Statistical Modeling I
Practical in R – Output

Practical in R – Output

This practical reminds you how to load the data from a .csv file and how to run a linear regression in R. Moreover, it gives you the opportunity to study the Q-Q plot and some tests.

Janka hardness is an important structural property of Australian timbers which is difficult to measure directly. However, it is related to the density of the timber which is comparatively easy to measure. Therefore it is desirable to fit a model enabling the Janka hardness to be predicted from the density. The Janka hardness and density of 36 Australian eucalyptus hardwoods are given in the table.

Density	Hardness	Density	Hardness	Density	Hardness
24.7	484	39.4	1210	53.4	1880
24.8	427	39.9	989	56.0	1980
27.3	413	40.3	1160	56.5	1820
28.4	517	40.6	1010	57.3	2020
28.4	549	40.7	1100	57.6	1980
29.0	648	40.7	1130	59.2	2310
30.3	587	42.9	1270	59.8	1940
32.7	704	45.8	1180	66.0	3260
35.6	979	46.9	1400	67.4	2700
38.5	914	48.2	1760	68.8	2890
38.8	1070	51.5	1710	69.1	3740
39.3	1020	51.5	2010	69.1	3140

The data is in a .csv file jankaNEW.csv on the QMplus page. Copy it to your home directory. In particular, the density values are in Column 1 and the hardness values in Column 2. In our scenario, the dependent variable (i.e. y) is the hardness, while the density is the regressor variable (x).

1. Load the data in R as follows: To begin you have to tell R where you have saved the data, which is known as your working directory. You set it by telling R where it is, by using the command:

   ```
   setwd("name_directory")
   ```

   (Keep attention at / if you are using a Mac/Linus computer or a Windows)

   You will have to put the drive and directory where you have put the jankaNEW.csv file.

   If you copy and paste the direction location in Windows you will get a single backslah and you need to change that.

   You can check if you are in the correct working directory by using the command

```
getwd()
```

Once you are in the correct directory, you can load and read the data that are in csv format, by using the read.csv command.

```
> janka <- read.csv("jankaNEW.csv")
```

The data have been read into R but they are stored in two different columns and we need to allocate the columns of the matrix to x and y by using the following commands

```
> x<- janka[ ,1]
> y<- janka[ ,2]
```

Or you can use

```
> x<- janka$Density
> y<- janka$Hardness
```

Check that the data has been correctly read in by looking at the first few rows using

```
> head(janka)
  Density Hardness
1    24.7      484
2    24.8      427
3    27.3      413
4    28.4      517
5    28.4      549
6    29.0      648
```

2. Plot y against x. Is there a linear relationship between y and x?

```
> plot(x,y)
```

In Figure 1.1, we can see the plot of the data and at first look, it seems that there is a relationship between y and x, but it is not clear if it is a linear relation.

3. Fit the simple linear regression model, look at the `summary` and `anova`. Write down the fitted model and look at the fitted line plot.

   First of all we look at the simple linear regression model, by considering the estimation of the parameters

```
> mody <- lm(y ~ x)
> summary(mody)

Call:
lm(formula = y ~ x)
```

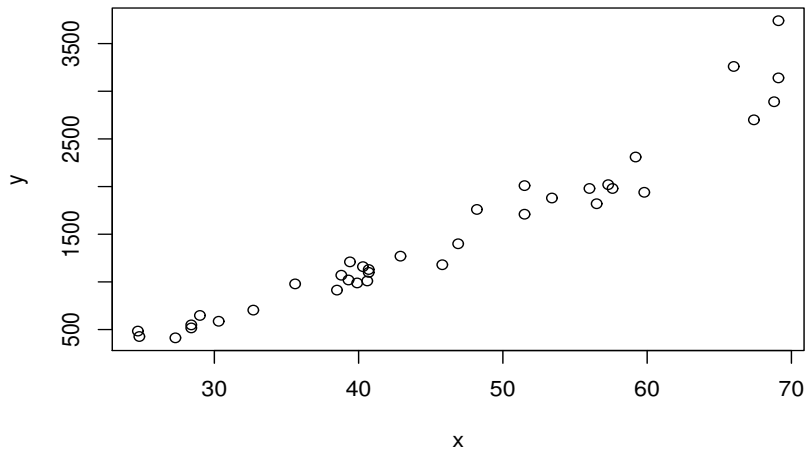

Figure 1.1: Plot of the data.

```
Residuals:
    Min      1Q  Median      3Q     Max
-417.10 -142.03  -13.83  103.70  814.42

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1298.282    139.496  -9.307  7.1e-11 ***
x              61.127      2.927  20.882  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 235.2 on 34 degrees of freedom
Multiple R-squared:  0.9277,Adjusted R-squared:  0.9255
F-statistic:   436 on 1 and 34 DF,  p-value: < 2.2e-16
```

Then we look at the ANOVA table

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

Response: y
          Df   Sum Sq  Mean Sq F value    Pr(>F)
x          1 24117743 24117743  436.04 < 2.2e-16 ***
Residuals 34  1880575    55311
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In conclusion, the fitted linear regression model is $y_i = -1298.282 + 61.127x_i$. The R^2 of the model is 92.77%, thus the model explains much of the variation. Looking at the

fitted line plot with respect to the data, in Figure 1.2, we can see that there is a positive relation among y and x.

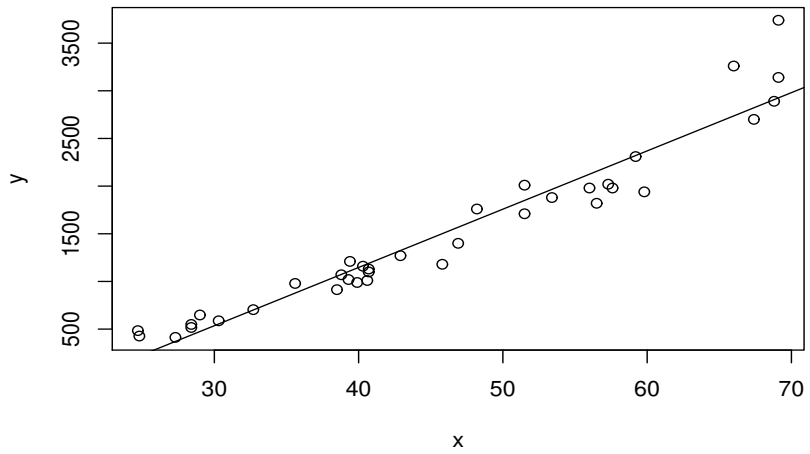

Figure 1.2: Plot of the data versus the fitted regression line.

4. Save the standardised residuals in `stdres` and fitted values in `fits`.

   We save the standardised residuals and the fitted values

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

5. Now let's look at the residual plots. First the Q-Q plot. Is there a possible problem? We can carry out a test of normality by `shapiro.test(stdres)` We run the Q-Q plot on the standardized residuals and the Shapiro-Wilk normality test as stated below

```
> qqnorm(stdres1, main="Q-Q Plot")
> qqline(stdres1)
> shapiro.test(stdres1)

Shapiro-Wilk normality test

data:  stdres1
W = 0.9051, p-value = 0.004718
```

   From Figure 1.3, we can see the QQ-plot, which shows two large outliers. This is confirmed also by the Shapiro-Wilk test, which have a small p-value 0.004, casting doubt on the assumption of normality. In fact a small p value means that the assumption of normality is not supported by the data.

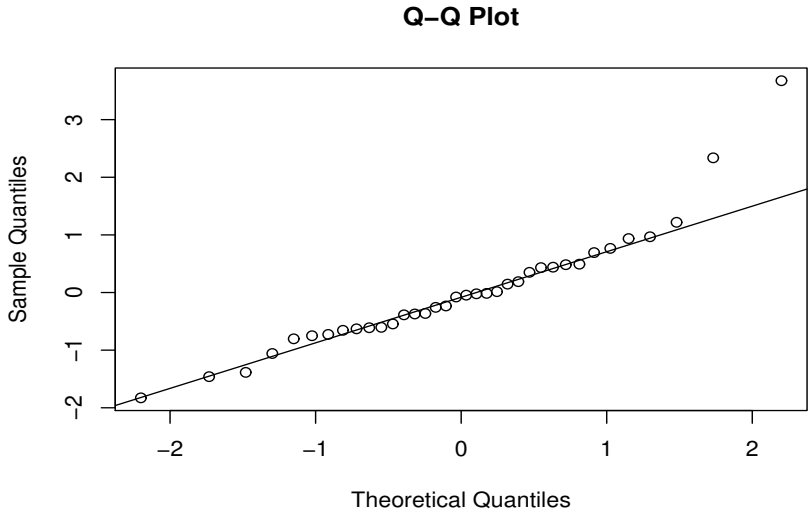

Figure 1.3: Q-Q plot of the standardized residuals of the model.

6. Next look at residuals versus fitted values. A random scatter here suggests that the assumption of equal variances is ok. A funnel shape suggests the variance is increasing with the mean. Is that what you see here?

   In conclusion, we plot the standardized residuals versus the data and versus the fitted values in order to see if the assumption of equal variance is correct or not.

```
> plot(x,stdres1, main="Std res vs x, Janka1")
> plot(fits1,stdres1, main="Std res vs fits, Janka1")
```

   Figure 1.4 shows the plots of the standardized residuals versus the data (left panel) and the standardized residuals versus the fitted values (right). In particular, we have that the plot has a funnel shape or more a trumpet shape in this case.

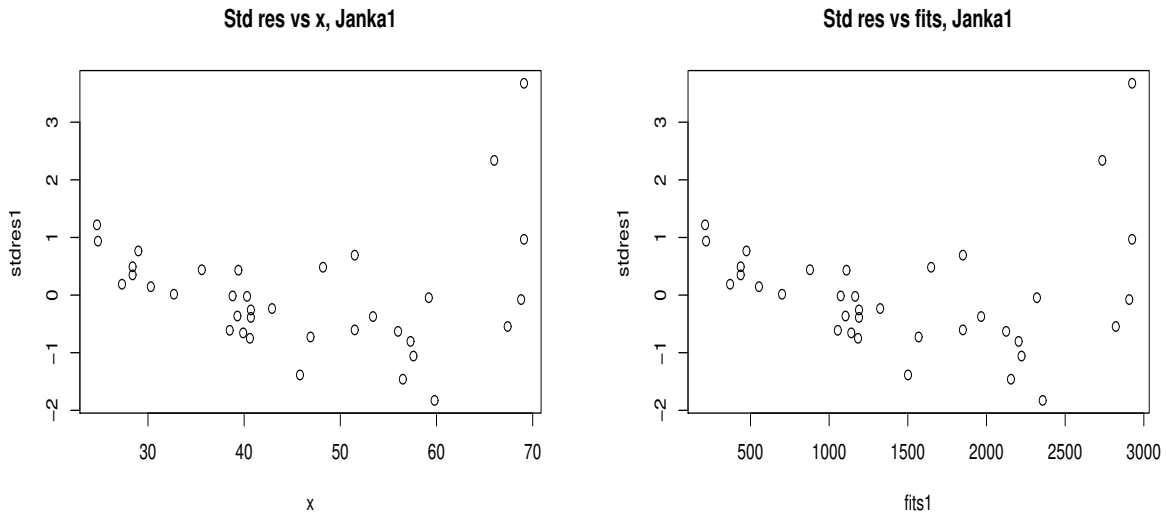

Figure 1.4: Plot of the standardized residuals versus x (left) and versus fitted values (right).