QUEEN MARY UNIVERSITY OF LONDON

MTH5120 Statistical Modelling I

Exercise Sheet 9

1. Based on the Hitters dataset available on the library ISLR, relative to Major League Baseball Data from the 1986 and 1987 seasons. We wish to predict a Baseball player's Salary on the basis of various statistics associated with performance in the previous year. Before working with the data, we need to clean them up, by deleting the missing values for some players:

```
>Hitters =na.omit(Hitters)
```

   (a) Use the `regsubsets` function for running the best model with the maximum number of predictors available, then state: which is the best model according to the adjusted R^2 and the Mallow's statistic? Show it graphically too.

   (b) show the coefficients for the two best models and show if they are statistically significant

   (c) after looking in details at the results, do you confirm the best model found in (a)?

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) Compute the R^2 and $adj(R^2)$.

   (b) In the same way, run a two dimensional model:

$$E[Y_i] = \beta + \beta_1 x_{1,i}$$

   to the same set of 25 observations and we have the following results:

$$\boldsymbol{X}^t\boldsymbol{X} = \begin{pmatrix} 25 & 219 \\ 219 & 3055 \end{pmatrix}, \qquad \boldsymbol{X}^t\boldsymbol{Y} = \begin{pmatrix} 559.60 \\ 7375.44 \end{pmatrix}$$

$$\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1} = \begin{pmatrix} 0.107517421 & -0.007707468 \\ -0.007707468 & 0.000879848 \end{pmatrix}$$

   Compute the R^2 and the adjusted R^2

(c) Which is the best model across the two models, the one with two explanatory variables or the one with one explanatory variable.