QUEEN MARY UNIVERSITY OF LONDON

MTH5120 Statistical Modelling I

Solution to Exercise Sheet 11

---

1. We use the Bridge.txt dataset available on QMPlus, where information from 45 bridge projects are compiled. The response and predictor variables are as follows:

   - Y: Time is the design time in person-days;
   - X_1: DArea is the deck area of bridge (000 sq ft);
   - X_2: CCost is the construction cost ($000);
   - X_3: Dwgs is the number of structural drawings;
   - X_4: Length is the length of bridge (ft);
   - X_5: Spans is the number of spans.

   Take the logarithm transformation of all the variables.

   (a) As in Coursework 10, before running the model, we need to take the logarithm of all the variables considered:

```
> data <- read.table("bridge.txt", header=TRUE)
> attach(data)
> Y<- log(data[,2])
> X1 <- log(data[,3])
> X2 <- log(data[,4])
> X3 <- log(data[,5])
> X4 <- log(data[,6])
> X5 <- log(data[,7])
```

   Then we define the model with all the explanatory variables and we run the backward elimination procedure:

```
> m1 <- lm(Y ~ X1 + X2 + X3 + X4 + X5)
> reduced.model <- step(m1, direction="backward")
Start:  AIC=-98.71
Y ~ X1 + X2 + X3 + X4 + X5

       Df Sum of Sq    RSS      AIC
- X4    1   0.00607 3.8497 -100.640
- X1    1   0.01278 3.8564 -100.562
<none>              3.8436  -98.711
- X2    1   0.18162 4.0252  -98.634
- X5    1   0.26616 4.1098  -97.698
- X3    1   1.45358 5.2972  -86.277

Step:  AIC=-100.64
```

```
Y ~ X1 + X2 + X3 + X5

       Df Sum of Sq    RSS      AIC
- X1    1   0.01958 3.8693 -102.412
<none>              3.8497 -100.640
- X2    1   0.18064 4.0303 -100.577
- X5    1   0.31501 4.1647  -99.101
- X3    1   1.44946 5.2991  -88.260

Step:  AIC=-102.41
Y ~ X2 + X3 + X5

       Df Sum of Sq    RSS      AIC
<none>              3.8693 -102.412
- X2    1   0.17960 4.0488 -102.370
- X5    1   0.29656 4.1658 -101.089
- X3    1   1.44544 5.3147  -90.128
```

Thus, backward elimination based on AIC chooses the model with the three predictors X_2, X_3 and X_5, which are the logarithm of the construction cost; of the number of structural drawings and of the number of spans.

Thus in conclusion the best model, called **M1**, is

$$Y = \beta_0 + \beta_1 X_3 + \beta_2 X_5 + \beta_3 X_2 + \varepsilon$$

where the variables are taken in logarithm. On the other hand, the second best model, called **M2**, is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_5 + \varepsilon$$

(b) Starting from M1, we need to define the model

```
> modfinal <- lm(Y ~ X3 + X5 + X2)
```

Then, we can find the leverage values and the Cook's distance values:

```
> hatvalues(modfinal)
         1          2          3          4          5
0.05597217 0.09329413 0.10592688 0.04644378 0.04758423
         6          7          8          9         10
0.06748107 0.09368962 0.02250127 0.03868009 0.13866675
        11         12         13         14         15
0.13038134 0.04398404 0.04844501 0.07297180 0.07297180
        16         17         18         19         20
0.04894325 0.13583933 0.04588027 0.04634131 0.06052667
        21         22         23         24         25
0.05170357 0.25375049 0.04590607 0.10698551 0.14842192
        26         27         28         29         30
```

```
0.09956467 0.09196830 0.13298453 0.13298453 0.04567849
        31         32         33         34         35
0.07166907 0.09749309 0.16660874 0.13446120 0.05356590
        36         37         38         39         40
0.05557333 0.06605840 0.19186747 0.17652561 0.04481103
        41         42         43         44         45
0.12231997 0.04459390 0.07533738 0.10159530 0.07104674
```

```
> i=(1:45)
> plot(i,hatvalues(modfinal),main="Leverage values, Bridge")
> plot(i,cooks.distance(modfinal),main="Cook's distance, Bridge")
```

Figure 1.1 shows the leverage values (left) and the Cook's distance values (right) for the model with three explanatory variables (X_3, X_5 and X_2). In our case, we

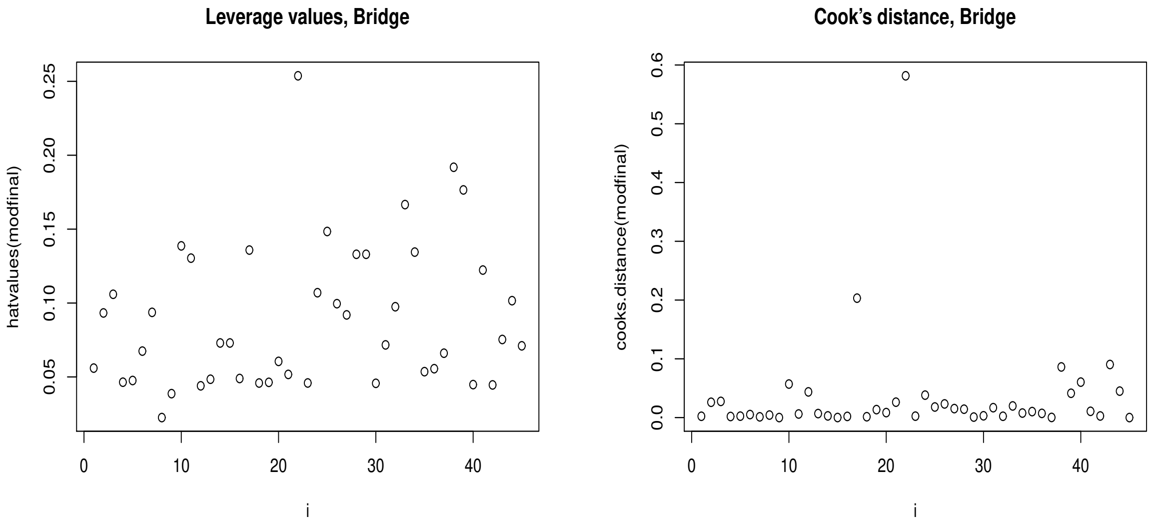

Figure 1.1: Plot of leverage values (left) and Cook's distance values (right) for the model with three explanatory variables.

have the number of observations, n, equal to 45 and number of regression values, p, equal to 4. Thus, a leverage values is larger if $2p/n$ and very large if $3p/n$ and in our case it means:

$$\frac{2p}{n} = \frac{2 \times 4}{45} = 0.178, \qquad \frac{3p}{n} = \frac{3 \times 4}{45} = 0.267$$

Looking at Figure 1.1, we have that there is one values very large related to observation (25) and a few bigger than the large value of 0.178 (the observation 38 and 39). Moving to the Cook's distance, the critical value is obtained as

```
> qf(p=0.50,df1=4,df2=41)
[1] 0.8532109
```

The right panel of Figure 1.1 shows the Cook's distance for all the observations and we can see that the highest Cook's distance for observation 22 is smaller than that, thus it is nevertheless more influential than any other states.

Moving to the second best model, **M2**, we have

```
> secmodfinal <- lm(Y ~ X1 + X2 + X3 + X5)
> summary(secmodfinal)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X5)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.67135 -0.17582 -0.02815  0.24654  0.67035 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.19011    0.47854   4.577 4.51e-05 ***
X1          -0.05431    0.12041  -0.451  0.65441    
X2           0.18389    0.13422   1.370  0.17832    
X3           0.85724    0.22089   3.881  0.00038 ***
X5           0.21252    0.11747   1.809  0.07795 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3102 on 40 degrees of freedom
Multiple R-squared:  0.7758,Adjusted R-squared:  0.7534 
F-statistic: 34.61 on 4 and 40 DF,  p-value: 1.694e-12
```

From this model, we can compute the leverage values:

```
> hatvalues(secmodfinal)
         1          2          3          4          5 
0.06785119 0.21042799 0.11005482 0.05709607 0.12333525 
         6          7          8          9         10 
0.09640295 0.09682564 0.02801548 0.11019164 0.19947899 
        11         12         13         14         15 
0.13043430 0.04418682 0.04894538 0.09391220 0.09391220 
        16         17         18         19         20 
0.05737107 0.15814656 0.10713544 0.07717193 0.06156026 
        21         22         23         24         25 
0.08045804 0.28030341 0.05045836 0.11249258 0.19784606 
        26         27         28         29         30 
0.10329039 0.16933385 0.14376263 0.14376263 0.07491489 
        31         32         33         34         35 
0.07195192 0.11966058 0.16663753 0.13861098 0.05683629 
        36         37         38         39         40 
0.05559312 0.06635602 0.19336401 0.19848587 0.05532075 
        41         42         43         44         45 
0.15725609 0.04466579 0.11554178 0.14773888 0.08290138 
```

In this case, the number of observations does not change, thus n is equal to 45, while the number of regressions moves to 5. The leverage values is larger if $(2 \times 5)/45 = 0.223$ and very large if $(3 \times 5)/45 = 0.334$. Figure 1.2 shows the leverage values (left) and the Cook's distance values (right).

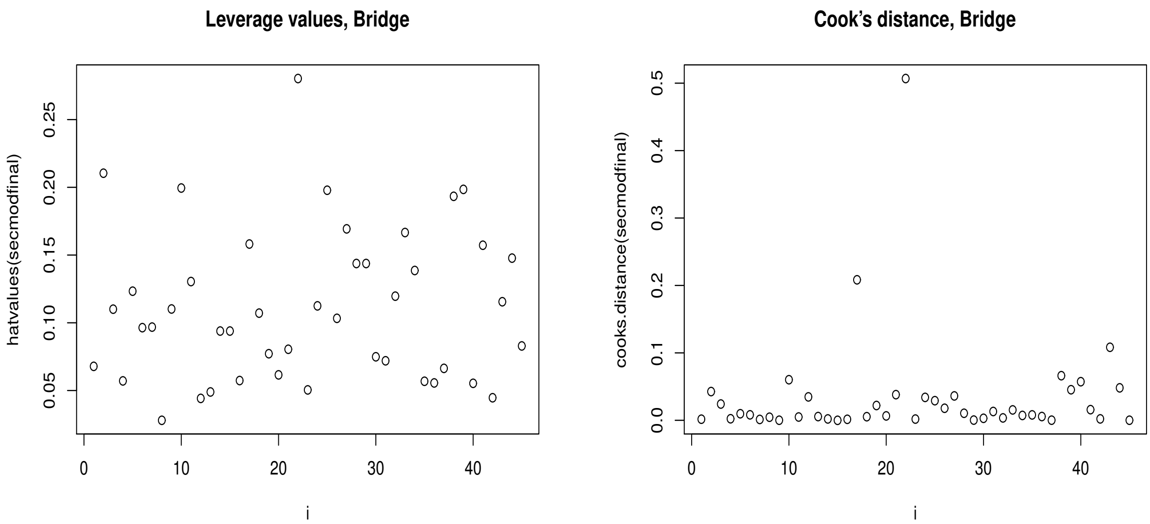

Figure 1.2: Plot of leverage values (left) and Cook's distance values (right) for the model with four explanatory variables.

From Figure 1.2, we see that the only observation with a very large value is related to observation 22, while for the Cook's distance, the critical value is 0.885 and the highest Cook's distance is for observation 22 but it is not greater than the critical value.

2. **Coursework component**
When fitting the model

$$E[Y_i] = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}$$

to a set of $n = 5$ observations, the following results were obtained using the general linear model notation:

$$\left(\boldsymbol{X}^t \boldsymbol{X}\right)^{-1} = \begin{pmatrix} 209.32 & -3.82 & -0.71 \\ -3.82 & 0.069 & 0.013 \\ -0.71 & 0.013 & 0.002 \end{pmatrix}$$

with variables:

Variable	1	2	3	4	5
Y	92.5	94.9	89.3	94.1	98.9
X_1	50.9	54.1	47.3	45.1	37.6
X_2	20.8	16.9	25.2	49.7	95.2

(a) In order to find the leverage values, we need to define $\boldsymbol{H}$, which is

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t = \begin{pmatrix} 0.302 & 0.336 & 0.263 & 0.158 & -0.059 \\ 0.336 & 0.805 & -0.191 & 0.080 & -0.030 \\ 0.263 & -0.191 & 0.775 & 0.245 & -0.092 \\ 0.158 & 0.080 & 0.245 & 0.227 & 0.290 \\ -0.059 & -0.030 & -0.092 & 0.290 & 0.891 \end{pmatrix}$$

Then we need to select the diagonal elements of this matrix and see if they are smaller of the threshold values. In our case the diagonal elements are

$$\begin{pmatrix} 0.302 & 0.805 & 0.775 & 0.227 & 0.891 \end{pmatrix}$$

So a leverage values is large if it bigger than $(2 \times 3)/5$ and very large if it is bigger than $(3 \times 3)/5$. In our case, no values are bigger than the thresholds.

(b) Let us consider the problem

$$E[Y_i] = \beta_0 + \beta_1 x_{1,i}$$

with Y and X_1 defined as above with the exception of $x_{1,5}$, which changes from 37.6 to 20. Thus the matrix $\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}$ becomes

$$\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1} = \begin{pmatrix} 2.77 & -0.06 \\ -0.06 & 0.001 \end{pmatrix}$$

In this case we need to compute the same matrix $\boldsymbol{H}$ as before

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t = \begin{pmatrix} 0.275 & 0.307 & 0.238 & 0.216 & -0.037 \\ 0.307 & 0.353 & 0.255 & 0.223 & -0.139 \\ 0.238 & 0.255 & 0.220 & 0.208 & 0.078 \\ 0.216 & 0.223 & 0.208 & 0.204 & 0.148 \\ -0.037 & -0.139 & 0.078 & 0.148 & 0.949 \end{pmatrix}$$

with diagonal elements equal to

$$\begin{pmatrix} 0.275 & 0.353 & 0.220 & 0.204 & 0.949 \end{pmatrix}$$

In this case, the threshold values change: the leverage is large if is bigger than $2 \times 2/5$ and very large if is bigger than $3 \times 2/5$ and in our case the last observation is bigger than the threshold value.