Weiterbildungs-Lehrgang in Angewandter Statistik ETHZ 2009/2011 14.12.09

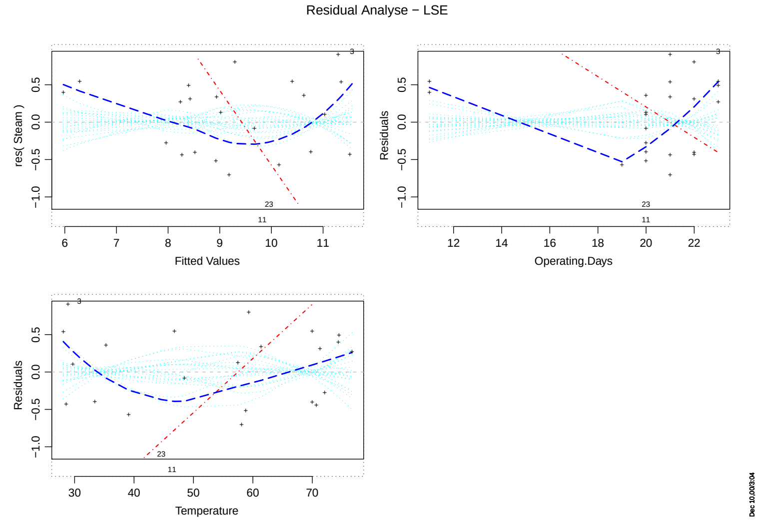
Angewandte Regression — Musterlösungen zur Serie 12

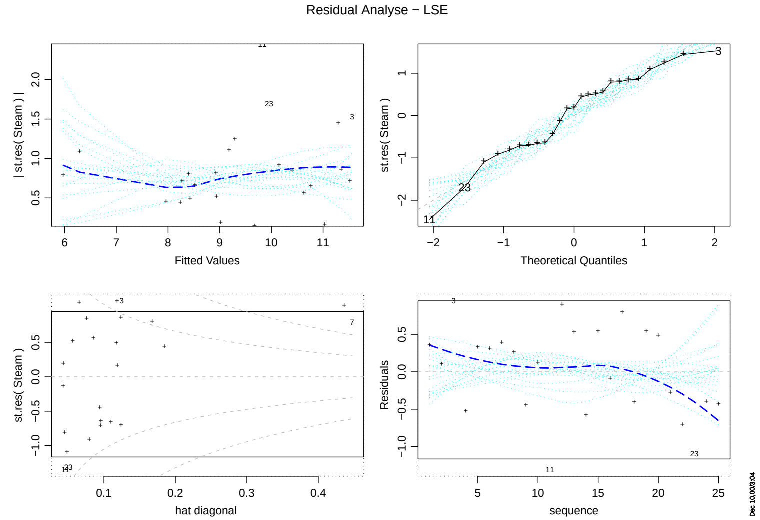
1. **a)**

```
r.steam.lse1 <- regr(Steam ~Operating.Days + Temperature, data=d.steam)
summary(r.steam.lse1)
r.steam.rob1 <- regr(Steam ~Operating.Days + Temperature, data=d.steam, method="rlm
summary(r.steam.rob1)

> summary(r.steam.lse1)
Call:
regr(formula = Steam ~ Operating.Days + Temperature, data = d.steam)
Fitting function  lm

Terms:
                  coef stcoef signif  R2.x df p.value
(Intercept)     9.1269  0.000   3.99    NA  1       0
Operating.Days  0.2028  0.375   2.14 0.022  1       0
Temperature    -0.0724 -0.767  -4.36 0.022  1       0

> summary(r.steam.rob1)
Call:
regr(formula = Steam ~ Operating.Days + Temperature, data = d.steam,
    method = "rlm")
Fitting function  rlm

Terms:
                  coef stcoef signif  R2.x df p.value
(Intercept)     3.0872  0.000   1.54    NA  1   0.004
Operating.Days  0.5147  0.953   6.17 0.022  1   0.000
Temperature    -0.0829 -0.877  -5.68 0.022  1   0.000
```

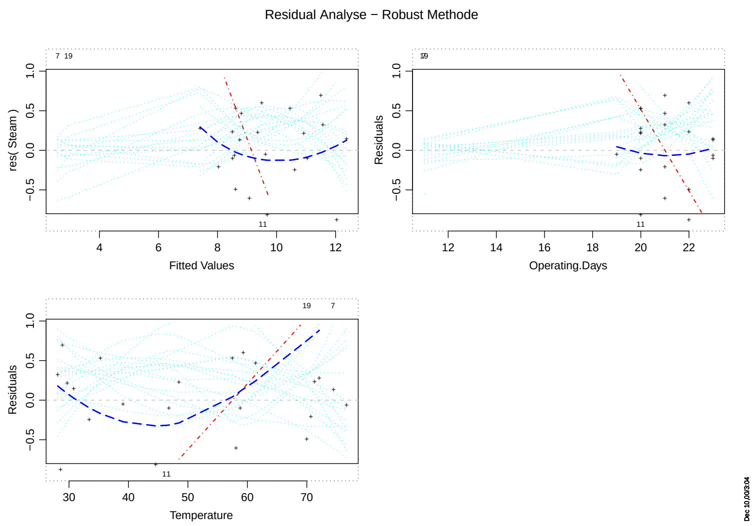
Beide Methoden zeigen hohe Signifikanz der Eingangsvariablen. Der Intercept und die Steigungen hingegen sind sehr verschieden.

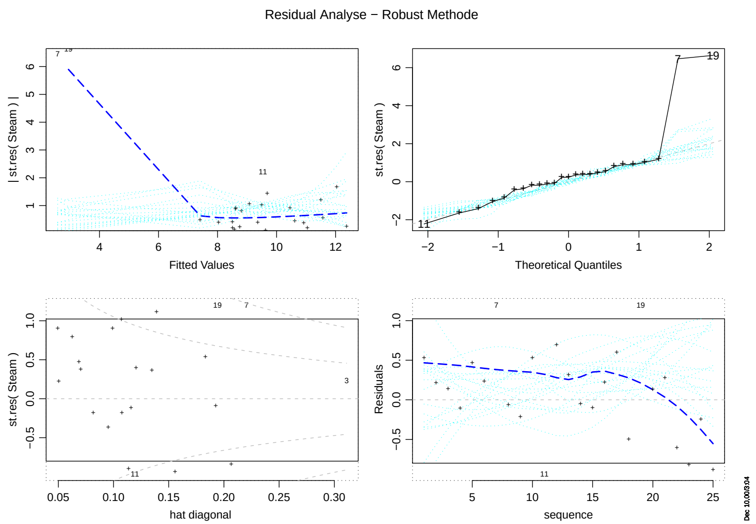
b) Residuenanalyse von der Least Square Methode: TA und die Variablen-Residuen-Plots zeigen sehr gekrümmtes Verhalten. Der QQ-Plot ist io. Die Varianz zeigt eine gute Konstanz. Die Residuen sind fast unabhgigig.





Residuenanalyse von der Robusten Methode: Der QQ-Plot zeigt zwei grosse Ausreisser. TA und die Variablen-Residuen-Plots zeigen ähnliches wie in der Least Square Methode, ausser dass die zwei Ausreisser nicht mitbetrachtet werden. Die konstante Varianz werden durch die beiden Ausreisser verletzt.





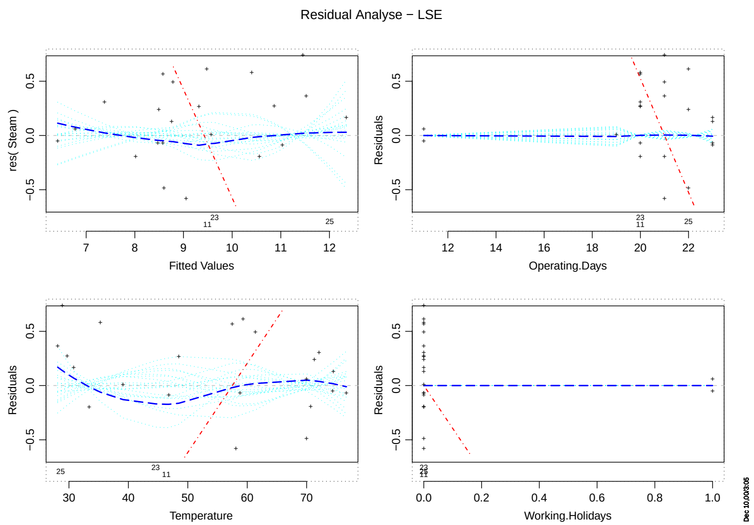
Ein Vergleich zeigt, dass die Residuenanalysen beider Methoden im wesentlichen Verschieden sind. Die Robuste Methode verschlechtert im allgemeinen die Annahmen und zeigt das Verhalten der beiden Ausreisser extremer. Die beiden Beobachtungen mit sehr kleiner Anzahl `Operating.Days` werden bei der robusten Regression als Ausreisser/Hebelpunkte erkannt und dementsprechend behandelt. Bei der gewöhnlichen Regression werden die beiden Hebelpunkte nicht als “bösartig” erkannt.

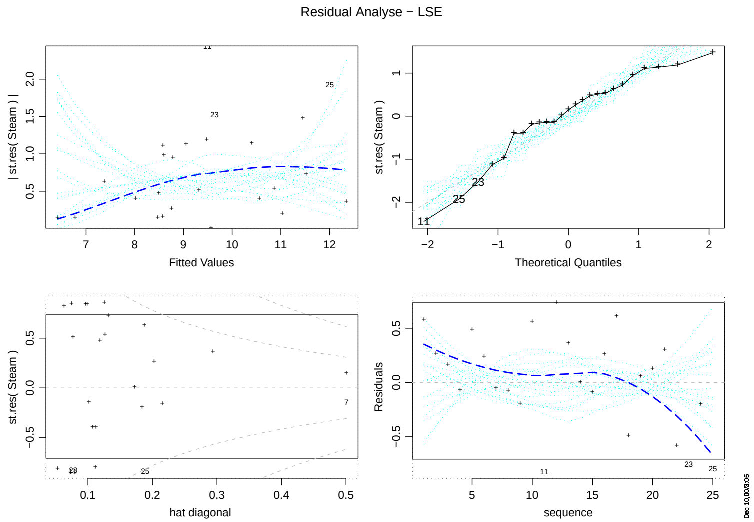
c)
```
r.steam.lse2 <- regr(Steam ~Operating.Days + Temperature + Working.Holidays, data=d
```



```
summary(r.steam.lse2)
r.steam.rob2 <- regr(Steam ~Operating.Days + Temperature + Working.Holidays, data=d
summary(r.steam.rob2)

> summary(r.steam.lse2)
Call:
regr(formula = Steam ~ Operating.Days + Temperature + Working.Holidays,
    data = d.steam)
Fitting function  lm

Terms:
                    coef  stcoef  signif  R2.x  df  p.value
(Intercept)       2.8122   0.000   0.686    NA   1    0.168
Operating.Days    0.5247   0.971   2.602 0.628   1    0.000
Temperature      -0.0822  -0.871  -5.646 0.100   1    0.000
Working.Holidays  3.9467   0.670   1.726 0.643   1    0.002

> summary(r.steam.rob2)
Call:
regr(formula = Steam ~ Operating.Days + Temperature + Working.Holidays,
    data = d.steam, method = "rlm")
Fitting function  rlm

Terms:
                    coef  stcoef  signif  R2.x  df  p.value
(Intercept)       3.2108   0.000   0.778    NA   1    0.121
Operating.Days    0.5107   0.945   2.516 0.628   1    0.000
Temperature      -0.0833  -0.882  -5.683 0.100   1    0.000
Working.Holidays  3.7816   0.642   1.643 0.643   1    0.001
```

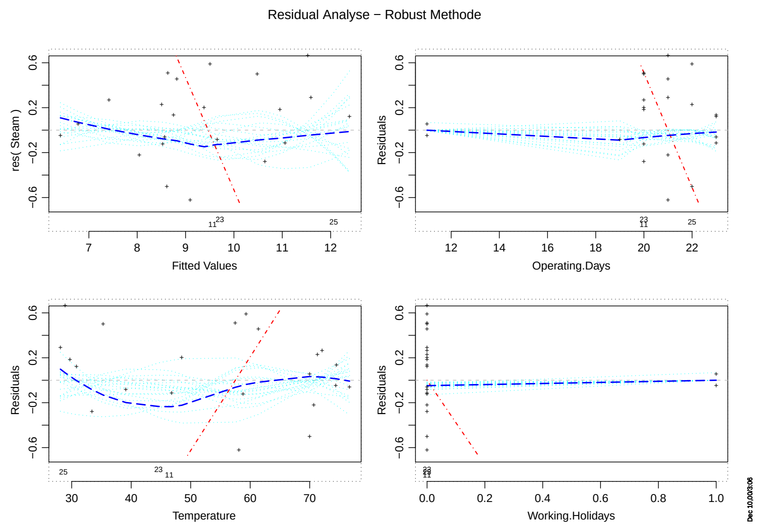
Alle Eingangsvariablen sind signifikant. Die Steigungen und die Intercepts stimmen besser über ein. Bemerke, dass die Koeffizienten des 2. Robusten Modells sich nicht viel vom 1. Robusten Modell unterschiedet.

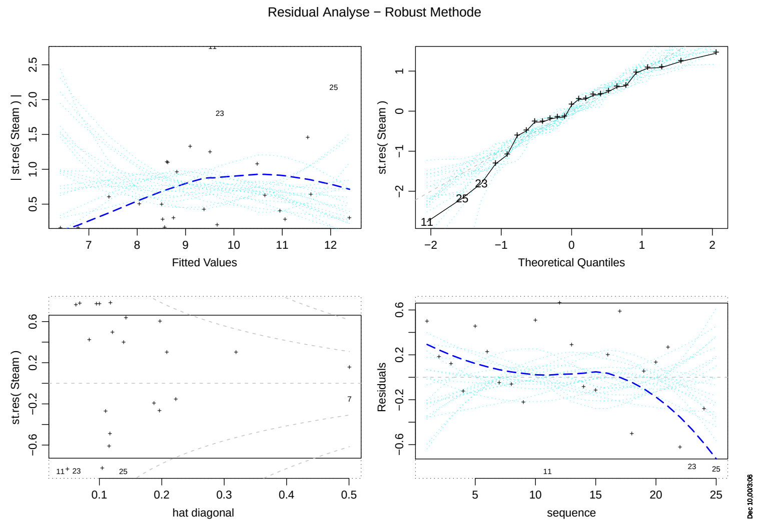




Residuenanalyse: Die Residuenanalyse beider Methoden sind sich gleich. TA und die Residuen-Variablen Plots sind io. Der QQ-Plot kann man als akzeptabel betrachten. Die Unabhängigkeit scheint leicht verletzt zu sein. Auch die konstante Varianz ist nich tso überzeugend, kann aber als akzeptabel betrachtet werden. In den QQ-Plots der beiden Anpassungen zeigen die Residuen aber immer noch eine ausgeprägte Linksschiefe. Einigen Beobachtungen müsste noch nachgegangen werden.
Aufgrund des Datensatzes muss auch mit korrelierten Residuen gerechnet werden. Die

Autokorrelationen und partiellen Autokorrelationen (vgl. Block: Zeitreihenanalyse) scheinen keine signifikanten zeitlichen Korrelationen aufzuweisen. Allerdings ist die Zeitreihe für eine sichere Aussage zu kurz.



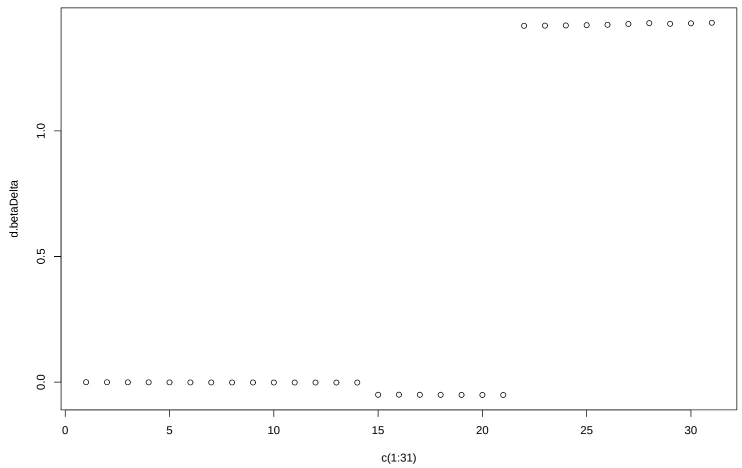


2. a)
```
d.asp <- read.table("ftp://stat.ethz.ch/Teaching/Datasets/WBL/asphalt.dat",
         header=TRUE)
r.asp <- regr(log10(RUT)~log10(VISC) + ASPH + VOIDS + RUN, d.asp, method="rlm")
```

b)
```
d.asp2 <- d.asp
d.asp2$VISC <- 10^6
d.asp2$RUT <- 10^6
```

c)
```
d.zusatz <- NULL
d.betaDelta <- NULL
for (i in 1:31){
  d.zusatz <- rbind(d.asp,d.asp2[1:i,])
  r.zusatz <- regr(formula(r.asp), d=d.zusatz, method="rlm")
  d.betaDelta <- cbind(d.betaDelta,r.zusatz$testcoef[2,1]-r.asp$testcoef[2,1])
}
```

d)
```
plot(c(1:31),d.betaDelta)
```



Wir sehen hier einen grossen Bruch bei 21.