Statistical Modeling I
Practical in R

Practical in R

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)`

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

   `modly<- lm(ly~x)` We have called it modly to remind me it is the model with `ly` as the dependent variable.

   Look at the `summary` and `anova`, how has the fitted model changed?

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.

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))
   ```

   Have a look at the summary and anova. The summary shows the t tests for the intercept, linear and quadratic terms. The anova table is testing whether at least one of the linear and quadratic terms are needed.

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?

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")
   ```

   Note that as there are now three parameters in the model the relevant figures for high and very high leverage are $\frac{6}{n} = \frac{2p}{n}$ and $\frac{9}{n} = \frac{3p}{n}$. Similarly the cut-off for Cook's distance is the 50% point of an F on 3 and 33 degrees of freedom.