1. $\hat{y} = 100 + 0.2x_1 + 4x_2$
2. $\hat{y} = 95 + 0.15x_1 + 3x_2 + 1x_1x_2$

Both models have been built over the range $20 \le x_1 \le 50$ (°C) and $0.5 \le x_2 \le 10$ (hours).

a. Using both models, what is the predicted value of conversion when $x_2 = 2$ in terms of x_1? Repeat this calculation for $x_2 = 8$. Draw a graph of the predicted values as a function of temperature for both conversion models. Comment on the effect of the interaction term in model 2.

b. Find the expected change in the mean conversion for a unit change in temperature x_1 for model 1 when $x_2 = 5$. Does this quantity depend on the specific value of reaction time selected? Why?

c. Find the expected change in the mean conversion for a unit change in temperature x_1 for model 2 when $x_2 = 5$. Repeat this calculation for $x_2 = 2$ and $x_2 = 8$. Does the result depend on the value selected for x_2? Why?

3.22 Show that an equivalent way to perform the test for significance of regression in multiple linear regression is to base the test on R^2 as follows: To test H_0: $\beta_1 = \beta_2 = \ldots = \beta_k$ versus H_1: at least one $\beta_j \neq 0$, calculate

$$F_0 = \frac{R^2(n-p)}{k(1-R^2)}$$

and to reject H_0 if the computed value of F_0 exceeds $F_{\alpha,k,n-p}$, where $p = k + 1$.

3.23 Suppose that a linear regression model with $k = 2$ regressors has been fit to $n = 25$ observations and $R^2 = 0.90$.

a. Test for significance of regression at $\alpha = 0.05$. Use the results of the previous problem.

b. What is the smallest value of R^2 that would lead to the conclusion of a significant regression if $\alpha = 0.05$? Are you surprised at how small this value of R^2 is?

3.24 Show that an alternate computing formula for the regression sum of squares in a linear regression model is

$$SS_R = \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2$$

3.25 Consider the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

Using the procedure for testing a general linear hypothesis, show how to test

a. H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta$

b. H_0: $\beta_1 = \beta_2, \beta_3 = \beta_4$

c. H_0 : $\beta_1 - 2\beta_2 = 4\beta_3$
$\beta_1 + 2\beta_2 = 0$

3.26 Suppose that we ha

y	x
y_1	x
y_2	x
⋮	⋮
y_{n_1}	x

Two models can be

y_i

a. Show how these

b. Using the result used to test the

c. Using the result used to test the

d. Using the result used to test tha

3.27 Show that $\mathrm{Var}(\hat{\mathbf{y}})$

3.28 Prove that the m $(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) =$

3.29 For the simple li matrix are

$$h_{ij} = \frac{1}{n}$$

Discuss the beha

3.30 Consider the mu least-squares esti

3.31 Show that the re $\mathbf{e} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$. [H

3.32 For the multiple

3.33 Prove that R^2 is

3.34 **Constrained leas** tor of $\boldsymbol{\beta}$ in the m say $\mathbf{T}\boldsymbol{\beta} = \mathbf{c}$. Sho

$$\tilde{\beta} = \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{T}'\left[\mathbf{T}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{T}\right]^{-1}\left(\mathbf{c} - \mathbf{T}\hat{\beta}\right)$$

where $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Discuss situations in which this constrained estimator might be appropriate. Find the residual sum of squares for the constrained estimator. Is it larger or smaller than the residual sum of squares in the unconstrained case?

3.35 Let $\mathbf{x}_j$ be the jth row of $\mathbf{X}$, and $\mathbf{X}_{-j}$ be the $\mathbf{X}$ matrix with the jth row removed. Show that

$$\text{Var}\left[\hat{\beta}_j\right] = \sigma^2\left[\mathbf{x}_j'\mathbf{x}_j - \mathbf{x}_j'\mathbf{X}_{-j}\left(\mathbf{X}_{-j}'\mathbf{X}_{-j}\right)^{-1}\mathbf{X}_{-j}'\mathbf{x}_j\right]$$

3.36 Consider the following two models where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$:

Model A: $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$

Model B: $\mathbf{y} = \mathbf{X}_1'\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$

Show that $R_A^2 \le R_B^2$.

3.37 Suppose we fit the model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$ when the true model is actually given by $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_2 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$. For both models, assume $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. Find the expected value and variance of the ordinary least-squares estimate, $\hat{\beta}_1$. Under what conditions is this estimate unbiased?

3.38 Consider a correctly specified regression model with p terms, including the intercept. Make the usual assumptions about $\boldsymbol{\varepsilon}$. Prove that

$$\sum_{i=1}^{n}\text{Var}\left(\hat{y}_i\right) = p\sigma^2$$

3.39 Let R_j^2 be the coefficient of determination when we regress the jth regressor on the other $k-1$ regressors. Show that the jth variance inflation factor may be expressed as

$$\frac{1}{1-R_j^2}$$

3.40 Consider the hypotheses for the general linear model, which are of the form

$$H_0: \mathbf{T}\boldsymbol{\beta} = \mathbf{c}, \quad H_1: \mathbf{T}\boldsymbol{\beta} \neq \mathbf{c}$$

where $\mathbf{T}$ is a $q \times p$ matrix of rank q. Derive the appropriate F statistic under both the null and alternative hypothesis.

3.41 Consider the 2016 major league baseball data in Table B.22. While team ERA was useful in predicting the number of games that a team wins, there are some other measures of team performance, including the number of strikeouts, the number of errors committed, and the number of runs allowed per game. Fit

a multiple lin
and runs allo
a. Test for sig
b. Is there an
number of
c. What actio

3.42 Table B.24 co
for 51 US cit
price as the r
tion as the pr
a. Test for sig
b. Find a 95%
c. Does this
in home re

3.43 You have fit
observations.
squares is 12
a. 15.1
b. 10
c. 13
d. 9.85
e. None of t

3.44 Consider the
R^2 is
a. 0.80
b. 0.75
c. 0.78
d. 0.65
e. None of t

3.45 Consider the
constructed
a. Using this
ERA of 3
b. Find a 95
a of this p
c. Use the
number o
d. Find a 95
c of this p
e. Compare
second p

Note that now $\hat{\beta}_1 = -1.222$, and a sign reversal has occurred. The reason is that $\hat{\beta}_1 = -1.222$ in the multiple regression model is a partial regression coefficient; it measures the effect of x_1 **given** that x_2 is also in the model.

The data from this example are plotted in Figure 3.16. The reason for the difference in sign between the partial and total regression coefficients is obvious from inspection of this figure. If we ignore the x_2 values, the apparent relationship between y and x_1 has a positive slope. However, if we consider the relationship between y and x_1 for constant values of x_2, we note that this relationship really has a negative slope. Thus, a wrong sign in a regression model may indicate that important regressors are missing. If the analyst can identify these regressors and include them in the model, then the wrong signs may disappear.

Multicollinearity can cause wrong signs for regression coefficients. In effect, severe multicollinearity inflates the variances of the regression coefficients, and this increases the probability that one or more regression coefficients will have the wrong sign. Methods for diagnosing and dealing with multicollinearity are summarized in Chapter 9.

Computational error is also a source of wrong signs in regression models. Different computer programs handle round-off or truncation problems in different ways, and some programs are more effective than others in this regard. Severe multicollinearity causes the $\mathbf{X}'\mathbf{X}$ matrix to be ill-conditioned, which is also a source of computational error. Computational error can cause not only sign reversals but regression coefficients to differ by several orders of magnitude. The accuracy of the computer code should be investigated when wrong-sign problems are suspected.

PROBLEMS

3.1 Consider the National Football League data in Table B.1.

a. Fit a multiple linear regression model relating the number of games won to the team's passing yardage (x_2), the percentage of rushing plays (x_7), and the opponents' yards rushing (x_8).

b. Construct the analysis-of-variance table and test for significance of regression.

c. Calculate t statistics for testing the hypotheses H_0: $\beta_2 = 0$, H_0: $\beta_7 = 0$, and H_0: $\beta_8 = 0$. What conclusions can you draw about the roles the variables x_2, x_7, and x_8 play in the model?

d. Calculate R^2 and R^2_{Adj} for this model.

e. Using the partial F test, determine the contribution of x_7 to the model. How is this partial F statistic related to the t test for β_7 calculated in part c above?

3.2 Using the results of Problem 3.1, show numerically that the square of the simple correlation coefficient between the observed values y_i and the fitted values $\hat{y}_i$ equals R^2.

3.3 Refer to Problem 3.1.

a. Find a 95% CI on β_7.

b. Find a 95% CI on the mean number of games won by a team when $x_2 = 2300$, $x_7 = 56.0$, and $x_8 = 2100$.

3.4 Reconsider the National Football League data from Problem 3.1. Fit a model to these data using only x_7 and x_8 as the regressors.

a. Test for significance of regression.

b. Calculate R^2 and R^2_{Adj}. How do these quantities compare to the values computed for the model in Problem 3.1, which included an additional regressor (x_2)?

c. Calculate a 95% CI on β_7. Also find a 95% CI on the mean number of games won by a team when $x_7 = 56.0$ and $x_8 = 2100$. Compare the lengths of these CIs to the lengths of the corresponding CIs from Problem 3.3.

d. What conclusions can you draw from this problem about the consequences of omitting an important regressor from a model?

3.5 Consider the gasoline mileage data in Table B.3.

a. Fit a multiple linear regression model relatmg gasoline mileage y (miles per gallon) to engine displacement x_1 and the number of carburetor barrels x_6.

b. Construct the analysis-of-variance table and test for significance of regression.

c. Calculate R^2 and R^2_{Adj} for this model. Compare this to the R^2 and the R^2_{Adj} for the simple linear regression model relating mileage to engine displacement in Problem 2.4.

d. Find a 95% CI for β_1.

e. Compute the t statistics for testing H_0: $\beta_1 = 0$ and H_0: $\beta_6 = 0$. What conclusions can you draw?

f. Find a 95% CI on the mean gasoline mileage when $x_1 = 275$ in.^{3} and $x_6 = 2$ barrels.

g. Find a 95% prediction interval for a new observation on gasoline mileage when $x_1 = 275$ in.^{3} and $x_6 = 2$ barrels.

3.6 In Problem 2.4 you were asked to compute a 95% CI on mean gasoline prediction interval on mileage when the engine displacement $x_1 = 275$ in.^{3} Compare the lengths of these intervals to the lengths of the confidence and prediction intervals from Problem 3.5 above. Does this tell you anything about the benefits of adding x_6 to the model?

3.7 Consider the house price data in Table B.4.

a. Fit a multiple regression model relating selling price to all nine regressors.

b. Test for significance of regression. What conclusions can you draw?

c. Use t tests to assess the contribution of each regressor to the model. Discuss your findings.

d. What is the contribution of lot size and living space to the model given that all of the other regressors are included?

e. Is multicollinearity a potential problem in this model?

3.8 The data in Table B.5 present the performance of a chemical process as a function of several controllable process variables.

a. Fit a multiple regression model relating CO_2 product (y) to total solvent (x_6) and hydrogen consumption (x_7).