1. Faraway (1st edition) problem 4.2

2. We will work with the `sat` data set from Faraway. Each observation is one state, with $n = 50$ total. We will consider regressing `total` (average total SAT score) on `expend` (expenditure, i.e. public school funding per student), `ratio` (student-to-teacher ratio in public schools), `salary` (teacher salary), and `takers` (what proportion of eligible students take the SAT). (We will also include an intercept in every model.)

   (a) Compare the coefficient on the `expend` covariate in the full model, against the coefficient if you regress `total` on `expend` only. Discuss what you see and explain intuitively what is happening in terms of the meaning of the variables.

   (b) Perform an F test, to test the full model against the model that regresses `total` on `takers` only. Do not use any R functions aside from `lm` and `summary(lm(...))`, and show all your calculations in R.

   (c) In the full model, draw the confidence region (i.e. the ellipse) for the coefficients $(\beta_{\texttt{salary}}, \beta_{\texttt{expend}})$. (You can use the `ellipse` library, see Faraway section 3.4 for an example—you can use the default confidence level of 95%). Explain the resulting ellipse shape that you see (hint: look at the correlations between the predictors).

3. In this problem we will do a simulation to examine how errors in our assumptions affect various calculations in regression.

   (a) Generate data for a simple linear regression as follows: use sample size $n = 100$, generate the covariate values from a Uniform$[0, 1]$ distribution and generate response values as

   $$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \text{ where } \epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$$

   with $\beta_0 = \beta_1 = 1$. Run this simulation 1000 times, each time storing (1) the estimate $\hat{\beta}_1$ and (2) the estimate of its (square root) variance, $\widehat{\text{SE}}(\hat{\beta}_1)$. (Hint: use `summary(model)$coef`, this matrix will have entries corresponding to $\hat{\beta}_1$ and its SE that you can easily extract for each run of your simulation). Compare (i) the empirical mean of $\hat{\beta}_1$ versus the target value β_1, and (ii) the empirical standard deviation of $\hat{\beta}_1$, versus the median value of $\widehat{\text{SE}}(\hat{\beta}_1)$. Is $\hat{\beta}_1$ unbiased? Does the estimated SE of $\hat{\beta}_1$ match the observed variation? Explain what you see.

   (b) Now repeat with a different data generating mechanism,

   $$Y_i = \beta_0 + \beta_1 X_i + (X_i)^4 \cdot \epsilon_i \text{ where } \epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$$

   which has a linear mean but heteroskedastic variance. Is $\hat{\beta}_1$ unbiased? Does the estimated SE of $\hat{\beta}_1$ match the observed variation? Explain what you see.

   (c) Next, using the same heteroskedastic-variance model, run your simulation again. For each run, use the fitted model construct a prediction interval at $x = 0.1$ at level $1 - \alpha = 0.9$. Next generate a Y value at this X value, drawn from the same model, and record whether or not it lands inside the prediction interval. What proportion of your trials succeed, i.e. what proportion of the time does the prediction interval contain the new Y value? Then repeat at $x = 0.9$. What proportion of the time does the prediction interval contain the new Y value? Explain what you see.

4. Suppose that we have a response $Y = (Y_1, \ldots, Y_n)$ and a single covariate $X = (X_1, \ldots, X_n)$. Our data set of size n is a mixture of data from two populations, labeled 0 and 1 arbitrarily, e.g. data from public schools and private schools. Let P_i be 0 or 1 to indicate which population the ith data point came from. Let n_0 and n_1 be the number of data points from each population, with $n_0 + n_1 = n$.

   If we're not sure whether the association between X & Y is the same within the two groups, we might do one of the following:

   - Option 1: Split the data into two parts—one data set of size n_0 containing all the data points from population 0, and the other of size n_1 containing all data points from population 1. We could then run two linear regressions

   $$\text{Data from population 0: } Y_i = \beta_0^{(0)} + \beta_1^{(0)} X_i + \text{noise}$$

and

$$\text{Data from population 1: } Y_i = \beta_0^{(1)} + \beta_1^{(1)} X_i + \text{noise}.$$

- Option 2: Run a single linear regression to fit the model:

$$\text{Combined data from both populations: } Y_i = \beta_0 + \beta_1 X_i + \beta_2 P_i + \beta_3 (X_i \cdot P_i) + \text{noise}.$$

(a) Are these two options the same or different in terms of what we're assuming about the mean of the response within this combined data set? Explain by comparing the coefficients $\beta_0^{(0)}, \beta_1^{(0)}, \beta_0^{(1)}, \beta_1^{(1)}$ from Option 1 with the coefficients $\beta_0, \beta_1, \beta_2, \beta_3$ in Option 2.

(b) Are these two options the same or different in terms of what we're assuming about the variance of the response within this combined data set? Explain.