Aufgabenblatt 1

(Teil 2)[1]

UK Erweiterungen des linearen Modells

Cordula Eggerth

Matrikelnummer: 00750881

Kursleiter:

Prof. Dr. Marcus Hudec &

Prof. Dr. Wilfried Grossmann

Sommersemester 2019

[1] Upload-Datenvolumen war begrenzt – daher musste ich 2 Teile anfertigen.

Aufgabe 3:

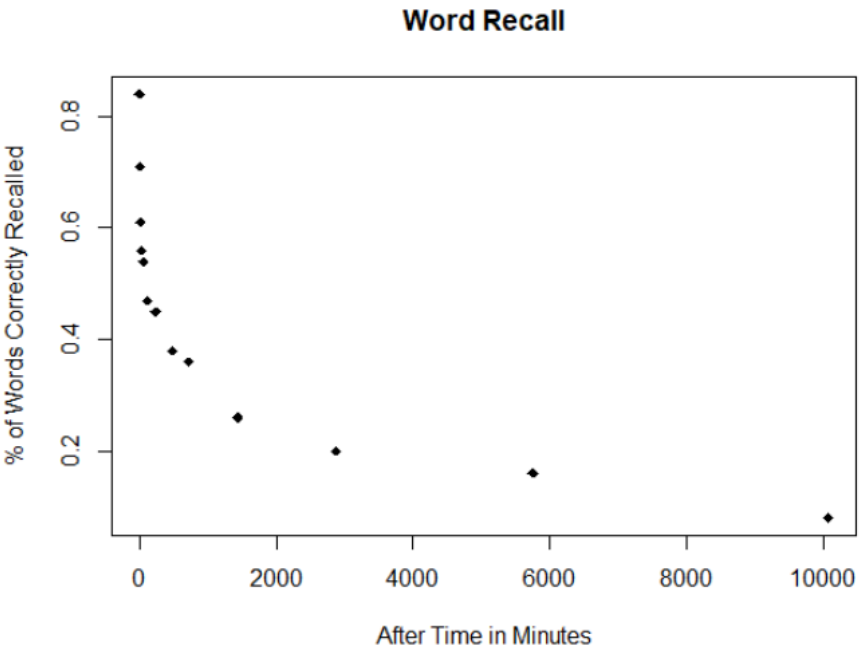
Im Excel-Sheet „*Some Datasets*“ finden Sie 5 kleine Datensätze. Führen Sie für die einzelnen Datensätze regressionsanalytische Auswertungen durch:

3.a.) *WordRecall*: Check for Linearity

Anmerkung aus der Angabe:

"Data stem from a memory retention experiment in which 13 subjects were asked to memorize a list of disconnected items. The subjects were then asked to recall the items at various times up to a week later. The proportion of items (y = prop) correctly recalled at various times (x = time, in minutes) since the list was memorized were recorded."

Bereits der Plot von x- und y-Variable zeigen, dass der Zusammenhang aller Ansicht nach nicht linear ist.



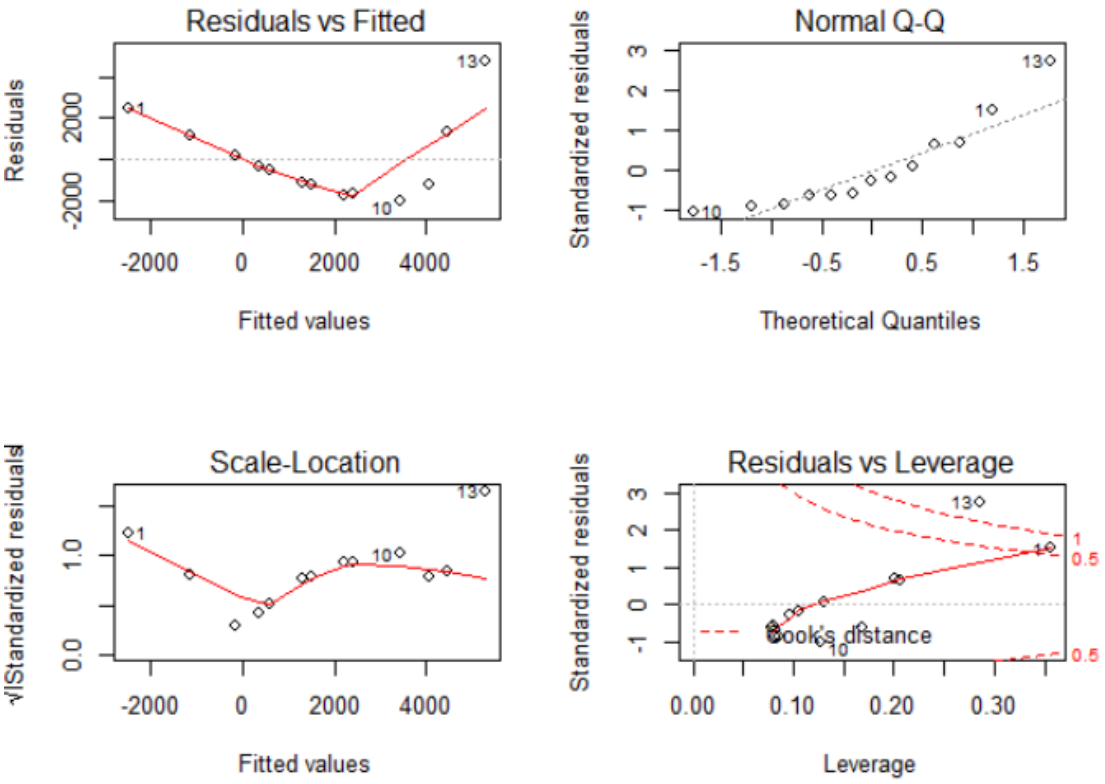
Die Korrelation zwischen *time* und *prop* ist außerdem stark negativ:

```
> cor(wordrecall_data$time, wordrecall_data$prop)
[1] -0.7555517
```

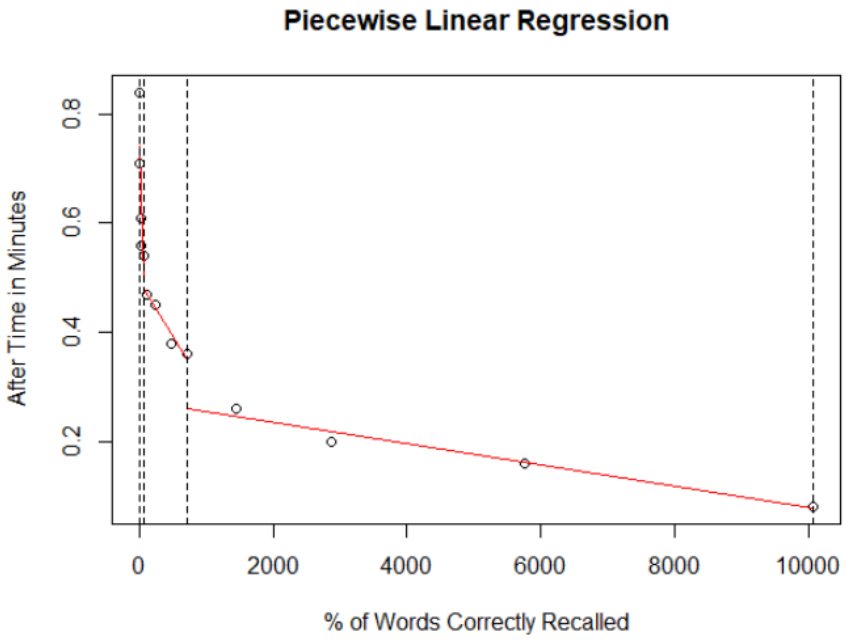
Laut linearer Regression mit lm() hat die Variable *prop* einen auf dem 0.01 Level signifikanten Einfluss auf die abhängige Variable:

```
Call:
lm(formula = wordrecall_data$time ~ wordrecall_data$prop)

Residuals:
    Min      1Q  Median      3Q     Max
-2004.8 -1258.0  -515.9  1170.9  4790.9

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)             6109       1292   4.729 0.000621 ***
wordrecall_data$prop   -10246       2678  -3.825 0.002817 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2065 on 11 degrees of freedom
Multiple R-squared:  0.5709,    Adjusted R-squared:  0.5318
F-statistic: 14.63 on 1 and 11 DF,  p-value: 0.002817
```

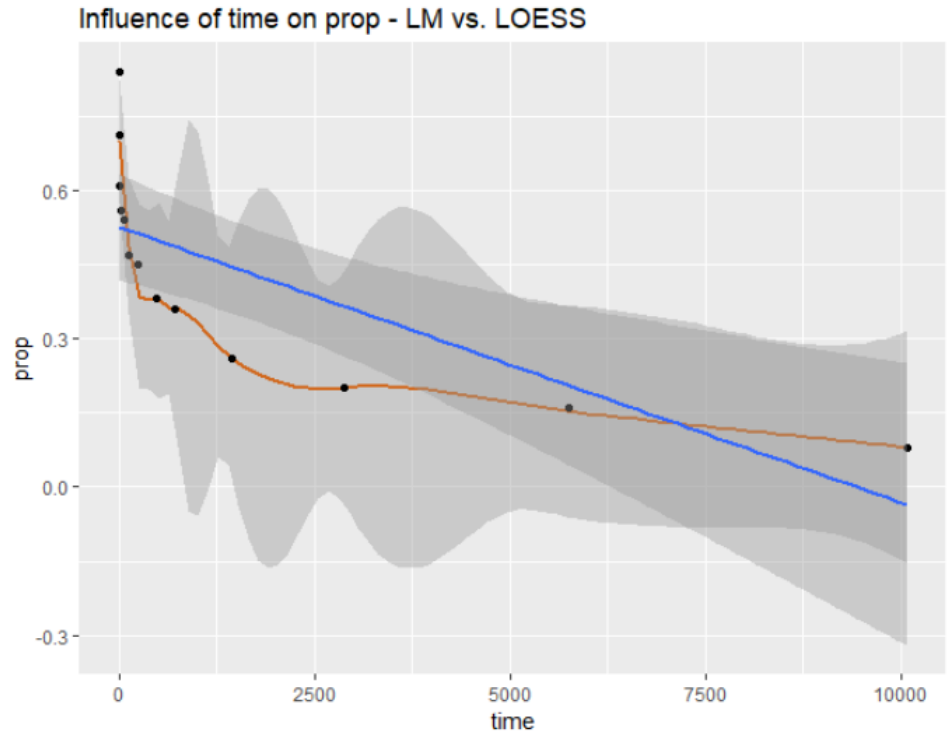
Wenn man die Diagnostic Plots begutachtet, sieht man, dass der Plot Residuals vs. Fitted einen nicht-linearen Zusammenhang andeutet, und der Plot Scale Location dies bestätigt, da die Residuen nicht zufällig als Punktewolke um die Nulllinie angeordnet sind:



Piecewise Linear Regression zeigt, dass die Steigung der Regressionsgeraden auf den einzelnen Intervallen sehr unterschiedlich ist, wodurch man sieht, dass kein linearer Zusammenhang vorliegt.

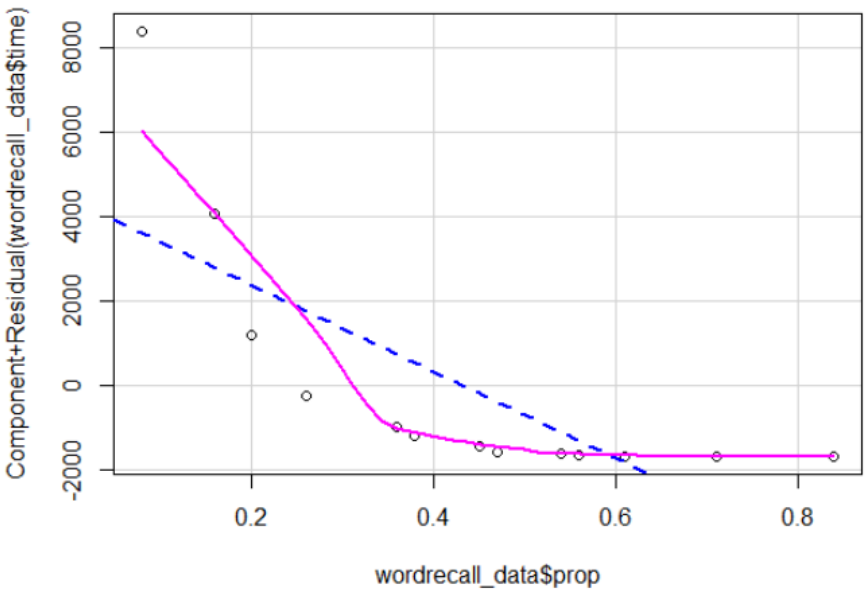


Auch der Vergleich des gewöhnlichen linearen Modells mit lokaler Regression (LOESS) ist sichtbar, dass die braune Linie (i.e. die LOESS-Linie) nicht linear ist und stark von der blauen Regressionsgerade abweicht. Daher kann man schließen, dass es keinen linearen Zusammenhang gibt.



Component+Residual Plot:

Drittens zeigt der C-R-Plot, dass die Geraden nicht übereinstimmen, was wiederum auf einen nicht linearen Zusammenhang hindeutet.

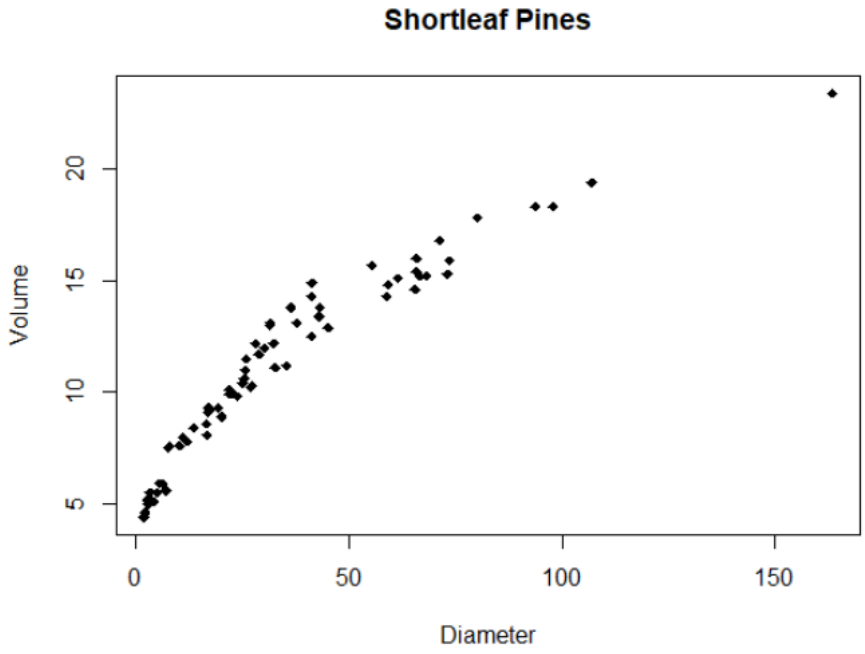


3.b.) *ShortLeaf*: Check for Linearity and Influential Observations

Anmerkung aus der Angabe:

"This is a classic data set — reported by C. Bruce and F. X. Schumacher in 1935 — concerning the diameter (x, in inches) and volume (y, in cubic feet) of n = 70 shortleaf pines. Data are used to predict the volume of the trees by means of the diameter."

Visualisierung des Datensatz:



Die Korrelation zwischen Volume und Diameter ist stark positiv:

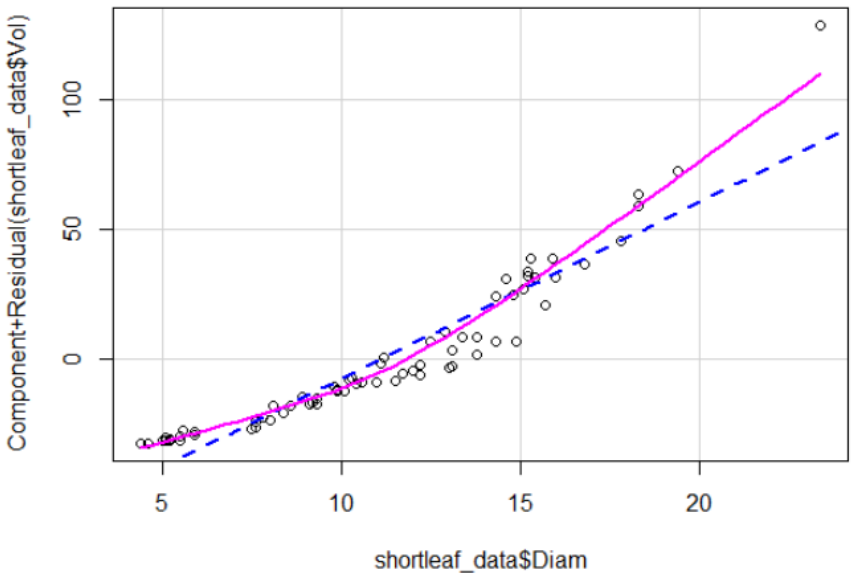
```
> cor(shortleaf_data$Vol, shortleaf_data$Diam)
[1] 0.9447509
```

Das gewöhnliche lineare Modell ergibt, dass der Regressor Diam signifikant ist auf dem 0.001 Level:

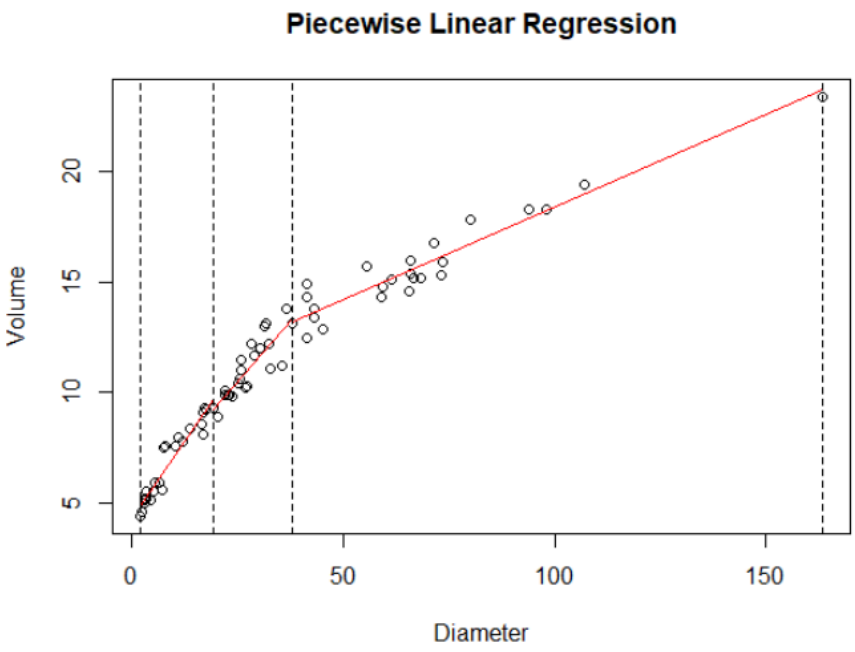
```
Call:
lm(formula = shortleaf_data$Vol ~ shortleaf_data$Diam)

Residuals:
    Min      1Q  Median      3Q     Max
-18.899  -4.768  -1.438   6.740  45.089

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -41.5681     3.4269  -12.13   <2e-16 ***
shortleaf_data$Diam   6.8367     0.2877   23.77   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.875 on 68 degrees of freedom
Multiple R-squared:  0.8926,    Adjusted R-squared:  0.891
F-statistic: 564.9 on 1 and 68 DF,  p-value: < 2.2e-16
```

Component+Residual Plot:

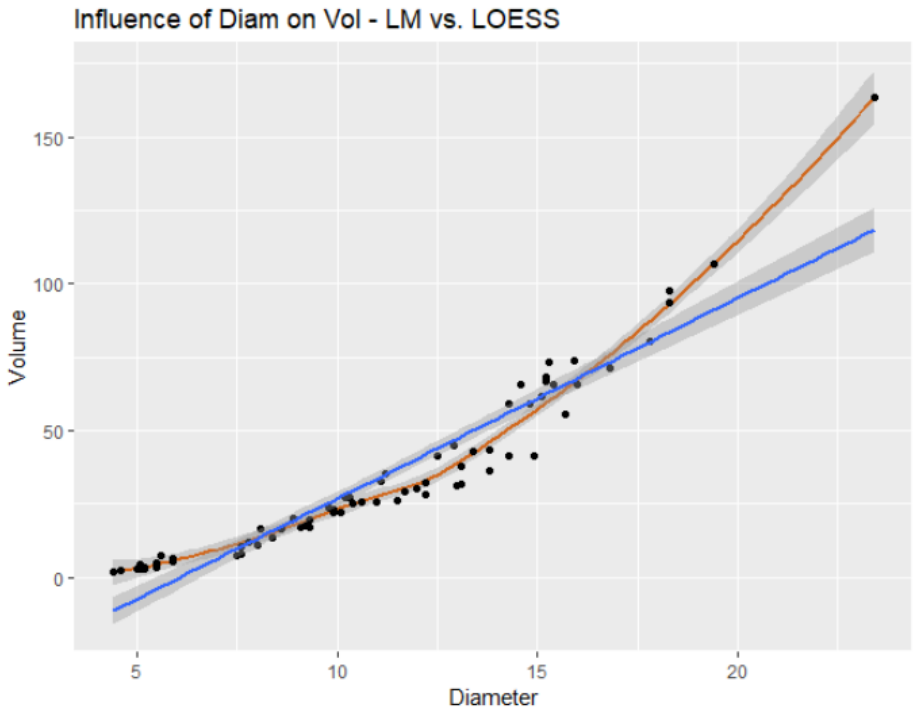
Der C-R-Plot zeigt, dass die Geraden nicht übereinstimmen, was wiederum auf einen nicht linearen Zusammenhang hindeutet.



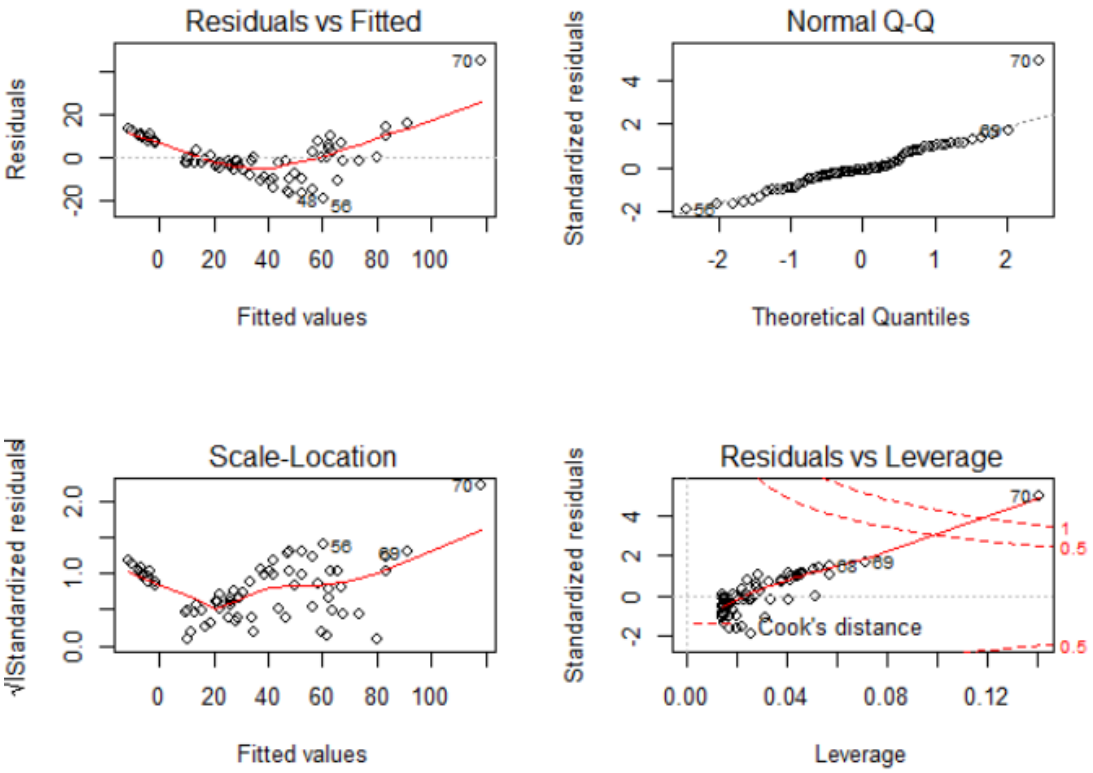
Piecewise Linear Regression zeigt, dass die Steigung der Regressionsgeraden auf den einzelnen Intervallen unterschiedlich ist, wodurch man sieht, dass kein linearer Zusammenhang vorliegt.



Auch der Vergleich des gewöhnlichen linearen Modells mit lokaler Regression (LOESS) ist sichtbar, dass die braune Linie (i.e. die LOESS-Linie) nicht linear ist und von der blauen Regressionsgerade abweicht. Daher kann man schließen, dass es keinen linearen Zusammenhang gibt.

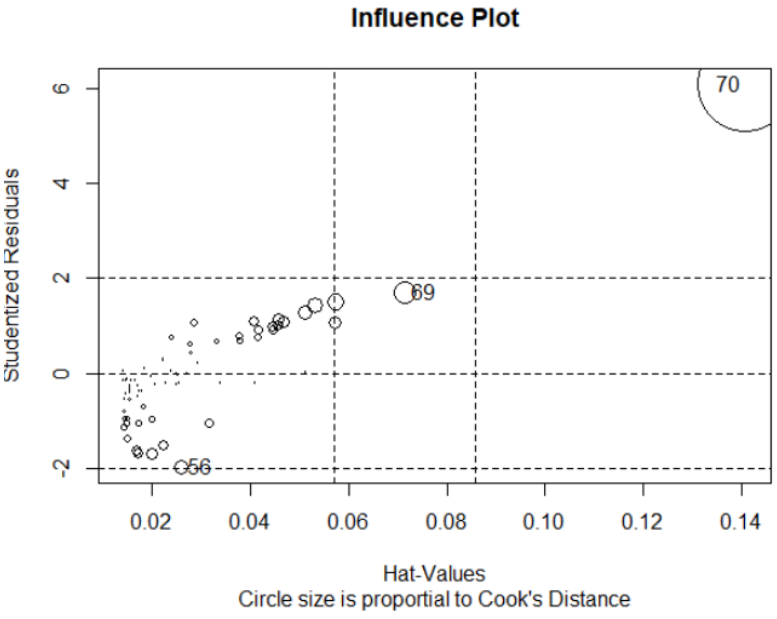


Die ***Regression Diagnostics Plots*** zu Residuals vs. Fitted und Sqrt(Standardized) Residuals vs. Fitted Values zeigen einen nicht linearen Zusammenhang. Gemäß Cook's Distance ist der Punkt 70 ein Influential Point und liegt außerhalb der Cook's Distance Linien.



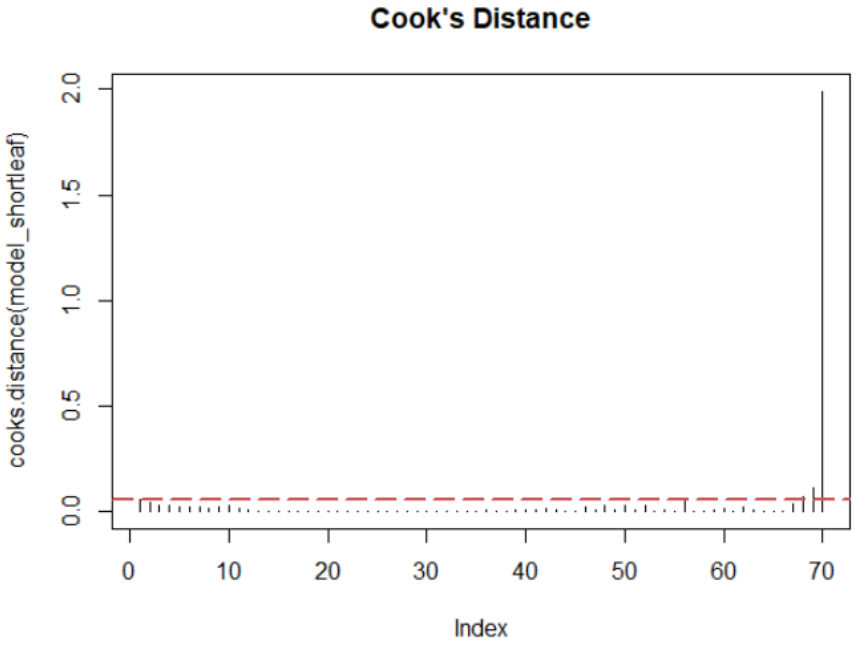
Influence Plot:

Der Influence Plot zeigt, dass die Punkte 70, 69 und 56 besonders hervorstechen und die Studentized Residuals beeinflussen.



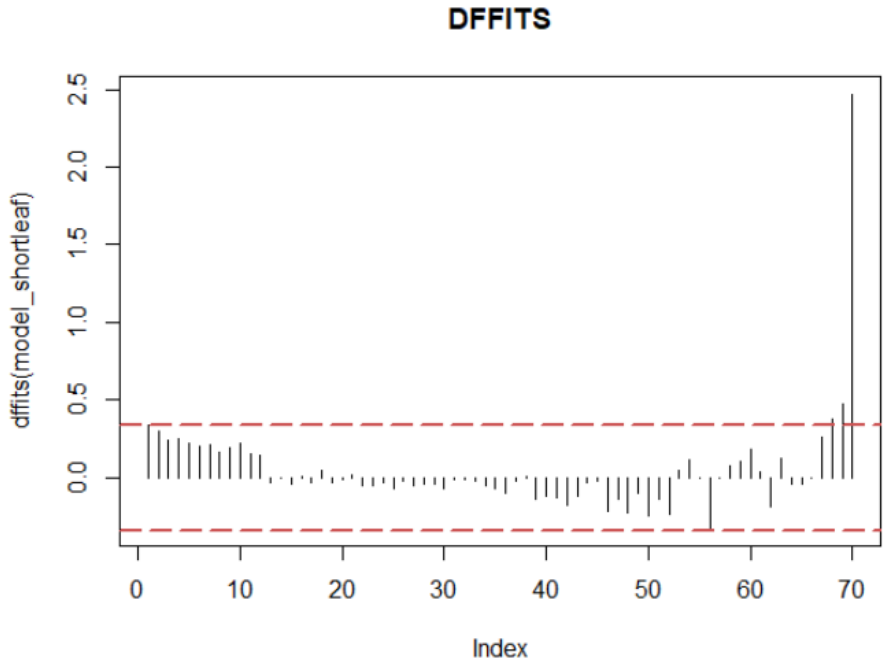
Cook's Distance Measure:

Durch Anwendung der Cook's Distance können rechts im untenstehenden Plot die einflussreichen Punkte ermittelt werden. Diese entsprechen auch u.a. den Punkten, die mittels Influence Plot oder DFFITS ermittelt werden. Hier sieht man ebenfalls, dass der Punkt 70 einflussreich ist, er im Plot stark über die strichlierte Linie hinausragt. Punkt 69 ragt leicht über die Linie hinaus.



DFFITS:

Bei DFFITS können einflussreiche Datenpunkte (hier: wieder Punkt 70 und 69) als jene Punkte, die außerhalb der strichlierten Linie liegen, festgestellt werden.



3.c.) *BirthWeight*: Use Indicator Variables

Anmerkung aus der Angabe:

"Researchers were interested in answering the research question if smoking behavior of the mother has an influence on the birth weight of a newborn child. They collected the following data (birthsmokers.txt) on a random sample of n = 32 births:

- Response (y): birth weight (Weight) in grams of baby
- Potential predictor (x1): Smoking status of mother (yes or no)
- Potential predictor (x2): length of gestation (Gest) in weeks."

Additives Modell mit Indikatorvariable:

Im additiven Modell mit Indikator für "smoke status" sind im Regressionsoutput die Variable Gestation und die Indikatorvariable I_smoke auf dem 0.001 Level signifikant. Der erklärte Anteil an der Gesamtvariabilität (R^2) liegt bei ca. 89.6%.

```
Call:
lm(formula = birthweight_data$Wgt ~ birthweight_data$Gest + I_smoke)

Residuals:
     Min       1Q   Median       3Q      Max
-223.693  -92.063   -9.365   79.663  197.507

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -2389.573    349.206  -6.843 1.63e-07 ***
birthweight_data$Gest  143.100      9.128  15.677 1.07e-15 ***
I_smoke                -244.544     41.982  -5.825 2.58e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 115.5 on 29 degrees of freedom
Multiple R-squared:  0.8964,    Adjusted R-squared:  0.8892
F-statistic: 125.4 on 2 and 29 DF,  p-value: 5.289e-15
```

Modell mit Interaktion:

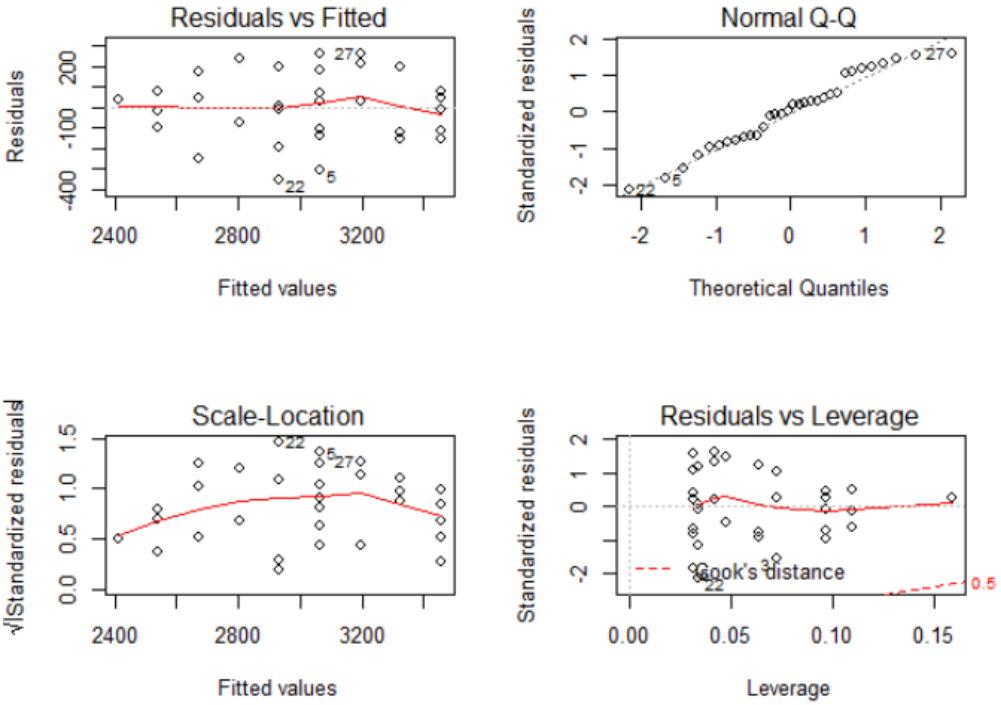
Im Modell mit Indikator für "smoke status" und Interaktion ist im Regressionsoutput die Variable Gestation auf dem 0.001 Level signifikant. Der erklärte Anteil an der Gesamtvariabilität (R^2) liegt bei ca. 89.7%, also in etwa gleich wie im additiven Modell zuvor.

```
Call:
lm(formula = birthweight_data$Wgt ~ birthweight_data$Gest * I_smoke)

Residuals:
     Min       1Q   Median       3Q      Max
-228.528  -89.560    0.273   83.629  184.529

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -2546.138    501.067  -5.081 2.22e-05 ***
birthweight_data$Gest          147.207     13.120  11.220 7.15e-12 ***
I_smoke                         71.574    716.950   0.100    0.921
birthweight_data$Gest:I_smoke   -8.178     18.515  -0.442    0.662
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 117.2 on 28 degrees of freedom
Multiple R-squared:  0.8971,    Adjusted R-squared:  0.8861
F-statistic: 81.37 on 3 and 28 DF,  p-value: 6.144e-14
```

Gesamtes Modell ohne Berücksichtigung des Indikators:

Im gesamten Modell liegt R^2, wenn man nur die Variable Gestation miteinbezieht, bei ca. 77.5%. Die Variable Gestation ist signifikant auf dem 0.001 Level.

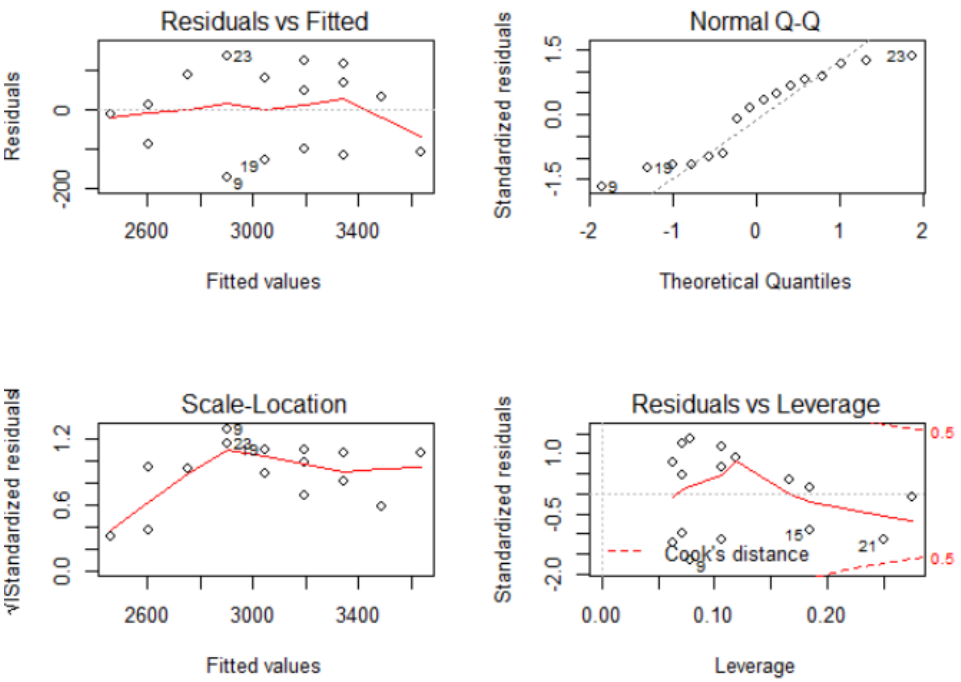
```
Call:
lm(formula = birthweight_data$Wgt ~ birthweight_data$Gest)

Residuals:
    Min      1Q  Median      3Q     Max
-354.03 -115.09   18.07  100.22  263.34

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -2037.00     498.11  -4.089 0.000298 ***
birthweight_data$Gest   130.82      12.86  10.170 3.09e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 167.3 on 30 degrees of freedom
Multiple R-squared:  0.7752,    Adjusted R-squared:  0.7677
F-statistic: 103.4 on 1 and 30 DF,  p-value: 3.085e-11
```



Teilmodell für die Gruppe der Nicht-Raucherinnen:

Die Variable Gestation ist auf dem 0.01 Level signifikant und R^2 liegt bei 91.5%.

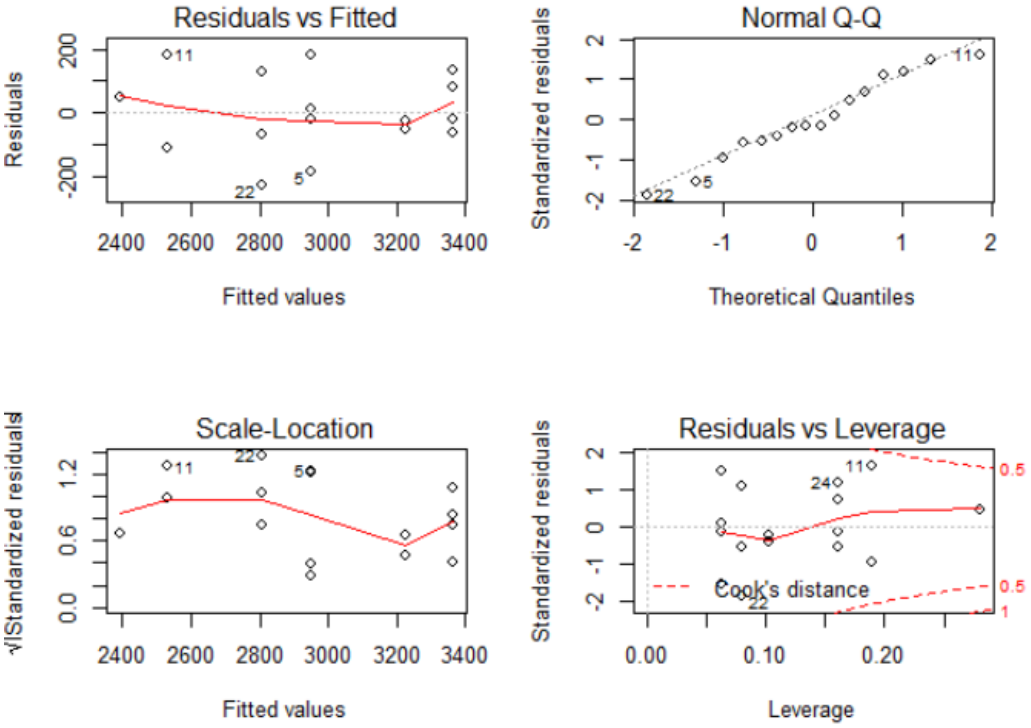
```
Call:
lm(formula = birthweight_data$Wgt ~ birthweight_data$Gest, subset = birthweight_data$Smoke ==
    "no")

Residuals:
    Min      1Q  Median      3Q     Max
-171.52 -101.59   23.28   83.63  139.48

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -2546.14     457.29  -5.568 6.93e-05 ***
birthweight_data$Gest   147.21      11.97  12.294 6.85e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 106.9 on 14 degrees of freedom
Multiple R-squared:  0.9152,    Adjusted R-squared:  0.9092
F-statistic: 151.1 on 1 and 14 DF,  p-value: 6.852e-09
```



Teilmodell für die Gruppe der Raucherinnen:

Die Variable Gestation ist auf dem 0.01 Level signifikant und R^2 liegt bei 87.4%.

```
Call:
lm(formula = birthweight_data$Wgt ~ birthweight_data$Gest, subset = birthweight_data$Smoke ==
    "yes")

Residuals:
    Min      1Q  Median      3Q     Max
-228.53  -64.86  -19.10   93.89  184.53

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -2474.56     553.97  -4.467 0.000532 ***
birthweight_data$Gest   139.03      14.11   9.851 1.12e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 126.6 on 14 degrees of freedom
Multiple R-squared:  0.8739,    Adjusted R-squared:  0.8649
F-statistic: 97.04 on 1 and 14 DF,  p-value: 1.125e-07
```



3.d.) *Anti-Depressiva*: Use Indicator Variables

Anmerkung aus der Angabe:

"Some researchers were interested in comparing the effectiveness of three treatments for severe depression. For the sake of simplicity, we denote the three treatments A, B, and C. The researchers collected the following data (depression.txt) on a random sample of n = 36 severely depressed individuals:

- y … measure of the effectiveness of the treatment for individual i
- possible predictor age (in years) of individual
- TRT the person has received."

Modell mit Indikatorvariablen (für TRT Variable; additiv):

In diesem Modell sind *age* und der Indikator für TRT signifikant und R^2 liegt bei 75.2%.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        36.0431     4.3678   8.252 1.57e-09 ***
antidepr_data$age   0.6659     0.0737   9.035 1.93e-10 ***
indikator_TRT      -5.1255     1.3020  -3.937 0.000403 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.374 on 33 degrees of freedom
Multiple R-squared:  0.7515,    Adjusted R-squared:  0.7364
F-statistic: 49.89 on 2 and 33 DF,  p-value: 1.057e-10
```

Modell mit Indikatorvariablen und Interaktion:

In diesem Modell sind die Interaktion zwischen *age* und Indikator und der Indikator für TRT signifikant und R^2 liegt bei 87.1%.

```
Call:
lm(formula = antidepr_data$y ~ antidepr_data$age * indikator_TRT)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8834 -2.1760 -0.3219  2.9215  8.5588

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     68.65977    6.77125  10.140 1.61e-11 ***
antidepr_data$age               -0.06779    0.14473  -0.468    0.643
indikator_TRT                  -20.71270    3.00840  -6.885 8.61e-08 ***
antidepr_data$age:indikator_TRT  0.35217    0.06448   5.461 5.19e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.657 on 32 degrees of freedom
Multiple R-squared:  0.8714,    Adjusted R-squared:  0.8593
F-statistic: 72.25 on 3 and 32 DF,  p-value: 2.434e-14
```

Gesamtes Modell:

In diesem Modell ist die Variable *age* signifikant und R^2 liegt bei 63.5%.

```
Call:
lm(formula = antidepr_data$y ~ antidepr_data$age)

Residuals:
     Min       1Q   Median       3Q      Max
-15.8916  -5.7463  -0.4105   4.7013  16.4607

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      25.33935    4.08258   6.207 4.65e-07 ***
antidepr_data$age  0.67619    0.08797   7.687 6.15e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.613 on 34 degrees of freedom
Multiple R-squared:  0.6347,    Adjusted R-squared:  0.624
F-statistic: 59.08 on 1 and 34 DF,  p-value: 6.155e-09
```

TRT Gruppe A:

In diesem Modell ist die Variable *age* signifikant und R^2 liegt bei 56.5%.

```
Call:
lm(formula = antidepr_data$y ~ antidepr_data$age, subset = antidepr_data$TRT ==
    "A")

Residuals:
    Min      1Q  Median      3Q     Max
-6.4223 -2.5643  0.4802  3.4463  6.3150

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       47.51559    4.30679  11.033 6.41e-07 ***
antidepr_data$age  0.33051    0.09175   3.602  0.00483 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.419 on 10 degrees of freedom
Multiple R-squared:  0.5648,    Adjusted R-squared:  0.5212
F-statistic: 12.98 on 1 and 10 DF,  p-value: 0.00483
```

TRT Gruppe B:

In diesem Modell ist die Variable *age* signifikant und R^2 liegt bei 79%.

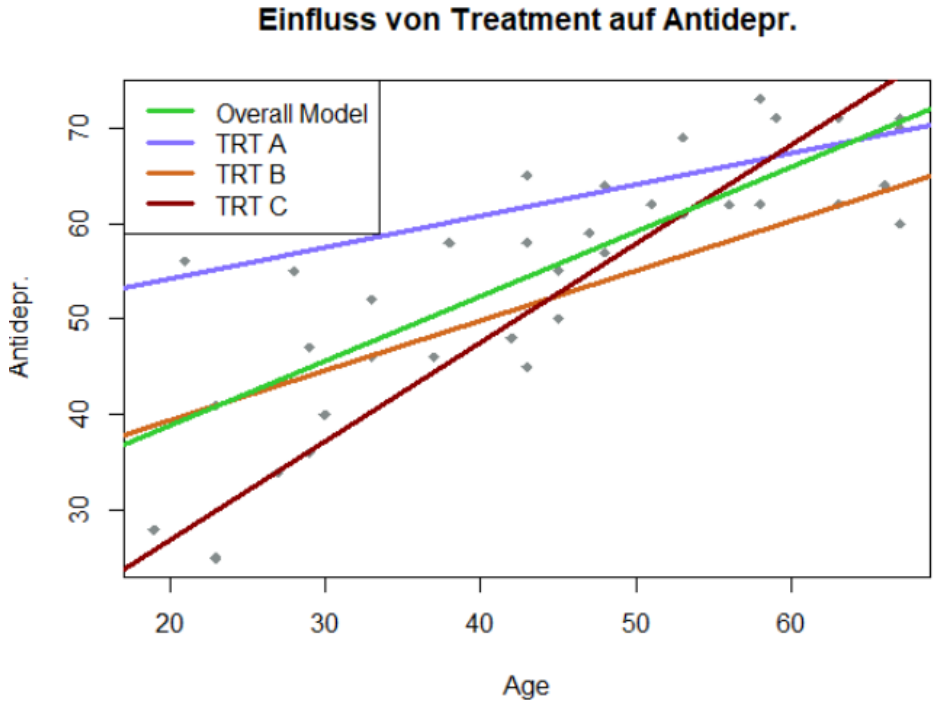
```
Call:
lm(formula = antidepr_data$y ~ antidepr_data$age, subset = antidepr_data$TRT ==
    "B")

Residuals:
    Min      1Q  Median      3Q     Max
-6.4366 -3.1860  0.2779  2.7548  6.5634

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       28.91821    3.92523   7.367  2.4e-05 ***
antidepr_data$age  0.52368    0.08539   6.133 0.000111 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.019 on 10 degrees of freedom
Multiple R-squared:   0.79,     Adjusted R-squared:  0.769
F-statistic: 37.61 on 1 and 10 DF,  p-value: 0.0001108
```

TRT Gruppe C:

In diesem Modell ist die Variable *age* signifikant und R^2 liegt bei 96.8%.

```
Call:
lm(formula = antidepr_data$y ~ antidepr_data$age, subset = antidepr_data$TRT ==
    "C")

Residuals:
    Min      1Q  Median      3Q     Max
-4.9794 -2.2394 -0.1463  2.3871  4.2192

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        6.21138    2.77048   2.242   0.0488 *
antidepr_data$age  1.03339    0.05982  17.275 8.94e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.246 on 10 degrees of freedom
Multiple R-squared:  0.9676,    Adjusted R-squared:  0.9643
F-statistic: 298.4 on 1 and 10 DF,  p-value: 8.94e-09
```

Überblick über die Regressionsgerade des Gesamtmodells und der einzelnen Gruppen:



R-Code zu Aufgabe 3:

```
#-----------------------------------------------------------------------
### AUFGABE 3 ### -------------------------------------------------------
#-----------------------------------------------------------------------
# Im Excel-Sheet „Some Datasets" finden Sie 5 kleine Datensätze.
# Führen Sie für die einzelnen Datensätze regressionsanalytische
# Auswertungen durch:

# a) WordRecall:      Check for Linearity

wordrecall_data <- read.xlsx(file=paste0(path,"/Some Datasets.xlsx"),
                             sheetName=1,
                             startRow=2, endRow=15,
                             colIndex=c(2:3),
                             colNames=TRUE, rowNames = FALSE)
wordrecall_data <- wordrecall_data[ ,-(3:4)]
wordrecall_data

plot(wordrecall_data$time, wordrecall_data$prop, pch=18,
     main="Word Recall", xlab="After Time in Minutes",
     ylab="% of Words Correctly Recalled")

cor(wordrecall_data$time, wordrecall_data$prop)

# lm
model_wordrecall <- lm(wordrecall_data$time ~ wordrecall_data$prop)
summary(model_wordrecall)
crPlots(model_wordrecall)

par(mfrow=c(2,2))
plot(model_wordrecall)

# piecewise linear regression
piecewise(wordrecall_data$time, wordrecall_data$prop, 3,
          xlab="% of Words Correctly Recalled",
          ylab="After Time in Minutes",
          main = "Piecewise Linear Regression")

# local regression (loess)
ggplot(wordrecall_data, aes(time, prop)) +
  stat_smooth(span=0.5,method=loess, col="chocolate") +
  geom_point() +
  ylab("prop") +
  xlab ("time") +
  ggtitle("Influence of time on prop - LM vs. LOESS") +
  stat_smooth(method=lm, se = TRUE)
```

```
# b) ShortLeaf: Check for Linearity and Influential Observations

shortleaf_data <- read.xlsx(file=paste0(path,"/Some Datasets.xlsx"),
                            sheetName=2,
                            startRow=1, endRow=71,
                            colIndex=c(1:2),
                            colNames=TRUE, rowNames = FALSE)
shortleaf_data <- shortleaf_data[ ,-(3:4)]
shortleaf_data

plot(shortleaf_data$Vol, shortleaf_data$Diam, pch=18,
     main="Shortleaf Pines", xlab="Diameter",
     ylab="Volume")

cor(shortleaf_data$Vol, shortleaf_data$Diam)

# lm
model_shortleaf <- lm(shortleaf_data$Vol ~ shortleaf_data$Diam)
summary(model_shortleaf)
crPlots(model_shortleaf)

# piecewise linear regression
piecewise(shortleaf_data$Vol, shortleaf_data$Diam, 3,
          xlab="Diameter",
          ylab="Volume",
          main = "Piecewise Linear Regression")

# local regression (loess)
ggplot(shortleaf_data, aes(Diam, Vol)) +
  stat_smooth(span=0.5,method=loess, col="chocolate") +
  geom_point() +
  ylab("Volume") +
  xlab ("Diameter") +
  ggtitle("Influence of Diam on Vol - LM vs. LOESS") +
  stat_smooth(method=lm, se = TRUE)

par(mfrow=c(2,2))
plot(model_shortleaf)
influence.measures(model_shortleaf)

par(mfrow=c(1,1))
influencePlot(model_shortleaf, id.method="identify", main="Influence Plot",
              sub="Circle size is proportial to Cook's Distance" )


# COOK'S DISTANCE MEASURE
plot(cooks.distance(model_shortleaf), type="h",
     main="Cook's Distance")
abline(h=4/length(influences), col="indianred3", lty=5, lwd=2)

# DDFITS
influences <- lm.influence(model_shortleaf)$hat
plot(dffits(model_shortleaf), type = "h", main = "DFFITS")
abline(h=2*sqrt(length(model_shortleaf$coef)/length(influences)),
       col="indianred3", lty=5, lwd=2)
abline(h=-2*sqrt(length(model_shortleaf$coef)/length(influences)),
       col="indianred3", lty=5, lwd=2)
```

```
# c) BirthWeight: Use Indicator Variables

birthweight_data <- read.xlsx(file=paste0(path,"/Some Datasets.xlsx"),
                              sheetName=3,
                              startRow=1, endRow=33,
                              colIndex=c(1:3),
                              colNames=TRUE, rowNames = FALSE)
birthweight_data <- birthweight_data[ ,-(4:5)]
birthweight_data

# additives modell mit indikator fuer "smoke status":
I_smoke <- ifelse(birthweight_data$Smoke == "yes", 1, 0)
model_I_smoke <- lm(birthweight_data$Wgt ~ birthweight_data$Gest + I_smoke)
summary(model_I_smoke)

# modell mit interaktion zwischen weight und indikator
model_interaktion_I <- lm(birthweight_data$Wgt ~ birthweight_data$Gest*I_smoke)
summary(model_interaktion_I)

# overall LM: (ohne beachtung von smoke status)
model_birthweight <- lm(birthweight_data$Wgt ~ birthweight_data$Gest)
summary(model_birthweight)
par(mfrow=c(2,2))
plot(model_birthweight)

# mit indikatorvariablen (fuer smoke status)
model_smokeYes <- lm(birthweight_data$Wgt ~ birthweight_data$Gest,
                     subset=birthweight_data$Smoke=="yes")
summary(model_smokeYes)
par(mfrow=c(2,2))
plot(model_smokeYes)

model_smokeNo <- lm(birthweight_data$Wgt ~ birthweight_data$Gest,
                    subset=birthweight_data$Smoke=="no")
summary(model_smokeNo)
par(mfrow=c(2,2))
plot(model_smokeNo)

par(mfrow=c(1,1))
plot(birthweight_data$Wgt ~ birthweight_data$Gest,
     pch=18, col="azure4",
     main="Einfluss von Smoke Status auf Weight",
     xlab="Gestation", ylab="Weight")
abline(model_birthweight, col="lightslateblue", lwd=3)
abline(model_smokeNo, col="chocolate", lwd=3)
abline(model_smokeYes, col="darkred", lwd=3)
legend("topleft", legend=c("Smoke: no", "Smoke: yes", "Overall"),
       col=c("chocolate", "darkred", "lightslateblue"), lwd=3)


# d) Anti-Depressiva: Use Indicator Variables

antidepr_data <- read.xlsx(file=paste0(path,"/Some Datasets.xlsx"),
                           sheetName=4,
                           startRow=1, endRow=37,
                           colIndex=c(1:3),
                           colNames=TRUE, rowNames = FALSE)
antidepr_data <- antidepr_data[ ,-(4:5)]
antidepr_data
```

```
# overall LM: (ohne beachtung von smoke status)
model_antidepr <- lm(antidepr_data$y ~ antidepr_data$age)
summary(model_antidepr)
par(mfrow=c(2,2))
plot(model_antidepr)

# mit indikatorvariablen (fuer smoke status)
model_antidepr_TRTA <- lm(antidepr_data$y ~ antidepr_data$age,
                          subset=antidepr_data$TRT=="A")
summary(model_antidepr_TRTA)
par(mfrow=c(2,2))
plot(model_antidepr_TRTA)

model_antidepr_TRTB <- lm(antidepr_data$y ~ antidepr_data$age,
                          subset=antidepr_data$TRT=="B")
summary(model_antidepr_TRTB)
par(mfrow=c(2,2))
plot(model_antidepr_TRTB)

# mit indikatorvariablen (fuer smoke status)
model_antidepr_TRTA <- lm(antidepr_data$y ~ antidepr_data$age,
                          subset=antidepr_data$TRT=="A")
summary(model_antidepr_TRTA)
par(mfrow=c(2,2))
plot(model_antidepr_TRTA)

model_antidepr_TRTB <- lm(antidepr_data$y ~ antidepr_data$age,
                          subset=antidepr_data$TRT=="B")
summary(model_antidepr_TRTB)
par(mfrow=c(2,2))
plot(model_antidepr_TRTB)

model_antidepr_TRTC <- lm(antidepr_data$y ~ antidepr_data$age,
                          subset=antidepr_data$TRT=="C")
summary(model_antidepr_TRTC)
par(mfrow=c(2,2))
plot(model_antidepr_TRTC)


par(mfrow=c(1,1))
plot(antidepr_data$y ~ antidepr_data$age,
     pch=18, col="azure4",
     main="Einfluss von Treatment auf Antidepr.",
     xlab="Age", ylab="Antidepr.")
abline(model_antidepr_TRTA, col="lightslateblue", lwd=3)
abline(model_antidepr_TRTB, col="chocolate", lwd=3)
abline(model_antidepr_TRTC, col="darkred", lwd=3)
abline(model_antidepr, col="limegreen", lwd=3)
legend("topleft", legend=c("Overall Model", "TRT A", "TRT B", "TRT C"),
       col=c("limegreen", "lightslateblue", "chocolate", "darkred"),
       lwd=3)

# indikatorvariable im additiven modell
indikator_TRT <- as.numeric(antidepr_data$TRT)
model_indik_TRT <- lm(antidepr_data$y ~ antidepr_data$age + indikator_TRT)
summary(model_indik_TRT)

# indikatorvariable im modell mit interaktion
model_indik_interaktion <- lm(antidepr_data$y ~ antidepr_data$age*indikator_TRT)
summary(model_indik_interaktion)
```

Literaturquellen:

- Folien und R-Codes zu den bisher vorgetragenen Kapiteln aus UK Erweiterungen des linearen Modells (Prof. Marcus Hudec).
- Kernel Regression Examples Using np (Jeffrey Racine, McMaster University Ontario (Canada), https://socialsciences.mcmaster.ca/racinej/Gallery/Regression.html).
- R Regression Diagnostics (Vik Paruchuri, DataQuest), http://www.vikparuchuri.com/blog/r-regression-diagnostics-part-1/).