Preface

This book is intended for persons with interest in analyzing financial data and with at least some knowledge of mathematics and statistics. No prior knowledge of finance is required, although a reader with experience in trading and in working with financial data may understand some of the discussion more readily. The reader with prior knowledge of finance may also come to appreciate some aspects of financial data from fresh perspectives that the book may provide. Financial data have many interesting properties, and a statistician may enjoy studying and analyzing the data just because of the challenges it presents.

There are many texts covering essentially the same material. This book differs from most academic texts because its perspective is that of a real-world trader who happens to be fascinated by data for its own sake. The emphasis in this book is on financial *data*. While some stale datasets at the book's website are provided for examples of analysis, the book shows how and where to get current financial data, and how to model and analyze it.

Understanding financial data may increase one's success in the markets, but the book does not offer investment advice.

The organization and development of the book are data driven. The book begins with a general description of the data-generating processes that yield financial data. In the first chapter, many sets of data are explored statistically with little discussion of the statistical methods themselves, how these exploratory analyses were performed, or how the data were obtained. The emphasis is on the *data-generating processes of finance:* types of assets and markets, and how they function.

The first chapter may seem overly long, but I feel that it is important to have a general knowledge of the financial data-generating processes. A financial data analyst must not only know relevant statistical methods of analysis, the analyst must also know, for example, the difference between a seasoned corporate and a T-Bill, and must understand why returns in a short index ETF are positively correlated with the VIX.

While reading Chapter 1 and viewing graphs and other analyses of the various datasets, the reader should ask, "Where could I get those or similar datasets, and how could I perform similar analyses?" For instance, "How could I obtain daily excess returns of the SPY ETF, as is used in the market model on page 61?".

These questions are addressed in an appendix to Chapter 1. Appendix A1, beginning on page 139, discusses computer methods used to obtain real finan-

cial data such as adjusted closing stock prices or T-Bill rates from the web, and to mung, plot, and analyze it.

The software used is R. Unless data can be obtained and put in a usable form, there can be no analysis. In the exercises for the appendix, the reader is invited to perform similar exploratory analyses of other financial data.

In the later chapters and associated exercises, the emphasis is on the *methods of analysis* of financial data. Specific datasets are used for illustration, but the reader is invited to perform similar analyses of other financial data.

I have expended considerable effort in attempting to make the Index complete and useful. The entries are a mix of terms relating to financial markets and terms relating to data science, statistics, and computer usage.

Financial Data

This book is about financial data and methods for analyzing it. Statisticians are fascinated by interesting data. Financial data is endlessly fascinating; it does not follow simple models. It is not predictable. There are no physical laws that govern it. It is "big data". Perhaps best of all, an endless supply of it is available freely for the taking.

Access to financial data, and the financial markets themselves, was very different when I first began participating in the market over fifty years ago. There was considerably more trading friction then; commissions were *very* significant. The options markets for the retail investor/trader were almost nonexistent; there were no listed options (those came in 1973). Hedging opportunities were considerably more limited. There were no ETFs. (An early form was released in 1989, but died a year later; finally, the first Spider came out in 1993.) Most mutual funds were "actively managed", and charged high fees.

It has been an interesting ride. The euphoria of the US markets of the '60s was followed by a long period of doldrums, with the exception of the great year of 1975. The Dow's first run on 1,000 fizzled to close 1968 at 943. Although it ended over 1,000 in two years (1972 and 1976), it was not until 1982 that it closed at and stayed over 1,000. Now, since 1999, it's been in five figures except for two crashes. The 80s and 90s went in one direction: up! The "dot.com crash" was a severe hiccup, when the Dow fell back to four figures. Not just the dot.coms, but almost everything else got smashed. Then, back on track until the crash caused by the financial crisis, when the Dow again fell to four figures. In early 2018, at historically low volatility, it broke through 26,000, and then within two years, 29,000. No one could understand the volatility, but lots of traders (especially the "smart money") made money trading volatility, until suddenly they lost money on vol trades, in many cases a *lot* of money (especially the "smart money"). Also, no one understood the

Christmas Eve crash of 2018 (although many analysts "explained" it), but the brave ones made lots of money in the new year.

My interest in playing with financial data as a statistician is much more recent than my participation in the market, and was not motivated by my trading. The only kind of "formal" analysis that I do for my own portfolios is to run a market model (equation (1.35)) weekly, for which I entered data manually and used Fortran until sometime in the 1990s; then I used spreadsheet programs until about 2000; then I used R. I don't enter data manually anymore. Until a few years ago, I used R directly, but now I often use a simple Shiny app I wrote to enter my time intervals and so on. I get most price data directly from Yahoo Finance using `quantmod` (see page 171), but data on options are still problematic. The data-generating process itself is interesting and fun to observe, and that's why I do it.

Like anyone else, I would like to believe that the process is rational, but like any other trader, I know that it is not. That just makes analysis of financial data even more interesting.

Outline

Chapter 1 is about exploratory data analysis (EDA) of financial data. It is less quantitative than the remaining chapters. The chapter ends with a summary of general "stylized facts" uncovered by the EDA in the chapter. Chapter 1 began as a rather brief and breezy overview of financial data, but grew in length as terms and topics came up in developing the later chapters.

I think that a data analyst, in addition to knowing some general characteristics of the data, should have at least a general understanding of the data-generating process, and a purpose of Chapter 1 is to provide that background knowledge. Chapter 1 introduces terms and concepts that relate to financial data and the markets that generate that data. (Among other guidelines for the content of this chapter, I have attempted to include most of the terms one is likely to hear on the CNBC daily episodes or on other financial programs in the mass media.) When I use data from Moody's Seasoned Baa Corporate Bonds in examples or exercises later in the book, I want the reader to know what "Moody's" is, what "seasoned" means, and what "Baa corporate bonds" are, and I do not want to have to intrude on the point of the examples in later chapters to explain those terms there. Those terms are defined or explained in Chapter 1.

The exercises for Chapter 1 are generally conceptual, and involve few computations, unlike the exercises for the appendix to Chapter 1 and later chapters.

The full R code for most of the graphs and computations in Chapter 1 is available at the website for the book. The appendix to Chapter 1 discusses

the R code and describes how the data were obtained. The main reason for placing the appendix at this point is that for all of the remaining chapters, R will be used extensively without much comment on the language itself, and many of the exercises in those chapters will require R, and will require internet access to real financial data. The exercises for the appendix involve the use of R, in some cases just to replicate illustrations in Chapter 1.

Instead of stale data on the website for the book or some opaque proprietary database, I want the reader to be able to get and analyze real, current data.

The remaining chapters are about statistical methods. The methods could be applied in other areas of application, but the motivation comes from financial applications.

Chapter 2 harks back to the exploratory data analyses of Chapter 1 and discusses general nonparametric and graphical methods for exploring data.

Chapter 3 covers random variables and probability distributions. Although the chapter does not address statistical analysis *per se*, these mathematical concepts underlie all of statistical inference. Distributional issues particularly relevant for financial data, such as heavy-tailed distributions and tail properties, are emphasized.

Chapter 3 also describes methods for computer generation of random numbers that simulate realizations from probability distributions. Some of the basic ideas of simulation are presented in Section 3.3, and Monte Carlo methods are used in later chapters and in the exercises.

Chapter 4 discusses the role that probability distributions play in statistical inference. It begins with a discussion of statistical models and how to fit them using data. The criteria for fitting models involve some form of optimization ("least" squares, "maximum" likelihood, etc.); therefore, Chapter 4 includes a small diversion into the general methods of optimization. The chapter continues with basic concepts of statistical inference: estimation, hypothesis testing, and prediction. Specific approaches, such as use of the bootstrap, and relevant applications, such as estimation of VaR, are described. Analysis of models of relationships among variables, particularly regression models, is discussed and illustrated.

In view of the recent browbeating by some statisticians about the use of the word "significant", I feel compelled to mention here that I use the term extensively; see the notes to Chapter 4, beginning on page 469.

Chapter 5 provides a brief introduction to the standard time series models, and a discussion of why these models do not work very well in practice. Time series models that account for certain types of heteroscedasticity (GARCH) are discussed, and methods of identifying and dealing with unit roots in autoregressive models are developed. Chapter 5 also addresses topics in vector autoregressive processes, in particular, cointegration of multiple series.

A couple of major topics that are not addressed are the analysis of fixed assets, such as bonds, and the pricing of derivative assets using continuous-time diffusion models. These topics are mentioned from time to time, however.

For any of the topics that are covered, there are many additional details that could be discussed. Some of these are alluded to in the "Notes and Further Reading" sections.

There are of course many smaller topics, such as processing streaming data and high-frequency trading, and the resulting market dynamics, that cannot be discussed because of lack of space.

Software and Programming

The software I use is R. Although I often refer to R in this book, and I give some examples of R code and require R in many exercises, the reader can use other software packages.

A reader with interest but even with no experience with R can quickly pick up enough R to produce simple plots and perform simple analyses. The best way to do this is to look at a few code fragments, execute the code, and then make small changes to it and observe the effects of those changes. The appendix to Chapter 1 displays several examples of R code, and the code to produce all of the graphs and computations shown in Chapter 1 is available at the website for the book.

If the objective is to be able to use R to perform some specific task, like making a graph, the objective can be achieved quickly by finding some R code that performs that kind of task, and then using it with the necessary modifications. (This is not "programming".)

R is a rich programming language. If the objective is to learn *to program* in R, it is my oft-stated belief that "the way to learn to program is to get started and then to program". That's the way I learned to program; what more can I say? That applies to other things also: the way to learn to type is to get started (get a keyboard and find out where the characters are) and then to type; the way to learn to swim is to get started (find some water that's not too deep) and then to swim.

Although I have programmed in many languages from Ada to APL, I don't believe it's desirable to be a programmer of all languages but master of none (to rephrase an old adage). I'd prefer to master just one (or three).

Prerequisites

The prerequisites for this text are minimal. Obviously some background in mathematics, including matrix algebra, is necessary. There are several books that can provide this background. I occasionally make reference to my own

books on these topics. That is not necessarily because they are the best for that purpose; it is because I know where the material is in those books.

Some background in statistics or data analysis and some level of scientific computer literacy are also required. I assume that the reader is generally familiar with the basic concepts of probability such as *random variable*, *distribution*, *expectation*, *variance*, and *correlation*. For more advanced concepts and theory, I refer the reader to one of my exhausting, unfinished labors of love, *Theory of Statistics*, at

`mason.gmu.edu/~jgentle/books/MathStat.pdf`

Mention of rather advanced mathematical topics is made occasionally in the text. If the reader does not know much about these topics, the material in this book should still be understandable, but if the reader is familiar with these topics, the mention of the topics should add to that reader's appreciation of the material.

No prior knowledge of finance is assumed, although a reader with some background in finance may understand some of the discussion more quickly.

In several places, I refer to computer programming, particularly in R. Some of the exercises require some simple programming, but most exercises requiring use of the computer do not involve programming.

Note on Examples and Exercises

The book uses real financial data in the examples, and it requires the reader to access real data to do the exercises. Some stale datasets are also available at the website for the book. The time periods for the data are generally the first couple of decades of the twenty-first century.

The book identifies internet repositories for getting *real* data and *interesting* data. For the exercises, the reader or the instructor of a course using the book is encouraged to substitute for "2017", "2018", or any other bygone time period, a more interesting time period.

In addition to real data, the book discusses methods of simulating artificial data following various models, and how to use simulated data in understanding and comparing statistical methods. Some exercises ask the reader to assess the performance of a statistical technique in various settings using simulated data.

Data preparation and cleansing are discussed, and some exercises require a certain amount of data wrangling.

The exercises in each chapter are not necessarily ordered to correspond to the ordering of topics within the chapter. Each exercise has a heading to indicate the topic of the exercise, but the reader is encouraged to read or skim the full chapter before attempting the exercises. The exercises vary considerably in difficulty and length. Some are quite long and involved.

Supplementary Material

The website for the book is

`mason.gmu.edu/~jgentle/books/StatFinBk/`

The website has a file of hints, comments, and/or solutions to selected exercises. Supplemental material at this site also includes R code used in producing examples in the text. Although I emphasize real, live data, the website also has a few financial datasets from the past.

The website has a list of known errors in the text that will be updated as errors are identified.

A complete solutions manual is available to qualified instructors on the book's webpage at

`www.crcpress.com`

Because adjusted asset prices change over time, the reader must be prepared to accept slight differences in the results shown from the results arising from data that the reader may access at a later time.

Acknowledgments

First, I thank John Chambers, Robert Gentleman, and Ross Ihaka for their foundational work on R. I thank the R Core Team and the many package developers and those who maintain the packages for continuing to make R more useful.

Jim Shine has read much of the book and has worked many of the exercises. I thank Jim for many helpful comments.

I thank the anonymous reviewers for helpful comments and suggestions.

I thank my editor John Kimmel. It has been a real pleasure working with John again, as it was working with him on previous books.

I thank my wife, María, to whom this book is dedicated, for everything.

I used TEX via LATEX 2_ε to write the book. I did all of the typing, programming, etc., myself, so all mistakes are mine. I would appreciate receiving notification of errors or suggestions for improvement.

Fairfax County, Virginia James E. Gentle
December 27, 2019