QUEEN MARY UNIVERSITY OF LONDON

MTH5120 Statistical Modelling I

Solution to Exercise Sheet 6

1. Based on the liver.csv dataset we have seen in the Practical session,

   (a) We fit the model with three explanatory variables defined as $\log_{10} Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$, where $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. We use the following R commands:

```
> data <- read.csv('liver.csv')
> x1 <- data$x1
> x2 <- data$x2
> x3 <- data$x3
> ly <- data$log10y
>
> modly3 <- lm(ly ~ x1 + x2 + x3)
> summary(modly3)
Call:
lm(formula = ly ~ x1 + x2 + x3)
Residuals:
      Min         1Q     Median         3Q        Max
-0.102004 -0.016222 -0.002609  0.011884  0.138314
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.4836209  0.0426287   11.35 1.95e-15 ***
x1          0.0095236  0.0003064   31.08  < 2e-16 ***
x2          0.0092945  0.0003825   24.30  < 2e-16 ***
x3          0.0692251  0.0040779   16.98  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.04687 on 50 degrees of freedom
Multiple R-squared:  0.9723,Adjusted R-squared:  0.9707
F-statistic:   586 on 3 and 50 DF,  p-value: < 2.2e-16
```

   The coefficients of the three explanatory variables jointly with the intercept are statistically significant with positive value. Moreover, looking at the R^2 there is an improvement with respect to the other models, since in this case it explains the 97% of the variation. Moreover, we define the standardized residuals and the fitted values

```
> stdres3 <-rstandard(modly3)
> fits3 <-fitted(modly3)
```

   (b) In order to assess the assumptions of normality and constant variance of the random errors, we run two different plots

```
> plot(fits3,stdres3, main="Std res vs fits, liver3")
> qqnorm(stdres3, main="Q-Q Plot, liver3")
> qqline(stdres3)
```

Figure 1.1 shows the standardized residuals versus the fitted values (left panel) and the QQ plot (right panel). The first figure show if there is any problem with the constant variance assumption, while the second one refers to the Normality assumption.

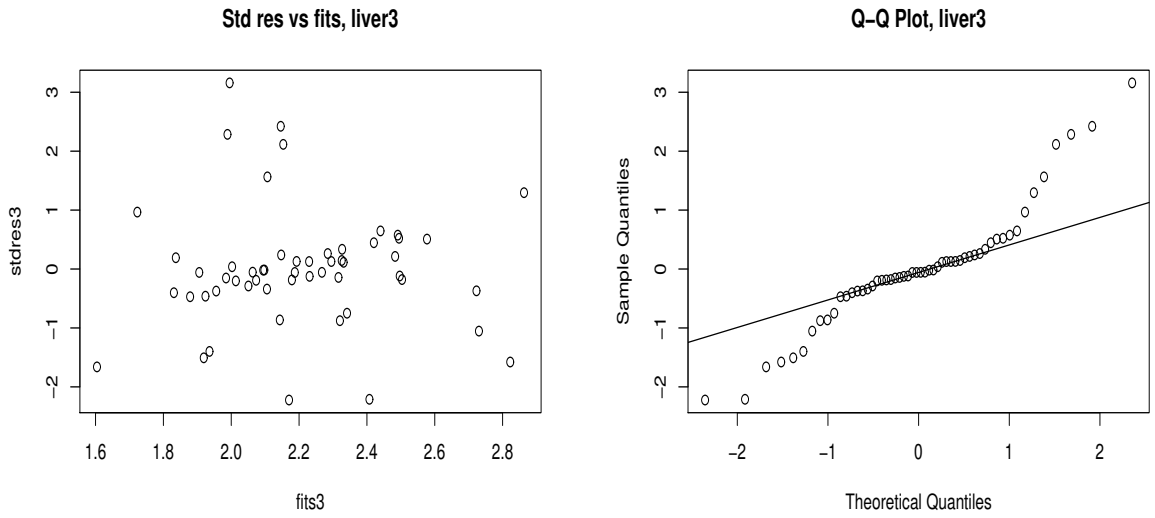

Figure 1.1: Plot of standardized residuals versus fitted values (left) and of the QQ plot for the model with three explanatory variables (right).

Regarding the constant variance it seems not be too many problems, while for the normality assumption, Figure 1.1 shows heavy tails in both part and thus we are ready to reject the normality assumption. Moreover, we run also the Shapiro-Wilk test for the normality assumption

```
> shapiro.test(stdres3)
Shapiro-Wilk normality test
data:  stdres3
W = 0.92082, p-value = 0.001605
```

From the test, we see a really small p-value (0.0016), thus we have strong evidence against the normality assumption.

(c) We compare the model 3 with the two models find in the Practical session. In particular, the model that fits the best is the second model for the following reasons:

- The model assumptions are approximately met for Model 1 (with 1 explanatory variable) and for Model 2 (with two explanatory variables), but not for Model 3. Thus Model 3, in the current form, should not be considered for inference.

- Model 2 gives considerably larger R^2 and in particular adjusted R^2 and a smaller values of s, the estimate of the square root of the error variance σ^2 than does Model 1
- All the parameters in Model 2 are significantly different from zero.

(d) We use Model 2 since it is the best model and we run the leverage values and the Cook's distance

```
> hat <- hatvalues(modly2)
> cook<-cooks.distance(modly2)
> i<-1:54
> plot(i,hat, main="Leverage values")
> plot(i,cook, main="Cooks distance values")
> qf(0.5, 3, 51)
[1] 0.7993137
```

Figure 1.2 shows the Leverage values and the Cook's distance. In both cases there are a couple of points higher than normal, but for the Cook's distance the higher values is 0.25, which is below the threshold point of 0.799. Thus the following results confirm that Model 2 fits well the data.

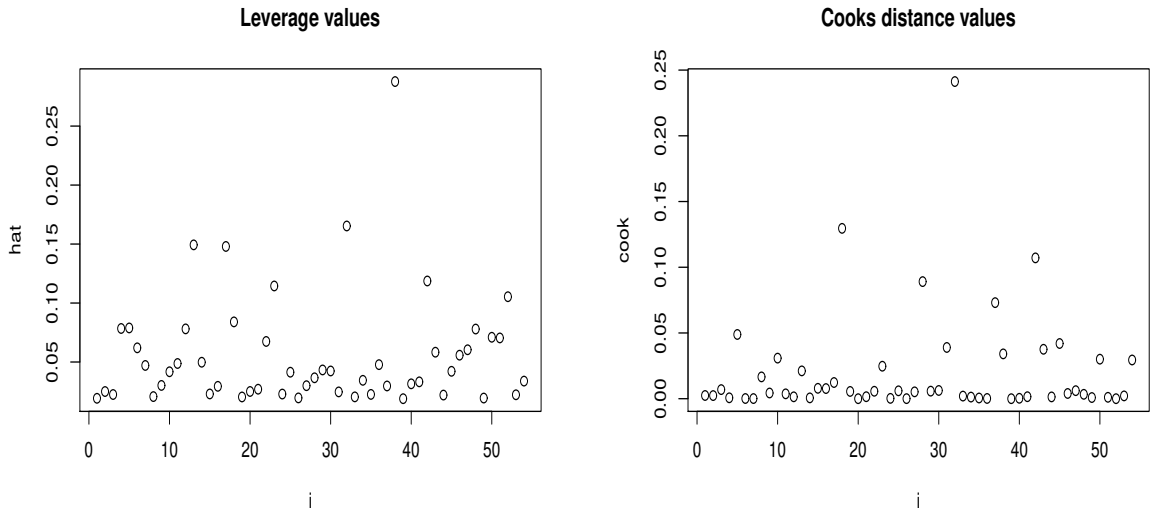

Figure 1.2: Plot of leverage values (left) and of the Cook's distance for the model with two explanatory variables (right).

2. **Coursework component**
Consider a set of equity returns from four different markets across 12 different periods. The data are available in the marketdata.txt. Define the forth variable as your response variable (Y). Define the following three models:

- Model 1 – $Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$;

- Model 2 – $Y_i = \beta_0 + \beta_2 x_{2i} + \varepsilon_i$
- Model 3 – $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

(a) We run the three different models in R through the following commands for the first model (Model 1)

```
> modl1 <- lm(Y ~ X1)
> summary(modl1)
Call:
lm(formula = Y ~ X1)
Residuals:
      Min        1Q    Median        3Q       Max
-0.011009 -0.007319  0.001411  0.005846  0.009434

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0001395  0.0022835  -0.061    0.952
X1           0.8733093  0.0657648  13.279 1.12e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007866 on 10 degrees of freedom
Multiple R-squared:  0.9463,Adjusted R-squared:  0.941
F-statistic: 176.3 on 1 and 10 DF,  p-value: 1.121e-07
```

In this case, the parameter of X_1 is statistically significant, while the intercept is not statistically significant. Moreover, the intercept has a small value, while the estimation of β_1 is strongly positive. The R^2 in this model is huge, around 94%, thus explaining a big amount of variation.

Moving to Model 2, we have:

```
> modl2 <- lm(Y ~ X2)
> summary(modl2)
Call:
lm(formula = Y ~ X2)
Residuals:
     Min       1Q   Median       3Q      Max
-0.03334 -0.01801 -0.00547  0.01595  0.04198
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0004339  0.0071658  -0.061   0.9529
X2           0.4400105  0.1462066   3.010   0.0131 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0246 on 10 degrees of freedom
Multiple R-squared:  0.4753,Adjusted R-squared:  0.4228
```

```
F-statistic: 9.057 on 1 and 10 DF,  p-value: 0.01313
```

The coefficient of X_2 is statistically significant but only at 0.05, while the intercept is not statistically significant. The R^2 is smaller than the previous model. Thus this model seems not to be the correct model. Moving to Model 3, we have

```
> modl3 <- lm(Y ~ X1 + X2)
> summary(modl3)
Call:
lm(formula = Y ~ X1 + X2)
Residuals:
      Min         1Q    Median         3Q        Max
-0.013399 -0.005104  0.000514  0.005249  0.008964
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.311e-05  2.289e-03   0.028    0.979
X1          8.163e-01  8.650e-02   9.436 5.79e-06 ***
X2          6.232e-02  6.150e-02   1.013    0.337
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.007855 on 9 degrees of freedom
Multiple R-squared:  0.9518,Adjusted R-squared:  0.9411
F-statistic: 88.92 on 2 and 9 DF,  p-value: 1.182e-06
```

In this case the estimated parameter of X_1 is statistically significant, while the other two parameters are not significant at all. Looking at the R^2 and it's adjusted form it seems that the model is not beating Model 1.

(b) Looking at the normality assumption, based on the QQ plot or at the Shapiro-Wilk test, we have:

```
> # Model 1
> shapiro.test(stdres1)

Shapiro-Wilk normality test

data:  stdres1
W = 0.89265, p-value = 0.1275

> # Model 2
> shapiro.test(stdres2)

Shapiro-Wilk normality test

data:  stdres2
W = 0.94658, p-value = 0.5876

> # Model 3
```

```
> shapiro.test(stdres3)

Shapiro-Wilk normality test

data:  stdres3
W = 0.93659, p-value = 0.4552
```

In all the models, the Shapiro-Wilk does not reject the null hypothesis, thus we are not rejecting the normality assumption in all the models.

(c) The best model across the three is Model 1 due to the parsimony principles. Looking at the R^2 or adjusted R^2, the best model is divided between Model 1 and Model 3. However, Model 3 does not have significant parameters except the β_1.

If the student argues correctly, we can consider also Model 3, the best model. It is a choice of the researcher in this case.