Assignment 1 - Group 27

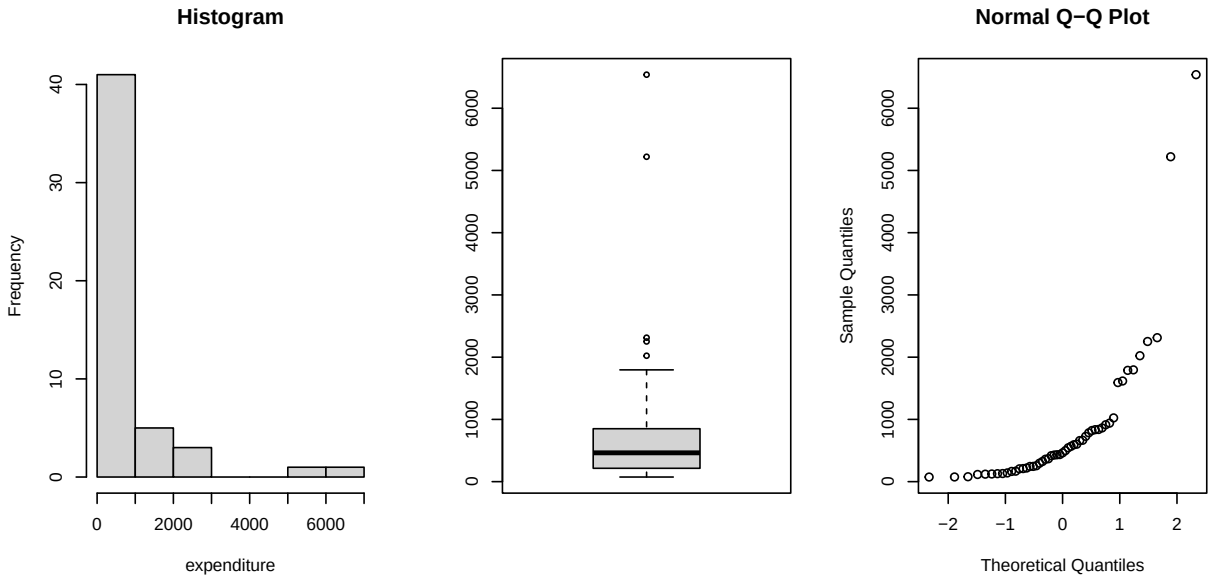
Joost Driessen, Emma van Lipzig, Rohan Zonneveld

2023-03-03

Excercise 1

Excercise 2

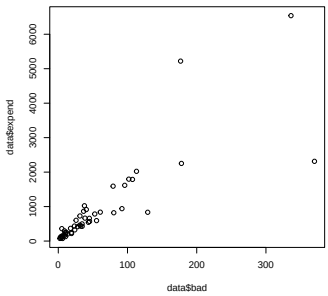
a

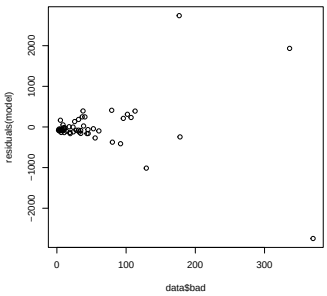


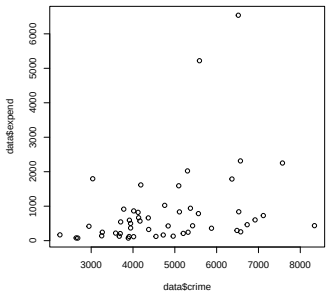
This dataset contains a number outliers.

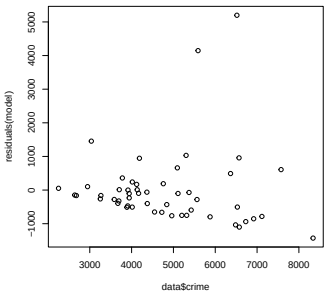
Influence Points

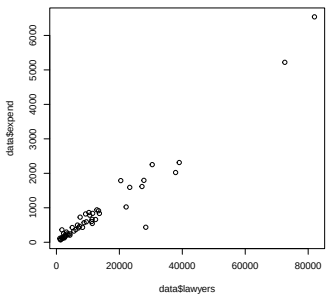
To investigate if these outliers are due to leverage points a scatter plot is made of the response variable expenditure against all individual explanatory variables. Also, the residuals of a linear regression model containing the corresponding explanatory variable is plotted.

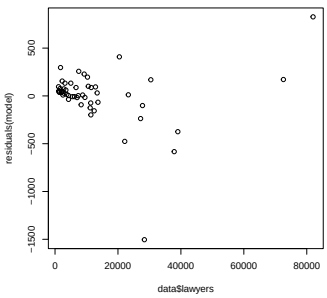

data$expend
data$bad

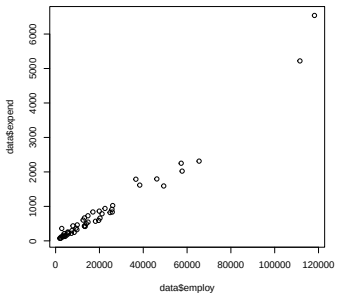

residuals(model)
data$bad

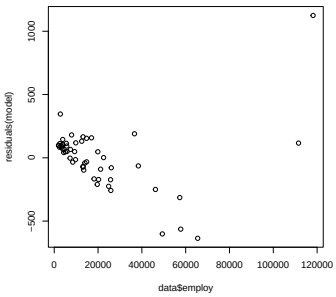

data$expend
data$crime

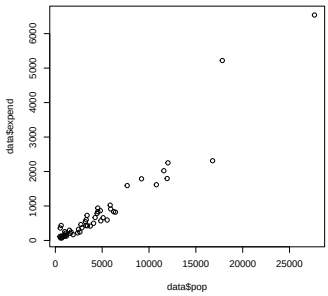

residuals(model)
data$crime

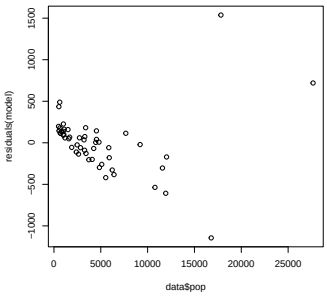

data$expend
data$lawyers


residuals(model)
data$lawyers


data$expend
data$employ


residuals(model)
data$employ


data$expend
data$pop


residuals(model)
data$pop

The outliers are most apparent in the residuals plots. Most points concentrate around zero, the points with high values appear on the right side of the residuals plot, which means they are likely due to a leverage point. To test if these points are influence points the Cook's distance is calculated.

$D_i = \frac{1}{(p+1)\hat{\sigma}^2} \sum_{j=1}^{n} (\hat{Y}_{(i),j} - \hat{Y}_j)^2.$

In words: the Cook's distance quantifes the influence of observation i on the prediction by calculating sum of squared differences between the predicted values with and without the i-th data point.

Collinearity

Excercise 3

Excercise 4