QUEEN MARY UNIVERSITY OF LONDON

MTH5120 Statistical Modelling I

Solution to Exercise Sheet 10

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) 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 run the model with all the explanatory variables:

```
> m1 <- lm(Y ~ X1 + X2 + X3 + X4 + X5)
> summary(m1)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68394 -0.17167 -0.02604  0.23157  0.67307

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.28590    0.61926   3.691 0.000681 ***
X1          -0.04564    0.12675  -0.360 0.720705
```

```
X2             0.19609    0.14445   1.358 0.182426
X3             0.85879    0.22362   3.840 0.000440 ***
X4            -0.03844    0.15487  -0.248 0.805296
X5             0.23119    0.14068   1.643 0.108349
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3139 on 39 degrees of freedom
Multiple R-squared:  0.7762,Adjusted R-squared:  0.7475
F-statistic: 27.05 on 5 and 39 DF,  p-value: 1.043e-11
```

The only statistically significant at 5% level variable is X_3 (number of structural drawing) and the intercept, while all the other variables are not significant at any level. This regression is overall significant with a F-statistic of 27.05 and a p-value smaller that 5%. The adjusted R^2 is higher with a value of 74.75%, thus explaining an high variation in the data. As a further results, we have a look at the VIF values

```
> vif(m1)
      X1       X2       X3       X4       X5
7.164619 8.483522 3.408900 8.014174 3.878397
```

All the values are smaller than 10, thus we do not have strong problems of multicollinearity.

Thus, we find the best reduced model by using the AIC procedure. We start with the backward elimination procedure:

```
> 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.

Based on the forward selection based on the AIC, arrives at the same model as backward elimination based on AIC.

```
> modyn <- lm(Y ~ 1)
> aic.forward.model <- step(modyn, scope=~X1 + X2 + X3 + X4 + X5,
 direction="forward")
Start:  AIC=-41.35
Y ~ 1

       Df Sum of Sq     RSS     AIC
+ X3    1   12.1765  4.9975 -94.898
+ X2    1   11.6147  5.5593 -90.104
+ X1    1   10.2943  6.8797 -80.514
+ X4    1   10.0120  7.1620 -78.704
+ X5    1    8.7262  8.4478 -71.274
<none>              17.1740 -41.347

Step:  AIC=-94.9
Y ~ X3

       Df Sum of Sq    RSS      AIC
+ X5    1   0.94866 4.0488 -102.370
+ X2    1   0.83170 4.1658 -101.089
+ X4    1   0.66914 4.3284  -99.366
+ X1    1   0.47568 4.5218  -97.399
<none>              4.9975  -94.898

Step:  AIC=-102.37
Y ~ X3 + X5

       Df Sum of Sq    RSS     AIC
+ X2    1  0.179598 3.8693 -102.41
<none>              4.0488 -102.37
+ X1    1  0.018535 4.0303 -100.58
+ X4    1  0.016924 4.0319 -100.56
```

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

       Df Sum of Sq    RSS     AIC
<none>              3.8693 -102.41
+ X1    1  0.019578 3.8497 -100.64
+ X4    1  0.012868 3.8564 -100.56
```

Thus in conclusion the best model is

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

where the variables are taken in logarithm.

(b) We run the best model in R by using the following commands:

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

and then we compute the VIF for this model:

```
> vif(modfinal)
      X3       X5       X2
3.245326 2.509206 4.905365
```

Also in this case, the three values are all smaller than 10, thus we do not have any problem of multicollinearity.

2. a-b Based on the best model selected, thus the one with X_3; X_5 and X_2 in logarithmic transformation, we have:

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

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.69415 -0.17456 -0.03566  0.22739  0.64945

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3317     0.3577   6.519  7.9e-08 ***
X3            0.8356     0.2135   3.914 0.000336 ***
X5            0.1963     0.1107   1.773 0.083710 .
X2            0.1483     0.1075   1.380 0.175212
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.3072 on 41 degrees of freedom
Multiple R-squared:  0.7747,Adjusted R-squared:  0.7582
F-statistic: 46.99 on 3 and 41 DF,  p-value: 2.484e-13
```

From the summary of the linear regression model, we have that the explanatory variable associated with the number of drawings remains statistically significant, while the variable related to the number of spans becomes statistically significant but only at 10% level. The overall regression is highly statistically significant with a F-statistic equal to 46.99 and a p-value really small. Looking at the adjusted R^2 of the model with all the 5 explanatory variables and the best AIC model, we have that it improves from 74.75% to 75.82%. It is a small improvement in term of adjusted R^2, but in term of AIC, it is an important improvement.

```
> anova(modfinal)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq  F value    Pr(>F)    
X3         1 12.1765 12.1765 129.0266 3.063e-14 ***
X5         1  0.9487  0.9487  10.0523  0.002878 ** 
X2         1  0.1796  0.1796   1.9031  0.175212    
Residuals 41  3.8693  0.0944                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Looking at the Anova table, we have that variable 2 (construction cost) is not statistically significant once variable 3 and 5 are included in the model. However, this variable is included in the model by the AIC.

Figure 1.1 shows the standardized residuals versus the fitted values (left) and the QQ plot (right). For the constant variance assumption, we do not have any problem, while for the normality assumption, the QQ plot shows some issues on both tails. In particular, the left tails seems out of the Normal distribution, but before making strong assumption, we had a look at the Shapiro-Wilk test:

```
> shapiro.test(stdresfinal)

Shapiro-Wilk normality test

data:  stdresfinal
W = 0.97452, p-value = 0.4175
```

The test gives a strong p-value of 0.42, thus we do not reject the null hypothesis of normality assumption of the standardized residuals.

3. When fitting the model

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

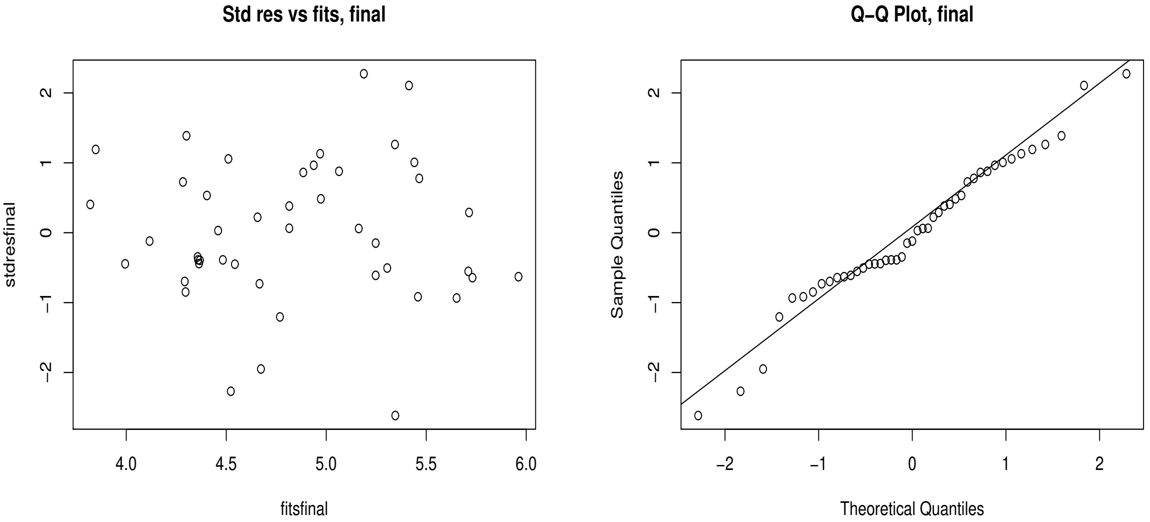

Figure 1.1: Plot of standardized residuals versus fitted values (left) and QQ plot (right) for the model with three explanatory variables.

to a set of $n = 25$ observations, the following results were obtained using the general linear model notation:

$$\boldsymbol{X}^t\boldsymbol{X} = \begin{pmatrix} 25 & 219 & 10232 \\ 219 & 3055 & 133899 \\ 10232 & 133899 & 6725688 \end{pmatrix}, \qquad \boldsymbol{X}^t\boldsymbol{Y} = \begin{pmatrix} 559.60 \\ 7375.44 \\ 337071.69 \end{pmatrix}$$

$$\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1} = \begin{pmatrix} 0.1132 & -0.0044 & -0.00008 \\ -0.0044 & 0.0027 & -0.00004 \\ -0.00008 & -0.00004 & 0.000001 \end{pmatrix}$$

Also $\boldsymbol{Y}^t\boldsymbol{Y} = 18310.63$ and $\bar{Y} = 22.384$.

(a) In order to compute the AIC criterion we need to find:

$$AIC = 2(p+1) - 2\log L$$

where

$$-2\log L = n(log 2\pi + \log\widehat{\sigma}^2 + 1)$$

and the MLE of σ^2 is $\widehat{\sigma}^2 = \frac{SS_E}{n}$. In our case, we have already compute the SS_E in the previous courseworks, thus

$$\begin{aligned} SS_E = SS_T - SS_R &= \left(\boldsymbol{Y}^t\boldsymbol{Y} - n\bar{y}^2\right) - \left(\widehat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{Y} - n\bar{y}^2\right) \\ &= 5784.54 - 5550.81 = 233.73 \end{aligned}$$

The MLE of σ^2 is

$$\widehat{\sigma}^2 = \frac{SS_E}{n} = \frac{233.73}{25} = 9.3492$$

In our scenario, we have the number of observations, n, equal to 25 and the number of regression parameters, p, equal to 3. Thus, the AIC is equal to:

$$AIC = 2(3+1) + 25(log 2\pi + \log 9.3492 + 1) = 8 + 126.8292 = 134.8292$$

Then we can compute the VIF, which is

$$VIF = \frac{1}{1-R^2} = 24.748$$

which is bigger than 10, thus we have some problems of multicollinearity in the data.

(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:

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

$$\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1} = \begin{pmatrix} 0.1075 & -0.0077 \\ -0.0077 & 0.00087 \end{pmatrix}$$

As stated in point (a), we need to run the AIC by using the usual formula, where in our case, p is changing to 2 and the $\widehat{\sigma}^2$ is changing too. In the two-dimensional problem, we have that

$$SS_R = 5382.409 \qquad SS_E = 402.1338$$

Thus the MLE estimator of σ^2 is equal to 16.0853 and the AIC is

$$AIC = 2(2+1) + 25(log 2\pi + \log 16.0853 + 1) = 6 + 140.3946 = 146.3946$$

Moving to the VIF, we have 14.38462, which is a bigger than 10 value.

(c) In conclusion, based on the AIC the best model results the one with two explanatory variables.