READING

7

Correlation and Regression

by Richard A. DeFusco, PhD, CFA, Dennis W. McLeavey, CFA, Jerald E. Pinto, PhD, CFA, and David E. Runkle, PhD, CFA

Richard A. DeFusco, PhD, CFA, is at the University of Nebraska-Lincoln (USA). Dennis W. McLeavey, CFA, is at the University of Rhode Island (USA). Jerald E. Pinto, PhD, CFA, is at CFA Institute (USA). David E. Runkle, PhD, CFA, is at Trilogy Global Advisors (USA).

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
☐	**a.** calculate and interpret a sample covariance and a sample correlation coefficient and interpret a scatter plot;
☐	**b.** describe limitations to correlation analysis;
☐	**c.** formulate a test of the hypothesis that the population correlation coefficient equals zero and determine whether the hypothesis is rejected at a given level of significance;
☐	**d.** distinguish between the dependent and independent variables in a linear regression;
☐	**e.** explain the assumptions underlying linear regression and interpret regression coefficients;
☐	**f.** calculate and interpret the standard error of estimate, the coefficient of determination, and a confidence interval for a regression coefficient;
☐	**g.** formulate a null and alternative hypothesis about a population value of a regression coefficient and determine the appropriate test statistic and whether the null hypothesis is rejected at a given level of significance;
☐	**h.** calculate the predicted value for the dependent variable, given an estimated regression model and a value for the independent variable;
☐	**i.** calculate and interpret a confidence interval for the predicted value of the dependent variable;
☐	**j.** describe the use of analysis of variance (ANOVA) in regression analysis, interpret ANOVA results, and calculate and interpret the *F*-statistic;
☐	**k.** describe limitations of regression analysis.

1 INTRODUCTION

As a financial analyst, you will often need to examine the relationship between two or more financial variables. For example, you might want to know whether returns to different stock market indexes are related and, if so, in what way. Or you might hypothesize that the spread between a company's return on invested capital and its cost of capital helps to explain the company's value in the marketplace. Correlation and regression analysis are tools for examining these issues.

This reading[1] is organized as follows. In Section 2, we present correlation analysis, a basic tool in measuring how two variables vary in relation to each other. Topics covered include the calculation, interpretation, uses, limitations, and statistical testing of correlations. Section 3 introduces basic concepts in regression analysis, a powerful technique for examining the ability of one or more variables (independent variables) to explain or predict another variable (the dependent variable).

2 CORRELATION ANALYSIS

We have many ways to examine how two sets of data are related. Two of the most useful methods are scatter plots and correlation analysis. We examine scatter plots first.

2.1 Scatter Plots

A **scatter plot** is a graph that shows the relationship between the observations for two data series in two dimensions. Suppose, for example, that we want to graph the relationship between long-term money growth and long-term inflation in six industrialized countries to see how strongly the two variables are related. Table 1 shows the average annual growth rate in the money supply and the average annual inflation rate from 1980 to 2012 for the six countries.

Table 1 Annual Money Supply Growth Rate and Inflation Rate by Country, 1980–2012

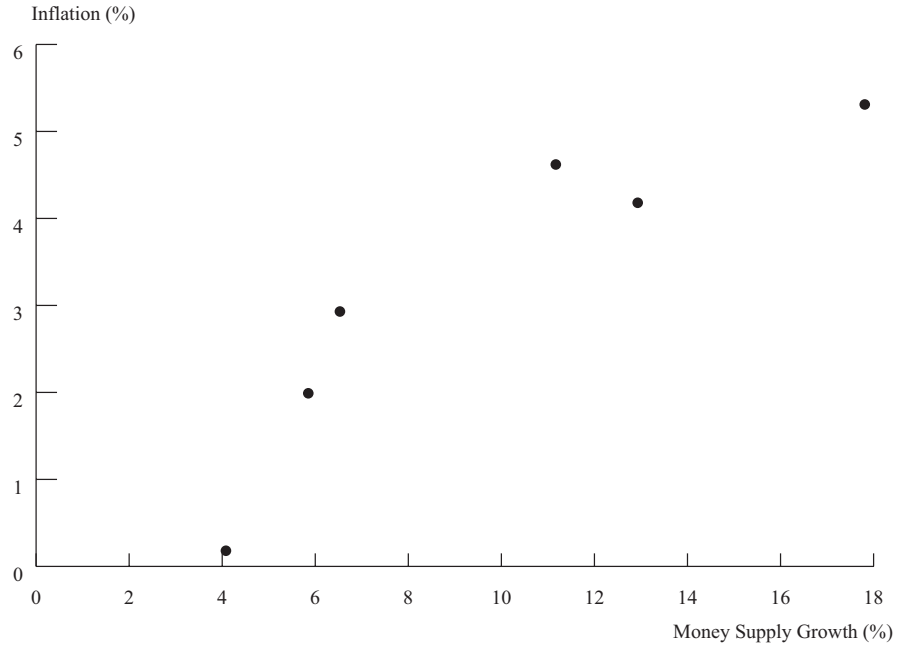
Country	Money Supply Growth Rate (%)	Inflation Rate (%)
Australia	11.17	4.62
Japan	4.08	0.18
South Korea	17.81	5.31
Switzerland	5.85	1.99
United Kingdom	12.93	4.18
United States	6.53	2.93
Average	9.73	3.20

Source: International Monetary Fund.

1 Examples in this reading were updated in 2014 by Professor Sanjiv Sabherwal of the University of Texas, Arlington.

To translate the data in Table 1 into a scatter plot, we use the data for each country to mark a point on a graph. For each point, the *x*-axis coordinate is the country's annual average money supply growth from 1980–2012, and the *y*-axis coordinate is the country's annual average inflation rate from 1980–2012. Figure 1 shows a scatter plot of the data in Table 1.

Figure 1 Scatter Plot of Annual Money Supply Growth Rate and Inflation Rate by Country, 1980–2012



Source: International Monetary Fund.

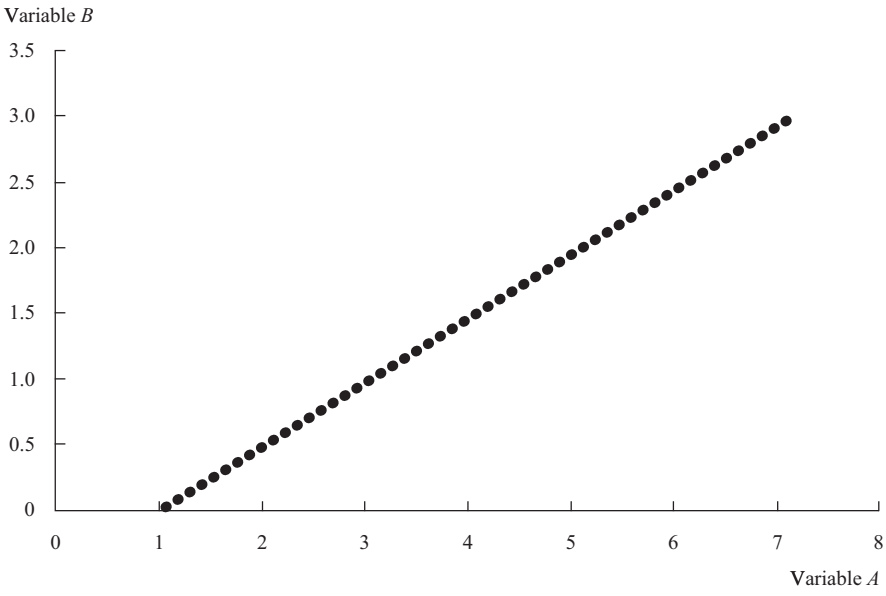
Note that each observation in the scatter plot is represented as a point, and the points are not connected. The scatter plot does not show which observation comes from which country; it shows only the actual observations of both data series plotted as pairs. For example, the rightmost point shows the data for South Korea. The data plotted in Figure 1 show a fairly strong linear relationship with a positive slope. Next we examine how to quantify this linear relationship.

2.2 Correlation Analysis

In contrast to a scatter plot, which graphically depicts the relationship between two data series, **correlation analysis** expresses this same relationship using a single number. The correlation coefficient is a measure of how closely related two data series are. In particular, the correlation coefficient measures the direction and extent of **linear association** between two variables. A correlation coefficient can have a maximum value of 1 and a minimum value of –1. A correlation coefficient greater than 0 indicates a positive linear association between the two variables: When one variable increases (or decreases), the other also tends to increase (or decrease). A correlation coefficient less than 0 indicates a negative linear association between the two variables: When

one increases (or decreases), the other tends to decrease (or increase). A correlation coefficient of 0 indicates no linear relation between the two variables.[2] Figure 2 shows the scatter plot of two variables with a correlation of 1.

Figure 2 Variables with a Correlation of 1



Note that all the points on the scatter plot in Figure 2 lie on a straight line with a positive slope. Whenever variable *A* increases by one unit, variable *B* increases by half a unit. Because all of the points in the graph lie on a straight line, an increase of one unit in *A* is associated with exactly the same half-unit increase in *B*, regardless of the level of *A*. Even if the slope of the line in the figure were different (but positive), the correlation between the two variables would be 1 as long as all the points lie on that straight line.

Figure 3 shows a scatter plot for two variables with a correlation coefficient of –1. Once again, the plotted observations fall on a straight line. In this graph, however, the line has a negative slope. As *A* increases by one unit, *B* decreases by half a unit, regardless of the initial value of *A*.

2 Later, we show that variables with a correlation of 0 can have a strong nonlinear relation.

Figure 3 Variables with a Correlation of –1

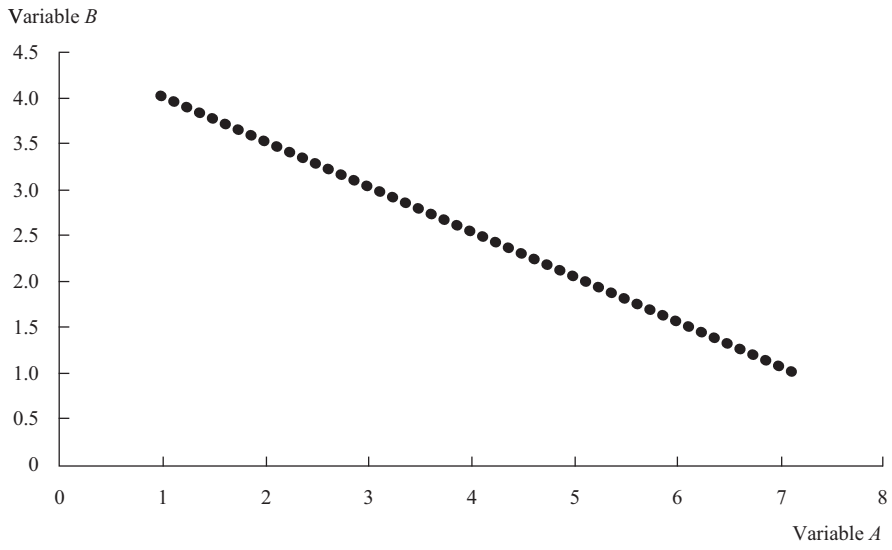
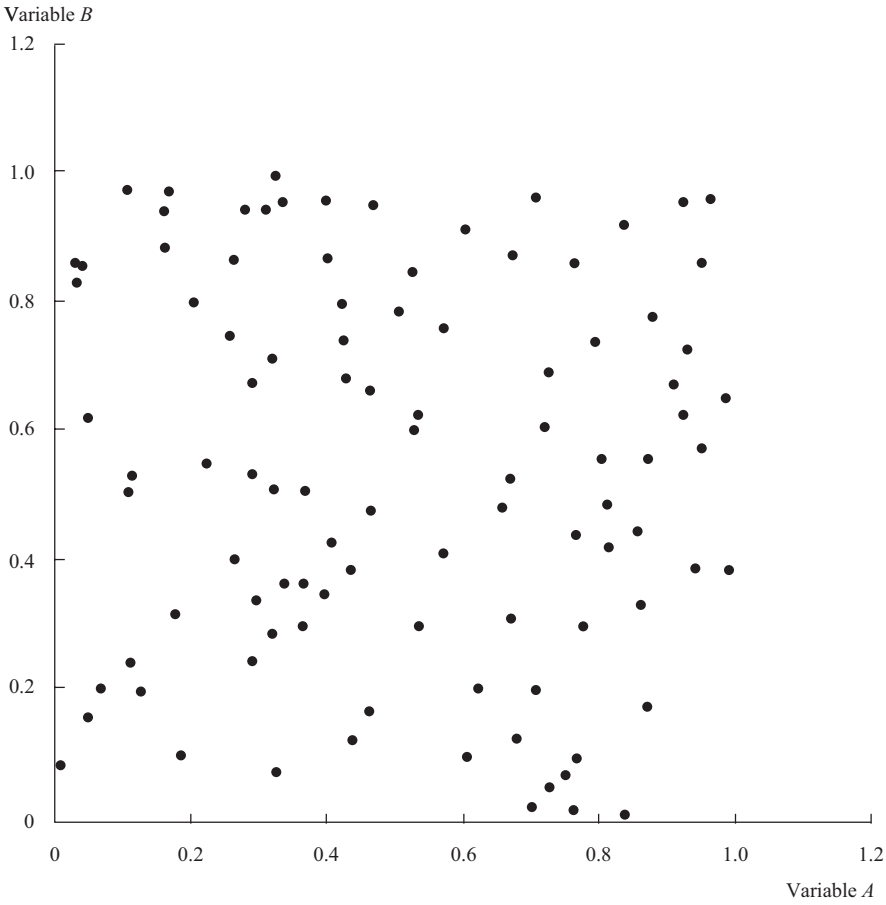


Figure 4 shows a scatter plot of two variables with a correlation of 0; they have no linear relation. This graph shows that the value of A tells us absolutely nothing about the value of B.

Figure 4 Variables with a Correlation of 0



2.3 Calculating and Interpreting the Correlation Coefficient

To define and calculate the correlation coefficient, we need another measure of linear association: covariance. At Level I of the CFA Program, covariance was defined as the expected value of the product of the deviations of two random variables from their respective population means. That was the definition of population covariance, which we would also use in a forward-looking sense. To study historical or sample correlations, we need to use sample covariance. The sample covariance of X and Y, for a sample of size n, is

$$\text{Cov}(X,Y) = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)\Big/(n-1) \quad \textbf{(1)}$$

The sample covariance is the average value of the product of the deviations of observations on two random variables from their sample means.[3] If the random variables are returns, the unit of covariance would be returns squared.

The sample correlation coefficient is much easier to explain than the sample covariance. To understand the sample correlation coefficient, we need the expression for the sample standard deviation of a random variable X. We need to calculate the sample

3 The use of $n - 1$ in the denominator is a technical point; it ensures that the sample covariance is an unbiased estimate of population covariance.

variance of X to obtain its sample standard deviation. The variance of a random variable is simply the covariance of the random variable with itself. The expression for the sample variance of X, s_X^2, is

$$s_X^2 = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \Big/ (n-1)$$

The sample standard deviation is the positive square root of the sample variance:

$$s_X = \sqrt{s_X^2}$$

Both the sample variance and the sample standard deviation are measures of the dispersion of observations about the sample mean. Standard deviation uses the same units as the random variable; variance is measured in the units squared.

The formula for computing the sample correlation coefficient is

$$r = \frac{\mathrm{Cov}(X,Y)}{s_X s_Y} \qquad (2)$$

The correlation coefficient is the covariance of two variables (X and Y) divided by the product of their sample standard deviations (s_X and s_Y). Like covariance, the correlation coefficient is a measure of linear association. The correlation coefficient, however, has the advantage of being a simple number, with no unit of measurement attached. It has no units because it results from dividing the covariance by the product of the standard deviations. Because we will be using sample variance, standard deviation, and covariance in this reading, we will repeat the calculations for these statistics.

Table 2 shows how to compute the various components of the correlation equation (Equation 2) from the data in Table 1.[4] The individual observations on countries' annual average money supply growth from 1980–2012 are denoted X_i, and individual observations on countries' annual average inflation rate from 1980–2012 are denoted Y_i. The remaining columns show the calculations for the inputs to correlation: the sample covariance and the sample standard deviations.

Table 2 Sample Covariance and Sample Standard Deviations: Annual Money Supply Growth Rate and Inflation Rate by Country, 1980–2012

Country	Money Supply Growth Rate X_i	Inflation Rate Y_i	Cross-Product $(X_i - \bar{X})(Y_i - \bar{Y})$	Squared Deviations $(X_i - \bar{X})^2$	Squared Deviations $(Y_i - \bar{Y})^2$
Australia	0.1117	0.0462	0.000204	0.000208	0.000201
Japan	0.0408	0.0018	0.001707	0.003190	0.000913
South Korea	0.1781	0.0531	0.001704	0.006531	0.000445
Switzerland	0.0585	0.0199	0.000470	0.001504	0.000147
United Kingdom	0.1293	0.0418	0.000313	0.001025	0.000096
United States	0.0653	0.0293	0.000087	0.001023	0.000007

(continued)

4 We have not used full precision in the table's calculations. We used the average value of the money supply growth rate of 0.5839/6 = 0.0973, rounded to four decimal places, in the cross-product and squared deviation calculations, and similarly, we used the mean inflation rate as rounded to 0.0320 in those calculations. We computed standard deviation as the square root of variance rounded to six decimal places, as shown in the table. Had we used full precision in all calculations, some of the table's entries would be slightly different but would not materially affect our conclusions.

Table 2 (Continued)

Country	Money Supply Growth Rate X_i	Inflation Rate Y_i	Cross-Product $(X_i - \bar{X})(Y_i - \bar{Y})$	Squared Deviations $(X_i - \bar{X})^2$	Squared Deviations $(Y_i - \bar{Y})^2$
Sum	0.5837	0.1921	0.004485	0.013482	0.001809
Average	0.0973	0.0320			
Covariance			0.000897		
Variance				0.002696	0.000362
Standard deviation				0.051926	0.019019

Notes:

1 Divide the cross-product sum by $n - 1$ (with $n = 6$) to obtain the covariance of X and Y.

2 Divide the squared deviations sums by $n - 1$ (with $n = 6$) to obtain the variances of X and Y.

Source: International Monetary Fund.

Using the data shown in Table 2, we can compute the sample correlation coefficient for these two variables as follows:

$$r = \frac{\text{Cov}(X,Y)}{s_x s_y} = \frac{0.000897}{(0.051926)(0.019019)} = 0.9083$$

The correlation coefficient of approximately 0.91 indicates a strong linear association between long-term money supply growth and long-term inflation for the countries in the sample. The correlation coefficient captures this strong association numerically, whereas the scatter plot in Figure 1 shows the information graphically.

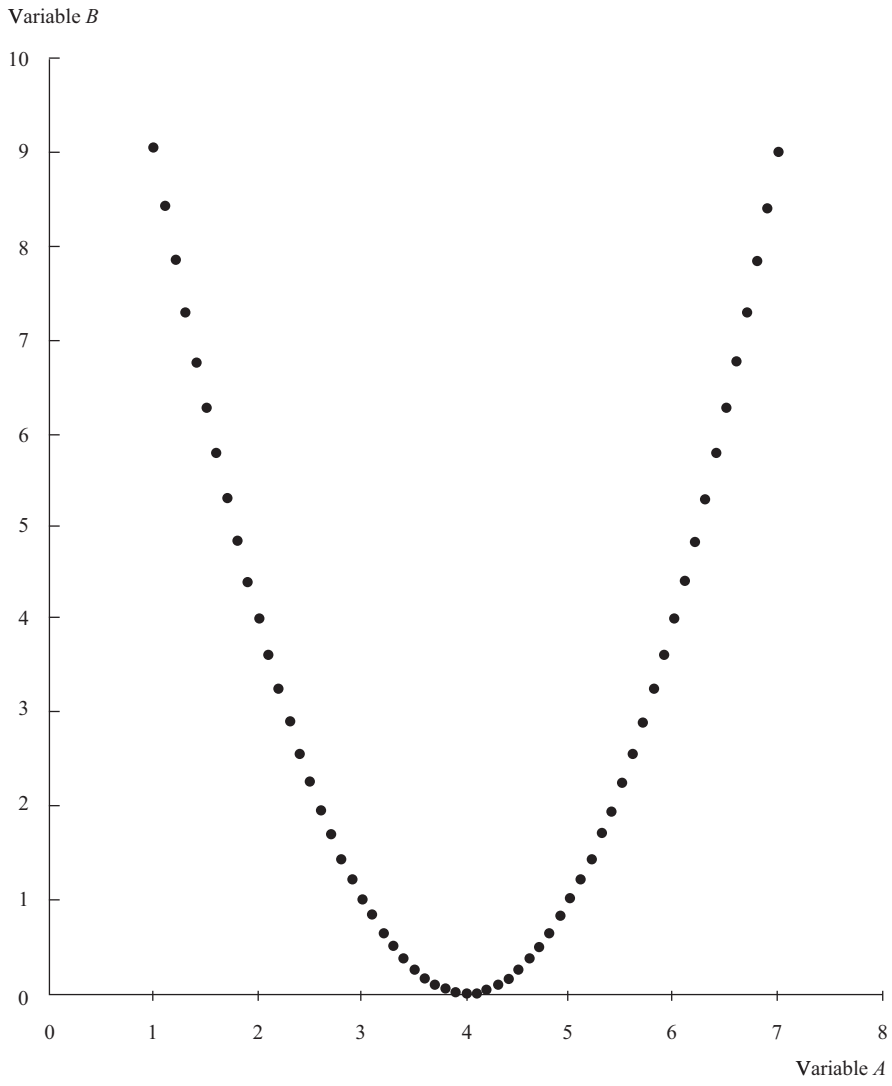
What assumptions are necessary to compute the correlation coefficient? Correlation coefficients can be computed validly if the means and variances of X and Y, as well as the covariance of X and Y, are finite and constant. Later, we will show that when these assumptions are not true, correlations between two different variables can depend greatly on the sample that is used.

2.4 Limitations of Correlation Analysis

Correlation measures the linear association between two variables, but it may not always be reliable. Two variables can have a strong **nonlinear relation** and still have a very low correlation. For example, the relation $B = (A - 4)^2$ is a nonlinear relation contrasted to the linear relation $B = 2A - 4$. The nonlinear relation between variables A and B is shown in Figure 5. Below a level of 4 for A, variable B decreases with increasing values of A. When A is 4 or greater, however, B increases whenever A increases. Even though these two variables are perfectly associated, the correlation between them is 0.[5]

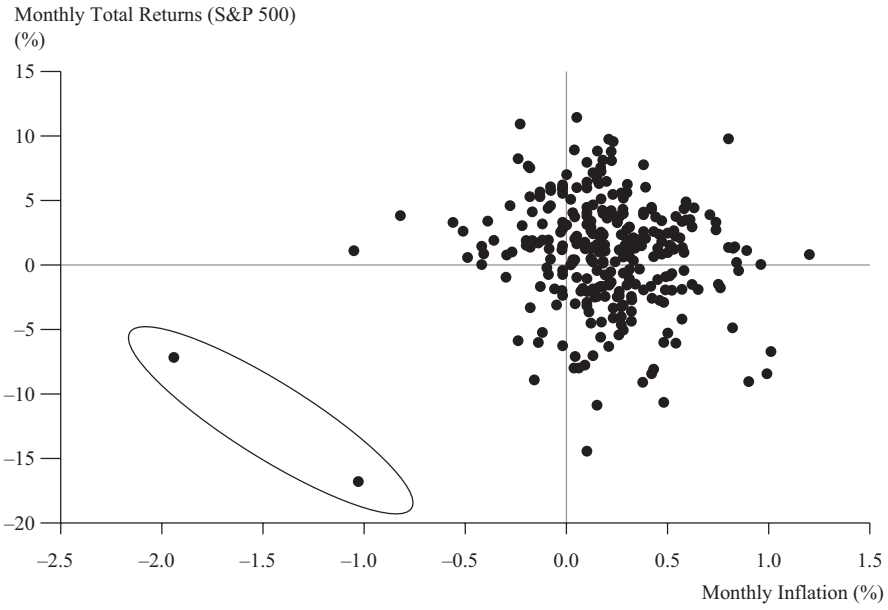
5 The perfect association is the quadratic relationship $B = (A - 4)^2$.

Figure 5 Variables with a Strong Nonlinear Association



Correlation also may be an unreliable measure when outliers are present in one or both of the series. Outliers are small numbers of observations at either extreme (small or large) of a sample. Figure 6 shows a scatter plot of the monthly returns to the Standard & Poor's 500 Index and the monthly inflation rate in the United States from January 1990 through December 2013.

Figure 6 US Inflation and Stock Returns: 1990–2013



Sources: Bureau of Labor Statistics and S&P Dow Jones Indices.

In the scatter plot in Figure 6, most of the data lie clustered together with little discernible relation between the two variables. Two cases, however (the two circled observations), stand out from the rest. In one of those cases, inflation was extremely low at almost –2 percent, and in the other case, stock returns were strongly negative at almost –17 percent. These observations are outliers. If we compute the correlation coefficient for the entire data sample, that correlation is –0.0350. If we eliminate the two outliers, however, the correlation is –0.1489.

The correlation in Figure 6 is quite sensitive to excluding only two observations. Does it make sense to exclude those observations? Are they noise or news? When the outliers are excluded, there seems to be a moderately negative correlation between inflation and stock returns. One possible partial explanation of this negative correlation is that whenever inflation was very high during a month, market participants became concerned that the Federal Reserve would raise interest rates, which would cause the value of stocks to decline. This story offers one plausible explanation for how investors reacted to large inflation announcements. When the two outliers are included, there is a noticeable decrease in the magnitude of the negative correlation. A closer examination of the monthly data used in the scatter plot reveals that the two outliers correspond to the months of October and November 2008 when bad news regarding the US economy and job market caused the stock market to decline sharply. During these two months, although the inflation was not high (in fact, inflation was negative in both months), stocks declined substantially. Therefore, inclusion of those two outliers reduces the magnitude of the negative correlation between inflation and stock returns. One could argue that while the data without the outliers provide a useful insight into the general relationship between inflation and stock returns, the outliers may provide information about the relationship during a period of market distress. Therefore, in this case, it would be reasonable to report the values of the correlation including and excluding the outliers.

As a general rule, we must determine whether a computed sample correlation changes greatly by removing a few outliers. But we must also use judgment to determine whether those outliers contain information about the two variables' relationship (and should thus be included in the correlation analysis) or contain no information (and should thus be excluded).

Keep in mind that correlation does not imply causation. Even if two variables are highly correlated, one does not necessarily cause the other in the sense that certain values of one variable bring about the occurrence of certain values of the other. Furthermore, correlations can be spurious in the sense of misleadingly pointing towards associations between variables.

The term **spurious correlation** has been used to refer to 1) correlation between two variables that reflects chance relationships in a particular data set, 2) correlation induced by a calculation that mixes each of two variables with a third, and 3) correlation between two variables arising not from a direct relation between them but from their relation to a third variable. As an example of the second kind of spurious correlation, two variables that are uncorrelated may be correlated if divided by a third variable. As an example of the third kind of spurious correlation, height may be positively correlated with the extent of a person's vocabulary, but the underlying relationships are between age and height and between age and vocabulary. Investment professionals must be cautious in basing investment strategies on high correlations. Spurious correlation may suggest investment strategies that appear profitable but actually would not be so, if implemented.

2.5 Uses of Correlation Analysis

In this section, we give examples of correlation analysis for investment. Because investors' expectations about inflation are important in determining asset prices, inflation forecast accuracy will serve as our first example.

EXAMPLE 1

Evaluating Economic Forecasts (1)

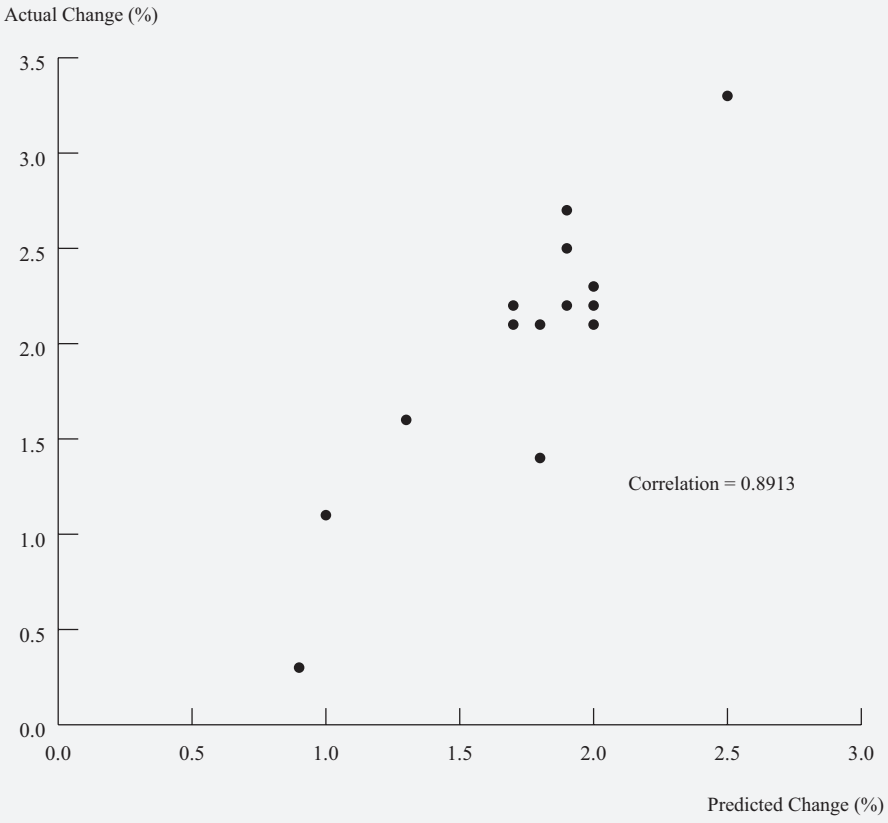
Investors closely watch economists' forecasts of inflation, but do these forecasts contain useful information? In the euro area, the Survey of Professional Forecasters (SPF) gathers professional forecasters' predictions about many economic variables.[6] Since 1999, SPF has gathered predictions on the euro area inflation rate using the change in the Harmonised Index of Consumer Prices (HICP) for the prices of consumer goods and services acquired by households to measure inflation. If these forecasts of inflation could perfectly predict actual inflation, the correlation between forecasts and inflation would be 1.

Figure 7 shows a scatter plot of the mean forecast made in the first quarter of a year for the percentage change in HICP during that year and the actual percentage change in HICP, from 1999 through 2013.[7] In this scatter plot, the forecast for each year is plotted on the x-axis and the actual change in the HICP is plotted on the y-axis.

6 The euro area survey is conducted by the European Central Bank (ECB). A survey of professional forecasters is also conducted by the Federal Reserve Bank of Philadelphia for the United States.

7 In this scatter plot, the actual inflation rate is from the Statistical Data Warehouse of the European Central Bank.

Figure 7 Actual Change in Euro Area HICP versus Predicted Change



Source: European Central Bank.

Discuss whether professional forecasters' predictions of the euro area inflation might be useful in investment decision-making.

Solution:

As Figure 7 shows, a fairly strong linear association exists between the forecast and the actual inflation rate, suggesting that professional forecasts of inflation might be useful in investment decision-making. In fact, the correlation between the two series is 0.8913. Although there is no causal relation here, there is a direct relation because forecasters assimilate information to forecast inflation.

One important issue in evaluating a portfolio manager's performance is determining an appropriate benchmark for the manager. Since the early 1990s, style analysis has been an important component of benchmark selection.[8]

8 See, for example, Sharpe (1992), Buetow, Johnson, and Runkle (2000), and Chan, Dimmock, and Lakonishok (2009).

EXAMPLE 2

Style Analysis Correlations

Portfolio managers using small-cap stocks in investment portfolios may favor a growth style, a value style, or neither.

In the United States, the Russell 2000 Growth Index and the Russell 2000 Value Index are often used as benchmarks for small-cap growth and small-cap value managers, respectively. Correlation analysis shows, however, that the returns to these two indexes are very closely associated with each other. For the 15 years ending in 2013 (January 1999 to December 2013), the correlation between the monthly returns to the Russell 2000 Growth Index and the Russell 2000 Value Index was 0.8249.

What conclusions can be drawn based on this result concerning the mean returns to small-cap growth and small-cap value investment styles? Explain your answer.

Solution:

The returns to the two indexes are highly positively correlated. But correlation does not provide information on variables' mean returns, only on how their returns covary. Here, for example, even a correlation of +1 to the returns to the two styles would not imply that the mean returns to the two styles are the same. Thus, the information given is not sufficient to reach a conclusion on the relative mean returns to the small-cap growth and small-cap investment styles.

The previous examples in this reading have examined the correlation between two variables. Often, however, investment managers need to understand the correlations among many asset returns. For example, investors who have any exposure to movements in exchange rates must understand the correlations of the returns to different foreign currencies and other assets in order to determine their optimal portfolios and hedging strategies.[9] In the following example, we see how a correlation matrix shows correlation between pairs of variables when we have more than two variables. We also see one of the main challenges to investment managers: Investment return correlations can change substantially over time.

EXAMPLE 3

Exchange Rate Return Correlations

The exchange rate return measures the periodic domestic currency return to holding foreign currency. Consider a British investor with British pounds (GBP) as her domestic currency. Suppose a change in inflation rates in Canada and the United Kingdom results in the pound price of a Canadian dollar changing from £0.50 to £0.45. If this change occurred in one month, the return in that month to holding Canadian dollars would be (0.45 – 0.50)/0.50 = –10 percent, in terms of pounds.

Table 3 shows a correlation matrix of monthly returns in British pounds to holding Canadian, Japanese, Swedish, or US currencies during two seven-year periods of 2000–2006 and 2007–2013. To interpret a correlation matrix, we first examine the top panel of this table.

9 See, for example, Campbell, Serfaty-de Medeiros, and Viceira (2010).

The first column of numbers of that panel shows the correlations between GBP returns to holding the Canadian dollar and GBP returns to holding Canadian, Japanese, Swedish, and US currencies during 2000–2006. Of course, any variable is perfectly correlated with itself, and so the correlation between GBP returns to holding the Canadian dollar and GBP returns to holding the Canadian dollar is 1. The second row of this column shows that the correlation between GBP returns to holding the Canadian dollar and GBP returns to holding the Japanese yen was 0.4552 during 2000–2006. The remaining correlations in the panel show how the GBP returns to other combinations of currency holdings were correlated during this period.

Table 3 Correlations of Monthly British Pound Returns to Selected Foreign Currency Returns

2000–2006	Canada	Japan	Sweden	United States
Canada	1.0000			
Japan	0.4552	1.0000		
Sweden	0.2686	0.1832	1.0000	
United States	0.6917	0.4360	0.0074	1.0000

2007–2013	Canada	Japan	Sweden	United States
Canada	1.0000			
Japan	0.3091	1.0000		
Sweden	0.5278	0.1742	1.0000	
United States	0.5263	0.7230	0.1862	1.0000

Source: www.oanda.com/currency/historical-rates/.

1 Explain why Table 3 omits many of the correlations.

Solution to 1:

The formula for correlation coefficient in the earlier equation (equation 2) shows that correlations are always symmetrical: The correlation between X and Y is always the same as the correlation between Y and X. Accordingly, duplicative coefficients are excluded in Table 3. For example, Column 2 of the panels omits the correlation between GBP returns to holding yen and GBP returns to holding Canadian dollars. This correlation is omitted because it is identical to the correlation between GBP returns to holding Canadian dollars and GBP returns to holding yen shown in Row 2 of Column 1. Similarly, other omitted correlations would also have been duplicative.

2 Compare the two panels of Table 3 and discuss whether the changes in correlations from 2000–2006 to 2007–2013 show a pattern.

Solution to 2:

A comparison of the two panels of Table 3 shows that that many of the currency return correlations changed dramatically between the periods of 2000–2006 and 2007–2013, but there is no pattern in these changes. During 2000–2006, for example, the correlation between the return to holding Canadian dollars and the return to holding Japanese yen (0.4552) was about the same as the correlation

between the return to holding yen and the return to holding US dollars (0.4360). During 2007–2013, however, the correlation between Canadian dollar returns and yen returns dropped substantially (to 0.3091), but the correlation between yen and US dollar returns increased substantially (to 0.7230). Some other correlations also increased markedly. For example, the correlation between Canadian dollar returns and Swedish krona returns almost doubled from 0.2686 to 0.5278 and the correlation between krona and US dollar returns increased from 0.0074 to 0.1862. In contrast, the correlation between Canadian and US dollars decreased from 0.6917 to 0.5263 and the correlation between yen and krona returns hardly changed (0.1832 to 0.1742).

Optimal asset allocation depends on expectations of future correlations. With less than perfect positive correlation between two assets' returns, there are potential risk-reduction benefits to holding both assets. Expectations of future correlation may be based on historical sample correlations, but the variability in historical sample correlations poses challenges. We discuss these issues in detail in the reading on portfolio concepts.

In the next example, we extend the discussion of the correlations of stock market indexes begun in Example 2 to indexes representing large-cap, small-cap, and broad-market returns. This type of analysis has serious diversification and asset allocation consequences because the strength of the correlations among the assets tells us how successfully the assets can be combined to diversify risk.

EXAMPLE 4

Correlations among Stock Return Series

Table 4 shows the correlation matrix of monthly returns to three UK stock indexes during the period January 1990 to December 2009 and in two subperiods (the 1990s and 2000s). The large-cap style is represented by the return to the FTSE 100 Index, the small-cap style is represented by the return to the FTSE Small Cap Excluding Investment Companies Index, and the broad-market returns are represented by the return to the FTSE All-Share Index.

Table 4 Correlations of Monthly Returns to Various UK Stock Indexes

1990–2009	FTSE 100	FTSE Small Cap	FTSE All-Share
FTSE 100	1.0000		
FTSE Small Cap	0.6914	1.0000	
FTSE All-Share	0.9906	0.7694	1.0000

1990–1999	FTSE 100	FTSE Small Cap	FTSE All-Share
FTSE 100	1.0000		
FTSE Small Cap	0.6553	1.0000	
FTSE All-Share	0.9873	0.7553	1.0000

(continued)

Table 4 (Continued)

2000–2009	FTSE 100	FTSE Small Cap	FTSE All-Share
FTSE 100	1.0000		
FTSE Small Cap	0.7245	1.0000	
FTSE All-Share	0.9937	0.7869	1.0000

Source: CompuSmart Global.

Discuss whether the correlation coefficients for the entire sample are consistent with your expectations.

Solution:

The first column of numbers in the top panel of Table 4 shows nearly perfect positive correlation between returns to the FTSE 100 and returns to the FTSE All-Share: The correlation between the two return series is 0.9906. This result should not be surprising, because both the FTSE 100 and the FTSE All-Share are value-weighted indexes, and large-cap stock returns receive most of the weight in both indexes. In fact, the companies that make up the FTSE 100 have more than 80 percent of the total market value of all companies included in the FTSE All-Share.

Small-cap stocks also have a reasonably high correlation with large stocks. In the total sample, the correlation between the FTSE 100 returns and the FTSE Small-Cap returns is 0.6914. The correlation between FTSE Small-Cap returns and returns to the FTSE All-Share is slightly higher (0.7694). This result is also not too surprising because the FTSE All-Share contains small-cap stocks and the FTSE 100 does not.

The second and third panels of Table 4 show that correlations among the various stock market return series show some variation from decade to decade. For example, the correlation between returns to the FTSE 100 and FTSE small-cap stocks increased from 0.6553 in the 1990s to 0.7245 in the 2000s.[10]

For asset allocation purposes, correlations among asset classes are studied carefully with a view toward maintaining appropriate diversification based on forecasted correlations.

EXAMPLE 5

Correlations of Debt and Equity Returns

Table 5 shows the correlation matrix for various US debt returns and US large and small company stock returns using monthly data from January 1926 to December 2012.

10 The correlation coefficient for the 1990s was not significantly different from that for the 1980s at the 0.10 significance level. A test for this type of hypothesis on the correlation coefficient can be conducted using Fisher's z-transformation. See Daniel and Terrell (1995) for information on this method.

Table 5 Correlations among US Stock and Debt Returns, 1926–2012

All	US Large Co. Stocks	US Small Co. Stocks	US Long-Term Corp.	US Long-Term Govt.	US T-Bills
US Large Co. Stocks	1.00				
US Small Co. Stocks	0.79	1.00			
US Long-Term Corp.	0.16	0.06	1.00		
US Long-Term Govt.	0.01	–0.08	0.89	1.00	
US T-bills	–0.01	–0.09	0.15	0.18	1.00

Source: Ibbotson Associates.

The first column of numbers, in particular, shows the correlations of US large company stock returns with small company stock returns and various debt returns. As expected, large and small company stocks have a high correlation (0.79). In contrast, large company stock returns are almost completely uncorrelated (–0.01) with Treasury bill returns for this period. Long-term corporate debt returns are somewhat more correlated (0.16) with large company stock returns.

Long-term government bonds, however, have a very low correlation (0.01) with large company stock returns. We expect some correlation between these variables because interest rate increases reduce the present value of future cash flows for both bonds and stocks. The low correlation between these two return series, however, shows that other factors affect the returns on stocks besides interest rates. Without these other factors, the correlation between bond and stock returns would be higher.

The third column of numbers in Table 5 shows that the correlation between long-term government bond and corporate bond returns is quite high (0.89) for this time period. Although this correlation is the highest in the entire matrix, it is not 1. The correlation is less than 1 because the default premium for long-term corporate bonds changes, whereas US government bonds do not incorporate a default premium. As a result, changes in required yields for government bonds have a correlation less than 1 with changes in required yields for corporate bonds, and return correlations between government bonds and corporate bonds are also below 1. Note also that T-bill returns have a very low correlation with all other return series.

In the next example, correlation is used in a financial statement setting to show that net income is an inadequate proxy for cash flow.

EXAMPLE 6

Correlations among Net Income, Cash Flow from Operations, and Free Cash Flow to the Firm

Net income (NI), cash flow from operations (CFO), and free cash flow to the firm (FCFF) are three measures of company performance that analysts often use to value companies. Differences in these measures for given companies would not cause differences in the relative valuation if the measures were highly correlated.

CFO equals net income plus the net noncash charges that were subtracted to obtain net income, minus the company's investment in working capital during the same time period. FCFF equals CFO plus net-of-tax interest expense, minus the company's investment in fixed capital over the time period. FCFF may be interpreted as the cash flow available to the company's suppliers of capital (debtholders and shareholders) after all operating expenses have been paid and necessary investments in working and fixed capital have been made.[11]

Some analysts base their valuations only on NI, ignoring CFO and FCFF. If the correlations among NI, CFO, and FCFF were very high, then an analyst's decision to ignore CFO and FCFF would be easy to understand because NI would then appear to capture everything one needs to know about cash flow.

Table 6 shows the correlations among NI, CFO, and FCFF for a group of six publicly traded US companies involved in retailing women's clothing for 2001. Before computing the correlations, we normalized all of the data by dividing each company's three performance measures by the company's revenue for the year.[12]

Table 6 Correlations among Performance Measures: US Women's Clothing Stores, 2001

	NI	CFO	FCFF
NI	1.0000		
CFO	0.6959	1.0000	
FCFF	0.4045	0.8217	1.0000

Source: Compustat.

Because CFO and FCFF include NI as a component (in the sense that CFO and FCFF can be obtained by adding and subtracting various quantities from NI), we might expect that the correlations between NI and CFO and between NI and FCFF would be positive. Table 6 supports that conclusion. These correlations with NI, however, are much smaller than the correlation between CFO and FCFF (0.8217). The lowest correlation in the table is between NI and FCFF (0.4045). This relatively low correlation shows that NI contained some

11 For more on these three measures and their use in equity valuation, see the Level II CFA Program curriculum reading "Free Cash Flow Valuation."

12 The results in this table are based on data for all women's clothing stores (US Occupational Health and Safety Administration Standard Industrial Classification 5621) with a market capitalization of more than $250 million at the end of 2001. The market-cap criterion was used to eliminate microcap firms, whose performance-measure correlations may be different from those of higher-valued firms.

but far from all the information in FCFF for these companies in 2001. Later in this reading, we will test whether the correlation between NI and FCFF is significantly different from zero.

The final example in this section introduces a growing area of activity in uncovering relationships among variables.

EXAMPLE 7

Analysis of Large Datasets—Big Data

Massive amounts of data containing information of potential value to investors are created and captured on a daily basis. These data include both structured data—such as order book data and security returns—and data lacking recognizable structure, generated by a vast number of activities on the internet and elsewhere. The term "big data" has been used to refer to massively large datasets. To acknowledge key features besides size, the term "alternative data" has also been used to refer to these data. In activities from marketing to investments, big, or alternative, data are being analyzed, using computational means, to discover patterns and associations that can afford profit or competitive advantage. Correlation analysis and linear regression as described in this reading are also concerned with associations, but typically are applied to structured data and make assumptions related to conducting statistical hypothesis tests. This sidebar introduces the use of big data, explaining the unstructured forms characteristic of most work in the area.

Existing in many formats and locations, unstructured data may be described as:

- open-source: data that are freely available for public consumption, such as the US government's open data project (approximately 193,000 databases)
- geospatial: data that contain a geographical component, such as store location data or satellite imagery
- sentiment based: data perceived to contain information that may indicate sentiment, including online mentions about a particular brand, concept, or product
- web-based content: data generated from the world wide web, such as internet search activity or online purchases
- micro-level: data at an individual or firm level, such as press releases or product sales prices
- macro-level: data at an aggregate or economy level, such as trade flows or bank lending amounts

Unstructured data often require transformation into a more useable, or structured, form before they can be analyzed. To transform and analyze big data, machine learning algorithms and advanced statistical methods are often used. An interesting application of big data was the evaluation of a proposed merger of UK betting shops.

Analysis of Ladbrokes–Gala Coral Merger by Schroders Asset Management

In 2015, Ladbrokes and Gala Coral, the second and third largest UK-based betting shops, announced their intention to merge. The merged entity would have a dominant online presence and own more than 4,000 land-based shops,

far more than the next largest competitor, William Hill, at 2,370 shops. The UK Competition & Markets Authority (CMA), concerned with the market concentration that would result from the merger, announced that a sell off of shops by the entities would be necessary to receive merger approval. Speculation ranged widely over how many shops would close.

By leveraging large amounts of unstructured data, UK-based asset manager Schroders was able to reach a more informed view of the proposed deal than would otherwise have been possible. Using geospatial analysis, store location data for the more than 4,000 stores, and CMA regulatory guidelines, Schroders estimated the number of shop closures that would be necessary to receive merger approval. Their estimate came very close to the required store closures later announced by the CMA. By identifying, collecting, and analyzing large amounts of unstructured, unconventional data, Schroders uncovered timely intelligence that enabled them to more accurately assess deal implications and potential valuation consequences of the Ladbrokes and Gala Coral merger.

Based on "Harnessing the Data Deluge," by Ben Wicks and Mark Ainsworth, in Schroders *Investment Horizons* (Issue 6, 2016), pages 2–5.

2.6 Testing the Significance of the Correlation Coefficient

Significance tests allow us to assess whether apparent relationships between random variables are the result of chance. If we decide that the relationships do not result from chance, we will be inclined to use this information in predictions because a good prediction of one variable will help us predict the other variable. Using the data in Table 2, we calculated 0.9083 as the sample correlation between long-term money growth and long-term inflation in six industrialized countries between 1980 and 2012. That estimated correlation seems high, but is it significantly different from 0? Before we can answer this question, we must know some details about the distribution of the underlying variables themselves. For purposes of simplicity, let us assume that both of the variables are normally distributed.[13]

We propose two hypotheses: the null hypothesis, H_0, that the correlation in the population is 0 ($\rho = 0$); and the alternative hypothesis, H_a, that the correlation in the population is different from 0 ($\rho \neq 0$).

The alternative hypothesis is a test that the correlation is not equal to 0; therefore, a two-tailed test is appropriate. As long as the two variables are distributed normally, we can test to determine whether the null hypothesis should be rejected using the sample correlation, r. The formula for the t-test is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{3}$$

This test statistic has a t-distribution with $n - 2$ degrees of freedom if the null hypothesis is true. One practical observation concerning Equation 3 is that the magnitude of r needed to reject the null hypothesis H_0: $\rho = 0$ decreases as sample size n increases, for two reasons. First, as n increases, the number of degrees of freedom increases and the absolute value of the critical value t_c decreases. Second, the absolute value of the numerator increases with larger n, resulting in larger-magnitude t-values. For example, with sample size $n = 12$, $r = 0.58$ results in a t-statistic of 2.252 that is just significant at the 0.05 level ($t_c = 2.228$). With a sample size $n = 32$, a smaller sample

13 Actually, we must assume that the variables come from a bivariate normal distribution. If two variables, X and Y, come from a bivariate normal distribution, then for each value of X the distribution of Y is normal. See, for example, Greene (2018).

correlation $r = 0.35$ yields a t-statistic of 2.046 that is just significant at the 0.05 level ($t_c = 2.042$); the $r = 0.35$ would not be significant with a sample size of 12 even at the 0.10 significance level. Another way to make this point is that sampling from the same population, a false null hypothesis H_0: $\rho = 0$ is more likely to be rejected as we increase sample size, all else equal.

EXAMPLE 8

Testing the Correlation between Money Supply Growth and Inflation

Earlier in this reading, we showed that the sample correlation between long-term money supply growth and long-term inflation in six industrialized countries was 0.9083 during the 1980–2012 period. Suppose we want to test the null hypothesis, H_0, that the true correlation in the population is 0 ($\rho = 0$) against the alternative hypothesis, H_a, that the correlation in the population is different from 0 ($\rho \neq 0$).

1 Calculate the test statistic to test the null hypothesis given above.

2 Determine whether the null hypothesis is rejected or not rejected at the 0.05 level of significance.

Solution to 1:

Recalling that this sample has six observations, we can compute the statistic for testing the null hypothesis as follows:

$$t = \frac{0.9083\sqrt{6-2}}{\sqrt{1-0.9083^2}} = 4.343$$

The value of the test statistic is 4.343.

Solution to 2:

As the table of critical values of the t-distribution for a two-tailed test shows, for a t-distribution with $n - 2 = 6 - 2 = 4$ degrees of freedom at the 0.05 level of significance, we can reject the null hypothesis (that the population correlation is equal to 0) if the value of the test statistic is greater than 2.776 or less than −2.776. The fact that we can reject the null hypothesis of no correlation based on only six observations is quite unusual; it further demonstrates the strong relation between long-term money supply growth and long-term inflation in these six countries.

EXAMPLE 9

Testing the Yen–Canadian Dollar Return Correlation

The data in Table 3 showed that the sample correlation between the GBP monthly returns to Japanese yen and Canadian dollar was 0.3091 for the period from January 2007 through December 2013.

Can we reject a null hypothesis that the underlying or population correlation equals 0 at the 0.05 level of significance?

Solution:

With 84 months from January 2007 through December 2013, we use the following statistic to test the null hypothesis, H_0, that the true correlation in the population is 0, against the alternative hypothesis, H_a, that the correlation in the population is different from 0:

$$t = \frac{0.3091\sqrt{84-2}}{\sqrt{1-0.3091^2}} = 2.9431$$

At the 0.05 significance level, the critical level for this test statistic is 1.99 (n = 84, degrees of freedom = 82). When the test statistic is either larger than 1.99 or smaller than −1.99, we can reject the hypothesis that the correlation in the population is 0. The test statistic is 2.9431, so we can reject the null hypothesis.

Note that the sample correlation coefficient in this case is significantly different from 0 at the 0.05 level, even though the coefficient is much smaller than that in the previous example. The correlation coefficient, though smaller, is still significant because the sample is much larger (84 observations instead of 6 observations).

The above example shows the importance of sample size in tests of the significance of the correlation coefficient. The following example also shows the importance of sample size and examines the relationship at the 0.01 level of significance as well as at the 0.05 level.

EXAMPLE 10

The Correlation between Bond Returns and T-Bill Returns

Table 5 showed that the sample correlation between monthly returns to US long-term government bonds and monthly returns to T-bills was 0.18 from January 1926 through December 2012.

Can we reject a null hypothesis that the underlying or population correlation coefficient equals 0 at the 0.05 and 0.01 levels of significance?

Solution:

There are 1,044 months during the period January 1926 to December 2012. Therefore, to test the null hypothesis, H_0 (that the true correlation in the population is 0), against the alternative hypothesis, H_a (that the correlation in the population is different from 0), we use the following test statistic:

$$t = \frac{0.18\sqrt{1{,}044-2}}{\sqrt{1-0.18^2}} = 5.9069$$

At the 0.05 significance level, the critical value for the test statistic is approximately 1.96. At the 0.01 significance level, the critical value for the test statistic is approximately 2.58. The test statistic is 5.9069, so we can reject the null hypothesis of no correlation in the population at both the 0.05 and 0.01 levels. This example shows that, in large samples, even relatively small correlation coefficients can be significantly different from zero.

In the final example of this section, we explore another situation of small sample size.

EXAMPLE 11

Testing the Correlation between Net Income and Free Cash Flow to the Firm

Earlier in this reading, we showed that the sample correlation between NI and FCFF for six women's clothing stores was 0.4045 in 2001. Suppose we want to test the null hypothesis, H_0, that the true correlation in the population is 0 ($\rho = 0$) against the alternative hypothesis, H_a, that the correlation in the population is different from 0 ($\rho \neq 0$). Recalling that this sample has six observations, we can compute the statistic for testing the null hypothesis as follows:

$$t = \frac{0.4045\sqrt{6-2}}{\sqrt{1-0.4045^2}} = 0.8846$$

With $n - 2 = 6 - 2 = 4$ degrees of freedom and a 0.05 significance level, we reject the null hypothesis that the population correlation equals 0 for values of the test statistic greater than 2.776 or less than −2.776. In this case, however, the *t*-statistic is 0.8846, so we cannot reject the null hypothesis. Therefore, for this sample of women's clothing stores, there is no **statistically significant** correlation between NI and FCFF, when each is normalized by dividing by sales for the company.[14]

The scatter plot creates a visual picture of the relationship between two variables, while the correlation coefficient quantifies the existence of any linear relationship. Large absolute values of the correlation coefficient indicate strong linear relationships. Positive coefficients indicate a positive relationship and negative coefficients indicate a negative relationship between two data sets. In Examples 9 and 10, we saw that relatively small sample correlation coefficients (0.3091 and 0.18, respectively) can be statistically significant and thus might provide valuable information about the behavior of economic variables.

Next we will introduce linear regression, another tool useful in examining the relationship between two variables.

LINEAR REGRESSION

3

Linear regression with one independent variable, sometimes called simple linear regression, models the relationship between two variables as a straight line. When the linear relationship between the two variables is significant, linear regression provides a simple model for forecasting the value of one variable, known as the dependent variable, given the value of the second variable, known as the independent variable. The following sections explain linear regression in more detail.

14 It is worth repeating that the smaller the sample, the greater the evidence in terms of the magnitude of the sample correlation needed to reject the null hypothesis of zero correlation. With a sample size of 6, the absolute value of the sample correlation would need to be greater than 0.81 (carrying two decimal places) for us to reject the null hypothesis. Viewed another way, the value of 0.4045 in the text would be significant if the sample size were 24, because $0.4045(24-2)^{1/2}/(1-0.4045^2)^{1/2} = 2.075$, which is greater than the critical *t*-value of 2.074 at the 0.05 significance level with 22 degrees of freedom.

3.1 Linear Regression with One Independent Variable

As a financial analyst, you will often want to understand the relationship between financial or economic variables, or to predict the value of one variable using information about the value of another variable. For example, you may want to know the impact of changes in the 10-year Treasury bond yield on the earnings yield of the S&P 500 (the earnings yield is the reciprocal of the price-to-earnings ratio). If the relationship between those two variables is linear, you can use linear regression to summarize it.

Linear regression allows us to use one variable to make predictions about another, test hypotheses about the relation between two variables, and quantify the strength of the relationship between the two variables. The remainder of this reading focuses on linear regression with a single independent variable. In the next reading, we will examine regression with more than one independent variable.

Regression analysis begins with the dependent variable (denoted Y), the variable that you are seeking to explain. The independent variable (denoted X) is the variable you are using to explain changes in the dependent variable. For example, you might try to explain small-stock returns (the dependent variable) based on returns to the S&P 500 (the independent variable). Or you might try to explain inflation (the dependent variable) as a function of growth in a country's money supply (the independent variable).

Linear regression assumes a linear relationship between the dependent and the independent variables. The following regression equation describes that relation:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i,\ i = 1, \ldots, n \quad \textbf{(4)}$$

This equation states that the **dependent variable**, Y, is equal to the intercept, b_0, plus a slope coefficient, b_1, times the **independent variable**, X, plus an **error term**, ε. The error term represents the portion of the dependent variable that cannot be explained by the independent variable. We refer to the intercept b_0 and the slope coefficient b_1 as the **regression coefficients**.

Regression analysis uses two principal types of data: cross-sectional and time series. Cross-sectional data involve many observations on X and Y for the same time period. Those observations could come from different companies, asset classes, investment funds, people, countries, or other entities, depending on the regression model. For example, a cross-sectional model might use data from many companies to test whether predicted earnings-per-share growth explains differences in price-to-earnings ratios (P/Es) during a specific time period. The word "explain" is frequently used in describing regression relationships. One estimate of a company's P/E that does not depend on any other variable is the average P/E. If a regression of a P/E on an independent variable tends to give more accurate estimates of P/E than just assuming that the company's P/E equals the average P/E, we say that the independent variable helps *explain* P/Es because using that independent variable improves our estimates. Finally, note that if we use cross-sectional observations in a regression, we usually denote the observations as $i = 1, 2, \ldots, n$.

Time-series data use many observations from different time periods for the same company, asset class, investment fund, person, country, or other entity, depending on the regression model. For example, a time-series model might use monthly data from many years to test whether US inflation rates determine US short-term interest rates.[15] If we use time-series data in a regression, we usually denote the observations as $t = 1, 2, \ldots, T$.[16]

15 A mix of time-series and cross-sectional data, also known as panel data, is now frequently used in financial analysis. The analysis of panel data is an advanced topic that Greene (2018) discusses in detail.

16 In this reading, we primarily use the notation $i = 1, 2, \ldots, n$ even for time series to prevent confusion that would be caused by switching back and forth between different notations.

Exactly how does linear regression estimate b_0 and b_1? Linear regression, also known as linear least squares, computes a line that best fits the observations; it chooses values for the intercept, b_0, and slope, b_1, that minimize the sum of the squared vertical distances between the observations and the regression line. Linear regression chooses the **estimated parameters** or **fitted parameters** $\hat{b}_0$ and $\hat{b}_1$ in Equation 4 to minimize[17]

$$\sum_{i=1}^{n}\left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)^2 \tag{5}$$

In this equation, the term $\left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)^2$ means (dependent variable – predicted value of dependent variable)^2. Using this method to estimate the values of $\hat{b}_0$ and $\hat{b}_1$, we can fit a line through the observations on X and Y that best explains the value that Y takes for any particular value of X.[18]

Note that we never observe the population parameter values b_0 and b_1 in a regression model. Instead, we observe only $\hat{b}_0$ and $\hat{b}_1$, which are estimates of the population parameter values. Thus predictions must be based on the parameters' estimated values, and testing is based on estimated values in relation to the hypothesized population values.

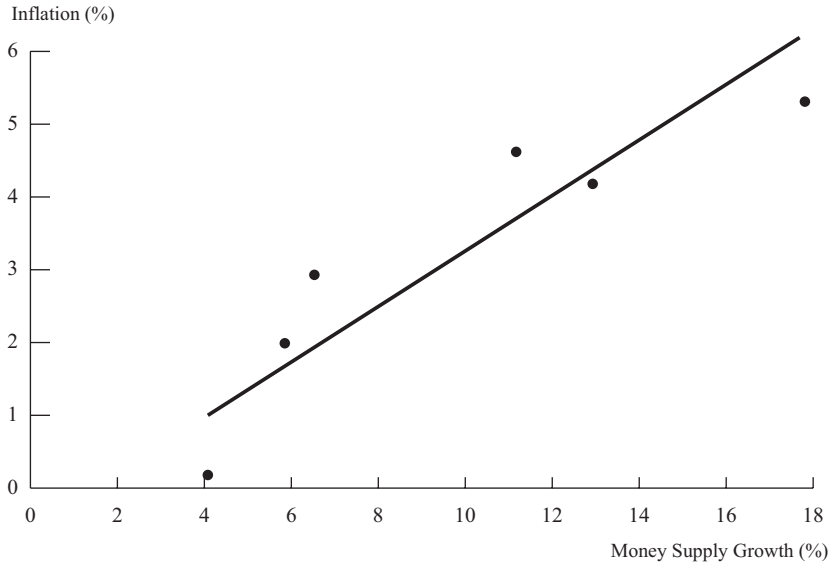
Figure 8 gives a visual example of how linear regression works. The figure shows the linear regression that results from estimating the regression relation between the annual rate of inflation (the dependent variable) and annual rate of money supply growth (the independent variable) for six industrialized countries from 1980 to 2012 ($n = 6$).[19] The equation to be estimated is Long-term rate of inflation = $b_0 + b_1$ (Long-term rate of money supply growth) + ε.

17 Hats over the symbols for coefficients indicate estimated values.

18 For a discussion of the precise statistical sense in which the estimates of b_0 and b_1 are optimal, see Greene (2018).

19 These data appear in Table 2.

Figure 8 Fitted Regression Line Explaining the Inflation Rate Using Growth in the Money Supply by Country, 1980–2012



Source: International Monetary Fund.

The distance from each of the six data points to the fitted regression line is the regression residual, which is the difference between the actual value of the dependent variable and the predicted value of the dependent variable made by the regression equation. Linear regression chooses the estimated coefficients $\hat{b}_0$ and $\hat{b}_1$ in Equation 4 such that the sum of the squared vertical distances is minimized. The estimated regression equation is Long-term inflation = –0.0003 + 0.3327 (Long-term money supply growth).[20]

According to this regression equation, if the long-term money supply growth is 0 for any particular country, the long-term rate of inflation in that country will be –0.03 percent. For every 1-percentage-point increase in the long-term rate of money supply growth for a country, the long-term inflation rate is predicted to increase by 0.3327 percentage points. In a regression such as this one, which contains one independent variable, the slope coefficient equals Cov(Y,X)/Var(X). We can solve for the slope coefficient using data from Table 2, excerpted here:

Table 2 (excerpted)

	Money Supply Growth Rate X_i	Inflation Rate Y_i	Cross-Product $(X_i - \bar{X})(Y_i - \bar{Y})$	Squared Deviations $(X_i - \bar{X})^2$	Squared Deviations $(Y_i - \bar{Y})^2$
Sum	0.5837	0.1921	0.004485	0.013481	0.001809
Average	0.0973	0.0320			
Covariance			**0.000897**		

20 We entered the monthly returns as decimals. Also, we used rounded numbers in the formulas discussed later to estimate the regression equation.

Table 2 (Continued)

	Money Supply Growth Rate X_i	Inflation Rate Y_i	Cross-Product $(X_i - \bar{X})(Y_i - \bar{Y})$	Squared Deviations $(X_i - \bar{X})^2$	Squared Deviations $(Y_i - \bar{Y})^2$
Variance				**0.002696**	0.000362
Standard deviation				0.051926	0.019019

$$\begin{aligned} \text{Cov}(Y,X) &= 0.000897 \\ \text{Var}(X) &= 0.002696 \\ \text{Cov}(Y,X)/\text{Var}(X) &= 0.000897/0.002696 \\ \hat{b}_1 &= 0.3327 \end{aligned}$$

In a linear regression, the regression line fits through the point corresponding to the means of the dependent and the independent variables. As shown in Table 1 (excerpted below), from 1980 to 2012, the mean long-term growth rate of the money supply for these six countries was 9.73 percent, whereas the mean long-term inflation rate was 3.20 percent.

Table 1 (excerpted)

	Money Supply Growth Rate	Inflation Rate
Average	9.73%	3.20%

Because the point (9.73, 3.20) lies on the regression line $\hat{b}_0 = \bar{Y} - \hat{b}_1\bar{X}$, we can solve for the intercept using this point as follows:

$$\hat{b}_0 = 0.0320 - 0.3327(0.0973) = -0.0003$$

We are showing how to solve the linear regression equation step by step to make the source of the numbers clear. Typically, an analyst will use the data analysis function on a spreadsheet or a statistical package to perform linear regression analysis. Later, we will discuss how to use regression residuals to quantify the uncertainty in a regression model.

3.2 Assumptions of the Linear Regression Model

We have discussed how to interpret the coefficients in a linear regression model. Now we turn to the statistical assumptions underlying this model. Suppose that we have n observations on both the dependent variable, Y, and the independent variable, X, and we want to estimate Equation 4:

$$Y_i = b_0 + b_1X_i + \varepsilon_i,\ i = 1, \ldots, n$$

To be able to draw valid conclusions from a linear regression model with a single independent variable, we need to make the following six assumptions, known as the classic normal linear regression model assumptions:

1. The relationship between the dependent variable, Y, and the independent variable, X is linear in the parameters b_0 and b_1. This requirement means that b_0 and b_1 are raised to the first power only and that neither b_0 nor b_1 is multiplied or divided by another regression parameter (as in b_0/b_1, for example). The requirement does not exclude X from being raised to a power other than 1.
2. The independent variable, X, is not random.[21]
3. The expected value of the error term is 0: $E(\varepsilon) = 0$.
4. The variance of the error term is the same for all observations: $E\left(\varepsilon_i^2\right) = \sigma_\varepsilon^2$, i = 1, …, n.
5. The error term, ε, is uncorrelated across observations. Consequently, $E(\varepsilon_i\varepsilon_j) = 0$ for all i not equal to j.[22]
6. The error term, ε, is normally distributed.[23]

Now we can take a closer look at each of these assumptions.

Assumption 1 is critical for a valid linear regression. If the relationship between the independent and dependent variables is nonlinear in the parameters, then estimating that relation with a linear regression model will produce invalid results. For example, $Y_i = b_0 e^{b_1 X_i} + \varepsilon_i$ is nonlinear in b_1, so we could not apply the linear regression model to it.[24]

Even if the dependent variable is nonlinear, linear regression can be used as long as the regression is linear in the parameters. So, for example, linear regression can be used to estimate the equation $Y_i = b_0 + b_1 X_i^2 + \varepsilon_i$.

Assumptions 2 and 3 ensure that linear regression produces the correct estimates of b_0 and b_1.

Assumptions 4, 5, and 6 let us use the linear regression model to determine the distribution of the estimated parameters $\hat{b}_0$ and $\hat{b}_1$ and thus test whether those coefficients have a particular value.

- Assumption 4, that the variance of the error term is the same for all observations, is also known as the homoskedasticity assumption. The reading on multiple regression discusses how to test for and correct violations of this assumption.
- Assumption 5, that the errors are uncorrelated across observations, is also necessary for correctly estimating the variances of the estimated parameters $\hat{b}_0$ and $\hat{b}_1$. The reading on multiple regression discusses violations of this assumption.

21 Although we assume that the independent variable in the regression model is not random, that assumption is clearly often not true. For example, it is unrealistic to assume that the monthly returns to the S&P 500 are not random. If the independent variable is random, then is the regression model incorrect? Fortunately, no. Econometricians have shown that even if the independent variable is random, we can still rely on the results of regression models given the crucial assumption that the error term is uncorrelated with the independent variable. The mathematics underlying this reliability demonstration, however, are quite difficult. See, for example, Greene (2018).

22 $\text{Var}(\varepsilon_i) = E[\varepsilon_i - E(\varepsilon_i)]^2 = E(\varepsilon_i - 0)^2 = E(\varepsilon_i)^2$. $\text{Cov}(\varepsilon_i, \varepsilon_j) = E\{[\varepsilon_i - E(\varepsilon_i)][\varepsilon_j - E(\varepsilon_j)]\} = E[(\varepsilon_i - 0)\ (\varepsilon_j - 0)]$ $= E(\varepsilon_i\varepsilon_j) = 0$.

23 If the regression errors are not normally distributed, we can still use regression analysis. Econometricians who dispense with the normality assumption use chi-square tests of hypotheses rather than F-tests. This difference usually does not affect whether the test will result in a particular null hypothesis being rejected.

24 For more information on nonlinearity in the parameters, see Gujarati and Porter (2011).

- Assumption 6, that the error term is normally distributed, allows us to easily test a particular hypothesis about a linear regression model.[25]

EXAMPLE 12

Evaluating Economic Forecasts (2)

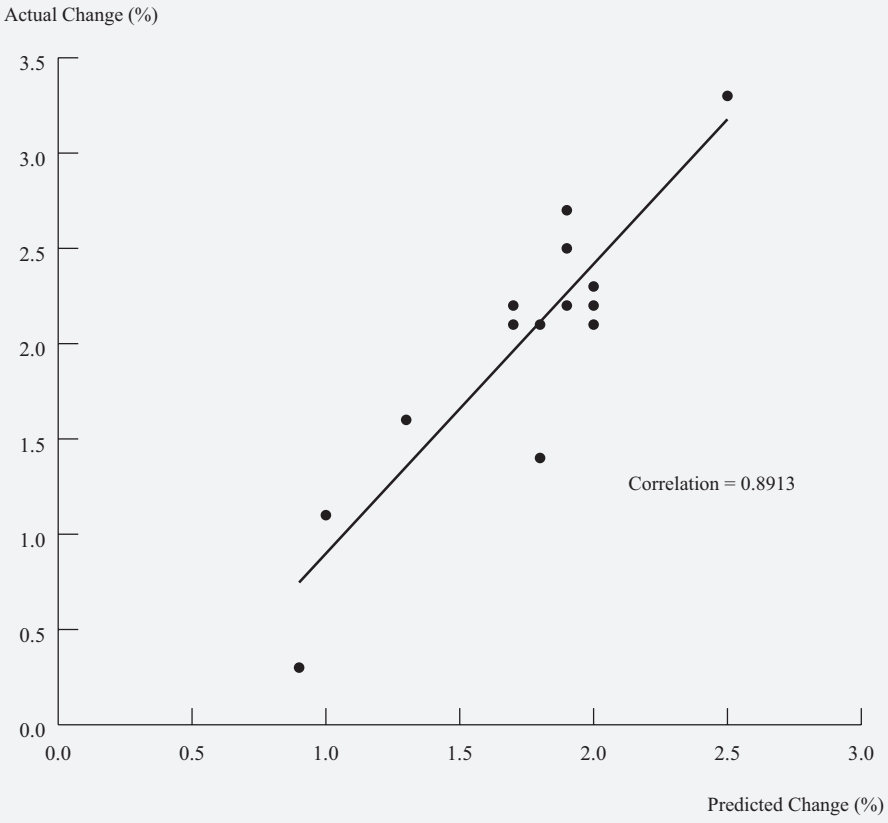
If economic forecasts were completely accurate, every prediction of change in an economic variable in a quarter would exactly match the actual change that occurs in that quarter. Even though forecasts can be inaccurate, we hope at least that they are unbiased—that is, that the expected value of the forecast error is zero. An unbiased forecast can be expressed as E(Actual change − Predicted change) = 0. In fact, most evaluations of forecast accuracy test whether forecasts are unbiased.[26]

Figure 9 repeats Figure 7 in showing a scatter plot of the mean forecast made in the first quarter of a year for the percentage change in HICP during that year and the actual percentage change in HICP, from 1999 through 2013, but it adds the fitted regression line for the equation Actual percentage change = $b_0 + b_1$ (Predicted percentage change) + ε. If the forecasts are unbiased, the intercept, b_0, should be 0 and the slope, b_1, should be 1. We should also find E(Actual change − Predicted change) = 0. If forecasts are actually unbiased, as long as $b_0 = 0$ and $b_1 = 1$, the error term [Actual change − b_0 − b_1(Predicted change)] will have an expected value of 0, as required by Assumption 3 of the linear regression model. With unbiased forecasts, any other values of b_0 and b_1 would yield an error term with an expected value different from 0.

25 For large sample sizes, we may be able to drop the assumption of normality by appeal to the central limit theorem; see Greene (2018). Asymptotic theory shows that, in many cases, the test statistics produced by standard regression programs are valid even if the error term is not normally distributed. Non-normality of some financial time series can be quite severe. With severe non-normality, even with a relatively large number of observations, invoking asymptotic theory to justify using test statistics from linear regression models may be inappropriate.

26 See, for example, Keane and Runkle (1990).

Figure 9 Actual Change in Euro Area HICP versus Predicted Change



Source: European Central Bank.

If $b_0 = 0$ and $b_1 = 1$, our best guess of actual change in HICP would be 0 if professional forecasters' predictions of change in HICP were 0. For every 1-percentage-point increase in the prediction of change by the professional forecasters, the regression model would predict a 1-percentage-point increase in actual change.

The fitted regression line in Figure 9 comes from the equation Actual change = −0.7006 + 1.5538(Predicted change). It seems that the estimated values of b_0 and b_1 are not particularly close to the values $b_0 = 0$ and $b_1 = 1$ that are consistent with unbiased forecasts. Later in this reading, we discuss how to test the hypotheses that $b_0 = 0$ and $b_1 = 1$.

3.3 The Standard Error of Estimate

The linear regression model sometimes describes the relationship between two variables quite well, but sometimes it does not. We must be able to distinguish between these two cases in order to use regression analysis effectively. Therefore, in this section and the next, we discuss statistics that measure how well a given linear regression model captures the relationship between the dependent and independent variables.

Figure 9, for example, shows a strong relation between predicted inflation and actual inflation. If we knew professional forecasters' predictions for inflation in a particular quarter, we would be reasonably certain that we could use this regression model to forecast actual inflation relatively accurately.

In other cases, however, the relation between the dependent and independent variables is not strong. Figure 10 adds a fitted regression line to the data on inflation and stock returns during 1990 to 2013 from Figure 6. In this figure, the actual observations are generally much farther from the fitted regression line than in Figure 9. Using the estimated regression equation to predict monthly stock returns assuming a particular level of inflation might result in an inaccurate forecast.

As noted, the regression relation in Figure 10 is less precise than that in Figure 9. The standard error of estimate (sometimes called the standard error of the regression) measures this uncertainty. This statistic is very much like the standard deviation for a single variable, except that it measures the standard deviation of $\hat{\varepsilon}_i$, the residual term in the regression.

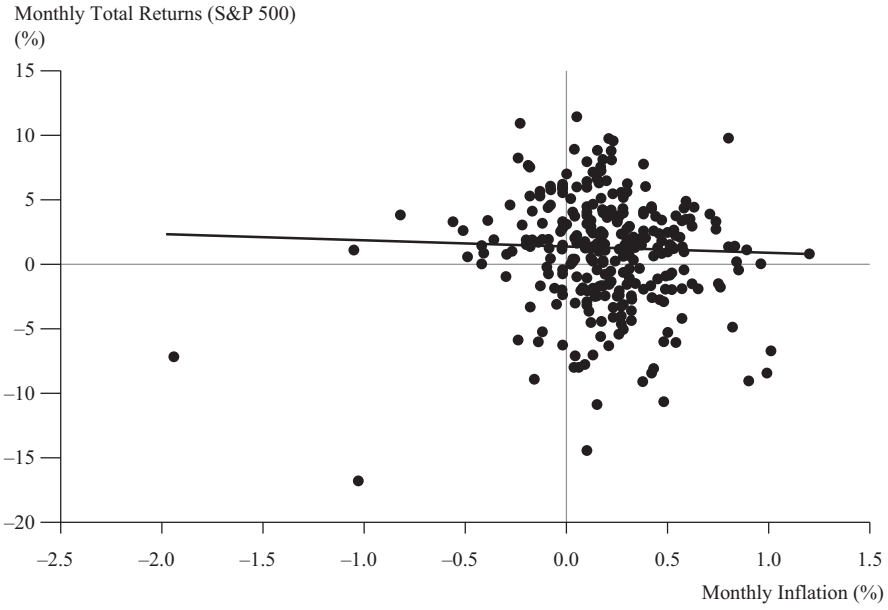
The formula for the standard error of estimate (SEE) for a linear regression model with one independent variable is

$$\text{SEE} = \left[\frac{\sum_{i=1}^{n}\left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)^2}{n-2}\right]^{1/2} = \left[\frac{\sum_{i=1}^{n}\left(\hat{\varepsilon}_i\right)^2}{n-2}\right]^{1/2} \quad \textbf{(6)}$$

In the numerator of this equation, we are computing the difference between the dependent variable's actual value for each observation and its predicted value $\left(\hat{b}_0 + \hat{b}_1 X_i\right)$ for each observation. The difference between the actual and predicted values of the dependent variable is the regression residual, $\hat{\varepsilon}_i$.

Equation 6 looks very much like the formula for computing a standard deviation, except that $n - 2$ appears in the denominator instead of $n - 1$. We use $n - 2$ because the sample includes n observations and the linear regression model estimates two parameters ($\hat{b}_0$ and $\hat{b}_1$); the difference between the number of observations and the number of parameters is $n - 2$. This difference is also called the degrees of freedom; it is the denominator needed to ensure that the estimated standard error of estimate is unbiased.

Figure 10 Fitted Regression Line Explaining Stock Returns by Inflation during 1990–2013



Sources: Bureau of Labor Statistics and S&P Dow Jones Indices.

EXAMPLE 13

Computing the Standard Error of Estimate

Recall that the estimated regression equation for the inflation and money supply growth data shown in Figure 8 was $Y_i = -0.0003 + 0.3327X_i$. Table 7 uses this estimated equation to compute the data needed for the standard error of estimate.

Table 7 Computing the Standard Error of Estimate

Country	Money Supply Growth Rate X_i	Inflation Rate Y_i	Predicted Inflation Rate $\hat{Y}_i$	Regression Residual $Y_i - \hat{Y}_i$	Squared Residual $\left(Y_i - \hat{Y}_i\right)^2$
Australia	0.1117	0.0462	0.0368	0.0094	0.000088
Japan	0.0408	0.0018	0.0132	–0.0114	0.000131
South Korea	0.1781	0.0531	0.0589	–0.0058	0.000034
Switzerland	0.0585	0.0199	0.0191	0.0008	0.000001
United Kingdom	0.1293	0.0418	0.0427	–0.0009	0.000001
United States	0.0653	0.0293	0.0214	0.0079	0.000063
Sum					0.000316

Source: International Monetary Fund.

The first and second columns of numbers in Table 7 show the long-term money supply growth rates, X_i, and long-term inflations rates, Y_i, for the six countries. The third column of numbers shows the predicted value of the dependent variable from the fitted regression equation for each observation. For the United States, for example, the predicted value of long-term inflation is–0.0003 + 0.3327(0.0653) = 0.0214 or 2.14 percent. The next-to-last column contains the regression residual, which is the difference between the actual value of the dependent variable, Y_i, and the predicted value of the dependent variable, $\left(\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i\right)$. So for the United States, the residual is equal to 0.0293 – 0.0214 = 0.0079 or 0.79 percent. The last column contains the squared regression residual. The sum of the squared residuals is 0.000316. Applying the formula for the standard error of estimate, we obtain

$$\left(\frac{0.000316}{6-2}\right)^{1/2} = 0.008895$$

Thus the standard error of estimate is about 0.89 percent.

Later, we will combine this estimate with estimates of the uncertainty about the parameters in this regression to determine confidence intervals for predicting inflation rates from money supply growth. We will see that smaller standard errors result in more accurate predictions.

3.4 The Coefficient of Determination

Although the standard error of estimate gives some indication of how certain we can be about a particular prediction of Y using the regression equation, it still does not tell us how well the independent variable explains variation in the dependent variable. The coefficient of determination does exactly this: It measures the fraction of the total variation in the dependent variable that is explained by the independent variable.

We can compute the coefficient of determination in two ways. The simpler method, which can be used in a linear regression with one independent variable, is to square the correlation coefficient between the dependent and independent variables. For example, recall that the correlation coefficient between the long-term rate of money growth and the long-term rate of inflation between 1980 and 2012 for six industrialized countries was 0.9083. Thus the coefficient of determination in the regression shown in Figure 8 is $(0.9083)^2 = 0.8250$. So in this regression, the long-term rate of money supply growth explains approximately 82.5 percent of the variation in the long-term rate of inflation across the countries between 1980 and 2012. (Relatedly, note that the square root of the coefficient of determination in a one-independent-variable linear regression, after attaching the sign of the estimated slope coefficient, gives the correlation coefficient between the dependent and independent variables.)

The problem with this method is that it cannot be used when we have more than one independent variable.[27] Therefore, we need an alternative method of computing the coefficient of determination for multiple independent variables. We now present the logic behind that alternative.

If we did not know the regression relationship, our best guess for the value of any particular observation of the dependent variable would simply be $\bar{Y}$, the mean of the dependent variable. One measure of accuracy in predicting Y_i based on $\bar{Y}$ is the sample variance of Y_i, $\sum_{i=1}^{n} \frac{\left(Y_i - \bar{Y}\right)^2}{n-1}$. An alternative to using $\bar{Y}$ to predict a particular obser-

27 We will discuss such models in the reading on multiple regression.

vation Y_i is using the regression relationship to make that prediction. In that case, our predicted value would be $\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$. If the regression relationship works well, the error in predicting Y_i using $\hat{Y}_i$ should be much smaller than the error in predicting Y_i using $\bar{Y}$. If we call $\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$ the total variation of Y and $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ the unexplained variation from the regression, then we can measure the explained variation from the regression using the following equation:

$$\text{Total variation} = \text{Unexplained variation} + \text{Explained variation} \quad \textbf{(7)}$$

The coefficient of determination is the fraction of the total variation that is explained by the regression. This gives us the relationship

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}} = 1 - \frac{\text{Unexplained variation}}{\text{Total variation}} \quad \textbf{(8)}$$

Note that total variation equals explained variation plus unexplained variation, as shown in Equation 7. Most regression programs report the coefficient of determination as R^2.[28]

EXAMPLE 14

Inflation Rate and Growth in the Money Supply

Using the data in Table 7, we can see that the unexplained variation from the regression, which is the sum of the squared residuals, equals 0.000316. Table 8 shows the computation of total variation in the dependent variable, the long-term rate of inflation.

Table 8 Computing Total Variation

Country	Money Supply Growth Rate X_i	Inflation Rate Y_i	Deviation from Mean $Y_i - \bar{Y}$	Squared Deviation $\left(Y_i - \bar{Y}\right)^2$
Australia	0.1117	0.0462	0.0142	0.000201
Japan	0.0408	0.0018	–0.0302	0.000913
South Korea	0.1781	0.0531	0.0211	0.000445
Switzerland	0.0585	0.0199	–0.0121	0.000147
United Kingdom	0.1293	0.0418	0.0098	0.000096
United States	0.0653	0.0293	–0.0027	0.000007
	Average:	0.0320	Sum:	0.001809

Source: International Monetary Fund.

28 As we illustrate in the tables of regression output later in this reading, regression programs also report multiple R, which is the correlation between the actual values and the forecast values of Y. The coefficient of determination is the square of multiple R.

The average inflation rate for this period is 3.20 percent. The next-to-last column shows the amount each country's long-term inflation rate deviates from that average; the last column shows the square of that deviation. The sum of those squared deviations is the total variation in Y for the sample (0.001809), shown in Table 8.

Compute the coefficient of determination for the regression.

Solution:

The coefficient of determination for the regression is

$$\frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}} = \frac{0.001809 - 0.000316}{0.001809} = 0.8250$$

Note that this method gives the same result that we obtained earlier. We will use this method again in the reading on multiple regression; when we have more than one independent variable, this method is the only way to compute the coefficient of determination.

3.5 Hypothesis Testing

In this section, we address testing hypotheses concerning the population values of the intercept or slope coefficient of a regression model. This topic is critical in practice. For example, we may want to check a stock's valuation using the capital asset pricing model; we hypothesize that the stock has a market-average beta or level of systematic risk. Or we may want to test the hypothesis that economists' forecasts of the inflation rate are unbiased (not overestimates or underestimates, on average). In each case, does the evidence support the hypothesis? Questions such as these can be addressed with hypothesis tests within a regression model. Such tests are often t-tests of the value of the intercept or slope coefficient(s). To understand the concepts involved in this test, it is useful to first review a simple, equivalent approach based on confidence intervals.

We can perform a hypothesis test using the confidence interval approach if we know three things: 1) the estimated parameter value, $\hat{b}_0$ or $\hat{b}_1$, 2) the hypothesized value of the parameter, b_0 or b_1, and 3) a confidence interval around the estimated parameter. A confidence interval is an interval of values that we believe includes the true parameter value, b_1, with a given degree of confidence. To compute a confidence interval, we must select the significance level for the test and know the standard error of the estimated coefficient.

Suppose we regress a stock's returns on a stock market index's returns and find that the slope coefficient ($\hat{b}_1$) is 1.5 with a standard error ($s_{\hat{b}_1}$) of 0.200. Assume we used 62 monthly observations in our regression analysis. The hypothesized value of the parameter (b_1) is 1.0, the market average slope coefficient. The estimated and the population slope coefficients are often called beta, because the population coefficient is often represented by the Greek symbol beta (β) rather than the b_1 we use in this reading. Our null hypothesis is that $b_1 = 1.0$ and $\hat{b}_1$ is the estimate for b_1. We will use a 95 percent confidence interval for our test, or we could say that the test has a significance level of 0.05.

Our confidence interval will span the range $\hat{b}_1 - t_c s_{\hat{b}_1}$ to $\hat{b}_1 + t_c s_{\hat{b}_1}$ or

$$\hat{b}_1 \pm t_c s_{\hat{b}_1} \quad (9)$$

where t_c is the critical t value.[29] The critical value for the test depends on the number of degrees of freedom for the t-distribution under the null hypothesis. The number of degrees of freedom equals the number of observations minus the number of parameters estimated. In a regression with one independent variable, there are two estimated parameters, the intercept term and the coefficient on the independent variable. For 62 observations and two parameters estimated in this example, we have 60 degrees of freedom (62 – 2). For 60 degrees of freedom, the table of critical values in the back of the book shows that the critical t-value at the 0.05 significance level is 2.00. Substituting the values from our example into Equation 9 gives us the interval

$$\begin{aligned}\hat{b}_1 \pm t_c s_{\hat{b}_1} &= 1.5 \pm 2.00(0.200) \\ &= 1.5 \pm 0.400 \\ &= 1.10 \text{ to } 1.90\end{aligned}$$

A 95% confidence interval is the interval, based on the sample value, that we would expect to include the population value with a 95% degree of confidence. Because we are testing the null hypothesis that $b_1 = 1.0$ and because our confidence interval does not include 1.0, we can reject the null hypothesis.

In practice, the most common way to test a hypothesis using a regression model is with a t-test of significance. To test the hypothesis, we can compute the statistic

$$t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \tag{10}$$

This test statistic has a t-distribution with $n - 2$ degrees of freedom because two parameters were estimated in the regression. We compare the absolute value of the t-statistic to t_c. If the absolute value of t is greater than t_c, then we can reject the null hypothesis. Substituting the values from the above example into this relationship gives the t-statistic associated with the test that the stock's beta equals 1.0 ($b_1 = 1.0$).

$$\begin{aligned}t &= \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \\ &= (1.5 - 1.0)/0.200 \\ &= 2.50\end{aligned}$$

Because $t > t_c$, we reject the null hypothesis that $b_1 = 1.0$.

The t-statistic in the example above is 2.50, and at the 0.05 significance level, $t_c = 2.00$; thus we reject the null hypothesis because $t > t_c$. This statement is equivalent to saying that we are 95 percent confident that the interval for the slope coefficient does not contain the value 1.0. If we were performing this test at the 0.01 level, however, t_c would be 2.66 and we would not reject the hypothesis because t would not be greater than t_c at this significance level. A 99 percent confidence interval for the slope coefficient does contain the value 1.0.

The choice of significance level is always a matter of judgment. When we use higher levels of confidence, the t_c increases. This choice leads to wider confidence intervals and to a decreased likelihood of rejecting the null hypothesis. Analysts often choose the 0.05 level of significance, which indicates a 5 percent chance of rejecting the null hypothesis when, in fact, it is true (a Type I error). Of course, decreasing the level of significance from 0.05 to 0.01 decreases the probability of Type I error, but it increases the probability of Type II error—failing to reject the null hypothesis when, in fact, it is false.

29 We use the t-distribution for this test because we are using a sample estimate of the standard error, s_b, rather than its true (population) value.

Often, financial analysts do not simply report whether or not their tests reject a particular hypothesis about a regression parameter. Instead, they report the *p*-value or probability value for a particular hypothesis. The *p*-value is the smallest level of significance at which the null hypothesis can be rejected. It allows the reader to interpret the results rather than be told that a certain hypothesis has been rejected or accepted. In most regression software packages, the *p*-values printed for regression coefficients apply to a test of null hypothesis that the true parameter is equal to 0 against the alternative that the parameter is not equal to 0, given the estimated coefficient and the standard error for that coefficient. For example, if the *p*-value is 0.005, we can reject the hypothesis that the true parameter is equal to 0 at the 0.5 percent significance level (99.5 percent confidence).

The standard error of the estimated coefficient is an important input for a hypothesis test concerning the regression coefficient (and for a confidence interval for the estimated coefficient). Stronger regression results lead to smaller standard errors of an estimated parameter and result in tighter confidence intervals. If the standard error ($s_{\hat{b}_1}$) in the above example were 0.100 instead of 0.200, the confidence interval range would be half as large and the *t*-statistic twice as large. With a standard error this small, we would reject the null hypothesis even at the 0.01 significance level because we would have $t = (1.5 - 1)/0.1 = 5.00$ and $t_c = 2.66$.

With this background, we can turn to hypothesis tests using actual regression results. The next three examples illustrate hypothesis tests in a variety of typical investment contexts.

EXAMPLE 15

Estimating Beta for Westport Innovations Stock

Westport Innovations Inc. (Westport) is a Canadian company that provides low-emission engine and fuel system technologies utilizing gaseous fuels. Its stock trades on the Toronto Stock Exchange. You are an investor in Westport's stock and want an estimate of its beta. As in the text example, you hypothesize that Westport has an average level of market risk and that its required return in excess of the risk-free rate is the same as the market's required excess return. One regression that summarizes these statements is

$$(R - R_F) = \alpha + \beta(R_M - R_F) + \varepsilon \quad \textbf{(11)}$$

where R_F is the periodic risk-free rate of return (known at the beginning of the period), R_M is the periodic return on the market, R is the periodic return to the stock of the company, and β measures the sensitivity of the required excess return to the excess return to market. Estimating this equation with linear regression provides an estimate of β, $\hat{\beta}$, which tells us the size of the required return premium for the security, given expectations about market returns.[30]

Suppose we want to test the null hypothesis, H_0, that $\beta = 1$ for Westport stock to see whether Westport stock has the same required return premium as the market as a whole. We need data on returns to Westport stock, a risk-free interest rate, and the returns to the market index. For this example, we use data

30 Beta (β) is typically estimated using 60 months of historical data, but the data-sample length sometimes varies. Although monthly data is typically used, some financial analysts estimate β using daily data. The expected excess return for Westport stock above the risk-free rate $(R - R_F)$ is $\beta(R_M - R_F)$, given a particular excess return to the market above the risk-free rate $(R_M - R_F)$. This result holds because we regress $(R - R_F)$ against $(R_M - R_F)$. For example, if a stock's beta is 1.5, its expected excess return is 1.5 times that of the market portfolio.

from January 2009 through December 2013 (n = 60). The return to Westport stock is R. The monthly return to 1 month Canadian Treasury bills is R_F. The return to the S&P/TSX Composite Index is R_M.[31] This index is the primary broad measure of the Canadian equity market. We are estimating two parameters, so the number of degrees of freedom is $n - 2 = 60 - 2 = 58$. Table 9 shows the results from the regression $(R - R_F) = \alpha + \beta\,(R_M - R_F) + \varepsilon$.

Table 9 Estimating Beta for Westport Stock

Regression Statistics	
Multiple R	0.3429
R-squared	0.1176
Standard error of estimate	0.1488
Observations	60

	Coefficients	Standard Error	*t*-Statistic
Alpha	0.0267	0.0273	0.9793
Beta	1.0788	0.3880	2.7800

Sources: Bank of Canada and ca.finance.yahoo.com.

1. Test the null hypothesis, H_0, that β for Westport equals 1 (β = 1) against the alternative hypothesis that β does not equal 1 (β ≠ 1) using the confidence interval approach.
2. Test the above hypothesis using a t-test.
3. How much of Westport stock's excess return variation can be attributed to company-specific risk?

Solution to 1:

The estimated $\hat{\beta}$ from the regression is 1.0788. The estimated standard error for that coefficient in the regression, $s_{\hat{\beta}}$ is 0.3880. The regression equation has 58 degrees of freedom (60 − 2), so the critical value for the test statistic is approximately $t_c = 2.00$ at the 0.05 significance level. Therefore, the 95 percent confidence interval for the data for any particular hypothesized value of β is shown by the range

$$\hat{\beta} \pm t_c s_{\hat{\beta}}$$

$$1.0788 \pm 2.00(0.3880)$$

$$0.3028 \text{ to } 1.8548$$

In this case, the hypothesized parameter value is β = 1, and the value 1 falls inside this confidence interval, so we cannot reject the hypothesis at the 0.05 significance level. This means that we cannot reject the hypothesis that Westport stock has the same systematic risk as the market as a whole.

31 Data on Westport stock returns and S&P/TSX Composite Index returns came from ca.finance.yahoo.com. Data on Canadian T-bill returns came from the Bank of Canada.

Solution to 2:

The *t*-statistic for the Westport beta hypothesized parameter can be computed using Equation 10:

$$t = \frac{\hat{\beta} - \beta}{s_{\hat{\beta}}} = \frac{1.0788 - 1.0}{0.3880} = 0.2031$$

This *t*-statistic is less than the critical *t*-value of 2.00. Therefore, neither approach allows us to reject the null hypothesis. Note that the *t*-statistic associated with $\hat{\beta}$ in the regression results in Table 9 is 2.7800. Given the significance level we are using, we cannot reject the null hypothesis that $\beta = 1$, but we can reject the hypothesis that $\beta = 0$.[32]

Solution to 3:

The R^2 in this regression is only 0.1176. This result suggests that only about 12 percent of the total variation in the excess return to Westport stock (the return to Westport above the risk-free rate) can be explained by excess return to the market portfolio. The remaining 88 percent of Westport stock's excess return variation is the nonsystematic component, which can be attributed to company-specific risk.

In the next example, we show a regression hypothesis test with a one-sided alternative.

EXAMPLE 16

Explaining Company Value Based on Returns to Invested Capital

Some financial analysts have argued that one good way to measure a company's ability to create wealth is to compare the company's return on invested capital (ROIC) to its weighted-average cost of capital (WACC). If a company has an ROIC greater than its cost of capital, the company is creating wealth; if its ROIC is less than its cost of capital, it is destroying wealth.[33]

Enterprise value (EV) is a market-price-based measure of company value defined as the market value of equity and debt minus the value of cash and investments. Invested capital (IC) is an accounting measure of company value defined as the sum of the book values of equity and debt. Higher ratios of EV to IC should reflect greater success at wealth creation in general. Mauboussin (1996) argued that the spread between ROIC and WACC helps explain the ratio of EV to IC. Using data on companies in the food-processing industry, we can test the relationship between EV/IC and (ROIC–WACC) using the regression model given in Equation 12.

$$\text{EV}_i/\text{IC}_i = b_0 + b_1(\text{ROIC}_i - \text{WACC}_i) + \varepsilon_i \quad \textbf{(12)}$$

32 The *t*-statistics for a coefficient automatically reported by statistical software programs assume that the null hypothesis states that the coefficient is equal to 0. If you have a different null hypothesis, as we do in this example ($\beta = 1$), then you must either construct the correct test statistic yourself or instruct the program to compute it.

33 See, for example, Sonkin and Johnson (2017).

where the subscript i is an index to identify the company. Our null hypothesis is H_0: $b_1 \leq 0$, and we specify a significance level of 0.05. If we reject the null hypothesis, we have evidence of a statistically significant relationship between EV/IC and (ROIC–WACC). Equation 12 is estimated using data from nine food-processing companies.[34] The results of this regression are displayed in Table 10 and Figure 11.

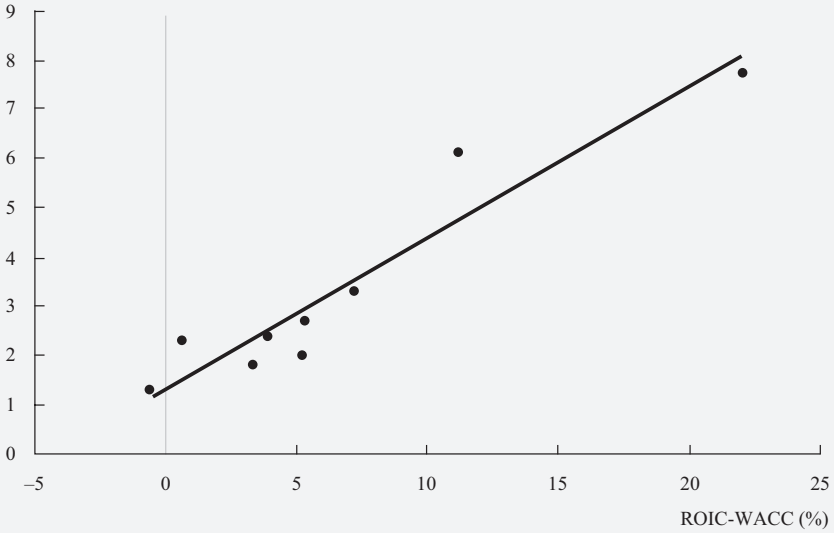
Table 10 Explaining Enterprise Value/Invested Capital by the ROIC–WACC Spread

Regression Statistics	
Multiple *R*	0.9469
R-squared	0.8966
Standard error of estimate	0.7422
Observations	9

	Coefficients	Standard Error	*t*-Statistic
Intercept	1.3478	0.3511	3.8391
Spread	30.0169	3.8519	7.7928

Source: Nelson, Moskow, Lee, and Valentine (2003).

Figure 11 Fitted Regression Line Explaining Enterprise Value/Invested Capital Using ROIC–WACC Spread for the Food Industry



Source: Nelson et al. (2003).

34 Our data come from Nelson, Moskow, Lee, and Valentine (2003) and relate to 2001. Many sell-side analysts use this type of regression. It is one of the most frequently used cross-sectional regressions in published analyst reports.

We reject the null hypothesis based on the t-statistic of approximately 7.79 on estimated slope coefficient. There is a strong positive relationship between the return spread (ROIC–WACC) and the ratio of EV to IC in our sample of companies. Figure 11 illustrates the strong positive relationship. The R^2 of 0.8966 indicates that the return spread explains about 90 percent of the variation in the ratio of EV to IC among the food-processing companies in the sample in 2001. The coefficient on the return spread of 30.0169 implies that the predicted increase in EV/IC is 0.01(30.0169) = 0.3002 or about 30 percent for a 1-percentage-point increase in the return spread, for our sample of companies.

In the final example of this section, the null hypothesis for a t-test of the slope coefficient is that the value of slope equals 1 in contrast to the null hypothesis that it equals 0 as in prior examples.

EXAMPLE 17

Testing whether Inflation Forecasts Are Unbiased

Example 12 introduced the concept of testing for bias in forecasts. That example showed that if a forecast is unbiased, its expected error is 0. We can examine whether a time-series of forecasts for a particular economic variable is unbiased by comparing the forecast at each date with the actual value of the economic variable announced after the forecast. If the forecasts are unbiased, then, by definition, the average realized forecast error should be close to 0. In that case, the value of b_0 (the intercept) should be 0 and the value of b_1 (the slope) should be 1, as discussed in Example 12.

Refer once again to Figure 9, which shows the mean forecast made by professional economic forecasters in the first quarter of a year for the percentage change in euro area HICP during that year and the actual percentage change from 1999 through 2013 (n = 14). To test whether the forecasts are unbiased, we must estimate the regression shown in Example 12. We report the results of this regression in Table 11. The equation to be estimated is

$$\text{Actual percentage change in HICP}_t = b_0 + b_1(\text{Predicted change}_t) + \varepsilon_t$$

This regression estimates two parameters (the intercept and the slope); therefore, the regression has $n - 2 = 14 - 2 = 12$ degrees of freedom.

Table 11 Testing whether Forecasts of Euro Area HICP Are Unbiased (Dependent Variable: CPI Change Expressed in Percent)

Regression Statistics	
Multiple R	0.9006
R-squared	0.8111
Standard error of estimate	0.3165
Observations	14

(continued)

Table 11 (Continued)

	Coefficients	Standard Error	*t*-Statistic
Intercept	–0.7006	0.3723	–1.8820
Forecast (slope)	1.5538	0.2079	7.4722

Source: European Central Bank.

We can now test two null hypotheses about the parameters in this regression. Our first null hypothesis is that the intercept in this regression is 0 (H_0: $b_0 = 0$). The alternative hypothesis is that the intercept does not equal 0 (H_a: $b_0 \neq 0$). Our second null hypothesis is that the slope coefficient in this regression is 1 (H_0: $b_1 = 1$). The alternative hypothesis is that the slope coefficient does not equal 1 (H_a: $b_1 \neq 1$).

To test the hypotheses about b_0 and b_1, we must first decide on a critical value based on a particular significance level and then construct the confidence intervals for each parameter. If we choose the 0.05 significance level, with 12 degrees of freedom, the critical value, t_c, is approximately 2.18. The estimated value of the parameter $\hat{b}_0$ is –0.7006, and the estimated value of the standard error for $\hat{b}_0\left(s_{\hat{b}_0}\right)$ is 0.3723. Let B_0 stand for any particular hypothesized value. Therefore, under the null hypothesis that $b_0 = B_0$, a 95 percent confidence interval for b_0 is

$$\hat{b}_0 \pm t_c s_{\hat{b}_0}$$
$$-0.7006 \pm 2.18(0.3723)$$
$$-1.5122 \text{ to } 0.1110$$

In this case, B_0 is 0. The value of 0 falls within this confidence interval, so we cannot reject the first null hypothesis that $b_0 = 0$. We will explain how to interpret this result shortly.

Our second null hypothesis is based on the same sample as our first null hypothesis. Therefore, the critical value for testing that hypothesis is the same as the critical value for testing the first hypothesis ($t_c = 2.18$). The estimated value of the parameter $\hat{b}_1$ is 1.5538, and the estimated value of the standard error for $\hat{b}_1$, $s_{\hat{b}_1}$, is 0.2079. Therefore, the 95 percent confidence interval for any particular hypothesized value of b_1 can be constructed as follows:

$$\hat{b}_1 \pm t_c s_{\hat{b}_1}$$
$$1.5538 \pm 2.18(0.2079)$$
$$1.1006 \text{ to } 2.0070$$

In this case, our hypothesized value of b_1 is 1. The value 1 falls outside this confidence interval, so we can reject the null hypothesis that $b_1 = 1$ at the 0.05 significance level. Because we did reject one of the two null hypotheses ($b_0 = 0$, $b_1 = 1$) about the parameters in this model, we can reject the hypothesis that the forecasts of HICP change were unbiased.[35]

35 Jointly testing the hypothesis $b_0 = 0$ and $b_1 = 1$ would require us to take into account the covariance of $\hat{b}_0$ and $\hat{b}_1$. For information on testing joint hypotheses of this type, see Greene (2018).

As an analyst, you often will need forecasts of **economic growth** to help you make recommendations about asset allocation, expected returns, and other investment decisions. The hypothesis tests just conducted suggest that you can reject the hypothesis that the HICP predictions in the Survey of Professional Forecasters are unbiased. If you need an unbiased forecast of future percentage change in HICP for your asset-allocation decision, you might not want to use these forecasts.

In view of the above concern, we further explored the inflation forecasts. Figure 9 suggests that the bottommost point in the plot of actual versus realized inflations is an outlier. This point corresponds to the year 2009 when macroeconomic volatility was exceptionally high due to the financial crisis. A study of forecasts in the European Central Bank Survey of Professional Forecasters by Genre, Kenny, Meyler, and Timmerman (2013) finds that the performance of inflation forecasts is lowered when the financial crisis period is included. We re-estimated the regression equation after excluding 2009. The new equation is Actual change = −0.2513 + 1.3209(Predicted change). Under the null hypothesis for the intercept that $b_0 = 0$, a 95 percent confidence interval for b_0 is −1.2116 to 0.7090. The value of 0 falls within this confidence interval, so we cannot reject the first null hypothesis that $b_0 = 0$. Under the null hypothesis for the slope that $b_1 = 1$, a 95 percent confidence interval for b_1 is 0.7984 to 1.8434. The value of 1 falls within this confidence interval, so we cannot reject the second null hypothesis that $b_1 = 1$. These hypothesis tests suggest that you cannot reject the hypothesis that the HICP predictions in the Survey of Professional Forecasters are unbiased.

3.6 Analysis of Variance in a Regression with One Independent Variable

Analysis of variance (ANOVA) is a statistical procedure for dividing the total variability of a variable into components that can be attributed to different sources.[36] In regression analysis, we use ANOVA to determine the usefulness of the independent variable or variables in explaining variation in the dependent variable. An important statistical test conducted in analysis of variance is the *F*-test. The *F*-statistic tests whether all the slope coefficients in a linear regression are equal to 0. In a regression with one independent variable, this is a test of the null hypothesis H_0: $b_1 = 0$ against the alternative hypothesis H_a: $b_1 \neq 0$.

To correctly determine the test statistic for the null hypothesis that the slope coefficient equals 0, we need to know the following:

- the total number of observations (n);
- the total number of parameters to be estimated (in a one-independent-variable regression, this number is two: the intercept and the slope coefficient);

36 In this reading, we focus on regression applications of ANOVA, the most common context in which financial analysts will encounter this tool. In this context, ANOVA is used to test whether all the regression slope coefficients are equal to 0. Analysts also use ANOVA to test a hypothesis that the means of two or more populations are equal. See Daniel and Terrell (1995) for details.

- the sum of squared errors or residuals, $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$, abbreviated SSE. This value is also known as the residual sum of squares; and
- the regression sum of squares, $\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$, abbreviated RSS. This value is the amount of total variation in Y that is explained in the regression equation. Total variation (TSS) is the sum of SSE and RSS.

The F-test for determining whether the slope coefficient equals 0 is based on an F-statistic, constructed using these four values. The F-statistic measures how well the regression equation explains the variation in the dependent variable. The F-statistic is the ratio of the average regression sum of squares to the average sum of the squared errors. The average regression sum of squares is computed by dividing the regression sum of squares by the number of slope parameters estimated (in this case, one). The average sum of squared errors is computed by dividing the sum of squared errors by the number of observations, n, minus the total number of parameters estimated (in this case, two: the intercept and the slope). These two divisors are the degrees of freedom for an F-test. If there are n observations, the F-test for the null hypothesis that the slope coefficient is equal to 0 is here denoted $F_{(\#\text{ slope parameters}),(n\ -\ \#\text{ parameters})} = F_{1,n-2}$, and the test has 1 and $n - 2$ degrees of freedom.

Suppose, for example, that the independent variable in a regression model explains none of the variation in the dependent variable. Then the predicted value for the regression model, $\hat{Y}_i$, is the average value of the dependent variable $\overline{Y}$. In this case, the regression sum of squares $\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$ is 0. Therefore, the F-statistic is 0. If the independent variable explains little of the variation in the dependent variable, the value of the F-statistic will be very small.

The formula for the F-statistic in a regression with one independent variable is

$$F = \frac{\text{RSS}/1}{\text{SSE}/(n-2)} = \frac{\text{Mean regression sum of squares}}{\text{Mean squared error}} \quad \textbf{(13)}$$

If the regression model does a good job of explaining variation in the dependent variable, then this ratio should be high. The explained regression sum of squares per estimated parameter will be high relative to the unexplained variation for each degree of freedom. Critical values for this F-statistic are given in Appendix D at the end of this volume.

Even though the F-statistic is commonly computed by regression software packages, analysts typically do not use ANOVA and F-tests in regressions with just one independent variable. Why not? In such regressions, the F-statistic is the square of the t-statistic for the slope coefficient. Therefore, the F-test duplicates the t-test for the significance of the slope coefficient. This relation is not true for regressions with two or more slope coefficients. Nevertheless, the one-slope coefficient case gives a foundation for understanding the multiple-slope coefficient cases.

Often, mutual fund performance is evaluated based on whether the fund has positive alpha—significantly positive excess risk-adjusted returns.[37] One commonly used method of risk adjustment is based on the capital asset pricing model. Consider the regression

$$(R_i - R_F) = \alpha_i + \beta_i(R_M - R_F) + \varepsilon_i \quad \textbf{(14)}$$

37 Note that the Greek letter alpha, α, is traditionally used to represent the intercept in Equation 14 and should not be confused with another traditional usage of α to represent a significance level.

where R_F is the periodic risk-free rate of return (known at the beginning of the period), R_M is the periodic return on the market, R_i is the periodic return to Mutual Fund i, and β_i is the fund's beta. A fund has zero risk-adjusted excess return if $\alpha_i = 0$. If $\alpha_i = 0$, then $(R_i - R_F) = \beta_i(R_M - R_F) + \varepsilon_i$ and taking expectations, $E(R_i) = R_F + \beta_i(R_M - R_F)$, implying that β_i completely explains the fund's mean excess returns. If, for example, $\alpha_i > 0$, the fund is earning higher returns than expected given its beta.

In summary, to test whether a fund has a positive alpha, we must test the null hypothesis that the fund has no risk-adjusted excess returns (H_0: $\alpha = 0$) against the alternative hypothesis of nonzero risk-adjusted returns (H_a: $\alpha \neq 0$).

EXAMPLE 18

Performance Evaluation: The Dreyfus Appreciation Fund

Table 12 presents results evaluating the excess return to the Dreyfus Appreciation Fund from January 2009 through December 2013. Note that the estimated beta in this regression, $\hat{\beta}_i$, is 0.8660. The Dreyfus Appreciation Fund was estimated to be almost 0.9 times as risky as the market as a whole.

Table 12 Performance Evaluation of Dreyfus Appreciation Fund, January 2009 to December 2013

Regression Statistics	
Multiple *R*	0.9633
R-squared	0.9279
Standard error of estimate	0.0111
Observations	60

ANOVA	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (MSS)	*F*
Regression	1	0.0925	0.0925	746.09
Residual	58	0.0072	0.0001	
Total	59	0.0997		

	Coefficients	Standard Error	*t*-Statistic
Alpha	–0.0012	0.0015	–0.8050
Beta	0.8660	0.0317	27.3147

Sources: Center for Research in Security Prices, University of Chicago; S&P Dow Jones Indices; and the Federal Reserve.

1. Test whether the fund had a significant excess return beyond the return associated with the market risk of the fund.
2. Based on the *t*-test, discuss whether the beta of the fund is likely to be zero.
3. Use Equation 13 to compute the *F*-statistic. Based on the *F*-test, determine whether the beta of the fund is likely to be zero.

Solution to 1:

The estimated alpha ($\hat{\alpha}$) in this regression is negative (–0.0012). The absolute value of the coefficient is less than the size of the standard error for that coefficient (0.0015), so the *t*-statistic for the coefficient is only –0.8050. Therefore, we cannot reject the null hypothesis ($\alpha = 0$) that the fund did not have a significant excess return beyond the return associated with the market risk of the fund. This result means that the returns to the fund were explained by the market risk of the fund and there was no additional statistical significance to the excess returns to the fund during this period.[38]

Solution to 2:

Because the *t*-statistic for the slope coefficient in this regression is 27.3147, the *p*-value for that coefficient is less than 0.0001 and is approximately zero. Therefore, the probability that the true value of this coefficient is actually 0 is microscopic.

Solution to 3:

The ANOVA portion of Table 12 provides the data we need to compute the *F*-statistic. In this case:

- the total number of observations (n) is 60;
- the total number of parameters to be estimated is 2 (intercept and slope);
- the sum of squared errors or residuals, SSE, is 0.0072; and
- the regression sum of squares, RSS, is 0.0925.

Therefore, the *F*-statistic to test whether the slope coefficient is equal to 0 is

$$\frac{0.0925/1}{0.0072/(60-2)} = 745.14$$

(The slight difference from the *F*-statistic in Table 12 is due to rounding.) The ANOVA output would show that the *p*-value for this *F*-statistic is less than 0.0001 and is exactly the same as the *p*-value for the *t*-statistic for the slope coefficient. Therefore, the *F*-test tells us nothing more than we already knew from the *t*-test. Note also that the *F*-statistic (746.09) is the square of the *t*-statistic (27.3147).

3.7 Prediction Intervals

Financial analysts often want to use regression results to make predictions about a dependent variable. For example, we might ask, "How fast will the sales of XYZ Corporation grow this year if real GDP grows by 4 percent?" But we are not merely interested in making these forecasts; we also want to know how certain we should be about the forecasts' results. For example, if we predicted that sales for XYZ Corporation would grow by 6 percent this year, our prediction would mean more if we were 95 percent confident that sales growth would fall in the interval from 5 percent to 7 percent, rather than only 25 percent confident that this outcome would occur. Therefore, we need to understand how to compute confidence intervals around regression forecasts.

38 This example introduces a well-known investment use of regression involving the capital asset pricing model. Researchers, however, recognize qualifications to the interpretation of alpha from a linear regression. The systematic risk of a managed portfolio is controlled by the portfolio manager. If, as a consequence, portfolio beta is correlated with the return on the market (as could result from market timing), inferences on alpha based on least-squares beta, as here, can be mistaken. This advanced subject is discussed in Dybvig and Ross (1985a) and (1985b).

We must take into account two sources of uncertainty when using the regression model $Y_i = b_0 + b_1X_i + \varepsilon_i$, $i = 1, \ldots, n$ and the estimated parameters, $\hat{b}_0$ and $\hat{b}_1$, to make a prediction. First, the error term itself contains uncertainty. The standard deviation of the error term, σ_ε, can be estimated from the standard error of estimate for the regression equation. A second source of uncertainty in making predictions about Y, however, comes from uncertainty in the estimated parameters $\hat{b}_0$ and $\hat{b}_1$.

If we knew the true values of the regression parameters, b_0 and b_1, then the variance of our prediction of Y, given any particular predicted (or assumed) value of X, would simply be s^2, the squared standard error of estimate. The variance would be s^2 because the prediction, $\hat{Y}$, would come from the equation $\hat{Y} = b_0 + b_1X$ and $\left(Y - \hat{Y}\right) = \varepsilon$.

Because we must estimate the regression parameters $\hat{b}_0$ and $\hat{b}_1$ however, our prediction of Y, $\hat{Y}$, given any particular predicted value of X, is actually $\hat{Y} = \hat{b}_0 + \hat{b}_1X$. The estimated variance of the prediction error, s_f^2 of Y, given X, is

$$s_f^2 = s^2\left[1 + \frac{1}{n} + \frac{\left(X - \bar{X}\right)^2}{(n-1)s_x^2}\right] \tag{15}$$

This estimated variance depends on:

- the squared standard error of estimate, s^2;
- the number of observations, n;
- the value of the independent variable, X, used to predict the dependent variable;
- the estimated mean, $\bar{X}$; and
- variance, s_x^2 of the independent variable.

Once we have this estimate of the variance of the prediction error, determining a prediction interval around the prediction is very similar to estimating a confidence interval around an estimated parameter, as shown earlier in this reading. We need to take the following four steps to determine the prediction interval for the prediction:

1. Make the prediction.
2. Compute the variance of the prediction error using Equation 15.
3. Choose a significance level, α, for the forecast. For example, the 0.05 level, given the degrees of freedom in the regression, determines the critical value for the forecast interval, t_c.
4. Compute the $(1 - \alpha)$ percent prediction interval for the prediction, namely $\hat{Y} \pm t_c s_f$.

EXAMPLE 19

Predicting the Ratio of Enterprise Value to Invested Capital

We continue with the example of explaining the ratio of enterprise value to invested capital among food-processing companies by the spread between the return to invested capital and the weighted-average cost of capital (ROIC–WACC). In Example 15, we estimated the regression given in Table 10.

Table 10 Explaining Enterprise Value/Invested Capital by the ROIC-WACC Spread (repeated)

Regression Statistics	
Multiple *R*	0.9469
R-squared	0.8966
Standard error of estimate	0.7422
Observations	9

	Coefficients	Standard Error	*t*-Statistic
Intercept	1.3478	0.3511	3.8391
Spread	30.0169	3.8519	7.7928

Source: Nelson, Moskow, Lee, and Valentine (2003).

You are interested in predicting the ratio of enterprise value to invested capital for a company if the return spread between ROIC and WACC is 10 percentage points. What is the 95 percent prediction interval for the ratio of enterprise value to invested capital for that company?

Using the data provided in Table 10, take the following steps:

1. Make the prediction: Expected EV/IC = 1.3478 + 30.0169(0.10) = 4.3495. This regression suggests that if the return spread between ROIC and WACC (X_i) is 10 percent, the EV/IC ratio will be 4.3495.
2. Compute the variance of the prediction error. To compute the variance of the forecast error, we must know:
   - the standard error of the estimate of the equation, s = 0.7422 (as shown in Table 10);
   - the mean return spread, $\bar{X}$ = 0.0647 (this computation is not shown in the table); and
   - the variance of the mean return spread in the sample, s_x^2 = 0.004641 (this computation is not shown in the table).

Using these data, you can compute the variance of the forecast error (s_f^2) for predicting EV/IC for a company with a 10 percent spread between ROIC and WACC.

$$s_f^2 = 0.7422^2\left[1 + \frac{1}{9} + \frac{(0.10 - 0.0647)^2}{(9-1)0.004641}\right]$$
$$= 0.630556$$

In this example, the variance of the forecast error is 0.630556, and the standard deviation of the forecast error is $s_f = (0.630556)^{1/2} = 0.7941$.

3. Determine the critical value of the *t*-statistic. Given a 95 percent confidence interval and 9 – 2 = 7 degrees of freedom, the critical value of the *t*-statistic, t_c, is 2.365 using the tables in the back of this volume.
4. Compute the prediction interval. The 95 percent confidence interval for EV/IC extends from 4.3495 – 2.365(0.7941) to 4.3495 + 2.365(0.7941), or 2.4715 to 6.2275.

In summary, if the spread between the ROIC and the WACC is 10 percent, the 95 percent prediction interval for EV/IC will extend from 2.4715 to 6.2275. The small sample size is reflected in the relatively large prediction interval.

3.8 Limitations of Regression Analysis

Although this reading has shown many of the uses of regression models for financial analysis, regression models do have limitations. First, regression relations can change over time, just as correlations can. This fact is known as the issue of **parameter instability**, and its existence should not be surprising as the economic, tax, regulatory, political, and institutional contexts in which financial markets operate change. Whether considering cross-sectional or time-series regression, the analyst will probably face this issue. As one example, cross-sectional regression relationships between stock characteristics may differ between growth-led and value-led markets. As a second example, the time-series regression estimating the beta often yields significantly different estimated betas depending on the time period selected. In both cross-sectional and time-series contexts, the most common problem is sampling from more than one population, with the challenge of identifying when doing so is an issue.

A second limitation to the use of regression results specific to investment contexts is that public knowledge of regression relationships may negate their future usefulness. Suppose, for example, an analyst discovers that stocks with a certain characteristic have had historically very high returns. If other analysts discover and act upon this relationship, then the prices of stocks with that characteristic will be bid up. The knowledge of the relationship may result in the relation no longer holding in the future.

Finally, if the regression assumptions listed in Section 3.2 are violated, hypothesis tests and predictions based on linear regression will not be valid. Although there are tests for violations of regression assumptions, often uncertainty exists as to whether an assumption has been violated. This limitation will be discussed in detail in the reading on multiple regression.

SUMMARY

- A scatter plot shows graphically the relationship between two variables. If the points on the scatter plot cluster together in a straight line, the two variables have a strong linear relation.
- The sample correlation coefficient for two variables X and Y is $r = \frac{\text{Cov}(X,Y)}{s_x s_y}$.
- If two variables have a very strong linear relation, then the absolute value of their correlation will be close to 1. If two variables have a weak linear relation, then the absolute value of their correlation will be close to 0.
- The squared value of the correlation coefficient for two variables quantifies the percentage of the variance of one variable that is explained by the other. If the correlation coefficient is positive, the two variables are directly related; if the correlation coefficient is negative, the two variables are inversely related.
- If we have n observations for two variables, we can test whether the population correlation between the two variables is equal to 0 by using a t-test. This test statistic has a t-distribution with $n - 2$ degrees of freedom if the null hypothesis of 0 correlation is true.

- Even one outlier can greatly affect the correlation between two variables. Analysts should examine a scatter plot for the variables to determine whether outliers might affect a particular correlation.
- Correlations can be spurious in the sense of misleadingly pointing toward associations between variables.
- The dependent variable in a linear regression is the variable that the regression model tries to explain. The independent variables are the variables that a regression model uses to explain the dependent variable.
- If there is one independent variable in a linear regression and there are n observations on the dependent and independent variables, the regression model is $Y_i = b_0 + b_1X_i + \varepsilon_i$, $i = 1, \ldots, n$, where Y_i is the dependent variable, X_i is the independent variable, and ε_i is the error term. In this model, the coefficient b_0 is the intercept. The intercept is the predicted value of the dependent variable when the independent variable has a value of zero. In this model, the coefficient b_1 is the slope of the regression line. If the value of the independent variable increases by one unit, then the model predicts that the value of the dependent variable will increase by b_1 units.
- The assumptions of the classic normal linear regression model are the following:
  - A linear relation exists between the dependent variable and the independent variable.
  - The independent variable is not random.
  - The expected value of the error term is 0.
  - The variance of the error term is the same for all observations (homoskedasticity).
  - The error term is uncorrelated across observations.
  - The error term is normally distributed.
- The estimated parameters in a linear regression model minimize the sum of the squared regression residuals.
- The standard error of estimate measures how well the regression model fits the data. If the SEE is small, the model fits well.
- The coefficient of determination measures the fraction of the total variation in the dependent variable that is explained by the independent variable. In a linear regression with one independent variable, the simplest way to compute the coefficient of determination is to square the correlation of the dependent and independent variables.
- To calculate a confidence interval for an estimated regression coefficient, we must know the standard error of the estimated coefficient and the critical value for the t-distribution at the chosen level of significance, t_c.
- To test whether the population value of a regression coefficient, b_1, is equal to a particular hypothesized value, B_1, we must know the estimated coefficient, $\hat{b}_1$, the standard error of the estimated coefficient, $s_{\hat{b}_1}$, and the critical value for the t-distribution at the chosen level of significance, t_c. The test statistic for this hypothesis is $\left(\hat{b}_1 - B_1\right)\Big/ s_{\hat{b}_1}$. If the absolute value of this statistic is greater than t_c, then we reject the null hypothesis that $b_1 = B_1$.

- In the regression model $Y_i = b_0 + b_1X_i + \varepsilon_i$, if we know the estimated parameters, $\hat{b}_0$ and $\hat{b}_1$, for any value of the independent variable, X, then the predicted value of the dependent variable Y is $\hat{Y} = \hat{b}_0 + \hat{b}_1X$.
- The prediction interval for a regression equation for a particular predicted value of the dependent variable is $\hat{Y} \pm t_c s_f$ where s_f is the square root of the estimated variance of the prediction error and t_c is the critical level for the t-statistic at the chosen significance level. This computation specifies a $(1 - \alpha)$ percent confidence interval. For example, if $\alpha = 0.05$, then this computation yields a 95 percent confidence interval.

REFERENCES

Buetow, Gerald W., Jr, Robert R. Johnson, and David E. Runkle. 2000. "The Inconsistency of Returns-Based Style Analysis." *Journal of Portfolio Management* 26 (3): 61–77.

Campbell, John Y., Karine Serfaty-de Medeiros, and Luis M. Viceira. 2010. "Global Currency Hedging." *Journal of Finance* 65 (1): 87–121.

Chan, Louis K. C., Stephen G. Dimmock, and Josef Lakonishok. 2009. "Benchmarking Money Manager Performance: Issues and Evidence." *Review of Financial Studies* 22 (11): 4553–99.

Daniel, Wayne W., and James C. Terrell. 1995. *Business Statistics for Management and Economics*. 7th ed. Boston: Houghton-Mifflin.

Dybvig, Philip H., and Stephen A. Ross. 1985a. "Differential Information and Performance Measurement Using a Security Market Line." *Journal of Finance* 40 (2): 383–99.

Dybvig, Philip H., and Stephen A. Ross. 1985b. "The Analytics of Performance Measurement Using a Security Market Line." *Journal of Finance* 40 (2): 401–16.

Genre, Veronique, Geoff Kenny, Aidan Meyler, and Allan Timmermann. 2013. "Combining expert forecasts: Can anything beat the simple average?" *International Journal of Forecasting* 29 (1): 108–21.

Greene, William H. 2018. Economic Analysis. 8th ed. Upper Saddle River, NJ: Prentice-Hall.

Keane, Michael P., and David E. Runkle. 1990. "Testing the Rationality of Price Forecasts: New Evidence from Panel Data." *American Economic Review* 80 (4): 714–35.

Nelson, David C., Robert B. Moskow, Tiffany Lee, and Gregg Valentine. 2003. Food Investor's Handbook. New York: Credit Suisse First Boston.

Sharpe, William F. 1992. "Asset Allocation: Management Style and Performance Measurement." *Journal of Portfolio Management* 18 (2): 7–19.

Sonkin, Paul D., and Paul Johnson. 2017. Pitch the Perfect Investment. New York: Wiley.

PRACTICE PROBLEMS

1 The following table shows the sample correlations between the monthly returns for four different mutual funds and the S&P 500. The correlations are based on 36 monthly observations. The funds are as follows:

Fund 1	Large-cap fund
Fund 2	Mid-cap fund
Fund 3	Large-cap value fund
Fund 4	Emerging markets fund
S&P 500	US domestic stock index

	Fund 1	Fund 2	Fund 3	Fund 4	S&P 500
Fund 1	1				
Fund 2	0.9231	1			
Fund 3	0.4771	0.4156	1		
Fund 4	0.7111	0.7238	0.3102	1	
S&P 500	0.8277	0.8223	0.5791	0.7515	1

Test the null hypothesis that each of these correlations, individually, is equal to zero against the alternative hypothesis that it is not equal to zero. Use a 5 percent significance level.

2 Julie Moon is an energy analyst examining electricity, oil, and natural gas consumption in different regions over different seasons. She ran a regression explaining the variation in energy consumption as a function of temperature. The total variation of the dependent variable was 140.58, the explained variation was 60.16, and the unexplained variation was 80.42. She had 60 monthly observations.

A Compute the coefficient of determination.

B What was the sample correlation between energy consumption and temperature?

C Compute the standard error of the estimate of Moon's regression model.

D Compute the sample standard deviation of monthly energy consumption.

3 You are examining the results of a regression estimation that attempts to explain the unit sales growth of a business you are researching. The analysis of variance output for the regression is given in the table below. The regression was based on five observations ($n = 5$).

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	1	88.0	88.0	36.667	0.00904
Residual	3	7.2	2.4		
Total	4	95.2			

A How many independent variables are in the regression to which the ANOVA refers?

B Define Total SS.

C Calculate the sample variance of the dependent variable using information in the above table.

D Define Regression SS and explain how its value of 88 is obtained in terms of other quantities reported in the above table.

E What hypothesis does the *F*-statistic test?

F Explain how the value of the *F*-statistic of 36.667 is obtained in terms of other quantities reported in the above table.

G Is the *F*-test significant at the 5 percent significance level?

4 An economist collected the monthly returns for KDL's portfolio and a diversified stock index. The data collected are shown below:

Month	Portfolio Return (%)	Index Return (%)
1	1.11	–0.59
2	72.10	64.90
3	5.12	4.81
4	1.01	1.68
5	–1.72	–4.97
6	4.06	–2.06

The economist calculated the correlation between the two returns and found it to be 0.996. The regression results with the KDL return as the dependent variable and the index return as the independent variable are given as follows:

Regression Statistics	
Multiple *R*	0.996
R-squared	0.992
Standard error	2.861
Observations	6

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	1	4101.62	4101.62	500.79	0
Residual	4	32.76	8.19		
Total	5	4134.38			

	Coefficients	Standard Error	*t*-Statistic	*p*-Value
Intercept	2.252	1.274	1.768	0.1518
Slope	1.069	0.0477	22.379	0

When reviewing the results, Andrea Fusilier suspected that they were unreliable. She found that the returns for Month 2 should have been 7.21 percent and 6.49 percent, instead of the large values shown in the first table. Correcting these values resulted in a revised correlation of 0.824 and the revised regression results shown as follows:

Regression Statistics	
Multiple *R*	0.824
R-squared	0.678
Standard error	2.062
Observations	6

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	1	35.89	35.89	8.44	0.044
Residual	4	17.01	4.25		
Total	5	52.91			

	Coefficients	Standard Error	*t*-Statistic	*p*-Value
Intercept	2.242	0.863	2.597	0.060
Slope	0.623	0.214	2.905	0.044

Explain how the bad data affected the results.

The following information relates to Questions 5–10

Kenneth McCoin, CFA, is a fairly tough interviewer. Last year, he handed each job applicant a sheet of paper with the information in the following table, and he then asked several questions about regression analysis. Some of McCoin's questions, along with a sample of the answers he received to each, are given below. McCoin told the applicants that the independent variable is the ratio of net income to sales for restaurants with a market cap of more than $100 million and the dependent variable is the ratio of cash flow from operations to sales for those restaurants. Which of the choices provided is the best answer to each of McCoin's questions?

Regression Statistics	
Multiple *R*	0.8623
R-squared	0.7436
Standard error	0.0213
Observations	24

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	1	0.029	0.029000	63.81	0
Residual	22	0.010	0.000455		
Total	23	0.040			

	Coefficients	Standard Error	*t*-Statistic	*p*-Value
Intercept	0.077	0.007	11.328	0
Slope	0.826	0.103	7.988	0

5 What is the value of the coefficient of determination?

A 0.8261.

B 0.7436.

C 0.8623.

6 Suppose that you deleted several of the observations that had small residual values. If you re-estimated the regression equation using this reduced sample, what would likely happen to the standard error of the estimate and the *R*-squared?

	Standard Error of the Estimate	*R*-Squared
A	Decrease	Decrease
B	Decrease	Increase
C	Increase	Decrease

7 What is the correlation between *X* and *Y*?

A –0.7436.

B 0.7436.

C 0.8623.

8 Where did the *F*-value in the ANOVA table come from?

A You look up the *F*-value in a table. The *F* depends on the numerator and denominator degrees of freedom.

B Divide the "Mean Square" for the regression by the "Mean Square" of the residuals.

C The *F*-value is equal to the reciprocal of the *t*-value for the slope coefficient.

9 If the ratio of net income to sales for a restaurant is 5 percent, what is the predicted ratio of cash flow from operations to sales?

A 0.007 + 0.103(5.0) = 0.524.

B 0.077 – 0.826(5.0) = –4.054.

C 0.077 + 0.826(5.0) = 4.207.

10 Is the relationship between the ratio of cash flow to operations and the ratio of net income to sales significant at the 5 percent level?

A No, because the *R*-squared is greater than 0.05.

B No, because the *p*-values of the intercept and slope are less than 0.05.

C Yes, because the *p*-values for *F* and *t* for the slope coefficient are less than 0.05.

The following information relates to Questions 11–16

Howard Golub, CFA, is preparing to write a research report on Stellar Energy Corp. common stock. One of the world's largest companies, Stellar is in the business of refining and marketing oil. As part of his analysis, Golub wants to evaluate the sensitivity of the stock's returns to various economic factors. For example, a client recently asked Golub whether the price of Stellar Energy Corporation stock has tended to rise following increases in retail energy prices. Golub believes the association between the two variables to be negative, but he does not know the strength of the association.

Golub directs his assistant, Jill Batten, to study the relationships between Stellar monthly common stock returns versus the previous month's percent change in the US Consumer Price Index for Energy (CPIENG), and Stellar monthly common stock returns versus the previous month's percent change in the US Producer Price Index for Crude Energy Materials (PPICEM). Golub wants Batten to run both a correlation and a linear regression analysis. In response, Batten compiles the summary statistics shown in Exhibit 1 for the 248 months between January 1980 and August 2000. All of the data are in decimal form, where 0.01 indicates a 1 percent return. Batten also

runs a regression analysis using Stellar monthly returns as the dependent variable and the monthly change in CPIENG as the independent variable. Exhibit 2 displays the results of this regression model.

Exhibit 1 Descriptive Statistics

	Monthly Return Stellar Common Stock	Lagged Monthly Change	
		CPIENG	PPICEM
Mean	0.0123	0.0023	0.0042
Standard Deviation	0.0717	0.0160	0.0534
Covariance, Stellar vs. CPIENG	–0.00017		
Covariance, Stellar vs. PPICEM	–0.00048		
Covariance, CPIENG vs. PPICEM	0.00044		
Correlation, Stellar vs. CPIENG	–0.1452		

Exhibit 2 Regression Analysis with CPIENG

Regression Statistics

Multiple *R*	0.1452
R-squared	0.0211
Standard error of the estimate	0.0710
Observations	248

	Coefficients	Standard Error	*t*-Statistic
Intercept	0.0138	0.0046	3.0275
Slope coefficient	–0.6486	0.2818	–2.3014

11 Batten wants to determine whether the sample correlation between the Stellar and CPIENG variables (–0.1452) is statistically significant. The critical value for the test statistic at the 0.05 level of significance is approximately 1.96. Batten should conclude that the statistical relationship between Stellar and CPIENG is:

A significant, because the calculated test statistic has a lower absolute value than the critical value for the test statistic.

B significant, because the calculated test statistic has a higher absolute value than the critical value for the test statistic.

C not significant, because the calculated test statistic has a higher absolute value than the critical value for the test statistic.

12 Did Batten's regression analyze cross-sectional or time-series data, and what was the expected value of the error term from that regression?

	Data Type	Expected Value of Error Term
A	Time-series	0
B	Time-series	ε_i
C	Cross-sectional	0

13 Based on the regression, which used data in decimal form, if the CPIENG *decreases* by 1.0 percent, what is the expected return on Stellar common stock during the next period?

A 0.0073 (0.73 percent).

B 0.0138 (1.38 percent).

C 0.0203 (2.03 percent).

14 Based on Batten's regression model, the coefficient of determination indicates that:

A Stellar's returns explain 2.11 percent of the variability in CPIENG.

B Stellar's returns explain 14.52 percent of the variability in CPIENG.

C Changes in CPIENG explain 2.11 percent of the variability in Stellar's returns.

15 For Batten's regression model, the standard error of the estimate shows that the standard deviation of:

A the residuals from the regression is 0.0710.

B values estimated from the regression is 0.0710.

C Stellar's observed common stock returns is 0.0710.

16 For the analysis run by Batten, which of the following is an *incorrect* conclusion from the regression output?

A The estimated intercept coefficient from Batten's regression is statistically significant at the 0.05 level.

B In the month after the CPIENG declines, Stellar's common stock is expected to exhibit a positive return.

C Viewed in combination, the slope and intercept coefficients from Batten's regression are not statistically significant at the 0.05 level.

The following information relates to Questions 17–26

Anh Liu is an analyst researching whether a company's debt burden affects investors' decision to short the company's stock. She calculates the short interest ratio (the ratio of short interest to average daily share volume, expressed in days) for 50 companies as of the end of 2016 and compares this ratio with the companies' debt ratio (the ratio of total liabilities to total assets, expressed in decimal form).

Liu provides a number of statistics in Exhibit 1. She also estimates a simple regression to investigate the effect of the debt ratio on a company's short interest ratio. The results of this simple regression, including the analysis of variance (ANOVA), are shown in Exhibit 2.

In addition to estimating a regression equation, Liu graphs the 50 observations using a scatterplot, with the short interest ratio on the vertical axis and the debt ratio on the horizontal axis.

Exhibit 1 Summary Statistics

Statistic	Debt Ratio X_i	Short Interest Ratio Y_i
Sum	19.8550	192.3000
Average	0.3971	3.8460
Sum of squared deviations from the mean	$\sum_{i=1}^{n}(X_i - \bar{X})^2 = 2.2225$	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 412.2042$
Sum of cross-products of deviations from the mean	$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = -9.2430$	

Exhibit 2 Regression of the Short Interest Ratio on the Debt Ratio

ANOVA	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)
Regression	1	38.4404	38.4404
Residual	48	373.7638	7.7867
Total	49	412.2042	

Regression Statistics	
Multiple R	0.3054
R^2	0.0933
Standard error of estimate	2.7905
Observations	50

	Coefficients	Standard Error	*t*-Statistic
Intercept	5.4975	0.8416	6.5322
Debt ratio	–4.1589	1.8718	–2.2219

Liu is considering three interpretations of these results for her report on the relationship between debt ratios and short interest ratios:

Interpretation 1	Companies' higher debt ratios cause lower short interest ratios.
Interpretation 2	Companies' higher short interest ratios cause higher debt ratios.
Interpretation 3	Companies with higher debt ratios tend to have lower short interest ratios.

She is especially interested in using her estimation results to predict the short interest ratio for MQD Corporation, which has a debt ratio of 0.40.

17 Based on Exhibits 1 and 2, if Liu were to graph the 50 observations, the scatterplot summarizing this relation would be *best* described as:

A horizontal.

B upward sloping.

C downward sloping.

18 Based on Exhibit 1, the sample covariance is *closest to*:

A –9.2430.

B –0.1886.

C 8.4123.

19 Based on Exhibit 1, the correlation between the debt ratio and the short interest ratio is *closest to*:

A –0.3054.

B 0.0933.

C 0.3054.

20 Which of the interpretations *best* describes Liu's findings for her report?

A Interpretation 1

B Interpretation 2

C Interpretation 3

21 The dependent variable in Liu's regression analysis is the:

A intercept.

B debt ratio.

C short interest ratio.

22 Based on Exhibit 2, the degrees of freedom for the *t*-test of the slope coefficient in this regression are:

A 48.

B 49.

C 50.

23 The upper bound for the 95% confidence interval for the coefficient on the debt ratio in the regression is *closest* to:

A –1.0199.

B –0.3947.

C 1.4528.

24 Which of the following should Liu conclude from these results shown in Exhibit 2?

A The average short interest ratio is 5.4975.

B The estimated slope coefficient is statistically significant at the 0.05 level.

C The debt ratio explains 30.54% of the variation in the short interest ratio.

25 Based on Exhibit 2, the short interest ratio expected for MQD Corporation is *closest to*:

A 3.8339.

B 5.4975.

C 6.2462.

26 Based on Liu's regression results in Exhibit 2, the *F*-statistic for testing whether the slope coefficient is equal to zero is *closest* to:

A –2.2219.

B 3.5036.

C 4.9367.

SOLUTIONS

1 The critical t-value for $n - 2 = 34$ df, using a 5 percent significance level and a two-tailed test, is 2.032. First, take the smallest correlation in the table, the correlation between Fund 3 and Fund 4, and see if it is significantly different from zero. Its calculated t-value is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.3102\sqrt{36-2}}{\sqrt{1-0.3102^2}} = 1.903$$

This correlation is not significantly different from zero. If we take the next lowest correlation, between Fund 2 and Fund 3, this correlation of 0.4156 has a calculated t-value of 2.664. So this correlation is significantly different from zero at the 5 percent level of significance. All of the other correlations in the table (besides the 0.3102) are greater than 0.4156, so they too are significantly different from zero.

2 A The coefficient of determination is

$$\frac{\text{Explained variation}}{\text{Total variation}} = \frac{60.16}{140.58} = 0.4279$$

B For a linear regression with one independent variable, the absolute value of correlation between the independent variable and the dependent variable equals the square root of the coefficient of determination, so the correlation is $\sqrt{0.4279} = 0.6542$. (The correlation will have the same sign as the slope coefficient.)

C The standard error of the estimate is

$$\left(\sum_{i=1}^{n}\frac{\left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)^2}{n-2}\right)^{1/2} = \left(\frac{\text{Unexplained variation}}{n-2}\right)^{1/2}$$

$$= \sqrt{\frac{80.42}{60-2}} = 1.178$$

D The sample variance of the dependent variable is

$$\sum_{i=1}^{n}\frac{\left(Y_i - \overline{Y}\right)^2}{n-1} = \frac{\text{Total variation}}{n-1} = \frac{140.58}{60-1} = 2.3827$$

The sample standard deviation is $\sqrt{2.3827} = 1.544$.

3 A The degrees of freedom for the regression is the number of slope parameters in the regression, which is the same as the number of independent variables in the regression. Because regression df = 1, we conclude that there is one independent variable in the regression.

B Total SS is the sum of the squared deviations of the dependent variable Y about its mean.

C The sample variance of the dependent variable is the total SS divided by its degrees of freedom ($n - 1 = 5 - 1 = 4$ as given). Thus the sample variance of the dependent variable is 95.2/4 = 23.8.

D The Regression SS is the part of total sum of squares explained by the regression. Regression SS equals the sum of the squared differences between predicted values of the Y and the sample mean of Y: $\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$. In terms of other values in the table, Regression SS is equal to Total SS minus Residual SS: 95.2 – 7.2 = 88.

E The F-statistic tests whether all the slope coefficients in a linear regression are equal to 0.

F The calculated value of F in the table is equal to the Regression MSS divided by the Residual MSS: 88/2.4 = 36.667.

G Yes. The significance of 0.00904 given in the table is the p-value of the test (the smallest level at which we can reject the null hypothesis). This value of 0.00904 is less than the specified significance level of 0.05, so we reject the null hypothesis. The regression equation has significant explanatory power.

4 The Month 2 data point is an outlier, lying far away from the other data values. Because this outlier was caused by a data entry error, correcting the outlier improves the validity and reliability of the regression. In this case, the true correlation is reduced from 0.996 to 0.824. The revised R-squared is substantially lower (0.678 versus 0.992). The significance of the regression is also lower, as can be seen in the decline of the F-value from 500.79 to 8.44 and the decline in the t-statistic of the slope coefficient from 22.379 to 2.905.

The total sum of squares and regression sum of squares were greatly exaggerated in the incorrect analysis. With the correction, the slope coefficient changes from 1.069 to 0.623. This change is important. When the index moves up or down, the original model indicates that the portfolio return goes up or down by 1.069 times as much, while the revised model indicates that the portfolio return goes up or down by only 0.623 times as much. In this example, incorrect data entry caused the outlier. Had it been a valid observation, not caused by a data error, then the analyst would have had to decide whether the results were more reliable including or excluding the outlier.

5 B is correct. The coefficient of determination is the same as R-squared.

6 C is correct. Deleting observations with small residuals will degrade the strength of the regression, resulting in an *increase* in the standard error and a *decrease* in R-squared.

7 C is correct. For a regression with one independent variable, the correlation is the same as the Multiple R with the sign of the slope coefficient. Because the slope coefficient is positive, the correlation is 0.8623.

8 B is correct. This answer describes the calculation of the F-statistic.

9 C is correct. To make a prediction using the regression model, multiply the slope coefficient by the forecast of the independent variable and add the result to the intercept.

10 C is correct. The p-value is the smallest level of significance at which the null hypotheses concerning the slope coefficient can be rejected. In this case the p-value is less than 0.05, and thus the regression of the ratio of cash flow from operations to sales on the ratio of net income to sales is significant at the 5 percent level.

11 B is correct because the calculated test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$
$$= \frac{-0.1452\sqrt{248-2}}{\sqrt{1-(-0.1452)^2}} = -2.3017$$

Because the absolute value of $t = -2.3017$ is greater than 1.96, the correlation coefficient is statistically significant. For a regression with one independent variable, the *t*-value (and significance) for the slope coefficient (which is −2.3014) should equal the *t*-value (and significance) of the correlation coefficient. The slight difference between these two *t*-values is caused by rounding error.

12 A is correct because the data are time series, and the expected value of the error term, $E(\varepsilon)$, is 0.

13 C is correct. From the regression equation, Expected return = 0.0138 + −0.6486(−0.01) = 0.0138 + 0.006486 = 0.0203, or 2.03 percent.

14 C is correct. *R*-squared is the coefficient of determination. In this case, it shows that 2.11 percent of the variability in Stellar's returns is explained by changes in CPIENG.

15 A is correct, because the standard error of the estimate is the standard deviation of the regression residuals.

16 C is the correct response, because it is a false statement. The slope and intercept are both statistically significant.

17 C is correct because the slope coefficient (Exhibit 2) and the cross-product (Exhibit 1) are negative.

18 B is correct. The sample covariance is calculated as

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = -9.2430 \div 49 = -0.1886$$

19 A is correct. The correlation coefficient equals the covariance between variables *X* and *Y* divided by the product of the standard deviations of variables *X* and *Y*, as follows:

$$\frac{\frac{-9.2430}{49}}{\sqrt{\frac{2.2225}{49}}\sqrt{\frac{412.2042}{49}}} = \frac{-0.1886327}{0.2130 \times 2.9004} = -0.3054$$

20 C is correct. Conclusions cannot be drawn regarding causation, only about association.

21 C is correct. Liu explains the short interest ratio using the debt ratio.

22 A is correct. The degrees of freedom are the number of observations minus the number of parameters estimated, which equals two in this case (the intercept and the slope coefficient). The number of degrees of freedom is 50 − 2 = 48.

23 B is correct. The calculation for the confidence interval is −4.1589 ± (2.011 × 1.8718). The upper bound is −0.3947. The 2.011 is the critical *t*-value for the 5% level of significance (2.5% in one tail) for 48 degrees of freedom.

24 B is correct. The *t*-statistic is –2.2219, which is outside of the bounds created by the critical *t*-values of ± 2.011 for a two-tailed test with a 5% significance level. The 2.011 is the critical *t*-value for the 5% level of significance (2.5% in one tail) for 48 degrees of freedom.

25 A is correct because Predicted value = 5.4975 + (–4.1589 × 0.40) = 5.4975 – 1.6636 = 3.8339.

26 C is correct because $F = \frac{\text{Mean regression sum of squares}}{\text{Mean squared error}} = \frac{38.4404}{7.7867} = 4.9367$.