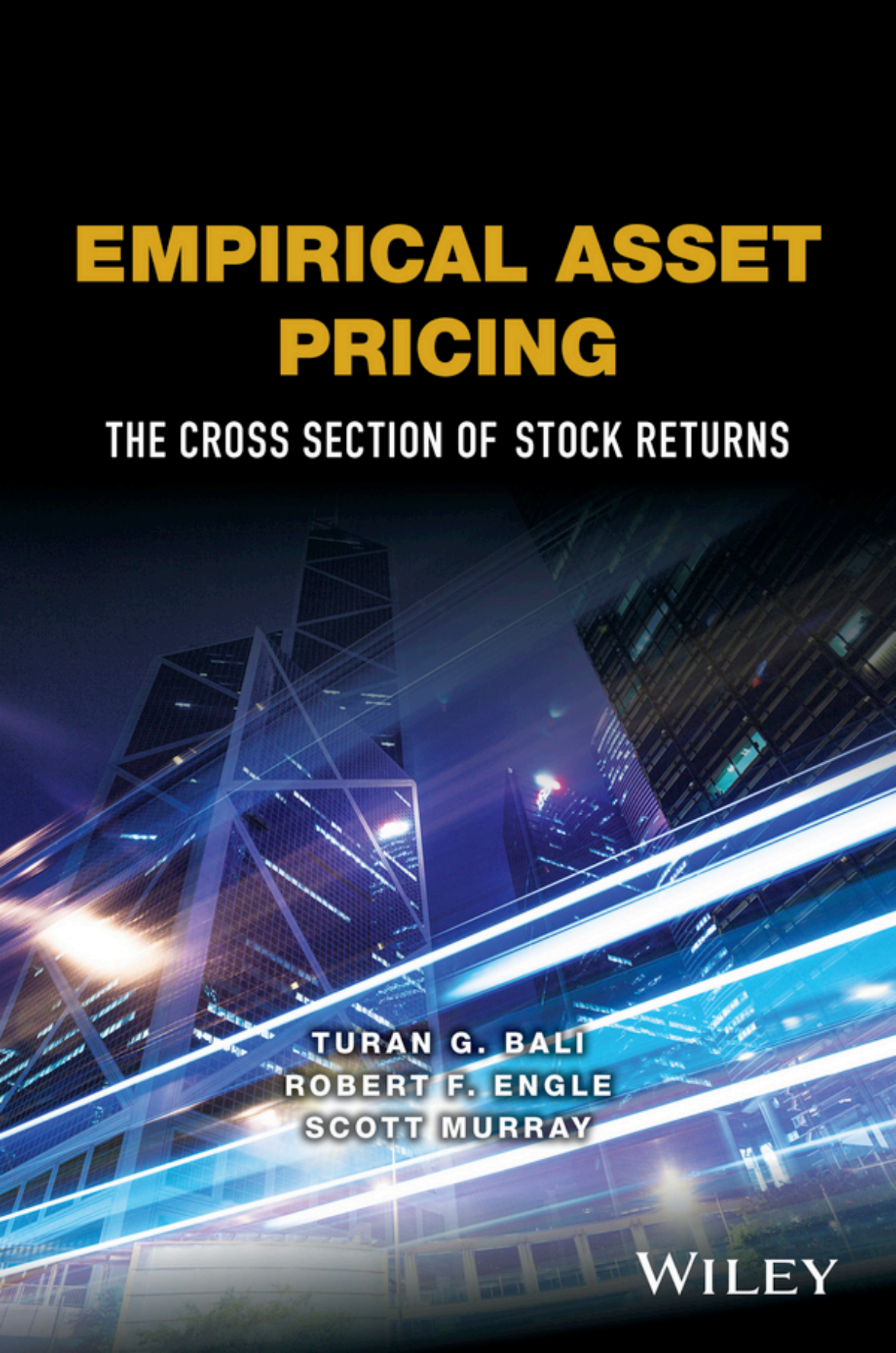
EMPIRICAL ASSET PRICING

THE CROSS SECTION OF STOCK RETURNS



TURAN G. BALI
ROBERT F. ENGLE
SCOTT MURRAY

WILEY

EMPIRICAL ASSET PRICING

EMPIRICAL ASSET PRICING

The Cross Section of Stock Returns

TURAN G. BALI

ROBERT F. ENGLE

SCOTT MURRAY

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data

Names: Bali, Turan G., author. | Engle, R. F. (Robert F.) author. | Murray, Scott, 1979- author.
Title: Empirical asset pricing : the cross section of stock returns / Turan G. Bali, Robert F. Engle, Scott Murray.
Description: Hoboken : Wiley, 2016. | Includes bibliographical references and index.
Identifiers: LCCN 2015036767 (print) | LCCN 2016003455 (ebook) | ISBN 9781118095041 (hardback) | ISBN 9781118589663 (ePub) | ISBN 9781118589472 (Adobe PDF)
Subjects: LCSH: Stocks–Prices. | Rate of return. | Stock exchanges. | BISAC: BUSINESS & ECONOMICS / Finance.
Classification: LCC HG4636 .B35 2016 (print) | LCC HG4636 (ebook) | DDC 332.63/221–dc23
LC record available at http://lccn.loc.gov/2015036767

Typeset in 10/12pt TimesLTStd by SPi Global, Chennai, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

"The empirical analysis of the cross section of stock returns is a monumental achievement of half a century of finance research. Both the established facts and the methods used to discover them have subtle complexities that can mislead casual observers and novice researchers. Bali, Engle, and Murray's clear and careful guide to these issues provides a firm foundation for future discoveries."

John Campbell, Morton L. and Carole S. Olshan Professor of Economics, Harvard University

"Bali, Engle, and Murray have produced a highly accessible introduction to the techniques and evidence of modern empirical asset pricing. This book should be read and absorbed by every serious student of the field, academic and professional."

Eugene Fama, Robert R. McCormick Distinguished Service Professor of Finance, University of Chicago

"Bali, Engle, and Murray provide clear and accessible descriptions of many of the most important empirical techniques and results in asset pricing."

Kenneth R. French, Roth Family Distinguished Professor of Finance, Tuck School of Business, Dartmouth College

"This exciting new book presents a thorough review of what we know about the cross section of stock returns. Given its comprehensive nature, systematic approach, and easy-to-understand language, the book is a valuable resource for any introductory PhD class in empirical asset pricing."

Lubos Pastor, Charles P. McQuaid Professor of Finance, University of Chicago

CONTENTS

PREFACE

The objective of this book is to provide an overview of the empirical research on the cross-section of expected stock returns. The book is intended for use in doctoral-level empirical asset pricing classes and by investors who are looking for a review of the most important predictors of future stock returns. A doctoral student reader should come away with a solid understanding of the most fundamental results in the field and a strong base upon which to pursue future research in empirical asset pricing. For the reader whose intention is to apply the results presented in this book to practice, our hope is that the book provides a basis upon which investment strategies can be constructed as well as a strong understanding of the most prevalent patterns of risk and returns in the cross-section of stocks.

It is assumed that the reader of this book has at least an MBA level understanding of theoretical asset pricing and a solid grasp of basic econometric techniques. Fantastic books on these topics have been written by Cochrane (2005), Campbell, Lo, and MacKinlay (1996), and Elton, Gruber, Brown, and Goetzmann (2014).[1] More in-depth knowledge in either of these areas is obviously a benefit. While all of the analyses in this book are statistical in nature, the book is not designed to be an econometrics or statistics reference. Our discussions of statistical concepts, therefore, will

[1] Several other books have been written on related topics. Ang (2014) gives an in-depth insight into factor investing. Factor analysis plays a large role in the empirical asset pricing literature and is used heavily throughout this book. Karolyi (2015) gives a comprehensive exposition of risks associated with investing in emerging markets. Pedersen (2015) provides a strong introduction into the trading strategies used by hedge funds, many of which have their roots in the phenomena documented throughout this book. Campbell (2015) provides a theoretical and empirical overview of empirical asset pricing research.

be primarily conceptual. For a more detailed discussion of the statistical theory underlying our methodologies, we suggest that the reader find an econometrics or statistics text appropriate for the reader's level of knowledge in this area.

This book is divided into two main parts. Part I is devoted to a discussion of the most widely used statistical methodologies in empirical asset pricing research. The objective of this section is to give readers a detailed understanding of how to conduct such analyses and how to interpret the results. In addition, we discuss how the results are summarized and presented in academic research articles. The techniques can, very generally, be separated into two groups. Techniques in the first group are designed to summarize the data upon which the research is based. Techniques in the second group are designed to assess relations between the variables used in a study. These are the tools used to investigate the cross-sectional relations between a set of variables and future stock returns. Analysis of such relations is the primary objective of this book and, more generally, the majority of empirical asset pricing research. That being said, these techniques can be used for other purposes as well.

The second, and by far most important, part of this book discusses the major findings in empirical asset pricing research. In presenting each of the findings, we begin by discussing in detail the calculation of the main variables used to capture the characteristic of the stock that is under investigation. We then apply the techniques discussed in Part I, with the main objective being to understand the relation between the characteristic being examined and expected stock returns. While there are literally hundreds of different variables that have been shown to be related to future stock returns, we focus on the most widely recognized and cited phenomena in the literature.

We would like to acknowledge substantial support from our colleagues at Georgetown University, Georgia State University, and New York University. We would like to specifically thank Viral Acharya, Vikas Agarwal, Yakov Amihud, Andrew Ang, Gurdip Bakshi, Hank Bessembinder, Jacob Boudoukh, Brian Boyer, Stephen Brown, Nusret Cakici, Fousseni Chabi-Yo, Peter Christoffersen, Martijn Cremers, Ozgur Demirtas, Elroy Dimson, Rory Ernst, Wayne Ferson, Fangjian Fu, Thomas Gilbert, Hui Guo, Umit Gurun, Cam Harvey, Bing Han, David Hirshleifer, Armen Hovakimian, Kris Jacobs, Andrew Karolyi, Haim Kassa, Haim Levy, Jonathan Lewellen, Lasse Pedersen, Lin Peng, Jeff Pontiff, Anna Scherbina, Rob Schoen, Robert Stambaugh, Avanidhar Subrahmanyam, Yi Tang, Raman Uppal, Grigory Vilkov, David Weinbaum, Robert Whitelaw, Liuren Wu, Yuhang Xing, Jianfeng Yu, Lu Zhang, Xiaoyan Zhang, Guofu Zhou, and Hao Zhou for their valuable feedback on both this book and on our previous research that has informed its writing. Your input has substantially improved the quality of this book. We are especially grateful to John Campbell, Gene Fama, Kenneth French, and Lubos Pastor for their meticulous reading and detailed feedback, as well as for writing valuable reviews of our book. The creation of this book would not have been possible without the help of Sari Friedman, Jon Gurstelle, Saleem Hameed, and Steve Quigley at Wiley and Sons, Inc. The efficiency and skill with which they executed all facets of the production of this book far surpassed any reasonable expectations. Finally, we would like to thank our wives and children, Marianne, Jordan, Lindsay, Mehtap, Kaan, and Dara, for their unwavering support. Your love, encouragement, and tolerance played

an integral role in our ability to produce Empirical Asset Pricing: The Cross Section of Stock Returns.

Turan G. Bali, Robert F. Engle, and Scott Murray New York, 2016.

REFERENCES

Ang, A. *Asset Management A Systematic Approach to Factor Investing*. Oxford University Press, Oxford, 2014.

Campbell, J. Y. *Financial Decisions and Markets*. Princeton University Press, Princeton, NJ, 2015, manuscript in preparation.

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. *The Econometrics of Financial Markets*. Princeton University Press, Princeton, NJ, 1996.

Cochrane, J. H. *Asset Pricing*. Princeton University Press, Princeton, NJ, 2005.

Elton, E. J., Gruber, M. J., Brown, S. J., and Goetzmann, W. N. *Modern Portfolio Theory and Investment Analysis*. John Wiley & Sons, Hoboken, NJ, 9th Edition, 2014.

Karolyi, G. A. *Cracking the Emerging Markets Enigma*. Oxford University Press, Oxford, 2015.

Pedersen, L. H. *Effficiently Inefficient: How Smart Money Invests & Market Prices Are Determined*. Princeton University Press, Princeton, NJ, 2015.

PART I

STATISTICAL METHODOLOGIES

1

PRELIMINARIES

In this chapter, we present a number of items that are essential components of the methodologies presented in (Part I) of this book. We present these elements here for several reasons. First, they are common to many of the different analyses that will be discussed. Second, being that they are common to many of the methodologies, there is no one logical alternative as to where to present this material. Thus, to avoid repetition, we present these items here and will assume them to be understood for the remainder of the book.

Specifically, in this chapter, we first introduce the type of sample, or data, required for each of the analyses presented in this part. We then discuss winsorization, a technique that is used to adjust data, in order to minimize the effect of outliers on statistical analyses. Finally, we explain Newey and West (1987)-adjusted standard errors, *t*-statistics, and *p*-values, which are commonly used to avoid problems with statistical inference associated with heteroscedasticity and autocorrelation in time-series data.

1.1 SAMPLE

Each of the statistical methodologies presented and used in this book is performed on a panel of data. Each entry in the panel corresponds to a particular combination of entity and time period. The entities are referred to using *i* and the time periods are referenced using *t*. In most asset pricing studies, the entities correspond to stocks,

Empirical Asset Pricing: The Cross Section of Stock Returns, First Edition.
Turan G. Bali, Robert F. Engle, and Scott Murray.

bonds, options, or firms. The time periods used in most studies are months, weeks, quarters, years, and in some cases days. Frequently, the data corresponding to any given time period are referred to as a cross section. Thus, for a fixed value of t, the set of entities i for which data are available in the given time period t is the cross section of entities in time t. In almost all cases, the sample is not a full panel, meaning that the set of entities included in the sample varies from time period to time period. For each entity and time period combination (i, t), the data include several variables. In general, the variable X for entity i during period t will be referred to as $X_{i,t}$. It is frequently the case that when the data contain more than one variable, for example, X and Y, for a given observation i, t, the value of $X_{i,t}$ is available but the value of $Y_{i,t}$ is not available. When this is the case, analyses that require values of both X and Y will not make use of the data point i, t. Most studies create their sample such that the main sample includes all data points for which values of the focal variables of the study are available. Analyses that use nonfocal or control variables will then use only the subset of observations for which the necessary data exist. This approach allows each analysis to be applied to the largest data set for which the required variables are available. However, in some cases, researchers prefer to restrict the sample used for all analyses to only those observations where valid values of each variable used in the entire study are available. The downside of this approach is that frequently a large number of observations are lost. The upside is that all analyses are performed on an identical sample, thus negating concerns related to the use of different data sets for each of the analyses.

In the remaining chapters of Part I, we will use a sample where each entity i corresponds to a stock and each time period t corresponds to a year. The sample covers a period of 25 years from 1988 through 2012 inclusive. For each year t, the sample includes all stocks i in the Center for Research in Security Prices (CRSP) database that are listed as U.S.-based common stocks on December 31 of the year t. Exactly how to determine which stocks are U.S.-based common stocks will be discussed later in the book. At this point, it suffices to say that the sample for each year t consists of U.S. common stocks that were traded on exchanges as of the end of the given year. We will use this sample to exemplify each of the methodologies that are discussed in the remainder of Part I. We use a short sample period and annual periodicity because having a small number of periods in the sample will facilitate presentation of the methodologies. We refer to this sample as the methodologies sample. In Part II of this book, which is devoted to the presentation of the main results in the empirical asset pricing literature, we use monthly data covering a much longer sample period.

For each observation in the methodologies sample, we calculate five variables. We should remind the reader that in many cases, one or more of the variables may be unavailable or missing for certain observations. This is one of the realities under which empirical asset pricing research is conducted. Here, we briefly describe these variables. Detailed discussions of exactly how these variables are calculated will be presented in later chapters.

We calculate the beta (β) of stock i in year t as the slope coefficient from a regression of the excess returns of the stock on the excess returns of the market portfolio using daily stock return data from all days during year t. We require a minimum of 200 days worth of valid daily return data to calculate β. Values of β for which

this criterion is not met are considered missing.[1] We define the market capitalization (*MktCap*) for stock *i* in year *t* as the number of shares outstanding times the price of the stock at the end of year *t* divided by one million. Thus, *MktCap* is measured in millions of dollars. We take *Size* to be the natural log of *MktCap*. As will be discussed in Chapter 2, the distribution of *MktCap* is highly skewed; thus, most researchers use *Size* instead of *MktCap* to measure the size of a firm.[2] The book-to-market ratio (*BM*) of a stock is calculated as the book value of the firm's equity divided by the market value of the firm's equity (*MktCap*).[3] Finally, the excess return of stock *i* in year *t* is calculated as the return of stock *i* in year *t* minus the return of the risk-free security in year *t*. All returns are recorded as percentages; thus, a value of 1.00 corresponds to a 1% return. Stock return, price, and shares outstanding data come from CRSP. The data used to calculate the book value of equity come from the Compustat database. Risk-free security return data come from Kenneth French's data library.[4]

1.2 WINSORIZATION AND TRUNCATION

Financial data are notoriously subject to outliers (extreme data points). In many statistical analyses, such data points may exert an undue influence on the results, making the results unreliable. Thus, if these outliers are not adjusted or accounted for, it is possible that they may lead to a failure to detect a phenomenon that does exist (a type II error), or even worse, results that indicate a phenomenon where no such phenomenon is actually present (a type I error). While there are several statistical methods that are designed to assess the effect of outliers or ameliorate their effect on results, empirical asset pricing researchers usually take a more ad hoc approach to dealing with the effect of outliers.

There are two techniques that are commonly used in empirical asset pricing research to deal with the effect of outliers. The first technique, known as winsorization, simply sets the values of a given variable that are above or below a certain cutoff to that cutoff. The second technique, known as truncation, simply takes values of a given variable that are deemed extreme to be missing. We discuss each technique in detail. In doing so, we assume that we are dealing with a variable X for which there are n different observations, which we denote $X_1, X_2, \ldots, X_n$.

Winsorization is performed by setting the values of X that are in the top h percent of all values of X to the 100-hth percentile of X. Similarly, values of X in the bottom l percent of X values are set to the lth percentile of X. For example, assume that we want to winsorize X on the high end at the 0.5% level ($h = 0.5$). We begin by calculating the 99.5th percentile of the values of X. We denote this value $Pctl_{99.5}(X)$. Then, we set all values of X that are higher than $Pctl_{99.5}(X)$ to $Pctl_{99.5}(X)$. Now, assume that we want to winsorize X on the low end at the 1.0% level ($l = 1.0$). This is done by

[1] The details of the calculation of β are discussed in Chapter 8.
[2] The details of the calculation of *MktCap* and *Size* are discussed in Chapter 9.
[3] The details of the calculation of *BM* are discussed in Chapter 10.
[4] Kenneth French's data library is found at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

calculating the first percentile value of X, $Pctl_1(X)$, and setting all values of X that are lower than $Pctl_{1\%}(X)$ to $Pctl_1(X)$. In most cases, the values of h and l are the same, and common values at which researchers winsorize are 0.5% and 1.0%. Throughout this book, we frequently say that we winsorize the data at the 0.5% level. What this means is that both h and l are 0.5, and that winsorization takes place at both the high and low ends of the variable. The level at which winsorization should be performed depends largely on the noise in the variable being winsorized, with more noisy variables being winsorized at higher levels.

Truncation is very similar to winsorization, except instead of setting the values of X above $Pctl_h(X)$ to $Pctl_h(X)$, we set them to missing or unavailable. Similarly, values of X that are less than $Pctl_l(X)$ are taken to be missing. Thus, the main difference between truncation and winsorization is that in truncation, observations with extreme values of a certain variable are effectively removed from the sample for analyses that use the variable X, whereas with winsorization, the extreme values of X are set to more moderate levels.

There are a few ways that winsorization or truncation can be implemented. The first is to winsorize or truncate using all values of the given variable X over all entities i and time periods t. The second is to winsorize or truncate X separately for each time period t. Which approach to winsorization is taken depends on the type of statistical analysis that will be conducted. If a single analysis will be performed on the entire panel of data, the first method of winsorization or truncation is most appropriate. However, most of the methodologies used throughout this book are performed in two stages. The first stage involves performing some analysis on each cross section (time period) in the sample. The second stage analyzes the results of each of these cross-sectional analyses. In this case, the second approach to winsorization or truncation is usually preferable. Throughout this book, when we perform winsorization, it is on a period-by-period basis (the second approach).

When to use winsorization or truncation is a difficult question to answer because some outliers are legitimate while others are data errors. In addition, researchers sometimes use simple functional forms that are not well suited for capturing outliers. In a statistical sense, one might argue that truncation should be used when the data points to be truncated are believed to be generated by a different distribution than the data points that are not to be truncated. Winsorization is perhaps preferable when the extreme data points are believed to indicate that the true values of the given variable for the entities whose values are to be winsorized are very high or very low, but perhaps not quite as extreme as is indicated by the calculated values. Most empirical asset pricing researchers choose to use winsorization instead of truncation. However, if the results of an analysis are substantially impacted by this choice, they should be viewed with skepticism.

1.3 NEWEY AND WEST (1987) ADJUSTMENT

As eluded to in Section 1.2, the methodologies presented in the remainder of Part I and used throughout this book are executed in two steps: a cross-sectional step and

a time-series step. In many cases, the values used during the time-series step may exhibit autocorrelation and/or heteroscedasticity. If this is the case, the standard errors and thus p-values and t-statistics used to test a null hypothesis may be inaccurate. To account for these issues in a time-series analysis, empirical asset pricing researchers frequently employ a methodology, developed by Newey and West (1987), that adjusts the standard errors of estimated values to account for the impact of autocorrelation and heteroscedasticity. In this section, we briefly describe implementation of this technique. The details can be found by reading Newey and West (1987).

In most empirical asset pricing research, the Newey and West (1987) adjustment is used when examining the time-series mean of a single variable. We refer to this variable measured at time t as A_t. Notice here that there is no entity dimension to A, as A represents a single time series. The basic idea is that if values of A_t are autocorrelated or heteroscedastic, then using a simple t-test to examine whether the mean of A is equal to some value specified by the null hypothesis (usually zero) may result in incorrect inference, as the autocorrelation and heteroscedasticity may deflate (or inflate) the standard error of the estimated mean. To adjust for this, instead of using a simple t-test, the time-series values of A_t are regressed on a unit constant. The result is that the estimated intercept coefficient is equal to the time-series mean of A and the regression residuals capture the time-series variation in A and thus A's autocorrelation and heteroscedasticity. The standard error of the estimated mean value of A is a function of these residuals. So far, this is not different from a standard t-test. Applying the Newey and West (1987) adjustment to the results of the regression, however, produces a new standard error for the estimated mean that is adjusted for autocorrelation and heteroscedasticity. The only input required for the Newey and West (1987) adjustment is the number of lags to use when performing the adjustment. As discussed in Newey and West (1994), the choice of lags is arbitrary. Frequently, econometrics software sets the number of lags to $4(T/100)^a$, where T is the number of periods in the time series, $a = 2/9$ when using the Bartlett kernel, and $a = 4/25$ when using the quadratic spectral kernel to calculate the autocorrelation and heteroscedasticity-adjusted standard errors.[5] A large proportion of empirical asset pricing studies use monthly samples covering the period from 1963 through the present (2012, or $T = 600$ months for the data used in this book). Plugging in the value $T = 600$ and taking a to be either $2/9$ or $4/25$ results in a value between five and six. Most studies, therefore, choose six as the number of lags. Once the Newey and West (1987)-adjusted standard error has been calculated, t-statistics and p-values can be adjusted to perform inference on the time-series mean of A. As is standard, the new t-statistic is the difference between the coefficient on the constant (same as the sample mean) and the null hypothesis mean divided by the adjusted standard error. The p-value can then be calculated using the adjusted t-statistic and the same number of degrees of freedom as would be used to calculate the unadjusted p-value.

The astute reader may have noticed that in the previous paragraph it was completely unnecessary to present the Newey and West (1987) adjustment within the

[5]See Newey and West (1987, 1994) and references therein for further discussion of the Bartlett and quadratic spectral kernels.

context of a regression, because regression on a unit constant simply produces an estimated coefficient equal to the mean value and residuals that represent variation in the time series of A. We present the Newey and West (1987) adjustment in this manner for two reasons. First, in most statistical software, the Newey and West (1987) adjustment is executed by appropriately setting a certain parameter or argument to the regression function. The second is that the Newey and West (1987) adjustment is actually much more general than described in the previous paragraph. In the general case, the Newey and West (1987) adjustment can be applied to any time-series regression. It is for this reason that statistical software implements the Newey and West (1987) adjustment within the context of regression analysis.

In its general form, the Newey and West (1987) adjustment can be used to adjust the standard errors on all estimated coefficients from a time-series regression for autocorrelation and heteroscedasticity in the regression residuals. The procedure to do so is exactly as described earlier, except that the time-series A is regressed on one or more additional time series and, in most cases, a constant as well. The Newey and West (1987) adjustment will then generate an adjusted variance–covariance matrix of the estimated regression coefficients that accounts for autocorrelation and heteroscedasticity in the residuals. The square roots of the diagonal entries of this adjusted variance–covariance matrix then serve as the standard errors of the estimated regression coefficients. These adjusted standard errors are used to calculate adjusted t-statistics and p-values. As in the univariate case, the researcher must determine the appropriate number of lags to use in the adjustment. While the Newey and West (1987) adjustment may seem a bit abstract at this point, its use will become much more clear in subsequent chapters. This nontrivial case of the Newey and West (1987) adjustment is commonly employed in factor regressions of portfolio excess returns on a set of common risk factors. This will be discussed in more detail in Section 5.1.7.

1.4 SUMMARY

In this chapter, we have presented three elements that are common to most of the empirical methodologies that will be discussed in the remainder of Part I and heavily employed in the analyses of Part II. We have also described the sample that will be used to exemplify the methodologies throughout the remainder of Part I, which we refer to as the methodologies sample. The reason for presenting these items here is to avoid repetition in the remaining chapters of Part I.

REFERENCES

Newey, W. K. and West, K. D. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica, 55(3), 703–708.

Newey, W. K. and West, K. D. 1994. Automatic lag selection in covariance matrix estimation. Review of Economic Studies, 61(4), 631–653.

2

SUMMARY STATISTICS

Perhaps one of the most important elements of conducting high-quality empirical research is to have a strong understanding of the data that are being used in the study. Similarly, for a reader of empirical research, to fully comprehend the results of the study and assess the applicability of these results beyond the scope of the study, it is important to have at least a cursory understanding of the data upon which the analyses presented in the article were performed. For these reasons, most empirical research papers present summaries of the data prior to discussing the main results. Frequently, the first table of a research paper presents such a summary.

In this chapter, we present the most commonly used approach in the empirical asset pricing literature to calculating and presenting summary statistics. Effective presentation of summary statistics represents a trade-off between showing enough results to give the reader a good sense of the important characteristics of the data and not presenting so much that the reader is overwhelmed. The optimal approach to presenting summary statistics depends greatly on the type of study being conducted. The approach presented in this chapter is most appropriate when the objective of the study is to understand a cross-sectional phenomenon of the entities (stocks, bonds, firms, etc.) being studied. The procedure, therefore, is geared toward understanding the cross-sectional distribution of the variables used in the study.

Empirical Asset Pricing: The Cross Section of Stock Returns, First Edition.
Turan G. Bali, Robert F. Engle, and Scott Murray.

2.1 IMPLEMENTATION

The summary statistics procedure consists of two steps. In the first step, for each time period t, certain characteristics of the cross-sectional distribution of the given variable, X, are calculated. In the second step, the time-series properties of the periodic cross-sectional characteristics are calculated. In most cases, the time-series property of interest is the mean, in which case the final results that are presented represent the average cross section, where the average is taken over all periods t during the sample period.

2.1.1 Periodic Cross-Sectional Summary Statistics

The details of the first step are as follows. For each time period t, we calculate the cross-sectional mean, standard deviation, skewness, excess kurtosis, minimum value, median value, maximum value, and selected additional percentiles of the distribution of the values of X, where each of these statistics is calculated over all available values of X in period t. We let $Mean_t$ be the mean, SD_t denote the sample standard deviation, $Skew_t$ represent the sample skewness, $Kurt_t$ be the sample excess kurtosis, Min_t be the minimum value, $Median_t$ denote the median value, and Max_t represent the maximum value of X in period t. In addition, we will record the fifth, 25th, 75th, and 95th percentiles of X in month t, which we denote $P5_t$, $P25_t$, $P75_t$, and $P95_t$, respectively. Depending on the data and the objective of the study, it may be desirable to include additional percentiles of the distribution. For example, if the study focuses on extreme values of X, then it may be valuable to record the first, second, third, fourth, 96th, 97th, 98th, and 99th percentiles of the distribution as well. Alternatively, calculating the minimum, maximum, fifth percentile, and 95th percentile of the data may not be necessary if the data are reasonably well behaved. Exactly which statistics to record and present is a decision made by the researcher, who, presumably, has a much deeper understanding of the data than could possibly be presented in a research article. In addition to these statistics describing the time t cross-sectional distribution of X, we also record the number of entities for which a valid value of X is available in period t and denote this number n_t.

In Table 2.1, we present the annual summary statistics for market beta (β) from our methodologies sample. The results show that, for example, in 1988, the average β of the stocks in the sample is 0.46; the cross-sectional standard deviation of the values of β is 0.48; the sample skewness of β is 0.17; and the sample excess kurtosis of β is 2.80. Furthermore, the minimum, fifth percentile, 25th percentile, median, 75th percentile, 95th percentile, and maximum values of β in 1988 are −4.29, −0.20, 0.13, 0.40, 0.75, 1.31, and 3.28, respectively. Finally, there are 5690 stocks with a valid value of β in 1988.

Table 2.1 presents a detailed account of the cross-sectional distribution of β on a period-by-period basis. In this case, presenting the periodic summary statistics in detail is possible because our sample consists of only 25 periods, and we only present summary statistics for one variable, β. While it is certainly valuable to present all of these statistics, in most empirical asset pricing studies, the sample has many more

TABLE 2.1 Annual Summary Statistics for β

This table presents summary statistics for β for each year during the sample period. For each year t, we calculate the mean ($Mean_t$), standard deviation (SD_t), skewness ($Skew_t$), excess kurtosis ($Kurt_t$), minimum (Min_t), fifth percentile ($P5_t$), 25th percentile ($P25_t$), median ($Median_t$), 75th percentile ($P75_t$), 95th percentile ($P95_t$), and maximum (Max_t) values of the distribution of β across all stocks in the sample. The sample consists of all U.S.-based common stocks in the Center for Research in Security Prices (CRSP) database as of the end of the given year t and covers the years from 1988 through 2012. The column labeled n_t indicates the number of observations for which a value of β is available in the given year.

t	$Mean_t$	SD_t	$Skew_t$	$Kurt_t$	Min_t	$P5_t$	$P25_t$	$Median_t$	$P75_t$	$P95_t$	Max_t	n_t
1988	0.46	0.48	0.17	2.80	−4.29	−0.20	0.13	0.40	0.75	1.31	3.28	5690
1989	0.46	0.53	0.15	1.88	−3.51	−0.27	0.11	0.40	0.79	1.38	3.63	5519
1990	0.58	0.59	0.23	1.14	−3.15	−0.24	0.16	0.51	0.96	1.61	3.66	5409
1991	0.57	0.61	0.23	1.96	−3.28	−0.29	0.17	0.52	0.95	1.62	5.29	5303
1992	0.65	0.83	0.34	6.10	−5.21	−0.50	0.17	0.59	1.09	2.05	9.90	5389
1993	0.62	0.77	−0.10	4.29	−4.70	−0.56	0.20	0.57	1.04	1.90	7.59	5670
1994	0.70	0.71	−0.17	6.59	−6.92	−0.32	0.27	0.67	1.07	1.89	6.50	6148
1995	0.64	0.84	0.30	5.17	−6.32	−0.49	0.19	0.56	1.02	2.15	8.77	6288
1996	0.67	0.64	0.46	1.97	−4.32	−0.20	0.26	0.59	1.01	1.89	3.98	6586
1997	0.53	0.48	0.39	1.46	−2.36	−0.13	0.21	0.48	0.80	1.38	3.20	6867
1998	0.71	0.51	0.49	0.95	−1.80	0.01	0.34	0.67	1.03	1.62	3.75	6608
1999	0.41	0.50	1.39	4.81	−2.21	−0.18	0.11	0.32	0.61	1.33	3.77	6097
2000	0.70	0.72	1.27	1.33	−1.10	−0.06	0.19	0.49	1.01	2.23	3.76	5901
2001	0.76	0.73	1.29	2.13	−1.48	−0.05	0.25	0.60	1.07	2.25	4.21	5508
2002	0.67	0.55	0.70	0.69	−1.19	−0.04	0.25	0.62	0.97	1.73	2.99	5099
2003	0.72	0.56	0.40	0.49	−2.17	−0.04	0.29	0.68	1.06	1.72	3.04	4737
2004	1.03	0.70	0.43	0.24	−1.75	0.01	0.53	0.99	1.46	2.30	4.02	4574
2005	0.95	0.64	0.00	−0.17	−1.60	−0.06	0.46	0.99	1.39	1.96	3.69	4495
2006	1.02	0.70	0.08	0.17	−3.71	−0.02	0.48	1.00	1.51	2.18	3.75	4453
2007	0.87	0.54	−0.04	−0.20	−1.50	0.01	0.45	0.91	1.26	1.72	3.06	4332
2008	0.87	0.53	0.17	0.06	−1.49	0.03	0.48	0.87	1.22	1.74	3.45	4264
2009	1.10	0.72	0.51	0.62	−1.74	0.09	0.55	1.03	1.57	2.36	5.31	3977
2010	1.04	0.54	−0.06	−0.15	−0.85	0.10	0.68	1.05	1.41	1.90	2.95	3805
2011	1.07	0.54	−0.14	−0.37	−0.62	0.14	0.70	1.13	1.45	1.93	3.03	3682
2012	1.04	0.57	0.04	0.48	−2.33	0.11	0.66	1.05	1.40	1.99	3.43	3545

periods than the 25 periods in the methodology sample. Presenting results such as those in Table 2.1 when there are a large number of periods will not only make it difficult to display the periodic summary statistics but will also make it difficult for the reader to get a general understanding of the characteristics of the data. These issues are magnified when, as in most studies, showing summary statistics for several variables is desirable. Thus, while there are certainly interesting patterns to be observed by presenting such a detailed account of each variable, doing so is usually not necessary to inform a reader about the most salient characteristics of the data,

and thus most articles present statistics that are substantially more summarized than the results in Table 2.1. We proceed now to describe how to further summarize the periodic cross-sectional summary statistics.

2.1.2 Average Cross-Sectional Summary Statistics

The second step in the summary statistics procedure is to calculate the time-series averages of the periodic cross-sectional values. For example, the average cross-sectional mean of the variable X, which we denote *Mean* (no subscript), is found by taking the time-series average of the values of $Mean_t$ over all periods t in the sample. Similarly, we calculate the times-series means of the other cross-sectional summary statistics.

For most studies, it is these time-series average values that are presented in the research article. These values describe the average cross section in the sample. This is appropriate when the objective of the study is to examine a cross-sectional phenomenon, as is the case for the analyses in this book. Table 2.2 presents the time-series averages of the annual cross-sectional summary statistics for β. The numbers in the table, therefore, represent the cross-sectional distribution of β for the average year in the methodologies sample. As can be seen, in the average year, the mean value of β is 0.75 and the median value of β is 0.71. Consistent with the mean being slightly greater than the median, in the average year, the skewness of the distribution of β of 0.34 is slightly positive. The cross-sectional distribution of β, in the average year, is leptokurtic because the average excess kurtosis of 1.78 is positive. The average cross-sectional standard deviation of β is 0.62. Finally, in the average year, there are 5198 stocks for which there is a valid value of β.

TABLE 2.2 Average Cross-Sectional Summary Statistics for β
This table presents the time-series averages of the annual cross-sectional summary statistics for β. The table presents the average mean (*Mean*), standard deviation (*SD*), skewness (*Skew*), excess kurtosis (*Kurt*), minimum (*Min*), fifth percentile (*P*5), 25th percentile (*P*25), median (*Median*), 75th percentile (*P*75), 95th percentile (*P*95), and maximum (*Max*) values of the distribution of β, where the average is taken across all years in the sample. The column labeled *n* indicates the average number of observations for which a value of β is available.

Mean	*SD*	*Skew*	*Kurt*	*Min*	*P*5	*P*25	*Median*	*P*75	*P*95	*Max*	*n*
0.75	0.62	0.34	1.78	−2.78	−0.13	0.33	0.71	1.12	1.85	4.40	5198

2.2 PRESENTATION AND INTERPRETATION

In most studies, there are many variables for which summary statistics should be presented. It is usually optimal to present the summary statistics for all variables in a single table. While each paper will present summary statistics in a slightly different manner, the approach we take in this book is to compile a table in which each row

(with the exception of the header row) presents summary statistics for one of the variables.

Table 2.3 gives an example of how we present summary statistics throughout this text. The first column indicates the variable whose summary statistics are presented in the given row. The subsequent columns present the time-series averages of the cross-sectional summary statistics.

The objectives in analyzing the summary statistics are twofold. First, the summary statistics are intended to give a basic overview of the cross-sectional properties of the variables that will be used in the study. This is useful for understanding the types of entities that comprise the sample. Second, the summary statistics can be used to identify any potential issues that may arise when using these variables in statistical analyses. We exemplify how the summary statistics can be used for each of these objectives in the following two paragraphs using the methodology sample and the results in Table 2.3.

The mean column in Table 2.3 can roughly be interpreted as indicating that the average stock in our sample has a β of 0.75, a market capitalization of just over \$2 billion, and a book-to-market ratio of 0.71. More precisely, the table indicates that in the average month, the cross-sectional means of the given variables are as indicated in the table, but we frequently adopt the simpler language used in the previous sentence. The average value of *Size*, which is the natural log of *MktCap*, is 5.04, and the average one-year-ahead excess return is 12.40%.

Table 2.3 shows that for β, the mean and the median are quite similar and, consistent with this, the skewness is quite small in magnitude and values of β are reasonably symmetric about the mean. The distribution of β is also slightly leptokurtic as the excess kurtosis of its cross-sectional distribution in the average year is 1.78.

The results for *MktCap* show that the distribution of market capitalization is highly positively skewed. This is driven by a small number of observations that have very large values of *MktCap*. The summary statistics therefore indicate that the sample is comprised predominantly of low-market capitalization stocks along with a few stocks that have very high market capitalizations. The median stock in the sample has a market capitalization of \$188 million, which is much smaller than the mean of more than \$2 billion. It is also worth noting that the smallest value of *MktCap* of 0, which means that the stock has market capitalization of less than \$0.5 million, is less than 0.02 standard deviations from the median and less than 0.1 standard deviations from the mean. This indicates that a very large portion of the variability of *MktCap* comes from extremely large values, consistent with the high positive skewness. The distribution of *MktCap* presents potential issues for statistical analyses, such as regression, that rely on the magnitude of the variables used, as the data points corresponding to the very large values may exert undesirably strong influence on the results of such analyses. Therefore, most empirical studies use *Size*, defined as the natural log of *MktCap*, in such analyses. Table 2.3 shows that the distribution of *Size* is much more symmetric than that of *MktCap*, as the average skewness is only 0.32. Furthermore, the excess kurtosis of -0.07 indicates that tails of the distribution of *Size* are, in the average year, very similar to those of a normal distribution. *Size*, therefore, appears much better suited for use in statistical analyses than *MktCap*.

TABLE 2.3 Summary Statistics for β, *MktCap*, and *BM*

This table presents summary statistics for our sample. The sample covers the years t from 1988 through 2012, inclusive, and includes all U.S.-based common stocks in the CRSP database. Each year, the mean (*Mean*), standard deviation (*SD*), skewness (*Skew*), excess kurtosis (*Kurt*), minimum (*Min*), fifth percentile (5%), 25th percentile (25%), median (*Median*), 75th percentile (75%), 95th percentile (95%), and maximum (*Max*) values of the cross-sectional distribution of each variable are calculated. The table presents the time-series means for each cross-sectional value. The column labeled n indicates the average number of stocks for which the given variable is available. β is the beta of a stock calculated from a regression of the excess stock returns on the excess market returns using all available daily data during year t. *MktCap* is the market capitalization of the stock calculated on the last trading day of year t and recorded in $millions. *Size* is the natural log of *MktCap*. *BM* is the ratio of the book value of equity to the market value of equity. r_{t+1} is the one-year-ahead excess stock return.

	Mean	*SD*	*Skew*	*Kurt*	*Min*	5%	25%	*Median*	75%	95%	*Max*	n
β	0.75	0.62	0.34	1.78	−2.78	−0.13	0.33	0.71	1.12	1.85	4.40	5198
MktCap	2030	10,230	14.20	282.85	0	9	48	188	802	7524	287,033	5550
Size	5.04	2.07	0.32	−0.07	−1.19	1.89	3.56	4.91	6.39	8.70	12.33	5550
BM	0.71	2.90	−9.49	1,226.68	−124.31	0.05	0.29	0.57	0.97	2.11	44.87	4273
r_{t+1}	12.40	80.83	5.94	125.33	−97.86	−67.46	−26.87	0.90	31.84	124.54	1,841.43	5381

As for the book-to-market ratio (*BM*), Table 2.3 shows that while the vast majority of the *BM* values fall between 0.05 (the fifth percentile) and 2.11 (the 95th percentile), the tails of the distribution are extremely long, as the kurtosis of *BM* is greater than 1226. Interestingly, despite the fact that the mean is greater than the median, in the average month, the distribution of *BM* is negatively skewed, as the average cross-sectional skewness of *BM* is −9.49.

Finally, the table indicates that the average one-year-ahead excess return (r_{t+1}) of the stocks in the methodology sample is 12.40% per year. The cross-sectional distribution of r_{t+1} is highly skewed and leptokurtic, with an average skewness of 5.94 and excess kurtosis of more than 125. This is driven by the fact that the minimum possible return is −100%, whereas there is no upper bound on the value that r_{t+1} can take. Table 2.3 shows that, in the average year, the maximum r_{t+1} is more than 1841%, with more than 5% of stock realizing excess returns greater than 100%.

There is one more aspect of the return data that is worth mentioning because it is not apparent in the presentation of the summary statistics. The latest data in the version of the CRSP database used to construct the methodology sample are from 2012. However, when t corresponds to year 2012, then r_{t+1} corresponds to excess returns from 2013. Unfortunately, return data for 2013 are not available. Thus, the summary statistics for r_{t+1} reported in Table 2.3 actually cover returns for the 24 years from 1989 through 2012, whereas the summary statistics for the other variables cover the 25 years from 1988 through 2012. While this detail of the summary statistics is not usually discussed in a research article because it rarely has a meaningful impact on the interpretation of the results, it is something that should be clearly understood by the researcher.

Although Table 2.3 is certainly expository, there are many characteristics of the data that are not captured in the highly summarized results. The most important drawback of summarizing the data in such a manner is that it does not indicate any time-series variation in the variables used in the study. For example, referring back to Table 2.1, it is evident that the mean and median values of β increase quite substantially over time. This feature of the data is not in any way captured in the summary presented in Table 2.3. Additionally, given that the values of market capitalization (*MktCap*) have not been adjusted for inflation, it is reasonable to assume that value of *MktCap* will exhibit generally increasing pattern over time as well. This is confirmed in unreported results. Furthermore, values of *MktCap* are likely to drop when the stock market experiences a large loss and increase as the stock market realizes gains. The opposite would be true for the book-to-market ratio (*BM*) as the market capitalization is the denominator of this variable, although in this case the increase in values of *BM* may be delayed due to the timing of the calculation of *BM*, which is discussed in detail in Chapter 10.

None of these characteristics of the data are captured in the summary statistics as presented in Table 2.3. In most cases, these details are not very important when interpreting and drawing conclusions from the results of subsequent analyses in the article. However, as a researcher, it is important to be aware of such patterns and to assess whether these patterns may have a significant impact on the main conclusions of the study. In many cases, this is done by subsample analyses aimed at examining

whether the main conclusions hold in both early periods of the study as well as in late periods. Frequently, it is also worthwhile to investigate whether the main results hold in periods of normal economic conditions as well as periods of deteriorating or poor economic conditions. This is especially the case if the summary statistics for the focal variables of the study are substantially different for these subperiods.

2.3 SUMMARY

In summary, the main objective of presenting summary statistics is to give the reader a sufficient but succinct understanding of the data being used and the characteristics of the entities that comprise the sample. In addition, the summary statistics can be used to identify and remedy any potential issues with using statistical analysis on the data. The approach that we have discussed presents the distribution of the given variables in the average cross section. While the results presented in the summary statistics table may be sufficient for a reader, they are likely not sufficient for the researcher. It is difficult to conduct high-quality research without having an in-depth understanding of the data. A good researcher will understand any potential issues with the data that are not evident in the summary statistics and address these issues in the statistical analyses presented in the research article.

3

CORRELATION

Summary statistics, discussed in Chapter 2, provide an overview of the univariate distributions of the variables used in a study. They do not, however, give any indication as to the relations between the variables. Understanding how the variables relate to each other is usually more important than understanding the variables' univariate characteristics, as in almost all cases, it is the relations that are the focus of the research. Therefore, in addition to presenting univariate summary statistics, researchers frequently present correlations between the main variables. Correlations provide a preliminary look at the bivariate relations between pairs of variables used in the study.

This chapter introduces a widely used methodology for calculating and presenting correlations. As with the summary statistics procedure presented in Chapter 2, the objective of the methodology discussed in this chapter is to understand the cross-sectional properties of the variables. This technique is therefore most appropriate when the economic phenomenon under investigation is cross-sectional in nature. While most studies present only Pearson product–moment correlations, here and in the remainder of this book, we will present both the Pearson product–moment correlations and the Spearman rank correlation.

The Pearson product–moment correlation is most applicable when the relation between the two variables, which we denote X and Y, is thought to be linear. If this is the case, the Pearson correlation can be roughly interpreted as the signed percentage of variation in X that is related to variation in Y, with the sign being positive if X

Empirical Asset Pricing: The Cross Section of Stock Returns, First Edition.
Turan G. Bali, Robert F. Engle, and Scott Murray.

tends to be high when Y is high, and the sign being negative when high values of X tend to correspond to low values of Y. The Pearson correlation can take values between -1 and 1, with -1 indicating a perfectly negative linear relation, 0 indicating no linear relation between the variables, and 1 indicating a perfectly positive linear relation.

The Spearman rank correlation is most applicable when the relation between the variables is thought to be monotonic, but not necessarily linear. The rank correlation, as the name implies, measures how closely related the ordering of X is to the ordering of Y, with no regard to the actual values of the variables. As with the product–moment correlation, the rank correlation can take on values between -1 and 1, with a Spearman correlation of 1 indicating that X and Y are perfectly monotonically increasing functions of each other and a value of -1 indicating that X and Y are perfectly monotonically decreasing functions of each other.

3.1 IMPLEMENTATION

Similar to the summary statistics procedure, the correlation procedure is executed in two steps. The first step is to calculate the cross-sectional correlation between the two variables in question, X and Y, for each period t. The second step is to take the time-series average of these cross-sectional correlations.

3.1.1 Periodic Cross-Sectional Correlations

In step one, for each time period t, we calculate the Pearson product–moment correlation and the Spearman rank correlation between X and Y. The Pearson product–moment correlation between X and Y for period t is defined as

$$\rho_t(X,Y) = \frac{\sum_{i=1}^{n_t}(X_{i,t} - \overline{X}_t)(Y_{i,t} - \overline{Y}_t)}{\sqrt{\sum_{i=1}^{n_t}(X_{i,t} - \overline{X}_t)^2}\sqrt{\sum_{i=1}^{n_t}(Y_{i,t} - \overline{Y}_t)^2}} \tag{3.1}$$

where each of the summations is taken over all entities i in the sample for which there are valid values of both X and Y in period t, and $\overline{X}_t$ and $\overline{Y}_t$ are the sample means of $X_{i,t}$ and $Y_{i,t}$, respectively, taken over the same set of entities. Here, n_t is the number of entities for which there are valid values of both X and Y in the given period t. In many cases, the values of X and Y are winsorized prior to calculating the Pearson product–moment correlation to minimize the effect of a small number of extreme observations. Winsorization is performed on a period-by-period basis using only entities for which valid values of both X and Y are available.

To calculate the Spearman rank correlation, one must first calculate the ranking for each entity i on each of X and Y. We let $x_{i,t}$ be the rank of $X_{i,t}$ calculated over all entities that have valid values of both X and Y during period t. Thus, if entity i has the lowest value of X, $x_{i,t}$ is 1. If entity i has the highest value of X, then $x_{i,t}$ is n_t. If there are multiple entities for which the value of X is the same, then each of

these entities is assigned a ranking equal to the average position of these entities in the ordered list of the entities when sorted on the variable X. The rankings for Y are calculated analogously and are denoted $y_{i,t}$. It should be noted that when calculating the Spearman rank correlation, the data should not be winsorized. For each entity i, the difference between the entity's ranking on X and it's ranking on Y is defined as $d_{i,t} = x_{i,t} - y_{i,t}$. Finally, the Spearman rank correlation between X and Y for period t is calculated as

$$\rho_t^S(X, Y) = 1 - \frac{6 \sum_{i=1}^{n_t} d_{i,t}^2}{n_t(n_t^2 - 1)}. \tag{3.2}$$

We exemplify the cross-sectional step of the correlation procedure by calculating both the Pearson product–moment correlation ($\rho_t(X, Y)$) and the Spearman rank correlation ($\rho_t^S(X, Y)$) between each pair of the variables β (beta), *Size* (log of market capitalization in $millions), *BM* (book-to-market ratio), and r_{t+1} (one-year-ahead excess return), for each year t during our sample period. Pearson product–moment correlations are calculated after winsorizing both of the variables at the 0.5% level using only data point for which both variables in the given calculation have valid values.

In Table 3.1, we present the Pearson product–moment and Spearman rank cross-sectional correlations between each pair of variables during each year t of our sample. The table shows that, in all years, β and *Size* are positively correlated, regardless of which measure of correlation is used. β and *BM* exhibit negative correlation in all years except for 2009, when this correlation is positive but small in magnitude. The relation between β and r_{t+1} varies substantially over time. *Size* and *BM* have a negative correlation in all time periods. This is not surprising given that market capitalization is the denominator of the calculation of *BM* and *Size* is the log of market capitalization. Thus, this effect is likely mechanical. The signs of the correlation between *Size* and r_{t+1}, as well as between *BM* and r_{t+1}, vary over time. Finally, it is worth noting that for year 2012 there are no correlations for pairs of variables that include r_{t+1}. This is because for $t = 2012$, r_{t+1} is the excess return in 2013, which is not available in the version of the Center for Research in Security Prices (CRSP) database used to generate the methodologies sample.

3.1.2 Average Cross-Sectional Correlations

Step two in the correlation procedure is to calculate the time-series averages of the periodic cross-sectional correlations between each pair of variables. These values represent the correlations in the average period. The time-series average correlations for each pair of variables used in the example are presented in Table 3.2. We denote these time-series averages as $\rho(X, Y)$ for the Pearson product–moment correlation and $\rho^S(X, Y)$ for the Spearman rank correlation. We therefore have

$$\rho(X, Y) = \frac{\sum_{t=1}^{N} \rho_t(X, Y)}{N} \tag{3.3}$$

TABLE 3.1 Annual Correlations for β, *Size*, *BM*, and r_{t+1}
This table presents the cross-sectional Pearson product–moment (ρ_t) and Spearman rank (ρ_t^S) correlations between pairs of β, *Size*, *BM*, and r_{t+1}. Each column presents either the Pearson or Spearman correlation for one pair of variables, indicated in the column header. Each row represents results from a different year, indicated in the column labeled *t*.

t	$\rho_t(\beta, Size)$	$\rho_t^S(\beta, Size)$	$\rho_t(\beta, BM)$	$\rho_t^S(\beta, BM)$	$\rho_t(\beta, r_{t+1})$	$\rho_t^S(\beta, r_{t+1})$	$\rho_t(Size, BM)$	$\rho_t^S(Size, BM)$	$\rho_t(Size, r_{t+1})$	$\rho_t^S(Size, r_{t+1})$	$\rho_t(BM, r_{t+1})$	$\rho_t^S(BM, r_{t+1})$
1988	0.47	0.45	−0.10	−0.12	0.04	0.06	−0.15	−0.11	0.13	0.24	0.04	0.04
1989	0.44	0.45	−0.15	−0.16	0.02	0.02	−0.14	−0.11	0.07	0.17	0.01	0.05
1990	0.43	0.45	−0.17	−0.23	0.07	0.15	−0.19	−0.16	−0.07	0.10	−0.04	−0.05
1991	0.45	0.49	−0.09	−0.16	−0.09	−0.09	−0.23	−0.22	−0.15	−0.03	0.13	0.19
1992	0.34	0.37	−0.20	−0.29	−0.10	−0.10	−0.20	−0.17	−0.14	−0.04	0.10	0.20
1993	0.36	0.38	−0.18	−0.25	−0.01	−0.03	−0.19	−0.17	−0.00	0.07	0.11	0.16
1994	0.31	0.35	−0.18	−0.22	0.03	0.02	−0.17	−0.11	−0.00	0.11	0.01	0.06
1995	0.30	0.32	−0.16	−0.21	−0.06	−0.08	−0.21	−0.19	−0.01	0.09	0.10	0.14
1996	0.30	0.32	−0.26	−0.36	−0.17	−0.19	−0.21	−0.17	0.04	0.10	0.12	0.21
1997	0.42	0.43	−0.23	−0.29	0.03	−0.00	−0.20	−0.18	0.08	0.18	0.03	0.07
1998	0.38	0.40	−0.25	−0.33	0.15	0.11	−0.24	−0.25	−0.09	−0.04	−0.04	−0.03
1999	0.48	0.47	−0.24	−0.32	−0.11	−0.10	−0.34	−0.38	0.04	0.07	0.03	0.08
2000	0.23	0.27	−0.39	−0.54	−0.20	−0.27	−0.27	−0.26	−0.19	−0.14	0.08	0.17
2001	0.32	0.38	−0.20	−0.34	−0.38	−0.44	−0.32	−0.40	−0.13	−0.09	0.17	0.25
2002	0.46	0.55	−0.23	−0.30	0.01	0.04	−0.27	−0.29	−0.23	−0.15	0.05	0.04
2003	0.51	0.59	−0.13	−0.17	−0.14	−0.13	−0.24	−0.31	−0.08	0.03	0.11	0.10
2004	0.32	0.40	−0.26	−0.28	−0.09	−0.08	−0.17	−0.13	0.06	0.14	0.08	0.13
2005	0.45	0.50	−0.16	−0.15	−0.01	0.03	−0.13	−0.11	−0.01	0.08	0.07	0.12
2006	0.41	0.47	−0.20	−0.22	0.07	0.06	−0.17	−0.15	0.12	0.19	−0.02	−0.04
2007	0.47	0.52	−0.12	−0.14	−0.01	−0.01	−0.17	−0.17	0.09	0.16	−0.01	0.00
2008	0.44	0.48	−0.09	−0.13	0.03	0.09	−0.24	−0.23	−0.18	−0.04	0.08	−0.06
2009	0.31	0.37	0.02	0.01	0.11	0.13	−0.28	−0.34	−0.00	0.09	0.03	0.04
2010	0.39	0.39	−0.18	−0.15	−0.10	−0.12	−0.31	−0.28	0.14	0.18	−0.01	0.00
2011	0.37	0.36	−0.26	−0.22	−0.05	−0.02	−0.29	−0.27	−0.04	0.04	0.12	0.12
2012	0.35	0.35	−0.17	−0.17			−0.32	−0.32				

and

$$\rho^S(X, Y) = \frac{\sum_{t=1}^{N} \rho_t^S(X, Y)}{N} \tag{3.4}$$

where N is the number of periods in the sample.

3.2 INTERPRETING CORRELATIONS

The correlations give preliminary indications of the nature of the cross-sectional relations between each pair of variables. If two variables that are measured

TABLE 3.2 Average Correlations for β, *Size*, *BM*, and r_{t+1}
This table presents the time-series averages of the annual cross-sectional Pearson product–moment (ρ) and Spearman rank (ρ^S) correlations between pairs of β, *Size*, *BM*, and r_{t+1}. Each column presents either the Pearson or Spearman correlation for one pair of variables, indicated in the column header.

$\rho(\beta, Size)$	$\rho^S(\beta, Size)$	$\rho(\beta, BM)$	$\rho^S(\beta, BM)$	$\rho(\beta, r_{t+1})$	$\rho^S(\beta, r_{t+1})$	$\rho(Size, BM)$	$\rho^S(Size, BM)$	$\rho(Size, r_{t+1})$	$\rho^S(Size, r_{t+1})$	$\rho(BM, r_{t+1})$	$\rho^S(BM, r_{t+1})$
0.39	0.42	−0.18	−0.23	−0.04	−0.04	−0.23	−0.22	−0.02	0.06	0.06	0.08

contemporaneously exhibit correlations that are very high in magnitude, this indicates that the information content of both variables is very similar and that the two variables are likely capturing the same characteristic of the entity. If variables that are not measured contemporaneously exhibit strong correlation, this is an indication that one variable (the variable measured chronologically earlier) may be a predictor of future values of the other variable. In making such a determination, it is important to ensure that such predictive power is not mechanical. To do so usually requires an in-depth understanding of exactly how the variables are calculated. If the correlation between a pair of variables is close to zero, this indicates that the variables contain completely different information regarding the underlying entities.

In addition to providing preliminary indications on the relations between the variables, correlation analysis can indicate potential issues associated with multivariate statistical analyses. For example, if two variables are very highly correlated, either positively or negatively, regression analyses that include both variables as independent variables in a regression specification may have difficulty distinguishing between the effect of one variable and the other on the dependent variable. This results in high standard errors on the regression coefficients. If the Spearman rank correlation is substantially larger in magnitude than the Pearson product–moment correlation, this likely indicates that there is a monotonic, but not linear, relation between the variables. This type of relation signals that linear regression analysis is a potentially problematic statistical technique to apply to the given variables if one of the variables is used as the dependent variable. If the Pearson product–moment correlation is substantially larger in magnitude than the Spearman rank correlation, this may indicate that there are a few extreme data points in one of the variables that are exerting a strong influence on the calculation of the Pearson product–moment correlation. In this case, it is possible that winsorizing one or both of the variables at a higher level will alleviate this issue. Finally, it is worth noting here that, because of the assumption of linearity in the calculation of the Pearson product–moment correlation, this measure is usually more indicative of results that will be realized using regression techniques such as Fama and MacBeth (1973) regression analysis (presented in Chapter 6). Because the Spearman rank correlation is based on the ordering of the variables, Spearman rank correlations are more likely indicative of the results of analyses that rely on

the ranking, or ordering, of the variables, such as portfolio analysis (presented in Chapter 5).

The average Pearson product–moment correlation of 0.39 between β and *Size* indicates that larger stocks tend to have higher betas. Stated alternatively, this correlation indicates that stocks with high betas tend to be larger. That being said, the correlation is not so high as to indicate that the two variables are capturing essentially the same information. There is certainly a substantial component of β that is orthogonal to *Size* and a substantial component of *Size* that is orthogonal to β. Thus, while there is an economically important relation between beta and size, they certainly cannot be seen as the same. The average Spearman rank correlation between β and *Size* of 0.42 is quite similar to the Pearson product–moment correlation. The results also indicate an economically important negative relation between β and *BM*, since the Pearson product–moment (Spearman rank) correlation between these variables is −0.18 (−0.23). The magnitude of these correlations indicates once again that while there is a substantial common component to these variables, there is also a very substantial component of each of these variables that is orthogonal to the other. The same conclusions hold when examining the correlations between *Size* and *BM*. Once again, the Pearson product–moment correlation of −0.23 and Spearman rank correlation of −0.22 are very similar in magnitude and indicate a moderate negative cross-sectional relation between *Size* and *BM*. Thus, while each of these pairs of variables exhibit some cross-sectional correlation, the correlations are low enough to alleviate concerns about potential statistical issues when several of these variables are included in multivariate statistical analyses. Furthermore, the Pearson product–moment and Spearman rank measures are similar enough to alleviate any serious concerns about potential data issues or severe lack of linearity in the relations between these variables. It is important to realize that β, *Size*, and *BM* are all measured contemporaneously; thus, in the analysis of these correlations, the primary objective is to assess the information content of each of these variables. It is also important to realize that just because the magnitudes of the pairwise correlations are not high enough to raise concern about subsequent statistical analysis, it remains possible that some combination of two of these variables is highly correlated with a third variable (multicollinearity). Correlation analysis cannot detect such issues.

The one-year-ahead excess return (r_{t+1}) is measured in the year subsequent to the time at which each of the other variables (β, *Size*, and *BM*) are calculated. Thus, correlation between r_{t+1} and any of these variables is likely indicative of a predictive relation. Furthermore, because each of β, *BM*, and *Size* are calculated using information that is readily available in year t, and r_{t+1} is calculated using only information that is generated during year $t + 1$, we are not concerned about a potential mechanical effect between r_{t+1} and any of the other variables. The results in Table 3.2 indicate a slightly negative average Pearson and Spearman correlations of −0.04 between β and r_{t+1}, indicating that, in the average year, high β stocks may generate lower excess returns than low β stocks. While this result is inconsistent with the predictions of the Capital Asset Pricing Model of Sharpe (1964), Lintner (1965), and Mossin (1966), we will postpone in-depth economic analysis of this result until the chapter that studies the relation between β and future stock returns in depth (Chapter 8). The Pearson

product–moment correlation between *Size* and r_{t+1} of −0.02 indicates almost no relation between *Size* and future excess stock returns, whereas the positive Spearman rank correlation of 0.06 indicates a slightly positive relation. While it is a little bit concerning that the two measures of correlation have, on average, the opposite sign, the magnitudes of these correlations are small enough so that we are not overly worried about this result. Finally, the results indicate a positive relation between *BM* and future stock returns, as the Pearson product–moment correlation of 0.06 and Spearman rank correlation of 0.08 are larger than any other correlation that includes r_{t+1}. It should be noted that, while the correlations between r_{t+1} and the other variables are all quite small in magnitude, as will be seen throughout the remainder of this text, what seems here to be only a minimal ability to predict future stock returns may be indicative of a very strong and important economic phenomenon.

3.3 PRESENTING CORRELATIONS

The standard way to present correlations is in a correlation matrix. Each row corresponds to one variable, indicated in the first column of the table. Similarly, each column corresponds to a variable, indicated in the first row of the table. The remaining entries in the table present the average cross-sectional correlations between the row and column variables. Diagonal entries, which represent the correlation between a variable and itself (equal to 1.00 by definition), are either left blank or the number 1.00 is displayed. In this book, we will leave these entries blank, as we feel that doing so makes for a cleaner presentation. If only the Pearson product–moment correlation is used, frequently only the entries below the diagonal or the entries above the diagonal entry are presented to avoid repetition. Here, and in the remainder of this book, we present both the average Pearson product–moment correlations and average Spearman rank correlations. The below-diagonal entries show the average Pearson product–moment correlations and the above-diagonal entries present the Spearman rank correlations. For the reasons discussed in the previous section, we feel it is valuable to present both types of correlations. Table 3.3 presents the average pairwise correlations between β, *Size*, *BM*, and r_{t+1} for our sample of stocks.

TABLE 3.3 Correlations Between β, *Size*, *BM*, and r_{t+1}
This table presents the time-series averages of the annual cross-sectional Pearson product–moment and Spearman rank correlations between pairs of β, *Size*, *BM*, and r_{t+1}. Below-diagonal entries present the average Pearson product–moment correlations. Above-diagonal entries present the average Spearman rank correlation.

	β	*Size*	*BM*	r_{t+1}
β		0.42	−0.23	−0.04
Size	0.39		−0.22	0.06
BM	−0.18	−0.23		0.08
r_{t+1}	−0.04	−0.02	0.06	

3.4 SUMMARY

In summary, correlation analysis gives us a first look at the relations between the variables used in a study. The procedure discussed in this chapter is designed to examine the cross-sectional correlation between pairs of variables, and the results presented are indicative of the relation between each pair of variables during the average period in the sample. We use two different measures of correlation. The first is the Pearson product–moment correlation, which is designed to indicate the strength of a linear relation between the two variables. The second is the Spearman rank correlation, which detects monotonicity in the relation between the two variables. Large differences between the two measures of correlation should be taken as indications that the data need to be examined in more depth to assess the cause of this difference.

REFERENCES

Fama, E. F. and MacBeth, J. D. 1973. Risk, return, and equilibrium: empirical tests. Journal of Political Economy, 81(3), 607.

Lintner, J. 1965. Security prices, risk, and maximal gains from diversification. Journal of Finance, 20(4), 687–615.

Mossin, J. 1966. Equilibrium in a capital asset market. Econometrica, 34(4), 768–783.

Sharpe, W. F. 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. Journal of Finance, 19(3), 425–442.

4

PERSISTENCE ANALYSIS

Many of the variables in empirical asset pricing research are intended to capture persistent characteristics of the entities in the sample. This means that the characteristic of the entity that is captured by the given variable is assumed to remain reasonably stable over time. Such variables are frequently estimated using historical data, and the value calculated from the historical data is assumed to be a good estimate of the given characteristic for the entity going forward. For example, the value of a stock's beta from the Capital Asset Pricing Model (Sharpe (1964), Lintner (1965), Mossin (1966)) is generally assumed to be a persistent characteristic of the stock, and it is frequently estimated from regressions of the stock's returns on the returns of the market portfolio using historical data. This is exactly how our variable β is calculated.

In this chapter, we discuss a technique that we call persistence analysis. We use persistence analysis to examine whether a given characteristic of the entities in our sample is in fact persistent. Persistence analysis can also be used to examine the ability of the variable in question to capture the desired characteristic of the entity. The basic approach is to examine the cross-sectional correlation between the given variable measured at two different points in time. If this correlation is high, this indicates that the variable is persistent, whereas low correlations indicate little or no persistence. This technique is not as widely used in the empirical asset pricing literature as the other techniques presented in Part I. We discuss it here and use it throughout this text because one of the objectives of this book is to provide a thorough understanding of the variables most commonly used throughout the empirical asset pricing literature.

Empirical Asset Pricing: The Cross Section of Stock Returns, First Edition.
Turan G. Bali, Robert F. Engle, and Scott Murray.

4.1 IMPLEMENTATION

As with the other methodologies presented in this text, implementation of persistence analysis is done in two steps. The first step involves calculating cross-sectional correlations between the given variable X measured a certain number of periods apart. The second step involves calculating the time-series average of each of these cross-sectional correlations.

4.1.1 Periodic Cross-Sectional Persistence

The first step in persistence analysis is to calculate the cross-sectional correlation between the variable under consideration, X, measured τ periods apart. This will be done for each time period t where both the time period t and the time period $t + \tau$ fall during the sample period. The entities used to calculate the cross-sectional correlation will be all entities i for which a valid value of the variable X is available for both period t and period $t + \tau$. For each time period t, we therefore define $\rho_{t,t+\tau}(X)$ as the cross-sectional Pearson product–moment correlation between X measured at time t and X measured at time $t + \tau$. Specifically, we have

$$\rho_{t,t+\tau}(X) = \frac{\sum_{i=1}^{n_t}[(X_{i,t} - \overline{X}_t)(X_{i,t+\tau} - \overline{X}_{t+\tau})]}{\sqrt{\sum_{i=1}^{n_t}(X_{i,t} - \overline{X}_t)^2}\sqrt{\sum_{i=1}^{n_t}(X_{i,t+\tau} - \overline{X}_{t+\tau})^2}} \tag{4.1}$$

where $\overline{X}_t$ is the mean value of $X_{i,t}$ and the summations and means are taken over all entities i for which a valid value of X is available in both periods t and $t + \tau$. n_t represents the number of such entities. Frequently, before the correlations are calculated, the values of X from month t are winsorized to remove the effect of outliers. The values of X from month $t + \tau$ are separately winsorized at the same level.

We illustrate this using β and values of τ between 1 and 5 inclusive. Our analysis will therefore examine the persistence of β measured one, two, three, four, and five years apart. Prior to calculating the cross-sectional correlations for each period t, the data are winsorized at the 0.5% level. To be perfectly clear, for each month t, we first find all entities that have valid values of β in both periods t and $t + \tau$. We then winsorize the corresponding values of β in each of the months t and $t + \tau$ separately. The annual values for these cross-sectional correlations are presented in Table 4.1. The year t is presented in the first column and the subsequent columns present the values of $\rho_{t,t+\tau}(\beta)$ for $\tau \in \{1, 2, 3, 4, 5\}$.

The results in Table 4.1 indicate that values of β measured one year apart ($\rho_{t,t+1}(\beta)$) exhibit cross-sectional correlations between 0.39 ($t = 1992$) and 0.80 ($t = 2008$). As might be expected, the correlations between β measured at longer lags τ tend to be lower than the correlations measured at shorter lags, although this is not always the case. When measured five years apart ($\rho_{t,t+5}(\beta)$), the table indicates that the correlation between β and its lagged counterpart is between 0.25 ($t = 2000$) and 0.56 ($t = 2006$). We withhold further interpretation of the results until later in the chapter.

TABLE 4.1 Annual Persistence of β

This table presents the cross-sectional Pearson product–moment correlations between β measured in year t and β measured in year $t+\tau$ for $\tau \in \{1,2,3,4,5\}$. The first column presents the year t. The subsequent columns present the cross-sectional correlations between β measured at time t and β measured at time $t+1$, $t+2$, $t+3$, $t+4$, and $t+5$.

t	$\rho_{t,t+1}(\beta)$	$\rho_{t,t+2}(\beta)$	$\rho_{t,t+3}(\beta)$	$\rho_{t,t+4}(\beta)$	$\rho_{t,t+5}(\beta)$
1988	0.50	0.48	0.47	0.39	0.34
1989	0.52	0.45	0.38	0.35	0.36
1990	0.55	0.45	0.42	0.40	0.37
1991	0.46	0.43	0.41	0.37	0.40
1992	0.39	0.37	0.36	0.41	0.38
1993	0.40	0.33	0.38	0.39	0.39
1994	0.38	0.39	0.38	0.37	0.33
1995	0.46	0.44	0.38	0.40	0.48
1996	0.53	0.48	0.43	0.52	0.53
1997	0.55	0.46	0.48	0.51	0.53
1998	0.51	0.50	0.53	0.53	0.50
1999	0.57	0.59	0.53	0.52	0.40
2000	0.79	0.58	0.56	0.58	0.25
2001	0.70	0.66	0.62	0.35	0.41
2002	0.79	0.64	0.50	0.49	0.38
2003	0.70	0.54	0.51	0.42	0.39
2004	0.62	0.60	0.45	0.38	0.34
2005	0.73	0.60	0.55	0.48	0.53
2006	0.67	0.56	0.50	0.56	0.56
2007	0.69	0.60	0.60	0.59	0.51
2008	0.80	0.69	0.64	0.60	
2009	0.73	0.65	0.59		
2010	0.76	0.70			
2011	0.79				

However, it is worth noting that for years t toward the end of the sample, in some cases the persistence values are missing. The reason for this is that, for example, in year 2009, to calculate the correlation between β measured in 2009 and β measured four years in the future ($\tau = 4$), we would need data from year 2013. As these data are not available in the version of the Center for Research in Security Prices (CRSP) database used to construct the methodology sample, we are unable to calculate this value. The reasons for the other missing entries are analogous.

4.1.2 Average Cross-Sectional Persistence

Although periodic cross-sectional persistence values such as those presented in Table 4.1 are quite informative, they are quite difficult to read and draw conclusions from. We therefore want to summarize these periodic values more succinctly. As with the other analyses we discuss, the main objective is to understand the persistence of the variable X in the average cross section. We therefore summarize the results by simply taking the time-series average of the periodic cross-sectional correlations. We denote these average persistence values using $\rho_\tau(X)$ where the subscript indicates the number of lags. Specifically, we have

$$\rho_\tau(X) = \frac{\sum_{t=1}^{N-\tau} \rho_{t,t+\tau}(X)}{N-\tau} \tag{4.2}$$

where N is the number of periods in the sample. Throughout the remainder of this book, we will refer to these values as the persistence of X at lag τ.

In Table 4.2, we present the persistence of β at lags of one, two, three, four, and five years. The results indicate that, consistent with what was observed in the annual persistence values presented in Table 4.1, the persistence of values of β measured one year apart, 0.62, is quite strong. The level of persistence drops off substantially as the amount of time between the measurement periods increases. When measured at a lag of 5 years, the average persistence of β has decreased to 0.42.

4.2 INTERPRETING PERSISTENCE

Interpreting the results of the persistence analyses is fairly straightforward. In general, a higher degree of time-lagged cross-sectional correlation in the given variable is indicative of higher persistence, although there are several caveats to this that must be understood to properly make use of this technique.

We begin our discussion of the interpretation of persistence analysis results by discussing potential causes of low persistence. Exactly what qualifies as low persistence is not perfectly well defined and depends on how long the lag is between the times of measurement (τ), how persistent the actual characteristic being captured by the variable is thought to be, and how accurately the variable is expected to capture the actual characteristic. There are generally two reasons that a variable may exhibit low or zero persistence. The first is that the characteristic being measured is in fact

TABLE 4.2 Average Persistence of β
This table presents the time-series averages of the cross-sectional Pearson product–moment correlations between β measured in year t and β measured in year $t+\tau$ for $\tau \in \{1, 2, 3, 4, 5\}$.

$\rho_1(\beta)$	$\rho_2(\beta)$	$\rho_3(\beta)$	$\rho_4(\beta)$	$\rho_5(\beta)$
0.61	0.53	0.48	0.46	0.42

not persistent. The second is that the variable used to proxy for the given characteristic does a poor job at measuring the characteristic under examination. In this case, even if the given characteristic of the entities in the sample is highly cross-sectionally persistent, the failure of the variable X to capture cross-sectional variation in this characteristic will cause the persistence analysis to generate a low value of $\rho_\tau(X)$. In this sense, the persistence analysis suffers from a dual hypothesis problem, as failure to find persistence does not necessarily indicate a lack of persistence in the characteristic under investigation. Low values of $\rho_\tau(X)$ may also indicate a failure of X to capture that characteristic. Thus, we must be careful when concluding that a certain characteristic of the entities in the sample is not cross-sectionally persistent based on the results of the persistence analysis. To reach such a conclusion, we must be highly confident that the variable X does in fact capture the characteristic under examination. On the other hand, if one is extremely confident, for reasons beyond the scope of the persistence analysis, that the characteristic in question is in fact highly cross-sectionally persistent, low values of $\rho_{t,t+\tau}(X)$ likely indicate that the variable X does a poor job at capturing cross-sectional variation in the characteristic. In the end, however, regardless of the reason for the lack of persistence in X, if X is intended to capture a persistent characteristic of a firm, but X does not exhibit persistence, then X is not a good measure of the characteristic of interest.

When the persistence analysis produces high values of $\rho_\tau(X)$, this very likely indicates both that the characteristic in question is in fact persistent and that the variable X does a good job at measuring the characteristic. There are two caveats with this statement that must be addressed. The first is that it is possible that the variable X is unintentionally capturing some persistent characteristic of the entities in the sample that is different from the characteristic that X is designed to capture. Thus, perhaps a more correct statement is that high values of $\rho_\tau(X)$ indicate that whatever characteristic is being captured by X is in fact persistent. If X does in fact capture the intended characteristic, then we can conclude that the characteristic is in fact persistent. Therefore, assuming sufficient effort has been devoted to designing the calculation of X such that it can reasonably be expected to capture the intended characteristic, high values of $\rho_\tau(X)$ are interpreted as indicating that the given characteristic is in fact persistent.

The second, and much more important, caveat associated with concluding that a characteristic is persistent is that in many cases there is a mechanical reason related to the calculation of X that would result in strong cross-sectional correlation between X_t and $X_{t+\tau}$ even if the characteristic in question is not persistent. In most cases, the reason for such a mechanical effect is that some subset of the data used to calculate X at times t and $t+\tau$ are the same. This is frequently the case when a variable is calculated from historical data covering more than τ periods. For example, if X_t is calculated using k periods of historical data up to and including period t, where $k > \tau$, then X_t and $X_{t+\tau}$ are calculated using some of the same data and are therefore likely to be correlated in the cross section as a result. For this reason, when X is calculated using k periods of historical data, persistence analysis is only effective when $\tau \geq k$.

In addition to examining whether a given characteristic of the entities in the sample is cross-sectionally persistent, persistence analysis can also be helpful in determining the optimal measurement period that should be used to calculate a given variable. Many of the variables used throughout the empirical asset pricing literature are calculated based on historical data. When calculating these variables, researchers are faced with the decision of how long a calculation period to use. Increasing the length of the calculation period means that more data are used in the calculation of the variable, which can increase the accuracy of the measurement. However, using longer calculation periods also means that when calculating the variable for time period t, data from many periods prior to t are used, and this data may no longer be reflective of the given characteristic of the entity at time t. For this reason, extending the calculation period too long may result in decreasing accuracy of measurement. How to optimally make this trade-off depends on the persistence of the characteristic being measured. While certainly none of the variables studied in asset pricing research are perfectly persistent, different variables exhibit different degrees of persistence.

To help determine the optimal calculation period for a variable calculated from historical data, we can examine the patterns in the persistence of the variable measured using different calculation periods. The main concept behind this application of persistence analysis is that the cross-sectional persistence of even the most persistent characteristic is likely to decay over time. Therefore, let us assume we have two variables X_1 and X_2 that measure the same characteristic using the same formula but applied to different calculation periods of length τ_1 and τ_2, respectively, and without loss of generality, let $\tau_1 > \tau_2$. Let us also make the assumption that X^1 and X^2 are equally accurate measures of the given characteristic.

If X^1 and X^2 are equally accurate measures of the characteristic, then based on the assumed decay in persistence as the value of τ increases, we would expect the persistence of X^1 measured at a lag of $\tau = \tau_1$ to be greater than the persistence of X^2 measured at lag $\tau = \tau_2$. Notice here that the lag at which the comparison of the persistence is done is such that neither analysis has the overlapping data issue discussed earlier. If the persistence of X_2 at lag $\tau = \tau_2$ is actually greater than that of X_1 at lag $\tau = \tau_1$, this is a contradiction of what would be expected if X_1 and X_2 were equally accurate measures. This therefore indicates that using τ_2 periods of data to calculate X provides a more accurate measure of the underlying characteristic than using τ_1 periods, as the additional amount of data used in the calculation apparently overcomes the decay in the persistence at longer lags τ.

If the persistence of X^2 at lag $\tau = \tau_2$ is less than that of X^1 at lag $\tau = \tau_1$, the results are a bit more challenging to interpret, but it can generally be taken to mean that the decay in the persistence over a period of $\tau_2 - \tau_1$ periods is substantial enough to overcome any additional benefit of using τ_2 periods of data, compared to τ_1, to calculate X. If this is the case, it may also be an indication that using a full τ_2 periods of data is too long a measurement period because the given characteristic of the firm does in fact change substantially over periods of length τ_2.

There is a practical consideration that may have an effect on using persistence to determine the optimal measurement period for the given variable X. This consideration is that the sample changes over time. The calculation of the value of $\rho_{t,t+\tau}(X)$ is

done using only those entities that are in the sample at both times t and $t + \tau$. In most cases, the set of entities in the sample at both time t and time $t + \tau_1$ is likely to be a superset of the set of entities in the sample at both time t and time $t + \tau_2$ ($\tau_2 > \tau_1$). Furthermore, in many cases, the set of entities that remain in the sample until time $t + \tau_2$ is likely to be more "well-behaved" than those that do not remain in the sample, where well-behaved can be taken to mean that the calculation of the variable X is a more accurate measure of the characteristic being examined for such entities than for entities that are not well-behaved. If these not well-behaved entities are more likely to enter and then drop out of the sample over a small number of periods, it is possible that using persistence analysis to examine the quality of a variable as described in this section may be misleading. That being said, for the analyses performed in this text this is unlikely to be a substantial issue, as the number of entities (stocks in this case) in each cross section is quite large relative to the number of stocks that drop out of the sample each period.

4.3 PRESENTING PERSISTENCE

Throughout this book, we will present the results of persistence analyses by displaying the values of $\rho_\tau(X)$. Each column in the tables that present the persistence analysis results will correspond to a given variable, indicated in the first row of the column. Each row will correspond to a given value of τ.

The results of persistence analyses for each of β, *Size*, and *BM* using lags of one, two, three, four, and five years are presented in Table 4.3. The results indicate that all three variables are highly persistent. The persistence of β measured at lag of one year ($\tau = 1$) is 0.61 and that of *Size* is 0.96, and for *BM* the persistence at lag of one year is 0.74. The results for each of these variables indicate fairly strong persistence at lags of up to five years. *Size* is very highly persistent, as the average cross-sectional

TABLE 4.3 Persistence of β, *Size*, and *BM*
This table presents the results of persistence analyses of β, *Size*, and *BM*. For each year t, the cross-sectional correlation between the given variable measured at time t and the same variable measured at time $t + \tau$ is calculated. The table presents the time-series averages of the annual cross-sectional correlations. The column labeled τ indicates the lag at which the persistence is measured.

τ	β	*Size*	*BM*
1	0.61	0.96	0.74
2	0.53	0.92	0.59
3	0.48	0.90	0.50
4	0.46	0.89	0.46
5	0.42	0.88	0.43

correlation between *Size* measured five years apart is 0.88, only slightly lower than when the persistence is measured at a lag of one year. The decay in the persistence of β and *BM* is substantially more pronounced, but even after five years, β and *BM* continue to exhibit substantial persistence.

4.4 SUMMARY

In this chapter, we have presented a methodology for examining the persistence of a given variable. The methodology has two primary applications. If we assume that the variable accurately measures the characteristic that it is intended to capture, then persistence analysis can be used to examine how persistent the given characteristic is in the cross section of the entities in the sample. If we assume the characteristic that the variable is intended to measure is in fact persistent, then we can use persistence analysis to examine the accuracy with which the variable captures the given characteristic and the optimal measurement period to use when calculating the variable. Of course, no characteristic is perfectly persistent and no variable perfectly captures the characteristic it is designed to measure. Despite these caveats, persistence analysis is a useful tool that we will employ throughout this text.

REFERENCES

Lintner, J. 1965. Security prices, risk, and maximal gains from diversification. Journal of Finance, 20(4), 687–615.

Mossin, J. 1966. Equilibrium in a capital asset market. Econometrica, 34(4), 768–783.

Sharpe, W. F. 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. Journal of Finance, 19(3), 425–442.