QUEEN MARY UNIVERSITY OF LONDON

MTH5120 Statistical Modelling I

Solution to 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) In order to identify the best model that contains a given number of predictor, we use the command `regsubsets` to perform the best subset selection.

```
> library(leaps)
> regfitAll.full <- regsubsets(Salary~.,Hitters,nvmax =19)
> regfitAll.full.summary <- summary(regfitAll.full)
> regfitAll.full.summary$outmat
          AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI
1  ( 1 )  " "   " "  " "   " "  " " " "   " "   " "    " "   " "    " "   "*" 
2  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*" 
3  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*" 
4  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*" 
5  ( 1 )  "*"   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*" 
6  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    " "   "*" 
7  ( 1 )  " "   "*"  " "   " "  " " "*"   " "   "*"    "*"   "*"    " "   " " 
8  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "    " "   "*"    "*"   " " 
9  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"   "*" 
10  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"   "*" 
11  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"   "*" 
12  ( 1 ) "*"   "*"  " "   "*"  " " "*"   " "   "*"    " "   " "    "*"   "*" 
13  ( 1 ) "*"   "*"  " "   "*"  " " "*"   " "   "*"    " "   " "    "*"   "*" 
14  ( 1 ) "*"   "*"  "*"   "*"  " " "*"   " "   "*"    " "   " "    "*"   "*" 
15  ( 1 ) "*"   "*"  "*"   "*"  " " "*"   " "   "*"    "*"   " "    "*"   "*" 
16  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   " "   "*"    "*"   " "    "*"   "*" 
17  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   " "   "*"    "*"   " "    "*"   "*" 
18  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   "*"   "*"    "*"   " "    "*"   "*" 
19  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   "*"   "*"    "*"   "*"    "*"   "*" 
          CWalks LeagueN DivisionW PutOuts Assists  Errors NewLeagueN
1  ( 1 )  " "    " "     " "       " "     " "      " "    " "       
2  ( 1 )  " "    " "     " "       " "     " "      " "    " "       
3  ( 1 )  " "    " "     " "       "*"     " "      " "    " "       
4  ( 1 )  " "    " "     "*"       "*"     " "      " "    " "       
5  ( 1 )  " "    " "     "*"       "*"     " "      " "    " "       
6  ( 1 )  " "    " "     "*"       "*"     " "      " "    " "       
7  ( 1 )  " "    " "     "*"       "*"     " "      " "    " "       
8  ( 1 )  "*"    " "     "*"       "*"     " "      " "    " "       
9  ( 1 )  "*"    " "     "*"       "*"     " "      " "    " "       
10  ( 1 ) "*"    " "     "*"       "*"     "*"      " "    " "       
11  ( 1 ) "*"    "*"     "*"       "*"     "*"      " "    " "       
12  ( 1 ) "*"    "*"     "*"       "*"     "*"      " "    " "       
13  ( 1 ) "*"    "*"     "*"       "*"     "*"      "*"    " "       
14  ( 1 ) "*"    "*"     "*"       "*"     "*"      "*"    " "       
15  ( 1 ) "*"    "*"     "*"       "*"     "*"      "*"    " "       
16  ( 1 ) "*"    "*"     "*"       "*"     "*"      "*"    " "       
17  ( 1 ) "*"    "*"     "*"       "*"     "*"      "*"    "*"       
18  ( 1 ) "*"    "*"     "*"       "*"     "*"      "*"    "*"       
19  ( 1 ) "*"    "*"     "*"       "*"     "*"      "*"    "*"       
```

An asterisk indicates that a given variable is included in the corresponding model and by default in R are included only the first eight-variable models. For instance, this output indicates that the best two-variable model contains only Hits and CRBI. For the best nineteen-variable models, the variable Hits is included in all the models except the first one, while the variable CRBI is included in seventeen of them.

(b) Now we move to see which is the best model and we need to look at the adjusted R^2 initially:

```
> regfit.full.summary$adjr2
 [1] 0.3188503 0.4208024 0.4450753 0.4672734 0.4808971
 0.4972001 0.5007849 0.5137083 0.5180572 0.5222606
 0.5225706 0.5217245 0.5206736 0.5195431 0.5178661
 0.5162219 0.5144464 0.5126097 0.5106270
```

Since we have 19 different models, it is difficult to see, which is the best model across them, thus we look at the max

```
> regfit.full.by.adjr2 <- which.max(regfit.full.summary$adjr2)
> regfit.full.by.adjr2
[1] 11
```

Thus the best model is the model which includes 11 variables and the second best model is the model that includes 10 variables. Moving to the Mallow's statistics, we look at the list of the metrics:

```
> regfit.full.summary$cp
 [1] 104.281319  50.723090  38.693127  27.856220  21.613011
 14.023870  13.128474   7.400719   6.158685   5.009317
  5.874113 7.330766   8.888112  10.481576  12.346193
  14.187546  16.087831  18.011425  20.000000
```

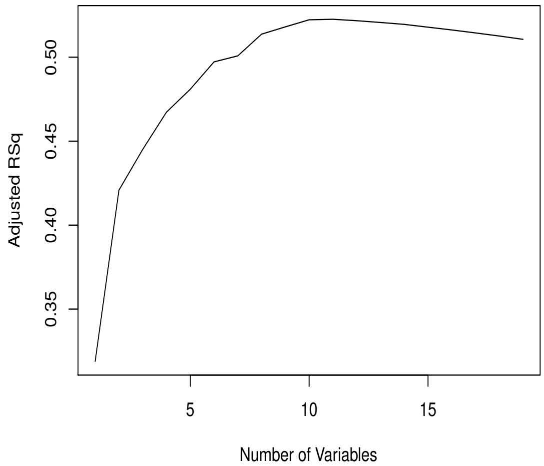
Also in this scenario, we look at the model with lowest Mallow's statistic,

```
> regfit.full.by.cp <- which.min(regfit.full.summary$cp)
> regfit.full.by.cp
[1] 10
```

Thus the best model is the model with 10 explanatory variables, followed by the model with 11 explanatory variables. These results are also confirmed graphically in Figure 1.1 on the left panel for the adjusted R^2 and on the right for the Mallow's statistic.

2. By using the Hitters data described in Question 1,

(a) We show the results for the model with 10 explanatory variables and with 11 explanatory variables. For the first model, we have



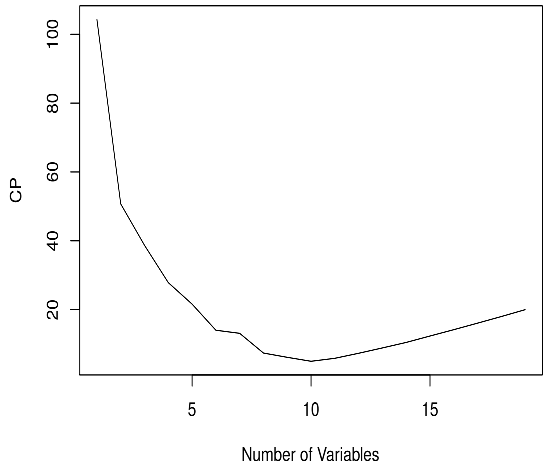


Figure 1.1: Plot of adjusted R^2 (left) and of the Mallow's C_k statistic (right) across the models.

```
> coef(regfit.full,10)
 (Intercept)        AtBat         Hits        Walks
 162.5354420   -2.1686501    6.9180175    5.7732246
CAtBat         CRuns          CRBI        CWalks    DivisionW
-0.1300798    1.4082490    0.7743122   -0.8308264 -112.3800575
PutOuts       Assists
0.2973726    0.2831680
```

Looking at the statistically significance of the coefficients, we run the linear regression model:

```
> mod10 <- lm(Salary~AtBat + Hits + Walks + CAtBat + CRuns
+ CRBI + CWalks + Division + PutOuts + Assists , Hitters)
> summary(mod10)

Call:
lm(formula = Salary ~ AtBat + Hits + Walks + CAtBat + CRuns +
    CRBI + CWalks + Division + PutOuts + Assists, data = Hitters)

Residuals:
    Min      1Q  Median      3Q     Max
-939.11 -176.87  -34.08  130.90 1910.55

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  162.53544   66.90784   2.429 0.015830 *
AtBat         -2.16865    0.53630  -4.044 7.00e-05 ***
Hits           6.91802    1.64665   4.201 3.69e-05 ***
Walks          5.77322    1.58483   3.643 0.000327 ***
```

```
CAtBat         -0.13008    0.05550  -2.344 0.019858 *
CRuns           1.40825    0.39040   3.607 0.000373 ***
CRBI            0.77431    0.20961   3.694 0.000271 ***
CWalks         -0.83083    0.26359  -3.152 0.001818 **
DivisionW    -112.38006   39.21438  -2.866 0.004511 **
PutOuts         0.29737    0.07444   3.995 8.50e-05 ***
Assists         0.28317    0.15766   1.796 0.073673 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 311.8 on 252 degrees of freedom
Multiple R-squared:  0.5405,Adjusted R-squared:  0.5223
F-statistic: 29.64 on 10 and 252 DF,  p-value: < 2.2e-16
```

In this case, almost all the coefficients are statistically significant, with a weaker significance for the Assists, the intercept and Number of times at bat during his career.
Moving to the 11 explanatory variables model, we have the following coefficients:

```
> coef(regfit.full,11)
 (Intercept)        AtBat         Hits        Walks
 135.7512195   -2.1277482    6.9236994    5.6202755
CAtBat        CRuns         CRBI       CWalks      LeagueN
-0.1389914    1.4553310    0.7852528   -0.8228559   43.1116152
DivisionW      PutOuts        Assists
-111.1460252    0.2894087    0.2688277
```

Looking at the statistically significance of the coefficients, we run the linear regression model:

```
> mod11 <- lm(Salary~AtBat + Hits + Walks + CAtBat + CRuns +
CRBI + CWalks + League + Division + PutOuts + Assists , Hitters)
> summary(mod11)

Call:
lm(formula = Salary ~ AtBat + Hits + Walks + CAtBat + CRuns +
    CRBI + CWalks + League + Division + PutOuts + Assists,
    data = Hitters)

Residuals:
   Min     1Q Median     3Q    Max
-932.2 -175.4  -29.2  130.4 1897.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  135.75122   71.34623   1.903 0.058223 .
AtBat         -2.12775    0.53746  -3.959 9.81e-05 ***
Hits           6.92370    1.64612   4.206 3.62e-05 ***
```

```
Walks            5.62028    1.59064   3.533 0.000488 ***
CAtBat          -0.13899    0.05609  -2.478 0.013870 *
CRuns            1.45533    0.39270   3.706 0.000259 ***
CRBI             0.78525    0.20978   3.743 0.000225 ***
CWalks          -0.82286    0.26361  -3.121 0.002010 **
LeagueN         43.11162   39.96612   1.079 0.281755
DivisionW     -111.14603   39.21835  -2.834 0.004970 **
PutOuts          0.28941    0.07478   3.870 0.000139 ***
Assists          0.26883    0.15816   1.700 0.090430 .
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 311.7 on 251 degrees of freedom
Multiple R-squared:  0.5426,Adjusted R-squared:  0.5226
F-statistic: 27.07 on 11 and 251 DF,  p-value: < 2.2e-16
```

In this case, the number of statistically significant variables is reduced with the respect to the previous model. Hence, in this case, the new variable League is not statistically significant and there is a confirmed weak significant for the Assists and the Number of times at bat during his career.

(b) Looking at the results previously described, I would suggest for the reasons of parsimony that the best model is the model with 10 explanatory variable, since the variable League is not statistically significance when included in the model.

3. 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) From CourseWork 8, we have that the $SS_R = 5550.811$ and $SS_T = 5784.543$, thus we can compute the R^2 as

$$R^2 = \frac{SS_R}{SS_T} = 0.9595937$$

Analogously, we can compute the adjusted R^2, which is:

$$adj(R^2) = \left(1 - (n-1)\frac{MS_E}{SS_T}\right) = \left(1 - (25-1)\cdot\frac{10.62417}{5784.543}\right) = 0.9559205$$

(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:

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

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

We find the least square estimator by using

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y} \\ &= \begin{pmatrix} 25 & 219 \\ 219 & 3055 \end{pmatrix}^{-1} \begin{pmatrix} 559.60 \\ 7375.44 \end{pmatrix} \\ &= \begin{pmatrix} 3.320780 \\ 2.176167 \end{pmatrix} \end{aligned}$$

Based on the previous results, we need to define

$$\begin{aligned} SS_R = \widehat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{Y} - n\bar{y}^2 &= \begin{pmatrix} 3.320780 & 2.176167 \end{pmatrix} \cdot \begin{pmatrix} 559.60 \\ 7375.44 \end{pmatrix} - 25 \cdot 22.384^2 \\ &= 17908.5 - 12526.09 = 5382.409 \end{aligned}$$

Moving to the SS_T, we have that

$$SS_T = \boldsymbol{Y}^t\boldsymbol{Y} - n\bar{y}^2 = 18310.63 - 12526.09 = 5784.54$$

Thus, we have that $SS_E = SS_T - SS_R = 5784.54 - 5382.409 = 402.1338$. Moving to S^2 or the so called MS_E, we have

$$S^2 = \frac{SS_E}{(25-2)} = \frac{402.1338}{23} = 17.48408$$

Thus, we can compute the R^2 and the adjusted R^2 as follows:

$$R^2 = \frac{SS_R}{SS_T} = \frac{5382.409}{5784.54} = 0.9304813$$

$$adj(R^2) = \left(1 - (25-1)\frac{MS_E}{SS_T}\right) = 0.9274588$$

(c) Looking at the adjusted R^2, we can conclude that the best model is the model with two explanatory variables (0.9559) with respect to the one explanatory variable (0.9274)