Statistical Modeling I
Practical in R – Output

Practical in R – Output

In this practical, we will work with the Janka dataset of last week. We will look at transformation of the variables.

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.

2. In the previous Practical, we have seen that the Shapiro-Wilk test had a small p-value, thus implying that the assumption of normality is not supported by the data. Moreover, the variance may be increasing suggests we should transform the dependent variable hardness. The first transformation we usually try is to take the logarithm of y.

```
> ly <- log(y)
> plot(x,ly)
```

We have plotted the x versus the logarithm transformation of the y. In Figure 1.1, we show the original plot between x and y and the corresponding with the $log(y)$. We can see the difference scale of the y axis moving from thousand to small values.

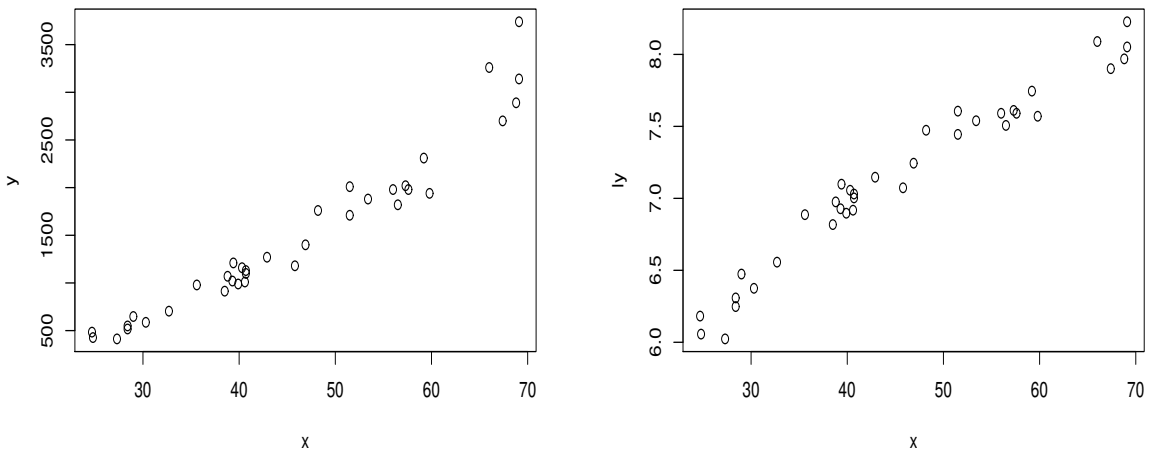

Figure 1.1: Plot of the original data (left) and of the transformed y (right).

3. Now fit a simple linear regression model with ly as the dependent variable.

```
> modly <- lm(ly ~ x)
> summary(modly)

Call:
lm(formula = ly ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32366 -0.09044  0.00305  0.07216  0.22764

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.165716   0.077762   66.43   <2e-16 ***
x           0.043274   0.001632   26.52   <2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1311 on 34 degrees of freedom
Multiple R-squared:  0.9539,Adjusted R-squared:  0.9525
F-statistic: 703.3 on 1 and 34 DF,  p-value: < 2.2e-16
```

The model with $log(y)$ is an improvement. The fitted model is $log(y_i) = 5.166 + 0.00433x_i$ with an R^2 equal to 95.39% thus explaining a bit more variation than the original model with x and y. This is also confirmed by the Anova table here below:

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

Response: ly
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 12.0875 12.0875  703.26 < 2.2e-16 ***
Residuals 34  0.5844  0.0172
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Save the fitted values and standardised residuals (you should probably call them something different). Look at the normal plot and the plot of residuals versus fitted values.

   We run the different plots and the Shapiro-Wilk test in order to see if there are problems with the normality assumption.

```
> stdres2 <-rstandard(modly)
> fits2<-fitted(modly)
> plot(x,stdres2, main="Std res vs x, Janka2")
> plot(fits2,stdres2, main="Std res vs fits, Janka2")
> qqnorm(stdres2, main="Q-Q Plot Janka 2")
> qqline(stdres2)
> shapiro.test(stdres2)

Shapiro-Wilk normality test

data:  stdres2
W = 0.97913, p-value = 0.7159
```

   From the Shapiro-Wilk test there is no evidence against the normality (p-value of 0.7159), which can be see also from the Q-Q plot (see Figure 1.3). However, looking at the plot of the standardized residuals against x (see left panel of Figure 1.2) shows curvature and suggests that fitting a quadratic term in x might be of value. This is confirmed by the plot of residuals versus fitted values (see right panel of Figure 1.2), which shows a distinct pattern. Small and large values of fitted values have mainly negative residuals and moderate values positive residuals.

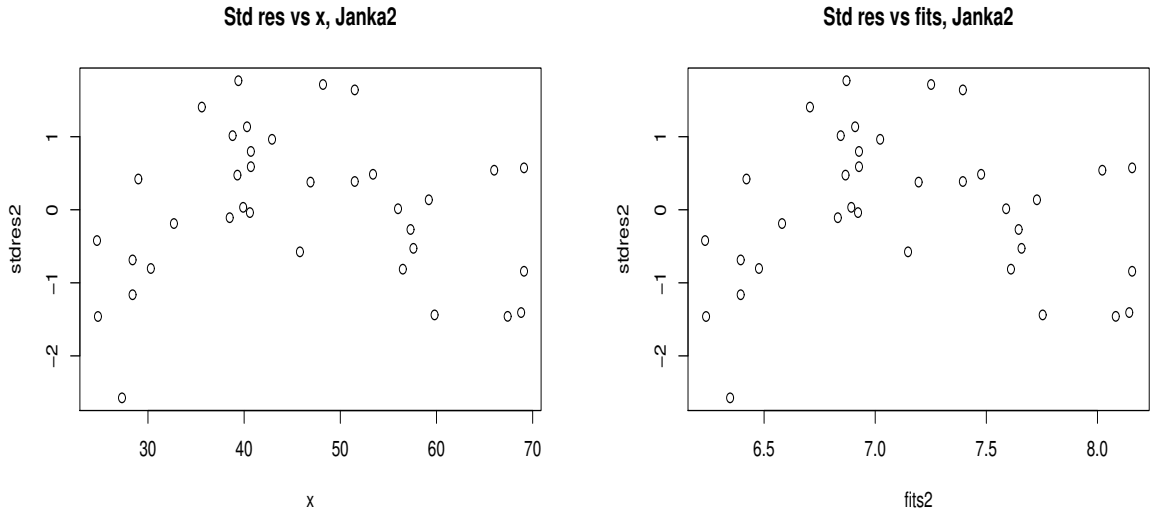

Figure 1.2: Plot of the residuals versus x (left) and versus fitted values (right).

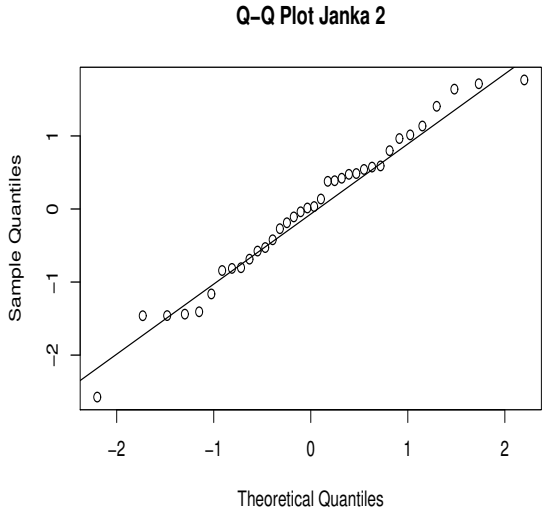

Figure 1.3: Plot of Q-Q plot.

5. It may happen that the plot of standardized residuals vs x suggested a quadratic model may be needed. To fit a polynomial model of degree 2, i.e. a quadratic model we use the command

   `modlyq <- lm(ly ~ poly(x,2,raw=TRUE))`

   We run a model with the logarithm of y and a quadratic model for the X. Thus the model is of the form $log(y_i) = \alpha + \beta_0 x_i + \beta_1 x_i^2 + \varepsilon_i$. In R the output for generating a quadratic component is shown above and here we report the summary

```
> modlyq <- lm(ly ~ poly(x,2,raw=TRUE))
> summary(modlyq)

Call:
```

```
lm(formula = ly ~ poly(x, 2, raw = TRUE))

Residuals:
     Min       1Q   Median       3Q      Max
-0.22989 -0.06121 -0.01134  0.08386  0.20051

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.2730665  0.2228136  19.178  < 2e-16 ***
poly(x, 2, raw = TRUE)1    0.0844374  0.0099349   8.499 8.04e-10 ***
poly(x, 2, raw = TRUE)2 -0.0004359  0.0001042  -4.181    2e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1076 on 33 degrees of freedom
Multiple R-squared:  0.9699,Adjusted R-squared:  0.968
F-statistic: 530.9 on 2 and 33 DF,  p-value: < 2.2e-16
```

The fitted model is $log(y) = 4.27 + 0.08x - 0.00044x^2$. All the parameters are highly significant. Moreover looking at the R^2 we have that the variation is explained mostly by this model, since the R^2 is equal to 96.99%. This is also confirmed by the anova table reported below

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

Response: ly
                          Df Sum Sq Mean Sq F value    Pr(>F)
poly(x, 2, raw = TRUE)  2 12.290  6.1450  530.86 < 2.2e-16 ***
Residuals              33  0.382  0.0116
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. Save the standardised residuals and fitted values for this model and check the residuals versus fitted values plot and the QQ-plot. Does this model seem to fit ok?

   We run the standardized residuals and we look at the different plots. Here are the command list:

```
> stdres3 <-rstandard(modlyq)
> fits3<-fitted(modlyq)
> plot(x,stdres3, main="Std res vs x, Janka3")
> plot(fits3,stdres3, main="Std res vs fits, Janka3")
> qqnorm(stdres3, main="Q-Q Plot Janka 3")
> qqline(stdres3)
> shapiro.test(stdres3)
Shapiro-Wilk normality test
```

```
data:  stdres3
W = 0.97904, p-value = 0.7129
```

Looking at the Shapiro-Wilk test, the assumption of the normality is not rejected. In fact the p-value is around 0.71 and this is confirmed by the Q-Q plot (see Figure 1.5). Looking at the standardized residuals versus x and versus the fitted values, it seems random and it looks okay (see Figure 1.4).

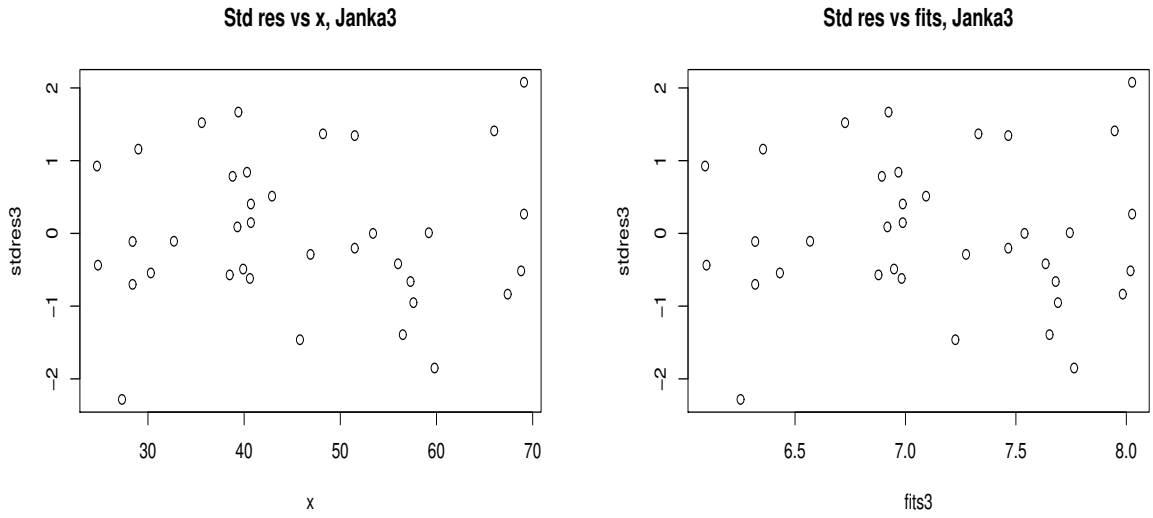

Figure 1.4: Plot of the residuals versus x (left) and versus fitted values (right).

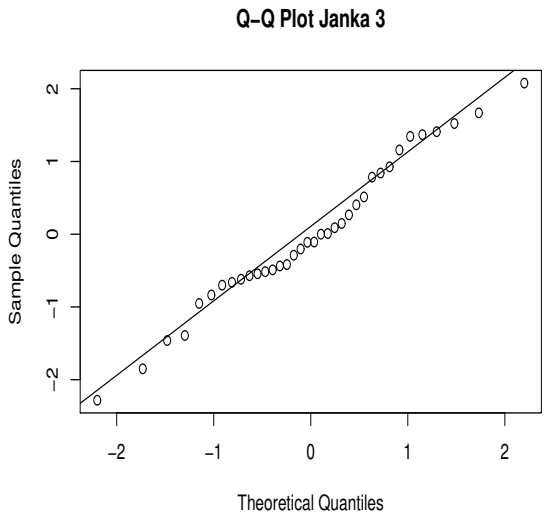

Figure 1.5: Plot of Q-Q plot.

7. Check the leverage values for this model.

```
> hat <- hatvalues(modlyq)
```

```
> cook<-cooks.distance(modlyq)
> i<-1:36
> plot(i,hat, main="Leverage values")
> plot(i,cook, main="Cooks distance values")
> qf(0.5, 3, 33)
[1] 0.8052067
```

Looking at the previous command, we have the largest leverage values are above $2*k/n$ but not $3*k/n$, where k is the number of estimated parameters. In our case, we have k equal to 3, the total number of observation n equal to 36, thus $6/n$ and $9/n$. In numbers, we have high leverage if it is greater than $6/36 = 0.16$ and very high leverage if greater than $9/36 = 0.25$. The largest Cook's distance is for observation 35 but at about 0.35 it is well below the value of 0.805 (which is the F value with 3 and 33 degrees of freedom), which would indicate a highly influential observation.

We can see these results in Figure 1.6, which shows the leverage values and Cook's distance. In conclusion, we can see that the model with quadratic term for x fits very well the data.

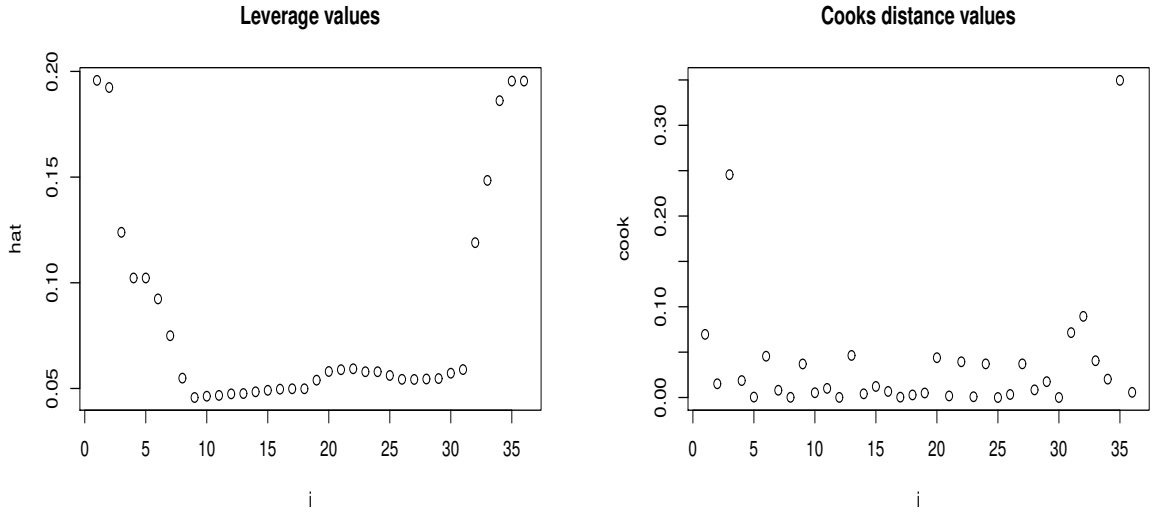

Figure 1.6: Plot of the Leverage values (left) and of the Cook's distance (right).