3.3 (Extra sum of squares test in terms of R^2) Let R_p^2 and R_q^2 denote the R^2's for the full model with p predictors and a partial model with and $q < p$ predictors. Show that the extra SS F-statistic equals

$$F = \frac{(R_p^2 - R_q^2)/(p-q)}{(1-R_p^2)/[n-(p+1)]}.$$

Suppose that $n = 26, q = 3$ and $p = 5$. Further suppose that $R_p^2 = 0.90$ and $R_q^2 = 0.80$. Test whether the increase in R^2 from the partial model to the full model is statistically significant at the 1% level.

3.10 (Matrix calculation by hand) This is a hand-calculation exercise to help you get an understanding of the matrix calculations in multiple regression. Consider the following small data set. We want to fit a straight line $y = \beta_0 + \beta_1 x$ to these data.

x	1	2	3	4	5
y	2	6	7	9	10

a) Write the $\boldsymbol{X}$ matrix and the $\boldsymbol{y}$ vector.

b) Calculate $\boldsymbol{X}'\boldsymbol{X}$ and its inverse. For a 2×2 matrix, the formula for the inverse is simple:

$$\begin{bmatrix} a & c \\ d & b \end{bmatrix}^{-1} = \frac{1}{ab-cd} \begin{bmatrix} b & -c \\ -d & a \end{bmatrix}.$$

Check that the product of the original matrix and its inverse equals the identity matrix.

c) Calculate the $\boldsymbol{X}'\boldsymbol{y}$ vector.

d) Finally calculate the LS estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ using the formula (3.7).

3.11 (Alternate coding of categorical variables) Refer to Example 3.15 and the data in Table 3.8. Suppose that the Gender is coded as $x_1 = -1$ for females and $x_1 = +1$ for males. Similarly, Race is coded as $x_2 = -1$ for non-Whites and $x_2 = +1$ for Whites. What are the new values of $\beta_0, \beta_1, \beta_2$ and β_3? Interpret them.

3.12 (Cobb-Douglas production function) Data on 569 European companies on their capital (x_1) measured as total fixed assets (in millions of euros) at the end of 1995, labor (x_2) measured as number of workers and output (y) measured as value added (in millions of euros) are available in file `cobbdouglas.csv`. The companies in this data set are from different industry sectors in which different Cobb-Douglas production functions may apply since their capital and labor requirements are different, but we ignore this problem.

a) Fit the Cobb-Douglas production function $y = \beta_0 x_1^{\beta_1} x_2^{\beta_2}$, where β_1 and β_2 are the capital and labor elasticities.

b) If $\beta_1 + \beta_2 = 1$ then it is easy to check that if capital and labor are changed by a common scaling factor then the output is changed by the same factor. In economics this is called the **constant returns to scale**. Test the null hypothesis of the constant returns to scale for these data by doing a t-test of $H_0 : \beta_1 + \beta_2 = 1$ using the estimates of $\text{Var}(\widehat{\beta}_1)$, $\text{Var}(\widehat{\beta}_2)$ and $\text{Cov}(\widehat{\beta}_1, \widehat{\beta}_2)$. (These estimates can be obtained in R by using the `vcov` function.)

c) The above null hypothesis can also be tested by using the extra SS method as follows. Since the response variable must be the same for both the full model and the partial model in the extra SS method so that we can validly compare the SSE's for the two models, subtract $\ln x_2$ from both sides of the model so that the new response variable is $\ln y - \ln x_2$ and the full model is $\ln y - \ln x_2 = \ln \beta_0 + \beta_1(\ln x_1 - \ln x_2) + \beta_3 \ln x_2 + \varepsilon$ where $\beta_3 = \beta_1 + \beta_2 - 1$. Test $H_0 : \beta_3 = 0$ using the extra SS F-test and compare the result obtained in Part (b).

3.15 (Salary data) File `salaries.csv` contains data on annual salaries of 46 employees of a company and possible predictors. The variables are defined in Table 3.10. Use $\log_{10}$(Salary) as the response variable.

a) Fit a prediction model using the given data. Check that the fitted equation using Male and Purchase as reference categories for Gender and Dept categorical variables is

$$\widehat{\log_{10}(\text{Salary})} = 4.429 + 0.0075\ \text{YrsEm} + 0.0017\ \text{PriorYr} + 0.0170\ \text{Educ} + 0.0004\ \text{Super} + 0.0231\ \text{Female} - 0.0388\ \text{Advert} - 0.00573\ \text{Engg} - 0.0938\ \text{Sales}.$$

b) If we use Female and Sales as reference categories, what will be the new coefficients for Male and for the other three departments?

c) The coefficient of Engg is highly nonsignificant with a P-value = 0.774 in the above regression. But if Sales is used as the reference category, the coefficient of Engg is highly significant with a P-value < 0.001. Interpret this result. If the coefficient of a dummy variable is nonsignificant, what does it tell you?

d) In the above regression, the coefficients of PriorYr and Super are nonsignificant with P-values of 0.395 and 0.631, respectively. The coefficient of Female is also nonsignificant with a P-value = 0.115 indicating nonsignificant gender difference. Drop these variables, refit the model and draw conclusions.