PRACTICE PROBLEMS

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

A Compute the coefficient of determination.

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

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

D Compute the sample standard deviation of monthly energy consumption.

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

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

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

B Define Total SS.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Explain how the bad data affected the results.

The following information relates to Questions 4–9

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

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

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

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

4 What is the value of the coefficient of determination?

A 0.8261.

B 0.7436.

C 0.8623.

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

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

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

A –0.7436.

B 0.7436.

C 0.8623.

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

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

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

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

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

A 0.007 + 0.103(5.0) = 0.524.

B 0.077 – 0.826(5.0) = –4.054.

C 0.077 + 0.826(5.0) = 4.207.

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

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

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

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

The following information relates to Questions 10–14

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

Golub directs his assistant, Jill Batten, to study the relationships between Stellar monthly common stock returns versus the previous month's percent change in the US Consumer Price Index for Energy (CPIENG), and Stellar monthly common stock returns versus the previous month's percent change in the US Producer Price Index for Crude Energy Materials (PPICEM). Golub wants Batten to run both a correlation and a linear regression analysis. In response, Batten compiles the summary statistics shown in Exhibit 1 for the 248 months between January 1980 and August 2000. All of the data are in decimal form, where 0.01 indicates a 1 percent return. Batten also runs a regression analysis using Stellar monthly returns as the dependent variable and the monthly change in CPIENG as the independent variable. Exhibit 2 displays the results of this regression model.

Exhibit 1 Descriptive Statistics

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

Exhibit 1 (Continued)

	Monthly Return Stellar Common Stock	Lagged Monthly Change	
		CPIENG	PPICEM
Covariance, Stellar vs. PPICEM	–0.00048		
Covariance, CPIENG vs. PPICEM	0.00044		
Correlation, Stellar vs. CPIENG	–0.1452		

Exhibit 2 Regression Analysis with CPIENG

Regression Statistics

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

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

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

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

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

A 0.0073 (0.73 percent).

B 0.0138 (1.38 percent).

C 0.0203 (2.03 percent).

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

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

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

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

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

A the residuals from the regression is 0.0710.

B values estimated from the regression is 0.0710.

C Stellar's observed common stock returns is 0.0710.

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

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

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

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

The following information relates to Questions 15–24

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

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

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

Exhibit 1 Summary Statistics

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

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

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

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

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

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

Interpretation 1 Companies' higher debt ratios cause lower short interest ratios.

Interpretation 2 Companies' higher short interest ratios cause higher debt ratios.

Interpretation 3 Companies with higher debt ratios tend to have lower short interest ratios.

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

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

A horizontal.

B upward sloping.

C downward sloping.

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

A –9.2430.

B –0.1886.

C 8.4123.

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

A –0.3054.

B 0.0933.

C 0.3054.

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

A Interpretation 1

B Interpretation 2

C Interpretation 3

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

A intercept.

B debt ratio.

C short interest ratio.

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

A 48.

B 49.

C 50.

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

A −1.0199.

B −0.3947.

C 1.4528.

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

A The average short interest ratio is 5.4975.

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

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

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

A 3.8339.

B 5.4975.

C 6.2462.

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

A −2.2219.

B 3.5036.

C 4.9367.

The following information relates to Questions 25–30

Elena Vasileva recently joined EnergyInvest as a junior portfolio analyst. Vasileva's supervisor asks her to evaluate a potential investment opportunity in Amtex, a multinational oil and gas corporation based in the US. Vasileva's supervisor suggests using regression analysis to examine the relation between Amtex shares and returns on crude oil.

Vasileva notes the following assumptions of regression analysis:

Assumption 1 The error term is uncorrelated across observations.

Assumption 2 The variance of the error term is the same for all observations.

Assumption 3 The expected value of the error term is equal to the mean value of the dependent variable.

Vasileva runs a regression of Amtex share returns on crude oil returns using the monthly data she collected. Selected data used in the regression are presented in Exhibit 1, and selected regression output is presented in Exhibit 2.

Exhibit 1 Selected Data for Crude Oil Returns and Amtex Share Returns

	Oil Return (X_i)	Amtex Return (Y_i)	Cross-Product $(X_i - \bar{X})(Y_i - \bar{Y})$	Predicted Amtex Return $(\hat{Y})$	Regression Residual $(Y_i - \hat{Y})$	Squared Residual $(Y_i - \hat{Y})^2$
Month 1	–0.032000	0.033145	–0.000388	0.002011	–0.031134	0.000969
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Month 36	0.028636	0.062334	0.002663	0.016282	–0.046053	0.002121
Sum			0.085598			0.071475
Average	–0.018056	0.005293				

Exhibit 2 Selected Regression Output Dependent Variable: Amtex Share Return

	Coefficient	Standard Error
Intercept	0.0095	0.0078
Oil return	0.2354	0.0760

Note: The critical *t*-value for a one-sided *t*-test at the 5% significance level is 1.691.

Vasileva expects the crude oil return next month, Month 37, to be –0.01. She computes the variance of the prediction error to be 0.0022.

25 Which of Vasileva's assumptions regarding regression analysis is *incorrect*?

A Assumption 1

B Assumption 2

C Assumption 3

26 Based on Exhibit 1, the standard error of the estimate is *closest* to:

A 0.044558.

B 0.045850.

C 0.050176.

27 Based on Exhibit 2, Vasileva should reject the null hypothesis that:

A the slope is less than or equal to 0.15.

B the intercept is less than or equal to 0.

C crude oil returns do not explain Amtex share returns.

28 Based on Exhibit 2, Vasileva should compute the:

A coefficient of determination to be 0.4689.

B 95% confidence interval for the intercept to be –0.0037 to 0.0227.

C 95% confidence interval for the slope coefficient to be 0.0810 to 0.3898.

29 Based on Exhibit 2 and Vasileva's prediction of the crude oil return for month 37, the estimate of Amtex share return for month 37 is *closest* to:

A –0.0024.

B 0.0071.

C 0.0119.

30 Using information from Exhibit 2, Vasileva should compute the 95% prediction interval for Amtex share return for month 37 to be:

A –0.0882 to 0.1025.

B –0.0835 to 0.1072.

C 0.0027 to 0.0116.

The following information relates to Question 31–33

Doug Abitbol is a portfolio manager for Polyi Investments, a hedge fund that trades in the United States. Abitbol manages the hedge fund with the help of Robert Olabudo, a junior portfolio manager.

Abitbol looks at economists' inflation forecasts and would like to examine the relationship between the US Consumer Price Index (US CPI) consensus forecast and actual US CPI using regression analysis. Olabudo estimates regression coefficients to test whether the consensus forecast is unbiased. Regression results are presented in Exhibit 1. Additionally, Olabudo calculates the 95% prediction interval of the actual CPI using a US CPI consensus forecast of 2.8.

Exhibit 1 Regression Output: Estimating US CPI

Regression Statistics	
Multiple *R*	0.9929
R-squared	0.9859
Standard error of estimate	0.0009
Observations	60

Exhibit 1 (Continued)

	Coefficients	Standard Error	*t*-Statistic
Intercept	0.0001	0.0002	0.5351
US CPI consensus forecast	0.9830	0.0155	63.6239

Notes:

1. The absolute value of the critical value for the *t*-statistic is 2.0 at the 5% level of significance.
2. The standard deviation of the US CPI consensus forecast is $s_x = 0.7539$.
3. The mean of US CPI consensus forecast is $\overline{X} = 1.3350$.

To conclude their meeting, Abitbol and Olabudo discuss the limitations of regression analysis. Olabudo notes the following limitations of regression analysis:

Limitation 1: Public knowledge of regression relationships may negate their future usefulness.

Limitation 2: Hypothesis tests and predictions based on linear regression will not be valid if regression assumptions are violated.

31 Based on Exhibit 1, Olabudo should:

A conclude that the inflation predictions are unbiased.

B reject the null hypothesis that the slope coefficient equals 1.

C reject the null hypothesis that the intercept coefficient equals 0.

32 Based on Exhibit 1, Olabudo should calculate a prediction interval for the actual US CPI *closest* to:

A 2.7506 to 2.7544.

B 2.7521 to 2.7529.

C 2.7981 to 2.8019.

33 Which of Olabudo's noted limitations of regression analysis is correct?

A Only Limitation 1

B Only Limitation 2

C Both Limitation 1 and Limitation 2

SOLUTIONS

1 A The coefficient of determination is

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

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

C The standard error of the estimate is

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

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

D The sample variance of the dependent variable is

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

16 B is correct. The sample covariance is calculated as

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

17 A is correct. For a regression with one independent variable, the correlation is the same as the Multiple *R* with the sign of the slope coefficient. Because the slope coefficient is negative, the correlation is –0.3054.

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

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

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

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

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

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

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

25 C is correct. The assumptions of the linear regression model are that the 1) the relationship between the dependent variable and the independent variable is linear in the parameters b_0 and b_1; 2) the independent variable is not random; 3) the expected value of the error term is 0; 4) the variance of the error term is the same for all observations; 5) the error term is uncorrelated across observations; and 6) the error term is normally distributed. Assumption 3 is incorrect because the expected value of the error term is assumed to be zero, not equal to the mean of the dependent variable.

26 B is correct. The standard error of the estimate (SEE) for a linear regression model with one independent variable is calculated as:

$$\begin{aligned}\text{SEE} &= \sqrt{\frac{\sum_{i=1}^{n}\left(Y - \hat{b}_0 - \hat{b}_1 X_i\right)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y - \hat{Y}\right)^2}{n-2}} \\ &= \sqrt{\frac{0.071475}{34}} \\ &= 0.045850\end{aligned}$$

27 C is correct. Crude oil returns explain the Amtex share returns if the slope coefficient is statistically different from zero. The slope coefficient is 0.2354 and is statistically different from zero because the absolute value of the *t*-statistic of 3.0974 is higher than the critical *t*-value of 2.032 (two-sided test for n – 2 = 34 degrees of freedom and a 5% significance level):

$$t\text{-statistic} = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{0.2354 - 0.0000}{0.0760} = 3.0974$$

Therefore, Vasileva should reject the null hypothesis that crude oil returns do not explain Amtex share returns because the slope coefficient is statistically different from zero.

28 C is correct. The confidence interval for the slope coefficient is calculated as:

$$\text{Confidence interval} = \hat{b}_1 \pm t_c s_{b_1}$$

Where $\hat{b}_1 = 0.2354$, $s_{b_1} = 0.0760$ and $t_c = 2.032$

The lower limit for the confidence interval = 0.2354 – (2.032 × 0.0760) = 0.0810

The upper limit for the confidence interval = 0.2354 + (2.032 × 0.0760) = 0.3898

29 B is correct. The predicted value of the dependent variable, Amtex share return, given the value of the independent variable, crude oil return, of –0.01, is calculated as:

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_i = 0.0095 + \left(0.2354 \times \left(-0.01\right)\right) = 0.0071$$

30 A is correct. The 95% prediction interval for the dependent variable given a certain value of the independent variable is calculated as:

$$\text{Prediction interval} = \hat{Y} \pm t_c s_f \text{ and the predicted value } \hat{Y} = \hat{b}_0 + \hat{b}_1 X_i$$

Therefore:

Predicted value = 0.0095 + (0.2354 × (–0.01)) = 0.0071

$s_f = (0.0022)^{0.5} = 0.0469$

$t_c = 2.032$

The lower limit for the prediction interval = 0.0071 – (2.032 × 0.0469) = –0.0882

The upper limit for the prediction interval = 0.0071 + (2.032 × 0.0469) = 0.1025

31 A is correct. If the consensus inflation forecast is unbiased, then the intercept, b_0, should equal 0, and the slope coefficient, b_1, should equal 1. The *t*-statistic for the intercept coefficient is 0.5351, which is less than the critical *t*-value of 2.0, so the intercept coefficient is not statistically different than 0. To test whether the slope coefficient equals 1, the *t*-statistic is calculated as:

$$t = \left(\hat{b}_1 - b_1\right) \Big/ s_{\hat{b}_1} = \left(0.9830 - 1\right)/0.0155 = -1.0968$$

Because the absolute value of the t-statistic of –1.0968 is less than the critical *t*-value of 2.0, the slope coefficient is not statistically different than 1. Therefore, Olabudo can conclude that the inflation forecasts are unbiased.

32 A is correct. The prediction interval for inflation is calculated in three steps:

Step 1 – Make the prediction given the US CPI forecast of 2.8:

$$\begin{aligned}\widehat{Y} &= b_0 + b_1 X \\ &= 0.0001 + \left(0.9830 \times 2.8\right) \\ &= 2.7525\end{aligned}$$

Step 2 – Compute the variance of the prediction error:

$$s_f^2 = s^2\left[1 + (1/n) + \left(\left(X - \bar{X}\right)^2\right) \Big/ \left((n-1) \times s_x^2\right)\right]$$

$$s_f^2 = 0.0009^2\left[1 + (1/60) + \left(\left(2.8 - 1.3350\right)^2\right) \Big/ \left((60-1) \times 0.7539^2\right)\right]$$

$$s_f^2 = 0.00000088$$

$$s_f = 0.0009$$

Step 3 – Compute the prediction interval:

$$\hat{Y} \pm t_c \times s_f$$

2.7525 ± (2.0 x 0.0009)

2.7525 – (2.0 x 0.0009) = 2.7506; lower bound

2.7525 + (2.0 x 0.0009) = 2.7544; upper bound

So, given the US CPI forecast of 2.8, the 95% prediction interval is 2.7506 to 2.7544.

33 C is correct. Public knowledge of regression relationships may negate their future usefulness in an investment context. Also, if regression assumptions are violated, hypothesis tests and predictions based on linear regression will not be valid.

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)

	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)

(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)

	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)

(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)

	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

(continued)

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

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

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

$$\left(\text{Analyst following}\right)_i = -0.0075 + 0.2648\text{Size}_i - 0.1829\left(\text{D/E}\right)_i + 0.4218\text{S\&P}_i + \varepsilon_i$$

Therefore the prediction for number of analysts following the indicated company that is not part of the S&P 500 Index is

$$\begin{aligned}(\text{Analyst following})_i &= -0.0075 + 0.2648(\ln 10{,}000) - 0.1829(2/3) + 0.4218(0)\\ &= -0.0075 + 2.4389 - 0.1219 + 0\\ &= 2.3095\end{aligned}$$

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

$$\begin{aligned}(\text{Analyst following})_i &= -0.0075 + 0.2648(\ln 10{,}000) - 0.1829(2/3) + 0.4218(1)\\ &= -0.0075 + 2.4389 - 0.1219 + 0.4218\\ &= 2.7313\end{aligned}$$

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.

$$\text{DW} = 1.8953 \approx 2(1 - r)$$

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:

$$\begin{aligned}\text{IR} &= 0.0477 + (0.0150 \times 6) + (0.435 \times 0.04) - (0.0009 \times 40) + (0.05 \times 0.70)\\ &= 0.1541\end{aligned}$$

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 $3.04 \pm 0.673 \times 1.96 = 3.04 \pm 1.31908$ 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.

PRACTICE PROBLEMS

Note: In the Problems and Solutions for this reading, we use the hat (ˆ) to indicate an estimate if we are trying to differentiate between an estimated and an actual value. However, we suppress the hat when we are clearly showing regression output.

1 The civilian unemployment rate (UER) is an important component of many economic models. Table 1 gives regression statistics from estimating a linear trend model of the unemployment rate: $\text{UER}_t = b_0 + b_1 t + \varepsilon_t$.

Table 1 Estimating a Linear Trend in the Civilian Unemployment Rate Monthly Observations, January 1996–December 2000

Regression Statistics	
R-squared	0.9314
Standard error	0.1405
Observations	60
Durbin–Watson	0.9099

	Coefficient	Standard Error	*t*-Statistic
Intercept	5.5098	0.0367	150.0363
Trend	–0.0294	0.0010	–28.0715

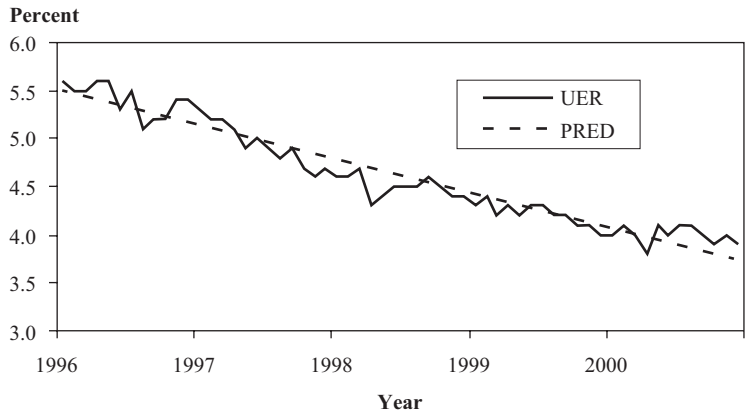
A Using the regression output in the above table, what is the model's prediction of the unemployment rate for July 1996?

B How should we interpret the Durbin–Watson (DW) statistic for this regression? What does the value of the DW statistic say about the validity of a *t*-test on the coefficient estimates?

2 Figure 1 compares the predicted civilian unemployment rate (PRED) with the actual civilian unemployment rate (UER) from January 1996 to December 2000. The predicted results come from estimating the linear time trend model $\text{UER}_t = b_0 + b_1 t + \varepsilon_t$.

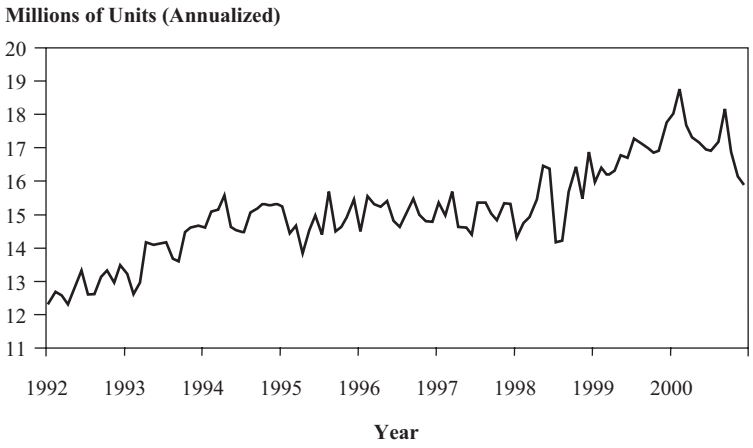
What can we conclude about the appropriateness of this model?

Figure 1 Predicted and Actual Civilian Unemployment Rates



3 You have been assigned to analyze automobile manufacturers and as a first step in your analysis, you decide to model monthly sales of lightweight vehicles to determine sales growth in that part of the industry. Figure 2 gives lightweight vehicle monthly sales (annualized) from January 1992 to December 2000.

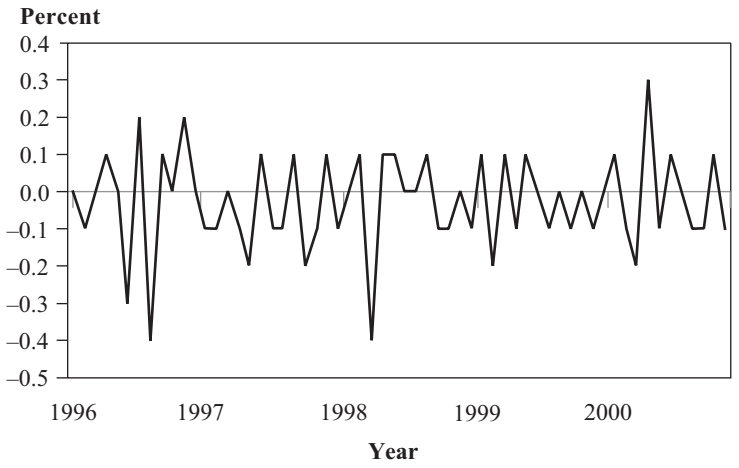
Figure 2 Lightweight Vehicle Sales



Monthly sales in the lightweight vehicle sector, Sales_t, have been increasing over time, but you suspect that the growth rate of monthly sales is relatively constant. Write the simplest time-series model for Sales_t that is consistent with your perception.

4 Figure 3 shows a plot of the first differences in the civilian unemployment rate (UER) between January 1996 and December 2000, $\Delta\text{UER}_t = \text{UER}_t - \text{UER}_{t-1}$.

Figure 3 Change in Civilian Unemployment Rate



A Has differencing the data made the new series, ΔUER_t, covariance stationary? Explain your answer.

B Given the graph of the change in the unemployment rate shown in the figure, describe the steps we should take to determine the appropriate autoregressive time-series model specification for the series ΔUER_t.

5 Table 2 gives the regression output of an AR(1) model on first differences in the unemployment rate. Describe how to interpret the DW statistic for this regression.

Table 2 Estimating an AR(1) Model of Changes in the Civilian Unemployment Rate Monthly Observations, March 1996–December 2000

Regression Statistics	
R-squared	0.2184
Standard error	0.1202
Observations	58
Durbin–Watson	2.1852

	Coefficient	Standard Error	*t*-Statistic
Intercept	–0.0405	0.0161	–2.5110
ΔUER_{t-1}	–0.4674	0.1181	–3.9562

6 Assume that changes in the civilian unemployment rate are covariance stationary and that an AR(1) model is a good description for the time series of changes in the unemployment rate. Specifically, we have $\Delta UER_t = -0.0405 - 0.4674\Delta UER_{t-1}$ (using the coefficient estimates given in the previous problem). Given this equation, what is the mean-reverting level to which changes in the unemployment rate converge?

7 Suppose the following model describes changes in the civilian unemployment rate: $\Delta UER_t = -0.0405 - 0.4674\Delta UER_{t-1}$. The current change (first difference) in the unemployment rate is 0.0300. Assume that the mean-reverting level for changes in the unemployment rate is –0.0276.

A What is the best prediction of the next change?

B What is the prediction of the change following the next change?

C Explain your answer to Part B in terms of equilibrium.

8 Table 3 gives the actual sales, log of sales, and changes in the log of sales of Cisco Systems for the period 1Q:2001 to 4Q:2001.

Table 3

Date Quarter: Year	Actual Sales ($ Millions)	Log of Sales	Changes in Log of Sales $\Delta\ln(\text{Sales}_t)$
1Q:2001	6,519	8.7825	0.1308
2Q:2001	6,748	8.8170	0.0345
3Q:2001	4,728	8.4613	–0.3557
4Q:2001	4,298	8.3659	–0.0954
1Q:2002			
2Q:2002			

Forecast the first- and second-quarter sales of Cisco Systems for 2002 using the regression $\Delta\ln(\text{Sales}_t) = 0.0661 + 0.4698\Delta\ln(\text{Sales}_{t-1})$.

9 Table 4 gives the actual change in the log of sales of Cisco Systems from 1Q:2001 to 4Q:2001, along with the forecasts from the regression model $\Delta\ln(\text{Sales}_t) = 0.0661 + 0.4698\Delta\ln(\text{Sales}_{t-1})$ estimated using data from 3Q:1991 to 4Q:2000. (Note that the observations after the fourth quarter of 2000 are out of sample.)

Table 4

Date	Actual Values of Changes in the Log of Sales $\Delta\ln(\text{Sales}_t)$	Forecast Values of Changes in the Log of Sales $\Delta\ln(\text{Sales}_t)$
1Q:2001	0.1308	0.1357
2Q:2001	0.0345	0.1299
3Q:2001	–0.3557	0.1271
4Q:2001	–0.0954	0.1259

A Calculate the RMSE for the out-of-sample forecast errors.

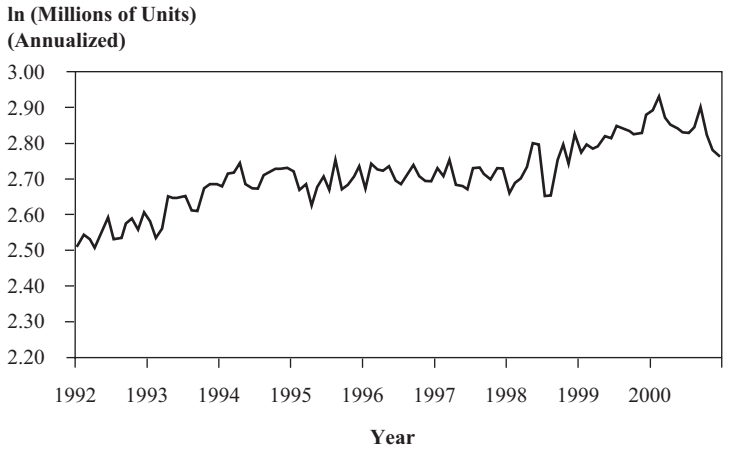
B Compare the forecasting performance of the model given with that of another model having an out-of-sample RMSE of 20 percent.

10 A The AR(1) model for the civilian unemployment rate, $\Delta UER_t = -0.0405 - 0.4674\Delta UER_{t-1}$, was developed with five years of data. What would be the drawback to using the AR(1) model to predict changes in the civilian unemployment rate 12 months or more ahead, as compared with one month ahead?

B For purposes of estimating a predictive equation, what would be the drawback to using 30 years of civilian unemployment data rather than only five years?

11 Figure 4 shows monthly observations on the natural log of lightweight vehicle sales, ln (Sales$_t$), for the period January 1992 to December 2000.

Figure 4 Lightweight Vehicle Sales

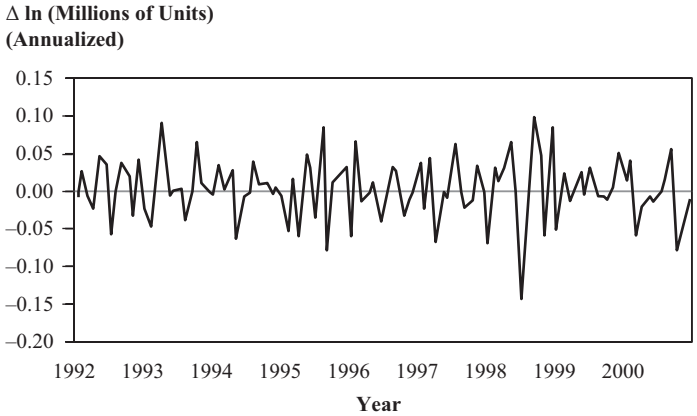


A Using the figure, comment on whether the specification ln (Sales$_t$) = b_0 + b_1[ln (Sales$_{t-1}$)] + ε_t is appropriate.

B State an appropriate transformation of the time series.

12 Figure 5 shows a plot of first differences in the log of monthly lightweight vehicle sales over the same period as in Problem 11. Has differencing the data made the resulting series, Δln (Sales$_t$) = ln (Sales$_t$) – ln (Sales$_{t-1}$), covariance stationary?

Figure 5 Change in Natural Log of Lightweight Vehicle Sales



13 Using monthly data from January 1992 to December 2000, we estimate the following equation for lightweight vehicle sales: Δln (Sales$_t$) = 2.7108 + 0.3987Δln (Sales$_{t-1}$) + ε_t. Table 5 gives sample autocorrelations of the errors from this model.

Table 5 Different Order Autocorrelations of Differences in the Logs of Vehicle Sales

Lag	Autocorrelation	Standard Error	t-Statistic
1	0.9358	0.0962	9.7247
2	0.8565	0.0962	8.9005
3	0.8083	0.0962	8.4001
4	0.7723	0.0962	8.0257
5	0.7476	0.0962	7.7696
6	0.7326	0.0962	7.6137
7	0.6941	0.0962	7.2138
8	0.6353	0.0962	6.6025
9	0.5867	0.0962	6.0968
10	0.5378	0.0962	5.5892
11	0.4745	0.0962	4.9315
12	0.4217	0.0962	4.3827

A Use the information in the table to assess the appropriateness of the specification given by the equation.

B If the residuals from the AR(1) model above violate a regression assumption, how would you modify the AR(1) specification?

14 Figure 6 shows the quarterly sales of Cisco Systems from 1Q:1991 to 4Q:2000.

Figure 6 Quarterly Sales at Cisco

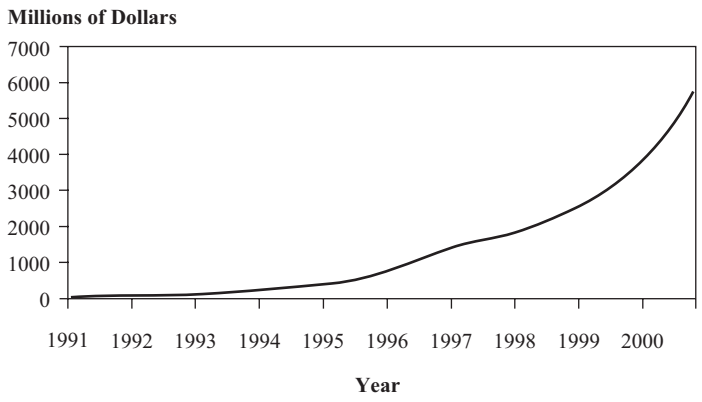


Table 6 gives the regression statistics from estimating the model $\Delta\ln(\text{Sales}_t) = b_0 + b_1\Delta\ln(\text{Sales}_{t-1}) + \varepsilon_t$.

Table 6 Change in the Natural Log of Sales for Cisco Systems Quarterly Observations, 3Q:1991–4Q:2000

Regression Statistics	
R-squared	0.2899
Standard error	0.0408
Observations	38
Durbin–Watson	1.5707

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0661	0.0175	3.7840
$\Delta \ln (\text{Sales}_{t-1})$	0.4698	0.1225	3.8339

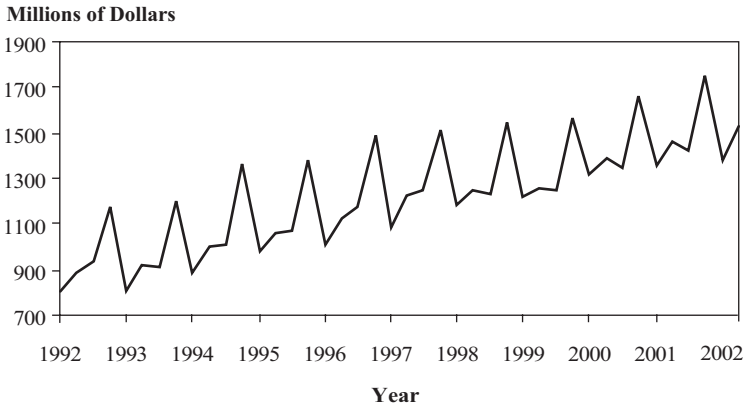
A Describe the salient features of the quarterly sales series.

B Describe the procedures we should use to determine whether the AR(1) specification is correct.

C Assuming the model is correctly specified, what is the long-run change in the log of sales toward which the series will tend to converge?

15 Figure 7 shows the quarterly sales of Avon Products from 1Q:1992 to 2Q:2002. Describe the salient features of the data shown.

Figure 7 Quarterly Sales at Avon



16 Table 7 below shows the autocorrelations of the residuals from an AR(1) model fit to the changes in the gross profit margin (GPM) of The Home Depot, Inc.

Table 7 Autocorrelations of the Residuals from Estimating the Regression $\Delta GPM_t = 0.0006 - 0.3330_1 \Delta GPM_{t-1} + \varepsilon_t$ 1Q:1992–4Q:2001 (40 Observations)

Lag	Autocorrelation
1	–0.1106
2	–0.5981

Table 7 (Continued)

Lag	Autocorrelation
3	–0.1525
4	0.8496
5	–0.1099

Table 8 shows the output from a regression on changes in the GPM for Home Depot, where we have changed the specification of the AR regression.

Table 8 Change in Gross Profit Margin for Home Depot 1Q:1992–4Q:2001

Regression Statistics	
R-squared	0.9155
Standard error	0.0057
Observations	40
Durbin–Watson	2.6464

	Coefficient	Standard Error	*t*-Statistic
Intercept	–0.0001	0.0009	–0.0610
ΔGPM_{t-1}	–0.0608	0.0687	–0.8850
ΔGPM_{t-4}	0.8720	0.0678	12.8683

A Identify the change that was made to the regression model.

B Discuss the rationale for changing the regression specification.

17 Suppose we decide to use an autoregressive model with a seasonal lag because of the seasonal autocorrelation in the previous problem. We are modeling quarterly data, so we estimate Equation 15: $(\ln \text{Sales}_t - \ln \text{Sales}_{t-1}) = b_0 + b_1(\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}) + b_2(\ln \text{Sales}_{t-4} - \ln \text{Sales}_{t-5}) + \varepsilon_t$. Table 9 shows the regression statistics from this equation.

Table 9 Log Differenced Sales: AR(1) Model with Seasonal Lag Johnson & Johnson Quarterly Observations, January 1985–December 2001

Regression Statistics	
R-squared	0.4220
Standard error	0.0318

(continued)

Table 9 (Continued)

Regression Statistics	
Observations	68
Durbin–Watson	1.8784

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0121	0.0053	2.3055
Lag 1	–0.0839	0.0958	–0.8757
Lag 4	0.6292	0.0958	6.5693

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	*t*-Statistic
1	0.0572	0.1213	0.4720
2	–0.0700	0.1213	–0.5771
3	0.0065	0.1213	–0.0532
4	–0.0368	0.1213	–0.3033

A Using the information in Table 9, determine if the model is correctly specified.

B If sales grew by 1 percent last quarter and by 2 percent four quarters ago, use the model to predict the sales growth for this quarter.

18 Describe how to test for autoregressive conditional heteroskedasticity (ARCH) in the residuals from the AR(1) regression on first differences in the civilian unemployment rate, $\Delta\text{UER}_t = b_0 + b_1\Delta\text{UER}_{t-1} + \varepsilon_t$.

19 Suppose we want to predict the annualized return of the five-year T-bill using the annualized return of the three-month T-bill with monthly observations from January 1993 to December 2002. Our analysis produces the data shown in Table 10.

Table 10 Regression with 3-Month T-Bill as the Independent Variable and 5-Year Treasury Bill as the Dependent Variable Monthly Observations, January 1993 to December 2002

Regression Statistics	
R-squared	0.5829
Standard error	0.6598
Observations	120
Durbin–Watson	0.1130

	Coefficient	Standard Error	*t*-Statistic
Intercept	3.0530	0.2060	14.8181
Three-month	0.5722	0.0446	12.8408

Can we rely on the regression model in Table 10 to produce meaningful predictions? Specify what problem might be a concern with this regression.

The following information relates to Questions 20–26

Angela Martinez, an energy sector analyst at an investment bank, is concerned about the future level of oil prices and how it might affect portfolio values. She is considering whether to recommend a hedge for the bank portfolio's exposure to changes in oil prices. Martinez examines West Texas Intermediate (WTI) monthly crude oil price data, expressed in US dollars per barrel, for the 181-month period from August 2000 through August 2015. The end-of-month WTI oil price was $51.16 in July 2015 and $42.86 in August 2015 (Month 181).

After reviewing the time-series data, Martinez determines that the mean and variance of the time series of oil prices are not constant over time. She then runs the following four regressions using the WTI time-series data.

- Linear trend model: Oil price$_t = b_0 + b_1 t + e_t$
- Log-linear trend model: ln Oil price$_t = b_0 + b_1 t + e_t$
- AR(1) model: Oil price$_t = b_0 + b_1$Oil price$_{t-1} + e_t$
- AR(2) model: Oil price$_t = b_0 + b_1$Oil price$_{t-1} + b_2$Oil price$_{t-2} + e_t$

Exhibit 1 presents selected data from all four regressions, and Exhibit 2 presents selected autocorrelation data from the AR(1) models.

Exhibit 1 Crude Oil Price per Barrel, August 2000–August 2015

	Regression Statistics (*t*-statistics for coefficients are reported in parentheses)			
	Linear	**Log-Linear**	**AR(1)**	**AR(2)**
R^2	0.5703	0.6255	0.9583	0.9656
Standard error	18.6327	0.3034	5.7977	5.2799
Observations	181	181	180	179
Durbin–Watson	0.10	0.08	1.16	2.08
RMSE			2.0787	2.0530
Coefficients:				
Intercept	28.3278	3.3929	1.5948	2.0017
	(10.1846)	(74.9091)	(1.4610)	(1.9957)
t (Trend)	0.4086	0.0075		
	(15.4148)	(17.2898)		
Oil Price$_{t-1}$			0.9767	1.3946
			(63.9535)	(20.2999)
Oil Price$_{t-2}$				−0.4249
				(−6.2064)

In Exhibit 1, at the 5% significance level, the lower critical value for the Durbin–Watson test statistic is 1.75 for both the linear and log-linear regressions.

Exhibit 2 Autocorrelations of the Residual from AR(1) Model

Lag	Autocorrelation	*t*-Statistic
1	0.4157	5.5768
2	0.2388	3.2045
3	0.0336	0.4512
4	–0.0426	–0.5712

Note: At the 5% significance level, the critical value for a *t*-statistic is 1.97.

After reviewing the data and regression results, Martinez draws the following conclusions.

Conclusion 1 The time series for WTI oil prices is covariance stationary.

Conclusion 2 Out-of-sample forecasting using the AR(1) model appears to be more accurate than that of the AR(2) model.

20 Based on Exhibit 1, the predicted WTI oil price for October 2015 using the linear trend model is *closest* to:

A $29.15.

B $74.77.

C $103.10.

21 Based on Exhibit 1, the predicted WTI oil price for September 2015 using the log-linear trend model is *closest* to:

A $29.75.

B $29.98.

C $116.50.

22 Based on the regression output in Exhibit 1, there is evidence of positive serial correlation in the errors in:

A the linear trend model but not the log-linear trend model.

B both the linear trend model and the log-linear trend model.

C neither the linear trend model nor the log-linear trend model.

23 Martinez's Conclusion 1 is:

A correct.

B incorrect because the mean and variance of WTI oil prices are not constant over time.

C incorrect because the Durbin–Watson statistic of the AR(2) model is greater than 1.75.

24 Based on Exhibit 1, the forecasted oil price in September 2015 based on the AR(2) model is *closest* to:

A $38.03.

B $40.04.

C $61.77.

25 Based on the data for the AR(1) model in Exhibits 1 and 2, Martinez can conclude that the:

A residuals are not serially correlated.

B autocorrelations do not differ significantly from zero.

C standard error for each of the autocorrelations is 0.0745.

26 Based on the mean-reverting level implied by the AR(1) model regression output in Exhibit 1, the forecasted oil price for September 2015 is *most likely* to be:

A less than $42.86.

B equal to $42.86.

C greater than $42.86.

The following information relates to Question 27–35

Max Busse is an analyst in the research department of a large hedge fund. He was recently asked to develop a model to predict the future exchange rate between two currencies. Busse gathers monthly exchange rate data from the most recent 10-year period and runs a regression based on the following AR(1) model specification:

Regression 1: $x_t = b_0 + b_1 x_{t-1} + \varepsilon_t$, where x_t is the exchange rate at time t.

Based on his analysis of the time series and the regression results, Busse reaches the following conclusions:

Conclusion 1 The variance of x_t increases over time.

Conclusion 2 The mean-reverting level is undefined.

Conclusion 3 b_0 does not appear to be significantly different from 0.

Busse decides to do additional analysis by first-differencing the data and running a new regression.

Regression 2: $y_t = b_0 + b_1 y_{t-1} + \varepsilon_t$, where $y_t = x_t - x_{t-1}$.

Exhibit 1 shows the regression results.

Exhibit 1 First-Differenced Exchange Rate AR(1) Model: Month-End Observations, Last 10 Years

Regression Statistics	
R^2	0.0017
Standard error	7.3336
Observations	118
Durbin–Watson	1.9937

	Coefficient	Standard Error	*t*-Statistic
Intercept	−0.8803	0.6792	−1.2960
$x_{t-1} - x_{t-2}$	0.0412	0.0915	0.4504

(continued)

Exhibit 1 (Continued)

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	*t*-Statistic
1	0.0028	0.0921	0.0300
2	0.0205	0.0921	0.2223
3	0.0707	0.0921	0.7684
4	0.0485	0.0921	0.5271

Note: The critical *t*-statistic at the 5% significance level is 1.98.

Busse decides that he will need to test the data for nonstationarity using a Dickey–Fuller test. To do so, he knows he must model a transformed version of Regression 1.

Busse's next assignment is to develop a model to predict future quarterly sales for PoweredUP, Inc., a major electronics retailer. He begins by running the following regression:

Regression 3: $\ln \text{Sales}_t - \ln \text{Sales}_{t-1} = b_0 + b_1(\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}) + \varepsilon_t$.

Exhibit 2 presents the results of this regression.

Exhibit 2 Log Differenced Sales: AR(1) Model PoweredUP, Inc., Last 10 Years of Quarterly Sales

Regression Statistics

R^2	0.2011
Standard error	0.0651
Observations	38
Durbin–Watson	1.9677

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0408	0.0112	3.6406
$\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}$	–0.4311	0.1432	–3.0099

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	*t*-Statistic
1	0.0146	0.1622	0.0903
2	–0.1317	0.1622	–0.8119
3	–0.1123	0.1622	–0.6922
4	0.6994	0.1622	4.3111

Note: The critical *t*-statistic at the 5% significance level is 2.02.

Because the regression output from Exhibit 2 raises some concerns, Busse runs a different regression. These regression results, along with quarterly sales data for the past five quarters, are presented in Exhibits 3 and 4, respectively.

Exhibit 3 Log Differenced Sales: AR(1) Model with Seasonal Lag PoweredUP, Inc., Last 10 Years of Quarterly Sales

Regression Statistics

R^2	0.6788
Standard error	0.0424
Observations	35
Durbin–Watson	1.8799

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0092	0.0087	1.0582
$\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}$	−0.1279	0.1137	−1.1252
$\ln \text{Sales}_{t-4} - \ln \text{Sales}_{t-5}$	0.7239	0.1093	6.6209

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	*t*-Statistic
1	0.0574	0.1690	0.3396
2	0.0440	0.1690	0.2604
3	0.1923	0.1690	1.1379
4	−0.1054	0.1690	−0.6237

Note: The critical *t*-statistic at the 5% significance level is 2.03.

Exhibit 4 Most Recent Quarterly Sales Data (in billions)

Dec 2015 (Sales_{t-1})	$3.868
Sept 2015 (Sales_{t-2})	$3.780
June 2015 (Sales_{t-3})	$3.692
Mar 2014 (Sales_{t-4})	$3.836
Dec 2014 (Sales_{t-5})	$3.418

After completing his work on PoweredUP, Busse is asked to analyze the relationship of oil prices and the stock prices of three transportation companies. His firm wants to know whether the stock prices can be predicted by the price of oil. Exhibit 5 shows selected information from the results of his analysis.

Exhibit 5 Analysis Summary of Stock Prices for Three Transportation Stocks and the Price of Oil

	Unit Root?	Linear or Exponential Trend?	Serial Correlation of Residuals in Trend Model?	ARCH(1)?	Comments
Company #1	Yes	Exponential	Yes	Yes	Not co-integrated with oil price
Company #2	Yes	Linear	Yes	No	Co-integrated with oil price

(continued)

Exhibit 5 (Continued)

	Unit Root?	Linear or Exponential Trend?	Serial Correlation of Residuals in Trend Model?	ARCH(1)?	Comments
Company #3	No	Exponential	Yes	No	Not co-integrated with oil price
Oil Price	Yes				

To assess the relationship between oil prices and stock prices, Busse runs three regressions using the time series of each company's stock prices as the dependent variable and the time series of oil prices as the independent variable.

27 Which of Busse's conclusions regarding the exchange rate time series is consistent with both the properties of a covariance-stationary time series and the properties of a random walk?

A Conclusion 1

B Conclusion 2

C Conclusion 3

28 Based on the regression output in Exhibit 1, the first-differenced series used to run Regression 2 is consistent with:

A a random walk.

B covariance stationarity.

C a random walk with drift.

29 Based on the regression results in Exhibit 1, the *original* time series of exchange rates:

A has a unit root.

B exhibits stationarity.

C can be modeled using linear regression.

30 In order to perform the nonstationarity test, Busse should transform the Regression 1 equation by:

A adding the second lag to the equation.

B changing the regression's independent variable.

C subtracting the independent variable from both sides of the equation.

31 Based on the regression output in Exhibit 2, what should lead Busse to conclude that the Regression 3 equation is not correctly specified?

A The Durbin–Watson statistic

B The *t*-statistic for the slope coefficient

C The *t*-statistics for the autocorrelations of the residual

32 Based on the regression output in Exhibit 3 and sales data in Exhibit 4, the forecasted value of quarterly sales for March 2016 for PoweredUP is *closest* to:

A $4.193 billion.

B $4.205 billion.

C $4.231 billion.

33 Based on Exhibit 5, Busse should conclude that the variance of the error terms for Company #1:

A is constant.

B can be predicted.

C is homoskedastic.

34 Based on Exhibit 5, for which company would the regression of stock prices on oil prices be expected to yield valid coefficients that could be used to estimate the long-term relationship between stock price and oil price?

A Company #1

B Company #2

C Company #3

35 Based on Exhibit 5, which single time-series model would *most likely* be appropriate for Busse to use in predicting the future stock price of Company #3?

A Log-linear trend model

B First-differenced AR(2) model

C First-differenced log AR(1) model

SOLUTIONS

1 A The estimated forecasting equation is $UER_t = 5.5098 - 0.0294(t)$. The data begin in January 1996, and July 1996 is period 7. Thus the linear trend model predicts the unemployment rate to be $UER_7 = 5.5098 - 0.0294\ (7) = 5.3040$ or approximately 5.3 percent.

B The DW statistic is designed to detect positive serial correlation of the errors of a regression equation. Under the null hypothesis of no positive serial correlation, the DW statistics is 2.0. Positive serial correlation will lead to a DW statistic that is less than 2.0. From the table in Problem 1, we see that the DW statistic is 0.9099. To see whether this result is significantly less than 2.0, refer to the Durbin–Watson table in Appendix E at the end of this volume, in the column marked $k = 1$ (one independent variable) and the row corresponding to 60 observations. We see that $d_l = 1.55$. Because our DW statistic is clearly less than d_l, we reject the null hypothesis of no serial correlation at the 0.05 significance level.

The presence of serial correlation in the error term violates one of the regression assumptions. The standard errors of the estimated coefficients will be biased downward, so we cannot conduct hypothesis testing on the coefficients.

2 The difference between UER and its forecast value, PRED, is the forecast error. In an appropriately specified regression model, the forecast errors are randomly distributed around the regression line and have a constant variance. We can see that the errors from this model specification are persistent. The errors tend first to be above the regression line and then, starting in 1997, they tend to be below the regression line until 2000 when they again are persistently above the regression line. This persistence suggests that the errors are positively serially correlated. Therefore, we conclude that the model is not appropriate for making estimates.

3 A log-linear model captures growth at a constant rate. The log-linear model $\ln(Sales_t) = b_0 + b_1 t + \varepsilon_t$ would be the simplest model consistent with a constant growth rate for monthly sales. Note that we would need to confirm that the regression assumptions are satisfied before accepting the model as valid.

4 A The plot of the series ΔUER_t seems to fluctuate around a constant mean; its volatility appears to be constant throughout the period. Our initial judgment is that the differenced series is covariance stationary.

B The change in the unemployment rate seems covariance stationary, so we should first estimate an AR(1) model and test to see whether the residuals from this model have significant serial correlation. If the residuals do not display significant serial correlation, we should use the AR(1) model. If the residuals do display significant serial correlation, we should try an AR(2) model and test for serial correlation of the residuals of the AR(2) model. We should continue this procedure until the errors from the final AR(p) model are serially uncorrelated.

5 The DW statistic cannot be appropriately used for a regression that has a lagged value of the dependent variable as one of the explanatory variables. To test for serial correlation, we need to examine the autocorrelations.

6 When a covariance-stationary series is at its mean-reverting level, the series will tend not to change until it receives a shock (ε_t). So, if the series ΔUER_t is at the mean-reverting level, $\Delta UER_t = \Delta UER_{t-1}$. This implies that $\Delta UER_t = -0.0405$

$-0.4674\Delta UER_t$, so that $(1 + 0.4674)\ \Delta UER_t = -0.0405$ and $\Delta UER_t = -0.0405/(1 + 0.4674) = -0.0276$. The mean-reverting level is –0.0276. In an AR(1) model, the general expression for the mean-reverting level is $b_0/(1 - b_1)$.

7 A The predicted change in the unemployment rate for next period is –5.45 percent, found by substituting 0.0300 into the forecasting model: $-0.0405 - 0.4674(0.03) = -0.0545$.

B If we substitute our one-period-ahead forecast of –0.0545 into the model (using the chain rule of forecasting), we get a two-period ahead forecast of –0.0150 or –1.5 percent.

C The answer to Part B is quite close to the mean-reverting level of – 0.0276. A stationary time series may need many periods to return to its equilibrium, mean-reverting level.

8 The forecast of sales is $4,391 million for the first quarter of 2002 and $4,738 million for the second quarter of 2002, as the following table shows.

Date	Sales ($ Millions)	Log of Sales	Actual Values of Changes in the Log of Sales Δln ($Sales_t$)	Forecast Values of Changes in the Log of Sales Δln ($Sales_t$)
1Q:2001	6,519	8.7825	0.1308	
2Q:2001	6,748	8.8170	0.0345	
3Q:2001	4,728	8.4613	–0.3557	
4Q:2001	4,298	8.3659	–0.0954	
1Q:2002	4,391	8.3872		0.0213
2Q:2002	4,738	8.4633		0.0761

We find the forecasted change in the log of sales for the first quarter of 2002 by inputting the value for the change in the log of sales from the previous quarter into the equation $\Delta\ln(Sales_t) = 0.0661 + 0.4698\Delta\ln(Sales_{t-1})$. Specifically, $\Delta\ln(Sales_t) = 0.0661 + 0.4698(-0.0954) = 0.0213$, which means that we forecast the log of sales in the first quarter of 2002 to be 8.3659 + 0.0213 = 8.3872.

Next, we forecast the change in the log of sales for the second quarter of 2002 as $\Delta\ln(Sales_t) = 0.0661 + 0.4698(0.0213) = 0.0761$. Note that we have to use our first-quarter 2002 estimated value of the change in the log of sales as our input for $\Delta\ln(Sales_{t-1})$ because we are forecasting past the period for which we have actual data.

With a forecasted change of 0.0761, we forecast the log of sales in the second quarter of 2002 to be 8.3872 + 0.0761 = 8.4633.

We have forecasted the log of sales in the first and second quarters of 2002 to be 8.3872 and 8.4633, respectively. Finally, we take the antilog of our estimates of the log of sales in the first and second quarters of 2002 to get our estimates of the level of sales: $e^{8.3872} = 4{,}391$ and $e^{8.4633} = 4{,}738$, respectively, for sales of $4,391 million and $4,738 million.

9 A The RMSE of the out-of-sample forecast errors is approximately 27 percent. Out-of-sample error refers to the difference between the realized value and the forecasted value of $\Delta\ln(Sales_t)$ for dates beyond the estimation period. In this case, the out-of-sample period is 1Q:2001 to 4Q:2001. These are the four quarters for which we have data that we did not use to obtain the estimated model $\Delta\ln(Sales_t) = 0.0661 + 0.4698\Delta\ln(Sales_{t-1})$.

The steps to calculate RMSE are as follows:

i. Take the difference between the actual and the forecast value. This is the error.

ii. Square the error.

iii. Sum the squared errors.

iv. Divide by the number of forecasts.

v. Take the square root of the average.

We show the calculations for RMSE in the table below.

Actual Values of Changes in the Log of Sales $\Delta\ln(\text{Sales}_t)$	Forecast Values of Changes in the Log of Sales $\Delta\ln(\text{Sales}_t)$	Error (Column 1 – Column 2)	Squared Error (Column 3 Squared)
0.1308	0.1357	–0.0049	0.0000
0.0345	0.1299	–0.0954	0.0091
–0.3557	0.1271	–0.4828	0.2331
–0.0954	0.1259	–0.2213	0.0490
		Sum	0.2912
		Mean	0.0728
		RMSE	0.2698

B The lower the RMSE, the more accurate the forecasts of a model in forecasting. Therefore, the model with the RMSE of 20 percent has greater accuracy in forecasting than the model in Part A, which has an RMSE of 27 percent.

10 A Predictions too far ahead can be nonsensical. For example, the AR(1) model we have been examining, $\Delta\text{UER}_t = -0.0405 - 0.4674\Delta\text{UER}_{t-1}$, taken at face value, predicts declining civilian unemployment into the indefinite future. Because the civilian unemployment rate will probably not go below 3 percent frictional unemployment and cannot go below 0 percent unemployment, this model's long-range forecasts are implausible. The model is designed for short-term forecasting, as are many time-series models.

B Using more years of data for estimation may lead to nonstationarity even in the series of first differences in the civilian unemployment rate. As we go further back in time, we increase the risk that the underlying civilian unemployment rate series has more than one regime (or true model). If the series has more than one regime, fitting one model to the entire period would not be correct. Note that when we have good reason to believe that a time series is stationary, a longer series of data is generally desirable.

11 A The graph of $\ln(\text{Sales}_t)$ appears to trend upward over time. A series that trends upward or downward over time often has a unit root and is thus not covariance stationary. Therefore, using an AR(1) regression on the undifferenced series is probably not correct. In practice, we need to examine regression statistics to confirm visual impressions such as this.

B The most common way to transform a time series with a unit root into a covariance-stationary time series is to difference the data—that is, to create a new series $\Delta\ln(\text{Sales}_t) = \ln(\text{Sales}_t) - \ln(\text{Sales}_{t-1})$.

12 The plot of the series $\Delta\ln(\text{Sales}_t)$ appears to fluctuate around a constant mean; its volatility seems constant throughout the period. Differencing the data appears to have made the time series covariance stationary.

13 A In a correctly specified regression, the residuals must be serially uncorrelated. We have 108 observations, so the standard error of the autocorrelation is $1/\sqrt{T}$, or in this case $1/\sqrt{108} = 0.0962$. The *t*-statistic for each lag is significant at the 0.01 level. We would have to modify the model specification before continuing with the analysis.

B Because the residuals from the AR(1) specification display significant serial correlation, we should estimate an AR(2) model and test for serial correlation of the residuals of the AR(2) model. If the residuals from the AR(2) model are serially uncorrelated, we should then test for seasonality and ARCH behavior. If any serial correlation remains in the residuals, we should estimate an AR(3) process and test the residuals from that specification for serial correlation. We should continue this procedure until the errors from the final AR(*p*) model are serially uncorrelated. When serial correlation is eliminated, we should test for seasonality and ARCH behavior.

14 A The series has a steady upward trend of growth, suggesting an exponential growth rate. This finding suggests transforming the series by taking the natural log and differencing the data.

B First, we should determine whether the residuals from the AR(1) specification are serially uncorrelated. If the residuals are serially correlated, then we should try an AR(2) specification and then test the residuals from the AR(2) model for serial correlation. We should continue in this fashion until the residuals are serially uncorrelated, then look for seasonality in the residuals. If seasonality is present, we should add a seasonal lag. If no seasonality is present, we should test for ARCH. If ARCH is not present, we can conclude that the model is correctly specified.

C If the model $\Delta\ln(\text{Sales}_t) = b_0 + b_1[\Delta\ln(\text{Sales}_{t-1})] + \varepsilon_t$ is correctly specified, then the series $\Delta\ln(\text{Sales}_t)$ is covariance stationary. So, this series tends to its mean-reverting level, which is $b_0/(1 - b_1)$ or $0.0661/(1 - 0.4698) = 0.1247$.

15 The quarterly sales of Avon show an upward trend and a clear seasonal pattern, as indicated by the repeated regular cycle.

16 A A second explanatory variable, the change in the gross profit margin lagged four quarters, ΔGPM_{t-4}, was added.

B The model was augmented to account for seasonality in the time series (with quarterly data, significant autocorrelation at the fourth lag indicates seasonality). The standard error of the autocorrelation coefficient equals 1 divided by the square root of the number of observations: $1/\sqrt{40}$ or 0.1581. The autocorrelation at the fourth lag (0.8496) is significant: $t = 0.8496/0.1581 = 5.37$. This indicates seasonality, and accordingly we added ΔGPM_{t-4}. Note that in the augmented regression, the coefficient on ΔGPM_{t-4} is highly significant. (Although the autocorrelation at second lag is also significant, the fourth lag is more important because of the rationale of seasonality. Once the fourth lag is introduced as an independent variable, we might expect that the second lag in the residuals would not be significant.)

17 A In order to determine whether this model is correctly specified, we need to test for serial correlation among the residuals. We want to test whether we can reject the null hypothesis that the value of each autocorrelation is 0 against the alternative hypothesis that each is not equal to 0. At the 0.05 significance level, with 68 observations and three parameters, this model has 65 degrees of freedom. The critical value of the *t*-statistic needed to reject the null hypothesis is thus about 2.0. The absolute value of the *t*-statistic for

each autocorrelation is below 0.60 (less than 2.0), so we cannot reject the null hypothesis that each autocorrelation is not significantly different from 0. We have determined that the model is correctly specified.

B If sales grew by 1 percent last quarter and by 2 percent four quarters ago, then the model predicts that sales growth this quarter will be 0.0121 − 0.0839 ln(1.01) + 0.6292 ln(1.02) = $e^{0.02372} - 1 = 2.40\%$.

18 We should estimate the regression $\Delta UER_t = b_0 + b_1 \Delta UER_{t-1} + \varepsilon_t$ and save the residuals from the regression. Then we should create a new variable, $\hat{\varepsilon}_t^2$, by squaring the residuals. Finally, we should estimate $\hat{\varepsilon}_t^2 = a_0 + a_1 \hat{\varepsilon}_{t-1}^2 + u_t$ and test to see whether a_1 is statistically different from 0.

19 To determine whether we can use linear regression to model more than one time series, we should first determine whether any of the time series has a unit root. If none of the time series has a unit root, then we can safely use linear regression to test the relations between the two time series. Note that if one of the two variables has a unit root, then our analysis would not provide valid results; if both of the variables have unit roots, then we would need to evaluate whether the variables are cointegrated.

20 C is correct. The predicted value for period t from a linear trend is calculated as $\hat{y}_t = \hat{b}_0 + \hat{b}_1(t)$.

October 2015 is the second month out of sample, or t = 183. So, the predicted value for October 2015 is calculated as

$$\hat{y}_t = 28.3278 + 0.4086(183) = \$103.10.$$

Therefore, the predicted WTI oil price for October 2015 based on the linear trend model is $103.10.

21 C is correct. The predicted value for period t from a log-linear trend is calculated as $\ln \hat{y}_t = \hat{b}_0 + \hat{b}_1(t)$.

September 2015 is the first month out of sample, or t = 182. So, the predicted value for September 2015 is calculated as follows:

$$\ln \hat{y}_t = 3.3929 + 0.0075(182)$$

$$\ln \hat{y}_t = 4.7579$$

$$\hat{y}_t = e^{4.7579} = \$116.50$$

Therefore, the predicted WTI oil price for September 2015, based on the log-linear trend model, is $116.50.

22 B is correct. The Durbin–Watson statistic for the linear trend model is 0.10 and, for the log-linear trend model, 0.08. Both of these values are below the critical value of 1.75. Therefore, we can reject the hypothesis of no positive serial correlation in the regression errors in both the linear trend model and the log-linear trend model.

23 B is correct. There are three requirements for a time series to be covariance stationary. First, the expected value of the time series must be constant and finite in all periods. Second, the variance of the time series must be constant and finite in all periods. Third, the covariance of the time series with itself for a fixed number of periods in the past or future must be constant and finite in

all periods. Martinez concludes that the mean and variance of the time series of WTI oil prices are not constant over time. Therefore, the time series is not covariance stationary.

24 B is correct. The last two observations in the WTI time series are July and August 2015, when the WTI oil price was $51.16 and $42.86, respectively. Therefore, September 2015 represents a one-period-ahead forecast. The one-period-ahead forecast from an AR(2) model is calculated as

$$\hat{x}_{t+1} = \hat{b}_0 + \hat{b}_1 x_t + \hat{b}_2 x_{t+1}$$

So, the one-period-ahead (September 2015) forecast is calculated as

$$\hat{x}_{t+1} = 2.0017 + 1.3946(\$42.86) - 0.4249(\$51.16) = \$40.04.$$

Therefore, the September 2015 forecast based on the AR(2) model is $40.04.

25 C is correct. The standard error of the autocorrelations is calculated as $\frac{1}{\sqrt{T}}$, where T represents the number of observations used in the regression. Therefore, the standard error for each of the autocorrelations is $\frac{1}{\sqrt{180}} = 0.0745$. Martinez can conclude that the residuals are serially correlated and are significantly different from zero because two of the four autocorrelations in Exhibit 2 have a t-statistic in absolute value that is greater than the critical value of 1.97.

Choices A and B are incorrect because two of the four autocorrelations have a t-statistic in absolute value that is greater than the critical value of the t-statistic of 1.97.

26 C is correct. The mean-reverting level from the AR(1) model is calculated as

$$\hat{x}_t = \frac{b_0}{1 - b_1} = \frac{1.5948}{1 - 0.9767} = \$68.45$$

Therefore, the mean-reverting WTI oil price from the AR(1) model is $68.45. The forecasted oil price in September 2015 will likely be greater than $42.86 because the model predicts that the price will rise in the next period from the August 2015 price of $42.86.

27 C is correct. A random walk can be described by the equation $x_t = b_0 + b_1 x_{t-1} + \varepsilon_t$, where $b_0 = 0$ and $b_1 = 1$. So $b_0 = 0$ is a characteristic of a random walk time series. A covariance-stationary series must satisfy the following three requirements:

1. The expected value of the time series must be constant and finite in all periods.
2. The variance of the time series must be constant and finite in all periods.
3. The covariance of the time series with itself for a fixed number of periods in the past or future must be constant and finite in all periods.

$b_0 = 0$ does not violate any of these three requirements and is thus consistent with the properties of a covariance-stationary time series.

28 B is correct. The critical t-statistic at a 5% confidence level is 1.98. As a result, neither the intercept nor the coefficient on the first lag of the first-differenced exchange rate in Regression 2 differs significantly from zero. Also, the residual autocorrelations do not differ significantly from zero. As a result, Regression 2 can be reduced to $y_t = \varepsilon_t$ with a mean-reverting level of $b_0/(1 - b_1) = 0/1 = 0$. Therefore, the variance of y_t in each period is $\text{Var}(\varepsilon_t) = \sigma^2$. The fact that the residuals are not autocorrelated is consistent with the covariance of the times

series, with itself being constant and finite at different lags. Because the variance and the mean of y_t are constant and finite in each period, we can also conclude that y_t is covariance stationary.

29 A is correct. If the exchange rate series is a random walk, then the first-differenced series will yield $b_0 = 0$ and $b_1 = 0$, and the error terms will not be serially correlated. The data in Exhibit 1 show that this is the case: Neither the intercept nor the coefficient on the first lag of the first-differenced exchange rate in Regression 2 differs significantly from zero because the *t*-statistics of both coefficients are less than the critical *t*-statistic of 1.98. Also, the residual autocorrelations do not differ significantly from zero because the *t*-statistics of all autocorrelations are less than the critical *t*-statistic of 1.98. Therefore, because all random walks have unit roots, the exchange rate time series used to run Regression 1 has a unit root.

30 C is correct. To conduct the Dickey–Fuller test, one must subtract the independent variable, x_{t-1}, from both sides of the original AR(1) model. This results in a change of the dependent variable (from x_t to $x_t - x_{t-1}$) and a change in the regression's slope coefficient (from b_1 to $b_1 - 1$) but not a change in the independent variable.

31 C is correct. The regression output in Exhibit 2 suggests there is serial correlation in the residual errors. The fourth autocorrelation of the residual has a value of 0.6994 and a *t*-statistic of 4.3111, which is greater than the *t*-statistic critical value of 2.02. Therefore, the null hypothesis that the fourth autocorrelation is equal to zero can be rejected. This indicates strong and significant seasonal autocorrelation, which means the Regression 3 equation is misspecified.

32 C is correct. The quarterly sales for March 2016 is calculated as follows:

$$\begin{aligned}
\ln \text{Sales}_t - \ln \text{Sales}_{t-1} &= b_0 + b_1(\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}) + b_2(\ln \text{Sales}_{t-4} \\
&\quad - \ln \text{Sales}_{t-5}).\\
\ln \text{Sales}_t - \ln 3.868 &= 0.0092 - 0.1279(\ln 3.868 - \ln 3.780) + 0.7239(\ln \\
&\quad 3.836 - \ln 3.418).\\
\ln \text{Sales}_t - 1.35274 &= 0.0092 - 0.1279(1.35274 - 1.32972) + \\
&\quad 0.7239(1.34443 - 1.22906).\\
\ln \text{Sales}_t &= 1.35274 + 0.0092 - 0.1279(0.02301) + \\
&\quad 0.7239(0.11538).\\
\ln \text{Sales}_t &= 1.44251.\\
\text{Sales}_t &= e^{1.44251} = 4.231.
\end{aligned}$$

33 B is correct. Exhibit 5 shows that the time series of the stock prices of Company #1 exhibits heteroskedasticity, as evidenced by the fact that the time series is ARCH(1). If a time series is ARCH(1), then the variance of the error in one period depends on the variance of the error in previous periods. Therefore, the variance of the errors in period $t + 1$ can be predicted in period t using the formula

$$\hat{\sigma}^2_{t+1} = \hat{a}_0 + \hat{a}_1 \hat{\varepsilon}^2_t$$

34 B is correct. When two time series have a unit root but are co-integrated, the error term in the linear regression of one time series on the other will be covariance stationary. Exhibit 5 shows that the series of stock prices of Company #2 and the oil prices both contain a unit root, and the two time series are co-integrated. As a result, the regression coefficients and standard errors are

consistent and can be used for hypothesis tests. Although the co-integrated regression estimates the long-term relation between the two series, it may not be the best model of the short-term relationship.

35 C is correct. As a result of the exponential trend in the time series of stock prices for Company #3, Busse would want to take the natural log of the series and then first-difference it. Because the time series also has serial correlation in the residuals from the trend model, Busse should use a more complex model, such as an autoregressive (AR) model.

© CFA Institute. For candidate use only. Not for distribution.

PRACTICE PROBLEMS

The following information relates to Questions 1–10

Alef Associates manages a long-only fund specializing in global smallcap equities. Since its founding a decade ago, Alef maintains a portfolio of 100 stocks (out of an eligible universe of about 10,000 stocks). Some of these holdings are the result of screening the universe for attractive stocks based on several ratios that use readily available market and accounting data; others are the result of investment ideas generated by Alef's professional staff of five securities analysts and two portfolio managers.

Although Alef's investment performance has been good, its Chief Investment Officer, Paul Moresanu, is contemplating a change in the investment process aimed at achieving even better returns. After attending multiple workshops and being approached by data vendors, Moresanu feels that data science should play a role in the way Alef selects its investments. He has also noticed that much of Alef's past outperformance is due to stocks that became takeover targets. After some research and reflection, Moresanu writes the following email to the Alef's CEO.

Subject: Investment Process Reorganization

I have been thinking about modernizing the way we select stock investments. Given that our past success has put Alef Associates in an excellent financial position, now seems to be a good time to invest in our future. What I propose is that we continue managing a portfolio of 100 global small-cap stocks but restructure our process to benefit from machine learning (ML). Importantly, the new process will still allow a role for human insight, for example, in providing domain knowledge. In addition, I think we should make a special effort to identify companies that are likely to be acquired. Specifically, I suggest following the four steps which would be repeated every quarter.

Step 1 We apply ML techniques to a model including fundamental and technical variables (features) to predict next quarter's return for each of the 100 stocks currently in our portfolio. Then, the 20 stocks with the lowest estimated return are identified for replacement.

Step 2 We utilize ML techniques to divide our investable universe of about 10,000 stocks into 20 different groups, based on a wide variety of the most relevant financial and non-financial characteristics. The idea is to prevent unintended portfolio concentration by selecting stocks from each of these distinct groups.

Step 3 For each of the 20 different groups, we use labeled data to train a model that will predict the five stocks (in any given group) that are most likely to become acquisition targets in the next one year.

(Continued)

Step 4 Our five experienced securities analysts are each assigned four of the groups, and then each analyst selects their one best stock pick from each of their assigned groups. These 20 "high-conviction" stocks will be added to our portfolio (in replacement of the 20 relatively underperforming stocks to be sold in Step 1).

A couple of additional comments related to the above:

Comment 1 The ML algorithms will require large amounts of data. We would first need to explore using free or inexpensive historical datasets and then evaluate their usefulness for the ML-based stock selection processes before deciding on using data that requires subscription.

Comment 2 As time passes, we expect to find additional ways to apply ML techniques to refine Alef's investment processes.

What do you think?
Paul Moresanu

---

1 The machine learning techniques appropriate for executing Step 1 are *most* likely to be based on:

A regression

B classification

C clustering

2 Assuming regularization is utilized in the machine learning technique used for executing Step 1, which of the following ML models would be *least* appropriate:

A Regression tree with pruning.

B LASSO with lambda (λ) equal to 0.

C LASSO with lambda (λ) between 0.5 and 1.

3 Which of the following machine learning techniques is *most* appropriate for executing Step 2:

A K-Means Clustering

B Principal Components Analysis (PCA)

C Classification and Regression Trees (CART)

4 The hyperparameter in the ML model to be used for accomplishing Step 2 is?

A 100, the number of small-cap stocks in Alef's portfolio.

B 10,000, the eligible universe of small-cap stocks in which Alef can potentially invest.

C 20, the number of different groups (i.e. clusters) into which the eligible universe of small-cap stocks will be divided.

5 The target variable for the labelled training data to be used in Step 3 is *most* likely which one of the following?

A A continuous target variable.

B A categorical target variable.

C An ordinal target variable.

6 Comparing two ML models that could be used to accomplish Step 3, which statement(s) *best* describe(s) the advantages of using Classification and Regression Trees (CART) instead of K-Nearest Neighbor (KNN)?

Statement I	For CART there is no requirement to specify an initial hyperparameter (like K).
Statement II	For CART there is no requirement to specify a similarity (or distance) measure.
Statement III	For CART the output provides a visual explanation for the prediction.

A Statement I only.

B Statement III only.

C Statements I, II and III.

7 Assuming a Classification and Regression Tree (CART) model is used to accomplish Step 3, which of the following is *most* likely to result in model overfitting?

A Using the k-fold cross validation method

B Including an overfitting penalty (i.e., regularization term).

C Using a fitting curve to select a model with low bias error and high variance error.

8 Assuming a Classification and Regression Tree (CART) model is initially used to accomplish Step 3, as a further step which of the following techniques is most likely to result in more accurate predictions?

A Discarding CART and using the predictions of a Support Vector Machine (SVM) model instead.

B Discarding CART and using the predictions of a K-Nearest Neighbor (KNN) model instead.

C Combining the predictions of the CART model with the predictions of other models – such as logistic regression, SVM, and KNN – via ensemble learning.

9 Regarding Comment #2, Moresanu has been thinking about the applications of neural networks (NNs) and deep learning (DL) to investment management. Which statement(s) *best* describe(s) the tasks for which NNs and DL are well-suited?

Statement I	NNs and DL are well-suited for image and speech recognition, and natural language processing.
Statement II	NNs and DL are well-suited for developing single variable ordinary least squares regression models.
Statement III	NNs and DL are well-suited for modelling non-linearities and complex interactions among many features.

A Statement II only.

B Statements I and III.

C Statements I, II and III.

10 Regarding neural networks (NNs) that Alef might potentially implement, which of the following statements is *least* accurate?

A NNs must have at least 10 hidden layers to be considered deep learning nets.

B The activation function in a node operates like a light dimmer switch since it decreases or increases the strength of the total net input.

C The summation operator receives input values, multiplies each by a weight, sums up the weighted values into the total net input, and passes it to the activation function.

SOLUTIONS

1 A is correct. The target variable (quarterly return) is continuous, hence this calls for a supervised machine learning based regression model.

B is incorrect, since classification uses categorical or ordinal target variables, while in Step 1 the target variable (quarterly return) is continuous.

C is incorrect, since clustering involves unsupervised machine learning so does not have a target variable.

2 B is correct. It is least appropriate because with LASSO, when $\lambda = 0$ the penalty (i.e., regularization) term reduces to zero, so there is no regularization and the regression is equivalent to an ordinary least squares (OLS) regression.

A is incorrect. With Classification and Regression Trees (CART), one way that regularization can be implemented is via pruning which will reduce the size of the regression tree—sections that provide little explanatory power are pruned (i.e., removed).

C is incorrect. With LASSO, when λ is between 0.5 and 1 the relatively large penalty (i.e., regularization) term requires that a feature makes a sufficient contribution to model fit to offset the penalty from including it in the model.

3 A is correct. K-Means clustering is an unsupervised machine learning algorithm which repeatedly partitions observations into a fixed number, *k*, of non-overlapping clusters (i.e., groups).

B is incorrect. Principal Components Analysis is a long-established statistical method for dimension reduction, not clustering. PCA aims to summarize or reduce highly correlated features of data into a few main, uncorrelated composite variables.

C is incorrect. CART is a supervised machine learning technique that is most commonly applied to binary classification or regression.

4 C is correct. Here, 20 is a hyperparameter (in the K-Means algorithm), which is a parameter whose value must be set by the researcher before learning begins.

A is incorrect, because it is not a hyperparameter. It is just the size (number of stocks) of Alef's portfolio.

B is incorrect, because it is not a hyperparameter. It is just the size (number of stocks) of Alef's eligible universe.

5 B is correct. To predict which stocks are likely to become acquisition targets, the ML model would need to be trained on categorical labelled data having the following two categories: "0" for "not acquisition target", and "1" for "acquisition target".

A is incorrect, because the target variable is categorical, not continuous.

C is incorrect, because the target variable is categorical, not ordinal (i.e., 1st, 2nd, 3rd, etc.).

6 C is correct. The advantages of using CART over KNN to classify companies into two categories ("not acquisition target" and "acquisition target"), include all of the following: For CART there are no requirements to specify an initial hyperparameter (like K) or a similarity (or distance) measure as with KNN, and CART provides a visual explanation for the prediction (i.e., the feature variables and their cut-off values at each node).

A is incorrect, because CART provides all of the advantages indicated in Statements I, II and III.

B is incorrect, because CART provides all of the advantages indicated in Statements I, II and III.

7 C is correct. A fitting curve shows the trade-off between bias error and variance error for various potential models. A model with low bias error and high variance error is, by definition, overfitted.

A is incorrect, because there are two common methods to reduce overfitting, one of which is proper data sampling and cross-validation. K-fold cross validation is such a method for estimating out-of-sample error directly by determining the error in validation samples.

B is incorrect, because there are two common methods to reduce overfitting, one of which is preventing the algorithm from getting too complex during selection and training, which requires estimating an overfitting penalty.

8 C is correct. Ensemble learning is the technique of combining the predictions from a collection of models, and it typically produces more accurate and more stable predictions than the best single model.

A is incorrect, because a single model will have a certain error rate and will make noisy predictions. By taking the average result of many predictions from many models (i.e., ensemble learning) one can expect to achieve a reduction in noise as the average result converges towards a more accurate prediction.

B is incorrect, because a single model will have a certain error rate and will make noisy predictions. By taking the average result of many predictions from many models (i.e., ensemble learning) one can expect to achieve a reduction in noise as the average result converges towards a more accurate prediction.

9 B is correct. NNs and DL are well-suited for addressing highly complex machine learning tasks, such as image classification, face recognition, speech recognition and natural language processing. These complicated tasks are characterized by non-linearities and complex interactions between large numbers of feature inputs.

A is incorrect, because NNs and DL are well-suited for addressing highly complex machine learning tasks, not simple single variable OLS regression models.

C is incorrect, because NNs and DL are well-suited for addressing highly complex machine learning tasks, not simple single variable OLS regression models.

10 A is correct. It is the least accurate answer because neural networks with many hidden layers—at least 3, but often more than 20 hidden layers—are known as deep learning nets.

B is incorrect, because the node's activation function operates like a light dimmer switch which decreases or increases the strength of the (total net) input.

C is incorrect, because the node's summation operator multiplies each (input) value by a weight and sums up the weighted values to form the total net input. The total net input is then passed to the activation function.

© CFA Institute. For candidate use only. Not for distribution.

PRACTICE PROBLEMS

The following information relates to Questions 1–15

Aaliyah Schultz is a fixed-income portfolio manager at Aries Investments. Schultz supervises Ameris Steele, a junior analyst.

A few years ago, Schultz developed a proprietary machine learning (ML) model that aims to predict downgrades of publicly-traded firms by bond rating agencies. The model currently relies only on structured financial data collected from different sources. Schultz thinks the model's predictive power may be improved by incorporating sentiment data derived from textual analysis of news articles and Twitter content relating to the subject companies.

Schultz and Steele meet to discuss plans for incorporating the sentiment data into the model. They discuss the differences in the steps between building ML models that use traditional structured data and building ML models that use textual big data. Steele tells Schultz:

Statement 1 The second step in building text-based ML models is text preparation and wrangling, whereas the second step in building ML models using structured data is data collection.

Statement 2 The fourth step in building both types of models encompasses data/text exploration.

Steele expresses concern about using Twitter content in the model, noting that research suggests that as much as 10%–15% of social media content is from fake accounts. Schultz tells Steele that she understands her concern but thinks the potential for model improvement outweighs the concern.

Steele begins building a model that combines the structured financial data and the sentiment data. She starts with cleansing and wrangling the raw structured financial data. Exhibit 1 presents a small sample of the raw dataset before cleansing: Each row represents data for a particular firm.

Exhibit 1 Sample of Raw Structured Data Before Cleansing

ID	Ticker	IPO Date	Industry (NAICS)	EBIT	Interest Expense	Total Debt
1	ABC	4/6/17	44	9.4	0.6	10.1
2	BCD	November 15, 2004	52	5.5	0.4	6.2
3	HIJ	26-Jun-74	54	8.9	1.2	15.8
4	KLM	14-Mar-15	72	5.7	1.5	0.0

After cleansing the data, Steele then preprocesses the dataset. She creates two new variables: an "Age" variable based on the firm's IPO date and an "Interest Coverage Ratio" variable equal to EBIT divided by interest expense. She also deletes the "IPO Date" variable from the dataset. After applying these transformations, Steele scales

the financial data using normalization. She notes that over the full sample dataset, the "Interest Expense" variable ranges from a minimum of 0.2 and a maximum of 12.2, with a mean of 1.1 and a standard deviation of 0.4.

Steele and Schultz then discuss how to preprocess the raw text data. Steele tells Schultz that the process can be completed in the following three steps:

Step 1 Cleanse the raw text data.

Step 2 Split the cleansed data into a collection of words for them to be normalized.

Step 3 Normalize the collection of words from Step 2 and create a distinct set of tokens from the normalized words.

With respect to Step 1, Steele tells Schultz:

> "I believe I should remove all html tags, punctuations, numbers, and extra white spaces from the data before normalizing them."

After properly cleansing the raw text data, Steele completes Steps 2 and 3. She then performs exploratory data analysis. To assist in feature selection, she wants to create a visualization that shows the most informative words in the dataset based on their term frequency (TF) values. After creating and analyzing the visualization, Steele is concerned that some tokens are likely to be noise features for ML model training; therefore, she wants to remove them.

Steele and Schultz discuss the importance of feature selection and feature engineering in ML model training. Steele tells Schultz:

> "Appropriate feature selection is a key factor in minimizing model overfitting, whereas feature engineering tends to prevent model underfitting."

Once satisfied with the final set of features, Steele selects and runs a model on the training set that classifies the text as having positive sentiment (Class "1" or negative sentiment (Class "0"). She then evaluates its performance using error analysis. The resulting confusion matrix is presented in Exhibit 2.

Exhibit 2 Confusion Matrix

		Actual Training Results	
		Class "1"	Class "0"
Predicted Results	Class "1"	TP = 182	FP = 52
	Class "0"	FN = 31	TN = 96

1 Which of Steele's statements relating to the steps in building structured data-based and text-based ML models is correct?

A Only Statement 1 is correct.

B Only Statement 2 is correct.

C Statement 1 and Statement 2 are correct.

2 Steele's concern about using Twitter data in the model *best* relates to:

A volume.

B velocity.

C veracity.

3 What type of error appears to be present in the IPO Date column of Exhibit 1?

A invalidity error.

B inconsistency error.

C non-uniformity error.

4 What type of error is most likely present in the last row of data (ID #4) in Exhibit 1?

A Inconsistency error

B Incompleteness error

C Non-uniformity error

5 During the preprocessing of the data in Exhibit 1, what type of data transformation did Steele perform during the data preprocessing step?

A Extraction

B Conversion

C Aggregation

6 Based on Exhibit 1, for the firm with ID #3, Steele should compute the scaled value for the "Interest Expense" variable as:

A 0.008.

B 0.083.

C 0.250.

7 Is Steele's statement regarding Step 1 of the preprocessing of raw text data correct?

A Yes.

B No, because her suggested treatment of punctuation is incorrect.

C No, because her suggested treatment of extra white spaces is incorrect.

8 Steele's Step 2 can be *best* described as:

A tokenization.

B lemmatization.

C standardization.

9 The output created in Steele's Step 3 can be *best* described as a:

A bag-of-words.

B set of n-grams.

C document term matrix.

10 Given her objective, the visualization that Steele should create in the exploratory data analysis step is a:

A scatter plot.

B word cloud.

C document term matrix.

11 To address her concern in her exploratory data analysis, Steele should focus on those tokens that have:

A low chi-square statistics.

B low mutual information (ML) values.

C very low and very high term frequency (TF) values.

12 Is Steele's statement regarding the relationship between feature selection/feature engineering and model fit correct?

A Yes.

B No, because she is incorrect with respect to feature selection.

C No, because she is incorrect with respect to feature engineering.

13 Based on Exhibit 2, the model's precision metric is *closest* to:

A 78%.

B 81%.

C 85%.

14 Based on Exhibit 2, the model's F1 score is *closest* to:

A 77%.

B 81%.

C 85%.

15 Based on Exhibit 2, the model's accuracy metric is *closest* to:

A 77%.

B 81%.

C 85%.

SOLUTIONS

1 B is correct. The five steps in building structured data-based ML models are: 1) conceptualization of the modeling task, 2) data collection, 3) data preparation and wrangling, 4) data exploration, and 5) model training. The five steps in building text-based ML models are: 1) text problem formulation, 2) data (text) curation, 3) text preparation and wrangling, 4) text exploration, and 5) model training. Statement 1 is incorrect: Text preparation and wrangling is the third step in building text ML models and occurs after the second data (text) curation step. Statement 2 is correct: The fourth step in building both types of models encompasses data/text exploration.

2 C is correct. Veracity relates to the credibility and reliability of different data sources. Steele is concerned about the credibility and reliability of Twitter content, noting that research suggests that as much as 10%–15% of social media content is from fake accounts.

3 C is correct. A non-uniformity error occurs when the data are not presented in an identical format. The data in the "IPO Date" column represent the IPO date of each firm. While all rows are populated with valid dates in the IPO Date column, the dates are presented in different formats (e.g., mm/dd/yyyy, dd/mm/yyyy).

4 A is correct. There appears to be an inconsistency error in the last row (ID #4). An inconsistency error occurs when a data point conflicts with corresponding data points or reality. In the last row, the interest expense data item has a value of 1.5, and the total debt item has a value of 0.0. This appears to be an error: Firms that have interest expense are likely to have debt in their capital structure, so either the interest expense is incorrect or the total debt value is incorrect. Steele should investigate this issue by using alternative data sources to confirm the correct values for these variables.

5 A is correct. During the data preprocessing step, Steele created a new "Age" variable based on the firm's IPO date and then deleted the "IPO Date" variable from the dataset. She also created a new "Interest Coverage Ratio" variable equal to EBIT divided by interest expense. Extraction refers to a data transformation where a new variable is extracted from a current variable for ease of analyzing and using for training an ML model, such as creating an age variable from a date variable or a ratio variable. Steele also performed a selection transformation by deleting the IPO Date variable, which refers to deleting the data columns that are not needed for the project.

6 B is correct. Steele uses normalization to scale the financial data. Normalization is the process of rescaling numeric variables in the range of [0, 1]. To normalize variable X, the minimum value (X_{min}) is subtracted from each observation (X_i), and then this value is divided by the difference between the maximum and minimum values of X ($X_{\text{max}} - X_{\text{min}}$):

$$X_{i\ (\text{normalized})} = \frac{X_i - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

The firm with ID #3 has an interest expense of 1.2. So, its normalized value is calculated as:

$$X_{i\ (\text{normalized})} = \frac{1.2 - 0.2}{12.2 - 0.2} = 0.083$$

7 B is correct. Although most punctuations are not necessary for text analysis and should be removed, some punctuations (e.g., percentage signs, currency symbols, and question marks) may be useful for ML model training. Such punctuations should be substituted with annotations (e.g., /percentSign/, /dollarSign/, and /questionMark/) to preserve their grammatical meaning in the text. Such annotations preserve the semantic meaning of important characters in the text for further text processing and analysis stages.

8 A is correct. Tokenization is the process of splitting a given text into separate tokens. This step takes place after cleansing the raw text data (removing html tags, numbers, extra white spaces, etc.). The tokens are then normalized to create the bag-of-words (BOW).

9 A is correct. After the cleansed text is normalized, a bag-of-words is created. A bag-of-words (BOW) is a collection of a distinct set of tokens from all the texts in a sample dataset.

10 B is correct. Steele wants to create a visualization for Schultz that shows the most informative words in the dataset based on their term frequency (TF, the ratio of the number of times a given token occurs in the dataset to the total number of tokens in the dataset) values. A word cloud is a common visualization when working with text data as it can be made to visualize the most informative words and their TF values. The most commonly occurring words in the dataset can be shown by varying font size, and color is used to add more dimensions, such as frequency and length of words.

11 C is correct. Frequency measures can be used for vocabulary pruning to remove noise features by filtering the tokens with very high and low TF values across all the texts. Noise features are both the most frequent and most sparse (or rare) tokens in the dataset. On one end, noise features can be stop words that are typically present frequently in all the texts across the dataset. On the other end, noise features can be sparse terms that are present in only a few text files. Text classification involves dividing text documents into assigned classes. The frequent tokens strain the ML model to choose a decision boundary among the texts as the terms are present across all the texts (an example of underfitting). The rare tokens mislead the ML model into classifying texts containing the rare terms into a specific class (an example of overfitting). Thus, identifying and removing noise features are critical steps for text classification applications.

12 A is correct. A dataset with a small number of features may not carry all the characteristics that explain relationships between the target variable and the features. Conversely, a large number of features can complicate the model and potentially distort patterns in the data due to low degrees of freedom, causing overfitting. Therefore, appropriate feature selection is a key factor in minimizing such model overfitting. Feature engineering tends to prevent underfitting in the training of the model. New features, when engineered properly, can elevate the underlying data points that better explain the interactions of features. Thus, feature engineering can be critical to overcome underfitting.

13 A is correct. Precision, the ratio of correctly predicted positive classes (true positives) to all predicted positive classes, is calculated as:

$$\text{Precision (P)} = \text{TP}/(\text{TP} + \text{FP}) = 182/(182 + 52) = 0.7778\ (78\%).$$

14 B is correct. The model's F1 score, which is the harmonic mean of precision and recall, is calculated as:

$$\text{F1 score} = (2 \times \text{P} \times \text{R})/(\text{P} + \text{R}).$$

$$\text{F1 score} = (2 \times 0.7778 \times 0.8545)/(0.7778 + 0.8545) = 0.8143\ (81\%).$$

15 A is correct. The model's accuracy, which is the percentage of correctly predicted classes out of total predictions, is calculated as:

Accuracy = (TP + TN)/(TP + FP + TN + FN).

Accuracy = (182 + 96)/(182 + 52 + 96 + 31) = 0.7701 (77%).

PRACTICE PROBLEMS

The following information relates to Questions 1–7

Alicia Maxwell, an analyst for a REIT, is evaluating the potential purchase of a hotel property. She plans to use simulation analysis to estimate the distribution of the property's annual operating cash flow for the next five years.

Revenue Construction

Maxwell recognizes that annual gross revenue for the property depends on the nightly room rate and the occupancy rate. She believes that the primary driver for the nightly room rate is the Employment Cost Index (ECI) and that the primary driver for the occupancy rate is the Consumer Sentiment Index (CSI). In the process of simulating revenues, she examines the ECI and the CSI quarterly over the past 20 years and their relation to the nightly room rate and occupancy rates of the REIT's existing properties, respectively. She estimates the following:

$$\text{Nightly room rate} = \$23 + 0.9(\text{ECI}_{t-1})$$

$$\text{Occupancy rate} = 0.25 + 0.7(\text{CSI}_{t-1})$$

Occupancy rates are assumed to be non-negative and cannot exceed 100%.

Maxwell generates 10,000 trials of the ECI and CSI based on the historical mean level of the indexes and their monthly standard deviations. Although the distribution of historical CSI is not symmetric, she assumes that both ECI and CSI are normally distributed. Maxwell is aware that if the inputs are correlated, this may present a problem. She also observes that the CSI and the ECI are correlated with one another and that the relation between the CSI and the occupancy rate is stronger than that between the ECI and nightly room rates. Maxwell estimates the corresponding nightly room rate and the occupancy rate based on these historical relations, multiplies these by the number of hotel nights in a year, and generates 10,000 estimates of annual gross revenue.

Expense Assumptions

Maxwell examines current expenses for the REIT's other hotel properties and selects the distributions for the simulation of operating expenses and management fees. Maxwell estimates operating expenses to be uniformly distributed between 68% and 70% of revenues and that the property management fee for the hotel is uniformly distributed between 5.9% and 6.1% of total annual revenue.

Simulation Results and Analysis

Maxwell has three concerns regarding the results from the simulation trials:

Concern 1: Although the distribution of historical CSI is slightly skewed, Maxwell uses the normal distribution to simulate the monthly CSI.

Concern 2: Property management firms may demand higher property management fees (that is, a higher percentage of revenue) when the CSI is lower.

Concern 3: When comparing the distribution of CSI over the past 20 years with those of the past 30 years, she notices a substantial difference in the mean and standard deviation of the CSI distribution.

In completing her analysis, Maxwell considers her choice of simulation analysis over alternative approaches, such as decision trees and scenario analysis, to be justified. Although scenario analysis and decision trees both consider possible outcomes, neither can be used easily when correlated variables are present, as is the case with CSI and ECI. Further, she notes that, compared with scenario analysis and decision trees, simulation is best suited for continuous risks, whether they be concurrent or sequential.

Based on the results of the simulation analysis, the REIT acquires the hotel. One year later, the REIT is considering the acquisition of another hotel, and Maxwell wants to use the same simulation model. Based on an analysis of the hotel industry, Maxwell notes that recent mergers in the industry have affected competition in the market in which this hotel operates. Consequently, Maxwell needs to update the simulation model.

1 With respect to forecasting annual gross revenue, Maxwell's *best* course of action to deal with the inputs problem should be to:

A simulate the ECI and the CSI independently.

B build the correlation explicitly into the simulation.

C estimate revenues using the ECI only and eliminate the CSI from the simulation.

2 With respect to the simulation, Maxwell should be concerned that:

A CSI does not follow a symmetric distribution.

B property management fees are not normally distributed.

C there is a lack of correlation between the CSI and the management fee percentage.

3 Maxwell's distribution assumption for the property management fee in the simulation is based on:

A historical data.

B simulation results.

C cross-sectional data.

4 Which of Maxwell's assumptions can be *best* described as a simulation constraint? The assumption about the:

A occupancy rate.

B operating expenses.

C property management fee.

5 Which of Maxwell's three concerns can be *best* described as being attributable to non-stationary distributions?

A Concern 1

B Concern 2

C Concern 3

6 Maxwell's justifications for her choice of simulation analysis are correct with respect to:

A correlated risks.

B continuous risks.

C both correlated risks and continuous risks.

7 Considering industry changes over the past year, one update that Maxwell should make to the simulation model is the:

A choice of the distribution of CSI and ECI.

B relation between CSI and occupancy rates.

C removal of constraints on occupancy rates.

SOLUTIONS

1 B is correct. Both the nightly room rate and the occupancy rate are dependent on inputs, ECI and CSI, which are correlated. If there are correlated inputs, there are two solutions to this problem. One is to allow only one of the two inputs to vary, emphasizing the one with the larger impact. The second solution is to the build the correlation explicitly into the simulation. Simulating ECI and CSI independently is not a remedy for the correlated inputs problem. Further, the relation between CSI and the occupancy rate is stronger than that between ECI and nightly room rates, which suggests that CSI should be kept if one of the two inputs is to be removed from the simulation. A is incorrect because the two input variables, ECI and CSI, are correlated, and simulating them independently is not appropriate. The remedies include dropping one of the probabilistic inputs or building the correlation into the simulation explicitly. C is incorrect because the relation between CSI and the occupancy rate is stronger than that between ECI and nightly room rates, which suggests that CSI should be kept if one of the two inputs is to be removed from the simulation.

2 A is correct. Using a normal distribution when the distribution is not normal may lead to misleading results. There is no limitation on the type of probability distribution (for example, normal or uniform); what is important is that the selected distribution reflect the likely distribution of future values. Further, there is no requirement that inputs be correlated with one another. In fact, correlated inputs present issues that must be overcome by either removing one of the probability inputs or explicitly building the correlation into the simulation. B is incorrect because probabilistic variables may follow different distributions, including the normal distribution and the uniform distribution. It is important to pick a statistical distribution that best captures the variability in the input and estimate the parameters for that distribution. C is incorrect because there is no requirement that inputs be correlated with one another. In fact, correlated inputs present issues that must be overcome by either removing one of the probability inputs or explicitly building the correlation into the simulation.

3 C is correct. Maxwell uses the distribution of property management fees for the REIT's other hotel properties to simulate the property management fee. Therefore, the property management fee distribution is based on differences in property management fees across a cross-section of the REIT's existing hotel properties that are similar to the hotel being analyzed. A is incorrect because Maxwell's distribution assumption about the property management fee is not based on historical data for the hotel. The property is new, and therefore Maxwell does not have a history of reliable data for the property management fee to use. B is incorrect because the property management fee distributional assumption is not based on the results of a simulation but rather is based on cross-sectional data. The property management fee is based on the property management fees for the REIT's other hotel properties. Once the distributional assumption is made (that is, the statistical distribution and parameters), Maxwell may then incorporate these into the trials for the management fee.

4 A is correct. Without Maxwell's assumption regarding the constraints that the occupancy rates be between 0% and 100%, the simulation could produce negative occupancy rates or rates above 100%. This assumption serves as a constraint on occupancy rates. Operating expenses are assumed to be a percentage of revenues and are not constrained; what is specified with respect to operating expenses is the distribution. Similarly, for the property management fee, the distribution is specified, but this is not a constraint. B is incorrect because the

assumption of a specific relation between operating expenses and revenues is not a constraint. C is incorrect because the assignment of a distribution, in this case a uniform distribution for the property management fee, is not a constraint. It is merely a step in a simulation in which the statistical distribution and parameters are specified.

5 C is correct. A substantial difference in the mean and standard deviation of CSI over the past 20 years relative to those of the past 30 years suggests a change in (i.e., non-stationary) distribution. Even when the data fits a statistical distribution or when historical data distributions are available, shifts in the market structure may lead to shifts in the distributions as well, as evidenced by the shift between the 20- and 30-year distributions of CSI. In some cases, such shifts can change the form of the distribution and in other cases, they can change the parameters of the distribution. A is incorrect because Concern 1 is a description of real data not fitting distributions. The problem with the real world is that the data seldom fits the stringent requirements of statistical distributions. Using probability distributions that bear little resemblance to the true distribution underlying an input variable will yield misleading results. B is incorrect because Concern 2 is an issue of correlation that is not incorporated into the simulation. Correlation across input variables can be modeled into simulations. This works only if the correlations remain stable and predictable, however. To the extent that correlations between input variables change over time, as is expressed in Concern 2, it becomes far more difficult to model them.

6 C is correct. Correlated risks are difficult to model in decision trees. In addition, adjusting scenario analysis for correlated risks is subjective. Scenario analysis and decision trees are generally built around discrete outcomes in risky events, whereas simulations are better suited for continuous risks. Further, if the various risks to which an investment is exposed are correlated, simulations allow for explicitly modeling these correlations if they can be estimated.

7 B is correct. Even when the data fits a statistical distribution or when historical data distributions are available, shifts in the market structure can lead to shifts in the distributions as well. In some cases, these shifts can change the form of the distribution, and in other cases, they can change the parameters of the distribution. Thus, relations from historical data for an input may change for the next period, affecting the relation between the nightly room rate and CSI, as well as the relation between the occupancy rate and ECI.

The constraints on the occupancy rates are necessary because without them, there may be unrealistic results (such as negative rates or rates exceeding 100%). Further, the choice of the distributions of CSI and ECI is not mentioned to be affected by the changes in the industry over the past year; rather, the relation between CSI and occupancy rates may change.

Economics

STUDY SESSION

Study Session 4 Economics

TOPIC LEVEL LEARNING OUTCOME

The candidate should be able to explain and demonstrate the use of economic concepts and methods in the determination and forecasting of currency exchange rates, the analysis of economic growth, and the analysis of business and financial market regulation.

A country's exchange rates, level of economic activity, and regulatory environment have significant implications for companies operating within its borders. Although predicting exchange rates is extremely difficult, exchange rate equilibrium relationships provide valuable insights for understanding the currency risks inherent in overseas operations and international investments.

© CFA Institute. For candidate use only. Not for distribution.