AdvStDaAn, Worksheet, Week 1

Micheal Lappert

31.03.2022

Contents

Exercise 1

Data Loading and Inspecting

```
path <- file.path('Datasets', 'Softdrink.dat')
df <- read.table(path, header=TRUE)

summary(df)
```

```
##       Time           volume         distance       location
##  Min.   : 8.00   Min.   : 2.00   Min.   : 10.8   Length:25
##  1st Qu.:13.75   1st Qu.: 4.00   1st Qu.: 45.0   Class :character
##  Median :18.11   Median : 7.00   Median : 99.0   Mode  :character
##  Mean   :22.38   Mean   : 8.76   Mean   :122.8
##  3rd Qu.:21.50   3rd Qu.:10.00   3rd Qu.:181.5
##  Max.   :79.24   Max.   :30.00   Max.   :438.0
```

```
head(df)
```

```
##    Time volume distance  location
## 1 16.68      7      168 San Diego
## 2 11.50      3       66 San Diego
## 3 12.03      3      102 San Diego
## 4 14.88      4       24 San Diego
## 5 13.75      6       45 San Diego
## 6 18.11      7       99 San Diego
```

```
tail(df)
```

```
##     Time volume distance    location
## 20 35.10     17    231.0      Austin
## 21 17.90     10     42.0      Austin
## 22 52.32     26    243.0      Austin
```

```
## 23 18.75      9    135.0      Austin
## 24 19.83      8    190.5 Minneapolis
## 25 10.75      4     45.0 Minneapolis
```

Data looks just fine.

Exercise 1.a)

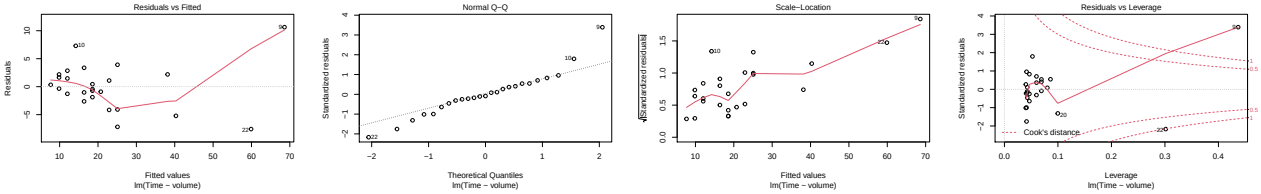
```
mod1 <- lm(Time ~ volume, data = df)
summary(mod1)
```

```
##
## Call:
## lm(formula = Time ~ volume, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5811 -1.8739 -0.3493  2.1807 10.6342
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.321      1.371   2.422   0.0237 *
## volume         2.176      0.124  17.546 8.22e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.181 on 23 degrees of freedom
## Multiple R-squared:  0.9305, Adjusted R-squared:  0.9275
## F-statistic: 307.8 on 1 and 23 DF,  p-value: 8.22e-15
```

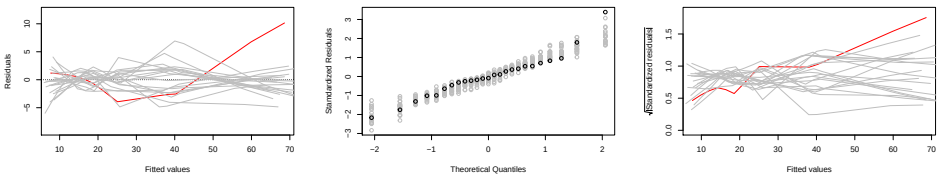
The model looks fine: - Volume is significant on the 5% niveau and the R-squared has a score of 0.93.

We have to do a residual and sensitivity analysis with stochastic simulation to investigate the correctness of the model.

```
plot(mod1)
```




```
plot.lmSim(mod1, SEED = 1)
```



Interpretation:

1. Tukey-Anscombe plot shows outlier with index i=9 which affects the smoother. In the simulation it is visible that the original curve is extreme. => The expected value of the residuals cannot be constant.
2. Scale-location plot shows a clear upwards trend. In the simulation it is visible that the original curve is extreme. => The scattering of the residuals is not constant.
3. q-q plot shows a slightly heavy tail and the outlier with index i=9 is again obvious. => Residuals are not normally distributed.

CONCLUSION: The fit is not satisfactory. Try transformations of response and explanatory variable.