivers? 5. How does one estimate the separate effect of each of these factors on sales? These questions can be answered using the multiple regression methods for cross-sectional data, discussed in Chapters 2–5, 9 and 10.

1.2.4 Loan default risk of a customer and the problem facing decision on a loan application[7]

When a customer submits an application to a bank for a loan he or she provides personal information in the application, and the person's case is then referred by the bank to a credit rating agency to get a credit rating based on his or her credit history. These two sets of data are used by the bank to determine the credit risk. The bank wishes to examine the past history of several such loan applications and the loan default histories in order to develop a risk score – the probability of default on a loan given the personal information and the information from the credit rating agency. It would also be interested in examining the effects of choosing different thresholds of credit risk score for rejecting the application.

[7] This example is based on a term paper submitted for an econometrics course at Indian Institute of Management-Bangalore, in March 2008 by Abhishek Agarwal, Amit Gupta, Dhilip Krishna, and S. Karthik.

1.2.4.1 Some data mining issues[8]

The actual data may pertain to several thousand applicants, and not all of them are similar. There can be information on more than a hundred variables. Actual data provided by the applicants could be of two types, one that can be easily verified with supporting documents and the other, that cannot be easily verified. One may regard some of those variables as variables that have some information on the default risk of the applicant and hence are signaling variables, while there are other variables that have no such information on loan default risk and hence are noisy variables. There may be some missing observations and there can be recording errors.

The first job of an analyst in this case is to clean the data for errors and decide on how to treat the missing data. If data were missing on one variable to throw away the entire observation is an inefficient way of using sample information. Another recommended procedure for replacing the missing value by means of a sample consisting of all non-missing values is also not an efficient way of using the information. One may instead replace the missing value by some kind of an appropriate mean. One way of doing it is to take all the observations that have no missing values and arrange them into data clusters with default risk being of 20 intervals between 0 and 1. Then one can arrange the missing value sample into similar clusters with the default risk being in the same 20 intervals. The missing values in each of these 20 clusters may then be replaced by the mean values observed in a matching cluster of default risk for the earlier sample that had no missing values.

The variables that have contributed to very little variation in default risk can be treated as noisy variables and dropped. The remaining variables can be treated as the signaling variables. Even then the number of variables could be too large, about 100, giving rise to difficulties in estimation due to correlations among such a large number of variables. This issue of problems associated with high correlations among the independent variables is discussed in detail in Section 2.4 of Chapter 2. The number of variables can be reduced through data reduction techniques such as **principal component analysis** discussed in detail in Chapter 9. Finally the model chosen must be the one that is best suited to dealing with binomial variable, default or no default. This is a special case of regression with a categorical dependent variable discussed in detail in Chapter 4. The data of the loan default example is provided in the Electronic References. Two alternate models were evaluated in terms of their performance in predicting the default risk with the historic data.

1.2.5 Panel data: performance of banks in India by the type of ownership after economic reforms

Several interesting questions arise with respect to the banking sector in India as a result of the financial sector reforms: 1. Do the private sector banks perform better than the public sector banks? 2. Are the public sector banks improving their performance relative to the

[8] Most of the issues mentioned here regarding the cleaning and editing of data and exploratory analysis are issues dealt with in data mining. Although data mining is not covered in detail in this book, given its importance we give some description of it in this example and in section 1.3. Section 1.1 of the Electronic References to this chapter and Chapter 9 provide some additional details about data mining. This book deals mostly with pattern recognition or statistical modeling of data that are already pre-cleaned and explored through exploratory data analysis. The reader may see Han and Kamber (2006) for more detail on data mining.

Table 1.4 Data of performance of banks in India.

Year	Panel	Size	ROA	NPA RATIO	Op Profit Ratio	CAR	Ownership
1999	_BOB	10.86	0.0066	7.70	1.95	13.30	0
1999	_BOI	10.90	0.0037	7.30	1.41	10.60	0
1999	__BOP	7.66	0.0153	3.70	2.36	14.60	1
1999	_BOR	8.16	–0.0202	9.50	–0.31	0.80	1
1999	_CUB	7.26	0.0087	8.00	1.79	14.30	1
1999	__CB	9.61	0.0117	2.00	2.28	13.20	0
1999	_DB	9.61	0.0055	7.70	1.59	11.10	0
1999	_DLB	7.23	0.0028	12.30	1.07	10.10	1
1999	_FB	9.00	0.0003	7.50	0.69	10.30	1
1999	_HDFC	8.38	0.0189	0.70	3.52	11.90	2
1999	_ICICI	8.85	0.0091	2.90	2.57	11.10	2
1999	_ISB	8.73	0.0056	7.20	2.29	15.20	2
1999	__JKB	8.93	0.0114	3.80	2.74	24.50	1
1999	_OBC	9.84	0.0123	4.50	2.30	14.10	0
1999	_SIB	8.18	0.0007	11.10	1.08	10.40	1
1999	_SBBJ	9.23	0.0090	10.50	1.79	12.30	0
1999	_SBI	12.31	0.0046	7.20	1.55	12.50	0
1999	_SBM	8.84	0.0049	10.60	1.79	10.20	0
1999	_SBT	9.29	0.0040	10.80	1.30	10.30	0
1999	_UTI	8.27	0.0069	6.30	2.14	11.60	2
1999	_UWB	8.27	0.0095	8.30	2.06	11.60	1
2000	_BOB	10.98	0.0086	7.00	1.93	12.10	0
2000	_BOI	10.93	0.0031	8.60	1.24	10.60	0

(ROA: return on Assets; NPA Ratio: Ratio of non-performing assets to all assets; Op Profit ratio: Operating profit divided by non-operating profit; CAR: Capital Adequacy Ratio, Ownership = 0 for scheduled public sector bank, = 1 for scheduled private sector bank, = 2 other kind of bank).

private sector banks after the introduction of financial sector reforms? 3. Is the performance of all banks improving after the introduction of financial sector reforms? In order to answer these questions one may acquire data from the official source, the Reserve Bank of India. Table 1.4 presents the data so collected. Complete data are provided in the Electronic References.

There are several public sector banks while there are only a few private banks. The data on banks' economic operations are available for several years. The data thus consists of a time series of cross-sections or is **panel data**. Regression models for such panel data have some special characteristics of their own and ordinary multiple regression models must be suitably modified so as to address the special features of the data. The statistical modeling of panel data using the **Stochastic Frontier Model** is discussed in Chapter 9 and Chapter 5, and using the **Self Modelling Regression Model** is discussed in the Electronic References for Chapter 5.

1.2.6 Single time series data: The Bombay Stock Exchange (BSE) index[9]

One of the areas where quantitative analysis has been used extensively in recent years is the field of finance. In one of its basic forms the efficiency hypothesis of the capital markets assumes that stock prices follow a random walk model. The six year daily BSE Index (Bombay Stock Exchange Index) data from April 2, 1996 until March 31, 2002 was used by Singhal (2005) to test this hypothesis. This data set is univariate time series data. Many financial time series come like this. Financial institutions require an econometric analysis of such a financial time series. Statistical analysis of univariate time series can be carried out if one can either assume that the series is stationary which means that the series has the same mean, variance, and other higher moments in different segments of time, or if one can find a deterministic transformation of the nonstationary series that will make it stationary. Modeling of time series is taken up in Chapters 7 and 8. Chapter 7 in particular deals with modeling a single time series or a univariate time series that is stationary, while Chapter 8 deals with multiple time series and nonstationary time series. If one plots the closing values of the BSE Sensex on a particular day against the closing value on the previous day in a scatter plot, the scatter does seem to confirm the random walk hypothesis. This is shown in Figures 1.1 and 1.2.

Figure 1.1 can be shown in Data Analysis and Statistical Computing (DASC for short) software by clicking the menu items just three times. Readers can substitute their own data or modify the data given in our example to gain experience with DASC and with this kind of example. The detailed method can be seen in Electronic References for this chapter.

Software for DASC and the Electronic References can be downloaded from the website http://public.whut.edu.cn/slx/English/Login1.htm.

We note that there are two pictures in Figure 1.1 which are drawn simultaneously by DASC. The user can select one of the two pictures to save. In fact, there are two figure systems in DASC for all models, but we will show only one of them in subsequent paragraphs.

Figure 1.2 above plots the daily difference in BSE Sensex against time. The raw data in the figures above give one the impression that the stock prices do follow a random walk and that there is little one can do to make gains in the stock market, contrary to the gains many people do make on the stock market. One very common problem with many econometric analyses is that they tend to model the series as given. The given data may have considerable noise built into them and it may be necessary to smooth the series through some kind of averaging so as to discern the patterns that might exist. This is illustrated by this example. It will be shown a little later in this chapter that a detailed exploratory data analysis using such averaging does provide a scope for making short-term gains in the Indian stock market through a strategy.

1.2.7 Multiple time series data: Stock prices in BRIC countries[10]

Four countries, Brazil, Russia, India, and China, nicknamed the BRIC countries, are gaining importance as possible destination countries for portfolio investment by investors in countries that had a head start in industrial development. Two economic questions are

[9] This example draws from Singhal (2005). We are grateful to Paras Singhal and the editor of the *IIMB Management Review* who provided us the raw data used by Singhal.

[10] This example is taken from the term paper submitted for an econometrics course at Indian Institute of Management-Bangalore in 2008 by Akash Agrawal, Hrishikesh Patil, Udayan Sarkar, and Vikram Balan.

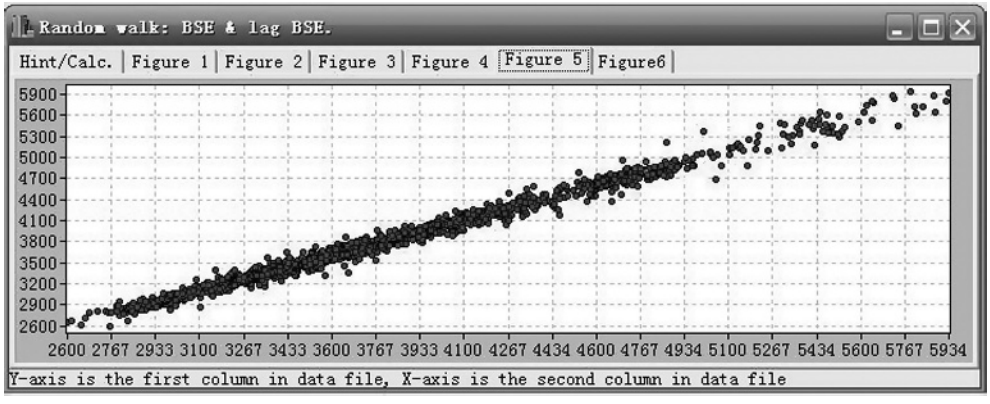
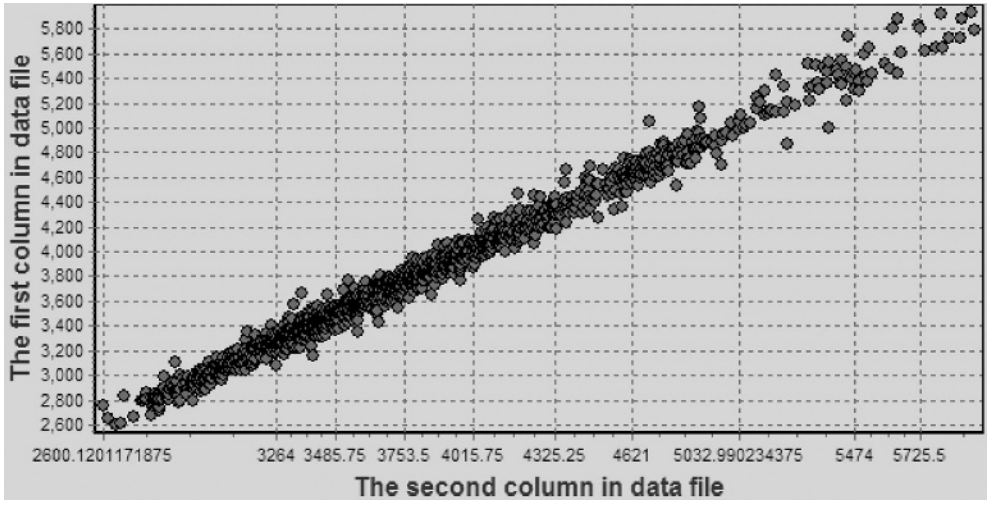


Figure 1.1 (a) Random walk: BSE & lag BSE.



(b) Random walk: BSE & lag BSE.

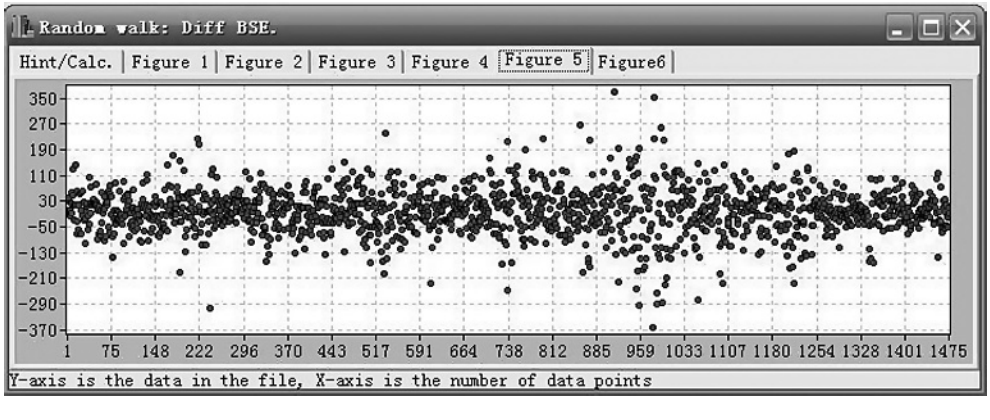


Figure 1.2 First Difference of BSE Index.

important in this context. First, are the stock markets in these four countries integrated with the stock markets of other industrially advanced countries? Second, how are the stock prices in these countries linked causally to those of other advanced countries? The data needed to answer these questions are time series data on stock price indices in these four countries and in other advanced industrialized countries. The data collected were the following weekly stock price indices: US (NYSE-100), UK (FTSE-100), Japan (Nikkei-225), India (BSE-Sensex), Brazil (Bovespa), China (SSE composite), and Russia (RTS). Modeling multiple time series is needed to answer the questions raised above. This subject is covered in detail in Chapter 8.

The study reported in the Electronic References for Chapter 8 reveals that whether the markets are integrated with advanced countries' markets or not depends on the period of study. The study shows that the Indian stock market is better integrated with the US and UK markets than those of the other BRIC countries. The Indian market is not integrated with the Japanese market. The statistical model and its analysis developed in Chapter 8 not only answer the two questions mentioned above, but also tell us what would be the impact on the stock prices in India if there was a shock to the NYSE100.

1.3 Pattern recognition and exploratory data analysis

1.3.1 Some basic issues in econometric modeling

In physical sciences the experimental data refer to observations from controlled experiments referring to a physical world that does not change much. In social sciences one deals with data generated by a non-experimental situation and refers to an ever-changing social environment with a lot of individual interaction and variation.[11] It is difficult to establish any universally applicable laws. One must determine, from the non-experimental data, the pattern that best fits the data for that social situation which generated the data. Let us illustrate the basic issues arising in such models, using the most commonly used econometric tool **regression,** and also with the simplest of such regression models, linear regression with one or more independent variables. We take observations from the independent variables $(X_1, X_2, ..., X_k)$ and the dependent variable (Y) and would like to determine a quantitative relationship between them that is best in some sense. We assume that the variables have a **joint probability distribution** and that the dependent variable has a **conditional probability distribution** given the independent variables. The regression model is supposed to be the conditional mean of the dependent variable given the independent variables.

The issues confronting the analyst in this situation can be summarized as:

(1) Should one use the raw data as given or should one use processed or derived (smooth) data?

(2) Do the observations come from the same population? Or does the sample seem to come from a mixture of two or more populations?

[11] This is what must have prompted Murray Gell-Man, a Nobel Laureate in physics, to remark: 'Imagine how hard physics would be if electrons could think' (quoted by Page, 1999).

(3) What transformation of variable X should one use? Linear in X, piece-wise linear in X, non-linear function of X, or Nonlinear in X with non-linearity appearing in parameters?

(4) Should one give equal or unequal importance to all the observations in the minimization of errors?

(5) If there are several possible models, how should one choose one from among them?

(6) Finally is the chosen model good or should one look for additional information?

The question we may ask is 'If our interest is the **conditional mean** of the distribution of the dependent variable, given the independent variables, what should be the most appropriate model we choose for it?' A model most appropriate with the entire sample may not be the one that is most appropriate if one is interested in a portion of that sample. The answer of course depends on what use we put the model to. If we want to explain the observed data, including the extreme values, we may include all observations in the sample. Even then the same pattern of relation may not fit all sections of the distribution of the dependent variable Y. If we are more interested in explaining the middle portions of the distributions of the variables we can use the standard multiple regression models discussed in detail in Chapters 2–5. If we are interested in different segments of the sample then fractile regression discussed in Chapters 9 and 10 will be useful.

One can say that whatever is the regression model such model can be regarded as a signal or pattern that we are trying to discover, and the rest is noise. The criterion for the best fitting model is maximizing the signal and minimizing the noise or maximizing the **signal to noise ratio** as the communication engineers say.[12] Thus, if there are alternative models the choice between them should be made using this criterion. In Chapter 10 we describe in greater detail how this is done. It is also possible that the same model or pattern may not fit equally well with all data points in the chosen sample. Different portions of the sample may have different patterns.

The application of statistics must give importance to an understanding of the phenomenon to which the statistics are applied. Hence statistical modeling necessarily requires an understanding of the domain of application that generated the data, economics in this case. In any model building we would encounter two types of drivers that determine the dependent variable. First, there are those factors that are quite general to the domain area and are suggested by the existing theories in the domain area, and others which are specific to the particular or specific situation that actually generated the data. The knowledge of those specific factors that affect the dependent variable must come from a thorough examination of the sample data itself. That is what we call **exploratory data analysis**. Exploratory data analysis is a special and important component of data mining. Again, as our focus is more on pattern recognition or statistical modeling or what is also called **predictive analytics** in business analytics, we cannot dwell much on data mining. However, given its importance, we are compelled to cover some basic features and refer the reader to the data mining book referred to earlier. Exploratory data analysis must precede identifying possible alternate models.

[12] It is said that Florence Nightingale, who was made an honorary member of the American Statistical Association, said that a statistician's work is the work of discovering God. To elaborate, by reducing our ignorance through statistics we improve our knowledge and get closer to the truth.

1.3.2 Exploratory data analysis using correlations and scatter diagrams: The relative importance of managerial function and labor

One might trace the origins of econometrics to exploring the quantitative relations between economic variables using correlations and **scatter diagrams** (Frisch, 1929).[13] Frisch suggested looking at all possible pairs of variables and drawing the scatter diagrams and calculating the correlation coefficients so as to understand the relations between variables. We would like to illustrate this with an example. A company was facing a situation where the workers' union was demanding a productivity-linked bonus year after year, attributing the increase in profits to their hard work. The management undertook a study of the relation between profits after taxes on three other variables. These were: 1. labour productivity, measured as output per unit labour; 2. managerial effectiveness, measured through a scale based on a battery of questions put to workers, managers, and managerial professionals outside the company (on the number of managerial decisions and their perception of whether they made any significant positive or negative impact on the company); and 3. cost of raw materials. The aim was to determine the best fitting statistical relationship between profit after taxes and the other three variables. Here, we are in search of a function that is linear in parameters and possibly involving nonlinear functions of the three explanatory variables that maximizes the explained variation in profits after taxes. As the regression coefficients of the linear regression model are related to the correlation and partial correlations, we can examine scatters and correlations to explore what model is to be chosen. Figures 1.3a to 1.3c provide the scatter plots of the three variables with profit after tax.

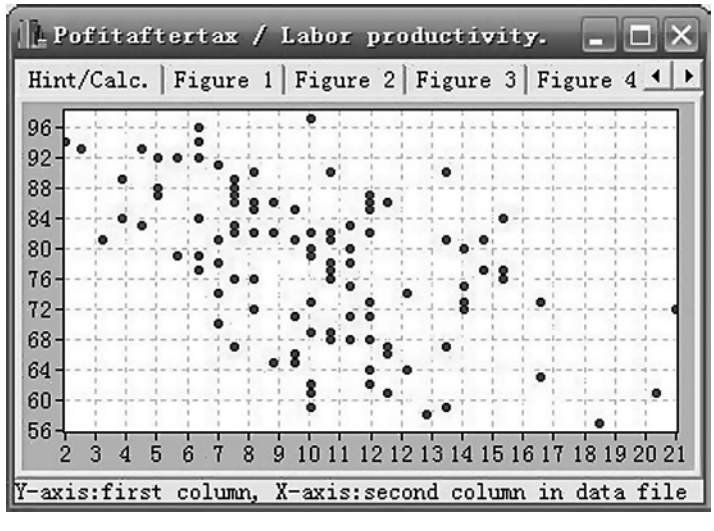
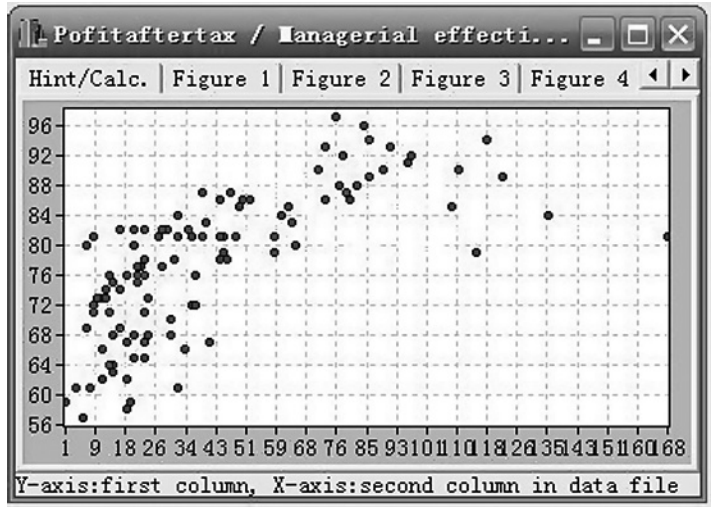
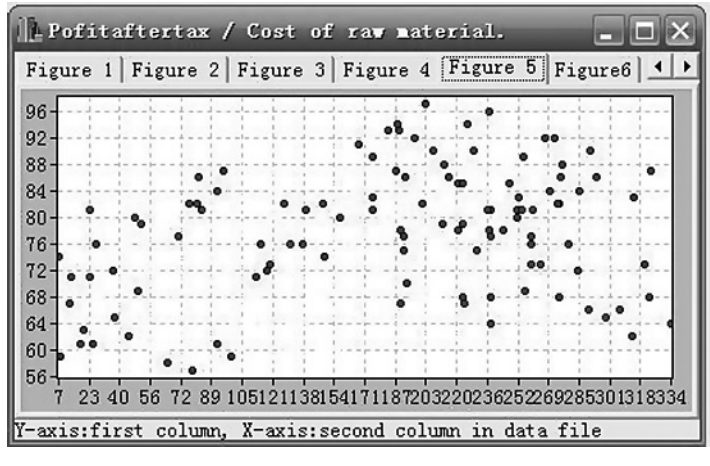


Figure 1.3 (a) Profit after tax/Labor productivity.

[13] Ragnar Frisch is one of the founders of the International Econometric Society that was founded in 1930. He is the one who coined the word econometrics. His first work on econometric, Frisch (1929), outlined the usefulness of such data exploration in determining the true structure of the data generating process.



(b) Profit after tax/Managerial effectiveness.



(c) Profit after tax/Cost of raw material.

Figures 1.3a to 1.3c can be shown in the DASC software. From these scatter diagrams we get the impression that profit after tax is positively related to managerial effectiveness and negatively related to labour productivity, and possibly not related to cost of raw material. From these scatter plots it is also apparent that managerial effectiveness has a nonlinear relationship with profit after tax. We now re-express the relationship through a scatter with the log of managerial effectiveness, and the square root of managerial effectiveness. We find that the scatter with log managerial effectiveness is still exhibiting nonlinearity and seems to indicate a quadratic relation.

We plotted the scatter with the square of the log of managerial effectiveness. The scatters of the square root and the square of the logarithm seem to be quite similar and good

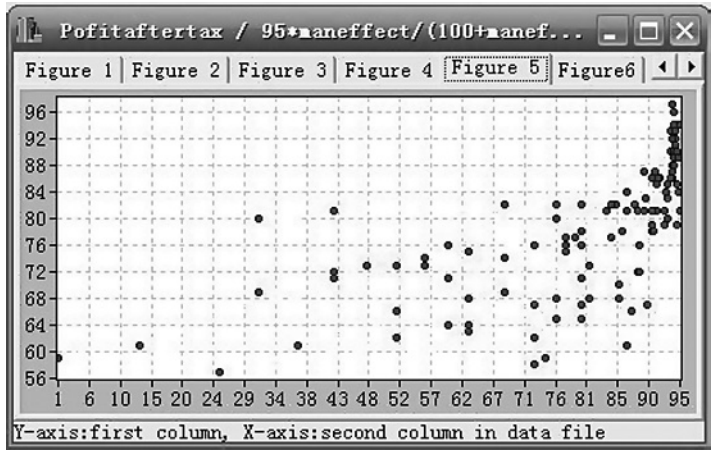


Figure 1.4 Profit after tax/95*maneffect/(100+maneffect).

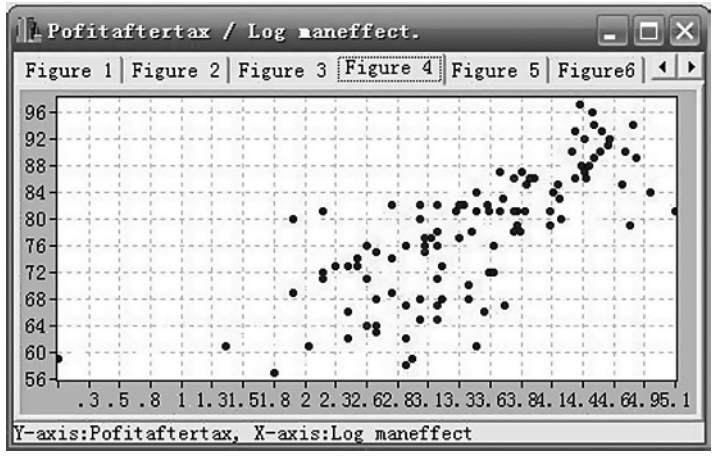


Figure 1.5 Pofitaftertax/Log maneffect.

suggesting that we should try these two re-expressions. These scatters are presented in Figures 1.4 to 1.7.[14]

Figure 1.7 is similar to the previous figure in visual appearance, but their *X*-axises are not the same.

We then calculated the **zero-order correlations** between profit after tax and these re-expressions of managerial effectiveness and the other two variables, and these are presented in Table 1.5. From this table it is clear that the square of log managerial effectiveness has the highest correlation with profit after tax. While the correlations of other variables are also significant we observe inter-correlations between them. So we wish to know if the other two variables are important after introducing square of log of managerial effectiveness as an

[14] Figure 1.4 shows the scatter with respect to a nonlinear transformation of variable maneffect = 95*maneffect/(100+maneffect).

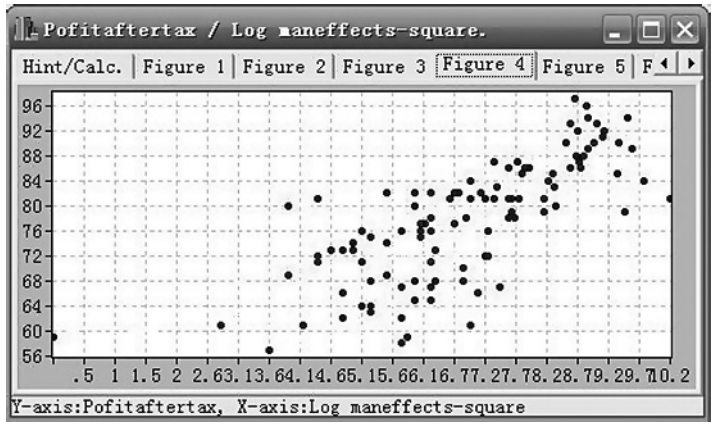


Figure 1.6 Pofitaftertax/Log maneffects-square.

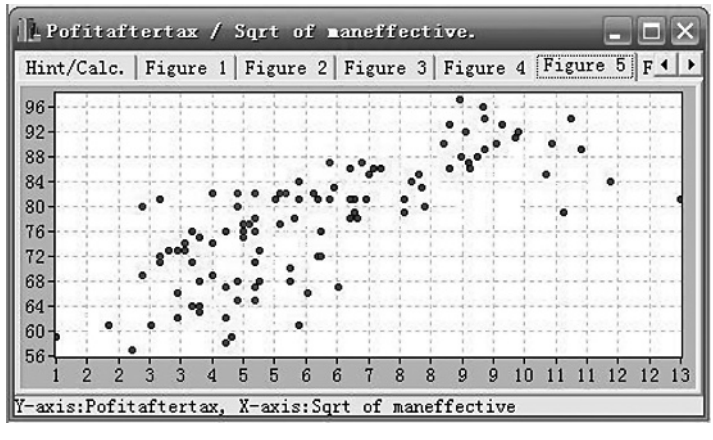


Figure 1.7 Pofitaftertax/Sqrt of maneffective.

explanatory variable. To answer this question we calculate the **partial correlations** after controlling for square of log managerial effectiveness.

These partial correlations are shown in Table 1.6 below.

From this table it is clear that labor productivity is the next most significant variable and that the cost of raw material is possibly not important. However, economic reasoning would suggest that the cost of raw materials must be an explanatory variable for profits after tax.

We are now ready to specify the **regression model** as:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \tag{1.1}$$

where X_1 is variable Ln Managerial effectiveness square, X_2 is variable Labor productivity, X_3 is variable Cost of raw material, and ε is the random errors. The results of the least squares estimation of the regression above are presented below in Table 1.7.

Table 1.5 Correlations between variables.

		Profit after tax	Labor productivity	Cost of raw material	In man effect sq	Sqrt man effect
Profit after tax	Pearson Correlation	1	−.497**	.294**	.755**	.746**
	Sig. (2-tailed)		.000	.002	.000	.000
	N	111	111	111	111	111
Labor productivity	Pearson Correlation	−.497**	1	−.127	−.602**	−.609**
	Sig. (2-tailed)	.000		.183	.000	.000
	N	111	111	111	111	111
Cost of raw material	Pearson Correlation	.294**	−.127	1	.421**	.403**
	Sig. (2-tailed)	.002	.183		.000	.000
	N	111	111	111	111	111
In man effect sq	Pearson Correlation	.755**	−.602**	.421**	1	.996**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	111	111	111	111	111
Sqrt man effect	Pearson Correlation	.746**	−.609**	.403**	.996**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	111	111	111	111	111

**Correlation is significant at the 0.01 level (2-tailed).

Table 1.6 Partial correlations.

Control Variables			Profit after tax	Cost of raw material	Labor productivity
In man effect sq	Profit after tax	Correlation	1.000	−.040	−.082
		Significance	.	.681	.396
		(2-tailed) df	0	108	108
	Cost of raw material	Correlation	−.040	1.000	.174
		Significance	.681	.	.069
		(2-tailed)df	108	0	108
	Labor productivity	Correlation	−.082	.174	1.000
		Significance	.396	.069	.
		(2-tailed) df	108	108	0

While the **adjusted R^2** (this term and its meaning will be explained in Chapter 2) was only 0.486 without using the transformation of the managerial effectiveness variable, after using lnmaneffectsq the adjusted R^2 has improved to 0.5677. As revealed by the partial correlations both labor productivity and cost of raw material have regression coefficients which are not significantly different from zero.

Table 1.7 Least squares estimation.

Variable	Regression Coefficient	Standard Error	t-statistic	Significance
Constant (Intercept)	63.2426	5.9168	10.6887	0.0000
Labor productivity	–0.1373	0.0931	1.4746	0.0716
Cost of raw material	–3.3053	4.7436	0.6968	0.2437
Ln Managerial effectiveness square	2.5481	1.1595	2.1975	0.0151
R^2	0.5795			
Adj R^2	0.5677			

Dependent variable: Profit after tax.

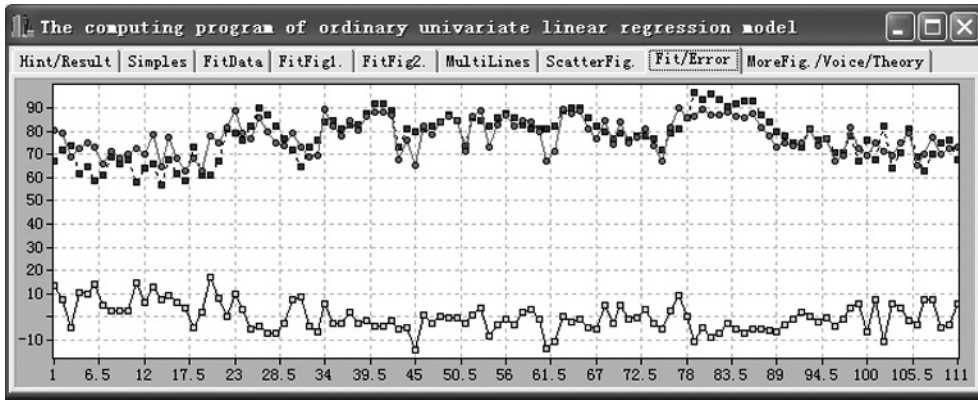


Figure 1.8 Actual and fitted values of profits after tax along with the residuals.

Once such correlations are calculated and scatter diagrams are drawn a potentially useful set of independent variables can be prepared. As we explain in greater detail, from that list one can select a final linear multiple regression using various variable selection methods. Thus, Chapter 2 also can be regarded as a major component of data mining or exploratory data analysis.

Often, we are tempted to specify a regression relation without examining or exploring the sample data to see what story it tells. It pays to look at the data more carefully. This is what we demonstrate next. Figure 1.8 presents the graph of actual and fitted values of profits after tax along with the residuals and can be shown in DASC software. The curve below represents the residual errors.

We note that the top part of figure shows the goodness of fit by plotting the actual and fitting values of the dependent variable, while the bottom part shows the plot of residual errors. Figure 1.9 is the same.

We note from the figure above that the estimated errors have a systematic pattern suggesting that for smaller companies there is an over-estimation and for larger companies there is an underestimation. We therefore introduced a new dummy variable labeled ‘large’ which is

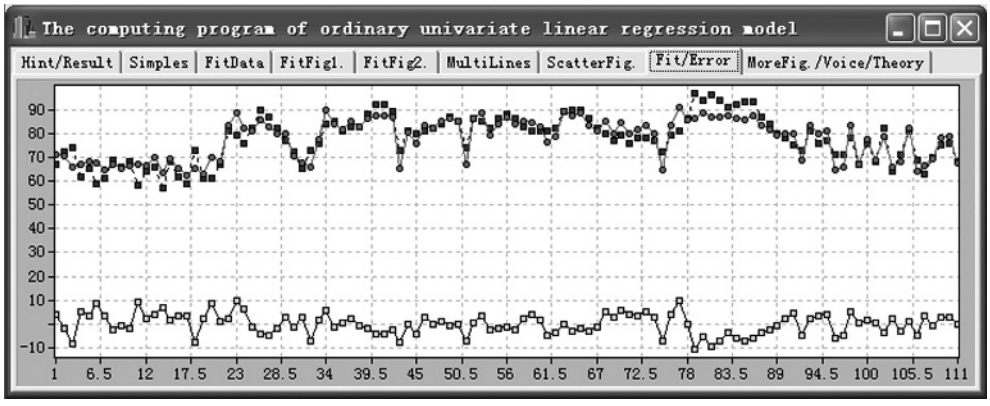


Figure 1.9 Fitted and actual values with a new dummy variable.

Table 1.8 Dependent variable: profit after tax.

Variable	Regression Coefficient	Standard Error	t-statistic	Significance
Constant	62.5065	4.1923	14.9099	0.0000
Labor productivity	–0.0109	0.0671	0.1626	0.4356
Cost of raw material	–0.8029	3.3692	0.2383	0.4061
Ln maneffect square	0.8708	0.8373	1.0401	0.1503
Large	11.6774	1.1279	10.3534	0.0000
R^2	0.7909			
Adj R^2	0.7830			

equal to 1 if the company has a profit after tax of Rs 75 Crores and zero otherwise. The adjusted R^2 has increased from 0.5677 to 0.783. The final results obtained are presented below in Table 1.8 followed by the graph of fitted and actual values and the estimated residuals in Figure 1.9.

What was shown above is the special case of a more general method. Here we have shown that the sample can be classified into two clusters, one for small firm with after tax profits of less than Rs. 75 Crores. A standard data mining technique is to see if the sample can be grouped into several clusters using cluster analysis so that one could try a different model specification for each cluster. Section 9.2.1 of Chapter 9 deals with discriminant analysis and cluster analysis. One can perform the scatter and correlation analysis separately for each cluster of samples.

1.3.3 Cleaning and reprocessing data to discover patterns: BSE index data

Singhal (2005) examined the overnight gains and losses in the stock prices on the Bombay Stock Exchange, along with day gains and losses. The BSE Index and individual stock prices did show the random walk type pattern. He asked two questions:

Table 1.9 BSE index data.

X	Y	X	Y
74.34	−31.67	2.53	2.24
27.67	−13.72	−0.39	−2.78
16.77	−9.32	−4.2	−5.89
10.68	5.94	−12.46	−8.18
6.34	−4.63	−58.34	11.27

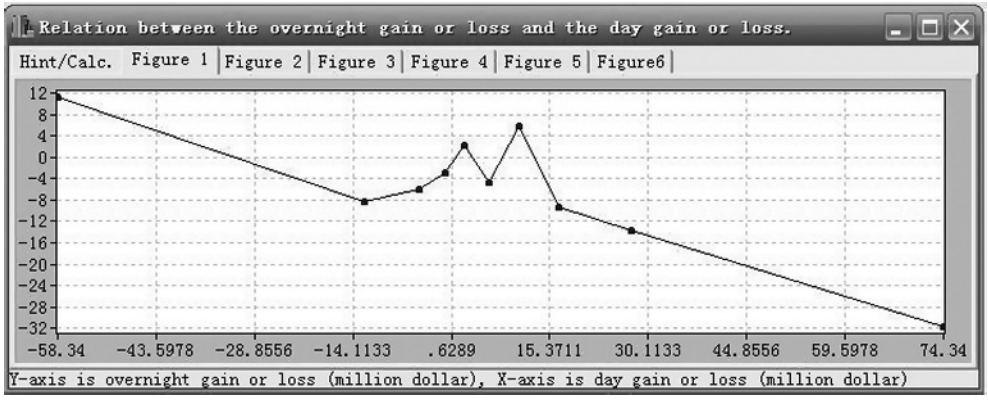


Figure 1.10 Relationship between the overnight gain or loss and the day gain or loss.

(1) Is it possible to predict the value of the day-time gain using the value of the over-night gain?

(2) Is it possible to get significantly more returns by over-night trading or day trading as compared to long-term trading?

He found that if the data were smoothed and some averages were derived for different ranges of gains and losses then the noise in individual daily series is removed and there emerges a definite pattern. That pattern can be used to make gains on the market from the knowledge that there was a substantial gain or loss overnight. The data were arranged in the increasing order of overnight gain and grouped into ten deciles, each of the deciles having about 148 points. The means of these deciles are presented below in Table 1.9.

Based on the processed data above we can see a relationship between the overnight gain or loss (Y) and the day gain or loss (X) as given in Figure 1.10.

The pattern observed with this processed data is different from the raw data observed earlier in this chapter (Section 1.2.6). The raw data only shows a random walk type of relationship between the overnight gain and day gain, the grouped data shows such random walk behavior only for the small ranges of gains or losses. When such gains or losses are substantial there is a negative relationship between them. Based on the pattern above Singhal developed a buy-sell strategy at the beginning of the day that takes into account the observed pattern above, and through simulation showed that short-run gains from the strategy exceeded

the gains from day trading (by 100 times) and long-term trading (by 35 times). He found similar gains in trading on individual stocks, the latter based on a similar observed patterns for individual stock prices.

The example given above is a special case of processing the data to form an aggregation of samples and data condensation. There are other methods of data mining (exploratory data analysis) such as reducing the number of variables into a smaller set. This aspect was already mentioned in the loan default example of Section 1.2.4. A more general approach is to use **principal components** and factor analysis methods to reduce the dimensions of the vector of independent variables. Under that approach the **multivariate analysis** of factor and principal components analysis, described more fully in Chapter 9, is used to determine fewer linear combinations of a large set of variables that are used to replace the larger set of variables.

We have thus given a brief tour of a more specialized subject of data mining (exploratory data analysis) which precedes specification of alternate models that constitute the main grist of the statistical modeling issues treated in this book.

1.4 Econometric modeling: The roadmap of this book

1.4.1 The econometric modeling strategy

Statistical modeling started with very general and simple problems with small samples and has now advanced to a stage where large data sets are being stored in data warehouses in the clouds and are used for statistical modeling. The statistical theories and methods also have advanced over the years and simple models are being replaced by those that are more complex. Models remained simple in the earlier years in order to keep the computation tasks manageable. With advances in computing statisticians were able to introduce more complexity into modeling. Unfortunately amongst a section of people who seek statistical applications an impression has been created that the more sophisticated or complex a model is the better it is. This is unfortunate. Model selection must be based not on sophistication or complexity but on its performance in obtaining a good fit to the sample data. Thus, the modeling strategy must consist of:

(1) Getting the state of the art of the domain knowledge so as to know a priori which variables could be related to which variables and how (or get the broad general structure of the model).

(2) Doing exploratory data analysis in order to refine the general model to suit the problem at hand (to refer to the population to which the sample belongs).

(3) Specifying alternate models and estimating (calibrating) them.

(4) Choosing one of the models as the best model based on some credible statistical criteria.

(5) Examining if the chosen model is acceptable.

(6) If not looking for new data that could have been omitted or doing further exploration of data or both to come up with better models, and repeating the process all over again.

The econometric modeling strategy above is described schematically in Figure 1.11.

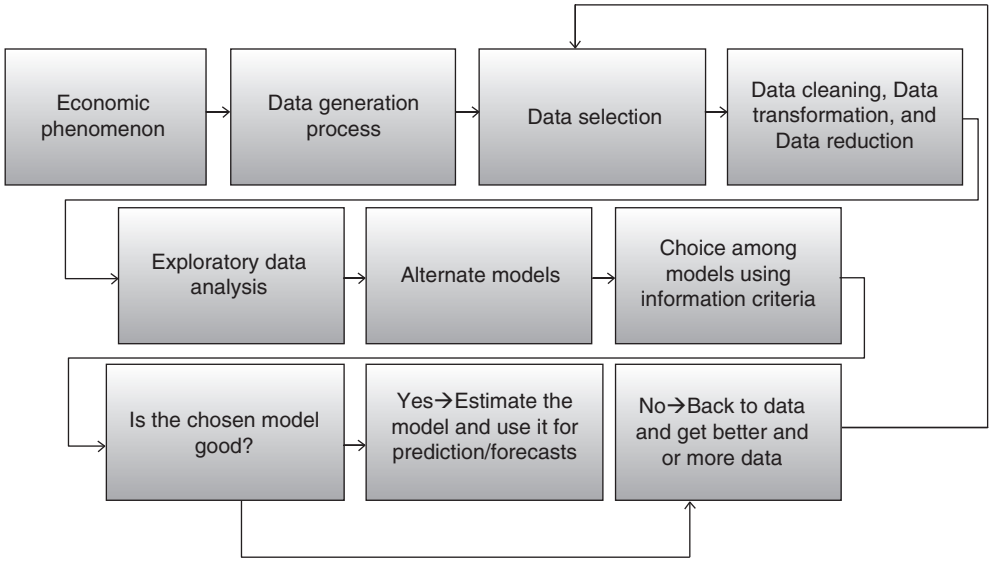


Figure 1.11 Econometric modeling process.

1.4.2 Plan of the book

After the exploratory data analysis we will arrive at one or more alternate specifications of the statistical model to represent the data generation process of the observations we have made. Each of those specifications will have, in general, a dependent variable and several dependent variables and an error in equations that has a probability distribution. The deterministic part is the signal and the equation error is the noise.

Strictly speaking the first part, signal, needs to be a computational procedure that will generate a unique value of the dependent variable, given the values assumed by each, of let us say, m independent variables. If that computational procedure is in terms of a mathematical function with possibly a few unknown constant parameters represented by a p-dimensional vector θ we can write the model as:

$$y_i = f(x_i;\theta) + \varepsilon_i, \quad i = 1,2,.....,n \tag{1.2}$$

If the function f is linear and y is a continuous variable and the error term has a distribution with mean zero and constant variance we get ordinary **linear multiple regression models** with homoscedasticity. Such models are discussed in Chapter 2. Chapter 2 has an extensive discussion on statistical criteria for the selection of variables in linear multiple regressions. It also deals with problems associated with correlated independent variables (multicollinearity and how to cope with that problem).

If the function f of equation (1.2) is linear and the error term has a mean zero but non-constant variance then we get linear multiple regression models with heteroscedasticity. These models are treated in Chapter 3. If the variables are not continuous but categorical then certain problems arise both in interpreting the regressions and in estimating them. These issues engage us in Chapter 4. If the function $f(x;\theta)$ is nonlinear in parameters we get some

other special issues in model specification and estimation. Nonlinear regression models of that nature are also covered in Chapter 4.

We mentioned at the beginning that the specification of the deterministic part of the model need not be a mathematical formula with unknown constant parameters. In fact there is a whole class of specifications that are nonparametric or a simple parametric combination of nonparametric functions (like a weighted average of two or more nonparametric functions). Models of that nature are dealt with in Chapter 5.

One of the basic features of econometrics is that several economic variables are mutually dependent and thus between some of them one cannot say which is dependent and which is independent. That situation is handled by specifying a model consisting of a system of regression equations with as many as there are economic variables that are explained by the model. The independent variables of such equations do not satisfy the assumptions of standard multiple linear regressions. They need a special treatment. The **systems of simultaneous economic equations** are the subject matter of Chapter 6. In economics the dynamic nature of economic relationships is usually captured by having economic relationships involving variables with time lags. These are also dealt with in Chapter 6.

When economic variables are time series then subsequent observations are correlated in some specific and interesting ways. Such economic time series can be modeled easily if they satisfy a property that any chunk of the time series taken as a distribution of a **stochastic process** has the same distribution as that of any other chunk. Such time series are called **stationary time series**. Economic modeling of a single stationary economic time series is the topic covered in Chapter 7. Reading Chapter 7 after reading Chapter 6 one might ask what if we have a system of simultaneous equations but with time series data with lags? This topic is covered in Chapter 8 under multiple time series analysis. Even when economic time series are not stationary it is possible that some transformations of them are stationary. If we know what those transformations are we can model the transformed time series as stationary time series, and retransform them to get the original series. This way of analysis of **non-stationary time series** is also covered in Chapter 8.

Chapter 9 covers some very important multivariate analysis tools such as Analysis of Variance of single and multifactor types. It also covers path analysis, factor analysis, principal component analysis, partial least squares. It has an interesting application of **structural equation modeling** using **path analysis** applied to analysis of the consumer satisfaction index. It also presents panel data analysis.

An inquisitive reader would perhaps look at all the statistical theories and methods ranging from classical inference, Bayesian inference, parametric regression, non-parametric regression, and quantile regression and wonder how they all fit together. To quench the thirst of such a reader the last chapter, Chapter 10, offers a unified treatment that is based primarily on the intuitive notion that statistical modeling covered in the book is based on harnessing the information contained in various types of information in the underlying structure of the data generating process.

Now here is a final word for the student reader. If you want to learn the theories and methods of econometrics, you can read this book, because this book offers statistical theories and methods in the form of a good collection of theorems. If you want to go deeper into the theories of statistics and mathematics in econometrics, you can read the Electronic References in this book. If you only want to use the methods of econometrics to deal with economic data, fit a suitable model to the data and make some predictions or forecasting, you can read the data examples in this book and use the computation software DASC, and do not hesitate to

skip the mathematical derivations. But such a reader is reminded that the scientific credibility of the results requires that the application of statistical tools be appropriate from the theoretical and methodological perspective. Then you may consult a statistician for advice.

Electronic references for Chapter 1

1-1 Data Mining
  1-1-1 Crucial Concepts in Data Mining
  1-1-2 Data Warehousing
  1-1-3 On-Line Analytic Processing (OLAP)
  1-1-4 Exploratory Data Analysis (EDA) and Data Mining Techniques
1-2 Guide to DASC
  1-2-1 How to Use DASC
  1-2-2 How to Calculate the Data Examples and Extended Contents by DASC
1-3 Brief Review of Linear Algebra
1-4 Brief Review of Probability Theory

References

Andersen, F.M., Celov D., Grinderslev D. & Kazlauskas A. (2005) A macro-econometric model of Lithuania LITMOD. *Economic Modeling* **22**, 707–19.

Anderson M.J. & Whitcomb P.J. (2000) *DOE Simplified: Practical Tools for Effective Experimentation*, Productivity Incorporated, Portland, Oregon, USA.

Beenstock M. (1995) An econometric model of the oil importing developing countries. *Economic Modeling* **12**, 3–14.

Bozdogan H. (1990) On the information-based method of covariance complexity and its applications to the evaluation of multivariate linear models. *Communications in Statistics Theory and Methods* **19**, 221–78.

Breeden J.L. (2007) Modeling data with multiple time dimensions. *Computational Statistics & Data Analysis* **51**, 4761–85.

Brownlees C.T. & Gallo G.M. (2006) Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational Statistics & Data Analysis* **51**, 2232–45.

Busemeyer J.R., Forsyth B. & Nozawa G. (1988) Comparisons of elimination by aspects and suppression of aspects choice models based on choice response time. *Journal of Mathematical Psychology* **32**, 341–49.

Chong J. (2005) The forecasting abilities of implied and econometric variance–covariance models across financial measures. *Journal of Economics and Business* **57**, 463–90.

Costantini M. & Destefanis S. (2009) Cointegration analysis for cross-sectionally dependent panels: The case of regional production functions. *Economic Modeling* **26**, 320–7.

D'Enza A.I., Palumbo F. & Greenacre M. (2008) Exploratory data analysis leading towards the most interesting simple association rules. *Computational Statistics & Data Analysis* **52**, 3269–81.

Dolk D.R. & Kridel D.J. (1991) An active modeling system for econometric analysis. *Decision Support Systems* **7**, 315–28.

Farebrother R.W. (1996) The role of chaotic processes in econometric models. *Journal of Statistical Planning and Inference* **49**, 163–76.

Frisch R. (1929) Correlation and scatter in statistical variables. *Nordic Statistical Journal* **1**, 36–108. Reproduced in Bjerkholt, Olav (1995) (Editor) *Foundations of Modern Econometrics: The Selected Essays of Ragnar Frisch* (in 2 Vols), Aldershot: Edward Elgar.

Han, J., & Kamber M. (2006), *Data Mining: Concepts and Techniques*, Elsevier, Second Edition.

Hendry D.F. (2001) Achievements and challenges in econometric methodology. *Journal of Econometrics* **100**, 7–10.

Kang I.-B. (2003) Multi-period forecasting using different models for different horizons: an application to U.S. economic time series data. *International Journal of Forecasting* **19**, 387–400.

Kumar T. Krishna, Sushanta K. Mallick & Jayarama H. (2009) Estimating consumption deprivation in India using survey data: A state-level rural-urban analysis before and during reform period, *Journal of Development Studies* **45**(4), 441–70.

Li Q. (1999) Consistent model specification tests for time series econometric models. *Journal of Econometrics* **92**, 101–47.

Li T. (2009) Simulation based selection of competing structural econometric models. *Journal of Econometrics* **148**, 114–23.

Lin K.-P. & Farley A.M. (1995) Causal reasoning in econometric models. *Decision Support Systems* **15**, 167–77.

Pagan A. (1997) Whatever happened to optimal control of econometric models. *Control Engineering Practice* **5**, 527–33.

Page, S.E. (1999) Computational models from A to Z. *Complexity* 5(1), 35–41.

Raghavarao, D., Wiley J.B., and Chitturi P. (2011) *Choice Based Conjoint Analysis: Models and Designs*, CRC Books, A Chapman and Hall Book, Boca Raton, London, and New York.

Ridder G. & Moffitt R. (2007) The econometrics of data combination. *Handbook of Econometrics* **6**, 5469–547.

Patterson K.D. (2003) Exploiting information in vintages of time-series data. *International Journal of Forecasting* **19**, 177–97.

Pesaran M.H. & Smith R.P. (1985) Evaluation of macroeconometric models. *Economic Modeling* **2**, 125–34.

Pesaran, M.H. & Smith R. (1995) The role of theory in econometrics. *Journal of Econometrics* **67**, 61–79.

Posse C. (1995) Projection pursuit exploratory data analysis. *Computational Statistics & Data Analysis* **20**, 669–87.

Sagan C. (1996) *Demon Haunted World: Science as a Candle in the Dark*. Ballantine, a Division of Random House, New York.

Sandiford P.J. and Seymour D. (2007) A discussion of qualitative data analysis in hospitality research with examples from an ethnography of English public houses. *International Journal of Hospitality Management* **26**, 724–42.

Singhal P. (2005) Inefficiencies in Indian capital markets: can overnight gain be used as a predictor of day gain? *IIMB Management Review* **17**, 23–30.

Thompson S.K. (2002) *Sampling*. John Wiley and Sons Inc, New York.

von Natzmer W. (1985) Econometric policy evaluation and expectations. *Economic Modeling* **2**, 52–8.

Wallbäcks L. (2007) Multivariate data analysis of multivariate populations. *Chemometrics and Intelligent Laboratory Systems* **86**, 10–16.

Wojciech W.C. (1991) Large econometric models of an East European economy: A critique of the methodology. *Economic Modeling* **8**, 45–62.

Zellner A. (1996) Past, present and future of econometrics. *Journal of Statistical Planning and Inference* **49**, 3–8.

Zheng X. (2008) Testing for discrete choice models. *Economics Letters* **98**, 176–84.