READING

8

Multiple Regression and Machine Learning

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

Sanjiv R. Das, PhD, is at Santa Clara University (USA). 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.** formulate a multiple regression equation to describe the relation between a dependent variable and several independent variables and determine the statistical significance of each independent variable;
☐	**b.** interpret estimated regression coefficients and their *p*-values;
☐	**c.** formulate a null and an alternative hypothesis about the population value of a regression coefficient, calculate the value of the test statistic, and determine whether to reject the null hypothesis at a given level of significance;
☐	**d.** interpret the results of hypothesis tests of regression coefficients;
☐	**e.** calculate and interpret 1) a confidence interval for the population value of a regression coefficient and 2) a predicted value for the dependent variable, given an estimated regression model and assumed values for the independent variables;
☐	**f.** explain the assumptions of a multiple regression model;
☐	**g.** calculate and interpret the *F*-statistic, and describe how it is used in regression analysis;
☐	**h.** distinguish between and interpret the R^2 and adjusted R^2 in multiple regression;
☐	**i.** evaluate how well a regression model explains the dependent variable by analyzing the output of the regression equation and an ANOVA table;
☐	**j.** formulate a multiple regression equation by using dummy variables to represent qualitative factors and interpret the coefficients and regression results;

(continued)

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
☐	**k.** explain the types of heteroskedasticity and how heteroskedasticity and serial correlation affect statistical inference;
☐	**l.** describe multicollinearity and explain its causes and effects in regression analysis;
☐	**m.** describe how model misspecification affects the results of a regression analysis and describe how to avoid common forms of misspecification;
☐	**n.** describe models with qualitative dependent variables;
☐	**o.** evaluate and interpret a multiple regression model and its results
☐	**p.** distinguish between supervised and unsupervised machine learning;
☐	**q.** describe machine learning algorithms used in prediction, classification, clustering, and dimension reduction;
☐	**r.** describe the steps in model training.

1 INTRODUCTION

As financial analysts, we often need to use more-sophisticated statistical methods than correlation analysis or regression involving a single independent variable. For example, a trading desk interested in the costs of trading NASDAQ stocks might want information on the determinants of the bid–ask spread on the NASDAQ. A mutual fund analyst might want to know whether returns to a technology mutual fund behaved more like the returns to a growth stock index or like the returns to a value stock index. An investor might be interested in the factors that determine whether analysts cover a stock. We can answer these questions using linear regression with more than one independent variable—multiple linear regression.

In Sections 2 and 3, we introduce and illustrate the basic concepts and models of multiple regression analysis. These models rest on assumptions that are sometimes violated in practice. In Section 4, we discuss three commonly occurring violations of regression assumptions. We address practical concerns such as how to diagnose an assumption violation and what remedial steps to take when a model assumption has been violated. Section 5 outlines some guidelines for building good regression models and discusses ways that analysts sometimes go wrong in this endeavor. In Section 6, we discuss a class of models whose dependent variable is qualitative in nature. These models are useful when the concern is over the occurrence of some event, such as whether a stock has analyst coverage or not.

2 MULTIPLE LINEAR REGRESSION

As investment analysts, we often hypothesize that more than one variable explains the behavior of a variable in which we are interested. The variable we seek to explain is called the dependent variable. The variables that we believe explain the dependent

variable are called the independent variables.[1] A tool that permits us to examine the relationship (if any) between the two types of variables is multiple linear regression. **Multiple linear regression** allows us to determine the effect of more than one independent variable on a particular dependent variable.

A **multiple linear regression model** has the general form

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki} + \varepsilon_i, i = 1, 2, \ldots n \quad (1)$$

where

Y_i = the ith observation of the dependent variable Y
X_{ji} = the ith observation of the independent variable X_j, $j = 1, 2, \ldots, k$
b_0 = the intercept of the equation
$b_1, \ldots, b_k$ = the slope coefficients for each of the independent variables
ε_i = the error term
n = the number of observations

A slope coefficient, b_j, measures how much the dependent variable, Y, changes when the independent variable, X_j, changes by one unit, holding all other independent variables constant. For example, if $b_1 = 1$ and all of the other independent variables remain constant, then we predict that if X_1 increases by one unit, Y will also increase by one unit. If $b_1 = -1$ and all of the other independent variables are held constant, then we predict that if X_1 increases by one unit, Y will decrease by one unit. Multiple linear regression estimates $b_0, \ldots, b_k$. In this reading, we will refer to both the intercept, b_0, and the slope coefficients, $b_1, \ldots, b_k$, as **regression coefficients**. As we proceed with our discussion, keep in mind that a regression equation has k slope coefficients and $k + 1$ regression coefficients.

Although Equation 1 may seem to apply only to cross-sectional data because the notation for the observations is the same ($i = 1, \ldots, n$), all of these results apply to time-series data as well. For example, if we analyze data from many time periods for one company, we would typically use the notation $Y_t, X_{1t}, X_{2t}, \ldots, X_{kt}$, in which the first subscript denotes the variable and the second denotes the tth time period.

In practice, we use software to estimate a multiple regression model. Example 1 presents an application of multiple regression analysis in investment practice. In the course of discussing a hypothesis test, Example 1 presents typical regression output and its interpretation.

EXAMPLE 1

Explaining the Bid–Ask Spread

As the manager of the trading desk at an investment management firm, you have noticed that the average bid–ask spreads of different NASDAQ-listed stocks can vary widely. When the ratio of a stock's bid–ask spread to its price is higher than for another stock, your firm's costs of trading in that stock tend to be higher. You have formulated the hypothesis that NASDAQ stocks' percentage bid–ask spreads are related to the number of market makers and the company's stock market capitalization. You have decided to investigate your hypothesis using multiple regression analysis.

You specify a regression model in which the dependent variable measures the percentage bid–ask spread and the independent variables measure the number of market makers and the company's stock market capitalization. The regression is estimated using data from 31 December 2013 for 2,587 NASDAQ-listed stocks.

1 Independent variables are also called explanatory variables or regressors.

Based on earlier published research exploring bid–ask spreads, you express the dependent and independent variables as natural logarithms, a so-called **log-log regression model**. A log-log regression model may be appropriate when one believes that proportional changes in the dependent variable bear a constant relationship to proportional changes in the independent variable(s), as we illustrate below. You formulate the multiple regression:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \varepsilon_i \quad \textbf{(2)}$$

where

Y_i = the natural logarithm of (Bid–ask spread/Stock price) for stock i
X_{1i} = the natural logarithm of the number of NASDAQ market makers for stock i
X_{2i} = the natural logarithm of the market capitalization (measured in millions of US$) of company i

In a log-log regression such as Equation 2, the slope coefficients are interpreted as elasticities, assumed to be constant. For example, a value of $b_2 = -0.75$ would mean that for a 1 percent increase in the market capitalization, we expect Bid–ask spread/Stock price to decrease by 0.75 percent, holding all other independent variables constant.[2]

Reasoning that greater competition tends to lower costs, you suspect that the greater the number of market makers, the smaller the percentage bid–ask spread. Therefore, you formulate a first null hypothesis (H_0) and alternative hypothesis (H_a):

H_0: $b_1 \geq 0$

H_a: $b_1 < 0$

The null hypothesis is the hypothesis that the "suspected" condition is not true. If the evidence supports rejecting the null hypothesis and accepting the alternative hypothesis, you have statistically confirmed your suspicion.[3]

You also believe that the stocks of companies with higher market capitalization may have more-liquid markets, tending to lower percentage bid–ask spreads. Therefore, you formulate a second null hypothesis and alternative hypothesis:

H_0: $b_2 \geq 0$

H_a: $b_2 < 0$

For both tests, we use a t-test, rather than a z-test, because we do not know the population variance of b_1 and b_2. Suppose that you choose a 0.01 significance level for both tests.

2 Note that $\Delta(\ln X) \approx \Delta X/X$, where Δ represents "change in" and $\Delta X/X$ is a proportional change in X. We discuss the model further in Example 11.

3 An alternative valid formulation is a two-sided test (H_0: $b_1 = 0$ versus H_a: $b_1 \neq 0$) which reflects the beliefs of the researcher less strongly. A two-sided test could also be conducted for the hypothesis on market capitalization that we discuss next.

Table 1 Results from Regressing ln(Bid–Ask Spread/Price) on ln(Number of Market Makers) and ln(Market Capitalization)

	Coefficient	Standard Error	*t*-Statistic
Intercept	1.5949	0.2275	7.0105
ln(Number of NASDAQ market makers)	–1.5186	0.0808	–18.7946
ln(Company's market capitalization)	–0.3790	0.0151	–25.0993

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	2	3,728.1334	1,864.0667	2,216.75	0.00
Residual	2,584	2,172.8870	0.8409		
Total	2,586	5,901.0204			
Residual standard error			0.9170		
Multiple *R*-squared			0.6318		
Observations			2,587		

Source: Center for Research in Security Prices, University of Chicago.

Table 1 shows the results of estimating this linear regression using data from 31 December 2013.

If the regression result is not significant, we follow the useful principle of not proceeding to interpret the individual regression coefficients. Thus the analyst might look first at the **analysis of variance (ANOVA)** section, which addresses the regression's overall significance.

- The ANOVA (analysis of variance) section reports quantities related to the overall explanatory power and significance of the regression. SS stands for sum of squares, and MSS stands for mean sum of squares (SS divided by df). The *F*-test reports the overall significance of the regression. For example, an entry of 0.01 for the significance of *F* means that the regression is significant at the 0.01 level. In Table 1, the regression is even more significant because the significance of *F* is 0 at two decimal places. Later in the reading, we will present more information on the *F*-test.

Having ascertained that the overall regression is highly significant, an analyst might turn to the first listed column in the first section of the regression output.

- The Coefficient column gives the estimates of the intercept, b_0, and the slope coefficients, b_1 and b_2. The estimated intercept is positive, but both estimated slope coefficients are negative. Are these estimated regression coefficients significantly different from zero? The Standard Error column gives the standard error (the standard deviation) of the estimated regression coefficients. The test statistic for hypotheses concerning the population value of a regression coefficient has the form (Estimated regression coefficient – Hypothesized population value of the regression coefficient)/(Standard error of the regression coefficient). This is a *t*-test. Under the null hypothesis, the hypothesized population value of the regression coefficient is 0. Thus (Estimated regression coefficient)/(Standard error of the regression coefficient) is the *t*-statistic given in the third column. For example, the *t*-statistic for the intercept is 1.5949/0.2275 = 7.0105. To

evaluate the significance of the *t*-statistic we need to determine a quantity called degrees of freedom (df).[4] The calculation is Degrees of freedom = Number of observations − (Number of independent variables + 1) = $n - (k + 1)$.

- The final section of Table 1 presents two measures of how well the estimated regression fits or explains the data. The first is the standard deviation of the regression residual, the residual standard error. This standard deviation is called the standard error of estimate (SEE). The second measure quantifies the degree of linear association between the dependent variable and all of the independent variables jointly. This measure is known as multiple R^2 or simply R^2 (the square of the correlation between predicted and actual values of the dependent variable).[5] A value of 0 for R^2 indicates no linear association; a value of 1 indicates perfect linear association. The final item in Table 1 is the number of observations in the sample (2,587).

Having reviewed the meaning of typical regression output, we can return to complete the hypothesis tests. The estimated regression supports the hypothesis that the greater the number of market makers, the smaller the percentage bid–ask spread: We reject H_0: $b_1 \geq 0$ in favor of H_a: $b_1 < 0$. The results also support the belief that the stocks of companies with higher market capitalization have lower percentage bid–ask spreads: We reject H_0: $b_2 \geq 0$ in favor of H_a: $b_2 < 0$.

To see that the null hypothesis is rejected for both tests, we can use *t*-test tables. For both tests, df = 2,587 − 3 = 2,584. The tables do not give critical values for degrees of freedom that large. The critical value for a one-tailed test with df = 200 at the 0.01 significance level is 2.345; for a larger number of degrees of freedom, the critical value would be even smaller in magnitude. Therefore, in our one-sided tests, we reject the null hypothesis in favor of the alternative hypothesis if

$$t = \frac{\hat{b}_j - b_j}{s_{\hat{b}_j}} = \frac{\hat{b}_j - 0}{s_{\hat{b}_j}} < -2.345$$

where

$\hat{b}_j$ = the regression estimate of b_j, $j = 1, 2$

b_j = the hypothesized value[6] of the coefficient (0)

$s_{\hat{b}_j}$ = the estimated standard error of $\hat{b}_j$

The *t*-values of −18.7946 and −25.0993 for the estimates of b_1 and b_2, respectively, are both less than −2.345.

Before proceeding further, we should address the interpretation of a prediction stated in natural logarithm terms. We can convert a natural logarithm to the original units by taking the antilogarithm. To illustrate this conversion, suppose that a particular stock has 20 NASDAQ market makers and a market capitalization of $100 million. The natural logarithm of the number of NASDAQ market makers is equal to ln 20 = 2.9957, and the natural logarithm of the company's market cap (in millions) is equal to ln 100 = 4.6052. With these values, the

4 To calculate the degrees of freedom lost in the regression, we add 1 to the number of independent variables to account for the intercept term.

5 Multiple R^2 is also known as the multiple coefficient of determination, or simply the coefficient of determination.

6 To economize on notation in stating test statistics, in this context we use b_j to represent the hypothesized value of the parameter (elsewhere we use it to represent the unknown population parameter).

regression model predicts that the natural log of the ratio of the bid–ask spread to the stock price will be 1.5949 + (−1.5186 × 2.9957) + (−0.3790 × 4.6052) = −4.6997. We take the antilogarithm of −4.6997 by raising e to that power: $e^{-4.6997}$ = 0.0091. The predicted bid–ask spread will be 0.91 percent of the stock price.[7] Later we state the assumptions of the multiple regression model; before using an estimated regression to make predictions in actual practice, we should assure ourselves that those assumptions are satisfied.

In Table 1, we presented output common to most regression software programs. Many software programs also report *p*-values for the regression coefficients.[8] For each regression coefficient, the *p*-value would be the smallest level of significance at which we can reject a null hypothesis that the population value of the coefficient is 0, in a two-sided test. The lower the *p*-value, the stronger the evidence against that null hypothesis. A *p*-value quickly allows us to determine if an independent variable is significant at a conventional significance level such as 0.05, or at any other standard we believe is appropriate.

Having estimated Equation 1, we can write

$$\begin{aligned}\hat{Y}_i &= \hat{b}_0 + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i} \\ &= 1.5949 - 1.5186 X_{1i} - 0.3790 X_{2i}\end{aligned}$$

where $\hat{Y}_i$ stands for the predicted value of Y_i, and $\hat{b}_0$, $\hat{b}_1$, and $\hat{b}_2$, stand for the estimated values of b_0, b_1, and b_2, respectively. How should we interpret the estimated slope coefficients −1.5186 and −0.3790?

Interpreting the slope coefficients in a multiple linear regression model is different than doing so in the one-independent-variable regressions explored in the reading on correlation and regression. Suppose we have a one-independent-variable regression that we estimate as $\hat{Y}_i = 0.50 + 0.75 X_{1i}$. The interpretation of the slope estimate 0.75 is that for every 1-unit increase in X_1, we expect Y to increase by 0.75 units. If we were to add a second independent variable to the equation, we would generally find that the estimated coefficient on X_1 is *not* 0.75 unless the second independent variable were uncorrelated with X_1. The slope coefficients in a multiple regression are known as **partial regression coefficients** or **partial slope coefficients** and need to be interpreted with care.[9] Suppose the coefficient on X_1 in a regression with the second independent variable was 0.60. Can we say that for every 1-unit increase in X_1, we expect Y to increase by 0.60 units? Not without qualification. For every 1-unit increase in X_1, we still expect Y to increase by 0.75 units when X_2 is not held constant. We would interpret 0.60 as the expected increase in Y for a 1-unit increase X_1 *holding the second independent variable constant.*

To explain what the shorthand reference "holding the second independent constant" refers to, if we were to regress X_1 on X_2, the residuals from that regression would represent the part of X_1 that is uncorrelated with X_2. We could then regress Y on those residuals in a one-independent-variable regression. We would find that the slope coefficient on the residuals would be 0.60; by construction, 0.60 would represent the expected effect on Y of a 1-unit increase in X_1 after removing the part of X_1 that is correlated with X_2. Consistent with this explanation, we can view 0.60

7 The operation illustrated (taking the antilogarithm) recovers the value of a variable in the original units as $e^{\ln X} = X$.

8 The entry 0.00 for the significance of *F* was a *p*-value for the *F*-test.

9 The terminology comes from the fact that they correspond to the partial derivatives of Y with respect to the independent variables. Note that in this usage, the term "regression coefficients" refers just to the slope coefficients.

as the expected net effect on Y of a 1-unit increase in X_1, after accounting for any effects of the other independent variables on the expected value of Y. To reiterate, a partial regression coefficient measures the expected change in the dependent variable for a 1-unit increase in an independent variable, holding all the other independent variables constant.

To apply this process to the regression in Table 1, we see that the estimated coefficient on the natural logarithm of market capitalization is −0.3790. Therefore, the model predicts that an increase of 1 in the natural logarithm of the company's market capitalization is associated with a −0.3790 change in the natural logarithm of the ratio of the bid–ask spread to the stock price, holding the natural logarithm of the number of market makers constant. We need to be careful not to expect that the natural logarithm of the ratio of the bid–ask spread to the stock price would differ by −0.3790 if we compared two stocks for which the natural logarithm of the company's market capitalization differed by 1, because in all likelihood the number of market makers for the two stocks would differ as well, which would affect the dependent variable. The value −0.3790 is the expected net effect of difference in log market capitalizations, net of the effect of the log number of market makers on the expected value of the dependent variable.

2.1 Assumptions of the Multiple Linear Regression Model

Before we can conduct correct statistical inference on a multiple linear regression model (a model with more than one independent variable estimated using ordinary least squares), we need to know the assumptions underlying that model.[10] Suppose we have n observations on the dependent variable, Y, and the independent variables, X_1, X_2, ..., X_k, and we want to estimate the equation $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + ... + b_kX_{ki} + \varepsilon_i$.

In order to make a valid inference from a multiple linear regression model, we need to make the following six assumptions, which as a group define the classical normal multiple linear regression model:

1. The relationship between the dependent variable, Y, and the independent variables, X_1, X_2, ..., X_k, is linear as described in Equation 1.
2. The independent variables (X_1, X_2, ..., X_k) are not random.[11] Also, no exact linear relation exists between two or more of the independent variables.[12]
3. The expected value of the error term, conditioned on the independent variables, is 0: $E(\varepsilon \mid X_1, X_2, \ldots, X_k) = 0$.
4. The variance of the error term is the same for all observations:[13] $E\left(\varepsilon_i^2\right) = \sigma_\varepsilon^2$.
5. The error term is uncorrelated across observations: $E(\varepsilon_i\varepsilon_j) = 0$, $j \neq i$.
6. The error term is normally distributed.

10 Ordinary least squares (OLS) is an estimation method based on the criterion of minimizing the sum of the squared residuals of a regression.

11 As discussed in the reading on correlation and regression, even though we assume that independent variables in the regression model are not random, often that assumption is clearly not true. For example, 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. Even if the independent variable is random but uncorrelated with the error term, we can still rely on the results of regression models. See, for example, Greene (2018) or Goldberger (1998).

12 No independent variable can be expressed as a linear combination of any set of the other independent variables. Technically, a constant equal to 1 is included as an independent variable associated with the intercept in this condition.

13 $\text{Var}(\varepsilon) = E(\varepsilon^2)$ and $\text{Cov}(\varepsilon_i\varepsilon_j) = E(\varepsilon_i\varepsilon_j)$ because $E(\varepsilon) = 0$.

Note that these assumptions are almost exactly the same as those for the single-variable linear regression model. Assumption 2 is modified such that no exact linear relation exists between two or more independent variables or combinations of independent variables. If this part of Assumption 2 is violated, then we cannot compute linear regression estimates.[14] Also, even if no exact linear relationship exists between two or more independent variables, or combinations of independent variables, linear regression may encounter problems if two or more of the independent variables or combinations thereof are highly correlated. Such a high correlation is known as multicollinearity, which we will discuss later in this reading. We will also discuss the consequences of conducting regression analysis premised on Assumptions 4 and 5 being met when, in fact, they are violated.

Although Equation 1 may seem to apply only to cross-sectional data because the notation for the observations is the same ($i = 1, ..., n$), all of these results apply to time-series data as well. For example, if we analyze data from many time periods for one company, we would typically use the notation $Y_t, X_{1t}, X_{2t}, ..., X_{kt}$, in which the first subscript denotes the variable and the second denotes the tth time period.

EXAMPLE 2

Factors Explaining the Valuations of Multinational Corporations

Kyaw, Manley, and Shetty (2011) examined which factors affect the valuation of a multinational corporation (MNC). Specifically, they wanted to know whether political risk, transparency, and geographic diversification affected the valuations of MNCs. They used data for 450 US MNCs from 1998 to 2003. The valuations of these corporations were measured using Tobin's q, a commonly used measure of corporate valuation that is calculated as the ratio of the sum of the market value of a corporation's equity and the book value of long-term debt to the sum of the book values of equity and long-term debt. The authors regressed Tobin's q of MNCs on variables representing political risk, transparency, and geographic diversification. The authors also included some additional variables that may affect company valuation, including size, leverage, and beta.[15] They used the equation

$$\text{Tobin's } q_{i,t} = b_0 + b_1(\text{Size}_{i,t}) + b_2(\text{Leverage}_{i,t}) + b_3(\text{Beta}_{i,t}) + b_4(\text{Political risk}_{i,t}) + b_5(\text{Transparency}_{i,t}) + b_6(\text{Geographic diversification}_{i,t}) + \varepsilon_{i,t}$$

14 When we encounter this kind of linear relationship (called perfect collinearity), we cannot compute the matrix inverse needed to compute the linear regression estimates. See Greene (2018) for a further description of this issue.

15 As mentioned in an earlier footnote, technically a constant equal to 1 is included as an independent variable associated with the intercept term in a regression. Because all the regressions reported in this reading include an intercept term, we will not separately mention a constant as an independent variable in the remainder of this reading.

where

$\text{Tobin's } q_{i,t}$ = the Tobin's q for MNC i in year t, with Tobin's q computed as (Market value of equity + Book value of long-term debt)/(Book value of equity + Book value of long-term debt)

$\text{Size}_{i,t}$ = the natural log of the total sales of MNC i in the year t in millions of US$

$\text{Leverage}_{i,t}$ = the ratio of total debt to total assets of MNC i in year t

$\text{Beta}_{i,t}$ = the beta of the stock of MNC i in year t

$\text{Political risk}_{i,t}$ = the at-risk-proportion of international operations of MNC i in year t, calculated as [1 – (number of safe countries/total number of foreign countries in which the firm has operations)], using national risk coding from *Euromoney*

$\text{Transparency}_{i,t}$ = the "transparency percent" (representing the level of disclosure) of MNC i in year t, using survey data from *S&P Transparency & Disclosure*

$\text{Geographic diviersification}_{i,t}$ = foreign sales of MNC i in year t expressed as a percentage of its total sales in that year

Table 2 shows the results of their analysis.[16]

Table 2 Results from Regressing Tobin's *q* on Factors Affecting the Value of Multinational Corporations

	Coefficient	Standard Error*	*t*-Statistic
Intercept	19.829	4.798	4.133
Size	–0.712	0.228	–3.123
Leverage	–3.897	0.987	–3.948
Beta	–1.032	0.261	–3.954
Political risk	–2.079	0.763	–2.725
Transparency	–0.129	0.050	–2.580
Geographic diversification	0.021	0.010	2.100

* This study combines time series observations with cross-sectional observations; such data are commonly referred to as panel data. In such a setting, the standard errors need to be corrected for bias by using a clustered standard error approach as in Petersen (2009). The standard errors reported in this table are clustered standard errors.
Source: Kyaw, Manley, and Shetty (2011).

16 Size is the natural log of total sales. A log transformation (either natural log or log base 10) is commonly used for independent variables that can take a wide range of values; company size and fund size are two such variables. One reason to use the log transformation is to improve the statistical properties of the residuals. If the authors had not taken the log of sales and instead used sales as the independent variable, the regression model probably would not have explained Tobin's q as well.

Suppose that we use the results in Table 2 to test the null hypothesis that the size of a multinational corporation has no effect on its value. Our null hypothesis is that the coefficient on the size variable equals 0 (H_0: $b_1 = 0$), and our alternative hypothesis is that the coefficient does not equal 0 (H_a: $b_1 \neq 0$). The *t*-statistic for testing that hypothesis is

$$t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{-0.712 - 0}{0.228} = -3.12$$

With 450 observations and seven coefficients, the *t*-statistic has 450 – 7 = 443 degrees of freedom. At the 0.05 significance level, the critical value for *t* is about 1.97. The absolute value of computed *t*-statistic on the size coefficient is 3.12, which suggests strongly that we can reject the null hypothesis that size is unrelated to MNC value. In fact, the critical value for *t* is about 2.6 at the 0.01 significance level.

Because $\text{Size}_{i,t}$ is the natural (base *e* or 2.72) log of sales, an increase of 1 in $\text{Size}_{i,t}$ is the same as a 2.72-fold increase in sales. Thus, the estimated coefficient of approximately –0.7 for $\text{Size}_{i,t}$ implies that every 2.72-fold increase in sales of the MNC (an increase of 1 in $\text{Size}_{i,t}$) is associated with an expected decrease of 0.7 in Tobin's $q_{i,t}$ of the MNC, *holding constant the other five independent variables in the regression*.

Now suppose we want to test the null hypothesis that geographic diversification is not related to Tobin's *q*; we want to test whether the coefficient on geographic diversification equals 0 (H_0: $b_6 = 0$) against the alternative hypothesis that the coefficient on geographic diversification does not equal 0 (H_a: $b_6 \neq 0$). The *t*-statistic to test this hypothesis is

$$t = \frac{\hat{b}_6 - b_6}{s_{\hat{b}_6}} = \frac{0.021 - 0}{0.010} = 2.10$$

The critical value of the *t*-test is 1.97 at the 0.05 significance level. Therefore, at the 0.05 significance level, we can reject the null hypothesis that geographic diversification has no effect on MNC valuation. We can interpret the coefficient on geographic diversification of 0.021 as implying that an increase of 1 in the percentage of MNC's sales that are foreign sales is associated with an expected 0.021 increase in Tobin's *q* for the MNC, holding all other independent variables constant.

EXAMPLE 3

Explaining Returns to the Fidelity Select Technology Portfolio

Suppose you are considering an investment in the Fidelity Select Technology Portfolio (FSPTX), a US mutual fund specializing in technology stocks. You want to know whether the fund behaves more like a large-cap growth fund or a large-cap value fund.[17] You decide to estimate the regression

$$Y_t = b_0 + b_1X_{1t} + b_2X_{2t} + \varepsilon_t$$

17 This regression is related to return-based style analysis, one of the most frequent applications of regression analysis in the investment profession. For more information, see Sharpe (1988), who pioneered this field, and Buetow, Johnson, and Runkle (2000).

where

Y_t = the monthly return to the FSPTX
X_{1t} = the monthly return to the S&P 500 Growth Index
X_{2t} = the monthly return to the S&P 500 Value Index

The S&P 500 Growth and S&P 500 Value indices represent predominantly large-cap growth and value stocks, respectively.

Table 3 shows the results of this linear regression using monthly data from January 2009 through December 2013. The estimated intercept in the regression is 0.0018. Thus, if both the return to the S&P 500 Growth Index and the return to the S&P 500 Value Index equal 0 in a specific month, the regression model predicts that the return to the FSPTX will be 0.18 percent. The coefficient on the large-cap growth index is 1.4697, and the coefficient on the large-cap value index return is –0.1833. Therefore, if in a given month the return to the S&P 500 Growth Index was 1 percent and the return to the S&P 500 Value Index was –2 percent, the model predicts that the return to the FSPTX would be 0.0018 + 1.4697(0.01) – 0.1833(–0.02) = 2.02 percent.

Table 3 Results from Regressing the FSPTX Returns on the S&P 500 Growth and S&P 500 Value Indices

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0018	0.0038	0.4737
S&P 500 Growth Index	1.4697	0.2479	5.9286
S&P 500 Value Index	–0.1833	0.2034	–0.9012

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	2	0.1653	0.0826	113.7285	1.27E-20
Residual	57	0.0414	0.0007		
Total	59	0.2067			
Residual standard error			0.0270		
Multiple *R*-squared			0.7996		
Observations			60		

Source: Bloomberg, finance.yahoo.com.

We may want to know whether the coefficient on the returns to the S&P 500 Value Index is statistically significant. Our null hypothesis states that the coefficient equals 0 (H_0: $b_2 = 0$); our alternative hypothesis states that the coefficient does not equal 0 (H_a: $b_2 \neq 0$).

Our test of the null hypothesis uses a *t*-test constructed as follows:

$$t = \frac{\hat{b}_2 - b_2}{s_{\hat{b}_2}} = \frac{-0.1833 - 0}{0.2034} = -0.9012$$

where

$\hat{b}_2$ = the regression estimate of b_2

b_2 = the hypothesized value[18] of the coefficient (0)

$s_{\hat{b}_2}$ = the estimated standard error of $\hat{b}_2$

This regression has 60 observations and three coefficients (two independent variables and the intercept); therefore, the *t*-test has 60 – 3 = 57 degrees of freedom. At the 0.05 significance level, the critical value for the test statistic is about 2.00. The absolute value of the test statistic is 0.9012. Because the test statistic's absolute value is less than the critical value (0.9012 < 2.00), we fail to reject the null hypothesis that $b_2 = 0$. (Note that the *t*-tests reported in Table 3, as well as the other regression tables, are tests of the null hypothesis that the population value of a regression coefficient equals 0.)

Similar analysis shows that at the 0.05 significance level, we cannot reject the null hypothesis that the intercept equals 0 (H_0: $b_0 = 0$) in favor of the alternative hypothesis that the intercept does not equal 0 (H_a: $b_0 \neq 0$). Table 3 shows that the *t*-statistic for testing that hypothesis is 0.4737, a result smaller in absolute value than the critical value of 2.00. However, at the 0.05 significance level we *can* reject the null hypothesis that the coefficient on the S&P 500 Growth Index equals 0 (H_0: $b_1 = 0$) in favor of the alternative hypothesis that the coefficient does not equal 0 (H_a: $b_1 \neq 0$). As Table 3 shows, the *t*-statistic for testing that hypothesis is 5.928, a result far above the critical value of 2.00. Thus multiple regression analysis suggests that returns to the FSPTX are very closely associated with the returns to the S&P 500 Growth Index, but they are not related to S&P 500 Value Index (the *t*-statistic of 0.9012 is not statistically significant).

2.2 Predicting the Dependent Variable in a Multiple Regression Model

Financial analysts often want to predict the value of the dependent variable in a multiple regression based on assumed values of the independent variables. We have previously discussed how to make such a prediction in the case of only one independent variable. The process for making that prediction with multiple linear regression is very similar.

To predict the value of a dependent variable using a multiple linear regression model, we follow these three steps:

1 Obtain estimates $\hat{b}_0$, $\hat{b}_1$, $\hat{b}_2$, ..., $\hat{b}_k$ of the regression parameters b_0, b_1, b_2, ..., b_k.

2 Determine the assumed values of the independent variables, $\hat{X}_{1i}$, $\hat{X}_{2i}$, ..., $\hat{X}_{ki}$.

3 Compute the predicted value of the dependent variable, $\hat{Y}_i$, using the equation

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1\hat{X}_{1i} + \hat{b}_2\hat{X}_{2i} + \ldots + \hat{b}_k\hat{X}_{ki} \quad \textbf{(3)}$$

Two practical points concerning using an estimated regression to predict the dependent variable are in order. First, we should be confident that the assumptions of the regression model are met. Second, we should be cautious about predictions based on values of the independent variables that are outside the range of the data on which the model was estimated; such predictions are often unreliable.

18 To economize on notation in stating test statistics, in this context we use b_2 to represent the hypothesized value of the parameter (elsewhere we use it to represent the unknown population parameter).

EXAMPLE 4

Predicting a Multinational Corporation's Tobin's *q*

In Example 2, we explained the Tobin's q for US multinational corporations (MNC) based on the natural log of sales, leverage, beta, political risk, transparency, and geographic diversification. To review the regression equation:

$$\text{Tobin's } q_{i,t} = b_0 + b_1(\text{Size}_{i,t}) + b_2(\text{Leverage}_{i,t}) + b_3(\text{Beta}_{i,t}) + b_4(\text{Political risk}_{i,t}) + b_5(\text{Transparency}_{i,t}) + b_6(\text{Geographic diversification}_{i,t}) + \varepsilon_i$$

Now we can use the results of the regression reported in Table 2 (excerpted here) to predict the Tobin's q for a US MNC.

Table 2 (excerpt)

	Coefficient
Intercept	19.829
Size	–0.712
Leverage	–3.897
Beta	–1.032
Political risk	–2.079
Transparency	–0.129
Geographic diversification	0.021

1 Suppose that a particular MNC has the following data for a given year.

- Total sales of $7,600 million. The natural log of total sales in millions of US$ equals ln(7,600) = 8.94.
- Leverage (Total debt/Total assets) of 0.45.
- Beta of 1.30.
- Political risk of 0.47, implying that the ratio of the number of safe countries to the total number of foreign countries in which the MNC has operations is 0.53.
- Transparency score of 65, indicating 65% "yes" answers to survey questions related to the corporation's transparency.
- Geographic diversification of 30, indicating that 30% of the corporation's sales are in foreign countries.

What is the predicted Tobin's q for the above MNC?

Solution to 1:

The predicted Tobin's q for the MNC, based on the regression, is:

$$19.829 + (-0.712 \times 8.94) + (-3.897 \times 0.45) + (-1.032 \times 1.30) + (-2.079 \times 0.47) + (-0.129 \times 65) + (0.021 \times 30) = 1.64$$

When predicting the dependent variable using a linear regression model, we encounter two types of uncertainty: uncertainty in the regression model itself, as reflected in the standard error of estimate, and uncertainty about the estimates of

the regression model's parameters. In the reading on correlation and regression, we presented procedures for constructing a prediction interval for linear regression with one independent variable. For multiple regression, however, computing a prediction interval to properly incorporate both types of uncertainty requires matrix algebra, which is outside the scope of this reading.[19]

2.3 Testing whether All Population Regression Coefficients Equal Zero

Earlier, we illustrated how to conduct hypothesis tests on regression coefficients individually. What if we now want to test the significance of the regression as a whole? As a group, do the independent variables help explain the dependent variable? To address this question, we test the null hypothesis that all the slope coefficients in a regression are simultaneously equal to 0. In this section, we further discuss ANOVA with regard to a regression's explanatory power and the inputs for an *F*-test of the above null hypothesis.

If none of the independent variables in a regression model helps explain the dependent variable, the slope coefficients should all equal 0. In a multiple regression, however, we cannot test the null hypothesis that *all* slope coefficients equal 0 based on *t*-tests that *each individual* slope coefficient equals 0, because the individual tests do not account for the effects of interactions among the independent variables. For example, a classic symptom of multicollinearity is that we can reject the hypothesis that all the slope coefficients equal 0 even though none of the *t*-statistics for the individual estimated slope coefficients is significant. Conversely, we can construct unusual examples in which the estimated slope coefficients are significantly different from 0 although jointly they are not.

To test the null hypothesis that all of the slope coefficients in the multiple regression model are jointly equal to 0 (H_0: $b_1 = b_2 = ... = b_k = 0$) against the alternative hypothesis that at least one slope coefficient is not equal to 0 we must use an *F*-test. The *F*-test is viewed as a test of the regression's overall significance.

To correctly calculate the test statistic for the null hypothesis, we need four inputs:

- total number of observations, n;
- total number of regression coefficients to be estimated, $k + 1$, where k is the number of slope coefficients;
- sum of squared errors or residuals, $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$, abbreviated SSE, also known as the residual sum of squares (unexplained variation);[20] and
- regression sum of squares, $\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$, abbreviated RSS.[21] This amount is the variation in Y from its mean that the regression equation explains (explained variation).

The *F*-test for determining whether the slope coefficients equal 0 is based on an *F*-statistic calculated using the four values listed above. The *F*-statistic measures how well the regression equation explains the variation in the dependent variable; it is the ratio of the mean regression sum of squares to the mean squared error.

19 For more information, see Greene (2018).

20 In a table of regression output, this is the number under "SS" column in the row "Residual."

21 In a table of regression output, this is the number under the "SS" column in the row "Regression."

We compute the mean regression sum of squares by dividing the regression sum of squares by the number of slope coefficients estimated, k. We compute the mean squared error by dividing the sum of squared errors by the number of observations, n, minus $(k + 1)$. The two divisors in these computations are the degrees of freedom for calculating an F-statistic. For n observations and k slope coefficients, the F-test for the null hypothesis that the slope coefficients are all equal to 0 is denoted $F_{k,n-(k+1)}$. The subscript indicates that the test should have k degrees of freedom in the numerator (numerator degrees of freedom) and $n - (k + 1)$ degrees of freedom in the denominator (denominator degrees of freedom).

The formula for the F-statistic is

$$F = \frac{\text{RSS}/k}{\text{SSE}/\left[n-(k+1)\right]} = \frac{\text{Mean regression sum of squares}}{\text{Mean squared error}} = \frac{\text{MSR}}{\text{MSE}} \quad \textbf{(4)}$$

where MSR is the mean regression sum of squares and MSE is the mean squared error. In our regression output tables, MSR and MSE are the first and second quantities under the MSS (mean sum of squares) column in the ANOVA section of the output. If the regression model does a good job of explaining variation in the dependent variable, then the ratio MSR/MSE will be large.

What does this F-test tell us when the independent variables in a regression model explain none of the variation in the dependent variable? In this case, each predicted value in the regression model, $\hat{Y}_i$, has the average value of the dependent variable, $\bar{Y}$, and the regression sum of squares, $\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$ is 0. Therefore, the F-statistic for testing the null hypothesis (that all the slope coefficients are equal to 0) has a value of 0 when the independent variables do not explain the dependent variable at all.

To specify the details of making the statistical decision when we have calculated F, we reject the null hypothesis at the α significance level if the calculated value of F is greater than the upper α critical value of the F distribution with the specified numerator and denominator degrees of freedom. Note that we use a one-tailed F-test.[22]

We can illustrate the test using Example 1, in which we investigated whether the natural log of the number of NASDAQ market makers and the natural log of the stock's market capitalization explained the natural log of the bid–ask spread divided by price. Assume that we set the significance level for this test to α = 0.05 (i.e., a 5 percent probability that we will mistakenly reject the null hypothesis if it is true). Table 1 (excerpted here) presents the results of variance computations for this regression.

Table 1 (excerpt)

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	2	3,728.1334	1,864.0667	2,216.7505	0.00
Residual	2,584	2,172.8870	0.8409		
Total	2,586	5,901.0204			

This model has two slope coefficients (k = 2), so there are two degrees of freedom in the numerator of this F-test. With 2,587 observations in the sample, the number of degrees of freedom in the denominator of the F-test is $n - (k + 1)$ = 2,587 − 3 =

22 We use a one-tailed test because MSR necessarily increases relative to MSE as the explanatory power of the regression increases.

2,584. The sum of the squared errors is 2,172.8870. The regression sum of squares is 3,728.1334. Therefore, the *F*-test for the null hypothesis that the two slope coefficients in this model equal 0 is

$$\frac{3{,}728.1334/2}{2{,}172.8870/2{,}584} = 2{,}216.7505$$

This test statistic is distributed as an $F_{2,2,584}$ random variable under the null hypothesis that the slope coefficients are equal to 0. In the table for the 0.05 significance level, we look at the second column, which shows *F*-distributions with two degrees of freedom in the numerator. Near the bottom of the column, we find that the critical value of the *F*-test needed to reject the null hypothesis is between 3.00 and 3.07.[23] The actual value of the *F*-test statistic at 2,216.75 is much greater, so we reject the null hypothesis that coefficients of both independent variables equal 0. In fact, Table 1 under "Significance *F*," reports a *p*-value of 0. This *p*-value means that the smallest level of significance at which the null hypothesis can be rejected is practically 0. The large value for this *F*-statistic implies a minuscule probability of incorrectly rejecting the null hypothesis (a mistake known as a Type I error).

2.4 Adjusted R^2

In the reading on correlation and regression, we presented the coefficient of determination, R^2, as a measure of the goodness of fit of an estimated regression to the data. In a multiple linear regression, however, R^2 is less appropriate as a measure of whether a regression model fits the data well (goodness of fit). Recall that R^2 is defined as

$$\frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}}$$

The numerator equals the regression sum of squares, RSS. Thus R^2 states RSS as a fraction of the total sum of squares, $\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$. If we add regression variables to the model, the amount of unexplained variation will decrease, and RSS will increase, if the new independent variable explains any of the unexplained variation in the model. Such a reduction occurs when the new independent variable is even slightly correlated with the dependent variable and is not a linear combination of other independent variables in the regression.[24] Consequently, we can increase R^2 simply by including many additional independent variables that explain even a slight amount of the previously unexplained variation, even if the amount they explain is not statistically significant.

Some financial analysts use an alternative measure of goodness of fit called **adjusted R^2**, or $\bar{R}^2$. This measure of fit does not automatically increase when another variable is added to a regression; it is adjusted for degrees of freedom. Adjusted R^2 is typically part of the multiple regression output produced by statistical software packages.

The relation between R^2 and $\bar{R}^2$ is

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\left(1 - R^2\right)$$

23 We see a range of values because the denominator has more than 120 degrees of freedom but less than an infinite number of degrees of freedom.

24 We say that variable *y* is a linear combination of variables *x* and *z* if $y = ax + bz$ for some constants *a* and *b*. A variable can also be a linear combination of more than two variables.

where n is the number of observations and k is the number of independent variables (the number of slope coefficients). Note that if $k \geq 1$, then R^2 is strictly greater than adjusted R^2. When a new independent variable is added, $\overline{R}^2$ can decrease if adding that variable results in only a small increase in R^2. In fact, $\overline{R}^2$ can be negative, although R^2 is always nonnegative.[25] If we use $\overline{R}^2$ to compare regression models, it is important that the dependent variable be defined the same way in both models and that the sample sizes used to estimate the models are the same.[26] For example, it makes a difference for the value of $\overline{R}^2$ if the dependent variable is GDP (gross domestic product) or ln(GDP), even if the independent variables are identical. Furthermore, we should be aware that a high $\overline{R}^2$ does not necessarily indicate that the regression is well specified in the sense of including the correct set of variables.[27] One reason for caution is that a high $\overline{R}^2$ may reflect peculiarities of the dataset used to estimate the regression. To evaluate a regression model, we need to take many other factors into account, as we discuss in Section 5.1.

3 USING DUMMY VARIABLES IN REGRESSIONS

Often, financial analysts need to use qualitative variables as independent variables in a regression. One type of qualitative variable, called a **dummy variable**, takes on a value of 1 if a particular condition is true and 0 if that condition is false.[28] For example, suppose we want to test whether stock returns were different in January than during the remaining months of a particular year. We include one independent variable in the regression, X_{1t}, that has a value of 1 for each January and a value of 0 for every other month of the year. We estimate the regression model

$$Y_t = b_0 + b_1 X_{1t} + \varepsilon_t$$

In this equation, the coefficient b_0 is the average value of Y_t in months other than January, and b_1 is the difference between the average value of Y_t in January and the average value of Y_t in months other than January.

We need to exercise care in choosing the number of dummy variables in a regression. The rule is that if we want to distinguish among n categories, we need $n - 1$ dummy variables. For example, to distinguish between *during January* and *not during January* above ($n = 2$ categories), we used one dummy variable ($n - 1 = 2 - 1 = 1$). If we want to distinguish between each of the four quarters in a year, we would include dummy variables for three of the four quarters in a year. If we make the mistake of including dummy variables for four rather than three quarters, we have violated Assumption 2 of the multiple regression model and cannot estimate the regression. The next example illustrates the use of dummy variables in a regression with monthly data.

25 When $\overline{R}^2$ is negative, we can effectively consider its value to be 0.

26 See Gujarati, Porter, and Gunasekar (2011). The value of adjusted R^2 depends on sample size. These points hold if we are using R^2 to compare two regression models.

27 See Mayer (1980).

28 Not all qualitative variables are simple dummy variables. For example, in a trinomial choice model (a model with three choices), a qualitative variable might have the value 0, 1, or 2.

EXAMPLE 5

Month-of-the-Year Effects on Japanese Small-Stock Returns

For many years, financial analysts have been concerned about seasonality in stock returns.[29] In particular, analysts have researched whether returns to small stocks differ during various months of the year. Suppose we want to test whether total returns to one small-stock index, the MSCI Japan Small Cap Index, differ by month. Using data from January 2001 (the first available date for these data) through the end of 2013, we can estimate a regression including an intercept and 11 dummy variables, one for each of the first 11 months of the year. The equation that we estimate is

$$\text{Returns}_t = b_0 + b_1\text{Jan}_t + b_2\text{Feb}_t + \ldots + b_{11}\text{Nov}_t + \varepsilon_t$$

where each monthly dummy variable has a value of 1 when the month occurs (e.g., $\text{Jan}_1 = \text{Jan}_{13} = 1$, as the first observation is a January) and a value of 0 for the other months. Table 4 shows the results of this regression.

The intercept, b_0, measures the average return for stocks in December because there is no dummy variable for December.[30] This equation estimates that the average return in December is 2.73 percent ($\hat{b}_0 = 0.0273$). Each of the estimated coefficients for the dummy variables shows the estimated difference between returns in that month and returns for December. So, for example, the estimated additional return in January is 2.13 percent lower than December ($\hat{b}_1 = -0.0213$). This gives a January return prediction of 0.60 percent (2.73 in December – 2.13 corresponding to the January coefficient).

Table 4 Results from Regressing MSCI Japan Small Cap Index Returns on Monthly Dummy Variables

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0273	0.0149	1.8322
January	–0.0213	0.0210	1.0143
February	–0.0112	0.0210	–0.5333
March	0.0101	0.0210	0.4810
April	–0.0012	0.0210	–0.0571
May	–0.0425	0.0210	–2.0238
June	–0.0065	0.0210	–0.3095
July	–0.0481	0.0210	–2.2905
August	–0.0367	0.0210	–1.7476
September	–0.0285	0.0210	–1.3571
October	–0.0429	0.0210	–2.0429
November	–0.0339	0.0210	–1.6143

(continued)

29 For a discussion of this issue, see Siegel (2014).

30 When $\text{Jan}_t = \text{Feb}_t = \ldots = \text{Nov}_t = 0$, the return is not associated with January through November so the month is December and the regression equation simplifies to $\text{Returns}_1 = b_0 + \varepsilon_t$. Because $E(\text{Returns}_1) = b_0 + E(\varepsilon_t) = b_0$, the intercept b_0 represents the mean return for December.

Table 4 (Continued)

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	11	0.0551	0.0050	1.7421	0.0698
Residual	144	0.4142	0.0029		
Total	155	0.4693			
Residual standard error		0.0536			
Multiple *R*-squared		0.1174			
Observations		156			

Source: Morgan Stanley Capital International.

The low R^2 in this regression (0.1174), however, suggests that a month-of-the-year effect in small-stock returns may not be very important for explaining small-stock returns. We can use the *F*-test to analyze the null hypothesis that jointly, the monthly dummy variables all equal 0 (H_0: $b_1 = b_2 = ... = b_{11} = 0$). We are testing for significant monthly variation in small-stock returns. Table 4 shows the data needed to perform an analysis of variance. The number of degrees of freedom in the numerator of the *F*-test is 11; the number of degrees of freedom in the denominator is [156 – (11 + 1)] = 144. The regression sum of squares equals 0.0551, and the sum of squared errors equals 0.4142. Therefore, the *F*-statistic to determine whether all of the regression slope coefficients are jointly equal to 0 is

$$\frac{0.0551/11}{0.4142/144} = 1.74$$

Appendix D (the *F*-distribution table) at the end of this volume shows the critical values for this *F*-test. If we choose a significance level of 0.05 and look in Column 11 (because the numerator has 11 degrees of freedom), we see that the critical value is 1.87 when the denominator has 120 degrees of freedom. The denominator actually has 144 degrees of freedom, so the critical value of the *F*-statistic is smaller than 1.87 (for df = 120) but larger than 1.79 (for an infinite number of degrees of freedom). The value of the test statistic is 1.74, so we cannot reject the null hypothesis that all of the coefficients jointly are equal to 0.

The *p*-value of 0.0698 shown for the *F*-test in Table 4 means that the smallest level of significance at which we can reject the null hypothesis is roughly 0.07, or 7 percent—which is above the conventional level of 5 percent. Among the 11 monthly dummy variables, May, July, and October have a *t*-statistic with an absolute value greater than 2. Although the coefficients for these dummy variables are statistically significant, we have so many insignificant estimated coefficients that we cannot reject the null hypothesis that returns are equal across the months. This test suggests that the significance of a few coefficients in this regression model may be the result of random variation. We may thus want to avoid portfolio strategies calling for differing investment weights for small stocks in different months.

EXAMPLE 6

Determinants of Short-Term Stock Return Performance in Mergers and Acquisitions by Chinese Companies

Bhabra and Huang (2013) examined short-term market reaction to mergers and acquisition deals initiated by Chinese companies listed on the Shanghai Stock Exchange and the Shenzhen Stock Exchange. They examined those deals during 1997 to 2007 in which the acquirer gained complete control of the target. As the measure of short-term stock return performance around the announcement day, they used the cumulative abnormal return on the acquirer's stock during a five-day window from day –2 to day +2, where day 0 is the acquisition announcement day. Cumulative abnormal return is the excess return achieved over a stated period measured in relation to the return expected given a security's risk. The independent variables in their model included the following firm- and deal-related factors that may affect short-term stock return performance:

- profit margin: ratio of acquiring firm's net income to revenue prior to the acquisition;
- sales growth: acquiring firm's annual sales growth rate prior to the acquisition;
- change in leverage: change in the ratio of debt to total assets for the acquiring firm due to the acquisition;
- firm value: natural logarithm of the market value of the acquirer in the announcement year;
- same industry: Dummy variable (1 = the acquiring and target firms are in the same industry, 0 = in different industries);
- state-owned enterprise: Dummy variable (1 = acquiring firm is a state-owned firm, 0 = not a state-owned firm);
- cash: Dummy variable (1 = the form of payment in the transaction is cash, 0 = other forms of payment);
- cross-border: Dummy variable (1 = cross-border deal, 0 = domestic deal);
- private: Dummy variable (1 = target firm is a stand-alone private firm, 0 = not a stand-alone private firm);
- missing method of payment: Dummy variable (1 = the form of payment information is not available, 0 = form of payment information is available).

Table 5 shows the authors' results.

Table 5 Multiple Regression Model of Cumulative Abnormal Returns for Chinese Acquisitions, 1997–2007

	Coefficient	*p*-Value
Intercept	–0.0543	0.2316
Profit margin	0.0000	0.1522
Sales growth	–0.0180	0.2774
Change in leverage	–0.0136	0.5887
Firm value	0.0024	0.2348
Same industry	0.0012	0.4296
State-owned enterprise	0.0333	0.0435

(continued)

Table 5 (Continued)

	Coefficient	*p*-Value
Cash	0.0041	0.7912
Cross-border	–0.0311	0.3376
Private	–0.0336	0.0204
Missing method of payment	0.0013	0.9399
R-squared	0.2194	
Observations	87	

Source: Bhabra and Huang (2013).

We can summarize Bhabra and Huang's findings as follows:

- The coefficient of state-owned enterprise in this regression model is positive and statistically significant at the 0.05 level as the *p*-value is less than 0.05. State-owned firms play a very important role in the Chinese economy. Non-state-owned firms are relatively new entrants in the Chinese market and are smaller firms. The statistically significant coefficient of state-owned enterprise suggests that the short-term increase in firm value is greater when the acquiring firm is state owned.
- The coefficient of private targets is also statistically significant at the 0.05 level. The sign of this coefficient is negative. Bhabra and Huang point out that the vast majority of target firms in Chinese M&As are unlisted firms, either stand-alone private firms or subsidiaries of listed firms. All the firms included in their sample are either subsidiaries or stand-alone private firms. The significantly negative coefficient of the dummy for private targets tends to result in lower cumulative abnormal returns compared to acquisitions of unlisted subsidiaries. The authors point out that a possible reason could be relatively limited data accessibility for stand-alone private firms as compared with subsidiaries of listed parents. Because of the challenges faced by acquirers when estimating the value and prospects of the private firms, acquisitions of subsidiaries elicit a more positive stock price response.
- Although none of the other coefficients are statistically significant in the above model, in some of the other models estimated in the study (not included in this reading), the authors find some evidence that the stock price response is more positive when the target is in the same industry as the acquirer and the change in leverage is low.

4 VIOLATIONS OF REGRESSION ASSUMPTIONS

In Section 2.1, we presented the assumptions of the multiple linear regression model. Inference based on an estimated regression model rests on those assumptions being satisfied. In applying regression analysis to financial data, analysts need to be able to

diagnose violations of regression assumptions, understand the consequences of violations, and know the remedial steps to take. In the following sections we discuss three regression violations: **heteroskedasticity**, serial correlation, and multicollinearity.

4.1 Heteroskedasticity

So far, we have made an important assumption that the variance of error in a regression is constant across observations. In statistical terms, we assumed that the errors were homoskedastic. Errors in financial data, however, are often **heteroskedastic**: the variance of the errors differs across observations. In this section, we discuss how heteroskedasticity affects statistical analysis, how to test for heteroskedasticity, and how to correct for it.

We can see the difference between homoskedastic and heteroskedastic errors by comparing two graphs. Figure 1 shows the values of the dependent and independent variables and a fitted regression line for a model with homoskedastic errors. There is no systematic relationship between the value of the independent variable and the regression residuals (the vertical distance between a plotted point and the fitted regression line). Figure 2 shows the values of the dependent and independent variables and a fitted regression line for a model with heteroskedastic errors. Here, a systematic relationship is visually apparent: On average, the regression residuals grow much larger as the size of the independent variable increases.

4.1.1 *The Consequences of Heteroskedasticity*

What are the consequences when the assumption of constant error variance is violated? Although heteroskedasticity does not affect the consistency[31] of the regression parameter estimators, it can lead to mistakes in inference. When errors are heteroskedastic, the *F*-test for the overall significance of the regression is unreliable.[32] Furthermore, *t*-tests for the significance of individual regression coefficients are unreliable because heteroskedasticity introduces bias into estimators of the standard error of regression coefficients. If a regression shows significant heteroskedasticity, the standard errors and test statistics computed by regression programs will be incorrect unless they are adjusted for heteroskedasticity.

31 Informally, an estimator of a regression parameter is consistent if the probability that estimates of a regression parameter differ from the true value of the parameter decreases as the number of observations used in the regression increases. The regression parameter estimates from ordinary least squares are consistent regardless of whether the errors are heteroskedastic or homoskedastic. For a more advanced discussion, see Greene (2018).

32 This unreliability occurs because the mean squared error is a biased estimator of the true population variance given heteroskedasticity.

Figure 1 Regression with Homoskedasticity

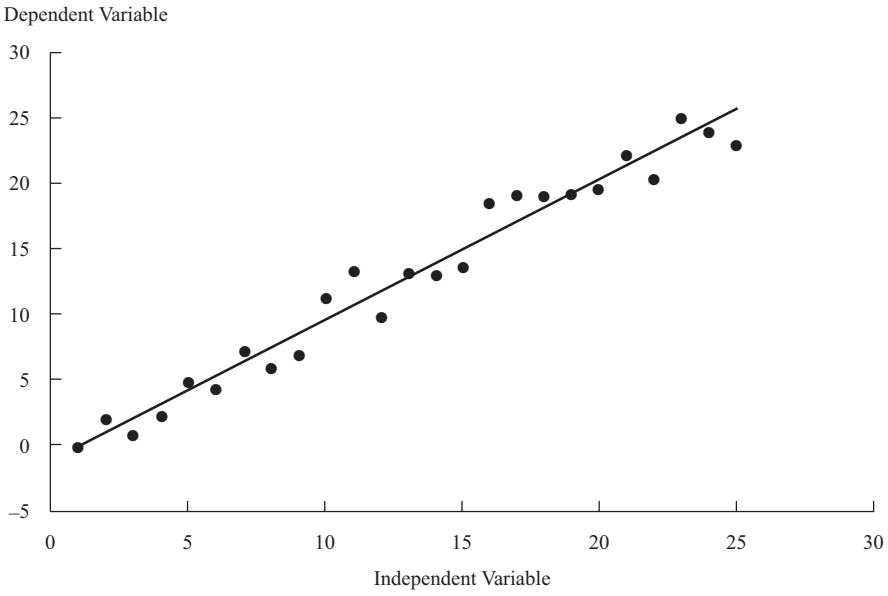
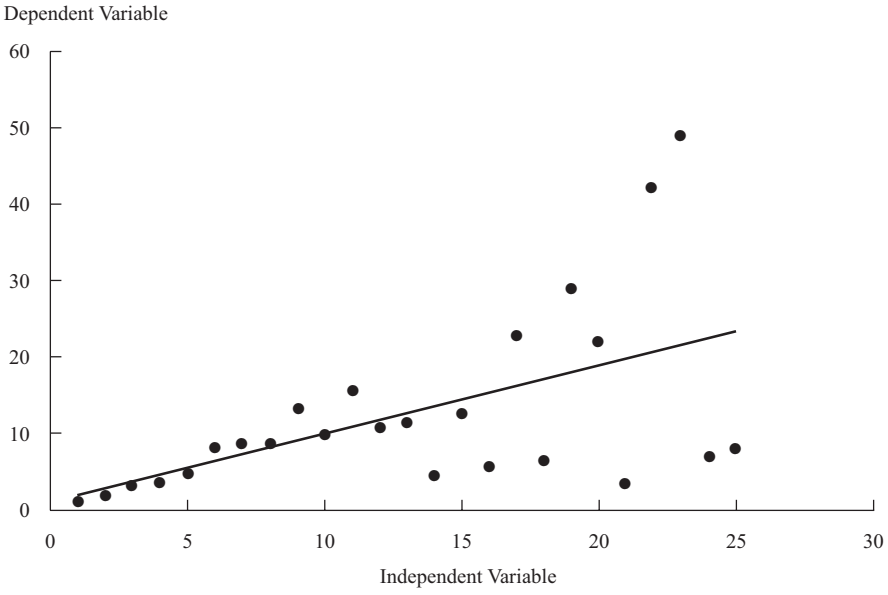


Figure 2 Regression with Heteroskedasticity



In regressions with financial data, the most likely result of heteroskedasticity is that the estimated standard errors will be underestimated and the *t*-statistics will be inflated. When we ignore heteroskedasticity, we tend to find significant relationships

where none actually exist.[33] The consequences in practice may be serious if we are using regression analysis in the development of investment strategies. As Example 7 shows, the issue impinges even on our understanding of financial models.

EXAMPLE 7

Heteroskedasticity and Tests of an Asset Pricing Model

MacKinlay and Richardson (1991) examined how heteroskedasticity affects tests of the capital asset pricing model (CAPM). These authors argued that if the CAPM is correct, they should find no significant differences between the risk-adjusted returns for holding small stocks versus large stocks. To implement their test, MacKinlay and Richardson grouped all stocks on the New York Stock Exchange and the American Stock Exchange (now called NYSE MKT) by market-value decile with annual reassignment. They then tested for systematic differences in risk-adjusted returns across market-capitalization-based stock portfolios. They estimated the following regression:

$$r_{i,t} = \alpha_i + \beta_i r_{m,t} + \varepsilon_{i,t}$$

where

$r_{i,t}$ = excess return (return above the risk-free rate) to portfolio i in period t

$r_{m,t}$ = excess return to the market as a whole in period t

The CAPM formulation hypothesizes that excess returns on a portfolio are explained by excess returns on the market as a whole. That hypothesis implies that $\alpha_i = 0$ for every portfolio i; on average, no excess return accrues to any portfolio after taking into account its systematic (market) risk.

Using data from January 1926 to December 1988 and a market index based on equal-weighted returns, MacKinlay and Richardson failed to reject the CAPM at the 0.05 level when they assumed that the errors in the regression model are normally distributed and homoskedastic. They found, however, that they could reject the CAPM when they corrected their test statistics to account for heteroskedasticity. They rejected the hypothesis that there are no size-based, risk-adjusted excess returns in historical data.[34]

We have stated that effects of heteroskedasticity on statistical inference can be severe. To be more precise about this concept, we should distinguish between two broad kinds of heteroskedasticity: unconditional and conditional.

Unconditional heteroskedasticity occurs when heteroskedasticity of the error variance is not correlated with the independent variables in the multiple regression. Although this form of heteroskedasticity violates Assumption 4 of the linear regression model, it creates no major problems for statistical inference.

The type of heteroskedasticity that causes the most problems for statistical inference is **conditional heteroskedasticity**—heteroskedasticity in the error variance that is correlated with (conditional on) the values of the independent variables in the regression. Fortunately, many statistical software packages easily test and correct for conditional heteroskedasticity.

33 Sometimes, however, failure to adjust for heteroskedasticity results in standard errors that are too large (and t-statistics that are too small).

34 MacKinlay and Richardson also show that when using value-weighted returns, one can reject the CAPM whether or not one assumes normally distributed returns and homoskedasticity.

4.1.2 *Testing for Heteroskedasticity*

Because of conditional heteroskedasticity's consequences on inference, the analyst must be able to diagnose its presence. The Breusch–Pagan test is widely used in finance research because of its generality.[35]

Breusch and Pagan (1979) suggested the following test for conditional heteroskedasticity: Regress the squared residuals from the estimated regression equation on the independent variables in the regression. If no conditional heteroskedasticity exists, the independent variables will not explain much of the variation in the squared residuals. If conditional heteroskedasticity is present in the original regression, however, the independent variables will explain a significant portion of the variation in the squared residuals. The independent variables can explain the variation because each observation's squared residual will be correlated with the independent variables if the independent variables affect the variance of the errors.

Breusch and Pagan showed that under the null hypothesis of no conditional heteroskedasticity, nR^2 (from the regression of the squared residuals on the independent variables from the original regression) will be a χ^2 random variable with the number of degrees of freedom equal to the number of independent variables in the regression.[36] Therefore, the null hypothesis states that the regression's squared error term is uncorrelated with the independent variables. The alternative hypothesis states that the squared error term is correlated with the independent variables. Example 8 illustrates the Breusch–Pagan test for conditional heteroskedasticity.

EXAMPLE 8

Testing for Conditional Heteroskedasticity in the Relation between Interest Rates and Expected Inflation

Suppose an analyst wants to know how closely nominal interest rates are related to expected inflation to determine how to allocate assets in a fixed income portfolio. The analyst wants to test the Fisher effect, the hypothesis suggested by Irving Fisher that nominal interest rates increase by 1 percentage point for every 1 percentage point increase in expected inflation.[37] The Fisher effect assumes the following relation between nominal interest rates, real interest rates, and expected inflation:

$$i = r + \pi^e$$

where

i = the nominal rate
r = the real interest rate (assumed constant)
π^e = the expected rate of inflation

To test the Fisher effect using time-series data, we could specify the following regression model for the nominal interest rate:

$$i_t = b_0 + b_1\pi_t^e + \varepsilon_t \quad \textbf{(5)}$$

35 Some other tests require more-specific assumptions about the functional form of the heteroskedasticity. For more information, see Greene (2018).

36 The Breusch–Pagan test is distributed as a χ^2 random variable in large samples. The constant 1 technically associated with the intercept term in a regression is not counted here in computing the number of independent variables. For more on the Breusch–Pagan test, see Greene (2018).

37 For more on the Fisher effect, see, for example, Mankiw (2015).

Noting that the Fisher effect predicts that the coefficient on the inflation variable is 1, we can state the null and alternative hypotheses as

H_0: $b_1 = 1$

H_a: $b_1 \neq 1$

We might also specify a 0.05 significance level for the test. Before we estimate Equation 5, we must decide how to measure expected inflation (π_t^e) and the nominal interest rate (i_t).

The Survey of Professional Forecasters (SPF) has compiled data on the quarterly inflation expectations of professional forecasters.[38] We use those data as our measure of expected inflation. We use three-month Treasury bill returns as our measure of the (risk-free) nominal interest rate.[39] We use quarterly data from the fourth quarter of 1968 to the fourth quarter of 2013 to estimate Equation 5. Table 6 shows the regression results.

To make the statistical decision on whether the data support the Fisher effect, we calculate the following *t*-statistic, which we then compare to its critical value.

$$t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{1.1744 - 1}{0.0761} = 2.29$$

With a 0.05 significance level and 181 – 2 = 179 degrees of freedom, the critical *t*-value is about 1.97. If we have conducted a valid test, we can reject at the 0.05 significance level the hypothesis that the true coefficient in this regression is 1 and that the Fisher effect holds. The *t*-test assumes that the errors are homoskedastic. Before we accept the validity of the *t*-test, therefore, we should test whether the errors are conditionally heteroskedastic. If those errors prove to be conditionally heteroskedastic, then the test is invalid.

Table 6 Results from Regressing T-Bill Returns on Predicted Inflation

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0116	0.0033	3.5152
Inflation prediction	1.1744	0.0761	15.4323
Residual standard error	0.0233		
Multiple *R*-squared	0.5708		
Observations	181		
Durbin–Watson statistic*	0.2980		

Note: The Durbin–Watson statistic will be explained in Section 4.2.2.
Source: Federal Reserve Bank of Philadelphia, US Department of Commerce.

We can perform the **Breusch–Pagan test** for conditional heteroskedasticity on the squared residuals from the Fisher effect regression. The test regresses the squared residuals on the predicted inflation rate. The R^2 in the squared residuals regression (not shown here) is 0.0666. The test statistic from this regression,

38 For this example, we use the annualized median SPF prediction of current-quarter growth in the GDP deflator (GNP deflator before 1992).

39 Our data on Treasury bill returns are based on three-month T-bill yields in the secondary market. Because those yields are stated on a discount basis, we convert them to a compounded annual rate so they will be measured on the same basis as our data on inflation expectations. These returns are risk-free because they are known at the beginning of the quarter and there is no default risk.

nR^2, is 181 × 0.0666 = 12.0546. Under the null hypothesis of no conditional heteroskedasticity, this test statistic is a χ^2 random variable with one degree of freedom (because there is only one independent variable).

We should be concerned about heteroskedasticity only for large values of the test statistic. Therefore, we should use a one-tailed test to determine whether we can reject the null hypothesis. The critical value of the test statistic for a variable from a χ^2 distribution with one degree of freedom at the 0.05 significance level is 3.84. The test statistic from the Breusch–Pagan test is 12.0546, so we can reject the hypothesis of no conditional heteroskedasticity at the 0.05 level. In fact, we can even reject the hypothesis of no conditional heteroskedasticity at the 0.01 significance level, because the critical value of the test statistic in the case is 6.63. As a result, we conclude that the error term in the Fisher effect regression is conditionally heteroskedastic. The standard errors computed in the original regression are not correct, because they do not account for heteroskedasticity. Therefore, we cannot accept the *t*-test as valid.

In Example 8, we concluded that a *t*-test that we might use to test the Fisher effect was not valid. Does that mean that we cannot use a regression model to investigate the Fisher effect? Fortunately, no. A methodology is available to adjust regression coefficients' standard error to correct for heteroskedasticity. Using an adjusted standard error for $\hat{b}_1$, we can reconduct the *t*-test. As we shall see in the next section, using this valid *t*-test we will not reject the null hypothesis in Example 8. That is, our statistical conclusion will change after we correct for heteroskedasticity.

4.1.3 *Correcting for Heteroskedasticity*

Financial analysts need to know how to correct for heteroskedasticity, because such a correction may reverse the conclusions about a particular hypothesis test—and thus affect a particular investment decision. In Example 7, for instance, MacKinlay and Richardson reversed their investment conclusions after correcting their model's significance tests for heteroskedasticity.

We can use two different methods to correct the effects of conditional heteroskedasticity in linear regression models. The first method, computing **robust standard errors**, corrects the standard errors of the linear regression model's estimated coefficients to account for the conditional heteroskedasticity. The second method, **generalized least squares**, modifies the original equation in an attempt to eliminate the heteroskedasticity. The new, modified regression equation is then estimated under the assumption that heteroskedasticity is no longer a problem.[40] The technical details behind these two methods of correcting for conditional heteroskedasticity are outside the scope of this reading.[41] Many statistical software packages can easily compute robust standard errors, however, and we recommend using them.[42]

Returning to the subject of Example 8 concerning the Fisher effect, recall that we concluded that the error variance was heteroskedastic. If we correct the regression coefficients' standard errors for conditional heteroskedasticity, we get the results shown in Table 7. In comparing the standard errors in Table 7 with those in Table 6, we see that the standard error for the intercept changes very little, but the standard error for the coefficient on predicted inflation (the slope coefficient) increases by

40 Generalized least squares requires econometric expertise to implement correctly on financial data. See Greene (2018), Hansen (1982), and Keane and Runkle (1998).

41 For more details on both methods, see Greene (2018).

42 Robust standard errors are also known as **heteroskedasticity-consistent standard errors** or **White-corrected standard errors**.

about 22 percent (from 0.0761 to 0.0931). Note also that the regression coefficients are the same in both tables, because the results in Table 7 correct only the standard errors in Table 6.

Table 7 Results from Regressing T-Bill Returns on Predicted Inflation (Standard Errors Corrected for Conditional Heteroskedasticity)

	Coefficients	Standard Error	*t*-Statistic
Intercept	0.0116	0.0034	3.4118
Inflation prediction	1.1744	0.0931	12.6144
Residual standard error	0.0233		
Multiple *R*-squared	0.5708		
Observations	181		

Source: Federal Reserve Bank of Philadelphia, US Department of Commerce.

We can now conduct a valid *t*-test of the null hypothesis that the slope coefficient has a true value of 1, using the robust standard error for $\hat{b}_1$. We find that $t = (1.1744 - 1)/0.0931 = 1.8733$. This number is smaller than the critical value of 1.97 needed to reject the null hypothesis that the slope equals 1.[43] So, we can no longer reject the null hypothesis that the slope equals 1. Thus, in this particular example, correcting for the statistically significant conditional heteroskedasticity had an effect on the result of the hypothesis test about the slope of the predicted inflation coefficient. Example 7 concerning tests of the CAPM is a similar case. In other cases, however, our statistical decision might not change based on using robust standard errors in the *t*-test.

4.2 Serial Correlation

A more common—and potentially more serious—problem than violation of the homoskedasticity assumption is the violation of the assumption that regression errors are uncorrelated across observations. Trying to explain a particular financial relation over a number of periods is risky, because errors in financial regression models are often correlated through time.

When regression errors are correlated across observations, we say that they are **serially correlated** (or autocorrelated). Serial correlation most typically arises in time-series regressions. In this section, we discuss three aspects of serial correlation: its effect on statistical inference, tests for it, and methods to correct for it.

4.2.1 *The Consequences of Serial Correlation*

As with heteroskedasticity, the principal problem caused by serial correlation in a linear regression is an incorrect estimate of the regression coefficient standard errors computed by statistical software packages. As long as none of the independent variables is a lagged value of the dependent variable (a value of the dependent variable from a previous period), then the estimated parameters themselves will be consistent and need not be adjusted for the effects of serial correlation. If, however, one of the independent variables is a lagged value of the dependent variable—for example, if the T-bill return from the previous month was an independent variable in the Fisher

43 Remember, this is a two-tailed test.

effect regression—then serial correlation in the error term will cause all the parameter estimates from linear regression to be inconsistent and they will not be valid estimates of the true parameters.[44]

In none of the regressions examined in this reading is an independent variable a lagged value of the dependent variable. Thus, in these regressions, any effect of serial correlation appears in the regression coefficient standard errors. We will examine here the positive serial correlation case, because that case is so common. **Positive serial correlation** is serial correlation in which a positive error for one observation increases the chance of a positive error for another observation. Positive serial correlation also means that a negative error for one observation increases the chance of a negative error for another observation.[45] In examining positive serial correlation, we make the common assumption that serial correlation takes the form of **first-order serial correlation**, or serial correlation between adjacent observations. In a time-series context, that assumption means the sign of the error term tends to persist from one period to the next.

Although positive serial correlation does not affect the consistency of the estimated regression coefficients, it does affect our ability to conduct valid statistical tests. First, the *F*-statistic to test for overall significance of the regression may be inflated because the mean squared error (MSE) will tend to underestimate the population error variance. Second, positive serial correlation typically causes the ordinary least squares (OLS) standard errors for the regression coefficients to underestimate the true standard errors. As a consequence, if positive serial correlation is present in the regression, standard linear regression analysis will typically lead us to compute artificially small standard errors for the regression coefficient. These small standard errors will cause the estimated *t*-statistics to be inflated, suggesting significance where perhaps there is none. The inflated *t*-statistics may, in turn, lead us to incorrectly reject null hypotheses about population values of the parameters of the regression model more often than we would if the standard errors were correctly estimated. This Type I error could lead to improper investment recommendations.[46]

4.2.2 *Testing for Serial Correlation*

We can choose from a variety of tests for serial correlation in a regression model,[47] but the most common is based on a statistic developed by Durbin and Watson (1951); in fact, many statistical software packages compute the Durbin–Watson statistic automatically. The equation for the Durbin–Watson test statistic is

$$DW = \frac{\sum_{t=2}^{T}\left(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1}\right)^2}{\sum_{t=1}^{T}\hat{\varepsilon}_t^2} \quad \textbf{(6)}$$

44 We address this issue in the reading on time-series analysis.

45 In contrast, with **negative serial correlation**, a positive error for one observation increases the chance of a negative error for another observation, and a negative error for one observation increases the chance of a positive error for another.

46 OLS standard errors need not be underestimates of actual standard errors if negative serial correlation is present in the regression.

47 See Greene (2018) for a detailed discussion of tests of serial correlation.

where $\hat{\varepsilon}_t$ is the regression residual for period t. We can rewrite this equation as

$$\frac{\frac{1}{T-1}\sum_{t=2}^{T}\left(\hat{\varepsilon}_t^2 - 2\hat{\varepsilon}_t\hat{\varepsilon}_{t-1} + \hat{\varepsilon}_{t-1}^2\right)}{\frac{1}{T-1}\sum_{t=1}^{T}\hat{\varepsilon}_t^2}$$
$$\approx \frac{Var\left(\hat{\varepsilon}_t\right) - 2\,\mathrm{Cov}\left(\hat{\varepsilon}_t,\hat{\varepsilon}_{t-1}\right) + \mathrm{Var}\left(\hat{\varepsilon}_{t-1}\right)}{\mathrm{Var}\left(\hat{\varepsilon}_t\right)}$$

If the variance of the error is constant through time, then we expect $\mathrm{Var}\left(\hat{\varepsilon}_t\right) = \hat{\sigma}_\varepsilon^2$ for all t, where we use $\hat{\sigma}_\varepsilon^2$ to represent the estimate of the constant error variance. If, in addition, the errors are also not serially correlated, then we expect $\mathrm{Cov}\left(\hat{\varepsilon}_t,\hat{\varepsilon}_{t-1}\right) = 0$. In that case, the Durbin–Watson statistic is approximately equal to

$$\frac{\hat{\sigma}_\varepsilon^2 - 0 + \hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\varepsilon^2} = 2$$

This equation tells us that if the errors are homoskedastic and not serially correlated, then the Durbin–Watson statistic will be close to 2. Therefore, we can test the null hypothesis that the errors are not serially correlated by testing whether the Durbin–Watson statistic differs significantly from 2.

If the sample is very large, the Durbin–Watson statistic will be approximately equal to $2(1 - r)$, where r is the sample correlation between the regression residuals from one period and those from the previous period. This approximation is useful because it shows the value of the Durbin–Watson statistic for differing levels of serial correlation. The Durbin–Watson statistic can take on values ranging from 0 (in the case of serial correlation of +1) to 4 (in the case of serial correlation of –1):

- If the regression has no serial correlation, then the regression residuals will be uncorrelated through time and the value of the Durbin–Watson statistic will be equal to $2(1 - 0) = 2$.
- If the regression residuals are positively serially correlated, then the Durbin–Watson statistic will be less than 2. For example, if the serial correlation of the errors is 1, then the value of the Durbin–Watson statistic will be 0.
- If the regression residuals are negatively serially correlated, then the Durbin–Watson statistic will be greater than 2. For example, if the serial correlation of the errors is –1, then the value of the Durbin–Watson statistic will be 4.

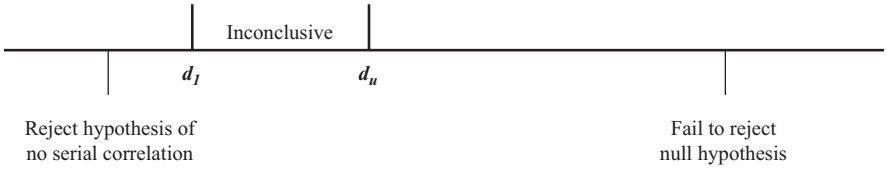
Returning to Example 8, which explored the Fisher effect, as shown in Table 6 the Durbin–Watson statistic for the OLS regression is 0.2980. This result means that the regression residuals are positively serially correlated:

$$\begin{aligned} DW &= 0.2980 \\ &\approx 2(1 - r) \\ r &\approx 1 - \mathrm{DW}/2 \\ &= 1 - 0.2980/2 \\ &= 0.8510 \end{aligned}$$

This outcome raises the concern that OLS standard errors may be incorrect because of positive serial correlation. Does the observed Durbin–Watson statistic (0.2980) provide enough evidence to warrant rejecting the null hypothesis of no positive serial correlation?

We should reject the null hypothesis of no serial correlation if the Durbin–Watson statistic is below a critical value, d^*. Unfortunately, Durbin and Watson also showed that, for a given sample, we cannot know the true critical value, d^*. Instead, we can determine only that d^* lies either between two values, d_u (an upper value) and d_l (a lower value), or outside those values. Figure 3 depicts the upper and lower values of d^* as they relate to the results of the Durbin–Watson statistic.

Figure 3 Value of the Durbin–Watson Statistic



From Figure 3, we learn the following:

- When the Durbin–Watson (DW) statistic is less than d_l, we reject the null hypothesis of no positive serial correlation.
- When the DW statistic falls between d_l and d_u, the test results are inconclusive.
- When the DW statistic is greater than d_u, we fail to reject the null hypothesis of no positive serial correlation.[48]

Returning to Example 8, the Fisher effect regression has one independent variable and 181 observations. The Durbin–Watson statistic is 0.2980. We can reject the null hypothesis of no correlation in favor of the alternative hypothesis of positive serial correlation at the 0.05 level because the Durbin–Watson statistic is far below d_l for $k = 1$ and $n = 100$ (1.65). The level of d_l would be even higher for a sample of 181 observations. This finding of significant positive serial correlation suggests that the OLS standard errors in this regression probably significantly underestimate the true standard errors.

4.2.3 *Correcting for Serial Correlation*

We have two alternative remedial steps when a regression has significant serial correlation. First, we can adjust the coefficient standard errors for the linear regression parameter estimates to account for the serial correlation. Second, we can modify the regression equation itself to eliminate the serial correlation. We recommend using the first method for dealing with serial correlation; the second method may result in inconsistent parameter estimates unless implemented with extreme care.

Two of the most prevalent methods for adjusting standard errors were developed by Hansen (1982) and Newey and West (1987). These methods are standard features in many statistical software packages.[49] An additional advantage of these methods is that they simultaneously correct for conditional heteroskedasticity.[50]

48 Of course, sometimes serial correlation in a regression model is negative rather than positive. For a null hypothesis of no serial correlation, the null hypothesis is rejected if DW < d_l (indicating significant positive serial correlation) or if DW > 4 – d_l (indicating significant negative serial correlation).

49 This correction is known by various names, including serial-correlation consistent standard errors, serial correlation and heteroskedasticity adjusted standard errors, and robust standard errors. The Hansen standard errors are also known as Hansen–White standard errors.

50 We do not always use Hansen's method or Newey–West method to correct for serial correlation and heteroskedasticity because sometimes the errors of a regression are not serially correlated.

Table 8 shows the results of correcting the standard errors from Table 6 for serial correlation and heteroskedasticity using the Newey–West method. Note that the coefficients for both the intercept and the slope are exactly the same as in the original regression. The robust standard errors are now much larger, however—more than twice the OLS standard errors in Table 6. Because of the severe serial correlation in the regression error, OLS greatly underestimates the uncertainty about the estimated parameters in the regression.

Note also that the serial correlation has not been eliminated, but the standard error has been corrected to account for the serial correlation.

Table 8 Results from Regressing T-Bill Returns on Predicted Inflation (Standard Errors Corrected for Conditional Heteroskedasticity and Serial Correlation)

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0116	0.0067	1.7313
Inflation prediction	1.1744	0.1751	6.7070
Residual standard error	0.0233		
Multiple *R*-squared	0.5708		
Observations	181		

Source: Federal Reserve Bank of Philadelphia, US Department of Commerce.

Now suppose we want to test our original null hypothesis (the Fisher effect) that the coefficient on the predicted inflation term equals 1 (H_0: $b_1 = 1$) against the alternative that the coefficient on the inflation term is not equal to 1 (H_a: $b_1 \neq 1$). With the corrected standard errors, the value of the test statistic for this null hypothesis is

$$\frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{1.1744 - 1}{0.1751} = 0.996$$

The critical values for both the 0.05 and 0.01 significance level are much larger than 0.996 (the *t*-test statistic), so we cannot reject the null hypothesis. This conclusion is the same as that reached in Example 7 where the correction was only for heteroskedasticity; but it is the opposite of the conclusion in Example 6 where there were no correlations.

This shows that for some hypotheses, serial correlation and conditional heteroskedasticity could have a big effect on whether we accept or reject those hypotheses.[51]

4.3 Multicollinearity

The second assumption of the multiple linear regression model is that no exact linear relationship exists between two or more of the independent variables. When one of the independent variables is an exact linear combination of other independent variables, it becomes mechanically impossible to estimate the regression. That case, known as perfect collinearity, is much less of a practical concern than

51 Serial correlation can also affect forecast accuracy.

multicollinearity.[52] **Multicollinearity** occurs when two or more independent variables (or combinations of independent variables) are highly (but not perfectly) correlated with each other. With multicollinearity we can estimate the regression, but the interpretation of the regression output becomes problematic. Multicollinearity is a serious practical concern because approximate linear relationships among financial variables are common.

4.3.1 *The Consequences of Multicollinearity*

Although the presence of multicollinearity does not affect the consistency of the OLS estimates of the regression coefficients, the estimates become extremely imprecise and unreliable. Furthermore, it becomes practically impossible to distinguish the individual impacts of the independent variables on the dependent variable. These consequences are reflected in inflated OLS standard errors for the regression coefficients. With inflated standard errors, *t*-tests on the coefficients have little power (ability to reject the null hypothesis).

4.3.2 *Detecting Multicollinearity*

In contrast to the cases of heteroskedasticity and serial correlation, we shall not provide a formal statistical test for multicollinearity. In practice, multicollinearity is often a matter of degree rather than of absence or presence.[53]

The analyst should be aware that using the magnitude of pairwise correlations among the independent variables to assess multicollinearity, as has occasionally been suggested, is generally not adequate. Although very high pairwise correlations among independent variables can indicate multicollinearity, it is not necessary for such pairwise correlations to be high for there to be a problem of multicollinearity.[54] Stated another way, high pairwise correlations among the independent variables are not a necessary condition for multicollinearity, and low pairwise correlations do not mean that multicollinearity is not a problem. The only case in which correlation between independent variables may be a reasonable indicator of multicollinearity occurs in a regression with exactly two independent variables.

The classic symptom of multicollinearity is a high R^2 (and significant F-statistic) even though the t-statistics on the estimated slope coefficients are not significant. The insignificant t-statistics reflect inflated standard errors. Although the coefficients might be estimated with great imprecision, as reflected in low t-statistics, the independent variables *as a group* may do a good job of explaining the dependent variable, and a high R^2 would reflect this effectiveness. Example 9 illustrates this diagnostic.

EXAMPLE 9

Multicollinearity in Explaining Returns to the Fidelity Select Technology Portfolio

In Example 3 we regressed returns to the Fidelity Select Technology Portfolio (FSPTX) on returns to the S&P 500 Growth Index and the S&P 500 Value Index. Table 9 shows the results of our regression, which uses data from January 2009 through December 2013. The t-statistic of 5.9286 on the growth index return is

52 To give an example of perfect collinearity, suppose we tried to explain a company's credit ratings with a regression that included net sales, cost of goods sold, and gross profit as independent variables. Because Gross profit = Net sales − Cost of goods sold by definition, there is an exact linear relationship between these variables. This type of blunder is relatively obvious (and easy to avoid).

53 See Kmenta (1986).

54 Even if pairs of independent variables have low correlation, there may be linear combinations of the independent variables that are very highly correlated, creating a multicollinearity problem.

greater than 2, indicating that the coefficient on the growth index differs significantly from 0 at standard significance levels. On the other hand, the *t*-statistic on the value index return is –0.9012 and thus is not statistically significant. This result suggests that the returns to the FSPTX are linked to the returns to the growth index and not closely associated with the returns to the value index. The coefficient on the growth index, however, is 1.4697. This result implies that returns on the FSPTX are more volatile than are returns on the growth index.

Table 9 Results from Regressing the FSPTX Returns on the S&P 500 Growth and Value Indices

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0018	0.0038	0.4737
S&P 500 Growth Index	1.4697	0.2479	5.9286
S&P 500 Value Index	–0.1833	0.2034	–0.9012

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	2	0.1653	0.0826	113.7285	1.27E-20
Residual	57	0.0414	0.0007		
Total	59	0.2067			
Residual standard error			0.0270		
Multiple *R*-squared			0.7996		
Observations			60		

Source: Bloomberg, finance.yahoo.com.

Note also that this regression explains a significant amount of the variation in the returns to the FSPTX. Specifically, the R^2 from this regression is 0.7996. Thus approximately 80 percent of the variation in the returns to the FSPTX is explained by returns to the S&P 500 Growth and S&P 500 Value indices.

Now suppose we run another linear regression that adds returns to the S&P 500 itself to the returns to the S&P 500 Growth and S&P 500 Value indices. The S&P 500 includes the component stocks of these two style indices, so we are introducing a severe multicollinearity problem.

Table 10 shows the results of that regression. Note that the R^2 in this regression has changed almost imperceptibly from the R^2 in the previous regression (increasing from 0.7996 to 0.8084), but now the standard errors of the coefficients of the independent variables are much larger. Adding the return to the S&P 500 to the previous regression does not explain any more of the variance in the returns to the FSPTX than the previous regression did, but now none of the coefficients is statistically significant. This is the classic case of multicollinearity mentioned in the reading.

Table 10 Results from Regressing the FSPTX Returns on Returns to the S&P 500 Growth and S&P 500 Value Indices and the S&P 500 Index

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0008	0.0038	0.2105
S&P 500 Growth Index	14.2444	7.9783	1.7854
S&P 500 Value Index	11.6955	7.4180	1.5766
S&P 500 Index	–24.6734	15.4022	–1.6019

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	3	0.1671	0.0557	78.7577	4.14E-20
Residual	56	0.0396	0.0007		
Total	59	0.2067			
Residual standard error			0.0266		
Multiple *R*-squared			0.8084		
Observations			60		

Source: Bloomberg, finance.yahoo.com, S&P Dow Jones Indices.

Multicollinearity may be a problem even when we do not observe the classic symptom of insignificant *t*-statistics but a highly significant *F*-test. Advanced textbooks provide further tools to help diagnose multicollinearity.[55]

4.3.3 *Correcting for Multicollinearity*

The most direct solution to multicollinearity is excluding one or more of the regression variables. In the example above, we can see that the S&P 500 total returns should not be included if both the S&P 500 Growth and S&P 500 Value indices are included, because the returns to the entire S&P 500 Index are a weighted average of the return to growth stocks and value stocks. In many cases, however, no easy solution is available to the problem of multicollinearity, and you will need to experiment with including or excluding different independent variables to determine the source of multicollinearity.

4.4 Heteroskedasticity, Serial Correlation, Multicollinearity: Summarizing the Issues

We have discussed some of the problems that heteroskedasticity, serial correlation, and multicollinearity may cause in interpreting regression results. These violations of regression assumptions, we have noted, all lead to problems in making valid inferences. The analyst should check that model assumptions are fulfilled before interpreting statistical tests.

Table 11 gives a summary of these problems, the effect they have on the linear regression results (an analyst can see these effects using regression software), and the solutions to these problems.

55 See Greene (2018).

Table 11 Problems in Linear Regression and Their Solutions

Problem	Effect	Solution
Heteroskedasticity	Incorrect standard errors	Use robust standard errors (corrected for conditional heteroskedasticity)
Serial correlation	Incorrect standard errors (additional problems if a lagged value of the dependent variable is used as an independent variable)	Use robust standard errors (corrected for serial correlation)
Multicollinearity	High R^2 and low t-statistics	Remove one or more independent variables; often no solution based in theory

MODEL SPECIFICATION AND ERRORS IN SPECIFICATION

5

Until now, we have assumed that whatever regression model we estimate is correctly specified. **Model specification** refers to the set of variables included in the regression and the regression equation's functional form. In the following, we first give some broad guidelines for correctly specifying a regression. Then we turn to three types of model misspecification: misspecified functional form, regressors that are correlated with the error term, and additional time-series misspecification. Each of these types of misspecification invalidates statistical inference using OLS; most of these misspecifications will cause the estimated regression coefficients to be inconsistent.

5.1 Principles of Model Specification

In discussing the principles of model specification, we need to acknowledge that there are competing philosophies about how to approach model specification. Furthermore, our purpose for using regression analysis may affect the specification we choose. The following principles have fairly broad application, however.

- *The model should be grounded in cogent economic reasoning.* We should be able to supply the economic reasoning behind the choice of variables, and the reasoning should make sense. When this condition is fulfilled, we increase the chance that the model will have predictive value with new data. This approach contrasts to the variable-selection process known as **data mining**. With data mining, the investigator essentially develops a model that maximally exploits the characteristics of a specific dataset. "Data mining" is used in the different sense of discovering patterns in large datasets in contexts discussed later in Section 7.
- *The functional form chosen for the variables in the regression should be appropriate given the nature of the variables.* As one illustration, consider studying mutual fund **market timing** based on fund and market returns alone. One might reason that for a successful timer, a plot of mutual fund returns against market returns would show curvature, because a successful timer would tend to

increase (decrease) beta when market returns were high (low). The model specification should reflect the expected nonlinear relationship.[56] In other cases, we may transform the data such that a regression assumption is better satisfied.

- *The model should be parsimonious.* In this context, "parsimonious" means accomplishing a lot with a little. We should expect each variable included in a regression to play an essential role.
- *The model should be examined for violations of regression assumptions before being accepted.* We have already discussed detecting the presence of heteroskedasticity, serial correlation, and multicollinearity. As a result of such diagnostics, we may conclude that we need to revise the set of included variables and/or their functional form.
- *The model should be tested and be found useful out of sample before being accepted.* The term "out of sample" refers to observations outside the dataset on which the model was estimated. A plausible model may not perform well out of sample because economic relationships have changed since the sample period. That possibility is itself useful to know. A second explanation, however, may be that relationships have not changed but that the model explains only a specific dataset.

Having given some broad guidance on model specification, we turn to a discussion of specific model specification errors. Understanding these errors will help an analyst develop better models and be a more informed consumer of investment research.

5.2 Misspecified Functional Form

Whenever we estimate a regression, we must assume that the regression has the correct functional form. This assumption can fail in several ways:

- One or more important variables could be omitted from regression.
- One or more of the regression variables may need to be transformed (for example, by taking the natural logarithm of the variable) before estimating the regression.
- The regression model pools data from different samples that should not be pooled.

First, consider the effects of omitting an important independent variable from a regression (omitted variable bias). If the true regression model was

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \varepsilon_i \quad (7)$$

but we estimate the model[57]

$$Y_i = a_0 + a_1X_{1i} + \varepsilon_i$$

then our regression model would be misspecified. What is wrong with the model?

If the omitted variable (X_2) is correlated with the remaining variable (X_1), then the error term in the model will be correlated with (X_1), and the estimated values of the regression coefficients a_0 and a_1 would be biased and inconsistent. In addition,

56 This example is based on Treynor and Mazuy (1966), an early regression study of mutual fund timing. To capture curvature, they included a term in the squared market excess return, which does not violate the assumption of the multiple linear regression model that relationship between the dependent and independent variables is linear *in the coefficients*.

57 We use a different regression coefficient notation when X_{2i} is omitted, because the intercept term and slope coefficient on X_{1i} will generally not be the same as when X_{2i} is included.

the estimates of the standard errors of those coefficients will also be inconsistent, so we can use neither the coefficients estimates nor the estimated standard errors to make statistical tests.

EXAMPLE 10

Omitted Variable Bias and the Bid–Ask Spread

In this example, we extend our examination of the bid–ask spread to show the effect of omitting an important variable from a regression. In Example 1, we showed that the natural logarithm of the ratio [(Bid–ask spread)/Price] was significantly related to both the natural logarithm of the number of market makers and the natural logarithm of the market capitalization of the company. We repeat Table 1 from Example 1 below.

Table 1 Results from Regressing ln(Bid–Ask Spread/Price) on ln(Number of Market Makers) and ln(Market Capitalization) (repeated)

	Coefficients	Standard Error	*t*-Statistic
Intercept	1.5949	0.2275	7.0105
ln(Number of NASDAQ market makers)	–1.5186	0.0808	–18.7946
ln(Company's market capitalization)	–0.3790	0.0151	–25.0993

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	2	3,728.1334	1,864.0667	2,216.7505	0.00
Residual	2,584	2,172.8870	0.8409		
Total	2,586	5,901.0204			
Residual standard error			0.9170		
Multiple *R*-squared			0.6318		
Observations			2,587		

Source: Center for Research in Security Prices, University of Chicago.

If we did not include the natural log of market capitalization as an independent variable in the regression, and we regressed the natural logarithm of the ratio [(Bid–ask spread)/Price] only on the natural logarithm of the number of market makers for the stock, the results would be as shown in Table 12.

Table 12 Results from Regressing ln(Bid–Ask Spread/Price) on ln(Number of Market Makers)

	Coefficients	Standard Error	*t*-Statistic
Intercept	5.0707	0.2009	25.2399
ln(Number of NASDAQ market makers)	–3.1027	0.0561	–55.3066

(continued)

Table 12 (Continued)

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	1	3,200.3918	3,200.3918	3,063.3655	0.00
Residual	2,585	2,700.6287	1.0447		
Total	2,586	5,901.0204			
Residual standard error			1.0221		
Multiple *R*-squared			0.5423		
Observations			2,587		

Source: Center for Research in Security Prices, University of Chicago.

Note that the coefficient on ln(Number of NASDAQ market makers) changed from −1.5186 in the original (correctly specified) regression to −3.1027 in the misspecified regression. Also, the intercept changed from 1.5949 in the correctly specified regression to 5.0707 in the misspecified regression. These results illustrate that omitting an independent variable that should be in the regression can cause the remaining regression coefficients to be inconsistent.

A second common cause of misspecification in regression models is the use of the wrong form of the data in a regression, when a transformed version of the data is appropriate. For example, sometimes analysts fail to account for curvature or nonlinearity in the relationship between the dependent variable and one or more of the independent variables, instead specifying a linear relation among variables. When we are specifying a regression model, we should consider whether economic theory suggests a nonlinear relation. We can often confirm the nonlinearity by plotting the data, as we will illustrate in Example 11 below. If the relationship between the variables becomes linear when one or more of the variables is represented as a proportional change in the variable, we may be able to correct the misspecification by taking the natural logarithm of the variable(s) we want to represent as a proportional change. Other times, analysts use unscaled data in regressions, when scaled data (such as dividing net income or cash flow by sales) are more appropriate. In Example 1, we scaled the bid–ask spread by stock price because what a given bid–ask spread means in terms of transactions costs for a given size investment depends on the price of the stock; if we had not scaled the bid–ask spread, the regression would have been misspecified.

EXAMPLE 11

Nonlinearity and the Bid–Ask Spread

In Example 1, we showed that the natural logarithm of the ratio [(Bid–ask spread)/Price] was significantly related to both the natural logarithm of the number of market makers and the natural logarithm of the company's market capitalization. But why did we take the natural logarithm of each of the variables in the regression? We began a discussion of this question in Example 1, which we continue now.

What does theory suggest about the nature of the relationship between the ratio (Bid–ask spread)/Price, or the percentage bid–ask spread, and its determinants (the independent variables)? Stoll (1978) builds a theoretical model of

the determinants of percentage bid–ask spread in a dealer market. In his model, the determinants enter multiplicatively in a particular fashion. In terms of the independent variables introduced in Example 1, the functional form assumed is

$$\left[\left(\text{Bid–ask spread}\right)/\text{Price}\right]_i = c\left(\text{Number of market makers}\right)_i^{b_1} \times \left(\text{Market capitalization}\right)_i^{b_2}$$

where c is a constant. The relationship of the percentage bid–ask spread with the number of market makers and market capitalization is not linear in the original variables.[58] If we take the natural log of both sides of the above model, however, we have a log-log regression that is linear in the transformed variables:[59]

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \varepsilon_i$$

where

Y_i = the natural logarithm of the ratio (Bid–ask spread)/Price for stock i
b_0 = a constant that equals $\ln(c)$
X_{1i} = the natural logarithm of the number of market makers for stock i
X_{2i} = the natural logarithm of the market capitalization of company i
ε_i = the error term

As mentioned in Example 1, a slope coefficient in the log-log model is interpreted as an elasticity, precisely, the partial elasticity of the dependent variable with respect to the independent variable ("partial" means holding the other independent variables constant).

We can plot the data to assess whether the variables are linearly related after the logarithmic transformation. For example Figure 4 shows a scatterplot of the natural logarithm of the number of market makers for a stock (on the X axis) and the natural logarithm of (Bid–ask spread)/Price (on the Y axis), as well as a regression line showing the linear relation between the two transformed variables. The relation between the two transformed variables is clearly linear.

58 The form of the model is analogous to the Cobb–Douglas production function in economics.

59 We have added an error term to the model.

Figure 4 Linear Regression When Two Variables Have a Linear Relation

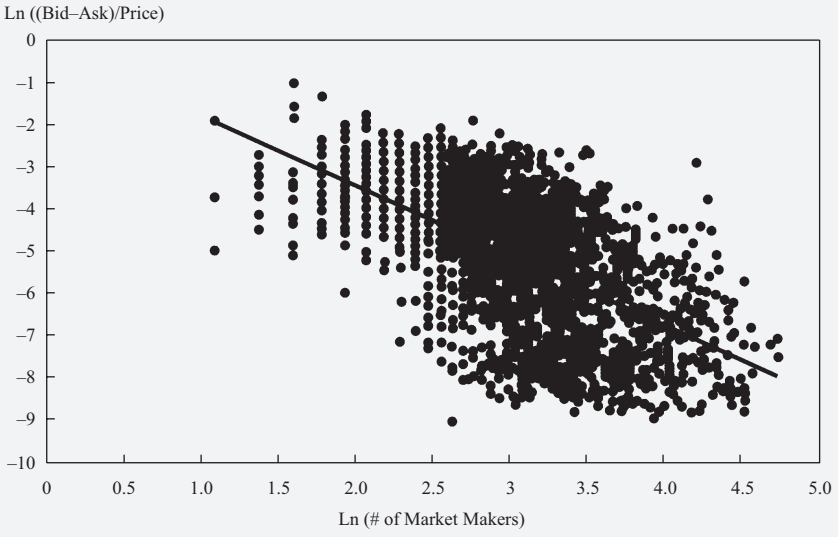
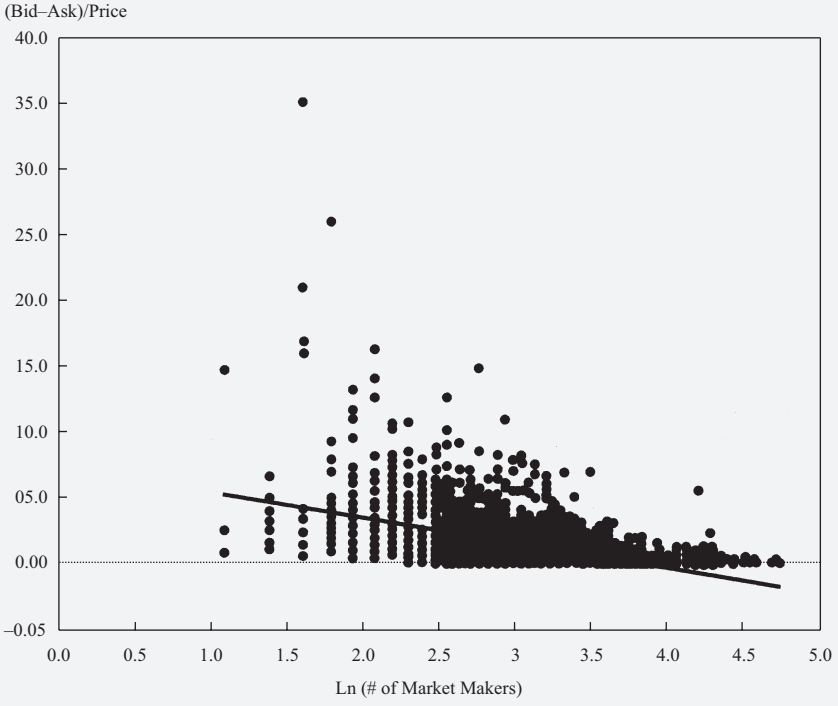


Figure 5 Linear Regression When Two Variables Have a Nonlinear Relation



If we do not take log of the ratio (Bid–ask spread)/Price, the plot is not linear. Figure 5 shows a plot of the natural logarithm of the number of market makers for a stock (on the *X* axis) and the ratio (Bid–ask spread)/Price expressed as a percentage (on the *Y* axis), as well as a regression line that attempts to show a linear relation between the two variables. We see that the relation between

the two variables is very nonlinear.[60] Consequently, we should not estimate a regression with (Bid–ask spread)/Price as the dependent variable. Consideration of the need to ensure that predicted bid–ask spreads are positive would also lead us to not use (Bid–ask spread)/Price as the dependent variable. If we use the non-transformed ratio (Bid–ask spread)/Price as the dependent variable, the estimated model could predict negative values of the bid–ask spread. This result would be nonsensical; in reality, no bid–ask spread is negative (it is hard to motivate traders to simultaneously buy high and sell low), so a model that predicts negative bid–ask spreads is certainly misspecified.[61] We illustrate the problem of negative values of the predicted bid–ask spreads now.

Table 13 shows the results of a regression with (Bid–ask spread)/ Price as the dependent variable and the natural logarithm of the number of market makers and the natural logarithm of the company's market capitalization as the independent variables.

Table 13 Results from Regressing Bid–Ask Spread/Price on ln(Number of Market Makers) and ln(Market Cap)

	Coefficients	Standard Error	*t*-Statistic
Intercept	0.0674	0.0035	19.2571
ln(Number of NASDAQ market makers)	−0.0142	0.0012	−11.8333
ln(Company's market cap)	−0.0016	0.0002	−8.0000

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	2	0.1539	0.0770	392.3338	0.00
Residual	2,584	0.5068	0.0002		
Total	2,586	0.6607			
Residual standard error			0.0140		
Multiple *R*-squared			0.2329		
Observations			2,587		

Source: Center for Research in Security Prices, University of Chicago.

1 Suppose that for a particular NASDAQ-listed stock, the number of market makers is 50 and the market capitalization is $6 billion. What is the predicted ratio of bid–ask spread to price for this stock based on the above model?

Solution to 1:

The natural log of the number of market makers equals ln 50 = 3.9120 and the natural log of the stock's market capitalization (in millions) is ln 6,000 = 8.6995. In this case, the predicted ratio of bid–ask spread to price is 0.0674 + (−0.0142 ×

60 The relation between (Bid–ask spread)/Price and ln(Market cap) is also nonlinear, while the relation between ln(Bid–ask spread)/Price and ln(Market cap) is linear. We omit these scatterplots to save space.

61 In our data sample, the bid–ask spread for each of the 2,587 companies is positive.

3.9120) + (−0.0016 × 8.6995) = −0.0021. Therefore, the model predicts that the ratio of bid–ask spread to stock price is −0.0021 or −0.21 percent of the stock price.

2 Does the predicted bid–ask spread for the above stock make sense? If not, how could this problem be avoided?

Solution to 2:

The predicted bid–ask spread is negative, which does not make economic sense. This problem could be avoided by using log of (Bid–ask spread)/Price as the dependent variable.[62]

Often, analysts must decide whether to scale variables before they compare data across companies. For example, in financial statement analysis, analysts often compare companies using **common size statements**. In a common size income statement, all the line items in a company's income statement are divided by the company's revenues. Common size statements make comparability across companies much easier. An analyst can use common size statements to quickly compare trends in gross margins (or other income statement variables) for a group of companies.

Issues of comparability also appear for analysts who want to use regression analysis to compare the performance of a group of companies. Example 12 illustrates this issue.

EXAMPLE 12

Scaling and the Relation between Cash Flow from Operations and Free Cash Flow

Suppose an analyst wants to explain free cash flow to the firm as a function of cash flow from operations in 2001 for 11 family clothing stores in the United States with market capitalizations of more than $100 million as of the end of 2001.

To investigate this issue, the analyst might use free cash flow as the dependent variable and cash flow from operations as the independent variable in single-independent-variable linear regression. Table 14 shows the results of that regression. Note that the *t*-statistic for the slope coefficient for cash flow from operations is quite high (6.5288), the significance level for the *F*-statistic for the regression is very low (0.0001), and the *R*-squared is quite high. We might be tempted to believe that this regression is a success and that for a family clothing store, if cash flow from operations increased by $1.00, we could confidently predict that free cash flow to the firm would increase by $0.3579.

Table 14 Results from Regressing the Free Cash Flow on Cash Flow from Operations for Family Clothing Stores

	Coefficients	Standard Error	*t*-Statistic
Intercept	0.7295	27.7302	0.0263
Cash flow from operations	0.3579	0.0548	6.5288

62 Whether the natural log of the percentage bid–ask spread, *Y*, is positive or negative, the percentage bid–ask spread found as e^Y is positive, because a positive number raised to any power is positive. The constant *e* is positive ($e \approx 2.7183$).

Table 14 (Continued)

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	1	245,093.7836	245,093.7836	42.6247	0.0001
Residual	9	51,750.3139	5,750.0349		
Total	10	296,844.0975			
Residual standard error			75.8290		
Multiple *R*-squared			0.8257		
Observations			11		

Source: Compustat.

But is this specification correct? The regression does not account for size differences among the companies in the sample.

We can account for size differences by using common size cash flow results across companies. We scale the variables by dividing cash flow from operations and free cash flow to the firm by the company's sales before using regression analysis. We will use (Free cash flow to the firm/Sales) as the dependent variable and (Cash flow from operations/Sales) as the independent variable. Table 15 shows the results of this regression. Note that the *t*-statistic for the slope coefficient on (Cash flow from operations/Sales) is 1.6262, so it is not significant at the 0.05 level. Note also that the significance level of the *F*-statistic is 0.1383, so we cannot reject at the 0.05 level the hypothesis that the regression does not explain variation in (Free cash flow/Sales) among family clothing stores. Finally, note that the *R*-squared in this regression is much lower than that of the previous regression.

Table 15 Results from Regressing the Free Cash Flow/Sales on Cash Flow from Operations/Sales for Family Clothing Stores

	Coefficient	Standard Error	*t*-Statistic
Intercept	–0.0121	0.0221	–0.5497
Cash flow from operations/Sales	0.4749	0.2920	1.6262

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	1	0.0030	0.0030	2.6447	0.1383
Residual	9	0.0102	0.0011		
Total	10	0.0131			
Residual standard error		0.0336			
Multiple *R*-squared		0.2271			
Observations		11			

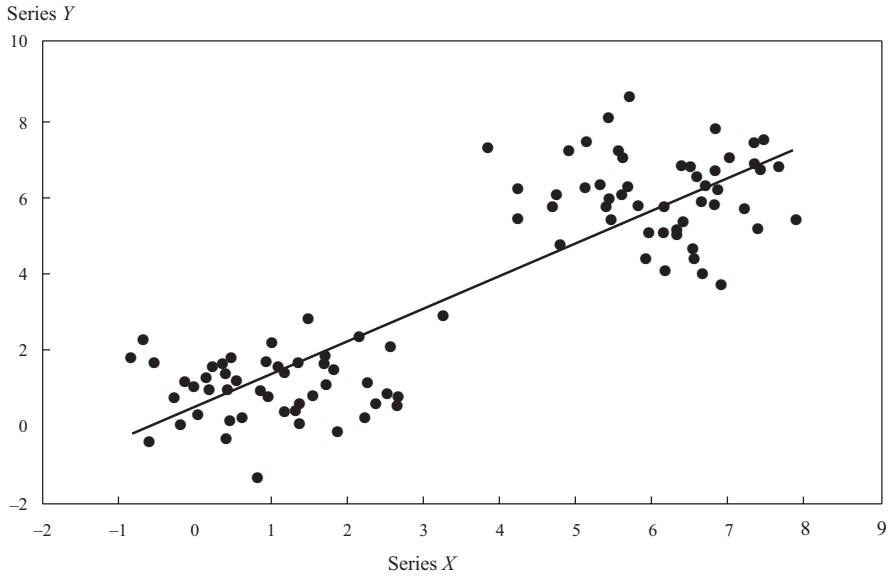
Source: Compustat.

Which regression makes more sense? Usually, the scaled regression makes more sense. We want to know what happens to free cash flow (as a fraction of sales) if a change occurs in cash flow from operations (as a fraction of sales).

Without scaling, the results of the regression can be based solely on scale differences across companies, rather than based on the companies' underlying economics.

A third common form of misspecification in regression models is pooling data from different samples that should not be pooled. This type of misspecification can best be illustrated graphically. Figure 6 shows two clusters of data on variables *X* and *Y*, with a fitted regression line. The data could represent the relationship between two financial variables at two different time periods, for example.

Figure 6 Plot of Two Series with Changing Means



In each cluster of data on *X* and *Y*, the correlation between the two variables is virtually 0. Because the means of both *X* and *Y* are different for the two clusters of data in the combined sample, *X* and *Y* are highly correlated. The correlation is spurious (misleading), however, because it reflects differences in the relationship between *X* and *Y* during two different time periods.

5.3 Time-Series Misspecification (Independent Variables Correlated with Errors)

In the previous section, we discussed the misspecification that arises when a relevant independent variable is omitted from a regression. In this section, we discuss problems that arise from the kinds of variables included in the regression, particularly in a time-series context. In models that use time-series data to explain the relations among different variables, it is particularly easy to violate Regression Assumption 3, that the error term has mean 0, conditioned on the independent variables. If this assumption is violated, the estimated regression coefficients will be biased and inconsistent.

Three common problems that create this type of time-series misspecification are:

- including lagged dependent variables as independent variables in regressions with serially correlated errors;

- including a function of a dependent variable as an independent variable, sometimes as a result of the incorrect dating of variables; and
- independent variables that are measured with error.

The next examples demonstrate these problems.

Suppose that an analyst includes the first lagged value of the dependent variable in a multiple regression that, as a result, has significant serial correlation in the errors. For example, the analyst might use the regression equation

$$Y_t = b_0 + b_1 X_{1t} + b_2 Y_{t-1} + \varepsilon_t \quad (8)$$

Because we assume that the error term is serially correlated, by definition the error term is correlated with the dependent variable. Consequently, the lagged dependent variable, Y_{t-1}, will be correlated with the error term, violating the assumption that the independent variables are uncorrelated with the error term. As a result, the estimates of the regression coefficients will be biased and inconsistent.

EXAMPLE 13

Fisher Effect with a Lagged Dependent Variable

In our discussion of serial correlation, we concluded from a test using the Durbin–Watson test that the error term in the Fisher effect equation (Equation 5) showed positive (first-order) serial correlation, using three-month T-bill returns as the dependent variable and inflation expectations of professional forecasters as the independent variable. Observations on the dependent and independent variables were quarterly. Table 16 modifies that regression by including the previous quarter's three-month T-bill returns as an additional independent variable.

Table 16 Results from Regressing T-Bill Returns on Predicted Inflation and Lagged T-Bill Returns

	Coefficient	Standard Error	*t*-Statistic
Intercept	–0.0005	0.0014	–0.3571
Inflation prediction	0.1843	0.0455	4.0505
Lagged T-bill return	0.8796	0.0295	29.8169
Residual standard error	0.0095		
Multiple *R*-squared	0.9285		
Observations	181		

Source: Federal Reserve Bank of Philadelphia, US Department of Commerce.

At first glance, these regression results look very interesting—the coefficient on the lagged T-bill return appears to be highly significant. But on closer consideration, we must ignore these regression results, because the regression is fundamentally misspecified. As long as the error term is serially correlated, including lagged T-bill returns as an independent variable in the regression will cause all the coefficient estimates to be biased and inconsistent. Therefore, this regression is not usable for either testing a hypothesis or for forecasting.

A second common time-series misspecification in investment analysis is to forecast the past. What does that mean? If we forecast the future (say we predict at time t the value of variable Y in period $t + 1$), we must base our predictions on information we knew at time t. We could use a regression to make that forecast using the equation

$$Y_{t+1} = b_0 + b_1X_{1t} + \varepsilon_{t+1} \quad \textbf{(9)}$$

In this equation, we predict the value of Y in time $t + 1$ using the value of X in time t. The error term, ε_{t+1}, is unknown at time t and thus should be uncorrelated with X_{1t}.

Unfortunately, analysts sometimes use regressions that try to forecast the value of a dependent variable at time $t + 1$ based on independent variable(s) that are functions of the value of the dependent variable at time $t + 1$. In such a model, the independent variable(s) would be correlated with the error term, so the equation would be misspecified. As an example, an analyst may try to explain the cross-sectional returns for a group of companies during a particular year using the market-to-book ratio and the market capitalization for those companies at the end of the year.[63] If the analyst believes that such a regression predicts whether companies with high market-to-book ratios or high market capitalizations will have high returns, the analyst is mistaken. This is because for any given period, the higher the return during the period, the higher the market capitalization and the market-to-book period will be at the end of the period. So in this case, if all the cross-sectional data come from period $t + 1$, a high value of the dependent variable (returns) actually causes a high value of the independent variables (market capitalization and the market-to-book ratio), rather than the other way around. In this type of misspecification, the regression model effectively includes the dependent variable on both the right-and left-hand sides of the regression equation.

The third common time-series misspecification arises when an independent variable is measured with error. Suppose a financial theory tells us that a particular variable X_t, such as expected inflation, should be included in the regression model. But we cannot directly observe X_t; instead, we can observe actual inflation, $Z_t = X_t + u_t$, where we assume u_t is an error term that is uncorrelated with X_t. Even in this best of circumstances, using Z_t in the regression instead of X_t will cause the regression coefficient estimates to be biased and inconsistent. To see why, assume we want to estimate the regression

$$Y_t = b_0 + b_1X_t + \varepsilon_t$$

but we substitute Z_t for X_t. Then we would estimate

$$Y_t = b_0 + b_1Z_t + (-b_1u_t + \varepsilon_t)$$

But $Z_t = X_t + u_t$, Z_t is correlated with the error term $(-b_1u_t + \varepsilon_t)$. Therefore, our estimated model violates the assumption that the error term is uncorrelated with the independent variable. Consequently, the estimated regression coefficients will be biased and inconsistent.

EXAMPLE 14

The Fisher Effect with Measurement Error

Recall from Example 8 on the Fisher effect that based on our initial analysis in which we did not correct for heteroskedasticity and serial correlation, we rejected the hypothesis that three-month T-bill returns moved one-for-one with expected inflation.

63 "Market-to-book ratio" is the ratio of price per share divided by book value per share.

Table 6 Results from Regressing T-Bill Returns on Predicted Inflation (repeated)

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0116	0.0033	3.5152
Inflation prediction	1.1744	0.0761	15.4323
Residual standard error	0.0223		
Multiple *R*-squared	0.5708		
Observations	181		
Durbin–Watson statistic	0.2980		

Source: Federal Reserve Bank of Philadelphia, US Department of Commerce.

What if we used actual inflation instead of expected inflation as the independent variable? Note first that

$$\pi = \pi^e + \nu$$

where

π = actual rate of inflation
π^e = expected rate of inflation
ν = the difference between actual and expected inflation

Because actual inflation measures expected inflation with error, the estimators of the regression coefficients using T-bill yields as the dependent variable and actual inflation as the independent variable will not be consistent.[64]

Table 17 shows the results of using actual inflation as the independent variable. The estimates in this table are quite different from those presented in the previous table. Note that the slope coefficient on actual inflation is much lower than the slope coefficient on predicted inflation in the previous regression. This result is an illustration of a general proposition: In a single-independent-variable regression, if we select a version of that independent variable that is measured with error, the estimated slope coefficient on that variable will be biased toward 0.[65]

Table 17 Results from Regressing T-Bill Returns on Actual Inflation

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0227	0.0034	6.6765
Actual inflation	0.8946	0.0761	11.7556
Residual standard error	0.0267		

(continued)

64 A consistent estimator is one for which the probability of estimates close to the value of the population parameter increases as sample size increases.

65 This proposition does not generalize to regressions with more than one independent variable. Of course, we ignore serially-correlated errors in this example, but because the regression coefficients are inconsistent (due to measurement error), testing or correcting for serial correlation is not worthwhile.

Table 17 (Continued)

	Coefficient	Standard Error	*t*-Statistic
Multiple *R*-squared	0.4356		
Observations	181		

Source: Federal Reserve Bank of Philadelphia, US Department of Commerce.

5.4 Other Types of Time-Series Misspecification

By far the most frequent source of misspecification in linear regressions that use time series from two or more different variables is nonstationarity. Very roughly, **nonstationarity** means that a variable's properties, such as mean and variance, are not constant through time. We will postpone our discussion about stationarity to the reading on time-series analysis, but we can list some examples in which we need to use stationarity tests before we use regression statistical inference.[66]

- Relations among time series with trends (for example, the relation between consumption and GDP).
- Relations among time series that may be **random walks** (time series for which the best predictor of next period's value is this period's value). Exchange rates are often random walks.

The time-series examples in this reading were carefully chosen such that nonstationarity was unlikely to be an issue for any of them. But nonstationarity can be a very severe problem for analyzing the relations among two or more time series in practice. Analysts must understand these issues before they apply linear regression to analyzing the relations among time series. Otherwise, they may rely on invalid statistical inference.

6 MODELS WITH QUALITATIVE DEPENDENT VARIABLES

Financial analysts often need to be able to explain the outcomes of a qualitative dependent variable. **Qualitative dependent variables** are dummy variables used as dependent variables instead of as independent variables.

For example, to predict whether or not a company will go bankrupt, we need to use a qualitative dependent variable (bankrupt or not) as the dependent variable and use data on the company's financial performance (e.g., return on equity, debt-to-equity ratio, or debt rating) as independent variables. Unfortunately, linear regression is not the best statistical method to use for estimating such a model. If we use the qualitative dependent variable bankrupt (1) or not bankrupt (0) as the dependent variable in a regression with financial variables as the independent variables, the predicted value of the dependent variable could be much greater than 1 or much lower than 0. Of course, these results would be invalid. The probability of bankruptcy (or of anything, for that matter) cannot be greater than 100 percent or less than 0 percent. Instead of a linear regression model, we should use probit, logit, or discriminant analysis for this kind of estimation.

66 We include both unit root tests and tests for cointegration in the term "stationarity tests."

Probit and logit models estimate the probability of a discrete outcome given the values of the independent variables used to explain that outcome. The **probit model**, which is based on the normal distribution, estimates the probability that $Y = 1$ (a condition is fulfilled) given the value of the independent variable X. The **logit model** is identical, except that it is based on the logistic distribution rather than the normal distribution.[67] Both models must be estimated using maximum likelihood methods.[68]

Another technique to handle qualitative dependent variables is **discriminant analysis**. In his Z-score and Zeta analysis, Altman (1968, 1977) reported on the results of discriminant analysis. Altman uses financial ratios to predict the qualitative dependent variable bankruptcy. Discriminant analysis yields a linear function, similar to a regression equation, which can then be used to create an overall score. Based on the score, an observation can be classified into the bankrupt or not bankrupt category.

Qualitative dependent variable models can be useful not only for portfolio management but also for business management. For example, we might want to predict whether a client is likely to continue investing in a company or to withdraw assets from the company. We might also want to explain how particular demographic characteristics might affect the probability that a potential investor will sign on as a new client, or evaluate the effectiveness of a particular direct-mail advertising campaign based on the demographic characteristics of the target audience. These issues can be analyzed with either probit or logit models.

EXAMPLE 15

Explaining Analyst Coverage

Suppose we want to investigate what factors determine whether at least one analyst covers a company. We can employ a probit model to address the question. The sample consists of 4,619 observations on public companies in 2013.

The variables in the probit model are as follows:

ANALYSTS = the discrete dependent variable, which takes on a value of 1 if at least one analyst covers the company and a value of 0 if no analysts cover the company

LNVOLUME = the natural log of the company's trading volume in the last month of the year

LNMV = the natural log of the market value of the company's equity

MATURITY = the mix of the company's earned and contributed capital, i.e., retained earnings as a proportion of total equity (RE/TE)

DIVPAYER = a dummy independent variable that takes on a value of 1 if the company paid a dividend

In this attempt to explain analyst coverage, we are examining whether more liquid companies, as captured by the trading volume in their shares, and larger companies, as captured by their market values, are more likely to be followed by at least one analyst. We also examine whether more mature and well established

67 The logistic distribution $e^{(b_0+b_1X)}/\left[1+e^{(b_0+b_1X)}\right]$ is easier to compute than the cumulative normal distribution. Consequently, logit models gained popularity when computing power was expensive.

68 For more on probit and logit models, see Greene (2018).

firms, as reflected in their mix of earned and contributed capital and their status as dividend payers, are more likely to be followed by an analyst. Table 18 shows the results of the probit estimation.

Table 18 Explaining Analyst Coverage Using a Probit Model

	Coefficient	Standard Error	*t*-Statistic
Intercept	2.5066	0.1005	24.9413
LNVOLUME	–0.0221	0.0173	–1.2775
LNMV	0.2441	0.0177	13.7910
MATURITY	0.0011	0.0007	1.5714
DIVPAYER	0.2798	0.0468	5.9786
Percent correctly predicted		78.00	

Source: I/B/E/S from Thomson Reuters, Center for Research in Security Prices at the University of Chicago, and S&P Capital IQ/Compustat.

As Table 18 shows, two coefficients (besides the intercept) have *t*-statistics with an absolute value greater than 2.0. The coefficient on LNMV has a *t*-statistic of 13.7910. That value is far above the critical value at the 0.05 level for the *t*-statistic (1.96), so we can reject at the 0.05 level of significance the null hypothesis that the coefficient on LNMV equals 0, in favor of the alternative hypothesis that the coefficient is not equal to 0. The second coefficient with an absolute value greater than 2 is DIVPAYER, which has a *t*-statistic of 5.9786. We can also reject at the 0.05 level of significance the null hypothesis that the coefficient on DIVPAYER is equal to 0, in favor of the alternative hypothesis that the coefficient is not equal to 0.

Neither of the two remaining independent variables is statistically significant at the 0.05 level in this probit analysis. That is, neither one reaches the critical value of 1.96 needed to reject the null hypothesis (that the associated coefficient is significantly different from 0). This result shows that once we take into account a company's market value and whether it pays dividends, the other factors—trading volume and maturity—have no power to explain whether at least one analyst will cover the company.

7 MACHINE LEARNING

Earlier sections explained the multiple linear regression model and showed how it can be applied to analyze relatively small and well-organized financial datasets. The variables in Example 3, for instance, were parallel time series of observations on returns for just three variables—an equity investment and two equity indexes. A set of tools broader than multiple linear regression is needed to extract information from the prodigious quantities of data being generated in real time by financial markets, businesses, governments, individuals, and sensors (e.g., by satellite imaging). Investors are leveraging such data—Big Data (a common label for complex and voluminous data)—to glean insights and information that can be used in investing.

The analysis of Big Data must deal with several challenges. Big Data comprises data generated from both traditional and non-traditional sources, and many of the data from non-traditional sources are typically unstructured—not recognizably organized—at

least when initially captured. The number and types of relationships among the data generally cannot be related to theory as with many examples from earlier sections. By applying computer algorithms and flexible models to Big Data, however, relationships may be discovered or learned. This section will provide a top-level orientation to computer-age techniques that come under the heading of *machine learning*. The emphasis will be on extracting information from data—data analysis, often referred to today as "data analytics."[69]

7.1 Major Focuses of Data Analytics

An understanding of the types of problems addressed by data analytics is helpful for a discussion of machine learning. Data analytics generally has one of six focuses:

1 *Measuring correlations.* Correlation focuses on understanding the contemporaneous relationship between variables. For example, an investment analyst may examine how pairs of financial variables tend to covary.

2 *Making predictions.* Prediction focuses on whether one or more variables can help forecast the value of some variable of interest. For example, can posts from social media predict the direction of a stock price? Identifying predictive relationships may give an investment edge if asset prices do not reflect such relationships.

3 *Making causal inferences.* **Causal inference** focuses on establishing that a change in an independent variable causes a change in the dependent variable. A causal relationship implies an underlying mechanism connecting the causal variable to the outcome variable and is a stronger relationship between two variables than that of correlation or prediction. In real-world contexts, causal inference is challenging because of the difficulty of controlling for confounding variables (variables that might influence both dependent and independent variables), among other problems. Questions such as whether financial shocks or government policy intervention can be linked in a causal manner to changes in real economic variables have been investigated using research methods inspired by the methods of controlled experiments.

4 *Classifying data.* **Classification** focuses on sorting observation into distinct categories. In a regression, when the dependent variable is categorical (not continuous), the econometric model relating the outcome to the independent variables is called a "classifier." Many classification models are binary classifiers, as in the case of fraud detection for credit card transactions (where the dependent variable is 1 if the transaction is fraudulent or 0 if non-fraudulent). Multicategory classification is not uncommon, as in the case of classifying firms into multiple credit rating categories. In assigning ratings, the outcome variable is ordinal, meaning the categories are ordered (e.g., from low to high creditworthiness). Ordinal variables are intermediate between categorical variables and continuous variables on a scale of measurement.

5 *Sorting data into clusters.* **Clustering** focuses on sorting observations into groups (clusters) such that observations in the same cluster are more similar to each other than they are to observations in other clusters. Groups are formed on the basis of a set of criteria that may or may not be prespecified. For

69 For further reading at an introductory level, see Theobald (2017).

example, clustering has been used by asset managers to sort companies into empirically determined groupings rather than conventional groupings based on sectors or countries.

6 *Reducing the dimension of data.* **Dimension reduction** focuses on reducing the number of independent variables while retaining variation across observations (to preserve the information contained in that variation). Dimension reduction may have several purposes. For prediction, simpler models tend to work better than complex models for out-of-sample forecasting.[70] Dimension reduction may be applied to data with a large number of attributes to produce a lower dimensional representation (i.e., with fewer attributes) that can fit, for example, on a computer screen. Dimension reduction is important in many quantitative investment and risk management contexts in which the objective is to identify the major factors underlying asset price movements.

Methodologies for addressing these types of problems have been part of the investment practitioner's statistical toolkit for decades and, in several instances, overlap with techniques employed in machine learning. Characteristically, however, machine learning methods address such problems as prediction, classification, clustering, and dimension reduction in large and unstructured datasets that contain many features (variables). To explain further, machine learning needs to be defined.

7.2 What Is Machine Learning?

Machine learning lies within the broader field of artificial intelligence, in which machines display intelligent decision-making ability through such activities as sensing, reasoning, and understanding and speaking languages. **Machine learning** (ML) comprises diverse approaches by which computers are programmed to improve performance in specified tasks with experience. A technical definition, germane to the scope of tools covered in this introduction, was provided in the classic text on the subject by Mitchell (1997):[71]

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.

As one example, ML has been applied to predict whether credit card transactions are fraudulent or legitimate. In this case, the terms may be defined as follows:

T (task): Predict whether credit card transactions are fraudulent or not

P (performance measure): Percentage of transactions correctly predicted[72]

E (experience): The database of credit card transactions that are labeled as actually fraudulent or not

The ML program can be said to learn from experience if the percentage of correctly predicted credit card transactions (P) increases as experience (E), the input from the growing database as described, increases.

70 This statement is an informal reference to the problem of overfitting, which complex models tend to be subject to. Overfitting will be explained later in the reading.

71 The term "machine learning" dates to Samuel (1959), who defined machine learning more narrowly, as algorithms that learn how to complete specific tasks without being explicitly programmed to do so (to paraphrase his definition).

72 The measure "percentage of transactions correctly predicted" is given for the sake of a simple illustration; more robust performance measures are available—for example, involving cross-tabulation of the actual (labeled) versus predicted (by ML) outcomes. Note too that although the objective in this example is to predict whether a credit card is being used fraudulently, this problem type would be called a "classification problem" in ML terminology, as discussed later.

7.3 Types of Machine Learning

With a focus on data analytics, machine learning is broadly divided into two distinct classes of techniques: supervised learning and unsupervised learning.

Supervised learning is machine learning that makes use of labeled training data. (The word "supervised" relates to supplying the labeled data.) In the credit card example, the ML program is given processed transactions labeled (tagged) "fraudulent" or "non-fraudulent" and uses them to train a model in predicting fraud more accurately in new transactions. A (usually non-linear) function relates attributes of inputs to the output (the target variable). More formally, *supervised learning is the process of training an algorithm to take a set of inputs X and find a model that best relates them to the output Y.* In the credit card example, the target variable is a binary variable, with a value of 1 for "fraudulent" or 0 for "non-fraudulent." However, the target can also be multicategory, ordinal, or continuous. Typical data analysis tasks associated with supervised learning are classification and prediction.

Unsupervised learning is machine learning that does not make use of labeled training data. More formally, in unsupervised learning, we have inputs X that are used for analysis without any targets Y being supplied. In unsupervised learning, because the ML program is not given tagged data, the ML program has to discover structure within the data themselves. Clustering is an example of data analytics to which unsupervised learning is applied. For example, we may take different firms' financial statement data and use an unsupervised ML program to cluster firms into groups based on their attributes. Each cluster will contain firms that have greater overall similarity to each other than they do to firms in other clusters.

Experts sometimes distinguish additional categories, such as deep learning, which will be briefly described later, and **reinforcement learning**, in which a computer learns from interacting with itself (or data generated by the same algorithm).[73]

Machine Learning Vocabulary

Machine learning comes with its own vocabulary, which differs from the terminology used earlier in this reading and in statistical modeling more generally. In regression analysis, the *Y* variable is known as the dependent variable, whereas in machine learning, it is called the **target variable** or **tag variable**. Whereas the *X* variables are known as independent variables or explanatory variables in regression analysis, they are called **features** in machine learning. Curating a dataset of features for ML processing is known as **feature engineering** by machine learning practitioners. In this section, we introduce machine learning vocabulary to a limited extent to control the amount of new vocabulary that must be learned.

73 An example of reinforcement learning is the program AlphaGoZero. Starting with only the rules of the game Go, by playing against itself the program learned to play Go at an exceptionally high master level. Because this class of learning techniques may not even use data, we exclude it from our discussion of ML in data analytics.

EXAMPLE 16

Machine Learning

1 Which of the following *best* describes machine learning? Machine learning:
   - **A** is a type of computer algorithm.
   - **B** is a set of computer-driven approaches that can be used to extract information from Big Data.
   - **C** is a set of computer-driven approaches adapted to extracting information from structured data.

2 Which of the following statements is *most* accurate? Machine learning:
   - **A** contrasts with human learning in relation to measuring performance on specific tasks.
   - **B** takes place when a computer program is programmed to perform specific tasks.
   - **C** takes place when a computer improves performance in a specific class of tasks as experience increases.

3 Which of the following statements is *most* accurate? When attempting to place data into groups based on their inherent similarities and differences:
   - **A** an unsupervised ML algorithm is used.
   - **B** an ML algorithm that is given tagged data is used.
   - **C** an ML algorithm that is given tagged data and untagged data is used.

4 Which of the following statements concerning supervised learning *best* distinguishes it from unsupervised learning? Supervised learning involves:
   - **A** training on labeled data.
   - **B** training on unlabeled data.
   - **C** learning from unlabeled data.

Solution to 1:

B is correct. A major application of machine learning is extracting information from Big Data. Choice A is not correct because although algorithms are used in machine learning, machine learning itself is not best described as a type of computer algorithm.

Solution to 2:

C is correct. ML takes place when a computer improves performance in a specific class of tasks as experience increases.

Solution to 3:

A is correct. The beginning of the statement that must be completed is a description of clustering. Unsupervised ML algorithms are used in clustering.

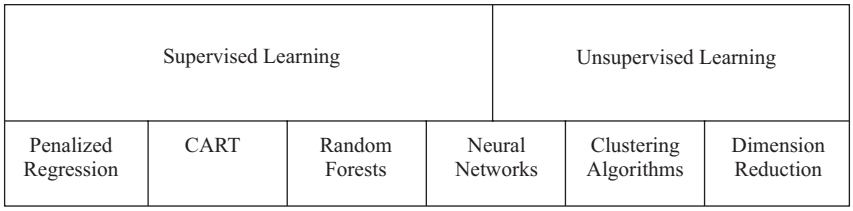
Solution to 4:

A is correct. Supervised learning computer programs are given labeled data in training, in contrast to unsupervised learning computer programs.

7.4 Machine Learning Algorithms

Having provided a top-level view of machine learning, including the distinction between supervised and unsupervised machine learning, the following sections provide a top-level description of a limited selection of important models and procedures. Figure 7 categorizes the methods selected. Neural networks are commonly used for (and are covered here under) supervised learning but are also important in reinforcement learning, which can be unsupervised.

Figure 7 Machine Learning Algorithms



7.4.1 *Supervised Learning*

Supervised learning can be divided into two categories: *regression* and *classification*. In the context of supervised learning, the distinction between regression and classification is determined by the nature of the Y variable. If the Y variable is continuous, then the task is one of regression (even if the ML technique used is not "regression" as covered earlier in the reading). If the Y variable is categorical or ordinal (e.g., determining a firm's rating), then it is a classification problem. Regression and classification use different ML techniques.

For regression problems, techniques include linear and non-linear models. In the following, we briefly describe an adaptation of linear models often used in machine learning for prediction problems. To illustrate classification, from the wide variety of classifiers—classification techniques—we have chosen classification and regression trees (CART), random forests, and neural networks for brief descriptions.

For brevity, in the following discussion, assume we have a number of observations of a target variable Y and n real-valued variables $X_1, ..., X_n$ that we may use to establish a relationship (regression or classification) between X (a vector of the X_i) and Y for each observation in our dataset. With this background established, we now provide brief descriptions of these selected techniques.

7.4.1.1 Penalized Regression Penalized regression is a computationally efficient technique used in prediction problems. As with multiple linear regression, penalized regression and many other forms of linear regression are subsumed as special cases of the generalized linear model (GLM). GLM refers to a flexible specification linear regression in which the modeler, by choice of parameters, can express preferences for model parsimony and how model fit is calibrated. As the special case of GLM to describe here, we take penalized regression, which has been found useful in practice for reducing a large number of independent variables to a manageable set and for making good predictions in a variety of large datasets. Our limited aim is to show how model complexity can be managed in the context of a linear regression model.

In a large dataset context, we may have many variables that potentially could be used to "explain" or model Y. When a model is fit to data, the model may so closely reflect the characteristics of those specific data that the model does not perform well on new data; included variables can reflect what is "noise" in a specific dataset that will not be present in future data. That is the problem of **overfitting**, and penalized regression could be described as a technique of regularization. **Regularization** "describes almost

any method that tamps down statistical variability in high-dimensional estimation or prediction problems."[74] In prediction, only out-of-sample performance matters, and relatively parsimonious models tend to work well because they are less subject to overfitting.

Conceptually, the idea underlying penalized regression is not complicated. Let us suppose we preprocess the data so the variables have a mean of 0 and a variance of 1. That standardization of variables will allow us to compare the magnitudes of regression coefficients for the independent variables.[75] In ordinary linear regression, the regression coefficients $\hat{b}_0$, $\hat{b}_1$, ..., $\hat{b}_K$ are chosen to minimize the sum of the squared residuals. In penalized regression, the regression coefficients are chosen to minimize the sum of squared residuals plus a penalty term that increases in size with the number of included variables with non-zero regression coefficients. Given this penalty, only the more important variables for explaining Y remain in the model. In one popular type of penalized regression, LASSO (least absolute shrinkage and selection operator), the penalty term has the following form, with $\lambda > 0$:

$$\text{Penalty term} = \lambda \sum_{k=1}^{K} \left| \hat{b}_k \right|.$$

This expression involves summing the absolute values of the regression coefficients. The greater the number of included variables (i.e., variables with non-zero coefficients), the larger the penalty, so in a penalized regression, an included variable has to make a sufficient contribution to model fit to offset the penalty from including it. All types of penalized regression involve a trade-off of this type.

7.4.1.2 Classification and Regression Trees CART is a common supervised ML technique that can be applied to predict either a categorical target variable (a classification problem, the "C" in the name), producing a classification tree, or a continuous outcome (a regression problem, the "R" in the name), producing a regression tree. The technique is computationally efficient and adaptable to datasets with complex structures.

Most commonly, CART is applied where the target is binary. We mentioned this kind of problem in Section 6, on regression with a qualitative dependent variable, where probit and logit regression were briefly described. CART is better adapted to classification problems than those models in cases in which there may be significant non-linear relationships among variables.

This algorithm produces decision trees with binary (i.e., two-way) branching to classify observations, using the feature set X. The approach uses a preclassified training dataset—for example, for the credit card transactions example, a dataset of features for transactions that were found to be fraudulent or not. At the top of the tree, the model picks one of the X variables and splits the sample into two using a cutoff value c, designating each group as belonging to one category. After searching over the variables in the feature set, the one that minimizes misclassification error (e.g., by a criterion such as mean-squared error) is chosen as the variable to use for splitting the observations into two groups.[76] The observations within each group have lower error within group than before. In Figure 8, two decision nodes are shown coming from the root node, reflecting this split. A similar partitioning is applied at each decision node, making smaller and smaller partitions and generating an entire "classification tree." At any level of the tree, when the classification error does not diminish much more from another split (bifurcation), the process stops, the node is a terminal node, and the category that is in the majority at that node is assigned to it.

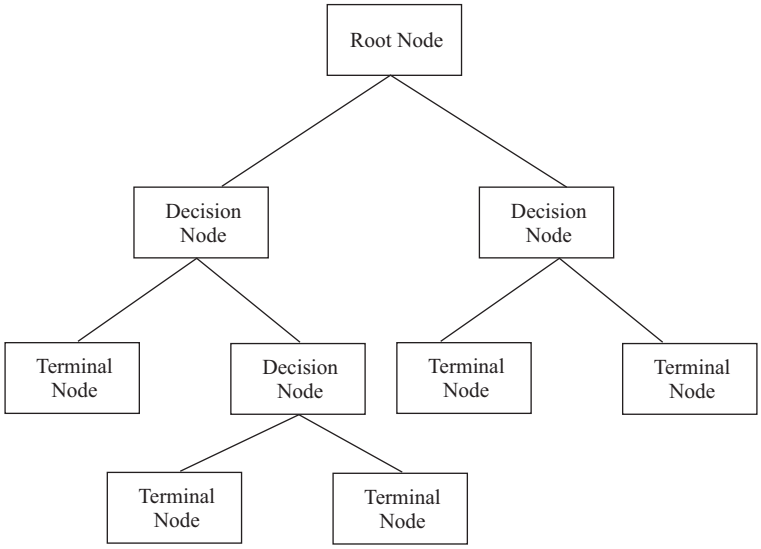
74 Efron and Hastie (2016, p. 101).

75 Standardization was introduced in the CFA Program Level I reading on probability distributions.

76 The cutoff values are similarly chosen to minimize classification error.

In Figure 8, the branching process stops after just two decisions on the right side of the tree but continues somewhat further on the left side. In general, certain regions of the tree may end up being denser than others, but overall, this method works very well and is a staple technique for classification problems. The model also produces a complete decision tree, which is very useful for interpreting how any observation is classified. When the goal of the model is classification not into discrete groups but into continuous values, the model is known as a "regression tree." At each terminal node in a regression tree, we have the mean predicted value of the target variable.

Figure 8 Classification Tree (Simple Case)



7.4.1.3 Random Forests A random forest classifier is a collection of classification trees. Rather than use just one classification tree, we build several, based on random selection of features (variables). At each node, when bifurcation is considered, rather than use all X variables as candidates, we select the best one from a random subset of the n features. Each tree is, therefore, slightly different from the others. For any new observation, we let all the classifier trees (the "random forest") undertake classification by majority vote, implementing a machine learning version of the "wisdom of crowds." The process involved in random forests tends to protect against overfitting on the training data. It also reduces the ratio of noise to signal because errors cancel out across the collection of classification trees.

Ensemble Learning

In making use of voting across classifier trees, random forests are an example of a class of techniques known as **ensemble learning**, in which incorporating the output of a collection of models produces classifications that have better signal-to-noise ratios than the individual classifiers. A good example of a credit card fraud detection problem comes from an open source dataset on Kaggle.[77] Here the data contained several anonymized variables that might be used to explain which transactions were fraudulent. The difficulty in the analysis arises from the fact that the rate of fraudulent transactions is very

77 See https://www.kaggle.com/dalpozz/creditcardfraud.

low; in a sample of 284,807 transactions, only 492 were fraudulent (0.17%). This is akin to finding a needle in a haystack. Applying a random forest classification algorithm with an oversampling technique—which involves increasing the proportional representation of fraudulent data in the data training set—does extremely well, delivering precision of 89% and recall of 82%, despite the lopsided sample.[78]

CART and random forests have been applied to classification problems where the information—such as positive and negative sentiment concerning a company—could be useful for investment or risk management. These techniques have also been used to forecast whether an IPO will be successful given the attributes of the IPO offering and the issuer, as well as in other problems with similar structures.

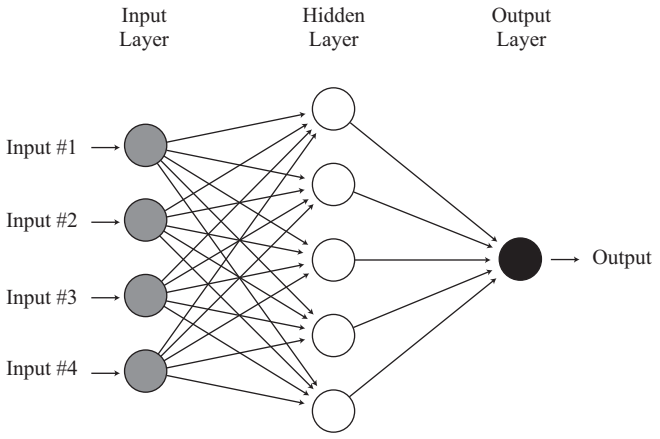
7.4.1.4 Neural Networks Neural networks (also called *artificial neural networks*, or ANNs) have been successfully applied to a variety of tasks characterized by non-linearities and interactions among variables. Neural networks consist of *nodes* (the circles in Figure 9) connected by *links* (the arrows connecting nodes in Figure 9). In the figure, the inputs would be scaled to account for differences in the units of the data. For example, if the inputs were positive numbers, each could be scaled by its maximum value so that their values lie between 0 and 1.

Neural networks have three types of layers: an input layer, hidden layers, and an output layer. The neural network shown has a single hidden layer and one node in the output layer, as might be relevant in the case of a supervised regression problem where the output is a real number (say, a predicted value). The input layer has four nodes, which could correspond to four features used for prediction—a dimension of four. The network shown has five hidden nodes. These three numbers—4, 5, and 1—for the neural network are an example of **hyperparameters**, constants that the human researcher determines to establish the structure of the network.

A link takes a number from a node and transmits it to another node. Now consider any of the nodes to the right of the input layer. These nodes are sometimes called "neurons" because they process information. Take the topmost hidden node. Four links connect to that node from the inputs, so the node gets four values transmitted by the links. The node weights each value received and adds up the weighted values, to which it applies a formula called an **activation function**, which is typically non-linear. In the context of a network in which all the features are interconnected with non-linear activation functions, complex, non-linear relationships among features can be modeled. The hidden layer is shown to feed an output node that generates the predicted value. Learning takes place through improvements in the weights applied at the nodes, as evidenced by improvements in some specified performance measure. For the type of problem described here, labeled training data would assist in training the network.

78 "Precision" is the percentage of cases marked as fraudulent that were actually fraudulent. "Recall" is the percentage of fraudulent cases that were marked as fraudulent. Thus both type I error and type II error are rather low.

Figure 9 A Neural Network with One Hidden Layer



Neural networks with many hidden layers (often more than 20)—known as **deep learning nets** (DLNs)—are the backbone of the artificial intelligence revolution. Advances in DLNs have driven developments in activities, such as image, pattern, and speech recognition.

Deep Learning Nets

To state the operation of DLNs succinctly, they take a set of inputs X from a feature set (the input layer), which are then passed to a layer of non-linear mathematical functions (neurons) with weights w_{ij} (for neuron i and input j), each of which usually produces a scaled number in the range (0, 1) or (–1, 1). These numbers are then passed to another layer of functions and into another, and so on, until the final layer produces a set of probabilities of the observation being in any of the target categories (each represented by a node in the output layer). The DLN assigns the category based on the category with the highest probability. The DLN is trained on large datasets: In training, the weights w_{ij} are determined so as to minimize a specified loss function. DLNs have been shown to be useful in general for pattern recognition problems (e.g., character and image recognition), credit card fraud detection, autonomous cars, and other applications. DLNs have become hugely successful because of a confluence of three developments: (1) advances in analytical methods for fitting these models,[79] (2) the availability of large quantities of data to train models, and (3) fast computers, especially new chips (in the graphics processing unit class) tailored for the type of calculations done on DLNs. Several financial firms are experimenting with DLNs for trading, as well as for automation of many of their internal processes. Hutchinson, Lo, and Poggio (1994) and Culkin and Das (2017) described how DLNs were trained to price options, mimicking the Black–Scholes–Merton equation. In Culkin and Das (2017), the DLN predicts option prices out of sample that are very close to actual option prices: A regression of predicted option prices on actual prices has an R^2 of 99.8%.

79 For more information, see Efron and Hastie (2016, pp. 356–358).

Text Analytics

ML analysis of text data captured from the internet—relating, for example, to consumption data, investor or consumer sentiment, regulatory filings, and social media posts—has been used in risk management and for getting an information edge for short-term trading.[80]

7.4.2 *Unsupervised Learning*

Recall that unsupervised learning is machine inference where there is no target to which we match the feature set. Succinctly, it is analysis on the *X* variables, and there is no *Y* target variable set. Many algorithms of this type of learning address clustering and dimension reduction.

7.4.2.1 Clustering Algorithms Clustering groups data objects solely on the basis of information found in the data. Clustering differs from classification. In classification, data are assigned to classes determined by the researcher (such as "fraudulent" or "non-fraudulent" credit card transactions); in clustering, the groups are determined by the data themselves.[81]

Approaches to clustering are often placed into one of two groups: bottom-up clustering and top-down clustering. Bottom-up clustering starts with each observation being its own cluster and then progressively groups the observations into larger, non-overlapping clusters according to some metric of closeness, such that each observation is always in exactly one cluster. With top-down clustering, we start with all observations belonging to a single cluster, which is then progressively partitioned into smaller and smaller clusters.

The choice of clustering algorithm depends on the nature of the data and the purpose of the analysis and can be evaluated by various metrics, as explained in the technical literature.

The K-means algorithm can be briefly described as an example of a clustering algorithm. This bottom-up algorithm involves two geometric ideas: the idea of a centroid, which is a type of point—one that averages the positions of all points in a set—and the idea of a metric for the distance between two points, such as the Euclidian ("straight-line") distance familiar from geometry. Suppose the researcher wants to group 100 firms into five groups based on two numerical measures of corporate governance quality.[82] These firms can be represented as 100 points scattered in a two-dimensional space. The five centroid points can be initially located randomly. The first step involves running through the 100 points to assign each one to the nearest centroid by the straight-line metric. The location of the centroid points is then recomputed on the basis of the average location of the included firms in each cluster. The firms are then reassigned to centroids, the centroids are recomputed again, and the algorithm continues to iterate until further iterations result in no movement of the centroids.

When it concludes, the K-means algorithm has located the five centroid points such that the average of the (squared) straight-line distances of the 100 points from their nearest centroid is at a (local) minimum. The five clusters accounting for the 100 firms are "optimal" in a sense roughly analogous to minimizing a variance (within

80 For a survey of text analytics in finance, see Das (2014).

81 Clustering can be referred to as *unsupervised classification*.

82 The algorithm involves averaging, so all attributes need to be numerical.

group), but the solution can depend on the initial locations seeded for the centroids.[83] The algorithm is fast and works well even on very large datasets with hundreds of millions of observations.[84]

Clustering is a very important application of machine learning because it uncovers potentially interesting structures in data without the addition of labeled data or any applicable theory. Among portfolio management applications, clustering has been used for improving portfolio diversification relative to using industry classifications.

7.4.2.2 Dimension Reduction The second type of unsupervised learning is dimension reduction. When there are many features in a dataset, representing the data visually or fitting models to the data may become unnecessarily complex and "noisy" in the sense of reflecting random influences specific to a dataset. In such cases, dimension reduction may be necessary. One long-established statistical method for dimension reduction is principal component analysis (PCA). PCA is used to summarize or reduce highly correlated features of data into a few main, uncorrelated composite variables (a **composite variable** is a variable that combines two or more variables that are statistically strongly related to each other). The first principal component is the most volatile: It represents the most important factor for explaining the volatility in the data. The second principal component extracts as much of the remaining volatility as possible, subject to the constraint that it is uncorrelated with the first principal component. Each subsequent principal component extracts remaining volatility, subject to the constraint that it is uncorrelated with the preceding principal components. The research objective decides the rule for how many principal components to retain; each successive component tends to have a lower ratio of information to noise. Attaching an interpretation to a principal component in terms of the features it is capturing is typically not possible, but PCA supplies a lower-dimensional view of the structure of the volatility in data. Historically, dimension reduction using PCA has been used to model the factors in stock market returns and yield curve dynamics. Dimension reduction methods are applied not only to numerical data but also often to textual data and visual data (e.g., in face recognition), which may be in irregular formats.[85]

EXAMPLE 17

Major Types of Machine Learning

1 As used in supervised machine learning, regression problems involve:
   - **A** binary target variables.
   - **B** continuous target variables.
   - **C** categorical target variables.

2 Which of the following *best* describes penalized regression? Penalized regression:
   - **A** is unrelated to multiple linear regression.
   - **B** involves a penalty term for the sum of squared residuals.
   - **C** is a category of general linear models that is used when the number of independent variables is a concern.

3 CART is *best* described as a type of:

83 This potential problem can be addressed by running the algorithm multiple times with different seeds and then choosing the best fit clustering.

84 See https://www.eecs.tufts.edu/~dsculley/papers/fastkmeans.pdf.

85 See Wang, Huang, Wang, and Wang (2014) for more information on dimension reduction techniques beyond standard PCA.

A unsupervised ML.

B a clustering algorithm based on decision trees.

C a supervised ML algorithm that accounts for non-linear relationships among the features.

4 Neural networks are *best* described as an ML technique for learning:

A exactly modeled on the human nervous system.

B based on layers of nodes when the relationships among the features are usually non-linear.

C based on a tree structure of nodes when the relationships among the features are non-linear.

5 Clustering is *best* described as a technique in which:

A the grouping of observations is unsupervised.

B features are grouped into clusters by a top-down algorithm.

C observations are classified according to predetermined labels.

6 Dimension reduction techniques are *best* described as means to reduce a set of features:

A to a manageable size.

B to a manageable size while controlling for variation in the data.

C to a manageable size while retaining as much of the variation in the data as possible.

Solution to 1:

B is correct. When the target variable is binary or categorical, the problem is a *classification problem* rather than a regression problem.

Solution to 2:

C is correct.

Solution to 3:

C is correct.

Solution to 4:

B is correct.

Solution to 5:

A is correct. Choice B is not the best choice because clustering algorithms can be either bottom up or top down.

Solution to 6:

C is correct.

7.5 Supervised Machine Learning: Training

Understanding training in machine learning serves to help one understand both ML and differences between ML and model specification as discussed in Section 5.1 of this reading with respect to multiple linear regression in financial economics contexts.

Model training in machine learning entails the choice of a specific algorithm to learn Y from X and then letting the algorithm fit some function of actual and predicted values through repeated rounds of passes over the data or subsets of the data.

Unlike the emphasis on connecting a model to cogent economic reasoning in many financial economic contexts (the first principle stated in Section 5.1), in ML contexts the emphasis is typically on improving accuracy in classification or prediction.

In a simplified view, the process to train ML models involves following steps:

1 Specify the ML technique/algorithm.

2 Specify the associated hyperparameters (values chosen before training begins); these may include the number of training cycles.

3 Divide data into a
   - **training sample** involving correctly labeled targets, which will be used to train or fit the algorithm, and a
   - **validation sample**, which is used to evaluate how well the model that is fit to the training sample works out of sample.[86]

4 Evaluate learning with performance measure P, using the validation sample, and adjust ("tune") the hyperparameters.

5 Repeat the training cycle the specified number of times or until the required performance level (e.g., level of accuracy) is obtained.

The output or artifact created by the training process is the "ML model."

In choosing how many training cycles to run, more cycles might result in overfitting the model on in-sample data, resulting in poor out-of-sample predictive performance.

In Step 3, training and validation are often accomplished in a repeated process of randomly splitting the dataset into training and validation samples. Following such a process, a data point might be used for training purposes in one split and for validation purposes in another split. Such a process, called **cross-validation**, is intended to control for bias in training data and is completely standard; no model validation is complete without cross-validation. Intuitively, the bigger the dataset, the less cross-validation is needed. With smaller datasets, a specific split of the data into training and validation samples may be biased, and cross-validation is a simple and effective way to address that concern.

SUMMARY

In this reading, we have presented the multiple linear regression model and discussed violations of regression assumptions, model specification and misspecification, and models with qualitative variables.

- The general form of a multiple linear regression model is $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki} + \varepsilon_i$
- We conduct hypothesis tests concerning the population values of regression coefficients using t-tests of the form

$$t = \frac{\hat{b}_j - b_j}{s_{\hat{b}_j}}$$

- The lower the p-value reported for a test, the more significant the result.
- The assumptions of classical normal multiple linear regression model are as follows:

86 When the data are divided, some observations may be set aside as a **hold-out sample** to test the final model arrived at by the training process.

1. A linear relation exists between the dependent variable and the independent variables.
2. The independent variables are not random. Also, no exact linear relation exists between two or more of the independent variables.
3. The expected value of the error term, conditioned on the independent variables, is 0.
4. The variance of the error term is the same for all observations.
5. The error term is uncorrelated across observations.
6. The error term is normally distributed.

- To make a prediction using a multiple linear regression model, we take the following three steps:
  1. Obtain estimates of the regression coefficients.
  2. Determine the assumed values of the independent variables.
  3. Compute the predicted value of the dependent variable.
- When predicting the dependent variable using a linear regression model, we encounter two types of uncertainty: uncertainty in the regression model itself, as reflected in the standard error of estimate, and uncertainty about the estimates of the regression coefficients.
- The *F*-test is reported in an ANOVA table. The *F*-statistic is used to test whether at least one of the slope coefficients on the independent variables is significantly different from 0.

$$F = \frac{\text{RSS}/k}{\text{SSE}/\left[n-\left(k+1\right)\right]} = \frac{\text{Mean regression sum of squares}}{\text{Mean squared error}}$$

  Under the null hypothesis that all the slope coefficients are jointly equal to 0, this test statistic has a distribution of $F_{k,n-(k+1)}$, where the regression has n observations and k independent variables. The *F*-test measures the overall significance of the regression.
- R^2 is nondecreasing in the number of independent variables, so it is less reliable as a measure of goodness of fit in a regression with more than one independent variable than in a one-independent-variable regression.

  Analysts often choose to use adjusted R^2 because it does not necessarily increase when one adds an independent variable.
- Dummy variables in a regression model can help analysts determine whether a particular qualitative independent variable explains the model's dependent variable. A dummy variable takes on the value of 0 or 1. If we need to distinguish among n categories, the regression should include $n - 1$ dummy variables. The intercept of the regression measures the average value of the dependent variable of the omitted category, and the coefficient on each dummy variable measures the average incremental effect of that dummy variable on the dependent variable.
- If a regression shows significant conditional heteroskedasticity, the standard errors and test statistics computed by regression programs will be incorrect unless they are adjusted for heteroskedasticity.
- One simple test for conditional heteroskedasticity is the Breusch–Pagan test. Breusch and Pagan showed that, under the null hypothesis of no conditional heteroskedasticity, nR^2 (from the regression of the squared residuals on the independent variables from the original regression) will be a χ^2 random variable with the number of degrees of freedom equal to the number of independent variables in the regression.

- The principal effect of serial correlation in a linear regression is that the standard errors and test statistics computed by regression programs will be incorrect unless adjusted for serial correlation. Positive serial correlation typically inflates the *t*-statistics of estimated regression coefficients as well as the *F*-statistic for the overall significance of the regression.
- The most commonly used test for serial correlation is based on the Durbin–Watson statistic. If the Durbin–Watson statistic differs sufficiently from 2, then the regression errors have significant serial correlation.
- Multicollinearity occurs when two or more independent variables (or combinations of independent variables) are highly (but not perfectly) correlated with each other. With multicollinearity, the regression coefficients may not be individually statistically significant even when the overall regression is significant as judged by the *F*-statistic.
- Model specification refers to the set of variables included in the regression and the regression equation's functional form. The following principles can guide model specification:
  - The model should be grounded in cogent economic reasoning.
  - The functional form chosen for the variables in the regression should be appropriate given the nature of the variables.
  - The model should be parsimonious.
  - The model should be examined for violations of regression assumptions before being accepted.
  - The model should be tested and be found useful out of sample before being accepted.
- If a regression is misspecified, then statistical inference using OLS is invalid and the estimated regression coefficients may be inconsistent.
- Assuming that a model has the correct functional form, when in fact it does not, is one example of misspecification. There are several ways this assumption may be violated:
  - One or more important variables could be omitted from the regression.
  - One or more of the regression variables may need to be transformed before estimating the regression.
  - The regression model pools data from different samples that should not be pooled.
- Another type of misspecification occurs when independent variables are correlated with the error term. This is a violation of Regression Assumption 3, that the error term has a mean of 0, and causes the estimated regression coefficients to be biased and inconsistent. Three common problems that create this type of time-series misspecification are:
  - including lagged dependent variables as independent variables in regressions with serially correlated errors;
  - including a function of dependent variable as an independent variable, sometimes as a result of the incorrect dating of variables; and
  - independent variables that are measured with error.
- Probit and logit models estimate the probability of a discrete outcome (the value of a qualitative dependent variable, such as whether a company enters bankruptcy) given the values of the independent variables used to explain that outcome. The probit model, which is based on the normal distribution,

estimates the probability that $Y = 1$ (a condition is fulfilled) given the values of the independent variables. The logit model is identical, except that it is based on the logistic distribution rather than the normal distribution.

- Supervised learning is machine learning that makes use of labelled training data and contrasts with unsupervised learning which does not make use of labelled data.
- Focuses of data analytics include correlation, prediction, causal inference, classification, clustering, and dimension reduction. Supervised ML is typically used for prediction and classification while unsupervised machine learning is used for clustering and dimension reduction.
- Penalized regression is a computationally efficient technique used in prediction problems. CART is a common supervised ML technique which can be applied to predict either a categorical or continuous target variable. Neural networks are applied to a variety of tasks characterized by nonlinearities and interactions among variables. Neural networks consist of three layers: an input layer, hidden layer(s), and an output layer. The K-means algorithm is a simple, bottom-up clustering algorithm based on concepts of geometric distance from points called centroids. PCA is an unsupervised learning algorithm that supplies a lower dimensional view of the structure of the volatility in data.
- The process to train ML models involves following steps:
  - Specify the ML technique/algorithm
  - Specify the associated hyperparameters
  - Divide data into training and validation samples
  - Evaluate learning with performance measure P, using the validation sample, and tune the hyperparameters
  - Repeat the training cycle the specified number of times or until the required performance level is obtained.

REFERENCES

Altman, Edward I. 1968. "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy." *Journal of Finance*, vol. 23: 589–609.

Altman, Edward I., R. Halderman, and P. Narayanan. 1977. "Zeta Analysis: A New Model to Identify Bankruptcy Risk of Corporations." *Journal of Banking & Finance*, vol. 1: 29–54.

Bhabra, Harjeet S., and Jiayin Huang. 2013. "An Empirical Investigation of Mergers and Acquisitions by Chinese Listed Companies, 1997–2007." *Journal of Multinational Financial Management*, vol. 23: 186–207.

Breusch, T., and A. Pagan. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica*, vol. 47: 1287–1294.

Buetow, Gerald W., Jr, Robert R. Johnson, and David E. Runkle. 2000. "The Inconsistency of Return-Based Style Analysis." *Journal of Portfolio Management*, vol. 26, no. 3: 61–77.

Culkin, Robert, and Sanjiv Das. 2017. "Machine Learning in Finance: The Case of Deep Learning for Option Pricing." *Journal of Investment Management* 15 (4).

Das, Sanjiv. 2014. "Text and Context: Language Analytics for Finance." *Foundations and Trends in Finance* 8 (3): 145–261.

Das, Sanjiv, Rong Fan, and Gary Geng."Bayesian Migration in Credit Ratings Based on Probabilities of Default," 2002, *Journal of Fixed Income* 12 (3): 17–23.

Durbin, J., and G.S. Watson. 1951. "Testing for Serial Correlation in Least Squares Regression, II." *Biometrika*, vol. 38: 159–178.

Efron, Bradley, and Trevor Hastie. 2016. *Computer Age Statistical Inference*. New York: Cambridge University Press.

Goldberger, Arthur S. 1998. *Introductory Econometrics*. Cambridge, MA: Harvard University Press.

Greene, William H. 2018. *Economic Analysis*, 8th edition. New York: Pearson Education.

Gujarati, Damodar N., Dawn C. Porter, and Sangeetha Gunasekar. 2011. *Basic Econometrics*, 5th edition. New York: McGraw-Hill Irwin.

Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica*, vol. 50, no. 4: 1029–1054.

Hutchinson, J. M., A. W. Lo, and T. Poggio. 1994. "A Nonparametric Approach to Pricing and Hedging Derivative Securities via Learning Networks." *Journal of Finance* 49 (3): 851–89.

Keane, Michael P., and David E. Runkle. 1998. "Are Financial Analysts' Forecasts of Corporate Profits Rational?" *Journal of Political Economy*, vol. 106, no. 4: 768–805.

Kmenta, Jan. 1986. *Elements of Econometrics*, 2nd edition. New York: Macmillan.

Kyaw, NyoNyo A., John Manley, and Anand Shetty. 2011. "Factors in Multinational Valuations: Transparency, Political Risk, and Diversification." *Journal of Multinational Financial Management*, vol. 21: 55–67.

MacKinlay, A. Craig, and Matthew P. Richardson. 1991. "Using Generalized Methods of Moments to Test Mean–Variance Efficiency." *Journal of Finance*, vol. 46, no. 2: 511–527.

Mankiw, N. Gregory. 2015. *Macroeconomics*, 9th edition. New York: Worth Publishers.

Mayer, Thomas. 1980. "Economics as a Hard Science: Realistic Goal or Wishful Thinking?" *Economic Inquiry*, vol. 18, no. 2: 165–178.

Mitchell, T. 1997. *Machine Learning*. McGraw-Hill.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, vol. 55, no. 3: 703–708.

Petersen, Mitchell A. 2009. "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches." *Review of Financial Studies*, vol. 22, no. 1: 435–480.

Samuel, A.July 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 3 (3): 210–29.

Sharpe, William F. 1988. "Determining a Fund's Effective Asset Mix." *Investment Management Review*, November/December:59–69.

Siegel, Jeremy J. 2014. *Stocks for the Long Run*, 5th edition. New York: McGraw-Hill.

Stoll, Hans R. 1978. "The Pricing of Security Dealer Services: An Empirical Study of Nasdaq Stocks." *Journal of Finance*, vol. 33, no. 4: 1153–1172.

Theobald, O. 2017. *Machine Learning for Absolute Beginners*. Independently published.

Tibshirani, R. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B. Methodological* 58 (1): 267–88.

Treynor, Jack L., and Kay Mazuy. 1966. "Can Mutual Funds Outguess the Market?" *Harvard Business Review*, vol. 44: 131–136.

Vapnik, V. N. 1989. *Statistical Learning Theory*. Wiley-Interscience.

Wang, W., Y. Huang, Y. Wang, and L. Wang. 2014. "Generalized Autoencoder: A Neural Network Framework for Dimensionality Reduction." *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*: 490–497.

PRACTICE PROBLEMS

1 With many US companies operating globally, the effect of the US dollar's strength on a US company's returns has become an important investment issue. You would like to determine whether changes in the US dollar's value and overall US equity market returns affect an asset's returns. You decide to use the S&P 500 Index to represent the US equity market.

A Write a multiple regression equation to test whether changes in the value of the dollar and equity market returns affect an asset's returns. Use the notations below.

R_{it} = return on the asset in period t
R_{Mt} = return on the S&P 500 in period t
ΔX_t = change in period t in the log of a trade-weighted index of the foreign exchange value of US dollar against the currencies of a broad group of major US trading partners.

B You estimate the regression for Archer Daniels Midland Company (NYSE: ADM). You regress its monthly returns for the period January 1990 to December 2002 on S&P 500 Index returns and changes in the log of the trade-weighted exchange value of the US dollar. The table below shows the coefficient estimates and their standard errors.

Coefficient Estimates from Regressing ADM's Returns: Monthly Data, January 1990–December 2002

	Coefficient	Standard Error
Intercept	0.0045	0.0062
R_{Mt}	0.5373	0.1332
ΔX_t	–0.5768	0.5121
n = 156		

Source: FactSet, Federal Reserve Bank of Philadelphia.

Determine whether S&P 500 returns affect ADM's returns. Then determine whether changes in the value of the US dollar affect ADM's returns. Use a 0.05 significance level to make your decisions.

C Based on the estimated coefficient on R_{Mt}, is it correct to say that "for a 1 percentage point increase in the return on the S&P 500 in period t, we expect a 0.5373 percentage point increase in the return on ADM"?

2 One of the most important questions in financial economics is what factors determine the cross-sectional variation in an asset's returns. Some have argued that book-to-market ratio and size (market value of equity) play an important role.

A Write a multiple regression equation to test whether book-to-market ratio and size explain the cross-section of asset returns. Use the notations below.

$(B/M)_i$ = book-to-market ratio for asset i

R_i = return on asset i in a particular month

Size_i = natural log of the market value of equity for asset i

B The table below shows the results of the linear regression for a cross-section of 66 companies. The size and book-to-market data for each company are for December 2001. The return data for each company are for January 2002.

Results from Regressing Returns on the Book-to-Market Ratio and Size

	Coefficient	Standard Error
Intercept	0.0825	0.1644
$(\text{B/M})_i$	−0.0541	0.0588
Size_i	−0.0164	0.0350
$n = 66$		

Source: FactSet.

Determine whether the book-to-market ratio and size are each useful for explaining the cross-section of asset returns. Use a 0.05 significance level to make your decision.

3 There is substantial cross-sectional variation in the number of financial analysts who follow a company. Suppose you hypothesize that a company's size (market cap) and financial risk (debt-to-equity ratios) influence the number of financial analysts who follow a company. You formulate the following regression model:

$$(\text{Analyst following})_i = b_0 + b_1\text{Size}_i + b_2(\text{D/E})_i + \varepsilon_i$$

where

$(\text{Analyst following})_i$ = the natural log of $(1 + n)$, where n_i is the number of analysts following company i

Size_i = the natural log of the market capitalization of company i in millions of dollars

$(\text{D/E})_i$ = the debt-to-equity ratio for company i

In the definition of Analyst following, 1 is added to the number of analysts following a company because some companies are not followed by any analysts, and the natural log of 0 is indeterminate. The following table gives the coefficient estimates of the above regression model for a randomly selected sample of 500 companies. The data are for the year 2002.

Coefficient Estimates from Regressing Analyst Following on Size and Debt-to-Equity Ratio

	Coefficient	Standard Error	*t*-Statistic
Intercept	−0.2845	0.1080	−2.6343
Size_i	0.3199	0.0152	21.0461

(continued)

(Continued)

	Coefficient	Standard Error	*t*-Statistic
$(D/E)_i$	−0.1895	0.0620	−3.0565
n = 500			

Source: First Call/Thomson Financial, Compustat.

A Consider two companies, both of which have a debt-to-equity ratio of 0.75. The first company has a market capitalization of $100 million, and the second company has a market capitalization of $1 billion. Based on the above estimates, how many more analysts will follow the second company than the first company?

B Suppose the p-value reported for the estimated coefficient on $(D/E)_i$ is 0.00236. State the interpretation of 0.00236.

4 In early 2001, US equity marketplaces started trading all listed shares in minimal increments (ticks) of $0.01 (decimalization). After decimalization, bid–ask spreads of stocks traded on the NASDAQ tended to decline. In response, spreads of NASDAQ stocks cross-listed on the Toronto Stock Exchange (TSE) tended to decline as well. Researchers Oppenheimer and Sabherwal (2003) hypothesized that the percentage decline in TSE spreads of cross-listed stocks was related to company size, the predecimalization ratio of spreads on NASDAQ to those on the TSE, and the percentage decline in NASDAQ spreads. The following table gives the regression coefficient estimates from estimating that relationship for a sample of 74 companies. Company size is measured by the natural logarithm of the book value of company's assets in thousands of Canadian dollars.

Coefficient Estimates from Regressing Percentage Decline in TSE Spreads on Company Size, Predecimalization Ratio of NASDAQ to TSE Spreads, and Percentage Decline in NASDAQ Spreads

	Coefficient	*t*-Statistic
Intercept	−0.45	−1.86
Size_i	0.05	2.56
$(\text{Ratio of spreads})_i$	−0.06	−3.77
$(\text{Decline in NASDAQ spreads})_i$	0.29	2.42
n = 74		

Source: Oppenheimer and Sabherwal (2003).

The average company in the sample has a book value of assets of C$900 million and a predecimalization ratio of spreads equal to 1.3. Based on the above model, what is the predicted decline in spread on the TSE for a company with these average characteristics, given a 1 percentage point decline in NASDAQ spreads?

5 The "neglected-company effect" claims that companies that are followed by fewer analysts will earn higher returns on average than companies that are followed by many analysts. To test the neglected-company effect, you have

collected data on 66 companies and the number of analysts providing earnings estimates for each company. You decide to also include size as an independent variable, measuring size as the log of the market value of the company's equity, to try to distinguish any small-company effect from a neglected-company effect. The small-company effect asserts that small-company stocks may earn average higher risk-adjusted returns than large-company stocks.

The table below shows the results from estimating the model $R_i = b_0 + b_1\text{Size}_i + b_2(\text{Number of analysts})_i + \varepsilon_i$ for a cross-section of 66 companies. The size and number of analysts for each company are for December 2001. The return data are for January 2002.

Results from Regressing Returns on Size and Number of Analysts

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0388	0.1556	0.2495
Size_i	–0.0153	0.0348	–0.4388
$(\text{Number of analysts})_i$	0.0014	0.0015	0.8995

ANOVA	df	SS	MSS
Regression	2	0.0094	0.0047
Residual	63	0.6739	0.0107
Total	65	0.6833	
Residual standard error	0.1034		
R-squared	0.0138		
Observations	66		

Source: First Call/Thomson Financial, FactSet.

A What test would you conduct to see whether the two independent variables are *jointly* statistically related to returns (H_0: $b_1 = b_2 = 0$)?

B What information do you need to conduct the appropriate test?

C Determine whether the two variables jointly are statistically related to returns at the 0.05 significance level.

D Explain the meaning of adjusted R^2 and state whether adjusted R^2 for the regression would be smaller than, equal to, or larger than 0.0138.

6 Some developing nations are hesitant to open their equity markets to foreign investment because they fear that rapid inflows and outflows of foreign funds will increase volatility. In July 1993, India implemented substantial equity market reforms, one of which allowed foreign institutional investors into the Indian equity markets. You want to test whether the volatility of returns of stocks traded on the Bombay Stock Exchange (BSE) increased after July 1993, when foreign institutional investors were first allowed to invest in India. You have collected monthly return data for the BSE from February 1990 to December 1997. Your dependent variable is a measure of return volatility of stocks traded on the BSE; your independent variable is a dummy variable that is coded 1 if foreign investment was allowed during the month and 0 otherwise.

You believe that market return volatility actually *decreases* with the opening up of equity markets. The table below shows the results from your regression.

Results from Dummy Regression for Foreign Investment in India with a Volatility Measure as the Dependent Variable

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0133	0.0020	6.5351
Dummy	−0.0075	0.0027	−2.7604
n = 95			

Source: FactSet.

A State null and alternative hypotheses for the slope coefficient of the dummy variable that are consistent with testing your stated belief about the effect of opening the equity markets on stock return volatility.

B Determine whether you can reject the null hypothesis at the 0.05 significance level (in a one-sided test of significance).

C According to the estimated regression equation, what is the level of return volatility before and after the market-opening event?

7 Both researchers and the popular press have discussed the question as to which of the two leading US political parties, Republicans or Democrats, is better for the stock market.

A Write a regression equation to test whether overall market returns, as measured by the annual returns on the S&P 500 Index, tend to be higher when the Republicans or the Democrats control the White House. Use the notations below.

R_{Mt} = return on the S&P 500 in period t

Party_t = the political party controlling the White House (1 for a Republican president; 0 for a Democratic president) in period t

B The table below shows the results of the linear regression from Part A using annual data for the S&P 500 and a dummy variable for the party that controlled the White House. The data are from 1926 to 2002.

Results from Regressing S&P 500 Returns on a Dummy Variable for the Party That Controlled the White House, 1926-2002

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.1494	0.0323	4.6270
Party_t	−0.0570	0.0466	−1.2242

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	1	0.0625	0.0625	1.4987	0.2247
Residual	75	3.1287	0.0417		
Total	76	3.1912			
Residual standard error		0.2042			

(Continued)

ANOVA	df	SS	MSS	*F*	Significance *F*
R-squared		0.0196			
Observations		77			

Source: FactSet.

Based on the coefficient and standard error estimates, verify to two decimal places the *t*-statistic for the coefficient on the dummy variable reported in the table.

C Determine at the 0.05 significance level whether overall US equity market returns tend to differ depending on the political party controlling the White House.

8 Problem 3 addressed the cross-sectional variation in the number of financial analysts who follow a company. In that problem, company size and debt-to-equity ratios were the independent variables. You receive a suggestion that membership in the S&P 500 Index should be added to the model as a third independent variable; the hypothesis is that there is greater demand for analyst coverage for stocks included in the S&P 500 because of the widespread use of the S&P 500 as a benchmark.

A Write a multiple regression equation to test whether analyst following is systematically higher for companies included in the S&P 500 Index. Also include company size and debt-to-equity ratio in this equation. Use the notations below.

$(\text{Analyst following})_i$ = natural log of (1 + Number of analysts following company i)
Size_i = natural log of the market capitalization of company i in millions of dollars
$(\text{D/E})_i$ = debt-to-equity ratio for company i
$\text{S\&P}_i$ = inclusion of company i in the S&P 500 Index (1 if included, 0 if not included)

In the above specification for analyst following, 1 is added to the number of analysts following a company because some companies are not followed by any analyst, and the natural log of 0 is indeterminate.

B State the appropriate null hypothesis and alternative hypothesis in a two-sided test of significance of the dummy variable.

C The following table gives estimates of the coefficients of the above regression model for a randomly selected sample of 500 companies. The data are for the year 2002. Determine whether you can reject the null hypothesis at the 0.05 significance level (in a two-sided test of significance).

Coefficient Estimates from Regressing Analyst Following on Size, Debt-to-Equity Ratio, and S&P 500 Membership, 2002

	Coefficient	Standard Error	*t*-Statistic
Intercept	–0.0075	0.1218	–0.0616
Size_i	0.2648	0.0191	13.8639

(continued)

(Continued)

	Coefficient	Standard Error	*t*-Statistic
$(D/E)_i$	−0.1829	0.0608	−3.0082
$S\&P_i$	0.4218	0.0919	4.5898
n = 500			

Source: First Call/Thomson Financial, Compustat.

D Consider a company with a debt-to-equity ratio of 2/3 and a market capitalization of $10 billion. According to the estimated regression equation, how many analysts would follow this company if it were not included in the S&P 500 Index, and how many would follow if it were included in the index?

E In Problem 3, using the sample, we estimated the coefficient on the size variable as 0.3199, versus 0.2648 in the above regression. Discuss whether there is an inconsistency in these results.

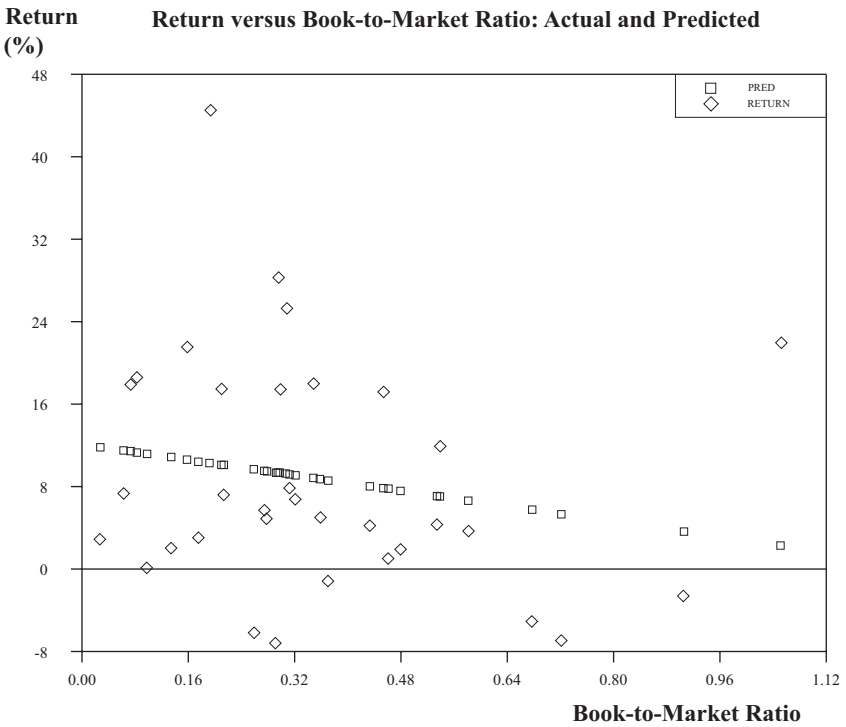
9 You believe there is a relationship between book-to-market ratios and subsequent returns. The output from a cross-sectional regression and a graph of the actual and predicted relationship between the book-to-market ratio and return are shown below.

Results from Regressing Returns on the Book-to-Market Ratio

	Coefficient	Standard Error	*t*-Statistic
Intercept	12.0130	3.5464	3.3874
$\left(\frac{\text{Book value}}{\text{Market value}}\right)_i$	−9.2209	8.4454	−1.0918

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	1	154.9866	154.9866	1.1921	0.2831
Residual	32	4162.1895	130.0684		
Total	33	4317.1761			
Residual standard error		11.4048			
R-squared		0.0359			
Observations		34			

(Continued)



A You are concerned with model specification problems and regression assumption violations. Focusing on assumption violations, discuss symptoms of conditional heteroskedasticity based on the graph of the actual and predicted relationship.

B Describe in detail how you could formally test for conditional heteroskedasticity in this regression.

C Describe a recommended method for correcting for conditional heteroskedasticity.

10 You are examining the effects of the January 2001 NYSE implementation of the trading of shares in minimal increments (ticks) of $0.01 (decimalization). In particular, you are analyzing a sample of 52 Canadian companies cross-listed on both the NYSE and the Toronto Stock Exchange (TSE). You find that the bid–ask spreads of these shares decline on both exchanges after the NYSE decimalization. You run a linear regression analyzing the decline in spreads on the TSE, and find that the decline on the TSE is related to company size, pre-decimalization ratio of NYSE to TSE spreads, and decline in the NYSE spreads. The relationships are statistically significant. You want to be sure, however, that the results are not influenced by conditional heteroskedasticity. Therefore, you regress the squared residuals of the regression model on the three independent variables. The R^2 for this regression is 14.1 percent. Perform a statistical test to determine if conditional heteroskedasticity is present.

11 You are analyzing if institutional investors such as mutual funds and pension funds prefer to hold shares of companies with less volatile returns. You have the percentage of shares held by institutional investors at the end of 1998 for a random sample of 750 companies. For these companies, you compute the standard deviation of daily returns during that year. Then you regress the institutional holdings on the standard deviation of returns. You find that the regression is

significant at the 0.01 level and the *F*-statistic is 12.98. The R^2 for this regression is 1.7 percent. As expected, the regression coefficient of the standard deviation of returns is negative. Its *t*-statistic is –3.60, which is also significant at the 0.01 level. Before concluding that institutions prefer to hold shares of less volatile stocks, however, you want to be sure that the regression results are not influenced by conditional heteroskedasticity. Therefore, you regress the squared residuals of the regression model on the standard deviation of returns. The R^2 for this regression is 0.6 percent.

A Perform a statistical test to determine if conditional heteroskedasticity is present at the 0.05 significance level.

B In view of your answer to Part A, what remedial action, if any, is appropriate?

12 In estimating a regression based on monthly observations from January 1987 to December 2002 inclusive, you find that the coefficient on the independent variable is positive and significant at the 0.05 level. You are concerned, however, that the *t*-statistic on the independent variable may be inflated because of serial correlation between the error terms. Therefore, you examine the Durbin–Watson statistic, which is 1.8953 for this regression.

A Based on the value of the Durbin–Watson statistic, what can you say about the serial correlation between the regression residuals? Are they positively correlated, negatively correlated, or not correlated at all?

B Compute the sample correlation between the regression residuals from one period and those from the previous period.

C Perform a statistical test to determine if serial correlation is present. Assume that the critical values for 192 observations when there is a single independent variable are about 0.09 above the critical values for 100 observations.

13 The book-to-market ratio and the size of a company's equity are two factors that have been asserted to be useful in explaining the cross-sectional variation in subsequent returns. Based on this assertion, you want to estimate the following regression model:

$$R_i = b_0 + b_1\left(\frac{\text{Book}}{\text{Market}}\right)_i + b_2\text{Size}_i + \varepsilon_i$$

where

R_i = Return of company *i*'s shares (in the following period)

$\left(\frac{\text{Book}}{\text{Market}}\right)_i$ = company *i*'s book-to-market ratio

Size_i = Market value of company *i*'s equity

A colleague suggests that this regression specification may be erroneous, because he believes that the book-to-market ratio may be strongly related to (correlated with) company size.

A To what problem is your colleague referring, and what are its consequences for regression analysis?

B With respect to multicollinearity, critique the choice of variables in the regression model above.

Regression of Return on Book-to-Market and Size

	Coefficient	Standard Error	*t*-Statistic
Intercept	14.1062	4.220	3.3427
$\left(\frac{\text{Book}}{\text{Market}}\right)_i$	−12.1413	9.0406	−1.3430
Size_i	−0.00005502	0.00005977	−0.92047
R-squared	0.06156		
Observations	34		

Correlation Matrix

	Book-to-Market Ratio	Size
Book-to-Market Ratio	1.0000	
Size	−0.3509	1.0000

C State the classic symptom of multicollinearity and comment on that basis whether multicollinearity appears to be present, given the additional fact that the *F*-test for the above regression is not significant.

14 You are analyzing the variables that explain the returns on the stock of the Boeing Company. Because overall market returns are likely to explain a part of the returns on Boeing, you decide to include the returns on a value-weighted index of all the companies listed on the NYSE, AMEX, and NASDAQ as an independent variable. Further, because Boeing is a large company, you also decide to include the returns on the S&P 500 Index, which is a value-weighted index of the larger market-capitalization companies. Finally, you decide to include the changes in the US dollar's value. To conduct your test, you have collected the following data for the period 1990–2002.

R_t = monthly return on the stock of Boeing in month t
R_{ALLt} = monthly return on a value-weighted index of all the companies listed on the NYSE, AMEX, and NASDAQ in month t
R_{SPt} = monthly return on the S&P 500 Index in month t
ΔX_t = change in month t in the log of a trade-weighted index of the foreign exchange value of the US dollar against the currencies of a broad group of major US trading partners

The following table shows the output from regressing the monthly return on Boeing stock on the three independent variables.

Regression of Boeing Returns on Three Explanatory Variables: Monthly Data, January 1990–December 2002

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0026	0.0066	0.3939
R_{ALLt}	−0.1337	0.6219	−0.2150

(continued)

(Continued)

	Coefficient	Standard Error	*t*-Statistic
R_{SPt}	0.8875	0.6357	1.3961
ΔX_t	0.2005	0.5399	0.3714

ANOVA	df	SS	MSS
Regression	3	0.1720	0.0573
Residual	152	0.8947	0.0059
Total	155	1.0667	
Residual standard error	0.0767		
R-squared	0.1610		
Observations	156		

Source: FactSet, Federal Reserve Bank of Philadelphia.

From the *t*-statistics, we see that none of the explanatory variables is statistically significant at the 5 percent level or better. You wish to test, however, if the three variables *jointly* are statistically related to the returns on Boeing.

A Your null hypothesis is that all three population slope coefficients equal 0—that the three variables *jointly* are statistically not related to the returns on Boeing. Conduct the appropriate test of that hypothesis.

B Examining the regression results, state the regression assumption that may be violated in this example. Explain your answer.

C State a possible way to remedy the violation of the regression assumption identified in Part B.

15 You are analyzing the cross-sectional variation in the number of financial analysts that follow a company (also the subject of Problems 3 and 8). You believe that there is less analyst following for companies with a greater debt-to-equity ratio and greater analyst following for companies included in the S&P 500 Index. Consistent with these beliefs, you estimate the following regression model.

$$(\text{Analysts following})_i = b_0 + b_1(\text{D/E})_i + b_2(\text{S\&P})_i + \varepsilon_i$$

where

$(\text{Analysts following})_i$ = natural log of (1 + Number of analysts following company i)
$(\text{D/E})_i$ = debt-to-equity ratio for company i
$\text{S\&P}_i$ = inclusion of company i in the S&P 500 Index (1 if included; 0 if not included)

In the preceding specification, 1 is added to the number of analysts following a company because some companies are not followed by any analysts, and the natural log of 0 is indeterminate. The following table gives the coefficient estimates of the above regression model for a randomly selected sample of 500 companies. The data are for the year 2002.

Coefficient Estimates from Regressing Analyst Following on Debt-to-Equity Ratio and S&P 500 Membership, 2002

	Coefficient	Standard Error	*t*-Statistic
Intercept	1.5367	0.0582	26.4038
$(D/E)_i$	−0.1043	0.0712	−1.4649
$S\&P_i$	1.2222	0.0841	14.5327
$n = 500$			

Source: First Call/Thomson Financial, Compustat.

You discuss your results with a colleague. She suggests that this regression specification may be erroneous, because analyst following is likely to be also related to the size of the company.

A What is this problem called, and what are its consequences for regression analysis?

B To investigate the issue raised by your colleague, you decide to collect data on company size also. You then estimate the model after including an additional variable, Size *i*, which is the natural log of the market capitalization of company *i* in millions of dollars. The following table gives the new coefficient estimates.

Coefficient Estimates from Regressing Analyst Following on Size, Debt-to-Equity Ratio, and S&P 500 Membership, 2002

	Coefficient	Standard Error	*t*-Statistic
Intercept	−0.0075	0.1218	−0.0616
$Size_i$	0.2648	0.0191	13.8639
$(D/E)_i$	−0.1829	0.0608	−3.0082
$S\&P_i$	0.4218	0.0919	4.5898
$n = 500$			

Source: First Call/Thomson Financial, Compustat.

What do you conclude about the existence of the problem mentioned by your colleague in the original regression model you had estimated?

16 You have noticed that hundreds of non-US companies are listed not only on a stock exchange in their home market but also on one of the exchanges in the United States. You have also noticed that hundreds of non-US companies are listed only in their home market and not in the United States. You are trying to predict whether or not a non-US company will choose to list on a US exchange. One of the factors that you think will affect whether or not a company lists in the United States is its size relative to the size of other companies in its home market.

A What kind of a dependent variable do you need to use in the model?

B What kind of a model should be used?

The following information relates to Questions 17–22

Gary Hansen is a securities analyst for a mutual fund specializing in small-capitalization growth stocks. The fund regularly invests in initial public offerings (IPOs). If the fund subscribes to an offer, it is allocated shares at the offer price. Hansen notes that IPOs frequently are underpriced, and the price rises when open market trading begins. The initial return for an IPO is calculated as the change in price on the first day of trading divided by the offer price. Hansen is developing a regression model to predict the initial return for IPOs. Based on past research, he selects the following independent variables to predict IPO initial returns:

Underwriter rank	=	1–10, where 10 is highest rank
Pre-offer price adjustment[a]	=	(Offer price – Initial filing price)/Initial filing price
Offer size ($ millions)	=	Shares sold × Offer price
Fraction retained[a]	=	Fraction of total company shares retained by insiders

[a]Expressed as a decimal

Hansen collects a sample of 1,725 recent IPOs for his regression model. Regression results appear in Exhibit 1, and ANOVA results appear in Exhibit 2.

Exhibit 1 Hansen's Regression Results Dependent Variable: IPO Initial Return (Expressed in Decimal Form, i.e., 1% = 0.01)

Variable	Coefficient (b_j)	Standard Error	*t*-Statistic
Intercept	0.0477	0.0019	25.11
Underwriter rank	0.0150	0.0049	3.06
Pre-offer price adjustment	0.4350	0.0202	21.53
Offer size	–0.0009	0.0011	–0.82
Fraction retained	0.0500	0.0260	1.92

Exhibit 2 Selected ANOVA Results for Hansen's Regression

	Degrees of Freedom (df)	Sum of Squares (SS)
Regression	4	51.433
Residual	1,720	91.436
Total	1,724	142.869

Multiple R-squared = 0.36

Hansen wants to use the regression results to predict the initial return for an upcoming IPO. The upcoming IPO has the following characteristics:

- underwriter rank = 6;
- pre-offer price adjustment = 0.04;

- offer size = $40 million;
- fraction retained = 0.70.

Because he notes that the pre-offer price adjustment appears to have an important effect on initial return, Hansen wants to construct a 95 percent confidence interval for the coefficient on this variable. He also believes that for each 1 percent increase in pre-offer price adjustment, the initial return will increase by less than 0.5 percent, holding other variables constant. Hansen wishes to test this hypothesis at the 0.05 level of significance.

Before applying his model, Hansen asks a colleague, Phil Chang, to review its specification and results. After examining the model, Chang concludes that the model suffers from two problems: 1) conditional heteroskedasticity, and 2) omitted variable bias. Chang makes the following statements:

Statement 1 "Conditional heteroskedasticity will result in consistent coefficient estimates, but both the *t*-statistics and *F*-statistic will be biased, resulting in false inferences."

Statement 2 "If an omitted variable is correlated with variables already included in the model, coefficient estimates will be biased and inconsistent and standard errors will also be inconsistent."

Selected values for the *t*-distribution and *F*-distribution appear in Exhibits 3 and 4, respectively.

Exhibit 3 Selected Values for the *t*-Distribution (df = ∞)

Area in Right Tail	*t*-Value
0.050	1.645
0.025	1.960
0.010	2.326
0.005	2.576

Exhibit 4 Selected Values for the *F*-Distribution (α = 0.01) (df1/df2: Numerator/Denominator Degrees of Freedom)

		df1	
		4	∞
df2	4	16.00	13.50
	∞	3.32	1.00

17 Based on Hansen's regression, the predicted initial return for the upcoming IPO is *closest* to:

A 0.0943.

B 0.1064.

C 0.1541.

18 The 95 percent confidence interval for the regression coefficient for the pre-offer price adjustment is *closest* to:

A 0.156 to 0.714.

B 0.395 to 0.475.

C 0.402 to 0.468.

19 The *most* appropriate null hypothesis and the *most* appropriate conclusion regarding Hansen's belief about the magnitude of the initial return relative to that of the pre-offer price adjustment (reflected by the coefficient b_j) are:

	Null Hypothesis	Conclusion about b_j (0.05 Level of Significance)
A	H_0: $b_j = 0.5$	Reject H_0
B	H_0: $b_j \geq 0.5$	Fail to reject H_0
C	H_0: $b_j \geq 0.5$	Reject H_0

20 The *most* appropriate interpretation of the multiple *R*-squared for Hansen's model is that:

A unexplained variation in the dependent variable is 36 percent of total variation.

B correlation between predicted and actual values of the dependent variable is 0.36.

C correlation between predicted and actual values of the dependent variable is 0.60.

21 Is Chang's Statement 1 correct?

A Yes.

B No, because the model's *F*-statistic will not be biased.

C No, because the model's *t*-statistics will not be biased.

22 Is Chang's Statement 2 correct?

A Yes.

B No, because the model's coefficient estimates will be unbiased.

C No, because the model's coefficient estimates will be consistent.

The following information relates to Questions 23–28

Adele Chiesa is a money manager for the Bianco Fund. She is interested in recent findings showing that certain business condition variables predict excess US stock market returns (one-month market return minus one-month T-bill return). She is also familiar with evidence showing how US stock market returns differ by the political party affiliation of the US President. Chiesa estimates a multiple regression model to predict monthly excess stock market returns accounting for business conditions and the political party affiliation of the US President:

$$\text{Excess stock market return}_t = a_0 + a_1\text{Default spread}_{t-1} + a_2\text{Term spread}_{t-1} + a_3\text{Pres party dummy}_{t-1} + e_t$$

Default spread is equal to the yield on Baa bonds minus the yield on Aaa bonds. Term spread is equal to the yield on a 10-year constant-maturity US Treasury index minus the yield on a 1-year constant-maturity US Treasury index. Pres party dummy is equal to 1 if the US President is a member of the Democratic Party and 0 if a member of the Republican Party.

Chiesa collects 432 months of data (all data are in percent form, i.e., 0.01 = 1 percent). The regression is estimated with 431 observations because the independent variables are lagged one month. The regression output is in Exhibit 1. Exhibits 2 through 5 contain critical values for selected test statistics.

Exhibit 1 Multiple Regression Output (the Dependent Variable Is the One-Month Market Return in Excess of the One-Month T-Bill Return)

	Coefficient	*t*-Statistic	*p*-Value
Intercept	−4.60	−4.36	<0.01
Default spread$_{t-1}$	3.04	4.52	<0.01
Term spread$_{t-1}$	0.84	3.41	<0.01
Pres party dummy$_{t-1}$	3.17	4.97	<0.01
Number of observations		431	
Test statistic from Breusch–Pagan (BP) test		7.35	
R^2		0.053	
Adjusted R^2		0.046	
Durbin–Watson (DW)		1.65	
Sum of squared errors (SSE)		19,048	
Regression sum of squares (SSR)		1,071	

An intern working for Chiesa has a number of questions about the results in Exhibit 1:

Question 1 How do you test to determine whether the overall regression model is significant?

Question 2 Does the estimated model conform to standard regression assumptions? For instance, is the error term serially correlated, or is there conditional heteroskedasticity?

Question 3 How do you interpret the coefficient for the Pres party dummy variable?

Question 4 Default spread appears to be quite important. Is there some way to assess the precision of its estimated coefficient? What is the economic interpretation of this variable?

After responding to her intern's questions, Chiesa concludes with the following statement: "Predictions from Exhibit 1 are subject to parameter estimate uncertainty, but not regression model uncertainty."

Exhibit 2 Critical Values for the Durbin–Watson Statistic (α = 0.05)

	K = 3	
N	d_l	d_u
420	1.825	1.854
430	1.827	1.855
440	1.829	1.857

Exhibit 3 Table of the Student's *t*-Distribution (One-Tailed Probabilities for df = ∞)

P	*t*
0.10	1.282
0.05	1.645
0.025	1.960
0.01	2.326
0.005	2.576

Exhibit 4 Values of χ^2

	Probability in Right Tail			
df	0.975	0.95	0.05	0.025
1	0.0001	0.0039	3.841	5.024
2	0.0506	0.1026	5.991	7.378
3	0.2158	0.3518	7.815	9.348
4	0.4840	0.7110	9.488	11.14

Exhibit 5 Table of the *F*-Distribution (Critical Values for Right-Hand Tail Area Equal to 0.05) Numerator: df1 and Denominator: df2

	df1				
df2	1	2	3	4	427
1	161	200	216	225	254
2	18.51	19.00	19.16	19.25	19.49
3	10.13	9.55	9.28	9.12	8.53

Exhibit 5 (Continued)

	df1				
df2	1	2	3	4	427
4	7.71	6.94	6.59	6.39	5.64
427	3.86	3.02	2.63	2.39	1.17

23 Regarding the intern's Question 1, is the regression model as a whole significant at the 0.05 level?

A No, because the calculated *F*-statistic is less than the critical value for *F*.

B Yes, because the calculated *F*-statistic is greater than the critical value for *F*.

C Yes, because the calculated χ^2 statistic is greater than the critical value for χ^2.

24 Which of the following is Chiesa's *best* response to Question 2 regarding serial correlation in the error term? At a 0.05 level of significance, the test for serial correlation indicates that there is:

A no serial correlation in the error term.

B positive serial correlation in the error term.

C negative serial correlation in the error term.

25 Regarding Question 3, the Pres party dummy variable in the model indicates that the mean monthly value for the excess stock market return is:

A 1.43 percent larger during Democratic presidencies than Republican presidencies.

B 3.17 percent larger during Democratic presidencies than Republican presidencies.

C 3.17 percent larger during Republican presidencies than Democratic presidencies.

26 In response to Question 4, the 95 percent confidence interval for the regression coefficient for the default spread is *closest* to:

A 0.13 to 5.95.

B 1.72 to 4.36.

C 1.93 to 4.15.

27 With respect to the default spread, the estimated model indicates that when business conditions are:

A strong, expected excess returns will be higher.

B weak, expected excess returns will be lower.

C weak, expected excess returns will be higher.

28 Is Chiesa's concluding statement correct regarding parameter estimate uncertainty and regression model uncertainty?

A Yes.

B No, predictions are not subject to parameter estimate uncertainty.

C No, predictions are subject to regression model uncertainty and parameter estimate uncertainty.

The following information relates to Questions 29–36

Doris Honoré is a securities analyst with a large wealth management firm. She and her colleague Bill Smith are addressing three research topics: how investment fund characteristics affect fund total returns, whether a fund rating system helps predict fund returns, and whether stock and bond market returns explain the returns of a portfolio of utility shares run by the firm.

To explore the first topic, Honoré decides to study US mutual funds using a sample of 555 large-cap US equity funds. The sample includes funds in style classes of value, growth, and blend (i.e., combining value and growth characteristics). The dependent variable is the average annualized rate of return (in percent) over the past five years. The independent variables are fund expense ratio, portfolio turnover, the natural logarithm of fund size, fund age, and three dummy variables. The multiple manager dummy variable has a value of 1 if the fund has multiple managers (and a value of 0 if it has a single manager). The fund style is indicated by a growth dummy (value of 1 for growth funds and 0 otherwise) and a blend dummy (value of 1 for blend funds and 0 otherwise). If the growth and blend dummies are both zero, the fund is a value fund. The regression output is given in Exhibit 1.

Exhibit 1 Multiple Regression Output for Large-Cap Mutual Fund Sample

	Coefficient	Standard Error	*t*-Statistic
Intercept	10.9375	1.3578	8.0551
Expense ratio (%)	–1.4839	0.2282	–6.5039
Portfolio turnover (%)	0.0017	0.0016	1.0777
ln (fund size in $)	0.1467	0.0612	2.3976
Manager tenure (years)	–0.0098	0.0102	–0.9580
Multiple manager dummy	0.0628	0.1533	0.4100
Fund age (years)	–0.0123	0.0047	–2.6279
Growth dummy	2.4368	0.1886	12.9185
Blend dummy	0.5757	0.1881	3.0611

ANOVA	df	SS	MSS
Regression	8	714.169	89.2712
Residual	546	1583.113	2.8995
Total	554	2297.282	
Multiple R	0.5576		
R^2	0.3109		
Adjusted R^2	0.3008		
Standard error (%)	1.7028		
Observations	555		

Based on the results shown in Exhibit 1, Honoré wants to test the hypothesis that all of the regression coefficients are equal to zero. For the 555 fund sample, she also wants to compare the performance of growth funds with the value funds.

Honoré is concerned about the possible presence of multicollinearity in the regression. She states that adding a new independent variable that is highly correlated with one or more independent variables already in the regression model, has three potential consequences:

1 The R^2 is expected to decline.
2 The regression coefficient estimates can become imprecise and unreliable.
3 The standard errors for some or all of the regression coefficients will become inflated.

Another concern for the regression model (in Exhibit 1) is conditional heteroskedasticity. Honoré is concerned that the presence of heteroskedasticity can cause both the *F*-test for the overall significance of the regression and the *t*-tests for significance of individual regression coefficients to be unreliable. She runs a regression of the squared residuals from the model in Exhibit 1 on the eight independent variables, and finds the R^2 is 0.0669.

As a second research project, Honoré wants to test whether including Morningstar's rating system, which assigns a one- through five-star rating to a fund, as an independent variable will improve the predictive power of the regression model. To do this, she needs to examine whether values of the independent variables in a given period predict fund return in the next period. Smith suggests three different methods of adding the Morningstar ratings to the model:

- Method 1: Add an independent variable that has a value equal to the number of stars in the rating of each fund.
- Method 2: Add five dummy variables, one for each rating.
- Method 3: Add dummy variables for four of the five ratings.

As a third research project, Honoré wants to establish whether bond market returns (proxied by returns of long-term US Treasuries) and stock market returns (proxied by returns of the S&P 500 Index) explain the returns of a portfolio of utility stocks being recommended to clients. Exhibit 2 presents the results of a regression of 10 years of monthly percentage total returns for the utility portfolio on monthly total returns for US Treasuries and the S&P 500.

Exhibit 2 Regression Analysis of Utility Portfolio Returns

	Coefficient	Standard Error	*t*-Statistic	*p*-Value
Intercept	–0.0851	0.2829	–0.3008	0.7641
US Treasury	0.4194	0.0848	4.9474	<0.0001
S&P 500	0.6198	0.0666	9.3126	<0.0001

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	2	827.48	413.74	46.28	<0.0001
Residual	117	1045.93	8.94		
Total	119	1873.41			
Multiple *R*	0.6646				
R^2	0.4417				
Adjusted R^2	0.4322				

(continued)

Exhibit 2 (Continued)

ANOVA	df	SS	MSS	*F*	Significance *F*
Standard error (%)	2.99				
Observations	120				

For the time-series model in Exhibit 2, Honoré says that positive serial correlation would not require that the estimated coefficients be adjusted, but that the standard errors of the regression coefficients would be underestimated. This issue would cause the *t*-statistics of the regression coefficients to be inflated. Honoré tests the null hypothesis that the there is no serial correlation in the regression residuals and finds that the Durbin–Watson statistic is equal to 1.81. The critical values at the 0.05 significance level for the Durbin–Watson statistic are $d_l = 1.63$ and $d_u = 1.72$.

Smith asks whether Honoré should have estimated the models in Exhibit 1 and Exhibit 2 using a probit or logit model instead of using a traditional regression analysis.

29 Considering Exhibit 1, the *F*-statistic is closest to:

A 3.22.

B 8.06.

C 30.79.

30 Based on Exhibit 1, the difference between the predicted annualized returns of a growth fund and an otherwise similar value fund is *closest* to:

A 1.86%.

B 2.44%.

C 3.01%.

31 Honoré describes three potential consequences of multicollinearity. Are all three consequences correct?

A Yes

B No, 1 is incorrect

C No, 2 is incorrect

32 Which of the three methods suggested by Smith would *best* capture the ability of the Morningstar rating system to predict mutual fund performance?

A Method 1

B Method 2

C Method 3

33 Honoré is concerned about the consequences of heteroskedasticity. Is she correct regarding the effect of heteroskedasticity on the reliability of the *F*-test and *t*-tests?

A Yes

B No, she is incorrect with regard to the *F*-test

C No, she is incorrect with regard to the *t*-tests

34 Is Honoré's description of the effects of positive serial correlation (in Exhibit 2) correct regarding the estimated coefficients and the standard errors?

A Yes

B No, she is incorrect about only the estimated coefficients

C No, she is incorrect about only the standard errors of the regression coefficients

35 Based on her estimated Durbin–Watson statistic, Honoré should:

A fail to reject the null hypothesis.

B reject the null hypothesis because there is significant positive serial correlation.

C reject the null hypothesis because there is significant negative serial correlation.

36 Should Honoré have estimated the models in Exhibit 1 and Exhibit 2 using probit or logit models instead of traditional regression analysis?

A Both should be estimated with probit or logit models.

B Neither should be estimated with probit or logit models.

C Only the analysis in Exhibit 1 should be done with probit or logit models.

The following information relates to Questions 37–45

Brad Varden, a junior analyst at an actively managed mutual fund, is responsible for research on a subset of the 500 large-cap equities the fund follows. Recently, the fund has been paying close attention to management turnover and to publicly available environmental, social, and governance (ESG) ratings. Varden is given the task of investigating whether any significant relationship exists between a company's profitability and either of these two characteristics. Colleen Quinni, a senior analyst at the fund, suggests that as an initial step in his investigation, Varden should perform a multiple regression analysis on the variables and report back to her.

Varden knows that Quinni is an expert at quantitative research, and she once told Varden that after you get an idea, you should formulate a hypothesis, test the hypothesis, and analyze the results. Varden expects to find that ESG rating is negatively related to ROE and CEO tenure is positively related to ROE. He considers a relationship meaningful when it is statistically significant at the 0.05 level. To begin, Varden collects values for ROE, CEO tenure, and ESG rating for a sample of 40 companies from the large-cap security universe. He performs a multiple regression with ROE (in percent) as the dependent variable and ESG rating and CEO tenure (in years) as the independent variables: $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \varepsilon_i$.

Exhibit 1 shows the regression results.

Exhibit 1 Regression Statistics

$\hat{Y}_i = 9.442 + 0.069X_{1i} + 0.681X_{2i}$

	Coefficient	Standard Error	*t*-Statistic	*p*-Value
Intercept	9.442	3.343	2.824	0.008
b_1 (ESG variable)	0.069	0.058	1.201	0.238
b_2 (Tenure variable)	0.681	0.295	2.308	0.027

(continued)

Exhibit 1 (Continued)

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	2	240.410	120.205	4.161	0.023
Residual	37	1069.000	28.892		
Total	39	1309.410			
Multiple *R*	0.428				
R^2	0.183				
Adjusted R^2	0.139				
Standard error (%)	5.375				
Observations	40				

DF Associates is one of the companies Varden follows. He wants to predict its ROE using his regression model. DF Associates' corporate ESG rating is 55, and the company's CEO has been in that position for 10.5 years.

Varden also wants to check on the relationship between these variables and the dividend growth rate (divgr), so he completes the correlation matrix shown in Exhibit 2.

Exhibit 2 Correlation Matrix

	ROE	ESG	Tenure	Divgr
ROE	1.0			
ESG	0.446	1.0		
Tenure	0.369	0.091	1.0	
Divgr	0.117	0.046	0.028	1.0

Investigating further, Varden determines that dividend growth is not a linear combination of CEO tenure and ESG rating. He is unclear about how additional independent variables would affect the significance of the regression, so he asks Quinni, "Given this correlation matrix, will both R^2 and adjusted R^2 automatically increase if I add dividend growth as a third independent variable?"

The discussion continues, and Quinni asks two questions.

1 What does your *F*-statistic of 4.161 tell you about the regression?

2 In interpreting the overall significance of your regression model, which statistic do you believe is most relevant: R^2, adjusted R^2, or the *F*-statistic?

Varden answers both questions correctly and says he wants to check two more ideas. He believes the following:

1 ROE is less correlated with the dividend growth rate in firms whose CEO has been in office more than 15 years, and

2 CEO tenure is a normally distributed random variable.

Later, Varden includes the dividend growth rate as a third independent variable and runs the regression on the fund's entire group of 500 large-cap equities. He finds that the adjusted R^2 is much higher than the results in Exhibit 1. He reports this

to Quinni and says, "Adding the dividend growth rate gives a model with a higher adjusted R^2. The three-variable model is clearly better." Quinni cautions, "I don't think you can conclude that yet."

37 Based on Exhibit 1 and given Varden's expectations, which is the *best* null hypothesis and conclusion regarding CEO tenure?

A $b_2 \leq 0$; reject the null hypothesis

B $b_2 = 0$; cannot reject the null hypothesis

C $b_2 \geq 0$; reject the null hypothesis

38 At a significance level of 1%, which of the following is the *best* interpretation of the regression coefficients with regard to explaining ROE?

A ESG is significant, but tenure is not.

B Tenure is significant, but ESG is not.

C Neither ESG nor tenure is significant.

39 Based on Exhibit 1, which independent variables in Varden's model are significant at the 0.05 level?

A ESG only

B Tenure only

C Neither ESG nor tenure

40 Based on Exhibit 1, the predicted ROE for DF Associates is *closest* to:

A 10.957%.

B 16.593%.

C 20.388%.

41 Based on Exhibit 2, Quinni's *best* answer to Varden's question about the effect of adding a third independent variable is:

A no for R^2 and no for adjusted R^2.

B yes for R^2 and no for adjusted R^2.

C yes for R^2 and yes for adjusted R^2.

42 Based on Exhibit 1, Varden's *best* answer to Quinni's question about the *F*-statistic is:

A both independent variables are significant at the 0.05 level.

B neither independent variable is significant at the 0.05 level.

C at least one independent variable is significant at the 0.05 level.

43 Varden's *best* answer to Quinni's question about overall significance is:

A R^2.

B adjusted R^2.

C the *F*-statistic.

44 If Varden's beliefs about ROE and CEO tenure are true, which of the following would violate the assumptions of multiple regression analysis?

A The assumption about CEO tenure distribution only

B The assumption about the ROE/dividend growth correlation only

C The assumptions about both the ROE/dividend growth correlation and CEO tenure distribution

45 The *best* rationale for Quinni's caution about the three-variable model is that the:

A dependent variable is defined differently.

B sample sizes are different in the two models.

C dividend growth rate is positively correlated with the other independent variables.

SOLUTIONS

1 A $R_{it} = b_0 + b_1 R_{Mt} + b_2 \Delta X_t + \varepsilon_{it}$

B We can test whether the coefficient on the S&P 500 Index returns is statistically significant. Our null hypothesis is that the coefficient is equal to 0 (H_0: $b_1 = 0$); our alternative hypothesis is that the coefficient is not equal to 0 (H_a: $b_1 \neq 0$). We construct the *t*-test of the null hypothesis as follows:

$$\frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{0.5373 - 0}{0.1332} = 4.0338$$

where

$\hat{b}_1$ = regression estimate of b_1

b_1 = the hypothesized value of the coefficient (here, 0)

$s_{\hat{b}_1}$ = the estimated standard error of $\hat{b}_1$

Because this regression has 156 observations and three regression coefficients, the *t*-test has 156 – 3 = 153 degrees of freedom. At the 0.05 significance level, the critical value for the test statistic is between 1.98 and 1.97. The absolute value of the test statistic is 4.0338; therefore, we can reject the null hypothesis that $b_1 = 0$.

Similarly, we can test whether the coefficient on the change in the value of the US dollar is statistically significant in this regression. Our null hypothesis is that the coefficient is equal to 0 (H_0: $b_2 = 0$); our alternative hypothesis is that the coefficient is not equal to 0 (H_a: $b_2 \neq 0$). We construct the *t*-test as follows:

$$\frac{\hat{b}_2 - b_2}{s_{\hat{b}_2}} = \frac{-0.5768 - 0}{0.5121} = -1.1263$$

As before, the *t*-test has 153 degrees of freedom, and the critical value for the test statistic is between 1.98 and 1.97 at the 0.05 significance level. The absolute value of the test statistic is 1.1263; therefore, we cannot reject the null hypothesis that $b_2 = 0$.

Based on the above *t*-tests, we conclude that S&P 500 Index returns do affect ADM's returns but that changes in the value of the US dollar do not affect ADM's returns.

C The statement is not correct. To make it correct, we need to add the qualification "holding ΔX constant" to the end of the quoted statement.

2 A $R_i = b_0 + b_1(\text{B/M})_i + b_2 \text{Size}_i + \varepsilon_i$

B We can test whether the coefficients on the book-to-market ratio and size are individually statistically significant using *t*-tests. For the book-to-market ratio, our null hypothesis is that the coefficient is equal to 0 (H_0: $b_1 = 0$); our alternative hypothesis is that the coefficient is not equal to 0 (H_a: $b_1 \neq 0$). We can test the null hypothesis using a *t*-test constructed as follows:

$$\frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{-0.0541 - 0}{0.0588} = -0.9201$$

where

$\hat{b}_1$ = regression estimate of b_1

b_1 = the hypothesized value of the coefficient (here, 0)

$s_{\hat{b}_1}$ = the estimated standard error of $\hat{b}_1$

This regression has 66 observations and three coefficients, so the *t*-test has 66 – 3 = 63 degrees of freedom. At the 0.05 significance level, the critical value for the test statistic is about 2.0. The absolute value of the test statistic is 0.9201; therefore, we cannot reject the null hypothesis that $b_1 = 0$. We can conclude that the book-to-market ratio is not useful in explaining the cross-sectional variation in returns for this sample.

We perform the same analysis to determine whether size (as measured by the log of the market value of equity) can help explain the cross-sectional variation in asset returns. Our null hypothesis is that the coefficient is equal to 0 (H_0: $b_2 = 0$); our alternative hypothesis is that the coefficient is not equal to 0 (H_a: $b_2 \neq 0$). We can test the null hypothesis using a *t*-test constructed as follows:

$$\frac{\hat{b}_2 - b_2}{s_{\hat{b}_2}} = \frac{-0.0164 - 0}{0.0350} = -0.4686$$

where

$\hat{b}_2$ = regression estimate of b_2

b_2 = the hypothesized value of the coefficient (here, 0)

$s_{\hat{b}_2}$ = the estimated standard error of $\hat{b}_2$

Again, because this regression has 66 observations and three coefficients, the *t*-test has 66 – 3 = 63 degrees of freedom. At the 0.05 significance level, the critical value for the test statistic is about 2.0. The absolute value of the test statistic is 0.4686; therefore, we cannot reject the null hypothesis that $b_2 = 0$. We can conclude that asset size is not useful in explaining the cross-sectional variation of asset returns in this sample.

3 A The estimated regression is $(\text{Analyst following})_i = -0.2845 + 0.3199\text{Size}_i - 0.1895(\text{D/E})_i + \varepsilon_i$. Therefore, the prediction for the first company is

$$\begin{aligned}(\text{Analyst following})_i &= -0.2845 + 0.3199(\ln 100) - 0.1895(0.75)\\ &= -0.2845 + 1.4732 - 0.1421 = 1.0466\end{aligned}$$

Recalling that $(\text{Analyst following})_i$ is the natural log of $(1 + n_i)$, where n_i is the number of analysts following company *i*; it follows that $1 + n_1 = e^{1.0466} = 2.848$, approximately. Therefore, $n_1 = 2.848 - 1 = 1.848$, or about two analysts. Similarly, the prediction for the second company is as follows:

$$\begin{aligned}(\text{Analyst following})_i &= -0.2845 + 0.3199(\ln 1{,}000) - 0.1895(0.75)\\ &= -0.2845 + 2.2098 - 0.1421\\ &= 1.7832\end{aligned}$$

Thus, $1 + n_2 = e^{1.7832} = 5.949$, approximately. Therefore, $n_2 = 5.949 - 1 = 4.949$, or about five analysts.

The model predicts that 5 – 2 = 3 more analysts will follow the second company than the first company.

B We would interpret the *p*-value of 0.00236 as the smallest level of significance at which we can reject a null hypothesis that the population value of the coefficient is 0, in a two-sided test. Clearly, in this regression the debt-to-equity ratio is a highly significant variable.

4 The estimated model is

$$\text{Percentage decline in TSE spread of company } i = -0.45 + 0.05\text{Size}_i - 0.06(\text{Ratio of spreads})_i + 0.29(\text{Decline in NASDAQ spreads})_i$$

Therefore, the prediction is

$$\begin{aligned}\text{Percentage decline in TSE spread} &= -0.45 + 0.05(\ln 900{,}000) - 0.06(1.3) + 0.29(1)\\ &= -0.45 + 0.69 - 0.08 + 0.29\\ &= 0.45\end{aligned}$$

The model predicts that for a company with average sample characteristics, the spread on the TSE declines by 0.45 percent for a 1 percent decline in NASDAQ spreads.

5 A To test the null hypothesis that all the slope coefficients in the regression model are equal to 0 (H_0: $b_1 = b_2 = 0$) against the alternative hypothesis that at least one slope coefficient is not equal to 0, we must use an *F*-test.

B To conduct the *F*-test, we need four inputs, all of which are found in the ANOVA section of the table in the statement of the problem:

i. total number of observations, n

ii. total number of regression coefficients to be estimated, $k + 1$

iii. sum of squared errors or residuals, $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ abbreviated SSE, and

iv. regression sum of squares, $\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$ abbreviated RSS

C The *F*-test formula is

$$F = \frac{\text{RSS}/k}{\text{SSE}/\left[n - (k+1)\right]} = \frac{0.0094/2}{0.6739/\left[66 - (2+1)\right]} = 0.4394$$

The *F*-statistic has degrees of freedom $F\{k, [n - (k + 1)]\} = F(2, 63)$. From the *F*-test table, for the 0.05 significance level, the critical value for $F(2, 63)$ is about 3.15, so we cannot reject the hypothesis that the slope coefficients are both 0. The two independent variables are jointly statistically unrelated to returns.

D Adjusted R^2 is a measure of goodness of fit that takes into account the number of independent variables in the regression, in contrast to R^2. We can assert that adjusted R^2 is smaller than $R^2 = 0.0138$ without the need to perform any calculations. (However, adjusted R^2 can be shown to equal –0.0175 using an expression in the text on the relationship between adjusted R^2 and R^2.)

6 A You believe that opening markets actually reduces return volatility; if that belief is correct, then the slope coefficient would be negative, $b_1 < 0$. The null hypothesis is that the belief is not true: H_0: $b_1 \geq 0$. The alternative hypothesis is that the belief is true: H_a: $b_1 < 0$.

B The critical value for the *t*-statistic with 95 – 2 = 93 degrees of freedom at the 0.05 significance level in a one-sided test is about 1.66. For the one-sided test stated in Part A, we reject the null hypothesis if the *t*-statistic on

the slope coefficient is less than –1.66. As the *t*-statistic of –2.7604 < –1.66, we reject the null. Because the dummy variable takes on a value of 1 when foreign investment is allowed, we can conclude that the volatility was lower with foreign investment.

C According to the estimated regression, average return volatility was 0.0133 (the estimated value of the intercept) before July 1993 and 0.0058 (= 0.0133 – 0.0075) after July 1993.

7 A The appropriate regression model is $R_{Mt} = b_0 + b_1 \text{Party}_t + \varepsilon_t$.

B The *t*-statistic reported in the table for the dummy variable tests whether the coefficient on Party_t is significantly different from 0. It is computed as follows:

$$\frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{-0.0570 - 0}{0.0466} = -1.22$$

where

$\hat{b}_1$ = regression estimate of b_1

b_1 = the hypothesized value of the coefficient (here, 0)

$s_{\hat{b}_1}$ = the estimated standard error of $\hat{b}_1$

To two decimal places, this value is the same as the *t*-statistic reported in the table for the dummy variable, as expected. The problem specified two decimal places because the reported regression output reflects rounding; for this reason, we often cannot exactly reproduce reported *t*-statistics.

C Because the regression has 77 observations and two coefficients, the *t*-test has 77 – 2 = 75 degrees of freedom. At the 0.05 significance level, the critical value for the two-tailed test statistic is about 1.99. The absolute value of the test statistic is 1.2242; therefore, we do not reject the null hypothesis that $b_1 = 0$. We can conclude that the political party in the White House does not, on average, affect the annual returns of the overall market as measured by the S&P 500.

8 A The regression model is as follows:

$$(\text{Analyst following})_i = b_0 + b_1\text{Size}_i + b_2(\text{D/E})_i + b_3\text{S\&P}_i + \varepsilon_i$$

where $(\text{Analyst following})_i$ is the natural log of (1 + number of analysts following company *i*); Size_i is the natural log of the market capitalization of company *i* in millions of dollars; $(\text{D/E})_i$ is the debt-to-equity ratio for company *i*, and $\text{S\&P}_i$ is a dummy variable with a value of 1 if the company *i* belongs to the S&P 500 Index and 0 otherwise.

B The appropriate null and alternative hypotheses are H_0: $b_3 = 0$ and H_a: $b_3 \neq 0$, respectively.

C The *t*-statistic to test the null hypothesis can be computed as follows:

$$\frac{\hat{b}_3 - b_3}{s_{\hat{b}_3}} = \frac{0.4218 - 0}{0.0919} = 4.5898$$

This value is, of course, the same as the value reported in the table. The regression has 500 observations and 4 regression coefficients, so the *t*-test has 500 – 4 = 496 degrees of freedom. At the 0.05 significance level, the critical value for the test statistic is between 1.96 and 1.97. Because the value of

the test statistic is 4.5898 we can reject the null hypothesis that $b_3 = 0$. Thus a company's membership in the S&P 500 appears to significantly influence the number of analysts who cover that company.

D The estimated model is

$$\begin{aligned}\left(\text{Analyst following}\right)_i &= -0.0075 + 0.2648\text{Size}_i - 0.1829\left(\text{D/E}\right)_i \\ &\quad + 0.4218\text{S\&P}_i + \varepsilon_i\end{aligned}$$

Therefore the prediction for number of analysts following the indicated company that is not part of the S&P 500 Index is

$$\begin{aligned}(\text{Analyst following})_i &= -0.0075 + 0.2648(\ln 10{,}000) - 0.1829(2/3) + \\ &\quad 0.4218(0) \\ &= -0.0075 + 2.4389 - 0.1219 + 0 \\ &= 2.3095\end{aligned}$$

Recalling that $(\text{Analyst following})_i$ is the natural log of $(1 + n_i)$, where n_i is the number of analysts following company i; it ensues (coding the company under consideration as 1) that $1 + n_1 = e^{2.3095} = 10.069$, approximately. Therefore, the prediction is that $n_1 = 10.069 - 1 = 9.069$, or about nine analysts.

Similarly, the prediction for the company that is included in the S&P 500 Index is

$$\begin{aligned}(\text{Analyst following})_i &= -0.0075 + 0.2648(\ln 10{,}000) - 0.1829(2/3) + \\ &\quad 0.4218(1) \\ &= -0.0075 + 2.4389 - 0.1219 + 0.4218 \\ &= 2.7313\end{aligned}$$

Coding the company that does belong to the S&P 500 as 2, $1 + n_2 = e^{2.7313} = 15.353$. Therefore, the prediction is that $n_2 = 15.353 - 1 = 14.353$, or about 14 analysts.

E There is no inconsistency in the coefficient on the size variable differing between the two regressions. The regression coefficient on an independent variable in a multiple regression model measures the expected net effect on the expected value of the dependent variable for a one-unit increase in that independent variable, after accounting for any effects of the other independent variables on the expected value of the dependent variable. The earlier regression had one fewer independent variable; after the effect of S&P 500 membership on the expected value of the dependent variable is taken into account, it is to be expected that the effect of the size variable on the dependent variable will change. What the regressions appear to indicate is that the net effect of the size variable on the expected analyst following diminishes when S&P 500 membership is taken into account.

9 A In a well-specified regression, the differences between the actual and predicted relationship should be random; the errors should not depend on the value of the independent variable. In this regression, the errors seem larger for smaller values of the book-to-market ratio. This finding indicates that we may have conditional heteroskedasticity in the errors, and consequently, the standard errors may be incorrect. We cannot proceed with hypothesis testing until we test for and, if necessary, correct for heteroskedasticity.

B A test for heteroskedasticity is to regress the squared residuals from the estimated regression equation on the independent variables in the regression. As seen in Section 4.1.2, Breusch and Pagan showed that, under the null hypothesis of no conditional heteroskedasticity, $n \times R^2$ (from the regression

of the squared residuals on the independent variables from the original regression) will be a χ^2 random variable, with the number of degrees of freedom equal to the number of independent variables in the regression.

C One method to correct for heteroskedasticity is to use robust standard errors. This method uses the parameter estimates from the linear regression model but corrects the standard errors of the estimated parameters to account for the heteroskedasticity. Many statistical software packages can easily compute robust standard errors.

10 The test statistic is nR^2, where n is the number of observations and R^2 is the R^2 of the regression of squared residuals. So, the test statistic is $52 \times 0.141 = 7.332$. Under the null hypothesis of no conditional heteroskedasticity, this test statistic is a χ^2 random variable. There are three degrees of freedom, the number of independent variables in the regression. Appendix C, at the end of this volume, shows that for a one-tailed test, the test statistic critical value for a variable from a χ^2 distribution with 3 degrees of freedom at the 0.05 significance level is 7.815. The test statistic from the Breusch–Pagan test is 7.332. So, we cannot reject the hypothesis of no conditional heteroskedasticity at the 0.05 level. Therefore, we do not need to correct for conditional heteroskedasticity.

11 A The test statistic is nR^2, where n is the number of observations and R^2 is the R^2 of the regression of squared residuals. So, the test statistic is $750 \times 0.006 = 4.5$. Under the null hypothesis of no conditional heteroskedasticity, this test statistic is a χ^2 random variable. Because the regression has only one independent variable, the number of degrees of freedom is equal to 1. Appendix C, at the end of this volume, shows that for a one-tailed test, the test statistic critical value for a variable from a χ^2 distribution with one degree of freedom at the 0.05 significance level is 3.841. The test statistic is 4.5. So, we can reject the hypothesis of no conditional heteroskedasticity at the 0.05 level. Therefore, we need to correct for conditional heteroskedasticity.

B Two different methods can be used to correct for the effects of conditional heteroskedasticity in linear regression models. The first method involves computing robust standard errors. This method corrects the standard errors of the linear regression model's estimated parameters to account for the conditional heteroskedasticity. The second method is generalized least squares. This method modifies the original equation in an attempt to eliminate the heteroskedasticity. The new, modified regression equation is then estimated under the assumption that heteroskedasticity is no longer a problem.

Many statistical software packages can easily compute robust standard errors (the first method), and we recommend using them.

12 A Because the value of the Durbin–Watson statistic is less than 2, we can say that the regression residuals are positively correlated. Because this statistic is fairly close to 2, however, we cannot say without a statistical test if the serial correlation is statistically significant.

B From January 1987 through December 2002, there are 16 years, or $16 \times 12 = 192$ monthly returns. Thus the sample analyzed is quite large. Therefore, the Durbin–Watson statistic is approximately equal to $2(1 - r)$, where r is the sample correlation between the regression residuals from one period and those from the previous period.

$$DW = 1.8953 \approx 2(1 - r)$$

So, $r \approx 1 - \text{DW}/2 = 1 - 1.8953/2 = 0.0524$. Consistent with our answer to Part A, the correlation coefficient is positive.

C Appendix E indicates that the critical values d_l and d_u for 100 observations when there is one independent variable are 1.65 and 1.69, respectively. Based on the information given in the problem, the critical values d_l and d_u for about 200 observations when there is one independent variable are about 1.74 and 1.78, respectively. Because the DW statistic of 1.8953 for our regression is above d_u, we fail to reject the null hypothesis of no positive serial correlation. Therefore, we conclude that there is no evidence of positive serial correlation for the error term.

13 A This problem is known as multicollinearity. When some linear combinations of the independent variables in a regression model are highly correlated, the standard errors of the independent coefficient estimates become quite large, even though the regression equation may fit rather well.

B The choice of independent variables presents multicollinearity concerns because market value of equity appears in both variables.

C The classic symptom of multicollinearity is a high R^2 (and significant F-statistic) even though the t-statistics on the estimated slope coefficients are insignificant. Here a significant F-statistic does not accompany the insignificant t-statistics, so the classic symptom is not present.

14 A To test the null hypothesis that all of the regression coefficients except for the intercept in the multiple regression model are equal to 0 (H_0: $b_1 = b_2 = b_3 = 0$) against the alternative hypothesis that at least one slope coefficient is not equal to 0, we must use an F-test.

$$F = \frac{\text{RSS}/k}{\text{SSE}/\left[n - (k+1)\right]} = \frac{0.1720/3}{0.8947/\left[156 - (3+1)\right]} = 9.7403$$

The F-statistic has degrees of freedom $F\{k, [n - (k + 1)]\} = F(3, 152)$. From the F-test table, the critical value for $F(3, 120) = 2.68$ and $F(3, 152)$ will be less than $F(3, 120)$, so we can reject at the 0.05 significance level the null hypothesis that the slope coefficients are all 0. Changes in the three independent variables are jointly statistically related to returns.

B None of the t-statistics are significant, but the F-statistic is significant. This suggests the possibility of multicollinearity in the independent variables.

C The apparent multicollinearity is very likely related to the inclusion of *both* the returns on the S&P 500 Index *and* the returns on a value-weighted index of all the companies listed on the NYSE, AMEX, and NASDAQ as independent variables. The value-weighting of the latter index, giving relatively high weights to larger companies such as those included in the S&P 500, may make one return series an approximate linear function of the other. By dropping one or the other of these two variables, we might expect to eliminate the multicollinearity.

15 A Your colleague is indicating that you have omitted an important variable from the regression. This problem is called the omitted variable bias. If the omitted variable is correlated with an included variable, the estimated values of the regression coefficients would be biased and inconsistent. Moreover, the estimates of standard errors of those coefficients would also be inconsistent. So, we cannot use either the coefficient estimates or the estimates of their standard errors to perform statistical tests.

B A comparison of the new estimates with the original estimates clearly indicates that the original model suffered from the omitted variable bias due to the exclusion of company size from that model. As the *t*-statistics of the new model indicate, company size is statistically significant. Further, for the debt-to-equity ratio, the absolute value of the estimated coefficient substantially increases from 0.1043 to 0.1829, while its standard error declines. Consequently, it becomes significant in the new model, in contrast to the original model, in which it is not significant at the 5 percent level. The value of the estimated coefficient of the S&P 500 dummy substantially declines from 1.2222 to 0.4218. These changes imply that size should be included in the model.

16 A You need to use a qualitative dependent variable. You could give a value of 1 to this dummy variable for a listing in the United States and a value of 0 for not listing in the United States.

B Because you are using a qualitative dependent variable, linear regression is not the right technique to estimate the model. One possibility is to use either a probit or a logit model. Both models are identical, except that the logit model is based on logistic distribution while the probit model is based on normal distribution. Another possibility is to use discriminant analysis.

17 C is correct. The predicted initial return (IR) is:

$$\begin{aligned} IR &= 0.0477 + (0.0150 \times 6) + (0.435 \times 0.04) - (0.0009 \times 40) + (0.05 \times 0.70) \\ &= 0.1541 \end{aligned}$$

18 B is correct. The 95% confidence interval is $0.435 \pm (0.0202 \times 1.96) = (0.395, 0.475)$.

19 C is correct. To test Hansen's belief about the direction and magnitude of the initial return, the test should be a one-tailed test. The alternative hypothesis is H_1: $b_j < 0.5$, and the null hypothesis is H_0: $b_j \geq 0.5$. The correct test statistic is: $t = (0.435 - 0.50)/0.0202 = -3.22$, and the critical value of the *t*-statistic for a one-tailed test at the 0.05 level is −1.645. The test statistic is significant, and the null hypothesis can be rejected at the 0.05 level of significance.

20 C is correct. The multiple *R*-squared for the regression is 0.36; thus, the model explains 36 percent of the variation in the dependent variable. The correlation between the predicted and actual values of the dependent variable is the square root of the *R*-squared or $\sqrt{0.36} = 0.60$.

21 A is correct. Chang is correct because the presence of conditional heteroskedasticity results in consistent parameter estimates, but biased (up or down) standard errors, *t*-statistics, and *F*-statistics.

22 A is correct. Chang is correct because a correlated omitted variable will result in biased and inconsistent parameter estimates and inconsistent standard errors.

23 B is correct.

The *F*-test is used to determine if the regression model as a whole is significant.

F = Mean square regression (MSR) ÷ Mean squared error (MSE)

MSE = SSE/[n − (k + 1)] = 19,048 ÷ 427 = 44.60

MSR = SSR/*k* = 1071 ÷ 3 = 357

F = 357 ÷ 44.60 = 8.004

The critical value for degrees of freedom of 3 and 427 with α = 0.05 (one-tail) is F = 2.63 from Exhibit 5. The calculated F is greater than the critical value, and Chiesa should reject the null hypothesis that all regression coefficients are equal to zero.

24 B is correct. The Durbin–Watson test used to test for serial correlation in the error term, and its value reported in Exhibit 1 is 1.65. For no serial correlation, DW is approximately equal to 2. If DW < d_l, *the error terms are positively serially correlated.* Because the DW = 1.65 is less than d_l = 1.827 for n = 431 (see Exhibit 2), Chiesa should reject the null hypothesis of no serial correlation and conclude that there is evidence of positive serial correlation among the error terms.

25 B is correct. The coefficient for the Pres party dummy variable (3.17) represents the increment in the mean value of the dependent variable related to the Democratic Party holding the presidency. In this case, the excess stock market return is 3.17 percent greater in Democratic presidencies than in Republican presidencies.

26 B is correct. The confidence interval is computed as $a_1 \pm s(a_1) \times t(95\%,\infty)$. From Exhibit 1, a_1 = 3.04 and $t(a_1)$ = 4.52, resulting in a standard error of $a_1 = s(a_1)$ = 3.04/4.52 = 0.673. The critical value for t from Exhibit 3 is 1.96 for p = 0.025. The confidence interval for a_1 is 3.04 ± 0.673 × 1.96 = 3.04 ± 1.31908 or from 1.72092 to 4.35908.

27 C is correct. The default spread is typically larger when business conditions are poor, i.e., a greater probability of default by the borrower. The positive sign for default spread (see Exhibit 1) indicates that expected returns are positively related to default spreads, meaning that excess returns are greater when business conditions are poor.

28 C is correct. Predictions in a multiple regression model are subject to both parameter estimate uncertainty and regression model uncertainty.

29 C is correct. The F-statistic is

$$F = \frac{\text{RSS}/k}{\text{SSE}/\left[n-(k+1)\right]} = \frac{714.169/8}{1583.113/546} = \frac{89.2712}{2.8995} = 30.79$$

Because F = 30.79 exceeds the critical F of 1.96, the null hypothesis that the regression coefficients are all 0 is rejected at the 0.05 significance level.

30 B is correct. The estimated coefficients for the dummy variables show the estimated difference between the returns on different types of funds. The growth dummy takes the value of 1 for growth funds and 0 for the value fund. Exhibit 1 shows a growth dummy coefficient of 2.4368. The estimated difference between the return of growth funds and value funds is thus 2.4368.

31 B is correct. The R^2 is expected to increase, not decline, with a new independent variable. The other two potential consequences Honoré describes are correct.

32 C is correct. Using dummy variables to distinguish among n categories would best capture the ability of the Morningstar rating system to predict mutual fund performance. We need n − 1 dummy variables to distinguish among n categories. In this case, there are five possible ratings and we need four dummy variables. Adding an independent variable that has a value equal to the number of stars in the rating of each fund is not appropriate because if the coefficient for this variable is positive, this method assumes that the extra return for a

two-star fund is twice that of a one-star fund, the extra return for a three-star fund is three times that of a one-star fund, and so forth, which is not a reasonable assumption.

33 A is correct. Heteroskedasticity causes the *F*-test for the overall significance of the regression to be unreliable. It also causes the *t*-tests for the significance of individual regression coefficients to be unreliable because heteroskedasticity introduces bias into estimators of the standard error of regression coefficients.

34 A is correct. The model in Exhibit 2 does not have a lagged dependent variable. Positive serial correlation will, for such a model, not affect the consistency of the estimated coefficients. Thus, the coefficients will not need to be corrected for serial correlation. Positive serial correlation will, however, cause the standard errors of the regression coefficients to be understated; thus, the corresponding *t*-statistics will be inflated.

35 A is correct. The critical Durbin–Watson (D–W) values are $d_l = 1.63$ and $d_u = 1.72$. Because the estimated D–W value of 1.81 is greater than $d_u = 1.73$ (and less than 2), she fails to reject the null hypothesis of no serial correlation.

36 B is correct. Probit and logit models are used for models with qualitative dependent variables, such as models in which the dependent variable can have one of two discreet outcomes (i.e., 0 or 1). The analysis in the two exhibits are explaining security returns, which are continuous (not 0 or 1) variables.

37 A is correct. Varden expects to find that CEO tenure is positively related to the firm's ROE. If he is correct, the regression coefficient for tenure, b_2, will be greater than zero ($b_2 > 0$) and statistically significant. The null hypothesis supposes that the "suspected" condition is not true, so the null hypothesis should state the variable is less than or equal to zero. The *t*-statistic for tenure is 2.308, significant at the 0.027 level, meeting Varden's 0.05 significance requirement. Varden should reject the null hypothesis.

38 C is correct. The *t*-statistic for tenure is 2.308, indicating significance at the 0.027 level but not the 0.01 level. The *t*-statistic for ESG is 1.201, with a *p*-value of 0.238, which means we fail to reject the null hypothesis for ESG at the 0.01 significance level.

39 B is correct. The *t*-statistic for tenure is 2.308, which is significant at the 0.027 level. The *t*-statistic for ESG is 1.201, with a *p*-value of 0.238. This result is not significant at the 0.05 level.

40 C is correct. The regression equation is as follows:

$$\hat{Y}_i = 9.442 + 0.069X_{1i} + 0.681X_{2i}$$

$$\begin{aligned}\text{ROE} &= 9.442 + 0.069(\text{ESG}) + 0.681(\text{Tenure})\\ &= 9.442 + 0.069(55) + 0.681(10.5)\\ &= 9.442 + 3.795 + 7.151\\ &= 20.388.\end{aligned}$$

41 B is correct. When you add an additional independent variable to the regression model, the amount of unexplained variance will decrease, provided the new variable explains any of the previously unexplained variation. This result occurs as long as the new variable is even slightly correlated with the dependent variable. Exhibit 2 indicates the dividend growth rate is correlated with the dependent variable, ROE. Therefore, R^2 will increase.

Adjusted R^2, however, may not increase and may even decrease if the relationship is weak. This result occurs because in the formula for adjusted R^2, the new variable increases k (the number of independent variables) in the denominator, and the increase in R^2 may be insufficient to increase the value of the formula.

$$\text{adjusted } R^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\left(1 - R^2\right)$$

42 C is correct. Exhibit 1 indicates that the *F*-statistic of 4.161 is significant at the 0.05 level. A significant *F*-statistic means at least one of the independent variables is significant.

43 C is correct. In a multiple linear regression (as compared with simple regression), R^2 is less appropriate as a measure of whether a regression model fits the data well. A high adjusted R^2 does not necessarily indicate that the regression is well specified in the sense of including the correct set of variables. The *F*-test is an appropriate test of a regression's overall significance in either simple or multiple regressions.

44 C is correct. Multiple linear regression assumes that the relationship between the dependent variable and each of the independent variables is linear. Varden believes that this is not true for dividend growth because he believes the relationship may be different in firms with a long-standing CEO. Multiple linear regression also assumes that the independent variables are not random. Varden states that he believes CEO tenure is a random variable.

45 B is correct. If we use adjusted R^2 to compare regression models, it is important that the dependent variable be defined the same way in both models and that the sample sizes used to estimate the models are the same. Varden's first model was based on 40 observations, whereas the second model was based on 500.