**TABLE 11.9** Summary of Bivariate and Multiple Regression

| | BIVARIATE REGRESSION | MULTIPLE REGRESSION | |
|---|---|---|---|
| Model | $E(Y) = \alpha + \beta X$ | $E(Y) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k$ | |
| Prediction equation | $\hat{Y} = a + bX$ | $\hat{Y} = a + b_1 X_1 + \cdots + b_k X_k$ | |
| | | Simultaneous effect of $X_1 \ldots, X_k$ | Partial effect of one $X_i$ |
| Properties of measures | $b$ = Slope<br>$r$ = Pearson correlation, standardized slope, $-1 \leq r \leq 1$, $r$ has the same sign as $b$ | $R$ = Multiple correlation, $0 \leq R \leq 1$ | $b_i$ = Partial slope<br>$b_i^*$ = Standardized regression coefficient |
| | $r^2$ = Coefficient of determination, PRE measure, $0 \leq r^2 \leq 1$ | $R^2$ = Coefficient of multiple determination, PRE measure, $0 \leq R^2 \leq 1$ | $r_{YX_i \cdot —}$ = Partial correlation, $-1 \leq r_{YX_i \cdot —} \leq 1$, same sign as $b_i$ and $b_i^*$, $r^2_{YX_i \cdot —}$ is PRE measure |
| Tests of no association | $H_0$: $\beta = 0$ or $H_0$: $\rho = 0$, $Y$ not associated with $X$ | $H_0$: $\beta_1 = \cdots = \beta_k = 0$ or $H_0$: $\mathbf{P} = 0$, $Y$ not associated with $X_1, \ldots X_k$ | $H_0$: $\beta_i = 0$, or $H_0$: $\rho_{YX_i \cdot —} = 0$, $Y$ not associated with $X_i$, controlling for other $X$ variables |
| Test statistic | $t = \frac{b}{\hat{\sigma}_b} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$<br>$df = n - 2$ | $F = \frac{\text{Regression mean square}}{\text{Error mean square}} = \frac{R^2/k}{(1-R^2)/[n-(k+1)]}$,<br>$df_1 = k$, $df_2 = n - (k+1)$ | $t = \frac{b_i}{\hat{\sigma}_{b_i}} = \frac{r_{YX_i \cdot —}}{\sqrt{\frac{1-r^2_{YX_i \cdot —}}{n-(k+1)}}}$<br>$df = n - (k+1)$ |

The model studied in this chapter is still somewhat restrictive in the sense that all the predictors are quantitative. The next chapter shows how to include qualitative predictors in the model.

## PROBLEMS

### *Practicing the Basics*

1. In Table 9.16 in Problem 9.24 regarding Florida counties, refer to the variables $Y$ = crime rate (number per 1000 residents), $X_1$ = median income (thousands of dollars), and $X_2$ = percent in urban environment.
   **a)** Figure 11.12 shows a SAS scatter diagram relating $Y$ to $X_1$. Predict the sign that the estimated effect of $X_1$ has in the prediction equation $\hat{Y} = a + bX_1$. Explain.
   **b)** Figure 11.13 shows a SAS partial regression plot relating $Y$ to $X_1$, controlling for $X_2$. Predict the sign that the estimated effect of $X_1$ has in the prediction equation $\hat{Y} = a + b_1 X_1 + b_2 X_2$. Explain.
   **c)** Table 11.10 shows part of a SAS printout for the bivariate and multiple regression models. Report the prediction equation relating $Y$ to $X_1$, and interpret the slope.

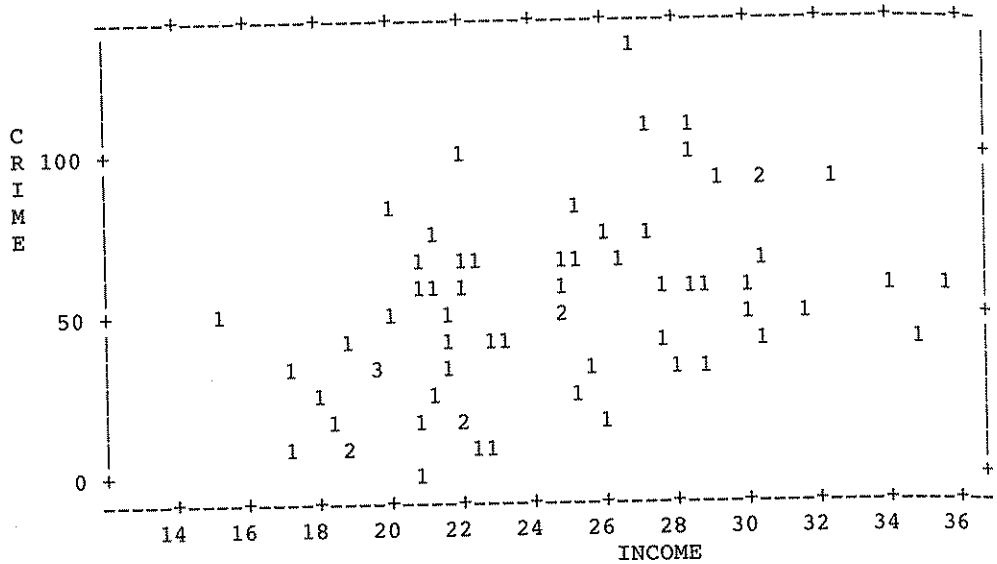


**Figure 11.12**

**d)** Report the prediction equation relating $Y$ to both $X_1$ and $X_2$. Interpret the coefficient of $X_1$, and compare to (c).
**e)** The Pearson correlations are $r_{YX_1} = .43$, $r_{YX_2} = .68$, $r_{X_1X_2} = .73$. Use these to explain why the $X_1$ effect seems so different in (c) and (d).
**f)** Report the prediction equations relating crime rate to income at urbanization levels of (i) 0, (ii) 50, (iii) 100. Interpret.

**TABLE 11.10**

| Variable | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|
| INTERCEP | -11.526 | 16.834 | -0.685 | 0.4960 |
| INCOME | 2.609 | 0.675 | 3.866 | 0.0003 |

| Variable | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|
| INTERCEP | 40.261 | 16.365 | 2.460 | 0.0166 |
| INCOME | -0.809 | 0.805 | -1.005 | 0.3189 |
| URBAN | 0.646 | 0.111 | 5.811 | 0.0001 |

2. For students at Walden University, the relationship between $Y$ = college GPA (with range 0–4.0) and $X_1$ = high school GPA (range 0–4.0) and $X_2$ = college board score (range 200–800) satisfies $E(Y) = .20 + .50X_1 + .002X_2$.
**a)** Find the mean college GPA for students having (i) high school GPA = 4.0 and college board score = 800, (ii) $X_1 = 3.0$ and $X_2 = 300$.
**b)** Show that the relationship between $Y$ and $X_1$ for those students with $X_2 = 500$ is $E(Y) = 1.2 + .5X_1$.

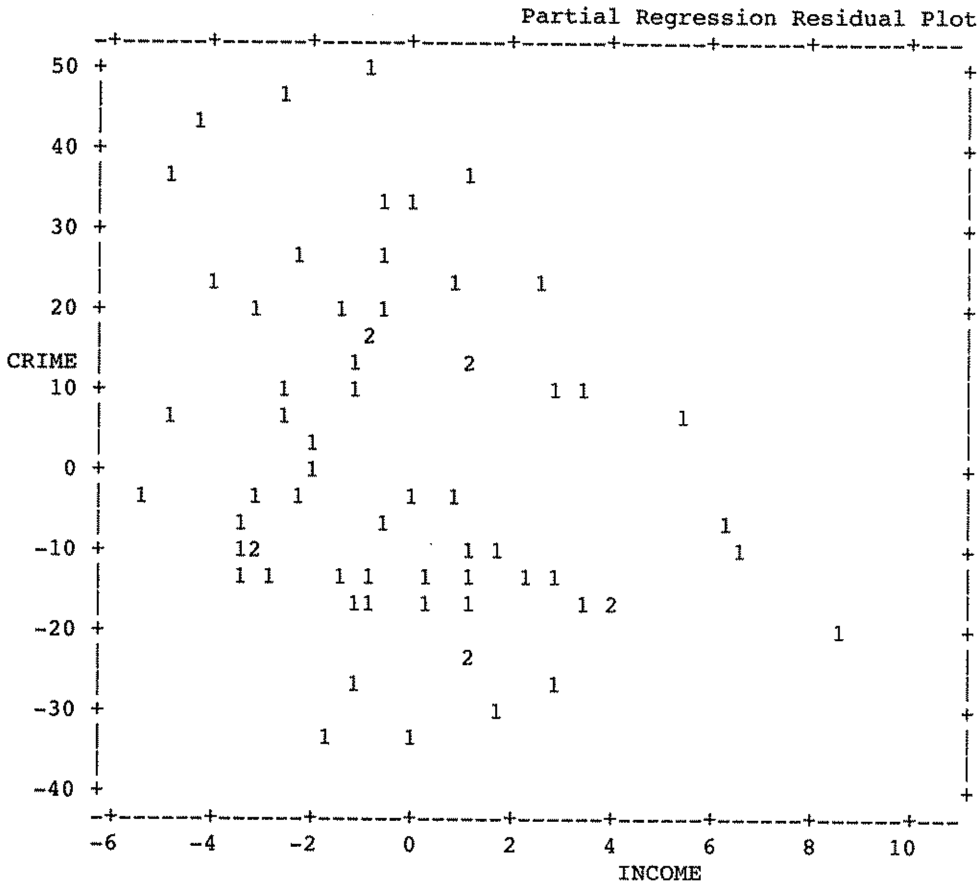


**Figure 11.13**

**c)** Show that when $X_2 = 600$, $E(Y) = 1.4 + .5X_1$. Thus, increasing $X_2$ by 100 shifts the line relating $Y$ to $X_1$ upward by $100\beta_2 = .2$ units.
**d)** Show that setting $X_1$ at a variety of values yields a collection of parallel lines, each having slope .002, relating the mean of $Y$ to $X_2$.
**e)** Since $\beta_1 = .50$ is larger than $\beta_2 = .002$, does this imply that $X_1$ has the greater partial effect on $Y$? Explain.

3. Refer to the data in Table 9.13 in Problem 9.17. Let $Y$ = crude birth rate, $X_1$ = women's economic activity, and $X_2$ = GNP. The least squares equation is $\hat{Y} = 34.53 - .13X_1 - .64X_2$.
**a)** Interpret the estimated regression coefficients.
**b)** Plot on a single graph the relationship between $Y$ and $X_1$ when $X_2 = 0$, $X_2 = 10$, and $X_2 = 20$. Interpret the results.
**c)** The bivariate prediction equation with $X_1$ is $\hat{Y} = 37.65 - .31X_1$. The Pearson correlations are $r_{YX_1} = -.58$, $r_{YX_2} = -.72$, and $r_{X_1X_2} = .58$. Explain why the coefficient of $X_1$

in the bivariate equation is so different from its value in the multiple predictor equation.
**d)** $R^2 = .56$. Interpret.

4. Refer to Example 11.1. Using computer software for the data in Problem 9.24 for those three variables:
**a)** Construct box plots for each variable and scatter diagrams and partial regression plots between $Y$ and each of $X_1$ and $X_2$. If available with your software, also construct coplots. Interpret these plots.
**b)** Find prediction equations for the bivariate models, using education to predict crime rate and using urbanization to predict crime rate. Interpret.
**c)** Construct the correlation matrix for these three variables. Interpret.
**d)** Find the prediction equation for the multiple regression model. Interpret.
**e)** Find $R^2$ for the multiple regression model, and show that it is not much larger than $r^2$ for the model using urbanization alone as the predictor. Interpret.

5. Table 11.11 shows a SAS printout from fitting the multiple regression model to the data from Table 9.1, excluding D.C., on $Y$ = violent crime rate, $X_1$ = poverty rate, and $X_2$ = percent living in metropolitan areas.

**TABLE 11.11**

| Source | DF | Sum of Squares | Mean Square | F Value | Prob>F |
|---|---|---|---|---|---|
| Model | 2 | 2448368.07 | 1224184.04 | 31.249 | 0.0001 |
| Error | 47 | 1841257.15 | 39175.68 | | |
| Total | 49 | 4289625.22 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 197.928 | R-square | 0.5708 |

| Variable | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| | Standard -ized Estimate |
|---|---|---|---|---|---|
| INTERCEP | -498.683 | 140.988 | -3.537 | 0.0009 | 0.000 |
| POVERTY | 32.622 | 6.677 | 4.885 | 0.0001 | 0.473 |
| METRO | 9.112 | 1.321 | 6.900 | 0.0001 | 0.668 |

Pearson Correlation Coefficients

| | VIOLENT | POVERTY | METRO |
|---|---|---|---|
| VIOLENT | 1.00000 | 0.36875 | 0.59396 |
| POVERTY | 0.36875 | 1.00000 | -0.15562 |
| METRO | 0.59396 | -0.15562 | 1.00000 |

**a)** Report the prediction equation, and interpret the estimated regression coefficients.
**b)** Find the predicted violent crime rate for Massachusetts. Find the residual, and interpret.
**c)** Interpret the fit by showing the prediction equation relating $Y$ and $X_1$ for states with (i) $X_2 = 0$, (ii) $X_2 = 50$, (iii) $X_2 = 100$. Interpret.
**d)** Interpret the correlation matrix.
**e)** Report $R^2$ and the multiple correlation, and interpret.
**f)** Find the partial correlation between violent crime rate and poverty, controlling for metropolitan rate. Interpret.
**g)** Refer to (f). Interpret the squared partial correlation.

h) Show how to construct the $F$ statistic for testing $H_0$: $\beta_1 = \beta_2 = 0$, report its $df$ values and $P$-value, and interpret.
i) Show how to construct the $t$ statistic for testing $H_0$: $\beta_1 = 0$, report its $df$ and its $P$-value for $H_a$: $\beta_1 \neq 0$, and interpret.
j) Construct a 95% confidence interval for $\beta_1$, and interpret.

6. Repeat the previous exercise using software to fit the model with murder rate in Table 9.1 as the response variable.

7. Refer to Problem 11.5. With $X_3$ = percentage of single-parent families also in the model, Table 11.12 shows results.
a) Report the prediction equation and interpret the coefficients.
b) Report $R^2$, and interpret.
c) Report the test statistic for $H_0$: $\beta_1 = \beta_2 = \beta_3 = 0$, find its $df$ values and $P$-value, and interpret.
d) Find the $t$ statistic for $H_0$: $\beta_1 = 0$ and its $df$ and $P$-value for $H_a$: $\beta_1 \neq 0$, and interpret.
e) Construct a 95% confidence interval for $\beta_1$, and interpret.
f) Without $X_3$ in the model, the coefficient of $X_1$ is 32.62. Why do you think it changes so much after $X_3$ is added?
g) Since Table 9.1 provides data for all states, what relevance, if any, do the inferences have in this exercise and Problem 11.5?

**TABLE 11.12**

| Variable | Coefficient | Std. Error |
|---|---|---|
| Intercept | −1197.538 | |
| Poverty | 18.283 | (6.136) |
| Metropolitan | 7.712 | (1.109) |
| Single-parent | 89.401 | (17.836) |
| $F$ | 39.9 | |
| $R^2$ | .722 | |
| $n$ | 50 | |

**TABLE 11.13**

| | Sum of Squares |
|---|---|
| Regression | 31.8 |
| Residual | 199.3 |
| Variable | B |
| MEDUC | −.24 |
| FSES | .02 |
| (Constant) | 5.25 |

8. Table 11.13 is part of a SPSS printout for fitting a regression model to the relationship between $Y$ = number of children in family, $X_1$ = mother's educational level (MEDUC) in years, and $X_2$ = father's socioeconomic status (FSES), for a random sample of 49 college students at Texas A&M University.
a) Write the prediction equation. Interpret parameter estimates.
b) For the first subject in the sample, $X_1 = 12$, $X_2 = 61$, and $Y = 5$. Find the predicted value of $Y$ and the residual.
c) Report SSE. Use it to explain the least squares property of this prediction equation.
d) Find the multiple correlation. Interpret.
e) Is it possible that $r_{YX_1 \cdot X_2} = .40$? Explain.
f) Can you tell from this printout whether $r_{YX_1}$ is positive or negative? Explain.

9. Refer to Problem 9.30 on feelings toward liberals, political ideology, and religious attendance. The sample size is small, but for illustrative purposes Table 11.14 shows results of fitting the multiple regression model with feelings toward liberals as the response, using the category numbers as scores for religion. Standard errors are shown in parentheses.
a) Report the prediction equation and interpret the estimates.

**b)** Report the predicted value and residual for the first observation, for which ideology = 7, religion = 9, and feelings = 10.
**c)** Report, and explain how to interpret, $R^2$.
**d)** Tables of this form often put * by an effect having $P < .05$, ** by an effect having $P < .01$, and *** by an effect having $P < .001$. Show how this was determined for the ideology effect, and discuss the disadvantage of summarizing in this manner.
**e)** Explain how the $F$ value was obtained, report its $df$ values, and explain how to interpret its result.
**f)** The estimated standardized regression coefficients are −.79 for ideology and −.23 for religion. Interpret.

**TABLE 11.14**

| Variable | Coefficient |
|---|---|
| Intercept | 135.31 |
| Ideology | −14.07 |
| | (3.16)** |
| Religion | −2.95 |
| | (2.26) |
| F | 13.93** |
| $R^2$ | .799 |
| Adj. $R^2$ | .742 |
| (n) | (10) |

**10.** Refer to Table 11.5. Test the null hypothesis $H_0$: $\beta_2 = 0$ that mental impairment is independent of SES, controlling for life events. Report the test statistic, and report and interpret the $P$-value for (a) $H_a$: $\beta_2 \neq 0$, (b) $H_a$: $\beta_2 < 0$.

**11.** Use software with Table 9.4 to conduct a multiple regression analysis of $Y$ = selling price of home, $X_1$ = size of home, $X_2$ = number of bedrooms, $X_3$ = number of bathrooms.
**a)** Use graphics to display the effects of the predictors. Interpret.
**b)** Show that the prediction equation is $\hat{Y} = -53.4 + 62.4X_1 + 1.64X_2 + 22.9X_3$. Interpret the estimates, and find the predicted selling price for a home with $X_1 = 2$, $X_2 = 4$, and $X_3 = 2$.
**c)** Inspect the correlation matrix, and report the variables having the (i) strongest association, (ii) weakest association.
**d)** Report $R^2$, and interpret.
**e)** Show how to calculate the $F$ statistic for testing the effect of the three predictors, report its $df$ values and its $P$-value, and interpret.
**f)** Show how to calculate the $t$ statistic for testing $H_0 : \beta_2 = 0$, report its $P$-value for $H_a : \beta_2 > 0$, and interpret. Why do you think this effect is not significant?
**g)** Fit the simpler model without size of home as a predictor. Now again test the partial effect of number of bedrooms, and interpret.
**h)** Interpret the fit in (g) by showing the prediction equation relating $Y$ and $X_2$ for homes with (i) one bathroom, (ii) two bathrooms.
**i)** Construct a 95% confidence interval for the coefficient of $X_2$ in the model in (g), and interpret.
**j)** Find the partial correlation between price and number of bedrooms, controlling for number of bathrooms. Compare it to the Pearson correlation, and interpret.

**TABLE 11.15**

| Variable | N | Mean | Std Dev |
|---|---|---|---|
| BIRTHS | 23 | 22.117 | 10.469 |
| ECON | 23 | 47.826 | 19.872 |
| LITER | 23 | 77.696 | 17.665 |

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0

| | BIRTHS | ECON | LITER |
|---|---|---|---|
| BIRTHS | 1.00000<br>0.0 | -0.61181<br>0.0019 | -0.81872<br>0.0001 |
| ECON | -0.61181<br>0.0019 | 1.00000<br>0.0 | 0.42056<br>0.0457 |
| LITER | -0.81872<br>0.0001 | 0.42056<br>0.0457 | 1.00000<br>0.0 |

| Source | DF | Sum of Squares | Mean Square | F Value | Prob>F |
|---|---|---|---|---|---|
| Model | 2 | 1825.969 | 912.985 | 31.191 | 0.0001 |
| Error | 20 | 585.424 | 29.271 | | |
| C Total | 22 | 2411.393 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 5.410 | R-square | 0.7572 |

| Variable | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|
| INTERCEP | 61.713 | 5.2453 | 11.765 | 0.0001 |
| ECON | -0.171 | 0.0640 | -2.676 | 0.0145 |
| LITER | -0.404 | 0.0720 | -5.616 | 0.0001 |

| Variable | Standardized Estimate |
|---|---|
| ECON | -0.325 |
| LITER | -0.682 |

**k)** Show how to calculate the estimated standardized regression coefficients for the model in (g), and interpret.
**l)** Write the prediction equation using standardized variables. Interpret.

**12.** Refer to the previous exercise and the model [in (g)] having numbers of bedrooms and bathrooms as predictors.
**a)** Fit the model allowing an interaction between these two predictors.
**b)** Interpret the fit by showing the prediction equation relating $Y$ and $X_2$ for homes with (i) two bathrooms, (ii) three bathrooms.
**c)** Use a $t$ test to analyze the significance of the interaction term. Interpret.

**d)** Use an $F$ test for complete and reduced models to analyze the significance of the interaction term. Interpret, and compare results to c).

**13.** Refer to Problem 11.5. Add an interaction term.
**a)** Report the prediction equation.
**b)** As the percentage living in metropolitan areas increases, does the effect of poverty rate tend to increase or decrease? Explain.
**c)** Show how to interpret the prediction equation graphically.
**d)** Describe the substantive effect of adding the interaction term by comparing $R^2$ values for the model with and without this term.

**14.** Refer to Table 9.13 and Problem 9.17. Table 11.15 shows a SAS printout from fitting the model $E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2$ to $Y$ = birth rate (BIRTHS), $X_1$ = women's economic activity (ECON), and $X_2$ = literacy rate (LITER). (We deleted observations for Germany, South Africa, and Vietnam.)
**a)** Report the value of each of the following:
(i) $r_{YX_1}$ (ii) $r_{YX_2}$ (iii) $R^2$
(iv) TSS (v) SSE (vi) MSE
(vii) $\hat{\sigma}$ (viii) $s_Y$ (ix) $\hat{\sigma}_{b_1}$
(x) $t$ for $H_0$: $\beta_1 = 0$
(xi) $P$ for $H_0$: $\beta_1 = 0$ against $H_a$: $\beta_1 \neq 0$
(xii) $P$ for $H_0$: $\beta_1 = 0$ against $H_a$: $\beta_1 < 0$
(xiii) $F$ for $H_0$: $\beta_1 = \beta_2 = 0$
(xiv) $P$ for $H_0$: $\beta_1 = \beta_2 = 0$
(xv) $df$ for $\hat{\sigma}$ (xvi) $P$ for $H_0$: $\rho_{YX_1} = 0$.
**b)** Report the prediction equation, and carefully interpret the three estimated regression coefficients.
**c)** Interpret the Pearson correlations $r_{YX_1}$ and $r_{YX_2}$.
**d)** Show how to calculate $R^2$, and interpret its value.
**e)** Calculate the multiple correlation, and interpret.
**f)** Though inference may not be relevant for these data, show how to construct the $F$ statistic for testing $H_0$: $\beta_1 = \beta_2 = 0$, report its $df$ values and $P$-value, and interpret.
**g)** Show how to construct the $t$ statistic for testing $H_0$: $\beta_1 = 0$, report its $df$ and $P$-value for $H_a$: $\beta_1 \neq 0$, and interpret.

**15.** Refer to the previous exercise.
**a)** Find the partial correlation between $Y$ and $X_1$, controlling for $X_2$. Interpret both the partial correlation and its square.
**b)** Show how to calculate the estimate of the conditional standard deviation, and interpret its value.
**c)** Show how to calculate the estimated standardized regression coefficient for $X_1$, and interpret its value.
**d)** Write the prediction equation using standardized variables. Interpret.
**e)** Find the predicted $z$-score for a country that is one standard deviation above the mean on both predictors. Interpret.

**16.** For a random sample of 66 state precincts, data are available on

$$\begin{aligned} Y &= \text{Percentage of adult residents who are registered to vote} \\ X_1 &= \text{Percentage of adult residents owning homes} \\ X_2 &= \text{Percentage of adult residents who are nonwhite} \end{aligned}$$

$X_3$ = Median family income (thousands of dollars)
$X_4$ = Median age of residents
$X_5$ = Percentage of residents who have lived in the precinct at least ten years

Table 11.16 shows a portion of the printout used to analyze the data.

**TABLE 11.16**

| | DF | Sum of Squares | Mean Square | F Value | Prob > F | R-Square |
|---|---|---|---|---|---|---|
| | | | | | | ---- |
| Model | --- | ---- | ---- | ---- | ---- | Root MSE |
| Error | --- | 2940.0 | ---- | | | ---- |
| Total | --- | 3753.3 | | | | |

| Variable | Parameter Estimate | Standard Error | T For H0: Parameter = 0 | Prob > \|T\| |
|---|---|---|---|---|
| Intercept | 70.0000 | | | |
| X1 | 0.1000 | .0450 | ---- | ---- |
| X2 | -0.1500 | .0750 | ---- | ---- |
| X3 | 0.1000 | .2000 | ---- | ---- |
| X4 | -0.0400 | .0500 | ---- | ---- |
| X5 | 0.1200 | .0500 | ---- | ---- |

**a)** Fill in all the missing values in the printout.
**b)** State the prediction equation and interpret the coefficient of $X_1$.
**c)** Do you think it is necessary to include all five explanatory variables in the model? Explain.
**d)** Interpret the "R-Square" value.
**e)** To what test does the "F Value" refer? Interpret the result of that test.
**f)** To what test does the $t$-value opposite X1 refer? Report and interpret the $P$-value.
**g)** Interpret the value listed under "Root MSE."
**h)** Find a 95% confidence interval for the change in the mean of $Y$ for a 1-unit increase in the percentage of adults owning homes, controlling for the other variables. Interpret.
**i)** Find a 95% confidence interval for the change in the mean of $Y$ for a 50-unit increase in the percentage of adults owning homes, controlling for the other variables. Interpret.

**17.** *Refer to the data in Table 9.4 on $Y$ = selling price of home, $X_1$ = size of home, and $X_2$ = whether the house is new (1 = yes, 0 = no). Chapters 12 and 13 show that one can incorporate qualitative predictors such as $X_2$ in regression models, and this exercise provides a preview. Table 11.17 shows part of the printout for the model using these predictors.
**a)** Report the prediction equation. By setting $X_2 = 0$ and then 1, construct the two separate lines for older and for new homes. Note that the model implies that the slope effect of size on selling price is the same for each.
**b)** Since $X_2$ takes only the values 0 and 1, explain why the coefficient of $X_2$ estimates the difference of mean selling prices between new and older homes, controlling for house size.
**c)** Test the significance of the partial effect of whether a house is new. Report the $P$-value, and interpret.

**d)** Construct a 95% confidence interval for the partial effect of whether a house is new, and interpret.

**TABLE 11.17**

| Variable | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|
| INTERCEP | -26.089 | 5.977 | -4.365 | 0.0001 |
| SIZE | 72.575 | 3.508 | 20.690 | 0.0001 |
| NEW | 19.587 | 3.995 | 4.903 | 0.0001 |

**18.** * Refer to the previous exercise. Table 13.16 (Chapter 13) is a SPSS printout showing the effect of adding an interaction cross-product variable (NEWSIZE) to the model.
**a)** Report the prediction equation. Interpret the fit by reporting the prediction equation between selling price and size of house separately for new homes ($X_2 = 1$) and for old homes ($X_2 = 0$). Interpret. (This fit is equivalent to fitting lines separately to the data for new homes and for old homes.)
**b)** Interpret the fit by reporting the difference between the predicted selling prices for new and old homes for houses with $X_1$ equal to (i) 1.5, (ii) 2.0, (iii) 2.5.
**c)** Test the significance of the interaction term. Report the $P$-value, and interpret.
**d)** Figure 13.14 (Chapter 13) shows a plot of these data, identifying the points by whether the home is new (1) or not (0). When the new home with largest price is removed from the data set and the model is re-fitted, Table 13.17 shows results. Again test for interaction, and note the large impact a potential outlier can have.

**19.** A study analyzes relationships among $Y$ = percentage vote for Democratic candidate, $X_1$ = percentage of registered voters who are Democrats, and $X_2$ = percentage of registered voters who vote in the election, for several congressional elections in 1996. The researchers expect interaction, since they expect a higher slope between $Y$ and $X_1$ at larger values of $X_2$ than at smaller values. They obtain the prediction equation $\hat{Y} = 20 + .30X_1 + .05X_2 + .005X_1X_2$. Does this equation support the direction of their prediction? Explain.

**20.** A multiple regression analysis investigates the relationship between $Y$ = college GPA and several explanatory variables, using a random sample of 195 students at Slippery Rock University. First, high school GPA and total SAT score are entered into the model. The sum of squared errors is SSE = 20. Next, parents' education and parents' income are added, to determine if they have an effect, controlling for high school GPA and SAT. For this expanded model SSE = 19. Test whether this complete model is significantly better than the one containing only high school GPA and SAT. Report and interpret the $P$-value.

**21.** Refer to Examples 11.1 and 11.8. Explain why the partial correlation between crime rate and high school graduation rate is so different from the Pearson correlation. (This is an example of *Simpson's paradox*; see Problem 10.8.)

**22.** For a group of 100 children of ages varying from 3 to 15, the Pearson correlation between vocabulary score on an achievement test and height of child is .65. The Pearson correlation between vocabulary score and age for this sample is .85, and the Pearson correlation between height and age is .75.
**a)** Show that the partial correlation between vocabulary and height, controlling for age, is .036. Interpret.
**b)** Test whether this partial correlation is significantly nonzero. Interpret.

**c)** Is it plausible that the relationship between height and vocabulary is spurious, in the sense that it is due to their joint dependence on age? Explain.

**23.** A multiple regression model describes the relationship among a collection of cities between $Y$ = murder rate (number of murders per 100,000 residents) and

$X_1$ = Number of police officers (per 100,000 residents)
$X_2$ = Median length of prison sentence given to convicted murderers (in years)
$X_3$ = Median income of residents of city (in thousands of dollars)
$X_4$ = Unemployment rate in city

These variables are measured in 1996 for a random sample of thirty cities with population size exceeding 35,000. For these cities, the least squares equation is $\hat{Y} = 30 - .02X_1 - .1X_2 - 1.2X_3 + .8X_4$, and $\bar{Y} = 15$, $\bar{X}_1 = 100$, $\bar{X}_2 = 15$, $\bar{X}_3 = 13$, $\bar{X}_4 = 7.8$, $s_Y = 8$, $s_{X_1} = 30$, $s_{X_2} = 10$, $s_{X_3} = 2$, $s_{X_4} = 2$.
**a)** Can you tell from the coefficients of the prediction equation which explanatory variable has the greatest partial effect on $Y$? Explain.
**b)** Find the standardized regression coefficients and interpret their values.
**c)** Write the prediction equation using standardized variables. Find the predicted $z$-score on murder rate for a city that is one standard deviation above the mean on $X_1$, $X_2$, and $X_3$, and one standard deviation below the mean on $X_4$. Interpret.

**24.** Refer to Problem 11.5. Report the estimated standardized regression coefficients, and interpret, and express the prediction equation using standardized variables.

### *Concepts and Applications*

**25.** Refer to the WWW data set (Problem 1.7). Using software, conduct a regression analysis using (i) $Y$ = political ideology and using predictors number of times per week of newspaper reading and religiosity, (ii) $Y$ = college GPA and predictors high school GPA and number of weekly hours of physical exercise. Prepare a report, summarizing your graphical analyses, bivariate models and interpretations, multiple regression models and interpretations, inferences, checks of effects of outliers, and overall summary of the relationships.

**26.** Refer to the data file you created in Problem 1.7. For variables chosen by your instructor, fit a multiple regression model and conduct descriptive and inferential statistical analyses. Interpret and summarize your findings.

**27.** Refer to data for the 50 states in Table 9.1. Using $Y$ = violent crime rate, $X_1$ = percentage of single-parent families, and $X_2$ = metropolitan rate, analyze these data using regression. Provide interpretations for all your analyses, and provide a paragraph summary of your conclusions at the end of your report.

**28.** Repeat the previous exercise using murder rate as the response variable.

**29.** Refer to Problem 11.27. Repeat this problem, adding $X_3$ = percentage white as an explanatory variable.

**30.** Refer to Problem 11.27. Repeat this problem, including the observation for D.C. Describe the effect on the various analyses of this observation.

31. Refer to Problem 9.17 and Table 9.13. Construct a multiple regression model containing two explanatory variables that provide good predictions for birth rate. How did you select this model? (*Hint*: One way is based on entries in the correlation matrix.)

32. For Example 11.2, Table 11.18 shows the result of adding religious attendance as a predictor, measured as the approximate number of times the subject attends a religious service over the course of a year. Write a short report, interpreting the information from this table.

**TABLE 11.18**

| Variable | Coefficient |
|---|---|
| Intercept | 27.422 |
| Life events | .0935 |
| | (.0313)** |
| SES | −.0958 |
| | (.0256)*** |
| Religious attendance | −.0370 |
| | (.0219) |
| $R^2$ | .358 |
| $(n)$ | (40) |

33. Refer to the variables in Problem 9.24. By constructing coplots, do you find any pairs of predictors that show evidence of interaction?

34. Describe a situation in which you would expect interaction. Describe the likely nature of the interaction; for example, as $X_2$ increases, would the slope of the relationship between $Y$ and $X_1$ tend to increase, or decrease?

35. For a linear model with two explanatory variables $X_1$ and $X_2$, which of the following must be incorrect? Why?
**a)** $r_{YX_1} = .01$, $r_{YX_2} = -.2$, $R = .75$
**b)** $r_{YX_1} = .01$, $r_{YX_2} = -.75$, $R = .2$
**c)** $r_{YX_1} = .4$, $r_{YX_2} = .4$, $R = .4$

36. Table 11.19 shows results of fitting various regression models to data on $Y$ = college GPA, $X_1$ = high school GPA, $X_2$ = mathematics entrance exam score, and $X_3$ = verbal entrance exam score.
Indicate which of the following statements are false. Give a reason for your answer.

**TABLE 11.19**

| | Model | | |
|---|---|---|---|
| Estimates | $E(Y) = \alpha + \beta X_1$ | $E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2$ | $E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ |
| Coefficient of $X_1$ | .450 | .400 | .340 |
| Coefficient of $X_2$ | | .003 | .002 |
| Coefficient of $X_3$ | | | .002 |
| $R^2$ | .25 | .34 | .38 |

**a)** The correlation between $Y$ and $X_1$ is positive.
**b)** A one-unit increase in $X_1$ corresponds to a change of .45 in the estimated mean of $Y$, controlling for $X_2$ and $X_3$.
**c)** The value of SSE increases as we add additional variables to the model.

**d)** It follows from the sizes of the estimates for the third model that $X_1$ has the strongest partial effect on $Y$.
**e)** The value of $r^2_{YX_3}$ is .40.
**f)** The partial correlation $r_{YX_1 \cdot X_2}$ is positive.
**g)** The partial correlation $r_{YX_1 \cdot X_3}$ could be negative.
**h)** Controlling for $X_1$, a 100-unit increase in $X_2$ corresponds to a predicted increase of .3 in college GPA.
**i)** For the first model, the estimated standardized regression coefficient equals .50.

**37.** In regression analysis, which of the following statements must be false? Why?
**a)** For the model $E(Y) = \alpha + \beta_1 X_1$, $Y$ is significantly related to $X_1$ at the .05 level, but when $X_2$ is added to the model, $Y$ is not significantly related to $X_1$ at the .05 level.
**b)** The estimated coefficient of $X_1$ is positive in the bivariate model, but negative in the multiple regression model.
**c)** When the model is refitted after $Y$ is multiplied by 10, $R^2$, $r_{YX_1}$, $r_{YX_1 \cdot X_2}$, $b_1^*$, the $F$ statistics and $t$ statistics do not change.
**d)** $r_{YX_2 \cdot X_1}$ cannot exceed $r_{YX_2}$.
**e)** The $F$ statistic for testing that all the regression coefficients equal 0 has $P < .05$, but none of the individual $t$ tests have $P < .05$.
**f)** If you compute the standardized regression coefficient for a bivariate model, you always get the Pearson correlation.
**g)** $r^2_{YX_1} = r^2_{YX_2} = .6$ and $R^2 = .6$.
**h)** $r^2_{YX_1} = r^2_{YX_2} = .6$ and $R^2 = 1.2$.
**i)** The Pearson correlation between $Y$ and $\hat{Y}$ equals $-.10$.
**j)** If $X_3$ is added to a model already containing $X_1$ and $X_2$, then if the prediction equation has $b_3 = 0$, $R^2$ stays the same.
**k)** For every $F$ test, there is an equivalent test using the $t$ distribution.

For Problems 11.38–11.42, select the correct answer(s) and indicate why the other responses are inappropriate.

**38.** If $\hat{Y} = 2 + 3X_1 + 5X_2 - 8X_3$, then controlling for $X_2$ and $X_3$, the predicted mean change in $Y$ when $X_1$ is increased from 10 to 20 equals
**a)** 3 **b)** 30 **c)** .3 **d)** Cannot be given—depends on specific values of $X_2$ and $X_3$.

**39.** If $\hat{Y} = 2 + 3X_1 + 5X_2 - 8X_3$,
**a)** The strongest Pearson correlation is between $Y$ and $X_3$.
**b)** The variable with the strongest partial influence on $Y$ is $X_2$.
**c)** The variable with the strongest partial influence on $Y$ is $X_3$, but one cannot tell from this equation which pair has the strongest Pearson correlation.
**d)** None of the above.

**40.** If $\hat{Y} = 2 + 3X_1 + 5X_2 - 8X_3$,
**a)** $r_{YX_3} < 0$
**b)** $r_{YX_3 \cdot X_1} < 0$
**c)** $r_{YX_3 \cdot X_1, X_2} < 0$
**d)** Insufficient information to answer.
**e)** Answers (a), (b), and (c) are all correct.

**41.** If $\hat{Y} = 2 + 3X_1 + 5X_2 - 8X_3$, and $H_0$: $\beta_3 = 0$ is rejected at the .05 level, then
**a)** $H_0$: $\rho_{YX_3 \cdot X_1, X_2} = 0$ is rejected at the .05 level.