QUEEN MARY UNIVERSITY OF LONDON

MTH5120 Statistical Modelling I

Exercise Sheet 8

1. **Coursework component**
   Based on the Boston dataset available on the library MASS, relative to Housing Values in Suburbs of Boston. The variables of interest are:

   - Y equal to $medv$ is median value of owner-occupied homes in $\$1000$.
   - X_1 equal to $lstat$ is the lower status of the population (percent)
   - X_2 equal to rm is the average number of rooms per dwelling
   - X_3 equal to age is the proportion of owner-occupied units built prior to 1940

   For Model 1: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$, where $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$:

   (a) test the hypothesis regarding the overall regression by using the F-test
   (b) test the hypothesis regarding the parameters β_j for $j = 0, 1, 2, 3$ by using the t-test

   For Model 2: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, where $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$:

   (c) test the hypothesis regarding overall regression and the parameters
   (d) Which is the best model?

2. When fitting the model
   $$E[Y_i] = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}$$
   to a set of $n = 25$ observations, the following results were obtained using the general linear model notation:
   $$\boldsymbol{X}^t\boldsymbol{X} = \begin{pmatrix} 25 & 219 & 10232 \\ 219 & 3055 & 133899 \\ 10232 & 133899 & 6725688 \end{pmatrix}, \qquad \boldsymbol{X}^t\boldsymbol{Y} = \begin{pmatrix} 559.60 \\ 7375.44 \\ 337071.69 \end{pmatrix}$$
   $$\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1} = \begin{pmatrix} 0.11321519 & -0.00444859 & -0.000083673 \\ -0.00444859 & 0.00274378 & -0.000047857 \\ -0.00008367 & -0.00004786 & 0.000001229 \end{pmatrix}$$
   Also $\boldsymbol{Y}^t\boldsymbol{Y} = 18310.63$ and $\bar{Y} = 22.384$.

   (a) Find the least squares estimated $\widehat{\boldsymbol{\beta}}$ and hence write down the fitted model;
   (b) Use the results to construct the Analysis of Variance Table (Remember that the regression sum of squares is $\widehat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{Y} - n\bar{y}^2$)

3. Based on the previous results:

   (a) Test the null hypothesis that the overall regression is non-significant using a significance level of 5%.
   (b) Find a 95% confidence interval for β_j with $j = 0, 1, 2$.