wasted when estimating unnecessary parameters. Therefore, it would be best to include only important variables in the model or variables that are clearly necessary. A solution to avoid over-fitting a model is to utilize a model-building procedure such as **stepwise regression**. Stepwise regression involves selecting independent variables using an automated procedure, which is beyond the scope of this text.

Some other issues with fitting multiple regression models are **extrapolation** (which was discussed in Chapter 13) and **correlated errors**. Extrapolation can be a concern when the regression model is used to predict values outside the range of the data used to estimate the model. Be sure to only use the model within an appropriate range of x-values. The problem with correlated errors arises when measurements of the dependent variable are correlated. That is, since the observations (the responses) of the regression model are assumed to be independent, it is problematic if there is a relationship between responses. One will often see this type of dependency with time series data. Current measurements are often dependent on measurements in the previous time period.

14.6 Exercises

Basic Concepts

1. Define multicollinearity. How can you detect if multicollinearity exists in a regression model?
2. Why is multicollinearity a concern when performing regression analysis?
3. How can you attempt to correct the problem of multicollinearity?
4. What is parameter estimability?
5. How can you alleviate concerns about parameter estimability?
6. Why is variable selection difficult when building a multiple regression model?
7. Does the R^2 value always increase as additional variables are added? Does this mean that adding additional variables always produces a more useful model? Explain.
8. What is extrapolation? Why is this a concern?
9. With what type of data do you often encounter issues with correlated errors?

Exercises

10. In Exercise 8 of Section 14.4, we modeled the relationship between total points and rushing yards, passing yards, and first downs.

    a. Using the correlation matrix below, discuss whether collinearity might play a role in estimating total points using rushing yards, passing yards, and first downs in the model.

Correlation Matrix			
	Rushing Yards	**Passing Yards**	**First Downs**
Rushing Yards	1.0000	–0.1943	0.3789
Passing Yards		1.0000	0.5744
First Downs			1.0000

    b. How would you determine if there is a relationship (and if so, the strength of such relationship) between the independent variables in the model?

    c. Given the multiple regression model that was fit in Exercise 8, what would the total points be if a team had 30 rushing yards, 100 passing yards, and 5 first downs?

    d. Should you have any concerns about the estimate in part **c.**? Explain your answer.

11. In Exercise 10 of Section 14.5, salary was modeled as a function of age, experience, and gender.

a. Discuss how collinearity might play a role in estimating salary using age, experience, and gender in the model.

b. How would you determine if there is a relationship (and if so, the strength of such relationship) between the independent variables in the model?

c. Given the multiple regression model that was fit in the problem, what would the expected salary be for a 60-year-old male employee with 25 years of experience?

d. Should you have any concerns about the estimate in part **c.**? Explain your answer.

12. In Exercise 11 of Section 14.5, we attempted to predict the number of crimes on a college/university campus based on the number of police, the enrollment at the university, and if it was a private institution.

a. Examine the correlation matrix below and discuss whether collinearity might play a role in estimating the number of crimes based on the number of police, enrollment, and if the institution is private.

Correlation Matrix			
	Police	**Enrollment**	**Private**
Police	1.0000	0.8042	–0.4942
Enrollment		1.0000	–0.8795
Private			1.0000

b. How would you determine if there is a relationship (and if so, the strength of such relationship) between the independent variables in the model?

c. Given the multiple regression model that was fit in the problem, what would the expected number of crimes be for a private university with a police force of 100 officers and an enrollment of 50,000?

d. Should you have any concerns about the estimate in part **c.**? Explain your answer.

13. Suppose you fit a multiple regression model of the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i$$

The correlation matrix for the pairs of independent variables is given in the following table. Discuss if you detect multicollinearity between any of the variables.

Correlation Matrix				
	x_1	x_2	x_3	x_4
x_1	1.00	0.18	0.86	0.45
x_2		1.00	0.35	0.22
x_3			1.00	0.50
x_4				1.00