FIGURE 4.20 Regression of Unemployment on One- and Two-Period Lagged Unemployment.

| Variable | Coefficient | Std Dev | T Stat | P Value |
|---|---|---|---|---|
| Intercept | 0.16764 | 0.04565 | 3.67 | 0.000 |
| UNEMPL1 | 0.89032 | 0.06497 | 13.70 | 0.000 |
| UNEMPL2 | 0.07842 | 0.06353 | 1.23 | 0.218 |

Standard Error = 0.151381 R-Sq = 98.7% R-Sq(adj) = 98.6%

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Stat | P Value |
|---|---|---|---|---|---|
| Regression | 2 | 395.55 | 197.77 | 8630.30 | 0.000 |
| Error | 235 | 5.39 | 0.02 | | |
| Total | 237 | 400.93 | | | |

# EXERCISES

5. **Mortgage Rates.** The file on the CD named MRATES4 contains monthly 30-year conventional mortgage rates from January 1985 to December 2002. (These figures are found on the web site www.economagic.com and are obtained from the Federal Home Mortgage Corporation).

Figure 4.21 provides a time-series plot of the rates. The following model is to be used to forecast mortgage rates:

$$y_i = \beta_0 + \beta_1 y_{i-1} + e_i$$

where $y_i$ is the mortgage rate at time $i$ and $y_{i-1}$ is the rate at time $i - 1$. The regression results are shown in Figure 4.22. Note that RATEL1 in the output represents the one-period lagged variable, $y_{i-1}$. Use the output to answer the following questions:

a. What is the estimated regression equation?

b. Is there a relationship between current and previous period mortgage rates? State the hypotheses to be tested, the decision rule, the test statistic, and your decision. Use a 5% level of significance.

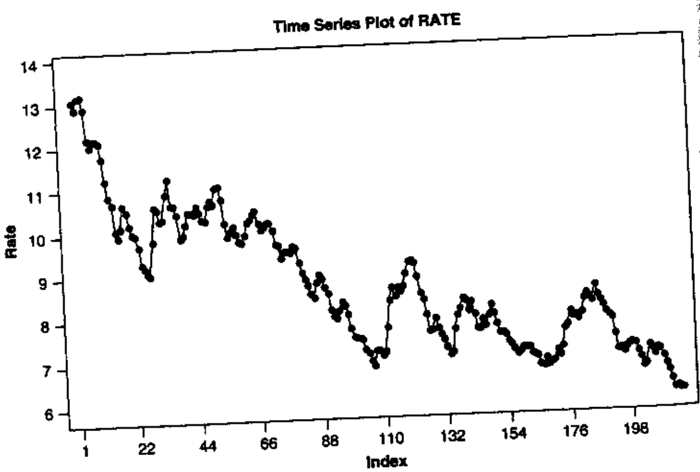


FIGURE 4.21 Time-Series Plot of Mortgage Rates.



FIGURE 4.22 Regression Results for Mortgage Rates Exercise.

| Variable | Coefficient | Std Dev | T Stat | P Value |
|---|---|---|---|---|
| Intercept | 0.16077 | 0.08854 | 1.82 | 0.071 |
| RATEL1 | 0.97774 | 0.01002 | 97.60 | 0.000 |

Standard Error = 0.235473 R-Sq = 97.8% R-Sq(adj) = 97.8%

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Stat | P Value |
|---|---|---|---|---|---|
| Regression | 1 | 528.13 | 528.13 | 9524.82 | 0.000 |
| Error | 213 | 11.81 | 0.06 | | |
| Total | 214 | 539.94 | | | |

c. What percentage of the variation in mortgage rates has been explained by the regression?

d. Use the estimated equation to produce a forecast of the mortgage rate in January 2003. Find out what the actual rate was and compare it to the forecast. How well did the equation do? Repeat this process for the remainder of 2003. Discuss any difficulties you encounter in forecasting more than 1 month ahead.

e. Test whether the intercept of the equation is equal to zero. State the hypotheses to be tested, the decision rule, the test statistic, and your decision. Use a 5% level of significance.

f. Test whether the slope of the equation is equal to 1. State the hypotheses to be tested, the decision rule, the test statistic, and your decision. Use a 5% level of significance.

# ADDITIONAL EXERCISES

6. Fill in the missing blanks on the following ANOVA table:

ANOVA

| Source | DF | SS | MS | F |
|---|---|---|---|---|
| Regression | 2 | | 250 | |
| Error (Residual) | | 400 | | |
| Total | 82 | | | |

7. Fill in the missing blanks on the following ANOVA table:

ANOVA

| Source | DF | SS | MS | F |
|---|---|---|---|---|
| Regression | | 300 | 100 | 10 |
| Error (Residual) | 27 | | | |
| Total | | | | |

8. **Wheat Exports.** The relationship between exchange rates, prices, and agricultural exports is of interest to agricultural economists. One such export of interest is wheat. The file named WHEAT4 on the CD contains data on the following variables:

$y$, U.S. wheat export shipments (SHIPMENT)
$x_1$, the real index of weighted-average exchange rates of the U.S. dollar (EXCHRATE)
$x_2$, the per-bushel real price of no. 1 red winter wheat (PRICE)

The dependent variable is U.S. wheat export shipments. The explanatory variables are exchange rate and price. The data are observed monthly from January 1974 through March 1985. The regression results are shown in Figure 4.23. Use the output to help answer the following questions:

a. What is the estimated regression equation relating SHIPMENT to EXCHRATE and PRICE?

b. Test the overall fit of the regression. State the hypotheses to be tested, the decision rule, the test statistic, and your decision. Use a 5% level of significance. What conclusion can be drawn from the result of the test?

c. After taking account of the effect of PRICE, are SHIPMENT and EXCHRATE related? Conduct a

This is referred to as the full model. The hypotheses to be tested are

$H_0$: $\beta_2 = \beta_3 = \beta_4 = 0$

$H_a$: At least one of the coefficients $\beta_2$, $\beta_3$, and $\beta_4$ is not zero

If the null hypothesis is not rejected, the seasonal indicator variables add nothing to the model and can be removed. In other words, there are no seasonal differences. In this case, the following reduced model is adopted:

$$y = \beta_0 + \beta_1 x + e$$

If the null hypothesis is rejected, then there are seasonal differences, and the quarterly indicators should remain in the model.

To conduct the test, the partial $F$ test statistic is computed as

$$F = \frac{(SSE_R - SSE_F)/(K - L)}{MSE_F}$$

where $SSE_R$ is the error sum of squares from the reduced model, $SSE_F$ is the error sum of squares from the full model, and $MSE_F$ is the mean square error from the full model. Because the hypothesis test determines whether three coefficients are equal to zero, a divisor of 3 is used in the numerator in place of $K - L$ ($K = 4$, $L = 1$). The decision rule for the test is

Reject $H_0$ if $F > F(\alpha; K - L, n - K - 1)$

Do not reject $H_0$ if $F \le F(\alpha; K - L, n - K - 1)$

**EXAMPLE 7.6** **ABX Company Sales (continued)** In Example 7.5, the following model was examined for ABX Company sales:

$$SALES = \beta_0 + \beta_1 TREND + \beta_2 Q1 + \beta_3 Q2 + \beta_4 Q3 + e$$

To determine whether there are seasonal components affecting sales, the following hypotheses should be tested:

$H_0$: $\beta_2 = \beta_3 = \beta_4 = 0$

$H_a$: At least one of the coefficients $\beta_2$, $\beta_3$, and $\beta_4$ is not zero

To test this hypothesis, the partial $F$ test is used. The test statistic is

$$F = \frac{(9622 - 1810)/3}{52} = 50.1$$

Note that the reduced model regression is in Figure 7.19, and the full model regression is in Figure 7.22. Using a 5% level of significance, the decision rule for the test is

Reject $H_0$ if $F > F(0.05; 3.35) = 2.92$ (approximately)

Do not reject $H_0$ if $F \le 2.92$

The null hypothesis is rejected. Thus, the conclusion is that seasonal components affect sales and should be taken into account, as was done in Example 7.5.



When computing forecasts with seasonal models, the coefficients of the seasonal indicators are used to adjust the level of the forecast in the appropriate time periods. The quarterly forecasts for the year 2004 are as follows, using the seasonal model with trend:

| Time Period | Point Forecast |
|---|---|
| 2004 Q1 | 211 + 2.57(41) + 3.75 = 320.12 |
| 2004 Q2 | 211 + 2.57(42) − 26.1 = 292.84 |
| 2004 Q3 | 211 + 2.57(43) − 25.8 = 295.71 |
| 2004 Q4 | 211 + 2.57(44) = 324.08 |

Throughout this section, the use of quarterly indicator variables has been discussed. If monthly instead of quarterly data are used, the applications are similar. Instead of four quarterly indicators, twelve monthly indicators are created. Eleven of the twelve monthly indicators are used in the regression with the estimated coefficients interpreted in a manner similar to those of the quarterly coefficients. Tests for seasonal variation involve the set of eleven indicator variable coefficients. Since the use of monthly indicators is so similar to quarterly indicators, the demonstration of their use is reserved for the exercises.

## EXERCISES

5. **Furniture Sales.** The file named FURNSALE7 on the CD contains data on monthly furniture sales (in millions of dollars) for the United States from January 1992 through December 2002. These data were obtained from the web site www.economagic.com. Figure 7.25 shows the time-series plot of the data. An extrapolative model to forecast furniture sales for each month in 2003 was developed. Figure 7.26 shows the regression results for the model. The model can be written as follows:

$$SALES = \beta_0 + \beta_1 TREND + \beta_2 LAGSALES + \beta_3 JAN + \beta_4 FEB + \beta_5 MARCH + \beta_6 APRIL + \beta_7 MAY + \beta_8 JUNE + \beta_9 JULY + \beta_{10} AUG + \beta_{11} SEPT + \beta_{12} OCT + \beta_{13} NOV + e$$

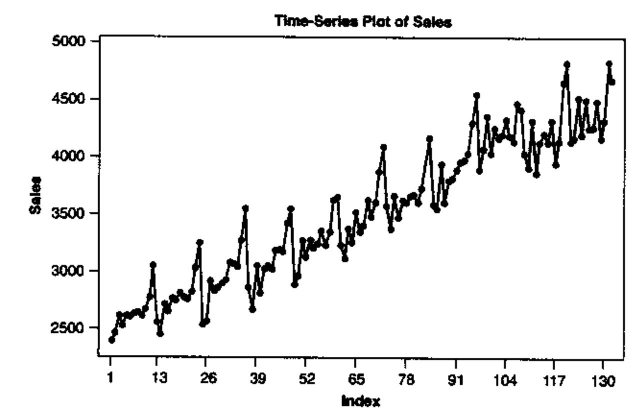


**FIGURE 7.25** Time-Series Plot of Furniture Sales.

**FIGURE 7.26** Regression Results with Seasonal Indicators for Furniture Sales Exercise.

| Variable | Coefficient | Std Dev | T Stat | P Value |
|---|---|---|---|---|
| Intercept | 1552.151 | 228.680 | 6.79 | 0.000 |
| TREND | 7.903 | 1.298 | 6.09 | 0.000 |
| LAGSALES | 0.484 | 0.083 | 5.84 | 0.000 |
| JAN | -643.717 | 42.576 | -15.12 | 0.000 |
| FEB | -404.481 | 53.915 | -7.50 | 0.000 |
| MARCH | -80.369 | 56.921 | -1.41 | 0.161 |
| APRIL | -457.714 | 42.852 | -10.68 | 0.000 |
| MAY | -178.202 | 53.127 | -3.35 | 0.001 |
| JUNE | -316.796 | 45.155 | -7.02 | 0.000 |
| JULY | -278.410 | 47.612 | -5.85 | 0.000 |
| AUG | -183.319 | 47.143 | -3.89 | 0.000 |
| SEPT | -357.504 | 43.251 | -8.27 | 0.000 |
| OCT | -239.408 | 48.319 | -4.95 | 0.000 |
| NOV | 6.826 | 45.791 | 0.15 | 0.882 |

Standard Error = 93.2329  R-Sq = 98.0%  R-Sq(adj) = 97.8%

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Stat | P Value |
|---|---|---|---|---|---|
| Regression | 13 | 49839825 | 3833833 | 441.06 | 0.000 |
| Error | 117 | 1017007 | 8692 | | |
| Total | 130 | 50856832 | | | |

**FIGURE 7.27** Regression Results with Seasonal Indicators for Furniture Sales Exercise.

| Variable | Coefficient | Std Dev | T Stat | P Value |
|---|---|---|---|---|
| Intercept | 1977.945 | 215.778 | 9.17 | 0.000 |
| TREND | 12.483 | 1.427 | 8.75 | 0.000 |
| LAGSALES | 0.201 | 0.087 | 2.32 | 0.022 |

Standard Error = 195.734  R-Sq = 90.4%  R-Sq(adj) = 90.2%

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Stat | P Value |
|---|---|---|---|---|---|
| Regression | 2 | 45952938 | 22976469 | 599.72 | 0.000 |
| Error | 128 | 4903894 | 38312 | | |
| Total | 130 | 50856832 | | | |

TREND is a linear trend variable, LAGSALES is SALES lagged one period, and JAN through NOV are 11 monthly seasonal indicators. Note that December has been used as the base-level month.

Figure 7.27 shows the regression results for the model without the seasonal indicators.

Use the regression results to help answer the following questions.

a. Is there seasonal variation in furniture sales? State the hypotheses to be tested, your decision rule, the test statistic, decision, and conclusion. Use a 5% level of significance.

b. Are the TREND and LAGSALES variables important to the equation explaining furniture sales? Test the coefficient of each of these variables individually to answer this question. State the hypotheses to be tested, your decision rule, the test statistic, decision, and conclusion. Use a 5% level of significance.

c. Based on the test result in part a, use either the full or reduced model (with or without seasonal indicators) to develop a forecast of furniture sales for each month in 2003.



# ADDITIONAL EXERCISES

6. **Absenteeism.** Data on 77 employees of the ABX Company have been collected. The dependent variable is absenteeism (ABSENT). The possible explanatory variables are

COMPLX = measure of job complexity
SENIOR = seniority
SATIS = response to "How satisfied are you with your foreman?"

These data are available in a file named ABSENT7 on the CD. In this exercise, use SENINV = 1/SENIOR, which is the reciprocal of the seniority variable, and COMPLX as two of the explanatory variables. The variable SATIS should be transformed into indicator variables as follows:

FS1 = 1 if SATIS = 1 (very dissatisfied)
= 0 otherwise
FS2 = 1 if SATIS = 2 (somewhat dissatisfied)
= 0 otherwise
FS3 = 1 if SATIS = 3 (neither satisfied nor dissatisfied)
= 0 otherwise
FS4 = 1 if SATIS = 4 (somewhat satisfied)
= 0 otherwise
FS5 = 1 if SATIS = 5 (very satisfied)
= 0 otherwise

Five indicator variables are created to represent all five supervisor satisfaction categories. Recall that only four need to be used in the regression. Run the regression with the explanatory variables described here. Answer the following questions:

a. Is there a difference in average absenteeism for employees in different supervisor satisfaction groups? Perform a hypothesis test to answer this question. State the hypotheses to be tested, the decision rule, the test statistic, and your decision. Use a 5% level of significance.

b. Using the model chosen in part a (and keeping the variables COMPLX and SENINV in the model), what would be your estimate of the average absenteeism rate for all employees with COMPLX = 60 and SENIOR = 30 who were very dissatisfied with their supervisor? What if they were very satisfied with their supervisor, but COMPLX and SENIOR were the same values?

c. How do you account for the differences in the estimates in part b?

d. How could this equation be used to help identify employees who might be prone to absenteeism?

7. **Work-Order Closing.** Management at the Texas Christian University (TCU) Physical Plant is interested in reducing the average time to completion of routine work orders. The time to completion is defined as the difference between the date of receipt of a work order and the date closing information is entered. The number of labor hours charged to each work order and the cost of materials are two variables believed to be related to the time to completion of the work order. Management wants to know if there is any difference in the time to completion of work orders, on average, for different types of buildings. Buildings are classified into four types on the TCU campus: residence halls, athletic, academic, and administrative. In answering the question, take into account the possible effect of labor hours charged and materials cost. The data for a random sample of 72 work orders (chosen from a population of 11,720) are available in a file named WORKORD7 on the CD. The variables are as follows:

$y$ = DAYS = number of days to complete each work order
$x_1$ = HOURS = number of hours of labor charged to each work order
$x_2$ = MATERIAL = cost of materials charged to each work order
$x_3$ = BUILDING = 1 for residence halls
2 for athletic buildings
3 for academic buildings
4 for administrative buildings

8. **Beer Production.** The file named BEER7 on the CD contains monthly U.S. beer production in millions of barrels for January 1982 through December 1991. Develop an extrapolative model for these data and use it to examine whether there is seasonal variation in beer production and whether beer production seems to be increasing, decreasing, or staying fairly constant over this time period. Use the model you select as best for

The $p$ value can also be used to conduct the test in the usual manner:

Decision rule: Reject $H_0$ if $p$ value $< 0.05$

Do not reject $H_0$ if $p$ value $\geq 0.05$

Test statistic: $p$ value $= 0.0007$

The Z test in MINITAB and the $\chi^2$ test in SAS are essentially equivalent, and serve the same purpose.

Above this table in the Testing Global Null Hypothesis: BETA = 0 table, information is provided for a test of whether all slopes are zero. This is similar to the overall-fit $F$ test in linear regression. There are three different test statistics provided (Likelihood Ratio, Score and Wald) which can be used to conduct the test. For any of the three options, the $p$ value provided can be used in the usual way to perform this test. The Likelihood Ratio appears to be one of the more dependable approaches for this test.

The remainder of the output is not discussed in this text. ■

When using logistic regression, the goal of the analysis may be to classify the observations into a particular group (as in discriminant analysis). In this case, some rule must be designed to help decide into which group each observation should be classified. Predicted values of the probability of group membership in the indicated group can be computed from the logistic regression. A rule using these predicted probabilities can then be designed. The form of such a rule is

Classify observations into the

$y = 0$ group if the predicted value is below the cutoff

$y = 1$ group otherwise

A cutoff value of 0.5 is reasonable when the 0 and 1 outcomes are equally likely and the costs of misclassification into each group are about equal. In other cases, a different cutoff may be considered superior.[3]

# EXERCISES

1. **Employee Classification.** Figure 10.8 shows summary logistic regression output for the employee classification data from Example 10.4 using both test results as explanatory variables. Use the output shown to answer the following questions. The data are in a file named TRAIN10 on the CD.
   a. Which variable or variables appear to be useful in the logistic regression function? Justify your answer. Use a 5% level of significance for any hypotheses tests.
   b. Using a cutoff probability of 0.5, classify potential employees whose test scores were as follows:
      1. TEST1 = 94 TEST2 = 88
      2. TEST1 = 80 TEST2 = 87
      3. TEST1 = 82 TEST2 = 74
      4. TEST1 = 90 TEST2 = 80

[3] For a more complete presentation of logistic regression, see C. E. Lunneborg, *Modeling Experimental and Observational Data*, Chapters 16, 17, and 18, or J. Neter, W. Wasserman, and M. Kutner, *Applied Linear Regression Models*, Chapter 16.



**FIGURE 10.8** Logistic Regression Output for Employee Classification Example Using TEST1 and TEST2 as Independent Variables.

Response Information

| Variable | Value | Count | |
|---|---|---|---|
| Group | 1 | 23 | (Event) |
| | 0 | 20 | |
| | Total | 43 | |

Logistic Regression Table

| Predictor | Coefficient | StDev | Z | P | Odds Ratio | 95% CI Lower | Upper |
|---|---|---|---|---|---|---|---|
| Constant | -56.1704 | 17.4516 | -3.22 | 0.001 | | | |
| TEST1 | 0.483314 | 0.157779 | 3.06 | 0.002 | 1.62 | 1.19 | 2.21 |
| TEST2 | 0.165218 | 0.102070 | 1.62 | 0.106 | 1.18 | 0.97 | 1.44 |

Log-Likelihood = -13.959

Test that all slopes are zero: G = 31.483, DF = 2, P-Value = 0.000

2. **Harris Salaries.** In Exercise 1 in Chapter 7, data from Harris Bank were examined to test for possible discrimination. In that exercise, the dependent variable was salary and one of the explanatory variables was an indicator variable to separate the employees into male and female groups. The coefficient of the indicator variable served as a measure of whether males earned more (or less), on average, than females.

   Another way of examining this problem might be to use the male/female indicator variable as the dependent variable and see if group membership can be predicted from knowledge of salary. This can be done with either discriminant analysis or logistic regression. Try discriminant analysis and/or logistic regression and see how well these methods do in predicting whether employees are male or female based only on knowledge of their salary. Does your result support the claim that Harris Bank discriminated by underpaying female employees?[4]

   The data are defined as follows and are in the file named HARRIS10 on the CD:

   $x_1$ = beginning salaries in dollars (SALARY)
   $x_2$ = years of schooling at the time of hire (EDUCAT)
   $x_3$ = number of months of previous work experience (EXPER)
   $x_4$ = number of months after January 1, 1969, that the individual was hired (MONTHS)
   $y$ = variable coded 1 for males and 0 for females (MALES)

   (*Source:* These data were obtained from D. Schafer, "Measurement-Error Diagnostics and the Sex Discrimination Problem," *Journal of Business and Economic Statistics*. Copyright 1987 by the American Statistical Association. Used with permission. All rights reserved.)

3. **Computer Purchase.** The Daleway Corporation owns stores throughout the United States that carry its brand of computer. Management would like some information on the purchasing practices of the American public. Specifically, it would like to know if certain factors affect the decision to upgrade a system with the purchase of a new computer. The company surveyed 40 recent customers who were considering upgrading and collected information on the following variables:

   PURCHASE, coded as 1 if the customer purchased a computer, 0 if the customer did not
   INCOME, the household income of the customer (in thousands of dollars)
   AGE, the age of the customer's current computer

[4] Note: Using discriminant analysis or logistic regression in this manner would probably not be the preferred method of examining this question in a legal proceeding. The regression approach discussed in Chapter 7 would be preferred, but this makes an interesting exercise.

Are either of the variables INCOME or AGE of use in predicting whether the customer will purchase a new computer? Justify your answer. Use a 5% level of significance for any hypothesis tests. The file containing these data is named COMPPURCH10 on the CD.

4. **March Madness (Men's Tournament).** Each March the games in the men's NCAA (National Collegiate Athletic Association) basketball tournament begin. Sixty-four teams are selected for the tournament. Each team is given a number from 1 to 16, called the *seed*. There are 4 teams assigned to each seed number (4 number 1s, 4 number 2s, etc). The lower the seed number, the higher the perceived quality of the team. In the first round of the tournament these 64 teams are paired according to their seed. Number 1 seeds play number 16 seeds; number 2 seeds play number 15 seeds; number 3 seeds play number 14 seeds and so on. The games are distributed in brackets with 16 teams (consisting of 1 through 16 seeds) in each bracket. The seed numbers are assigned by a committee formed by the NCAA. It is a commonly accepted fact that the seed numbers are of use in predicting the winner of the games. For example, a number 16 seed has never beaten a number 1 seed. This seems reasonable. Number 1 seeds, the teams of highest quality, should have a high probability of beating number 16 seeds. But do the seed numbers contain all the information useful in predicting the winners, or are there other team characteristics that might be of use? The file NCAAMEN10 on the CD contains data on the following variables:

WIN = 1 if the higher seed won the game
= 0 otherwise

DIFF = lower seed number–higher seed number; note that DIFF will always be negative (or zero if two teams of equal seed play)

PCTHIGH = winning percentage for the higher-seeded team

PCTLOW = winning percentage for the lower-seeded team

There is one additional column in the file labeled ROUND. This column indicates in which round of the tournament the game took place. Although it might be of interest as an explanatory variable, it was included primarily for information.

ROUND = 1, first-round game (32 total games)
= 2, second-round game (16 games pairing the 32 teams who survived the first round)
= 3, third-round games (8 games—these 16 teams are often referred to as the Sweet 16)
= 4, fourth-round games (4 games—these 8 teams are often referred to as the Elite 8 or the Great 8)
= 5, fifth-round games (2 games—Final 4 teams)
= 6, championship game

Use logistic regression to determine whether DIFF, PCTHIGH, or PCTLOW are useful in predicting the winners of games in the NCAA basketball tournament. Use a 5% level of significance in any hypothesis tests. Are there other variables that might be useful in addition to ones given in this problem?

5. **March Madness (Women's Tournament).** Each March the games in the women's NCAA (National Collegiate Athletic Association) basketball tournament begin. Sixty-four teams are selected for the tournament. Each team is given a number from 1 to 16, called the *seed*. There are 4 teams assigned to each seed number (4 number 1s, 4 number 2s etc). The lower the seed number, the higher the perceived quality of the team. In the first round of the tournament these 64 teams are paired according to their seed. Number 1 seeds play number 16 seeds; number 2 seeds play number 15 seeds; number 3 seeds play number 14 seeds, and so on. The games are distributed in brackets with 16 teams (consisting of 1 through 16 seeds) in each bracket. The seed numbers are assigned by a committee formed by the NCAA. It is a commonly accepted fact that the seed numbers are of use in predicting the winner of the games. But do the seed numbers contain all the information useful in predicting the winners, or are there other team characteristics that might be of use? The file NCAAWOMEN10 on the CD contains data on the following variables:

WIN = 1 if the higher seed won the game
= 0 otherwise

DIFF = lower seed number–higher seed number; note that DIFF will always be negative (or zero if two teams of equal seed play)

PCTHIGH = winning percentage for the higher-seeded team

PCTLOW = winning percentage for the lower-seeded team

There is one additional column in the file labeled ROUND. This column indicates in which round of the tournament the game took place. Although it might be of interest as an explanatory variable, it was included primarily for information.

ROUND = 1, first-round game (32 total games)
= 2, second-round game (16 games pairing the 32 teams who survived the first round)
= 3, third-round games (8 games—these 16 teams are often referred to as the Sweet 16)
= 4, fourth-round games (4 games—these 8 teams are often referred to as the Elite 8 or the Great 8)
= 5, fifth-round games (2 games—Final 4 teams)
= 6, championship game

Use logistic regression to determine whether DIFF, PCTHIGH, or PCTLOW are useful in predicting the winners of games in the NCAA basketball tournament. Use a 5% level of significance in any hypothesis tests. Are there other variables that might be useful in addition to ones given in this problem?

6. **Loan Performance.** The National Bank of Fort Worth, Texas, wants to examine methods for predicting sub-par payment performance on loans. It has data on unsecured consumer loans made over a 3-day period in October 1994 with a final maturity of 2 years. There are a total of 348 observations in the sample. The data, which have been transformed to provide confidentiality, include the following:

PASTDUE: Coded as 1 if the loan payment is past due and zero otherwise.
CBSCORE: Score generated by the CSC Credit Reporting Agency. Values range from 400 to 8390, with higher values indicating a better credit rating.
DEBT: This is a debt ratio calculated by taking required monthly payments on all debt and dividing it by the gross monthly income of the applicant and coapplicant. This ratio represents the amount of the applicant's income that will go toward repayment of debt.
GROSSINC: Gross monthly income of the applicant and coapplicant.
LOANAMT: Loan amount.

You have been asked to examine the feasibility of predicting past-due loan payment. Describe your results to the bank in a two-part report. The report should include an executive summary with a brief nontechnical description of your results and an accompanying technical report with the details of your analysis. The data are in a file named LOAN10 on the CD.



# USING THE COMPUTER

The Using the Computer section in each chapter describes how to perform the computer analyses in the chapter using Excel, MINITAB, and SAS. For further detail on Excel, MINITAB, and SAS, see Appendix C.

## EXCEL

### *Discriminant Analysis and Logistic Regression*

There are currently no options for either discriminant analysis or logistic regression in Excel.