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.