

DIRECTIONS:

- Some exercises include special instructions which modify or clarify textbook instructions.
 - Use a 10% significance level for tests unless noted otherwise.
 - Data sets for all exercises are available on the MINITAB Data Sets link.
-

Textbook exercises are *grouped together* around the same word problem and labeled A, B, C, etc.

A. Computer Science Students**Directions:**

Before working Exercise 11.45 turn to Topic 9 Part 1 Example 5 on page 110 in the Notes. Open the data file CSDATA and reproduce MINITAB output for both the full and reduced models on page 113. Were you successful?

- Exercise 11.45 (p. 648)
 - (a) Use Four Steps and 5% significance.

Add part (b) :

- (b) Example 5 shows that SAT scores are not important predictors if used *in addition* to high school grades. Are SAT scores important if used by *themselves*? Support your answer with appropriate evidence.

B. Price-Fixing Litigation

- Exercise 11.108 (p. 678)

Note: If the actual price can be explained by supply and demand, there is no evidence of price-fixing. (Use 95% certainty in part (a).)

C. A Taste of Cheese**Directions:**

1. There's some quick background reading for this one. Please read the one-paragraph description of Case Study 1 at the bottom of page 681. Then read more details about the **Cheese** data set on page A-1 in the Data Appendix near the back of the book. (Assume that Taste is measured in points and that acid concentrations in cheese are measured as percents.)

2. Make a matrix plot for all of the variables. (See Topic 9 Part 1 Example 1 on page 83 in the Notes for the MINITAB steps.) Also get a matrix of correlations to go along with the plots:

Stat > Basic Statistics > Correlation > (Select all variables) > OK

3. Try all of the different models (combinations of predictor variables), using
Regression > Regression

Observe which models have significant t tests at the 10% level. (There are 7 different models to check.)

Then answer these questions:

- (a) Describe the correlations (positive or negative?) between all four variables by looking at patterns in the Matrix Plot.
- (b) Which predictor variable is most highly correlated with Taste? Which two predictors are most highly correlated with each other?
- (c) What does the F test from the *full* model indicate?
- (d) Identify a predictor variable which is ineffective when used in combination with any other predictor. (Use 10% significance.)
- (e) Which regression models (i.e. which combinations of predictors) are possible *candidates* for the “best conservative” regression model, as defined in the Topic 9 Notes?
- (f) Choose the best conservative model. Write down the sample regression equation from MINITAB output.
- (g) Make a residuals plot for your chosen model. (Create a “residuals vs. fits” graph as part of the regression.) Does it appear that the model can be safely used? Why or why not?
- (h) Use the chosen model to interpret the slopes for each of the predictor variables in the data set.
- (i) Is Taste related to Acetic? If so, provide an interpretation.
- (j) Predict with 99% certainty the taste of a single batch of cheese which has 5.4% Acetic, 6.9% H₂S, and 1.5% Lactic. (Use the chosen model.) Interpret the answer.
- (k) Suppose that your business client insists that the variable H₂S be included in the model used for prediction. Write down the equation of the modified best conservative model (at the 10% significance level.)
- (l) Suppose that your business client insists that the variable Acetic be included in the model used for prediction. Write down the equation of the modified best conservative model (at the 10% significance level.)

D. Bank Wages

Look back at Topic 9 Part 2 Example 6 (Bank wages) on page 137 in the Notes.

First: To “get your feet wet” open Table 10.1 in MINITAB and follow the steps on page 138 to make the special scatterplot, create the variable *Newsize*, and run the regression. Did you successfully recreate the graph and output on pages 138-139?

Our target population is all female customer-service representatives at Indiana banks.

The *response* variable is

$$y = \text{Weekly Wages (dollars)}$$

The *predictor* variables are

$$\begin{aligned}x_1 &= \text{Length of Service (months)} \\x_2 &= \text{Bank Size} = \begin{cases} 0 & \text{if small} \\ 1 & \text{if large} \end{cases}\end{aligned}$$

The variable x_2 is a *binary* variable since it assumes only the two values 0 and 1.

Note: Such variables are also called *dummy* variables or *indicator* variables since they use the “dummy” numbers 0 and 1 to “indicate” different categories.

The population regression equation is

$$\mu_{\text{Wages}} = \beta_0 + \beta_1 \times \text{LOS} + \beta_2 \times \text{Size}$$

Or equivalently:

$$\text{Wages} = \beta_0 + \beta_1 \times \text{LOS} + \beta_2 \times \text{Size} + \varepsilon$$

Questions:

- Write the theoretical regression equation for the subpopulation of women who work at small banks. (Hint: What is the value of x_2 ?)
- Identify the starting wage at small banks in the population.
- Find the sample estimate of the starting wage at small banks.
- Write the theoretical regression equation for the subpopulation of women who work at large banks.
- Identify the starting wage at large banks in the population.
- Find the sample estimate of the starting wage at large banks.

E. Hourly Wages

Directions:

1. Read the brief description of Case Study 3 on pages 681-682, then read the note about the Hourly database on page A-4 in the Data Appendix.
2. Open the MINITAB data file and create three binary dummy variables:

$$x_1 = \text{Newgender} \quad x_2 = \text{Newstatus} \quad x_3 = \text{Newrace}$$

original value of Gender	Newgender	original value of Status	Newstatus	original value of Race	Newrace
F	0	PT	0	Black	0
M	1	FT	1	White	1
				Other	0

Questions:

- (a) How many workers are there in the database?
- (b) Write the population regression equation.
- (c) Write the equation for the subpopulation of whites.
- (d) Write the equation for the subpopulation of nonwhites.
- (e) Consider the two previous answers. What does the slope β_3 represent?
- (f) The case study mentioned concern about a lawsuit. Does there appear to be earnings discrimination by gender? Explain.
- (g) Does there appear to be earnings discrimination by race? Explain.
- (h) Choose the best conservative model at 10% significance. Use this model to answer all remaining questions. Write the sample regression equation.
- (i) Interpret the slopes of all predictors variables which are available in the database.
- (j) Estimate mean earnings for all female, nonwhite, part-time employees with 95% certainty. (Hint: The numbers to enter under Options are all either 0 or 1.)
- (k) Estimate mean earnings for all female, white, full-time employees with 95% certainty.

F. Home Prices

Directions:

1. Read the one-paragraph description of Case 11.3 in the middle of page 654. Look at Table 11.5 on page 655. The objective is to model Price based on a home's many characteristics.
2. Initially we'll try linear and quadratic regression with the predictor variable SqFt. Just as we did in Topic 9 Part 1 Example 4, we'll need to create a new variable $(\text{SqFt})^2$ in a MINITAB column.

Open Table 11.5 in MINITAB, name a new column "SqFt Squared" and use the MINITAB calculator to put the squared values of SqFt into the column. Then create a matrix plot for the three variables Price, SqFt, SqFt Squared.

Questions:

- (a) From the matrix plot, does the relationship of Price to SqFt seem to be more of a straight line (good for linear regression) or a curve (good for quadratic regression)?
- (b) Describe the relationship between SqFt and SqFt Squared.
- (c) Try both linear regression and quadratic regression using SqFt. Which of these two models would you recommend, and why?
- (d) Now consider multiple regression, beginning with all available predictors: SqFt, Bedrooms, Baths, Garage. Find the best conservative model at 10% significance. Write down the sample regression equation.
- (e) How much does R^2 improve by using the chosen model rather than using simple regression on SqFt alone?
- (f) Interpret all four slopes. (Use the chosen model.)
- (g) Estimate the mean price for all homes which have 1200 square feet, 2 bedrooms, 1.5 bathrooms, and garages with room for 2 cars. Then repeat the estimate, but this time with 90% certainty. (Use the chosen model.)

G. Self-Concept

Read the one-paragraph description of Case Study 2 on page 681. Briefly review the **Concept** data on Page A-2 in the Data Appendix.

Directions: Apply the Drop Method to these data to choose a model, using 10% significance.

Questions:

- (a) List the sequence of dropped variables, in the order in which they are dropped.
- (b) Write the theoretical (population) regression equation for the chosen model.
- (c) Write the fitted (sample) regression equation for the chosen model.
- (d) Suppose that our client is a famous psychologist who theorizes that the Piers-Harris Children's Self-Concept scale (SC) is an indispensable factor in determining GPA. Do the data support this theory?
- (e) Interpret the slopes of all potential predictor variables, using the chosen model.
- (f) Use the chosen model to predict with 95% certainty the GPA for a student who has the following profile:

<u>Variable</u>	<u>Value</u>
IQ	100
Age	12
SC	55
Sex	2
C1	15
C2	14
C3	13
C4	12
C5	11
C6	10

- (g) Suppose our client insists that the predictor SC is so important that it must be included in the model. Perform a “modified Drop Method” to find such a model. Write the sample equation.
- (h) In hindsight, was the client wise to insist on a “modified” model? Specifically, does the modified model do a better job of explaining the variability of GPA in the sample data?

H. Predicting Blood Pressure

The data in the file `Peru` were collected in a study conducted by anthropologists to determine the long-term effects of a change in environment on blood pressure. The anthropologists measured both *Systolic* blood pressure and *Diastolic* blood pressure for a number of Indians who had migrated from a very primitive environment, high in the Andes mountains of Peru, into the mainstream of Peruvian society, at a much lower altitude. (These two variables are both considered response variables.)

Predictor variables measured were Age (years), Height (millimeters), Weight (kilograms), Pulse Rate (beats per minute), Years since migration, and three skin-fold measurements (in millimeters) meant to measure obesity in the chin, forearm, and calf, respectively.

Questions:

- (a) Suppose one of the goals of the anthropologists is to obtain the best-possible understanding of the relationship between Systolic blood pressure and Weight. Which model should be used? Interpret the best-available slope. (**Hint: Re-read “Three Recommended Models” on page 120 in the Notes.**)
- (b) Can the three skin-fold measurements (as a group) be dropped from the full regression model to predict Systolic blood pressure, using 5% significance? (Show Four Steps.)
- (c) Can the standard body measurements Height and Weight be dropped (as a group) when predicting Systolic blood pressure, using 5% significance?
- (d) Now apply the Drop Method to choose a regression model to predict Systolic blood pressure, using 5% significance. (**Use this model to answer all remaining questions.**) List the sequence of dropped variables in the order in which they are dropped.
- (e) Give the fitted equation.
- (f) Interpret the slopes of all potential predictor variables.
- (g) Estimate with 95% certainty the Systolic blood pressure of a Peruvian Indian who has the following profile:

Age	26
Years	7
Weight	62.0
Height	1549
Chin	6.3
Forearm	11.5
Calf	3.8
Pulse	69

I. Predicting Corn Yield

Directions:

1. Briefly read Exercises 11.114, 11.115, 11.116 on pages 679-680 to see what questions are being asked but do not actually answer them! Glance at U.S. Corn Yield and U.S. Soybean yield in Table 11.10 and Table 11.11 on page 679. (The measurements are in bushels per acre.)
2. Open both Table 11.10 and Table 11.11 in MINITAB. Copy and paste the “SoyBeanYield” column from Table 11.11 into a column in Table 11.10 so that you have data for Year, CornYield, and SoyBeanYield together in one worksheet. (Scroll down the worksheet to check that you successfully copied 44 years of data.)
3. Make a Matrix Plot for the three variables.

Questions:

- (a) Suppose we wish to predict CornYield from Year and/or SoyBeanYield. Which model do you recommend and why? Write down the fitted (sample) equation for your recommended model. What percentage of variation in CornYield is explained by the model?
- (b) If past trends continue into the future, by approximately how much can we expect corn yield to improve each year, on average?
- (c) If past trends continue into the future, by approximately how much can we expect corn yield to improve each year on average, even if soybean yield remains flat (constant)?

(end of assignment)