Report1

Seif Ben Mustapha, Mohamed Dhraief, Alice Patin, Julian Schnitzler

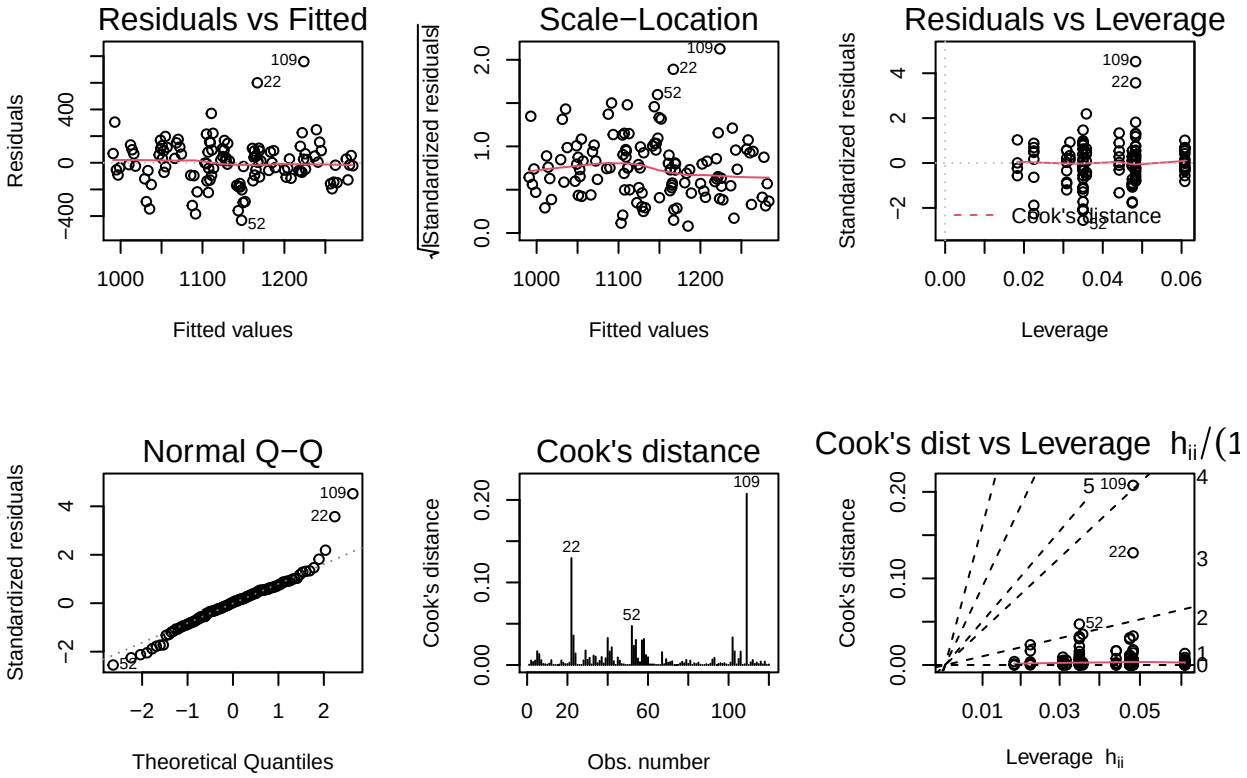
Introduction

Exploratory Data Analysis

Model Fitting

Model Assessment

In the following, we assess whether the assumptions underlying our model hold, i.e. that the errors have mean 0, the errors are homoscedastic, the errors are uncorrelated and are normally distributed.



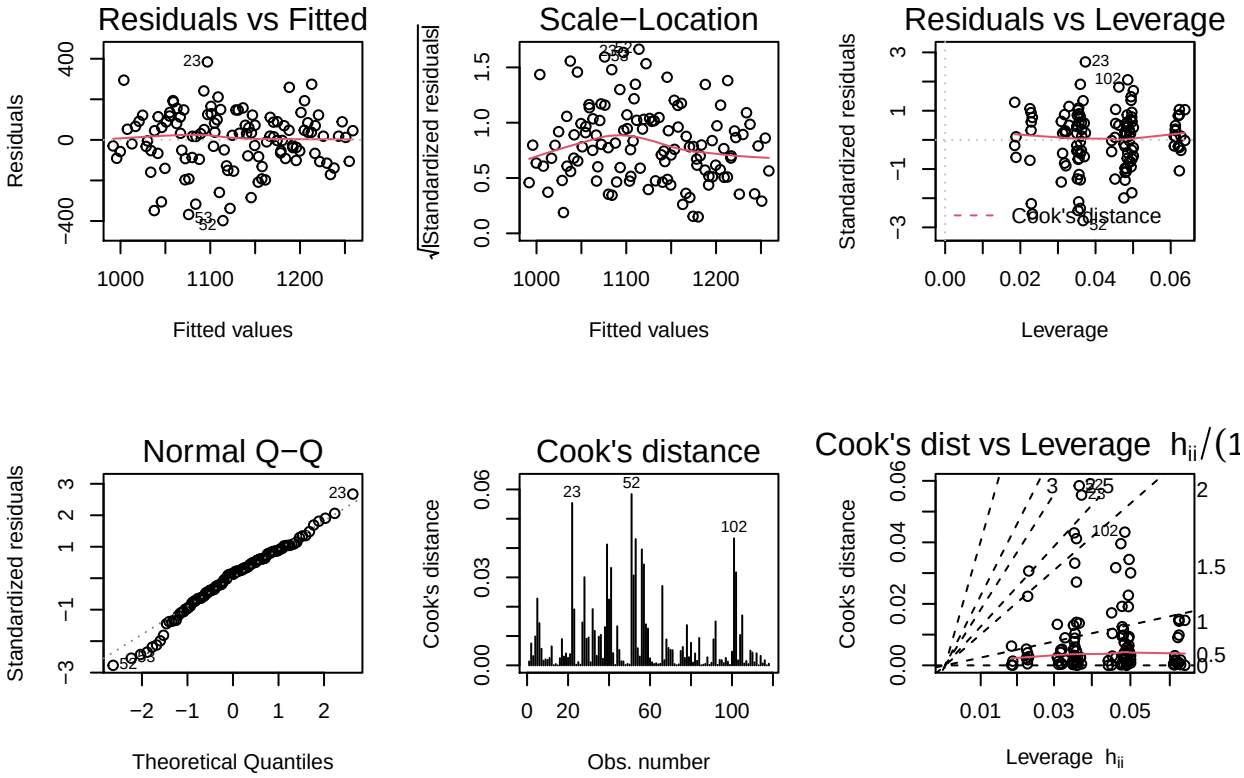
We observe that there are two outliers that might harm our model assumptions: Rows 22 and

109. Since these two points are the only two points that appear as outliers in all 4 graphs, we will remove those two from our dataset. For all other observations, it seems that they follow the model assumptions approximately.

In the Residuals vs Fitted graph, we can observe that the errors are homoscedastic, i.e. have the same variance, since the datapoints distribute in an similar manner above and underneath the line, independent of the fitted value. We can see an exception for 109 and 22.

The design of the experiment as written in the original paper indicates that the samples were chosen randomly, so we can consider all features to be uncorrelated.

From the Normal Q-Q plot, we can see that the residuals follow the line closely, i.e. the standard residuals do not exceed 3, except for rows 109 and 22. Hence, except for those two points, we consider the assumption of normally distributed errors to hold.



Indeed, removing those outliers seem to make the whole dataset support the model assumptions.

Conclusions

Two observations of the dataset seem to be outliers.

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

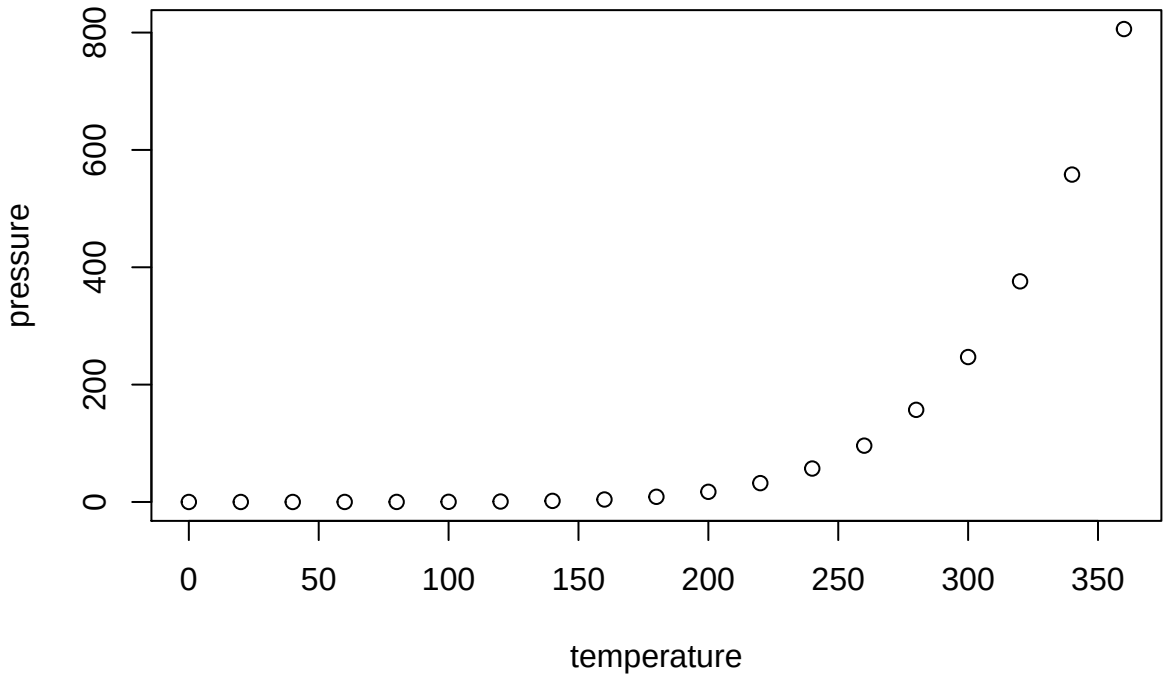
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.