

Statistical Modeling I

Practical in R

Practical in R

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/Linux 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 backslash 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)
```

2. Plot y against x . Is there a linear relationship between y and x ?
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.
4. Save the standardised residuals in `stdres` and fitted values in `fits`.
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)`
A small p value means that the assumption of normality is not supported by the data.
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?