Statistical Modeling I
Practical in R – Output

Practical in R – Output

In this practical, we will work with the Liver dataset (liver.csv). We will look at two different models and their analysis.

A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random selection of 54 patients was available for analysis. From each patient record, the following information was extracted from the pre-operation evaluation and are reported in the liver.csv file in the following order:

- X_3: blood clotting score
- X_2: prognostic index
- X_1: enzyme function test score
- $log_{10}y$: the base 10 logarithm transformation of the response variable (Y), which is the number of weeks the patients survived after the operation.

1. We load the data by using the read.csv data and then we renominate them:

```
> data <- read.csv('liver.csv')
> x1 <- data$x1
> x2 <- data$x2
> x3 <- data$x3
> ly <- data$log10y
```

2. We fit the Model 1 as $\log_{10} Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$, where $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and we use the following R command

```
> modly1 <- lm(ly ~ x1)
> summary(modly1)
Call:
lm(formula = ly ~ x1)
Residuals:
     Min       1Q   Median       3Q      Max
-0.51859 -0.12908 -0.00951  0.14817  0.44233

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.545465   0.106628  14.494  < 2e-16 ***
```

```
x1           0.008568   0.001334   6.423 4.11e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.2064 on 52 degrees of freedom
Multiple R-squared:  0.4424,Adjusted R-squared:  0.4317
F-statistic: 41.25 on 1 and 52 DF,  p-value: 4.111e-08
```

As one can see from the summary of the linear regression, the coefficients of the intercept and of the x_1 are both statistically significant and they have both positive coefficients. Looking at the R^2, we can see that the model explains a 44% of the total variation. Moving to the fitted values and the standardised residuals, we can define them as

```
> stdres1 <-rstandard(modly1)
> fits1<-fitted(modly1)
```

3. In order to assess the assumption of normality and constant variance, we need to run the Shapiro-Wilk test and the plots of the standardized residuals versus the fitted values and the relative QQ plot. Starting from the plots, we have the following commands:

```
> plot(fits1,stdres1, main="Std res vs fits, liver1")
> qqnorm(stdres1, main="Q-Q Plot, liver1")
> qqline(stdres1)
```

which produce Figure 1.1.

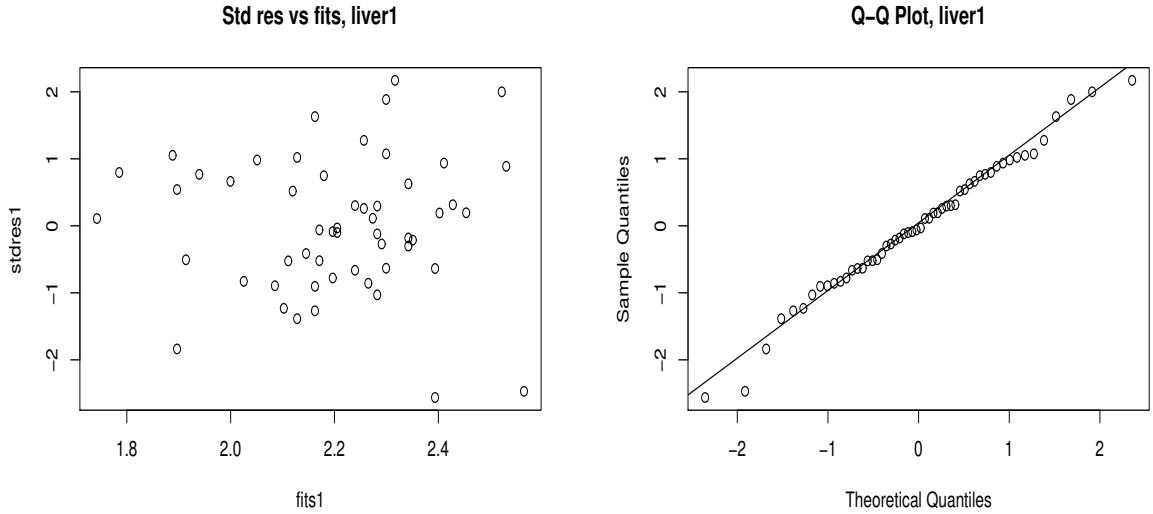

Figure 1.1: Plot of standardized residuals versus fitted values (left) and of the QQ plot for the model with one explanatory variable (right).

From Figure 1.1, we have no clear evidence against of constant variance of the random errors (left panel) and of the assumption of normality (right panel), which is also confirmed from the Shapiro-Wilk test:

```
> shapiro.test(stdres1)
Shapiro-Wilk normality test
data:  stdres1
W = 0.98639, p-value = 0.7954
```

From the p-value of the test (0.7954), there is no evidence against the normality.

4. We obtain the scatter plot of the standardized residuals previously find against the second explanatory variable, X_2, which is the prognostic index. We use the following command:

```
> plot(x2,stdres1, main="Std res vs x2, liver1")
```

Figure 1.2 shows the scatterplot and it indicates that the residuals increase when the values of X_2 increase. It may be possible that some relationship between the residuals and the explanatory variables. Thus in the next part, we will study a new model.

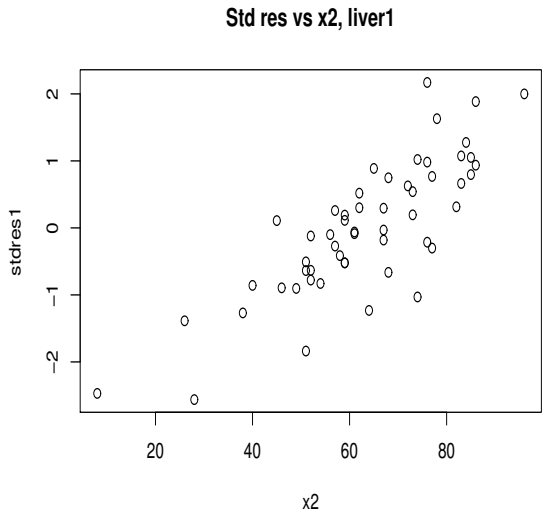

Figure 1.2: Plot of standardized residuals versus x_2.

5. We fit the novel model with two explanatory variables: $\log_{10} Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, where $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and we use the following R commands

```
> modly2 <- lm(ly ~ x1 + x2)
> summary(modly2)
Call:
lm(formula = ly ~ x1 + x2)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.31817 -0.05522  0.00751  0.07153  0.31409
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.9074230  0.0889753   10.20 6.57e-14 ***
x1          0.0087530  0.0007803   11.22 2.22e-15 ***
x2          0.0098633  0.0009812   10.05 1.08e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.1207 on 51 degrees of freedom
Multiple R-squared:  0.813,Adjusted R-squared:  0.8056
F-statistic: 110.8 on 2 and 51 DF,  p-value: < 2.2e-16
```

Adding one explanatory variable, X_2 leads to improvements in the R^2 from 44% to 81%. This is also confirmed from the significance of the parameters. In model 2, the three coefficients are all statistically significant and they are all positive. Then we also save the standardized residuals and the fitted values.

```
> stdres2 <-rstandard(modly2)
> fits2 <-fitted(modly2)
```

6. As for the previous model, we assess the assumption of constant volatility and of normality by using the plots of the standardized residuals versus the fitted values in the novel model 2 as stated below:

```
> plot(fits2,stdres2, main="Std res vs fits, liver2")
> qqnorm(stdres2, main="Q-Q Plot, liver2")
> qqline(stdres2)
```

Figure 1.3 shows the standardized residuals versus the fitted values in the left panel, while the QQ plot is shown in the right panel.

From the left panel, we can see that there is no problem with the constant variance assumption, regarding the normality assumption, looking at the QQ plot, it seems that the assumption of normality is not contradicted. However, looking at the distribution it seems that there is some slightly heavier tails than we might expect. Thus, we run the Shapiro-Wilk test

```
> shapiro.test(stdres2)
Shapiro-Wilk normality test
data:  stdres2
W = 0.97326, p-value = 0.2673
```

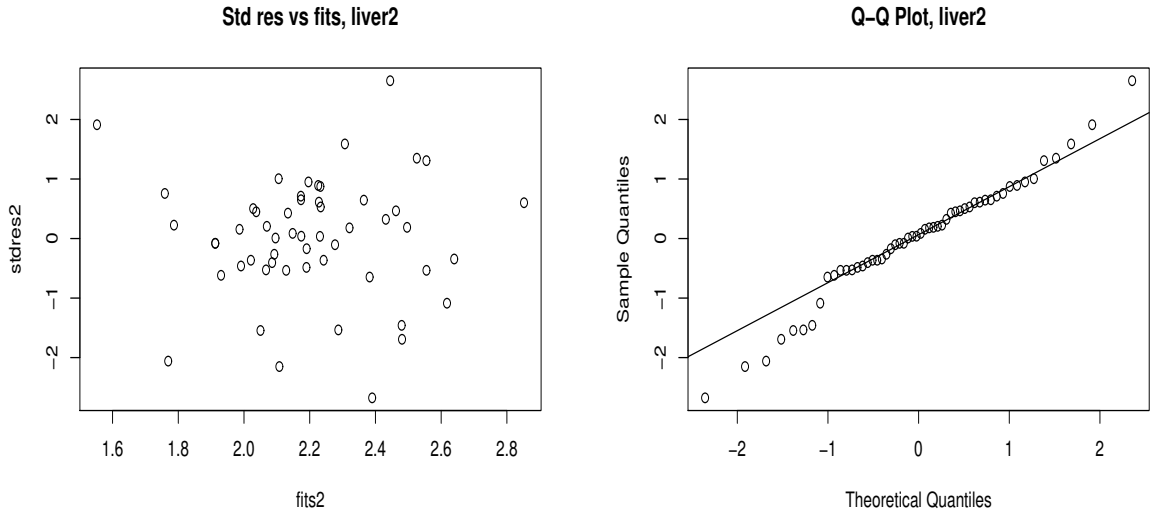

Figure 1.3: Plot of standardized residuals versus fitted values (left) and of the QQ plot for the model with two explanatory variables (right).

which has a p-value of 0.2673 greater than the threshold, thus confirming that there are no evidence against the normality assumption.

7. We obtain the scatter plot of the standardized residuals from the model including X_1 and X_2 (as in point 5) versus the explanatory variable X_3. We use the following R command:

```
> plot(x3,stdres2, main="Std res vs x3, liver2")
```

Figure 1.4 shows the scatterplot and it indicates that the residuals increase when the values of X_3 increase although not linearly. It may be possible that some relationship between the residuals and the explanatory variables.

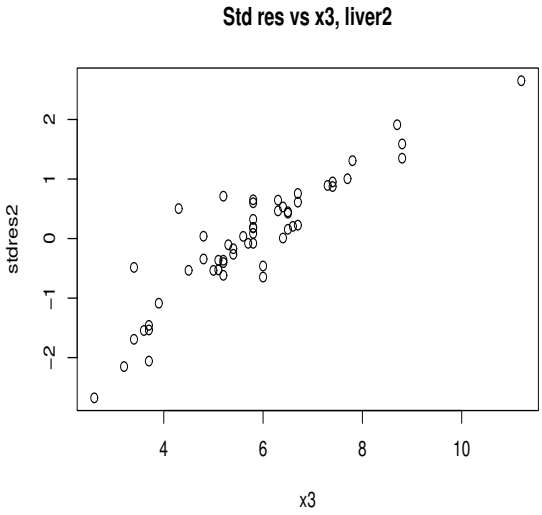

Figure 1.4: Plot of standardized residuals versus x_3.