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.
© 2016 John Wiley & Sons, Inc. Published 2016 by John Wiley & Sons, Inc.

Copyright © 2016. John Wiley & Sons, Incorporated. All rights reserved.

Bali, Turan G., et al. Empirical Asset Pricing : The Cross Section of Stock Returns, John Wiley & Sons, Incorporated, 2016.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=4451896.
Created from oxford on 2021-09-26 19:28:48.

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

Copyright © 2016. John Wiley & Sons, Incorporated. All rights reserved.

Bali, Turan G., et al. Empirical Asset Pricing : The Cross Section of Stock Returns, John Wiley & Sons, Incorporated, 2016.

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

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

Copyright © 2016. John Wiley & Sons, Incorporated. All rights reserved.

Bali, Turan G., et al. Empirical Asset Pricing : The Cross Section of Stock Returns, John Wiley & Sons, Incorporated, 2016.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=4451896.
Created from oxford on 2021-09-26 19:28:48.

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

Copyright © 2016. John Wiley & Sons, Incorporated. All rights reserved.

Bali, Turan G., et al. Empirical Asset Pricing : The Cross Section of Stock Returns, John Wiley & Sons, Incorporated, 2016.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=4451896.
Created from oxford on 2021-09-26 19:28:48.

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

Copyright © 2016. John Wiley & Sons, Incorporated. All rights reserved.

Bali, Turan G., et al. Empirical Asset Pricing : The Cross Section of Stock Returns, John Wiley & Sons, Incorporated, 2016.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=4451896.
Created from oxford on 2021-09-26 19:28:48.

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

Copyright © 2016. John Wiley & Sons, Incorporated. All rights reserved.

Bali, Turan G., et al. Empirical Asset Pricing : The Cross Section of Stock Returns, John Wiley & Sons, Incorporated, 2016.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=4451896.
Created from oxford on 2021-09-26 19:28:48.

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	

Copyright © 2016. John Wiley & Sons, Incorporated. All rights reserved.

Bali, Turan G., et al. Empirical Asset Pricing : The Cross Section of Stock Returns, John Wiley & Sons, Incorporated, 2016.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=4451896.
Created from oxford on 2021-09-26 19:28:48.

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.

Copyright © 2016. John Wiley & Sons, Incorporated. All rights reserved.

Bali, Turan G., et al. Empirical Asset Pricing : The Cross Section of Stock Returns, John Wiley & Sons, Incorporated, 2016.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=4451896.
Created from oxford on 2021-09-26 19:28:48.