

Aufgabe 09

Gruppe 01

30.06.2020

Aufgabe 20

20 a)

Berechnen Sie - soweit möglich - geeignete Korrelationskoeffizienten für die austauschbaren Ca-Ionen (evtl. transformiert) mit sämtlichen Reliefgrößen

```
SPDF <- l.jz
```

```
#Wie geht das eleganter als nachfolgend?
```

```
cor(x = SPDF@data$Ca_exch, y = SPDF@data$yingtan_elevation)
```

```
## [1] -0.3230832
```

```
cor(x = SPDF@data$Ca_exch, y = SPDF@data$tri)
```

```
## [1] -0.1272219
```

```
cor(x = SPDF@data$Ca_exch, y = SPDF@data$tpi)
```

```
## [1] -0.02697608
```

```
cor(x = SPDF@data$Ca_exch, y = SPDF@data$roughness)
```

```
## [1] -0.1337503
```

```
cor(x = SPDF@data$Ca_exch, y = SPDF@data$slope)
```

```
## [1] -0.1351794
```

```
cor(x = SPDF@data$Ca_exch, y = SPDF@data$aspect)
```

```
## [1] -0.04609819
```

```
cor(x = SPDF@data$Ca_exch, y = SPDF@data$flowdir)
```

```
## [1] 0.1844461
```

```
cor(x = SPDF@data$Ca_exch, y = SPDF@data$CONVG2)
```

```
## [1] -0.2155261
```

```
cor(x = SPDF@data$Ca_exch, y = SPDF@data$SAGAWI)
```

```
## [1] 0.3242175
```

Die Variable Ca wurde, obwohl sie nicht normalverteilt ist, nicht transformiert, da durch eine Transformierung keine wesentliche Annäherung an die Normalverteilung erreicht wurde. Die Korrelationskoeffizienten zeigen zu den meisten Reliefparametern keinen signifikanten Zusammenhang der austauschbaren Ca-Ionen zu den

Reliefparametern. Zu der Demographie besteht ein leichter negativer Zusammenhang (-0,32) und zu der relativen Bodenfeuchte (SAGAWI) besteht ein leichter positiver Zusammenhang (0,32).

20 b)

Erstellen Sie mit Hilfe der Methode `scatterplotMatrix` eine Scatterplot-Matrix mit den am besten korrelierenden Kovariablen. Stellen Sie die jeweiligen Histogramme in der Diagonalen dar.

```
#install.packages("car")
library(car)

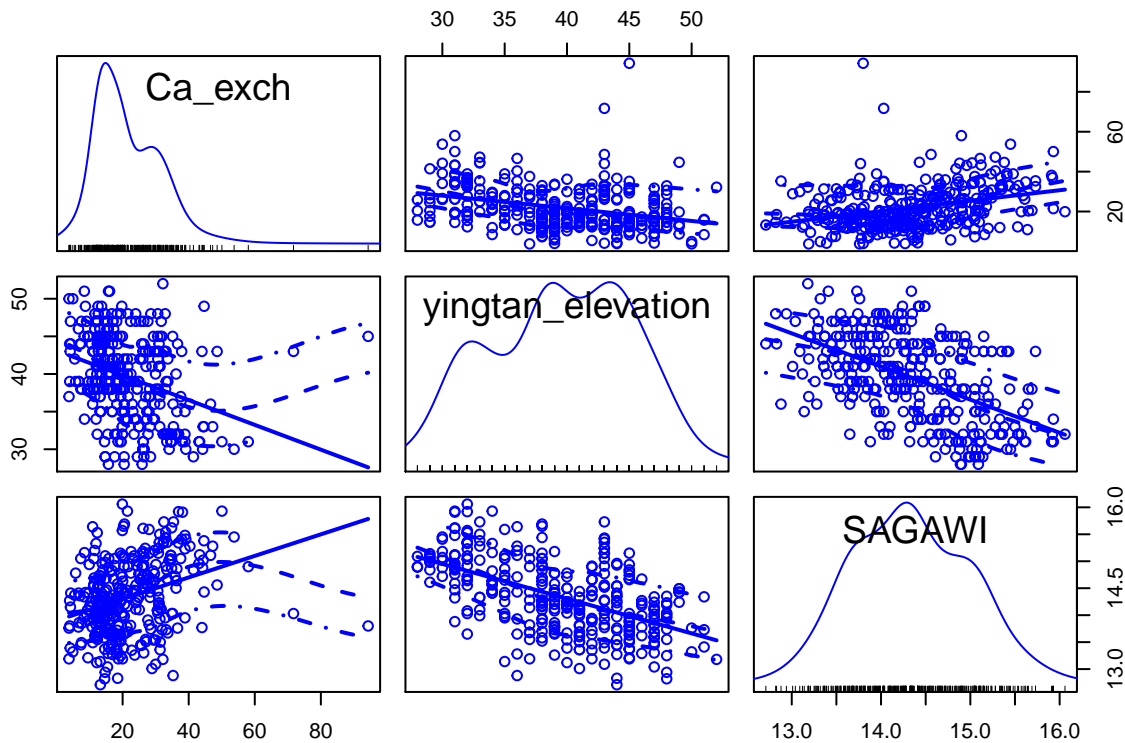
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##      recode

## The following object is masked from 'package:purrr':
##
##      some

##?scatterplotMatrix()
scatterplotMatrix(~Ca_exch + yingtian_elevation + SAGAWI,
                  #Ca_exch~yingtian_elevation+SAGAWI,
                  data=SPDF,
                  diagonal = TRUE,
                  #regLine=TRUE
                  #groups=,
                  #use=
                  )
```



20 c)

Beurteilen Sie abschließend, welche Reliefparameter als potentielle Kovariablen für eine Modellierung in Frage kommen und welche sich eher nicht eignen. Begründen Sie ihre Einschätzung anhand der bisherigen Ergebnisse dieses Aufgabenblattes.

Die Höhe steht in einer negativen Korrelation zur Zielvariablen. Je mehr die Höhe zunimmt, desto geringer ist der Gehalt an Ca_exch. Der Saga Wetness Index hingegen korreliert positiv mit der Zielvariable. Je stärker der Index, umso höher der Gehalt an Ca_exch. Der Konvergenz/ Divergenz Index ist steht ebenfalls in einer negativen Korrelation zur Zielvariable. Ebenfalls auffällig ist die Normalverteilung der Werte für SAGAWI & CONVG2.

Da sich die Werte der austauschbaren Ca-Ionen kaum bezüglich ihrer räumlichen Distanz (bezogen auf ihre geographische Lage zueinander) kaum ähneln, eignen sich die Koordinaten nicht für eine potenzielle Modellierung. Es scheint einen leichten Zusammenhang zwischen den austauschbaren Ca-Ionen und dem Höhenmodell sowie der relativen Bodenfeuchte (SAGAWI) zu geben. Deshalb eignen sich diese beiden Parameter als Kovariablen für eine Modellierung.

Die Höhe steht in einer negativen Korrelation zur Zielvariablen. Je mehr die Höhe zunimmt, desto geringer ist der Gehalt an Ca_exch. Der Saga Wetness Index hingegen korreliert positiv mit der Zielvariable. Je stärker der Index, umso höher der Gehalt an Ca_exch. Der Konvergenz/ Divergenz Index ist steht ebenfalls in einer negativen Korrelation zur Zielvariable. Ebenfalls auffällig ist die Normalverteilung der Werte für SAGAWI & CONVG2.

Aufgabe 21

21 a)

Führen Sie eine Variablenauswahl durch. Nutzen Sie die Rückwärtselimination der step-Funktion und beginnen Sie mit den Einflussgrößen, für die Sie sich in Aufg. 20c) entschieden haben.

```
##?lm()
linear <- lm(formula = Ca_exch~yingtan_elevation+SAGAWI,
  data = SPDF
  #na.action = ,
  #method = ,
  #model = ,
  #x = ,
  #y = ,
  #qr =
  )

##?step()
step(object = linear, #lm model
  #scope = , #defines the range of models examined in the stepwise search. This should be either a s
  #scale = , default 0 --> scale estimated: see extractAIC
  direction = "backward",
  trace = TRUE, #info printed during running
  #keep = ,
  #steps = , #maximum number of steps to be considered, default 1000
  #k = ,
  )
```

```
## Start: AIC=1570.16
## Ca_exch ~ yingtan_elevation + SAGAWI
##
##              Df Sum of Sq  RSS   AIC
## <none>                  35712 1570.2
## - yingtan_elevation    1    1306.1 37019 1580.2
## - SAGAWI                1    1336.4 37049 1580.5
##
## Call:
## lm(formula = Ca_exch ~ yingtan_elevation + SAGAWI, data = SPDF)
##
## Coefficients:
##      (Intercept)  yingtan_elevation          SAGAWI
##          -11.2069           -0.4121             3.4462
```

21 b)

Wenden Sie die Methode summary auf das lm-Objekt aus Aufgabe a) an und beschreiben Sie in knappen Worten, wofür die dargestellten Statistiken stehen. Was sagen die Werte konkret aus?

```
summary(linear)

##
## Call:
## lm(formula = Ca_exch ~ yingtan_elevation + SAGAWI, data = SPDF)
##
## Residuals:
```

```
##      Min      1Q  Median      3Q      Max
## -18.573  -6.927  -1.489   5.430  76.505
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -11.2069     16.9843  -0.660 0.509816
## yingtang_elevation -0.4121      0.1183  -3.485 0.000559 ***
## SAGAWI           3.4462      0.9777   3.525 0.000483 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.37 on 332 degrees of freedom
## Multiple R-squared:  0.1367, Adjusted R-squared:  0.1315
## F-statistic: 26.28 on 2 and 332 DF,  p-value: 2.533e-11
```

Der Call zeigt an, welche Kovariablen verwendet wurden. Wir haben den Ca_exch mit Elevation und SAGAWI korreliert.

Residuals: Fehler zwischen Modellierung und tatsächlich beobachteten Werten. Minimumwert, Maximumwert, median Wert und das erste und dritte Quartil der Fehler wurden ausgegeben. Es beträgt der Interquartilabstand 12,357. Im Zusammenhang mit dem Maximalwert 76,5, könnte man annehmen, dass es besser wäre, wenn der Maximalwert geringer wäre.

Coefficients: Estimate beschreibt den zu verwendenden Faktor in unserer Regressionsgleichung: $f(x) = -11,2069 - 0,4121(\text{elevation}) + 3,4462*(\text{SAGAWI})$

Den Achsenschnittpunkt mit der Y-Achse beschreibt intercept: -11,2069. Der Wetness-Index ist stärker gewichtet als die elevation.

Std. Error: Std.-Error beschreibt die jeweilige Standardabweichung (hoch bei intercept).

t value: t-value gibt den Ergebnisswert des t-Testes an.

Pr(>|t|): Pr-Wert ist zur Überprüfung der Nullhypothese. Das dazu zu verwendende Signifikanzniveau wird durch die * angezeigt.

Residual standard error: er Residual Standard Error ist die Standardabweichung der Residuen. Hier wird auch die Anzahl der Freiheitsgrade angegeben (332).

degrees of freedom:

Multiple R-squared/Adjusted R-squared: Die Werte multiple R-squared und adjusted R-squared geben die jeweiligen Werte für das zentrierte und das unzentrierte R^2 an.

F-statistic:

p-value:

Notieren Sie das Bestimmtheitsmaß der resultierenden Regression. Aus welchen Kovariablen setzt sich das abschließende Regressionsmodell zusammen und sind diese signifikant?

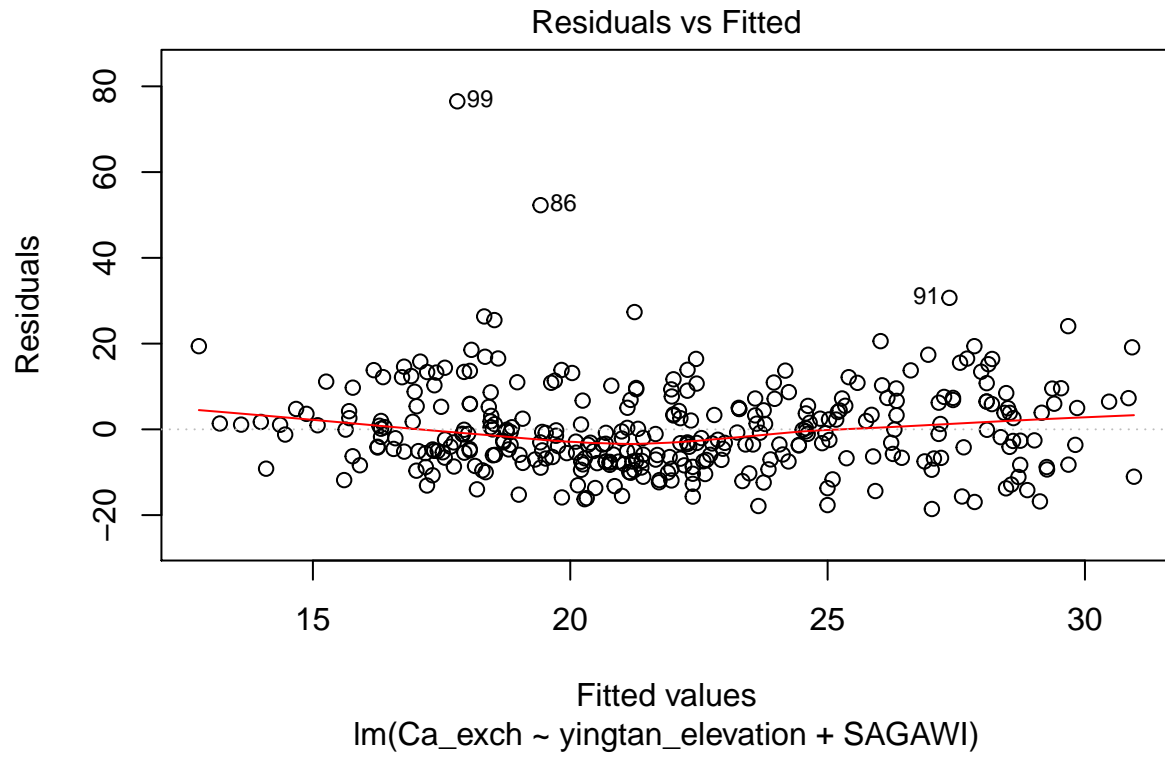
Bestimmtheitsmaß: R-squared, R^2

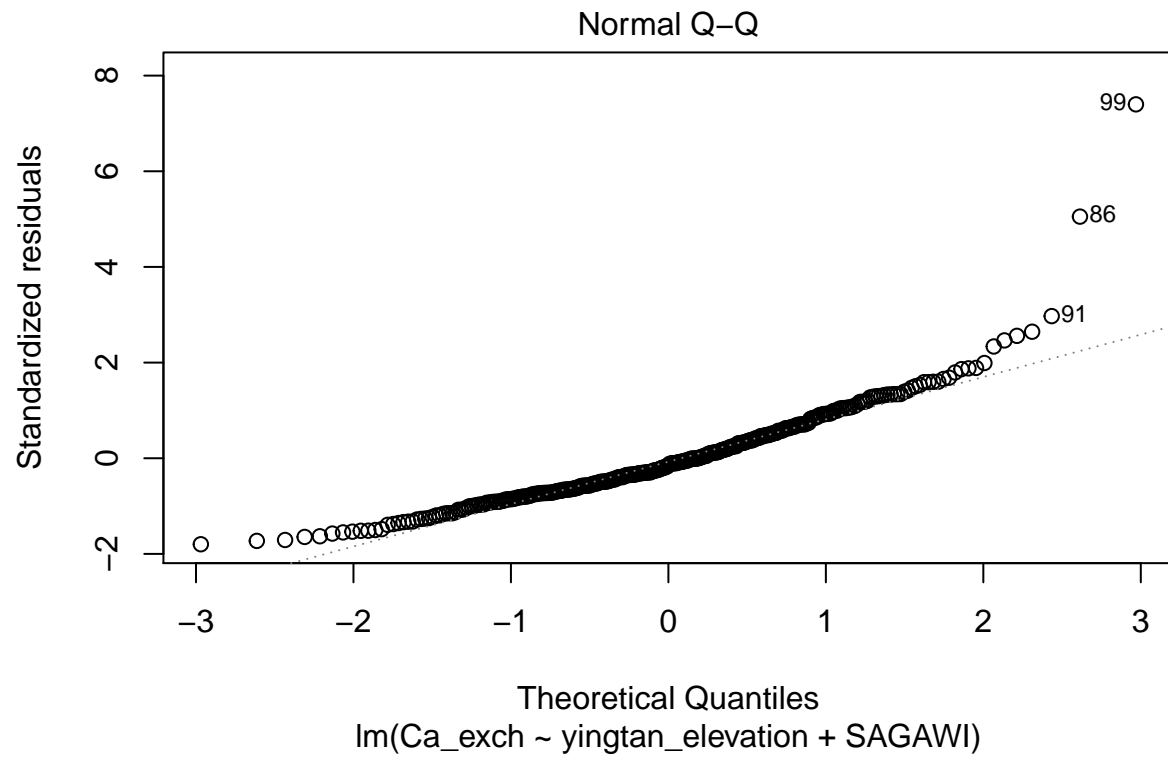
Kovariablen des abschließenden Regressionsmodells: Signifikanz: Ja, da P-Wert < 0.05

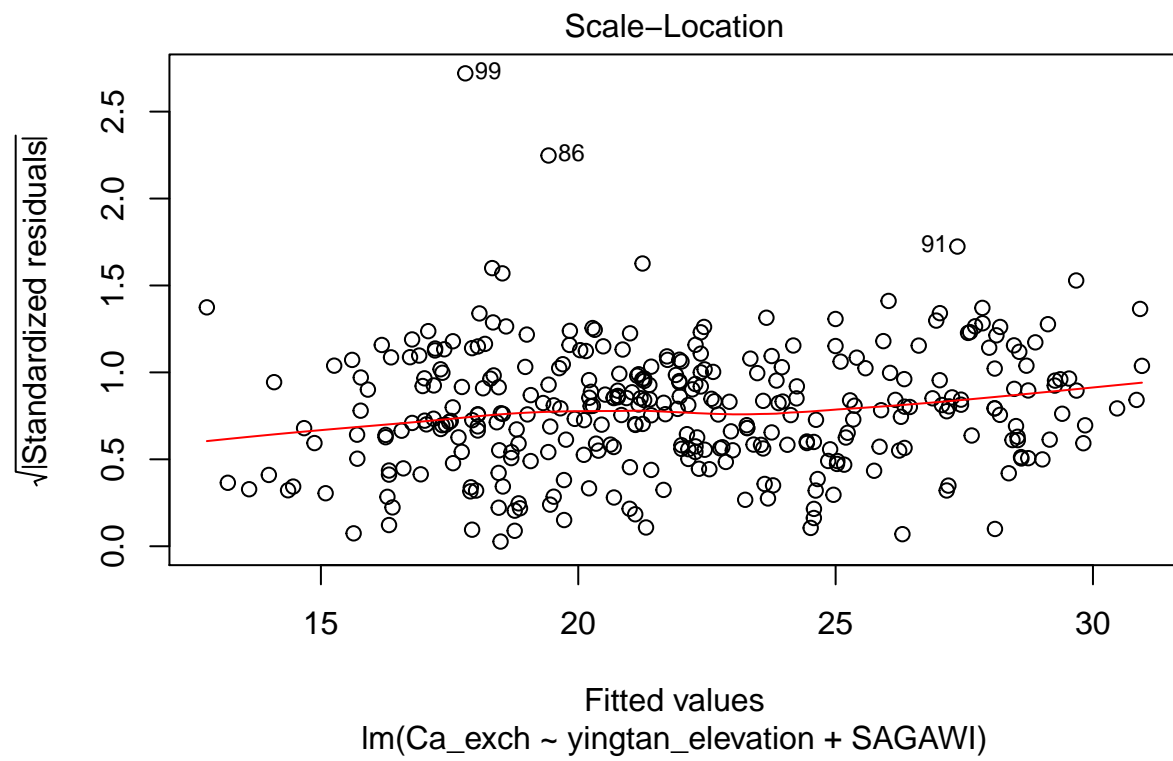
21 c)

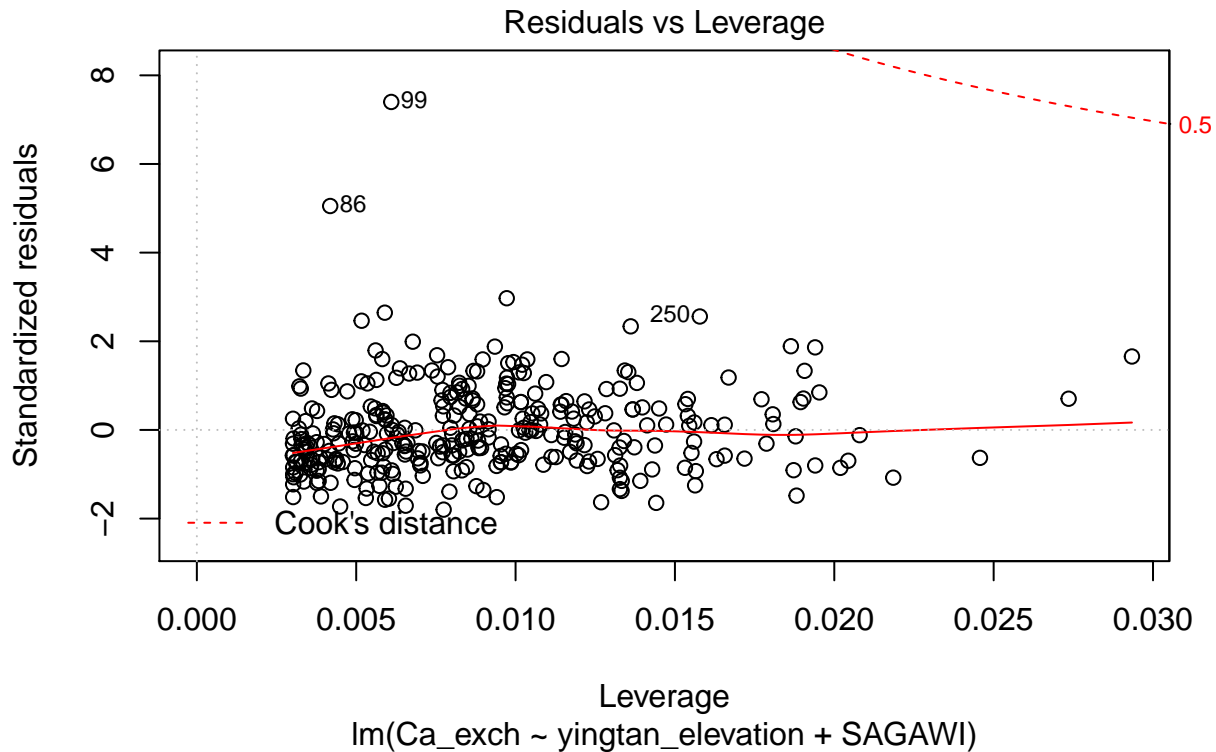
Untersuchen Sie ihren Fit aus a) hinsichtlich • fehlender Normalverteilung der Residuen • Ausreißern und high-leverage points • Heteroskedastizität • nicht-linearer Regression. Nennen Sie alle Annahmeverletzungen, die Sie finden können. Begründen Sie ihre Auswahl

```
plot(linear)
```









```
outlierTest(linear)
```

```
##      rstudent unadjusted p-value Bonferroni p
## 99 8.084562      1.1867e-14    3.9756e-12
## 86 5.249971      2.7257e-07    9.1310e-05
```

- fehlender Normalverteilung der Residuen

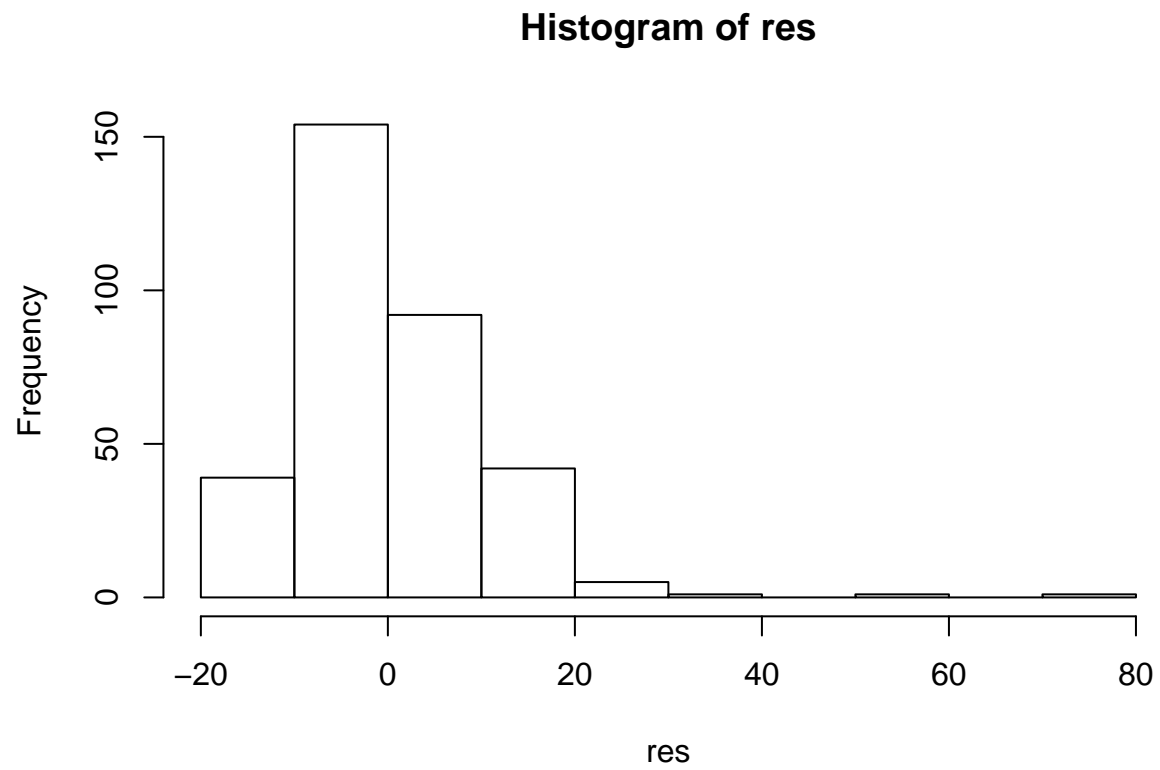
```
# Residuen
res <- resid(linear)
# Normalverteilung der Residuen
shapiro.test(res)
```

```
##
## Shapiro-Wilk normality test
##
## data:  res
## W = 0.89278, p-value = 1.327e-14
```

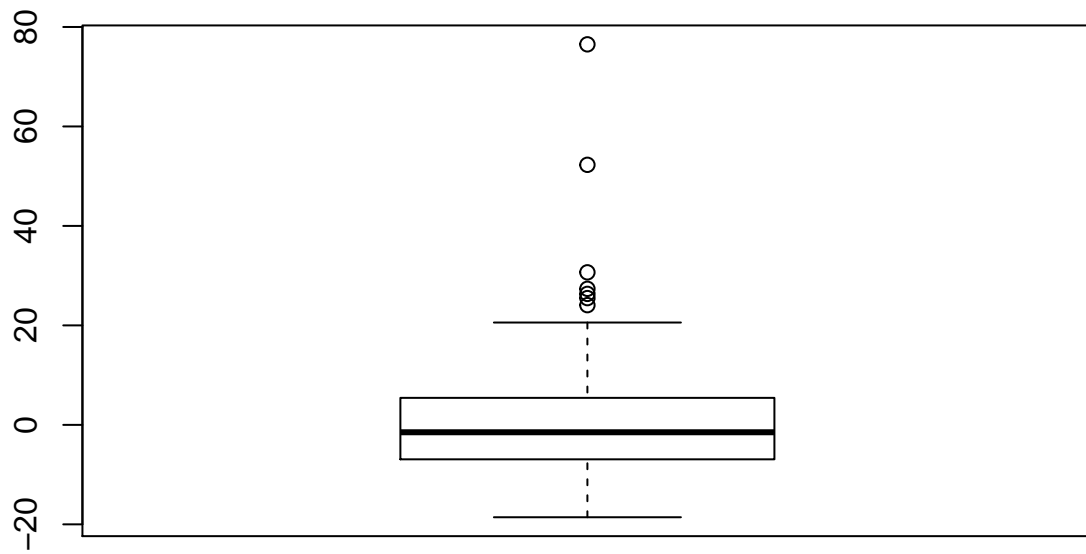
Der W-Wert ist mit 0,89 zwar dicht an 1, aber der p-Wert ist wesentlich kleiner 0,05. Die Residuen sind nicht normalverteilt.

- Ausreißern und high-leverage points

```
# Verteilung der Residuen  
hist(res)
```

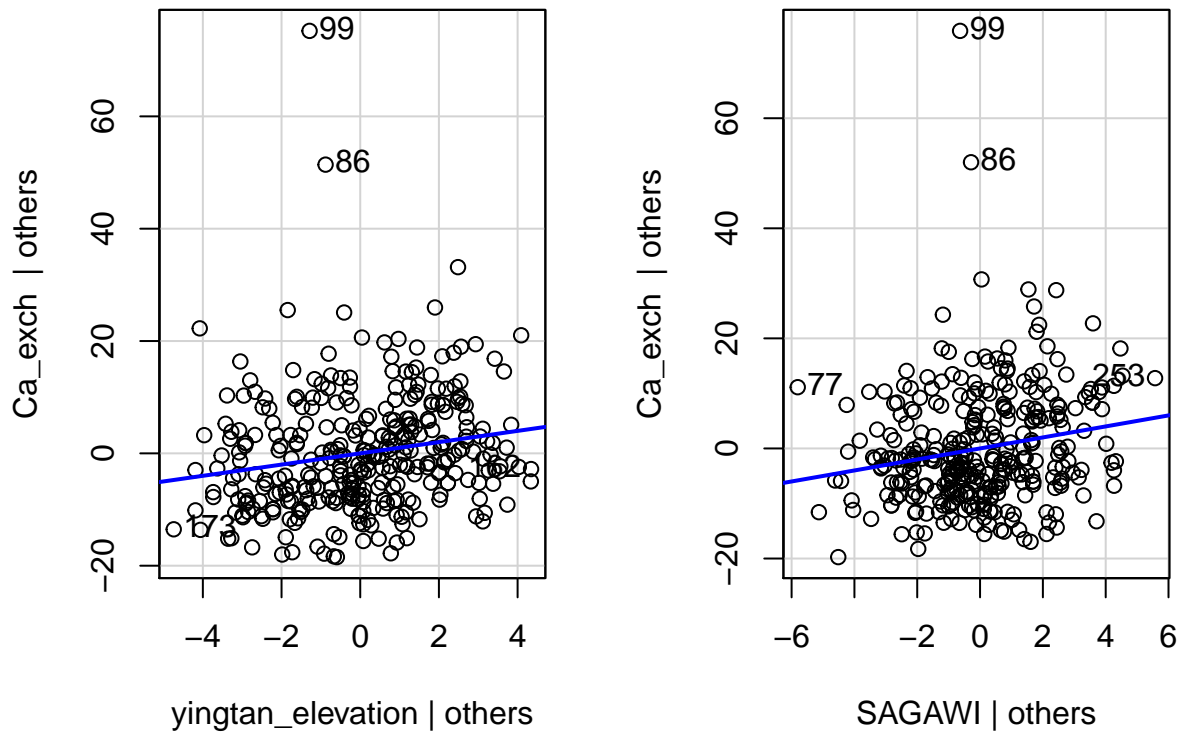


```
boxplot(res)
```



```
# Ausreißer  
out <- ifelse(res < (mean(res) - 2*sd(res)) | res > (mean(res) + 2*sd(res)), 1, 0)  
#out  
# high-leverage points  
leveragePlots(model = linear)
```

Leverage Plots



Werte > 20 werden als Ausreißer dargestellt. Die Werte 86 und 99 sind high leverage points.

- Heteroskedastizität

```
# Breusch-Pagan Test
#install.packages("AER")
library(AER)
```

```
## Loading required package: lmtest
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##   as.Date, as.Date.numeric
## Loading required package: sandwich
## Loading required package: survival
bptest(linear)
```

```
##
## studentized Breusch-Pagan test
##
## data: linear
## BP = 0.57065, df = 2, p-value = 0.7518
```

```
ncvTest(linear)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.417208, Df = 1, p = 0.23386
```

- nicht-linearer Regression.

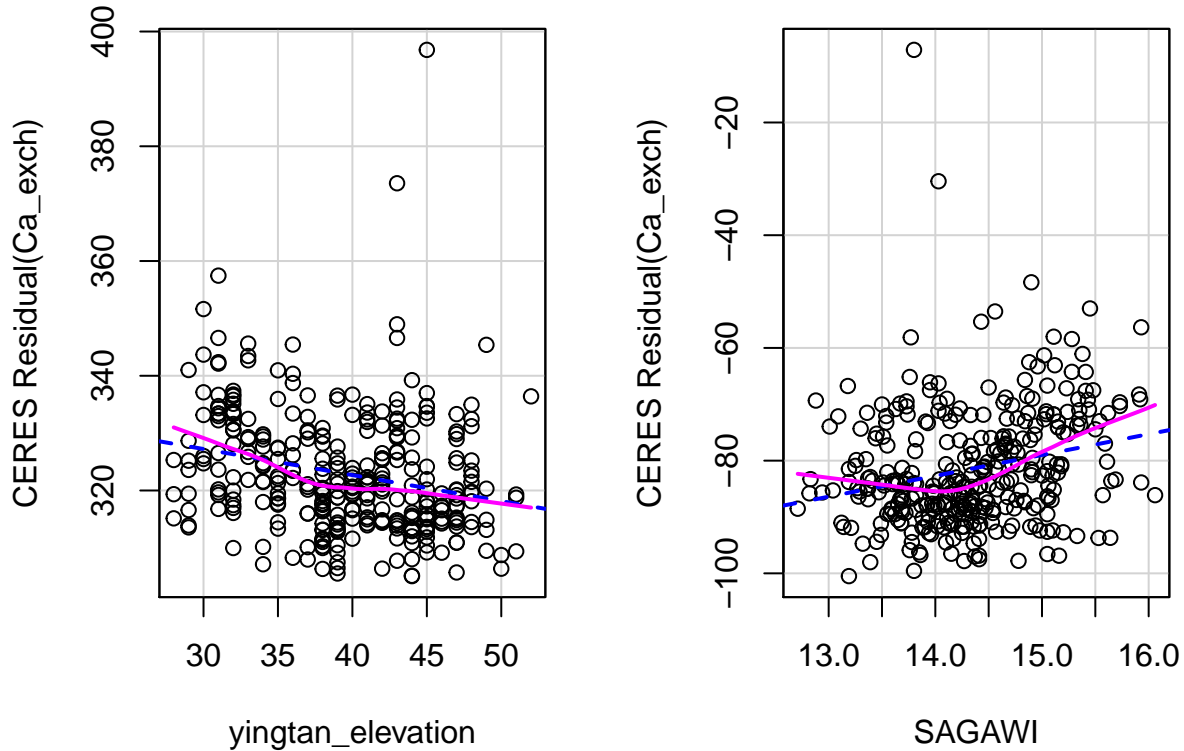
```
# nicht-lineare Regression
```

```
summary(linear)
```

```
##
## Call:
## lm(formula = Ca_exch ~ yingtang_elevation + SAGAWI, data = SPDF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.573  -6.927  -1.489   5.430  76.505
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -11.2069     16.9843  -0.660 0.509816
## yingtang_elevation -0.4121      0.1183  -3.485 0.000559 ***
## SAGAWI          3.4462      0.9777   3.525 0.000483 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.37 on 332 degrees of freedom
## Multiple R-squared:  0.1367, Adjusted R-squared:  0.1315
## F-statistic: 26.28 on 2 and 332 DF,  p-value: 2.533e-11
```

```
ceresPlots(linear)
```

CERES Plots



#<<https://www.statmethods.net/stats/riagnostics.html>>

$R^2 = 0,14$ -> mit dem Modell werden knapp 14% der Werte für austauschbare Ca-Ionen erklärt. Es scheint noch weitere Parameter zu geben, die Einfluss auf die Zielgröße haben.

21 d)

Führen Sie nun eine Vorhersage mit dem generierten Regressionsmodell durch. Nutzen Sie das Objekt „terrain“ des geladenen Workspace als ZielGrid. Ermitteln Sie anschließend den RMSE dieser Methode, indem Sie eine LOO-Kreuzvalidierung durchführen.

```
var <- variogram(Ca_exch~yingtian_elevation+SAGAWI,
  SPDF,
  cutoff=2202, # ca. 50% der Max. Distanz
  width= 100
)

m <- vgm(#psill = 110,
  model = "Exp",
  #model = c("Nug", "Exp", "Log", "Gau"),
  #range = 300,
  #kappa= 2,
  #nugget = 2,
  cutoff= 2202)

v_fit <- fit.variogram(object = var,
  model = m,
```

```

        #fit.sills = TRUE,
        #fit.ranges = TRUE,
        fit.method = 7, #vgl. Gstat user's manual,
                        #p. 42, tab. 4.2,
                        #<http://www.gstat.org/gstat.pdf>
                        # 7 ordinary least squares
        #fit.kappa = TRUE,
    )

krig <- gstat::krige(Ca_exch~yingtan_elevation+SAGAWI,
                    locations = SPDF,
                    newdata = terrain,
                    model = v_fit)

```

```
## [using universal kriging]
```

```
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO"): Discarded datum WGS_1984 in CRS d
## but +towgs84= values preserved
```

```

LOOCV <- krige.cv(Ca_exch~yingtan_elevation+SAGAWI, #statt SPDF@data$Ca_exch~1
                 locations = SPDF,
                 model= v_fit,
                 #maxdist = 2202
                 )

```

```
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO"): Discarded datum WGS_1984 in CRS d
## but +towgs84= values preserved
```

```
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO"): Discarded datum WGS_1984 in CRS d
## but +towgs84= values preserved
```

```
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO"): Discarded datum WGS_1984 in CRS d
## but +towgs84= values preserved
```

```
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO"): Discarded datum WGS_1984 in CRS d
## but +towgs84= values preserved
```

```
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO"): Discarded datum WGS_1984 in CRS d
## but +towgs84= values preserved
```

```
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO"): Discarded datum WGS_1984 in CRS d
## but +towgs84= values preserved
```

```
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO"): Discarded datum WGS_1984 in CRS d
## but +towgs84= values preserved
```

```
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO"): Discarded datum WGS_1984 in CRS d
## but +towgs84= values preserved
```

```
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO"): Discarded datum WGS_1984 in CRS d
## but +towgs84= values preserved
```

```
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO"): Discarded datum WGS_1984 in CRS d
## but +towgs84= values preserved
```



```
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO"): Discarded datum WGS_1984 in CRS d
## but +towgs84= values preserved
```

```
RMSE <- function(residuals){  
  sqrt(sum((residuals)^2)/length(residuals))  
}
```

```
RMSE(residuals = LOOCV@data$residual)
```

```
## [1] 9.298445
```

```
bubble(LOOCV, "residual", main = "ordinary kriging Ca-exch LOOCV")
```

