Aufgabenblatt 2

UK Erweiterungen des linearen Modells

Cordula Eggerth

Matrikelnummer: 00750881

Kursleiter:

Prof. Dr. Marcus Hudec &

Prof. Dr. Wilfried Grossmann

Sommersemester 2019

Aufgabe 1:

Führen Sie mit dem Datensatz realestate.txt eine regressionsanalytische Modellierung durch. Evaluieren Sie die erzielte Vorhersage-Güte mittels des PRESS-criterion (prediction sum of squares – siehe Skriptum).

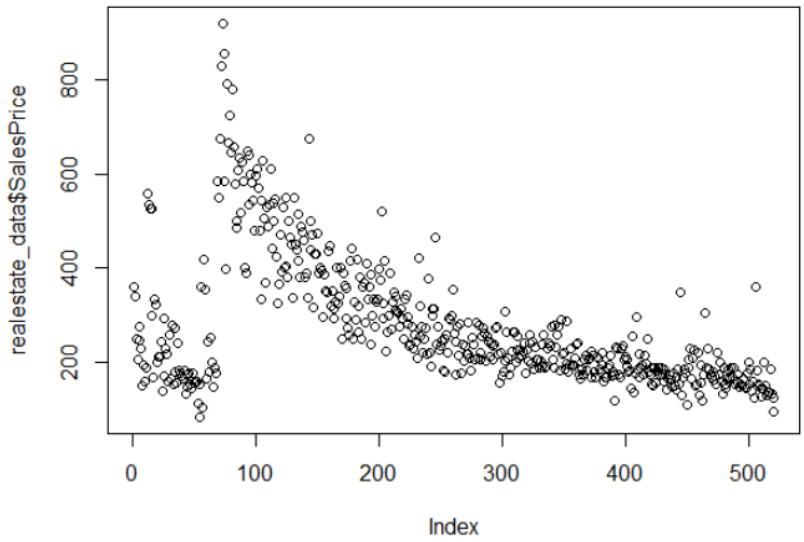
Im Rahmen dieser Aufgabe wird eine regressionsanalytische Modellierung mittels linearen Modellen und „Analysis of Variance“ vorgenommen, und die Vorhersagegüte der Modelle mittels PRESS-Kriterium ermittelt und verglichen.

Der Datensatz *realestate* enthält 521 Datenpunkte zu je 12 Variablen. Im Laufe der Analyse wird die abhängige Variable *SalesPrice* durch die verbleibenden Variablen erklärt. Untenstehend befindet sich ein Ausschnitt aus dem Datensatz:

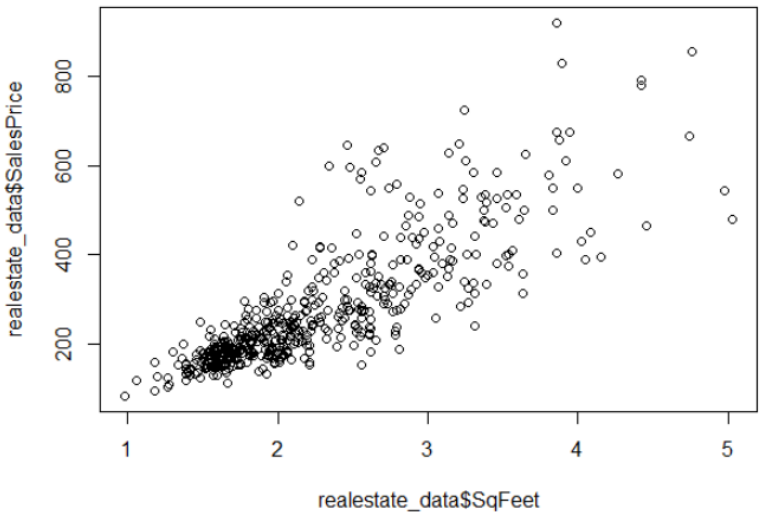
```
head(realestate_data, n=20)
```

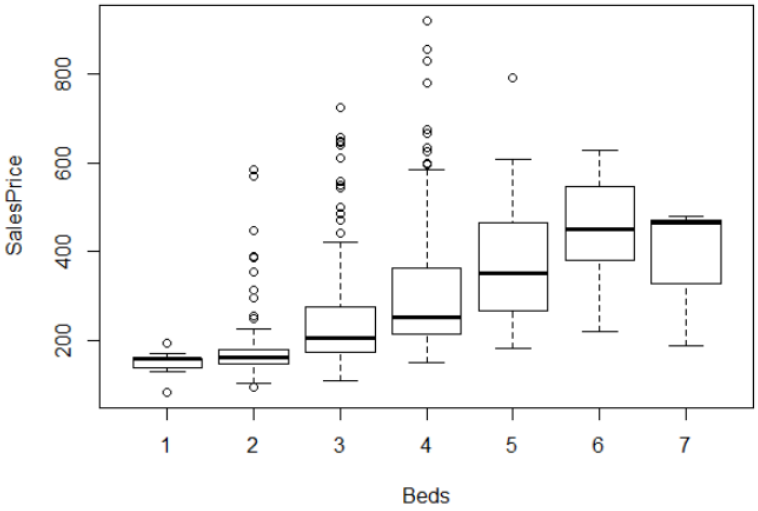
V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
SalesPrice	SqFeet	Beds	Baths	AirCond	Garage	Pool	Year	Quality	Style	Lot	Highway
360	3.032	4	4	1	2	0	1972	2	1	22.221	0
340	2.058	4	2	1	2	0	1976	2	1	22.912	0
250	1.78	4	3	1	2	0	1980	2	1	21.345	0
205.5	1.638	4	2	1	2	0	1963	2	1	17.342	0
275.5	2.196	4	3	1	2	0	1968	2	7	21.786	0
248	1.966	4	3	1	5	1	1972	2	1	18.902	0
229.9	2.216	3	2	1	2	0	1972	2	7	18.639	0
150	1.597	2	1	1	1	0	1955	2	1	22.112	0
195	1.622	3	2	1	2	0	1975	3	1	14.321	0
160	1.976	3	3	0	1	0	1918	3	1	32.358	0

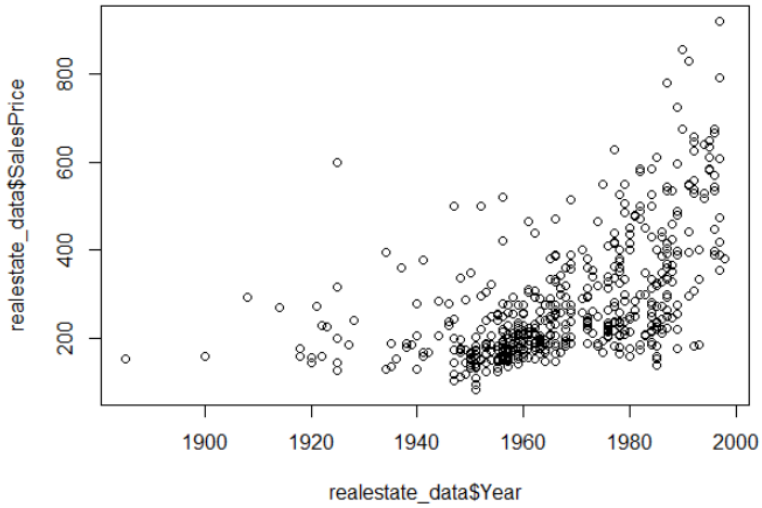
Die Variable *SalesPrice* in ihrer Verteilung in der Reihenfolge des Vorkommens im Datensatz:

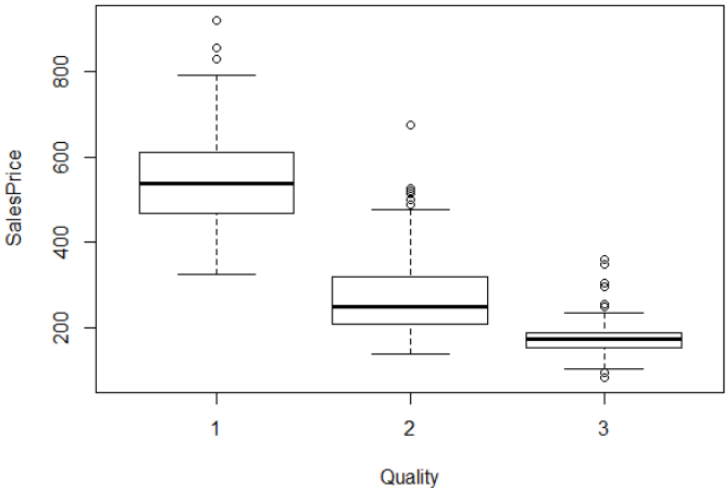


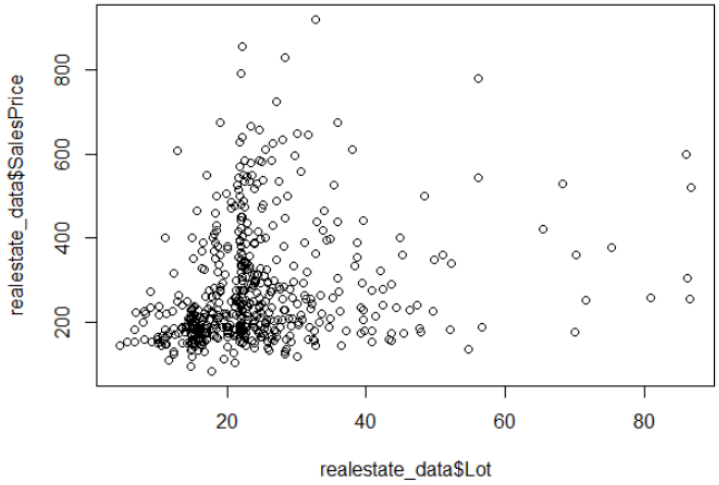
Es folgen weitere deskriptive Plots zur Darstellung der abhängigen Variable *SalesPrice* durch die verbleibenden Variablen:


realestate_data$SalesPrice
200
400
600
800
1
2
3
4
5
realestate_data$SqFeet


SalesPrice
200
400
600
800
1
2
3
4
5
6
7
Beds


realestate_data$SalesPrice
200
400
600
800
1900
1920
1940
1960
1980
2000
realestate_data$Year


SalesPrice
200
400
600
800
1
2
3
Quality


realestate_data$SalesPrice
200
400
600
800
20
40
60
80
realestate_data$Lot

Modell 1:

- Multiple lineare Regression
- Additives Modell mit allen Regressoren
- d.h. Testen der Variablen in Anwesenheit der anderen Variablen
- Multiple R-squared liegt in diesem Modell bei 79.22%. Die Variablen SqFeet, Year, Quality, Style, Lot sind hierbei auf dem 0.001 Alpha-Level signifikant.
- Modell 1 hat ein PRESS-Kriterium als Maß der Vorhersagegüte von 2180184. Ziel der Verwendung des PRESS-Kriteriums ist es, mehrere Modelle zu vergleichen, wobei jenes mit dem niedrigsten Wert das beste im untersuchten Kontext ist. Die „Leave-One-Out"-Strategie wird gewählt, um Overfitting zu vermeiden.

```
Call:
lm(formula = SalesPrice ~ SqFeet + Beds + Baths + AirCond + Garage +
    Pool + Year + Quality + Style + Lot + Highway)

Residuals:
    Min      1Q  Median      3Q     Max
-186.25  -37.55   -2.36   31.70  293.79

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2351.8910   433.9996  -5.419 9.26e-08 ***
SqFeet        129.7452     7.6675  16.921  < 2e-16 ***
Beds           -8.2357     3.5350  -2.330   0.0202 *
Baths           4.7172     4.6687   1.010   0.3128
AirCond       -13.5533     8.5834  -1.579   0.1150
Garage         13.5320     5.4562   2.480   0.0135 *
Pool            8.8900    11.2356   0.791   0.4292
Year            1.2410     0.2188   5.671 2.38e-08 ***
Quality       -46.6914     7.4423  -6.274 7.54e-10 ***
Style          -9.4818     1.4305  -6.628 8.66e-11 ***
Lot             1.1536     0.2579   4.473 9.51e-06 ***
Highway       -37.4620    19.6072  -1.911   0.0566 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 63.41 on 509 degrees of freedom
Multiple R-squared:  0.7922,    Adjusted R-squared:  0.7877
F-statistic: 176.4 on 11 and 509 DF,  p-value: < 2.2e-16
```

Die Beiträge der Variablen können auch sequenziell beurteilt werden mittels der Funktion *anova*, wobei aber der Nachteil entsteht, dass die p-Werte abhängig von der Reihenfolge der Variablen sind:

```
Analysis of Variance Table

Response: SalesPrice
           Df  Sum Sq Mean Sq   F value    Pr(>F)
SqFeet      1 6666447 6666447 1658.0875 < 2.2e-16 ***
Beds        1   14478   14478    3.6011  0.058307 .
Baths       1  166243  166243   41.3482 2.938e-10 ***
AirCond     1   27206   27206    6.7666  0.009558 **
Garage      1  179440  179440   44.6305 6.254e-11 ***
Pool        1      99      99    0.0245  0.875662
Year        1  216490  216490   53.8456 8.628e-13 ***
Quality     1  232916  232916   57.9311 1.327e-13 ***
Style       1  208099  208099   51.7587 2.259e-12 ***
Lot         1   75302   75302   18.7292 1.814e-05 ***
Highway     1   14677   14677    3.6505  0.056614 .
Residuals 509 2046467    4021
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mittels *Anova* kann der Beitrag der Variablen unabhängig von der Reihenfolge des Vorkommens in der Regression beurteilt werden:[1]

```
Anova Table (Type II tests)

Response: SalesPrice
           Sum Sq  Df  F value    Pr(>F)
SqFeet    1151216   1 286.3319 < 2.2e-16 ***
Beds        21822   1   5.4277   0.02021 *
Baths        4105   1   1.0209   0.31279
AirCond     10025   1   2.4933   0.11495
Garage      24730   1   6.1509   0.01346 *
Pool         2517   1   0.6261   0.42917
Year       129318   1  32.1640 2.378e-08 ***
Quality    158253   1  39.3608 7.538e-10 ***
Style      176650   1  43.9367 8.664e-11 ***
Lot         80454   1  20.0106 9.507e-06 ***
Highway     14677   1   3.6505   0.05661 .
Residuals 2046467 509
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Das PRESS-Kriterium wurde auf zwei Arten ermittelt.

<u>Methode 1:</u>

(vereinfachte Formel mit Hilfe der Diagonalelemente der Hat-Matrix aus dem Foliensatz „Multiple Regression“ (S. 36) für lineare Modelle)

```
PRESScriterion <- function(linmod) {
  predictiveResid <- residuals(linmod)/(1 - lm.influence(linmod)$hat)
  sum(predictiveResid^2) # PRESS formula form slide 36
}

> PRESS_lm1_method1
[1] 2180184
```

<u>Methode 2:</u>

[1] Anmerkung: Die Ergebnisse von *anova* und *Anova* sind nur dieselben, wenn alle Variablen zueinander orthogonal stehen. Im vorliegenden Fall sind die Ergebnisse unterschiedlich.

(Verwendung des Package *qpcR*)

```
PRESS_lm1_method2 <- PRESS(lm1)$stat
```

```
> PRESS_lm1_method2
[1] 2180184
```

Beide Berechnungsmethoden ergeben schließlich dasselbe PRESS-Kriterium für das betrachtete Modell – die Überprüfung ist daher geglückt.

Modell 2:

- Multiple lineare Regression
- Modell mit allen Regressoren und mit Interaktion
- Multiple R-squared liegt in diesem Modell bei 82.14%. Die Variablen SqFeet, Year, Quality, Style, Lot und die Interaktionen SqFeet:Beds sowie SqFeet:Year sind hierbei auf dem 0.001 Alpha-Level signifikant.
- Modell 2 hat ein PRESS-Kriterium als Maß der Vorhersagegüte von 1968112. Dieses Modell ist bezüglich PRESS also besser als Modell 1.

```
Call:
lm(formula = SalesPrice ~ SqFeet + Beds + Baths + AirCond + Garage +
    Pool + Year + Quality + Style + Lot + Highway + SqFeet:Beds +
    SqFeet:Year + SqFeet:Pool + Quality:Style + Beds:Pool)

Residuals:
     Min       1Q   Median       3Q      Max
-161.341  -35.091    0.368   27.201  235.535

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    7303.0860  1197.7774   6.097 2.15e-09 ***
SqFeet        -3952.1133   475.3625  -8.314 8.67e-16 ***
Beds             29.3936     9.1257   3.221 0.001360 **
Baths             4.8009     4.3913   1.093 0.274789
AirCond          -5.9893     8.1266  -0.737 0.461463
Garage            9.1684     5.1496   1.780 0.075612 .
Pool             75.6688    42.6239   1.775 0.076457 .
Year             -3.7211     0.6110  -6.090 2.24e-09 ***
Quality         -55.2365     8.7254  -6.331 5.41e-10 ***
Style           -15.7215     3.9919  -3.938 9.36e-05 ***
Lot               1.1590     0.2428   4.773 2.38e-06 ***
Highway         -27.8175    18.3196  -1.518 0.129527
SqFeet:Beds     -13.9417     3.5924  -3.881 0.000118 ***
SqFeet:Year       2.0971     0.2427   8.641  < 2e-16 ***
SqFeet:Pool      -3.8040    15.4329  -0.246 0.805405
Quality:Style     3.2488     1.8367   1.769 0.077525 .
Beds:Pool       -12.4189    10.8623  -1.143 0.253455
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.07 on 504 degrees of freedom
Multiple R-squared:  0.8214,    Adjusted R-squared:  0.8157
F-statistic: 144.9 on 16 and 504 DF,  p-value: < 2.2e-16
```

```
> PRESS_lm2_method1
[1] 1968112
> PRESS_lm2_method2
[1] 1968112
```

Modell 3:

- Multiple lineare Regression
- Modell mit weniger Regressoren und mit Interaktion
- Multiple R-squared liegt in diesem Modell bei 81.69%. Die Variablen wurden auf jene reduziert, die auf dem 0.001 Alpha-Level signifikant sind.
- Modell 3 hat ein PRESS-Kriterium als Maß der Vorhersagegüte von 1927748. Dieses Modell ist bezüglich PRESS also besser als Modell 1 und 2.

```
Call:
lm(formula = SalesPrice ~ SqFeet + Beds + Year + Quality + Style +
    Lot + SqFeet:Beds + SqFeet:Year + Quality:Style)

Residuals:
     Min       1Q   Median       3Q      Max
-171.604  -34.455   -0.615   27.751  231.538

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    7411.5992  1173.3834   6.316 5.83e-10 ***
SqFeet        -4016.8594   470.2097  -8.543  < 2e-16 ***
Beds             34.2492     8.7632   3.908 0.000105 ***
Year             -3.7734     0.5977  -6.313 5.93e-10 ***
Quality         -59.5600     8.4643  -7.037 6.37e-12 ***
Style           -15.7839     3.9888  -3.957 8.66e-05 ***
Lot               1.1552     0.2369   4.877 1.44e-06 ***
SqFeet:Beds     -16.0141     3.4069  -4.700 3.34e-06 ***
SqFeet:Year       2.1371     0.2400   8.904  < 2e-16 ***
Quality:Style     3.2966     1.8322   1.799 0.072572 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.41 on 511 degrees of freedom
Multiple R-squared:  0.8169,    Adjusted R-squared:  0.8136
F-statistic: 253.2 on 9 and 511 DF,  p-value: < 2.2e-16
```

```
> PRESS_lm3_method1
[1] 1927748
> PRESS_lm3_method2
[1] 1927748
```

Modell 4:

- Multiple lineare Regression
- Modell mit noch weniger Regressoren und mit Interaktion
- Multiple R-squared liegt in diesem Modell bei 80.64%.
- Modell 4 hat ein PRESS-Kriterium als Maß der Vorhersagegüte von 2003169. Dieses Modell ist bezüglich PRESS also schlechter als Modell 3.

```
Call:
lm(formula = SalesPrice ~ SqFeet + Beds + Year + Quality + Style +
    SqFeet:Beds + SqFeet:Year)

Residuals:
     Min       1Q   Median       3Q      Max
-184.456  -36.910   -0.812   28.491  270.409

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7464.3354  1176.8218   6.343 4.96e-10 ***
SqFeet      -3814.5160   469.5500  -8.124 3.39e-15 ***
Beds           42.0945     8.7000   4.838 1.73e-06 ***
Year           -3.8110     0.6014  -6.337 5.14e-10 ***
Quality       -52.9145     6.8421  -7.734 5.57e-14 ***
Style         -10.0263     1.3456  -7.451 3.96e-13 ***
SqFeet:Beds   -18.6503     3.3963  -5.491 6.29e-08 ***
SqFeet:Year     2.0416     0.2400   8.505  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.96 on 513 degrees of freedom
Multiple R-squared:  0.8064,    Adjusted R-squared:  0.8038
F-statistic: 305.3 on 7 and 513 DF,  p-value: < 2.2e-16


> PRESS_lm4_method1
[1] 2003169
> PRESS_lm4_method2
[1] 2003169
```

Abschließend ist Modell 3 also hier das beste Modell, da es den kleinsten PRESS-Wert hat.

R-Code zu Aufgabe 1:

```
rm(list=ls())

install.packages("qpcR")
installpackages("DAAG")
library(car)
library(qpcR)
library(MASS)
library(DAAG)

path <- "C:/Users/Coala/Desktop/A2_ERWEIT"


#*********************************************************************************************
# AUFGABE 1
#*********************************************************************************************
# 1. Fuehren Sie mit dem Datensatz realestate.txt eine regressionsanalytische
#    Modellierung durch. Evaluieren Sie die erzielte Vorhersage-Guete mittels
#    des PRESS-criterion (prediction sum of squares - siehe Skriptum).

# daten einlesen
realestate_data <- read.table(file="C:/Users/Coala/Desktop/A2_ERWEIT/realestate.txt",
                              header=FALSE, stringsAsFactors=FALSE)
head(realestate_data, n=20)
variablenames <- c("SalesPrice","SqFeet","Beds","Baths","AirCond","Garage","Pool",
                   "Year","Quality","Style","Lot","Highway")
realestate_data <- data.frame(realestate_data[2:nrow(realestate_data), ])
colnames(realestate_data) <- variablenames
for(i in 1:ncol(realestate_data)){
  realestate_data[ ,i] <- as.numeric(realestate_data[ ,i])
}


# deskriptive statistiken
nrow(realestate_data)
plot(realestate_data$SalesPrice) # abhaengige variable
plot(realestate_data$SqFeet, realestate_data$SalesPrice)
plot(as.factor(realestate_data$Beds), realestate_data$SalesPrice,
     xlab="Beds", ylab="SalesPrice")
plot(realestate_data$Year, realestate_data$SalesPrice)
plot(as.factor(realestate_data$Quality), realestate_data$SalesPrice,
     xlab="Quality", ylab="SalesPrice")
plot(realestate_data$Lot, realestate_data$SalesPrice)
summary(realestate_data)

# erklaere SalesPrice (V1) durch (kombi der) verbleibenden variablen
attach(realestate_data)


#---------------------------------------------------------------------------
# MODELL 1
#---------------------------------------------------------------------------
# multiple lineare regression (modell mit allen regressoren, additiv)
# (i.e. variablen in der anwesenheit anderer variablen testen)
lm1 <- lm(SalesPrice ~ SqFeet + Beds + Baths + AirCond + Garage + Pool
          + Year + Quality + Style + Lot + Highway)
summary(lm1)

par(mfrow=c(2,2))
plot(lm1) # diagnostic plots

par(mfrow=c(1,1))

# anova
# (beitrag der variablen sequenziell beurteilen -
#  nachteil: p-werte abhaengig von der reihenfolge der variablen)
anova_lm1 <- anova(lm1)

# Anova
# (beitrag der variablen unabhaengig von der reihenfolge beurteilen)
# (nur falls die variablen alle orthogonal zu einander sind, dann
#  ergeben anova() und Anova() dasselbe)
Anova_lm1 <- Anova(lm1)
```

```
# Method 1: PRESS (see sources 1-3, Hudec chapter "Multiple Regression" slide 36)
#           simplified formula for linear models
PRESScriterion <- function(linmod) {
  predictiveResid <- residuals(linmod)/(1 - lm.influence(linmod)$hat)
  sum(predictiveResid^2) # PRESS formula form slide 36
}

PRESS_lm1_method1 <- PRESScriterion(lm1) # PRESS for MODELL 1

# Method 2: PRESS (using package qpcR:
# https://www.rdocumentation.org/packages/qpcR/versions/1.4-1/topics/PRESS)
PRESS_lm1_method2 <- PRESS(lm1)$stat

# beide methoden ergeben dasselbe PRESS criterion, daher ueberpruefung in ordnung


#-------------------------------------------------------------------------------
# MODELL 2
#-------------------------------------------------------------------------------
# multiple lineare regression (modell mit allen regressoren, mit interaktion)
lm2 <- lm(SalesPrice ~ SqFeet + Beds + Baths + AirCond + Garage + Pool
          + Year + Quality + Style + Lot + Highway + SqFeet:Beds +
            SqFeet:Year + SqFeet:Pool + Quality:Style + Beds:Pool)
summary(lm2)

# PRESS
PRESS_lm2_method1 <- PRESScriterion(lm2)
PRESS_lm2_method2 <- PRESS(lm2)$stat


#-------------------------------------------------------------------------------
# MODELL 3
#-------------------------------------------------------------------------------
# multiple lineare regression (modell mit weniger variablen, mit interaktion)
lm3 <- lm(SalesPrice ~ SqFeet + Beds + Year + Quality + Style + Lot +
          SqFeet:Beds + SqFeet:Year + Quality:Style)
summary(lm3)

# PRESS
PRESS_lm3_method1 <- PRESScriterion(lm3)
PRESS_lm3_method2 <- PRESS(lm3)$stat


#-------------------------------------------------------------------------------
# MODELL 4
#-------------------------------------------------------------------------------
# multiple lineare regression (modell mit weniger variablen, mit interaktion)
lm4 <- lm(SalesPrice ~ SqFeet + Beds + Year + Quality + Style + SqFeet:Beds
          + SqFeet:Year)
summary(lm4)

# PRESS
PRESS_lm4_method1 <- PRESScriterion(lm4)
PRESS_lm4_method2 <- PRESS(lm4)$stat

# CONCLUSIO: modell 3 ist hier das beste, da es den kleinsten PRESS-wert,
#            und somit die hoechste vorhersageguete der betrachteten 4
#            modelle hat
```

Aufgabe 2:

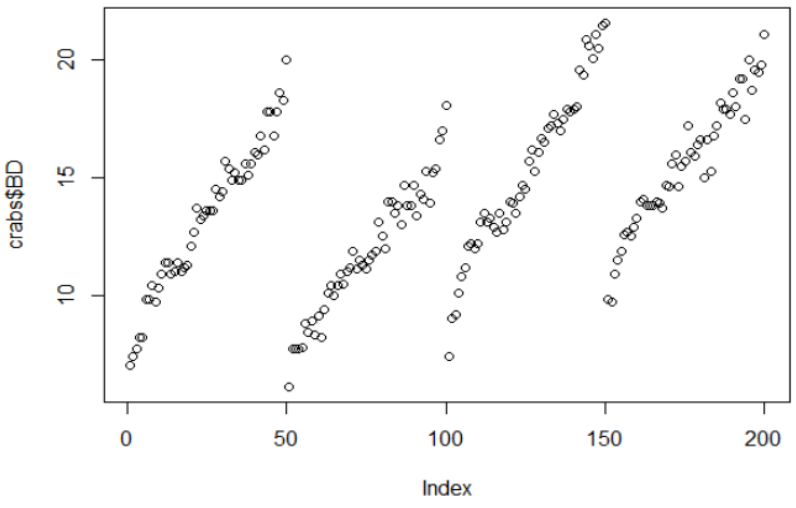
Der Datensatz crabs aus der library (MASS) enthält die Daten von 50 weiblichen und männlichen Tieren. Untersuche den linearen Zusammenhang zwischen BD (abhängige Variable und den anderen Variablen). Modelliere den Geschlechtseffekt mittels Indikatorvariablen und diskutiere die Ergebnisse.

Der Datensatz umfasst 200 Datenpunkte. Untenstehend ein Überblick über den Datensatz *crabs*:

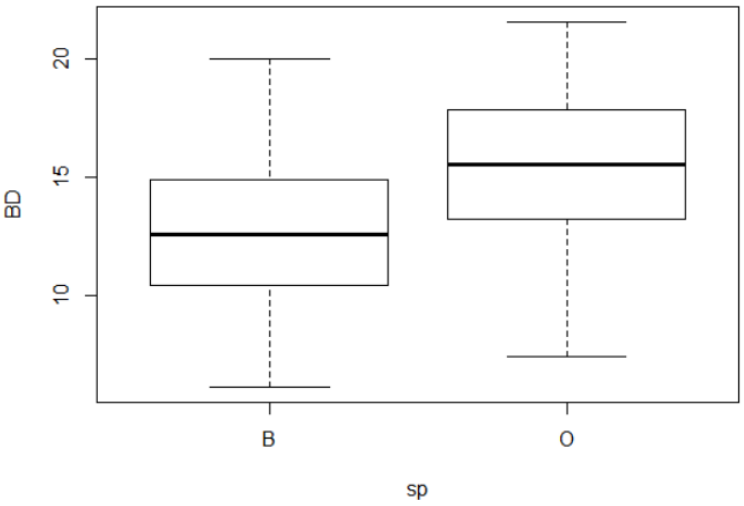
```
> head(crabs,n=20)
   sp sex index   FL   RW   CL   CW   BD
1   B   M     1  8.1  6.7 16.1 19.0  7.0
2   B   M     2  8.8  7.7 18.1 20.8  7.4
3   B   M     3  9.2  7.8 19.0 22.4  7.7
4   B   M     4  9.6  7.9 20.1 23.1  8.2
5   B   M     5  9.8  8.0 20.3 23.0  8.2
6   B   M     6 10.8  9.0 23.0 26.5  9.8
7   B   M     7 11.1  9.9 23.8 27.1  9.8
8   B   M     8 11.6  9.1 24.5 28.4 10.4
9   B   M     9 11.8  9.6 24.2 27.8  9.7
10  B   M    10 11.8 10.5 25.2 29.3 10.3
11  B   M    11 12.2 10.8 27.3 31.6 10.9
12  B   M    12 12.3 11.0 26.8 31.5 11.4
13  B   M    13 12.6 10.0 27.7 31.7 11.4
14  B   M    14 12.8 10.2 27.2 31.8 10.9
15  B   M    15 12.8 10.9 27.4 31.5 11.0
```

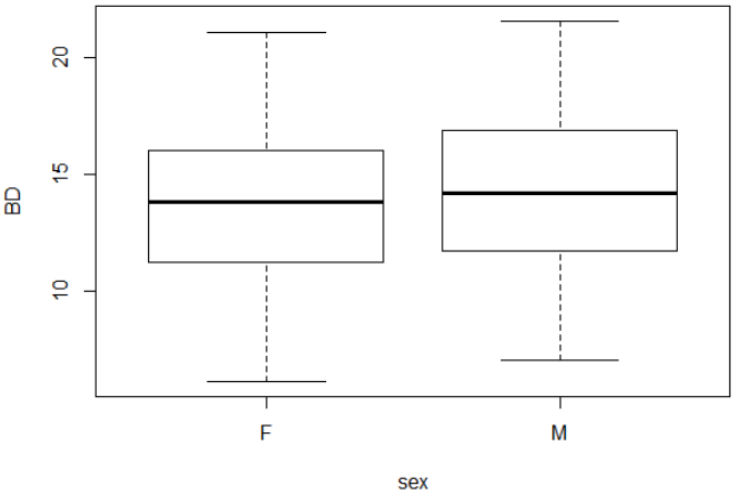
```
> summary(crabs)
 sp       sex         index            FL              RW
 B:100   F:100   Min.   : 1.0   Min.   : 7.20   Min.   : 6.50
 O:100   M:100   1st Qu.:13.0   1st Qu.:12.90   1st Qu.:11.00
                 Median :25.5   Median :15.55   Median :12.80
                 Mean   :25.5   Mean   :15.58   Mean   :12.74
                 3rd Qu.:38.0   3rd Qu.:18.05   3rd Qu.:14.30
                 Max.   :50.0   Max.   :23.10   Max.   :20.20

       CL              CW              BD
 Min.   :14.70   Min.   :17.10   Min.   : 6.10
 1st Qu.:27.27   1st Qu.:31.50   1st Qu.:11.40
 Median :32.10   Median :36.80   Median :13.90
 Mean   :32.11   Mean   :36.41   Mean   :14.03
 3rd Qu.:37.23   3rd Qu.:42.00   3rd Qu.:16.60
 Max.   :47.60   Max.   :54.60   Max.   :21.60
```

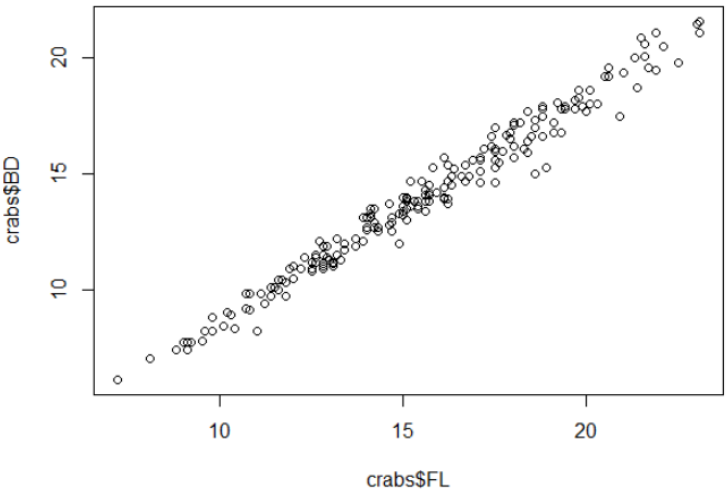
Der folgende Plot zeigt die abhängige Variable BD nach Indexreihenfolge:

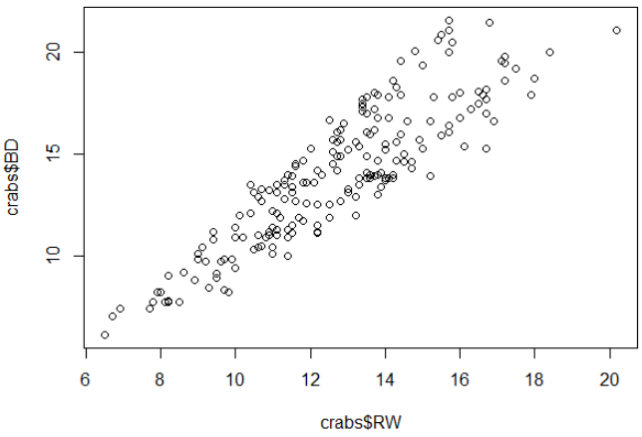


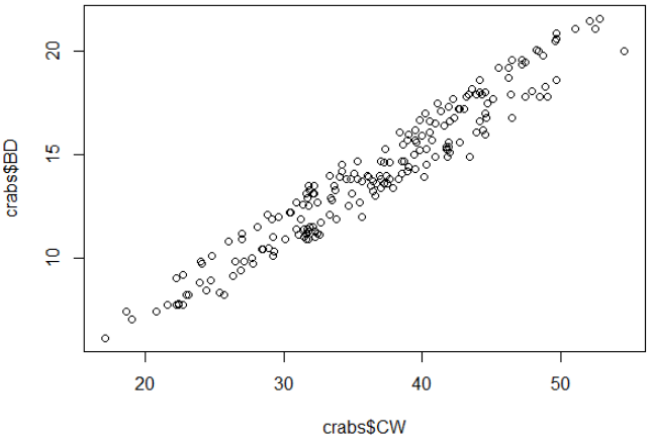
In den folgenden Plots wird der Zusammenhang von BD mit den weiteren Variablen dargestellt:







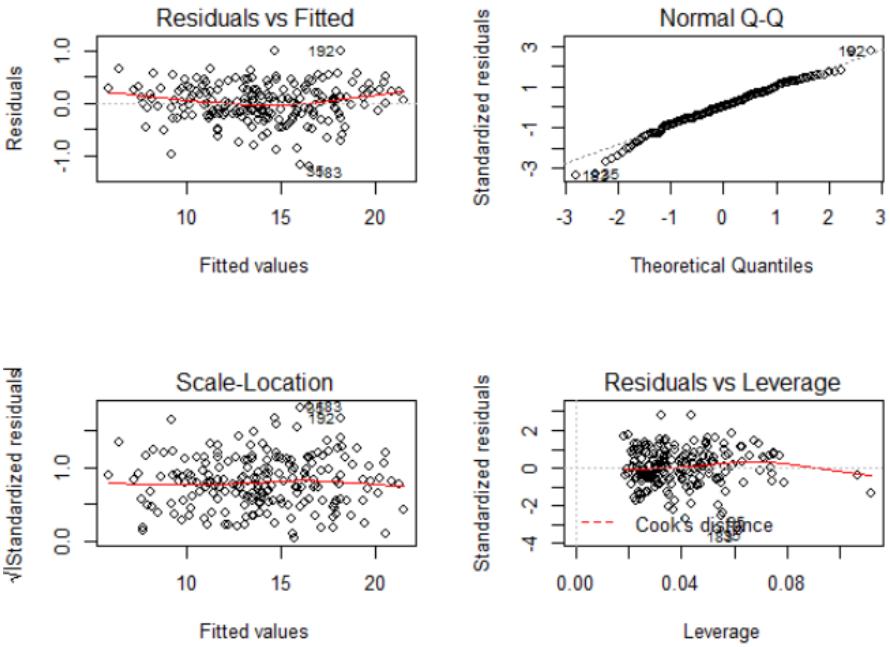




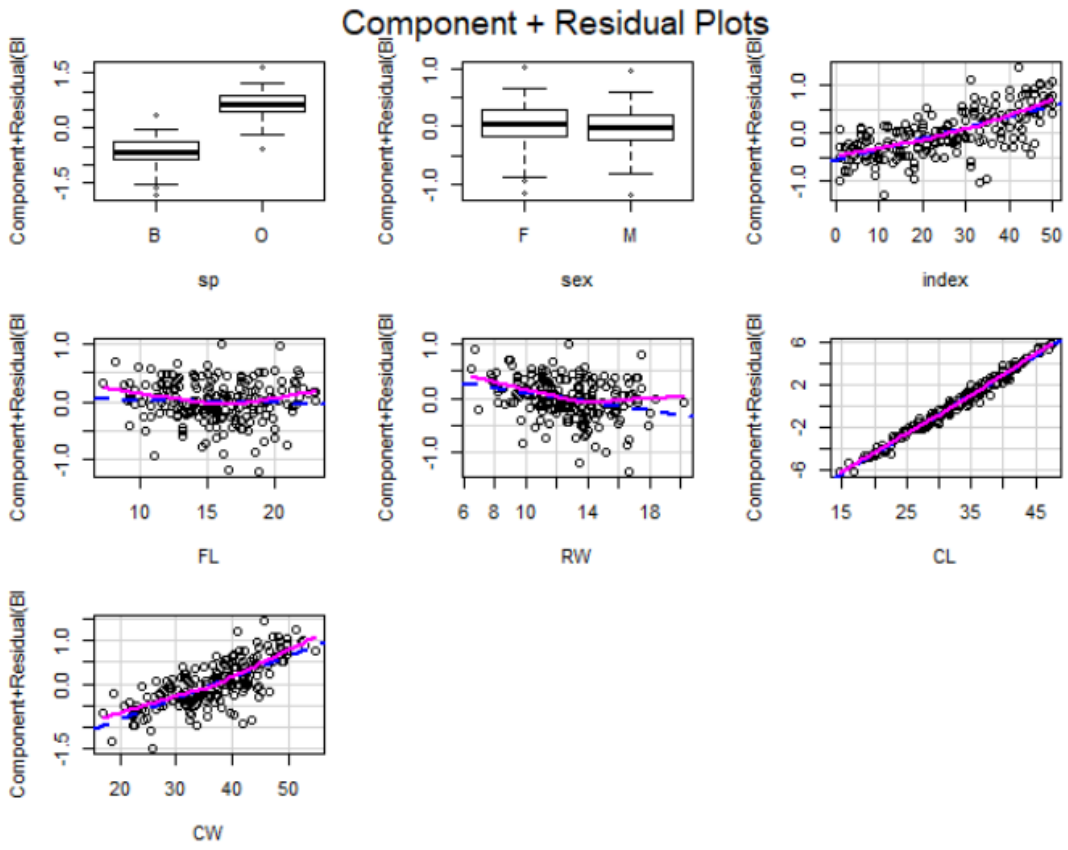
Modell 1:

- Multiple lineare Regression
- Additives Modell mit allen Regressoren
- Das Multiple R-squared liegt bei 98.9%.

```
Call:
lm(formula = BD ~ sp + sex + index + FL + RW + CL + CW)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1961 -0.2043  0.0119  0.2428  0.9977

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.510742   0.338062   1.511   0.1325
sp1         -0.656504   0.083306  -7.881 2.37e-13 ***
sex1         0.026431   0.052771   0.501   0.6170
index        0.022823   0.006699   3.407   0.0008 ***
FL          -0.006127   0.071947  -0.085   0.9322
RW          -0.042263   0.047348  -0.893   0.3732
CL           0.366607   0.068911   5.320 2.87e-07 ***
CW           0.049472   0.061715   0.802   0.4238
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3661 on 192 degrees of freedom
Multiple R-squared:  0.989,     Adjusted R-squared:  0.9886
F-statistic:  2460 on 7 and 192 DF,  p-value: < 2.2e-16
```



Bei Betrachtung der obigen Diagnostic Plots wird ersichtlich, dass die Residuen bzw. die Wurzel der standardisierten Residuen im Plot „Residuals vs. Fitted“ bzw. „Scale-Location“ zufällig um die Nulllinie liegen, was auf einen (zumindest annähernd) linearen Zusammenhang in den Daten hindeutet. In den untenstehenden C+R-Plots wird ebenfalls der lineare Zusammenhang ersichtlich.



Modell mit Indikatorvariablen für Geschlecht:

- ***Mittels Contrast Treatment:***

```
> crabs <- within(crabs, {
+          sex.ct <- C(sex.f, treatment)
+          print(attributes(sex.ct))
+          })
$`levels`
[1] "F" "M"

$class
[1] "factor"

$contrasts
[1] "contr.treatment"


Call:
lm(formula = BD ~ sex.ct, data = crabs)

Residuals:
   Min     1Q Median     3Q    Max
-7.624 -2.449  0.076  2.463  7.376

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.7240     0.3420  40.134   <2e-16 ***
sex.ctM       0.6130     0.4836   1.268    0.206
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.42 on 198 degrees of freedom
Multiple R-squared:  0.00805,   Adjusted R-squared:  0.00304
F-statistic: 1.607 on 1 and 198 DF,  p-value: 0.2064
```

- ***Mittels Helmert Coding:***

```
> crabs <- within(crabs, {
+          sex.ch <- C(sex.f, helmert)
+          print(attributes(sex.ch))
+          })
$`levels`
[1] "F" "M"

$class
[1] "factor"

$contrasts
[1] "contr.helmert"


Call:
lm(formula = BD ~ sex.ch, data = crabs)

Residuals:
   Min     1Q Median     3Q    Max
-7.624 -2.449  0.076  2.463  7.376
```

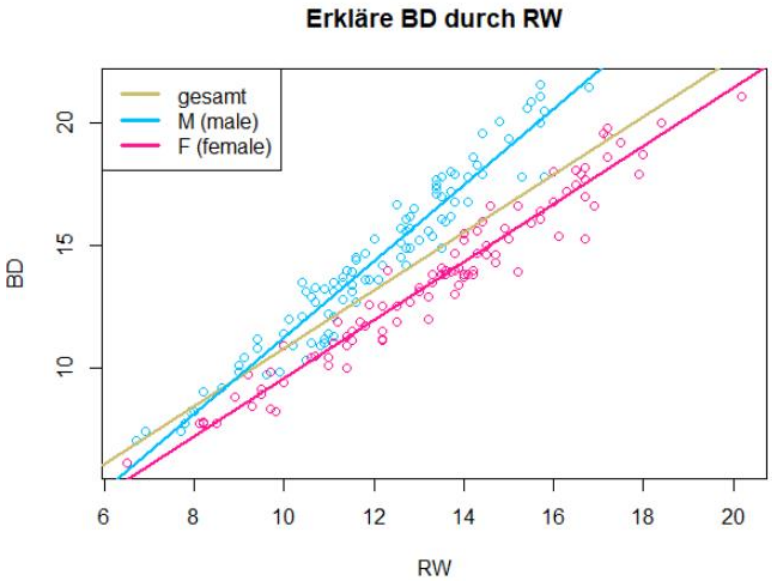
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.0305     0.2418  58.025   <2e-16 ***
sex.ch1       0.3065     0.2418   1.268    0.206
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.42 on 198 degrees of freedom
Multiple R-squared:  0.00805,   Adjusted R-squared:  0.00304
F-statistic: 1.607 on 1 and 198 DF,  p-value: 0.2064
```

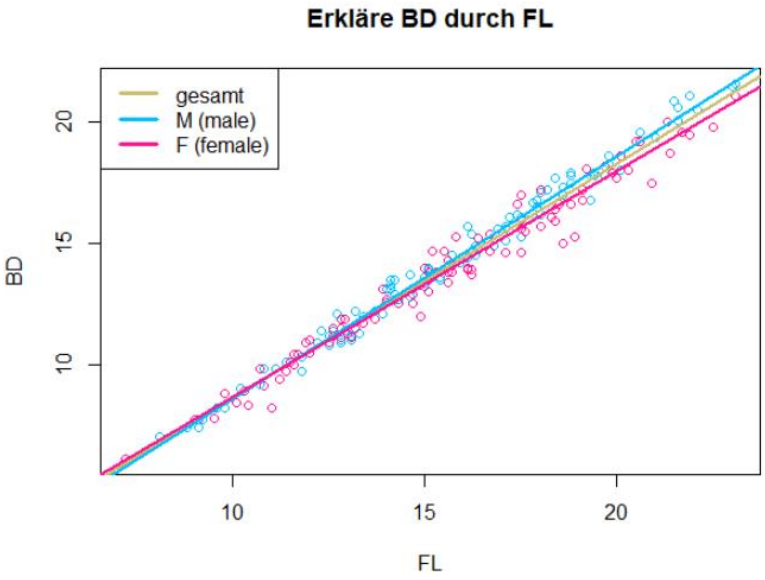
- ***Mittels contrasts():***

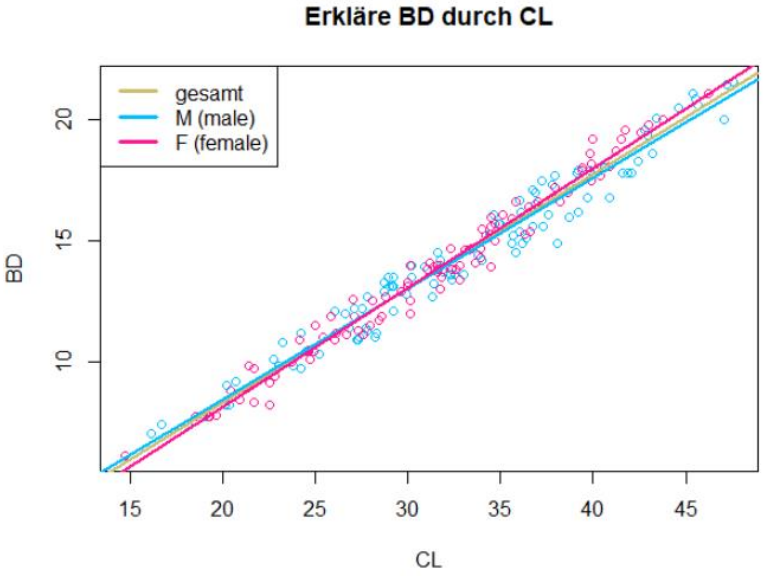
```
contrasts(crabs$sex.f) <- contr.treatment(2, base=1)
summary(lm(BD ~ sex.f, data=crabs))
```

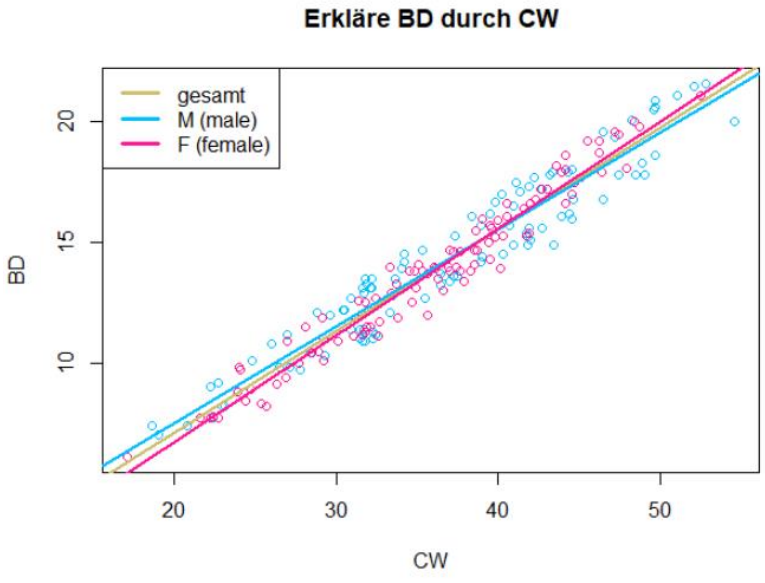
```
Call:
lm(formula = BD ~ sex.f, data = crabs)

Residuals:
   Min     1Q Median     3Q    Max
-7.624 -2.449  0.076  2.463  7.376

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.7240     0.3420  40.134   <2e-16 ***
sex.f2        0.6130     0.4836   1.268    0.206
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.42 on 198 degrees of freedom
Multiple R-squared:  0.00805,   Adjusted R-squared:  0.00304
F-statistic: 1.607 on 1 and 198 DF,  p-value: 0.2064
```

Die folgenden Plots stellen in grün die gesamte Regressionsgerade, in blau die Regressionsgerade nur für Männer, und in rosarot die Regressionsgerade nur für Frauen dar und erklären die abhängige Variable BD durch die jeweiligen weiteren Variablen:




Erkläre BD durch FL
gesamt
M (male)
F (female)
BD
FL
10
15
20


Erkläre BD durch CL
gesamt
M (male)
F (female)
BD
CL
15
20
25
30
35
40
45


Erkläre BD durch CW
gesamt
M (male)
F (female)
BD
CW
20
30
40
50

R-Code zu Aufgabe 2:

```
#*******************************************************************************
# AUFGABE 2
#*******************************************************************************
# 2. Der Datensatz crabs aus der library (MASS) enthält die Daten von
#    50 weiblichen und männlichen Tieren. Untersuche den linearen Zusammenhang
#    zwischen BD (abhängige Variable und den anderen Variablen).
#    Modelliere den Geschlechtseffekt mittels Indikatorvariablen und diskutiere
#    die Ergebnisse.

# deskriptive statistiken
attach(crabs)
head(crabs,n=20)
nrow(crabs)
summary(crabs)

plot(crabs$BD) # abhaengige variable
plot(crabs$sp, crabs$BD,
     xlab="sp", ylab="BD")
plot(crabs$sex, crabs$BD,
     xlab="sex", ylab="BD")
plot(crabs$FL, crabs$BD)
plot(crabs$RW, crabs$BD)
plot(crabs$CW, crabs$BD)


# MODELL 1
# multiple lineare regression (modell mit allen regressoren, additiv)
lm_crabs1 <- lm(BD ~ sp + sex + index + FL + RW + CL + CW)
summary(lm_crabs1)

par(mfrow=c(2,2))
plot(lm_crabs1) # diagnostic plots
crPlots(lm_crabs1)
# conclusion: linearer zshg sichtbar, da in den plots "residulas vs.
#             fitted" und "scale-location" die residuen bzw. stand-
#             ardisierten residuen zufaellig um die nulllinie liegen.
#             auch in den C+R-Plots (Component+Residual-Plots) wird
#             der lineare zshg ersichtlich.

par(mfrow=c(1,1))


# MODELL mit INDIKATORVARIABLEN fuer geschlecht
# mittels contrast treatment:
crabs$sex.f <- factor(crabs$sex)
is.factor(crabs$sex.f)

crabs <- within(crabs, {
         sex.ct <- C(sex.f, treatment)
         print(attributes(sex.ct))
         })

summary(lm(BD ~ sex.ct, data=crabs))


# mittels helmert coding:
crabs <- within(crabs, {
         sex.ch <- C(sex.f, helmert)
         print(attributes(sex.ch))
         })

summary(lm(BD ~ sex.ch, data=crabs))
```

```
# mittels contrasts():
contrasts(crabs$sex.f) <- contr.treatment(2, base=1)
summary(lm(BD ~ sex.f, data=crabs))

# plots der regressionsgeraden gesamt und
# unter beachtung des geschlechts:

colors <- c("deeppink1", "deepskyblue")


# erklaerung von BD durch RW:
plot(BD ~ RW, main="Erkläre BD durch RW", col=colors[sex])
# gesamte regressionsgerade:
abline(lm(BD ~ RW), col="lightgoldenrod3", lwd=2)
# regressionsgerade fuer M (maennlich):
abline(lm(BD ~ RW, subset=(sex=="M")), col="deepskyblue", lwd=2)
# regressionsgerade fuer F (weiblich):
abline(lm(BD ~ RW, subset=(sex=="F")), col="deeppink1", lwd=2)
# legend
legend("topleft", legend=c("gesamt", "M (male)", "F (female)"),
       col=c("lightgoldenrod3", "deepskyblue", "deeppink1"), lwd=3)


# erklaerung von BD durch FL:
plot(BD ~ FL, main="Erkläre BD durch FL", col=colors[sex])
# gesamte regressionsgerade:
abline(lm(BD ~ FL), col="lightgoldenrod3", lwd=2)
# regressionsgerade fuer M (maennlich):
abline(lm(BD ~ FL, subset=(sex=="M")), col="deepskyblue", lwd=2)
# regressionsgerade fuer F (weiblich):
abline(lm(BD ~ FL, subset=(sex=="F")), col="deeppink1", lwd=2)
# legend
legend("topleft", legend=c("gesamt", "M (male)", "F (female)"),
       col=c("lightgoldenrod3", "deepskyblue", "deeppink1"), lwd=3)


# erklaerung von BD durch CL:
plot(BD ~ CL, main="Erkläre BD durch CL", col=colors[sex])
# gesamte regressionsgerade:
abline(lm(BD ~ CL), col="lightgoldenrod3", lwd=2)
# regressionsgerade fuer M (maennlich):
abline(lm(BD ~ CL, subset=(sex=="M")), col="deepskyblue", lwd=2)
# regressionsgerade fuer F (weiblich):
abline(lm(BD ~ CL, subset=(sex=="F")), col="deeppink1", lwd=2)
# legend
legend("topleft", legend=c("gesamt", "M (male)", "F (female)"),
       col=c("lightgoldenrod3", "deepskyblue", "deeppink1"), lwd=3)


# erklaerung von BD durch CW:
plot(BD ~ CW, main="Erkläre BD durch CW", col=colors[sex])
# gesamte regressionsgerade:
abline(lm(BD ~ CW), col="lightgoldenrod3", lwd=2)
# regressionsgerade fuer M (maennlich):
abline(lm(BD ~ CW, subset=(sex=="M")), col="deepskyblue", lwd=2)
# regressionsgerade fuer F (weiblich):
abline(lm(BD ~ CW, subset=(sex=="F")), col="deeppink1", lwd=2)
# legend
legend("topleft", legend=c("gesamt", "M (male)", "F (female)"),
       col=c("lightgoldenrod3", "deepskyblue", "deeppink1"), lwd=3)
```

Aufgabe 3:

Analysiere den Datensatz leaftemp aus der library (DAAG). Untersuche den linearen Zusammenhang zwischen tempDiff (abhängige Variable und den anderen Variablen). Modelliere den Einfluss von CO2level mittels Indikatorvariablen sowohl mit contrast.treatment als auch mit contrast.sum und diskutiere die Ergebnisse.

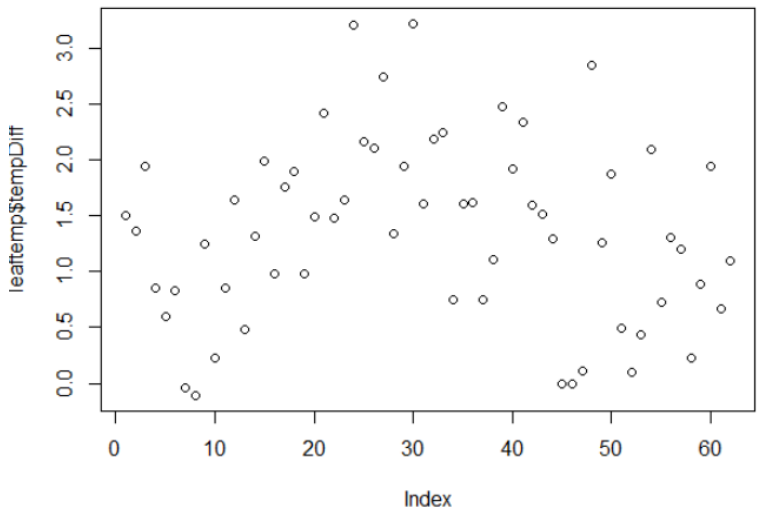
Der Datensatz *leaftemp* (siehe untenstehender Ausschnitt) umfasst 62 Datenpunkte.

```
> head(leaftemp,n=20)
   CO2level vapPress tempDiff BtempDiff
1      high     2.56     1.50      1.84
2       low     1.88     1.36      1.54
3    medium     2.38     1.94      1.96
4      high     2.55     0.85      0.93
5       low     2.20     0.60      0.68
6    medium     2.72     0.83      0.89
7      high     2.17    -0.04      0.02
8    medium     2.21    -0.11     -0.06
9      high     1.64     1.25      1.10
10      low     1.75     0.23      0.50
11   medium     1.67     0.85      0.65
12     high     1.67     1.64      1.27
13      low     1.85     0.48      0.65
14   medium     1.67     1.32      1.05
15     high     1.81     1.99      1.72
```

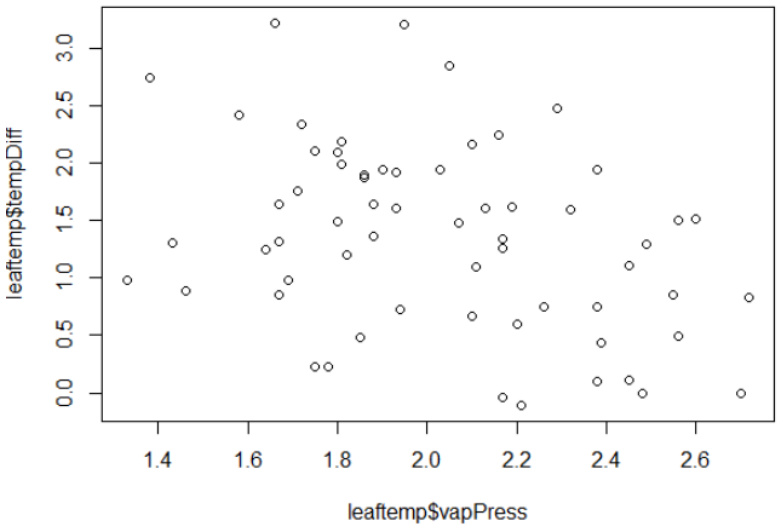
Deskriptive Statistiken zum Datensatz:

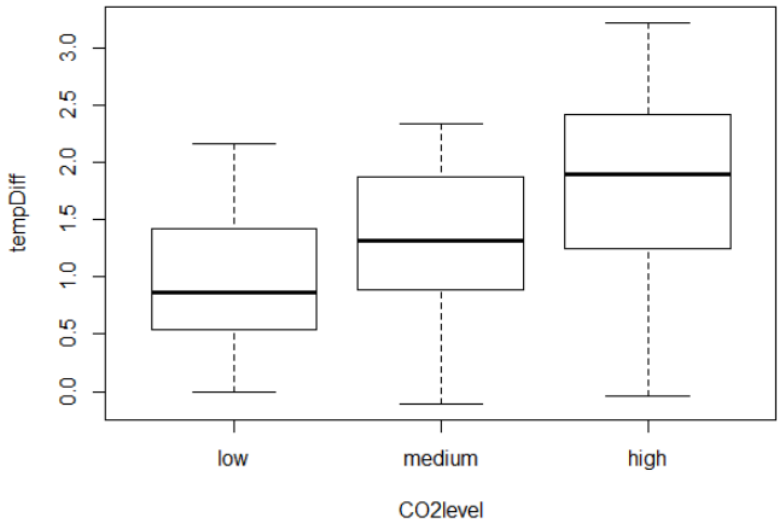
```
> summary(leaftemp)
   CO2level     vapPress        tempDiff        BtempDiff
 low   :20   Min.   :1.330   Min.   :-0.110   Min.   :-0.0600
 medium:21   1st Qu.:1.785   1st Qu.: 0.770   1st Qu.: 0.6725
 high  :21   Median :1.990   Median : 1.350   Median : 1.0650
             Mean   :2.028   Mean   : 1.360   Mean   : 1.1450
             3rd Qu.:2.283   3rd Qu.: 1.935   3rd Qu.: 1.7175
             Max.   :2.720   Max.   : 3.220   Max.   : 2.3400
```

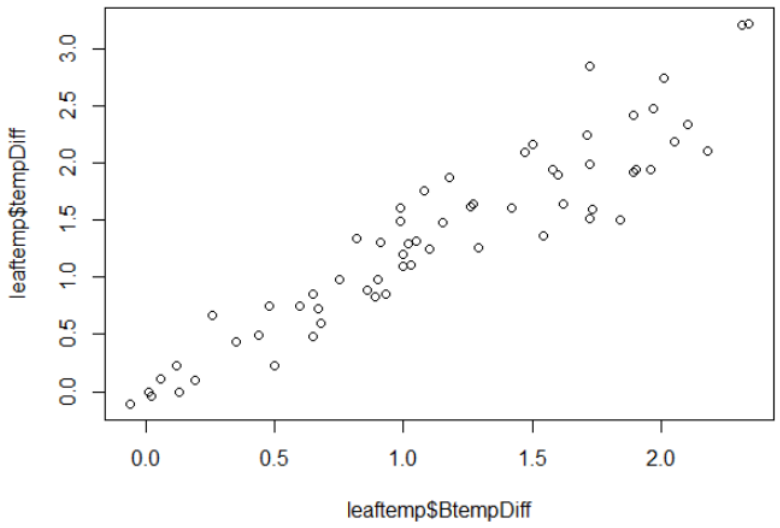
Im folgenden Plot wird die Variable *tempDiff* in der Reihenfolge ihres Vorkommens im Datensatz dargestellt:



Die weiteren drei Plots zeigen den Zusammenhang der abhängigen Variable *tempDiff* mit den Variablen *vapPress*, *CO2level* und *BtempDiff*:



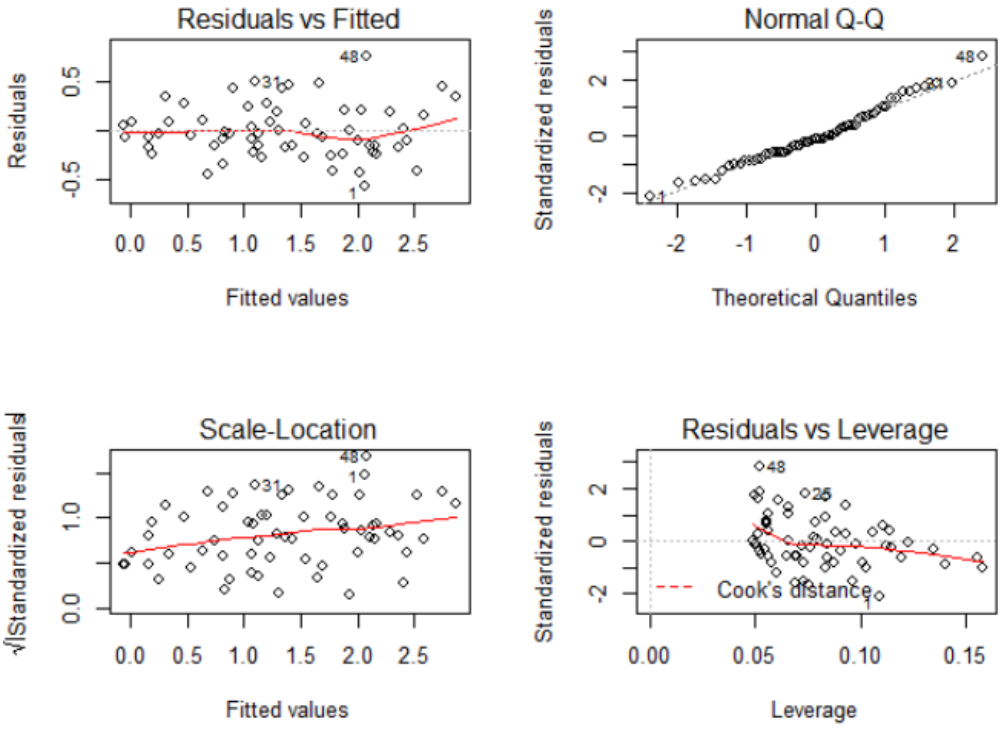


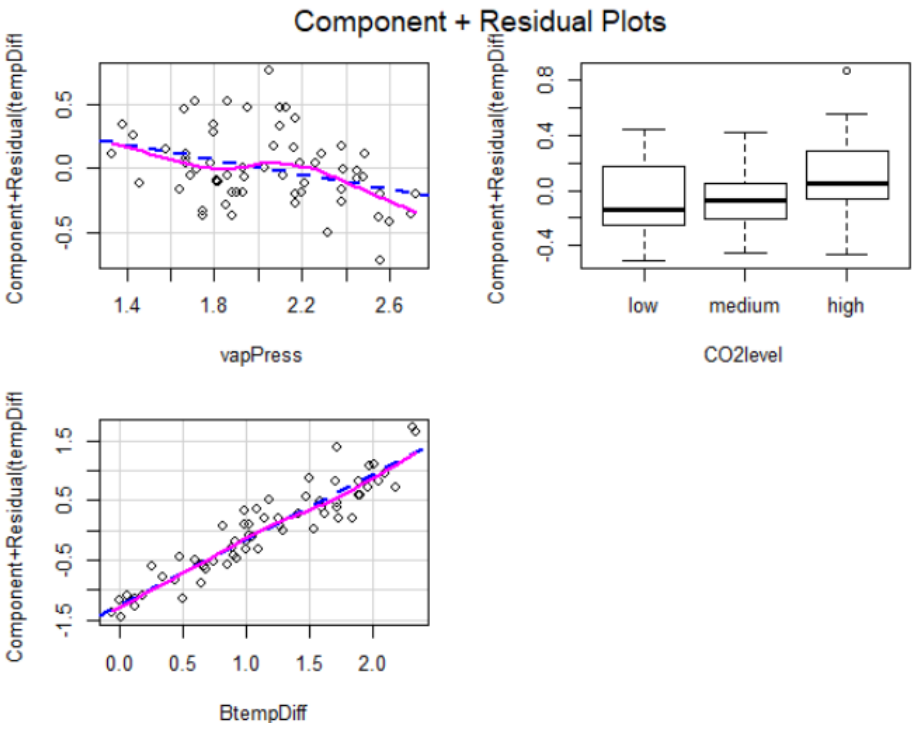


Multiples lineares Regressionsmodell mit allen Variablen (additiv):

Das Modell hat ein Multiple R-square von 88.85% und gemäß Residual vs. Fitted sowie Scale-Location Plot sowie den C+R-Plots ist ersichtlich, dass es sich um einen (annähernd) linearen Zusammenhang handelt.

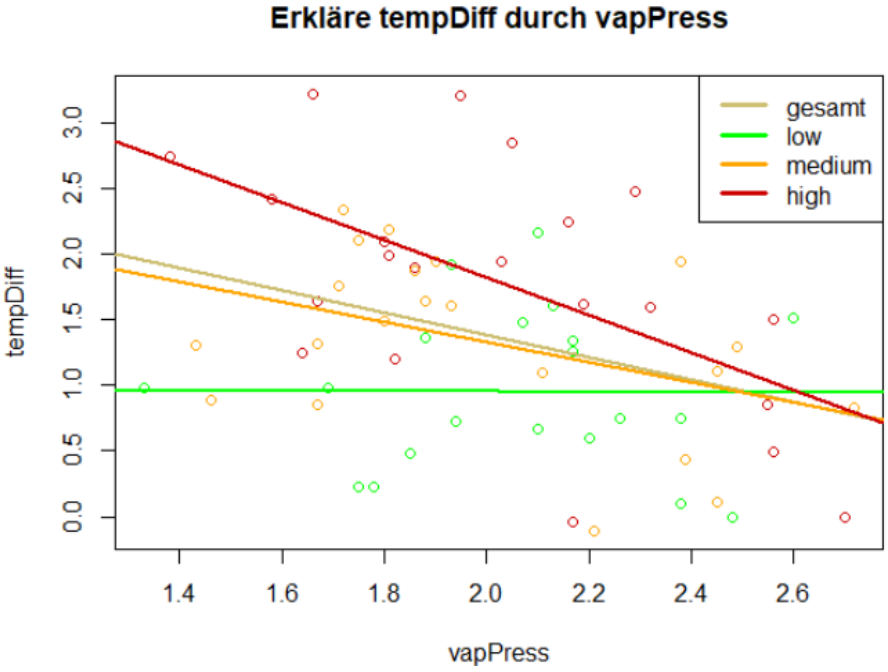
```
Call:
lm(formula = tempDiff ~ vapPress + CO2level + BtempDiff)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56633 -0.16463 -0.03188  0.17651  0.76942

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.68888    0.25465   2.705  0.00899 **
vapPress    -0.28404    0.10946  -2.595  0.01201 *
CO2level1   -0.05675    0.05387  -1.053  0.29658
CO2level2   -0.04520    0.05020  -0.900  0.37177
BtempDiff    1.08839    0.06250  17.415  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2797 on 57 degrees of freedom
Multiple R-squared:  0.8885,    Adjusted R-squared:  0.8807
F-statistic: 113.6 on 4 and 57 DF,  p-value: < 2.2e-16
```

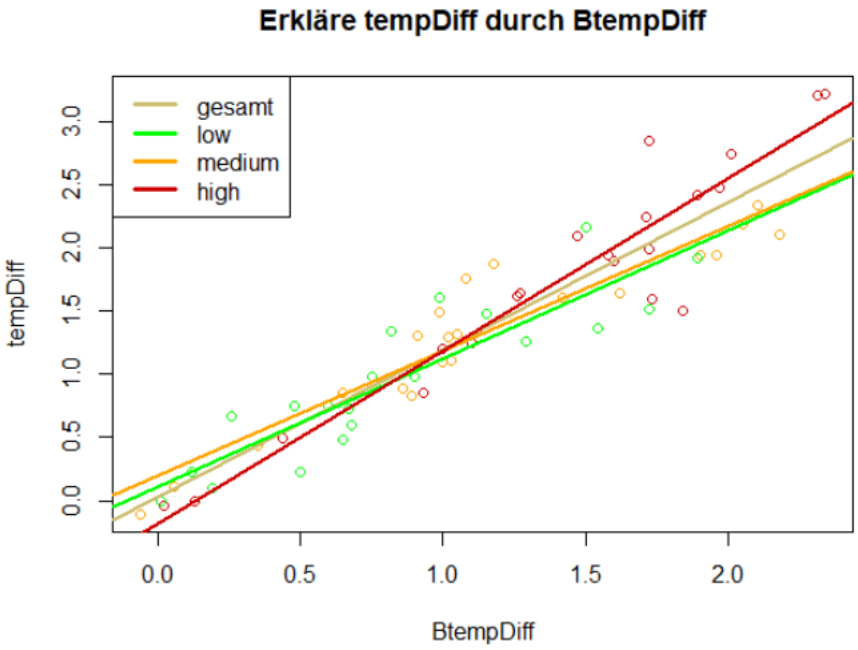




Das Modell hat ein Multiple R-squared von 88.85% und gemäß Residual vs. Fitted sowie Scale-Location Plot sowie den C+R-Plots ist ersichtlich, dass es sich um einen (annähernd) linearen Zusammenhang handelt.

Die folgenden Plots stellen in ockerfarben die gesamte Regressionsgerade, in grün die Regressionsgerade nur für „low *CO2level*“, in orange die Regressionsgerade nur für „medium *CO2level*“ und in rot die Regressionsgerade nur für „high *CO2level*“ dar und erklären die abhängige Variable *tempDiff* durch die jeweiligen weiteren Variablen unter Berücksichtigung des CO2-Levels:





<u>Modellmatrix für die Modelle mit Indikatorvariablen für CO2level:</u>

- ***Additives Modell:***

Ausschnitt aus der *model.matrix*:

```
>   # additiv (ohne interaktion)
> model.matrix(tempDiff ~ vapPress + BtempDiff + CO2level)
   (Intercept) vapPress BtempDiff CO2level1 CO2level2
1            1     2.56      1.84        -1        -1
2            1     1.88      1.54         1         0
3            1     2.38      1.96         0         1
4            1     2.55      0.93        -1        -1
5            1     2.20      0.68         1         0
6            1     2.72      0.89         0         1
7            1     2.17      0.02        -1        -1
8            1     2.21     -0.06         0         1
9            1     1.64      1.10        -1        -1
10           1     1.75      0.50         1         0
11           1     1.67      0.65         0         1
12           1     1.67      1.27        -1        -1
13           1     1.85      0.65         1         0
14           1     1.67      1.05         0         1
15           1     1.81      1.72        -1        -1
```

- ***Modell mit Interaktion:***

Ausschnitt aus der *model.matrix*:

```
> model.matrix(tempDiff ~ vapPress * CO2level * BtempDiff)
   (Intercept) vapPress CO2level1 CO2level2 BtempDiff vapPress:CO2level1 vapPress:CO2level2 vapPress:BtempDiff
1            1     2.56        -1        -1      1.84              -2.56              -2.56             4.7104
2            1     1.88         1         0      1.54               1.88               0.00             2.8952
3            1     2.38         0         1      1.96               0.00               2.38             4.6648
4            1     2.55        -1        -1      0.93              -2.55              -2.55             2.3715
5            1     2.20         1         0      0.68               2.20               0.00             1.4960
6            1     2.72         0         1      0.89               0.00               2.72             2.4208
7            1     2.17        -1        -1      0.02              -2.17              -2.17             0.0434
8            1     2.21         0         1     -0.06               0.00               2.21            -0.1326
9            1     1.64        -1        -1      1.10              -1.64              -1.64             1.8040
10           1     1.75         1         0      0.50               1.75               0.00             0.8750
11           1     1.67         0         1      0.65               0.00               1.67             1.0855
12           1     1.67        -1        -1      1.27              -1.67              -1.67             2.1209
13           1     1.85         1         0      0.65               1.85               0.00             1.2025
14           1     1.67         0         1      1.05               0.00               1.67             1.7535
15           1     1.81        -1        -1      1.72              -1.81              -1.81             3.1132
```

<u>Anwendung verschiedener Kontraststrategien:</u>

- ***Additives Modell (contr.treatment):***

Bei der Verwendung von *contrast treatment* wird eine Non-Treatment-Gruppe als Baseline gefittet, und der gruppenspezifische Effekt wird als Differenz zur Baseline dargestellt. Die Kontraste werden also folgendermaßen dargestellt:

```
> contrasts(CO2level) # baseline hier: "low"
       medium high
low         0    0
medium      1    0
high        0    1
```

Zugehöriges Modell:

```
Call:
lm(formula = tempDiff ~ vapPress + BtempDiff + CO2level)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56633 -0.16463 -0.03188  0.17651  0.76942

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.63213    0.25360   2.493   0.0156 *
vapPress      -0.28404    0.10946  -2.595   0.0120 *
BtempDiff      1.08839    0.06250  17.415   <2e-16 ***
CO2levelmedium 0.01156    0.08946   0.129   0.8977
CO2levelhigh   0.15870    0.09471   1.676   0.0993 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2797 on 57 degrees of freedom
Multiple R-squared:  0.8885,    Adjusted R-squared:  0.8807
F-statistic: 113.6 on 4 and 57 DF,  p-value: < 2.2e-16
```

- ***Interaktives Modell (contr.treatment):***

Zugehöriges Modell:

```
Call:
lm(formula = tempDiff ~ vapPress * CO2level * BtempDiff)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46558 -0.16123 -0.01313  0.13540  0.74679

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -0.21498    0.80268  -0.268    0.790
vapPress                         0.15093    0.37210   0.406    0.687
CO2levelmedium                   1.43937    1.13766   1.265    0.212
CO2levelhigh                    -0.18233    1.24555  -0.146    0.884
BtempDiff                        1.39832    0.86327   1.620    0.112
vapPress:CO2levelmedium         -0.62898    0.52889  -1.189    0.240
vapPress:CO2levelhigh           -0.03137    0.55205  -0.057    0.955
vapPress:BtempDiff              -0.17909    0.39715  -0.451    0.654
CO2levelmedium:BtempDiff        -0.87244    1.11887  -0.780    0.439
CO2levelhigh:BtempDiff           0.69059    1.03956   0.664    0.510
vapPress:CO2levelmedium:BtempDiff  0.38283  0.52781   0.725    0.472
vapPress:CO2levelhigh:BtempDiff  -0.20024    0.47390  -0.423    0.674

Residual standard error: 0.2659 on 50 degrees of freedom
Multiple R-squared:  0.9116,    Adjusted R-squared:  0.8922
F-statistic: 46.88 on 11 and 50 DF,  p-value: < 2.2e-16
```

Wenn man nun die beiden Modelle beispielsweise hinsichtlich PRESS-Kriterium vergleicht, ist das additive Modell das bessere:

```
> press_res_treat_1
[1] 5.209631
> press_res_treat_2
[1] 6.252792
```

- ***Additives Modell (contr.sum):***

Bei der Verwendung von *contrast sum* wird eine „durchschnittliche“ Gerade gefittet, und die Abbildung des gruppenspezifischen Effekts erfolgt als Abweichung vom „Durchschnitt“. Die Kontraste werden also folgendermaßen dargestellt:

```
> contrasts(CO2level)
       [,1] [,2]
low       1    0
medium    0    1
high     -1   -1
```

Zugehöriges Modell:

```
Call:
lm(formula = tempDiff ~ vapPress + BtempDiff + CO2level)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56633 -0.16463 -0.03188  0.17651  0.76942

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.68888    0.25465   2.705  0.00899 **
vapPress    -0.28404    0.10946  -2.595  0.01201 *
BtempDiff    1.08839    0.06250  17.415  < 2e-16 ***
CO2level1   -0.05675    0.05387  -1.053  0.29658
CO2level2   -0.04520    0.05020  -0.900  0.37177
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2797 on 57 degrees of freedom
Multiple R-squared:  0.8885,    Adjusted R-squared:  0.8807
F-statistic: 113.6 on 4 and 57 DF,  p-value: < 2.2e-16
```

- ***Modell mit Interaktion (contr.sum):***

Zugehöriges Modell:

```
Call:
lm(formula = tempDiff ~ vapPress * CO2level * BtempDiff)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46558 -0.16123 -0.01313  0.13540  0.74679

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.20404    0.49457   0.413  0.68170
vapPress                    -0.06918    0.22262  -0.311  0.75726
CO2level1                   -0.41901    0.67776  -0.618  0.53923
CO2level2                    1.02035    0.67916   1.502  0.13929
BtempDiff                    1.33771    0.41996   3.185  0.00249 **
vapPress:CO2level1           0.22012    0.30937   0.711  0.48008
vapPress:CO2level2          -0.40886    0.31088  -1.315  0.19445
vapPress:BtempDiff          -0.11823    0.19592  -0.603  0.54893
CO2level1:BtempDiff          0.06061    0.65175   0.093  0.92627
CO2level2:BtempDiff         -0.81182    0.58758  -1.382  0.17322
vapPress:CO2level1:BtempDiff -0.06086    0.30159  -0.202  0.84089
vapPress:CO2level2:BtempDiff  0.32197    0.28048   1.148  0.25647
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2659 on 50 degrees of freedom
Multiple R-squared:  0.9116,    Adjusted R-squared:  0.8922
F-statistic: 46.88 on 11 and 50 DF,  p-value: < 2.2e-16
```

Wenn man nun die beiden Modelle beispielsweise hinsichtlich PRESS-Kriterium vergleicht, ist das additive Modell das bessere:

```
> press_res_sum_1
[1] 5.209631
> press_res_sum_2
[1] 6.252792
```

Die verschiedenen Kontraststrategien sind Darstellungsformen, die schließlich dasselbe Ergebnis liefern (z.B. hinsichtlich Multiple R-squared).

R-Code zu Aufgabe 3:

```
#*******************************************************************************
# AUFGABE 3
#*******************************************************************************
# 3. Analysiere den Datensatz leaftemp aus der library (DAAG). Untersuche den
#    linearen Zusammenhang zwischen tempDiff (abhängige Variable und den anderen
#    Variablen). Modelliere den Einfluss von CO2level mittels Indikatorvariablen
#    sowohl mit contrast.treatment als auch mit contrast.sum und diskutiere die
#    Ergebnisse.

# deskriptive statistiken
attach(leaftemp)
head(leaftemp,n=20)
nrow(leaftemp)
summary(leaftemp)

plot(leaftemp$tempDiff) # abhaengige variable
plot(leaftemp$vapPress, leaftemp$tempDiff)
plot(leaftemp$CO2level, leaftemp$tempDiff,
     xlab="CO2level", ylab="tempDiff")
plot(leaftemp$BtempDiff, leaftemp$tempDiff)


# MODELL 1
# multiple lineare regression (modell mit allen regressoren, additiv)
lm_leaf1 <- lm(tempDiff ~ vapPress + CO2level + BtempDiff)
summary(lm_leaf1)

par(mfrow=c(2,2))
plot(lm_leaf1) # diagnostic plots
crPlots(lm_leaf1)
# conclusion: linearer zshg sichtbar, da in den plots "residuals vs.
#             fitted" und "scale-location" die residuen bzw. stand-
#             ardisierten residuen zufaellig um die nulllinie liegen.
#             auch in den C+R-Plots (Component+Residual-Plots) wird
#             der lineare zshg ersichtlich.

par(mfrow=c(1,1))

# plots der regressionsgeraden gesamt und
# unter beachtung des CO2level:

colors <- c("green", "orange", "red3")


# erklaerung von tempDiff durch vapPress:
plot(tempDiff ~ vapPress, col=colors[leaftemp$CO2level],
     main="Erkläre tempDiff durch vapPress")
# gesamte regressionsgerade:
abline(lm(tempDiff ~ vapPress), col="lightgoldenrod3", lwd=2)
# regressionsgerade fuer low:
abline(lm(tempDiff ~ vapPress, subset=(CO2level=="low")),
       col="green", lwd=2)
# regressionsgerade fuer medium:
abline(lm(tempDiff ~ vapPress, subset=(CO2level=="medium")),
       col="orange", lwd=2)
# regressionsgerade fuer high:
abline(lm(tempDiff ~ vapPress, subset=(CO2level=="high")),
       col="red3", lwd=2)
# legend
legend("topright", legend=c("gesamt", "low", "medium", "high"),
       col=c("lightgoldenrod3", "green", "orange", "red3"), lwd=3)
```

```
# erklaerung von tempDiff durch BtempDiff:
plot(tempDiff ~ BtempDiff, main="Erkläre tempDiff durch BtempDiff",
     col=colors[CO2level])
# gesamte regressionsgerade:
abline(lm(tempDiff ~ BtempDiff), col="lightgoldenrod3", lwd=2)
# regressionsgerade fuer low:
abline(lm(tempDiff ~ BtempDiff, subset=(CO2level=="low")),
       col="green", lwd=2)
# regressionsgerade fuer medium:
abline(lm(tempDiff ~ BtempDiff, subset=(CO2level=="medium")),
       col="orange", lwd=2)
# regressionsgerade fuer high:
abline(lm(tempDiff ~ BtempDiff, subset=(CO2level=="high")),
       col="red3", lwd=2)
# legend
legend("topleft", legend=c("gesamt", "low", "medium", "high"),
       col=c("lightgoldenrod3", "green", "orange", "red3"), lwd=3)


# MODELL mit INDIKATORVARIABLEN fuer geschlecht
# mittels contrast treatment:
leaftemp$co2.f <- factor(leaftemp$CO2level)
is.factor(leaftemp$co2.f)

leaftemp <- within(leaftemp, {
  co2.ct <- C(co2.f, treatment)
  print(attributes(co2.ct))
})

summary(lm(tempDiff ~ vapPress + co2.f + BtempDiff, data=leaftemp))


# weitere modelle mit indikatorvariablen:
  # additiv (ohne interaktion)
model.matrix(tempDiff ~ vapPress + BtempDiff + CO2level)
  # mit interaktion
model.matrix(tempDiff ~ vapPress * CO2level * BtempDiff)

# mittels contrast.treatment:
# (i.e. fitten auf non-treatment gruppe als baseline,
# und hinzugeben des gruppenspezifischen effekts als
# differenz zur baseline)
options(contrasts=c("contr.treatment", "contr.poly"))
contrasts(CO2level) # baseline hier: "low"
  # contrast.treatment bei additivem modell:
summary(res_treat_1 <- lm(tempDiff ~ vapPress + BtempDiff + CO2level))
  # contrast.treatment bei modell mit interaktion:
summary(res_treat_2 <- lm(tempDiff ~ vapPress * CO2level * BtempDiff))


par(mfrow=c(2,2))
plot(res_treat_1)
plot(res_treat_2)
par(mfrow=c(1,1))

press_res_treat_1 <- PRESScriterion(res_treat_1)
press_res_treat_2 <- PRESScriterion(res_treat_2)
# vorhersageguete von modell res_treat_1 ist besser laut PRESS

# mittels contrast.sum:
# (i.e. fitten einer "durchschnittlichen" gerade, und
# abbildung des gruppenspezifischen effekts als
# abweichung vom "durchschnitt")
options(contrasts=c("contr.sum", "contr.poly"))
contrasts(CO2level)
# contrast.sum bei additivem modell:
summary(res_sum_1 <- lm(tempDiff ~ vapPress + BtempDiff + CO2level))
# contrast.sum bei modell mit interaktion:
summary(res_sum_2 <- lm(tempDiff ~ vapPress * CO2level * BtempDiff))
```

```
par(mfrow=c(2,2))
plot(res_sum_1)
plot(res_sum_2)
par(mfrow=c(1,1))

press_res_sum_1 <- PRESScriterion(res_sum_1)
press_res_sum_2 <- PRESScriterion(res_sum_2)

# vorhersageguete von modell res_treat_1 ist besser laut PRESS
```

Literaturquellen:

- Folien und R-Codes zu den bisher vorgetragenen Kapiteln aus UK Erweiterungen des linearen Modells (Prof. Marcus Hudec).
- Model Selection Criteria and Predictive Power of Regression (Github) – PRESS (Tom Hopper, 2018); https://gist.github.com/tomhopper/8c204d978c4a0cbcb8c0; Zugriff am 12.04.2019.
- PRESS Diagnostic (2016); https://stats.stackexchange.com/questions/248603/how-can-one-compute-the-press-diagnostic; Zugriff am 12.04.2019.
- Coding for Categorical Variables in Regression Models; UCLA Institute for Digital Research & Education (2018); https://stats.idre.ucla.edu/r/modules/coding-for-categorical-variables-in-regression-models/; Zugriff am 12.04.2019.