Aufgabe 09

Gruppe 01

28 6 2020

Laden Sie den Workspace yingtan_20_ueb9.Rdata sowie das Paket car. Sie werden feststellen, dass der Workspace Variablen enthält, die denen gleichen, welche Sie in der letzen Sitzung erzeugt haben.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.2
```

```
## 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
```

```
load("data/yingtan_20_ueb9.Rdata")
```

Aufgabe 20 Korrelation

Korrelation beschreibt die Richtung und Stärke des Zusammenspiels zweier kontinuierlicher Variablen.

a) Berechnen Sie - soweit möglich - geeignete Korrelationskoeffizienten für die austauschbaren Ca-Ionen (evtl. transformiert) mit sämtlichen Reliefgrößen. (1 Punkt)

```
# Variablen
Ca <- ljz$Ca_exch
```

```
## Loading required package: sp
```

```
elev <- ljz$yingtan_elevation
tri <- ljz$tri
```

```r
tpi <- ljz$tpi
rough <- ljz$roughness
slope <- ljz$slope
aspect <- ljz$aspect
flowdir <- ljz$flowdir
convg2 <- ljz$CONVG2
sagawi <- ljz$SAGAWI

# Korrelationskoeffiezienten
cor(Ca, elev)
```

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

```r
cor(Ca, tri)
```

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

```r
cor(Ca, tpi)
```

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

```r
cor(Ca, rough)
```

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

```r
cor(Ca, slope)
```

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

```r
cor(Ca, aspect)
```

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

```r
cor(Ca, flowdir)
```

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

```r
cor(Ca, convg2)
```

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

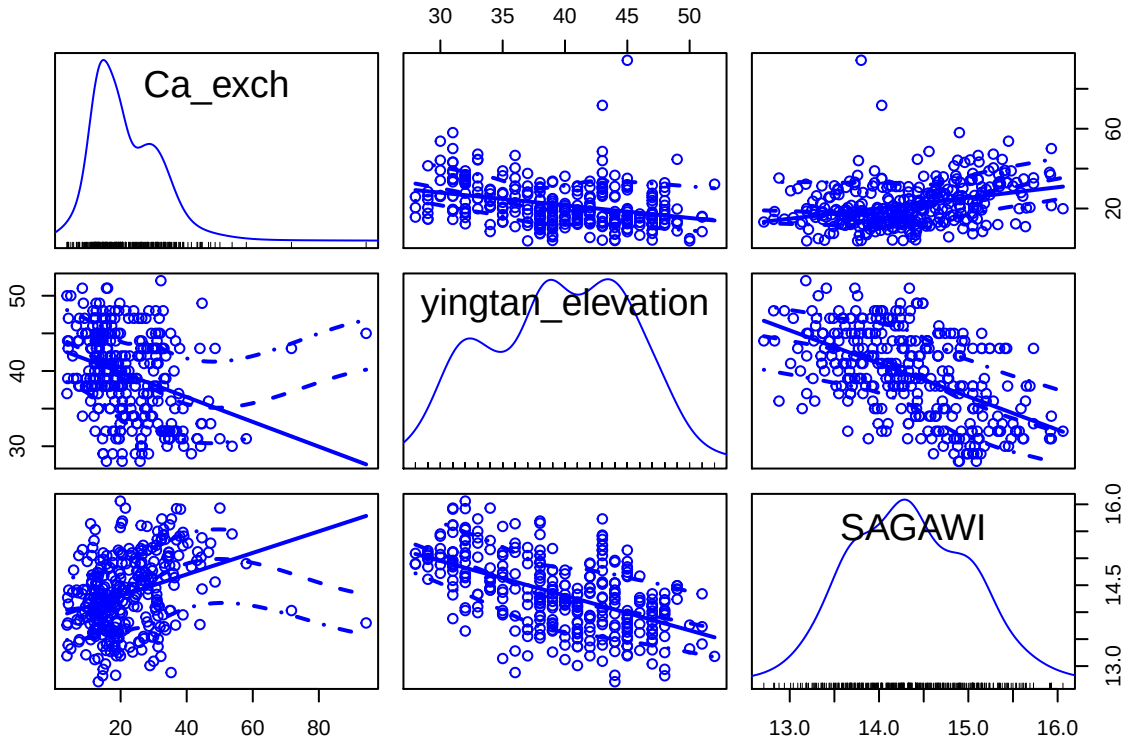
```r
cor(Ca, sagawi)
```

```
## [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).

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. (1 Punkt)

```r
scatterplotMatrix(~Ca_exch + yingtan_elevation + SAGAWI,
                  data = ljz,
                  diagonal = TRUE,)
```



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. (2 Punkte)

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 Multiple Linear Regression

Mit dem Modell der Linearen Regression lassen sich Zielgrößen durch einen oder mehrere Einflussparameter abbilden:

Wesentliche Arbeitsschritte bei der multiplen linearen Regression sind die Auswahl geeigneter Kovariablen sowie die Überprüfung der Modellannahmen im Anschluss an die Modellierung.

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. (1 Punkt)

```
model <- lm(Ca ~ elev + sagawi)
step(object = model,
     direction = "backward")
```

```
## Start:  AIC=1570.16
## Ca ~ elev + sagawi
##
##          Df Sum of Sq   RSS    AIC
```

```
## <none>                35712 1570.2
## - elev    1    1306.1 37019 1580.2
## - sagawi  1    1336.4 37049 1580.5
```

```
##
## Call:
## lm(formula = Ca ~ elev + sagawi)
##
## Coefficients:
## (Intercept)         elev       sagawi
##    -11.2069      -0.4121       3.4462
```

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? Notieren Sie das Bestimmtheitsmaß der resultierenden Regression. Aus welchen Kovariablen setzt sich das abschließende Regressionsmodell zusammen und sind diese signifikant? (2 Punkte)

```
summary(model)
```

```
##
## Call:
## lm(formula = Ca ~ elev + sagawi)
##
## 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
## elev         -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
```

Call: 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(elevation) +3,4462*(SAGAWI)

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

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

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

Pr(>|t|): Das in der Modellausgabe gefundene Akronym Pr(>t) bezieht sich auf die Wahrscheinlichkeit, dass ein beliebiger Wert gleich oder größer als t beobachtet wird. (https://feliperego.github.io/blog/2015/10/23/I

nterpreting-Model-Output-In-R)

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

degrees of freedom: Die Anzahl der Werte in der endgültigen Berechnung einer Statistik, die frei variieren können. Eine gebräuchliche Art und Weise, sich Freiheitsgrade vorzustellen, ist die Anzahl der unabhängigen Werte, die zur Schätzung eines anderen Wertes zur Verfügung stehen. Genauer gesagt ist die Anzahl der Freiheitsgrade die Anzahl der unabhängigen Beobachtungen in einer Stichprobe von Daten, die zur Schätzung eines Parameters der Grundgesamtheit, aus der diese Stichprobe gezogen wird, zur Verfügung stehen. (https://www.uni-regensburg.de/wirtschaftswissenschaften/vwl-tschernig/medien/mitarbeiter/rameseder/moe_interpretationroutput.pdf)

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

F-statistic: Dieser Wert gibt an, ob ein Zusammenhang zwischen Ca_exch und den beiden Kovariablen besteht. Der Wert sollte möglichst weit von 1 entfernt sein und muss in Zusammenhang mit den Datensätzen und dem p-value betrachtet werden. Hier ist der Wert 26,28 und der p-Wert ist $< 0{,}05$. Es scheint einen leichten Zusammenhang zu den Kovariablen zu geben.

p-value: Überschreitungswahrscheinlichkeit, oder Signifikanzwert. Die beste Wahrscheinlichkeit, Testergebnisse zu erhalten, die mindestens so extrem sind wie die tatsächlich beobachteten Ergebnisse, unter der Annahme, dass die Nullhypothese richtig ist. Ein kleinerer p-Wert bedeutet, dass es stärkere Evidenz zugunsten der Alternativhypothese gibt. Hier scheint es einen Zusammenhang zwischen den Ca-exch-Werten und den Kovariablen zu geben.

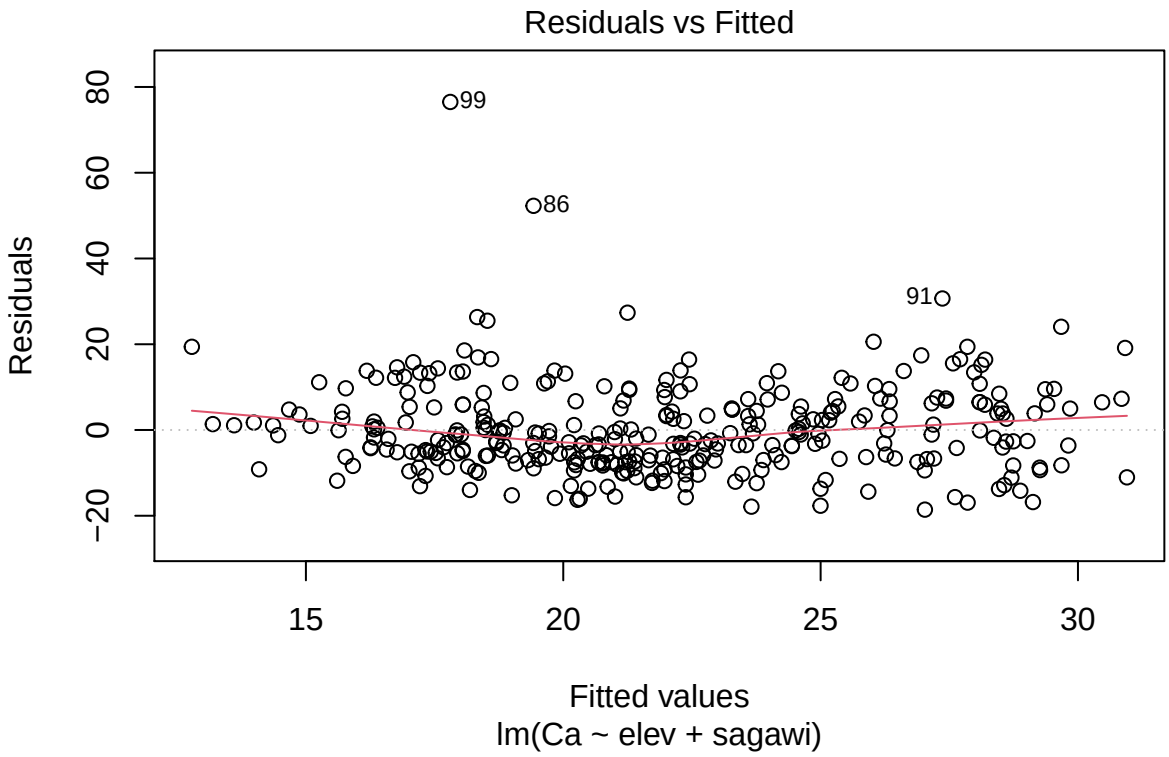
Notieren Sie das Bestimmtheitsmaß der resultierenden Regression. Aus welchen Kovariablen setzt sich das abschließende Regressionsmodell zusammen und sind diese signifikant?

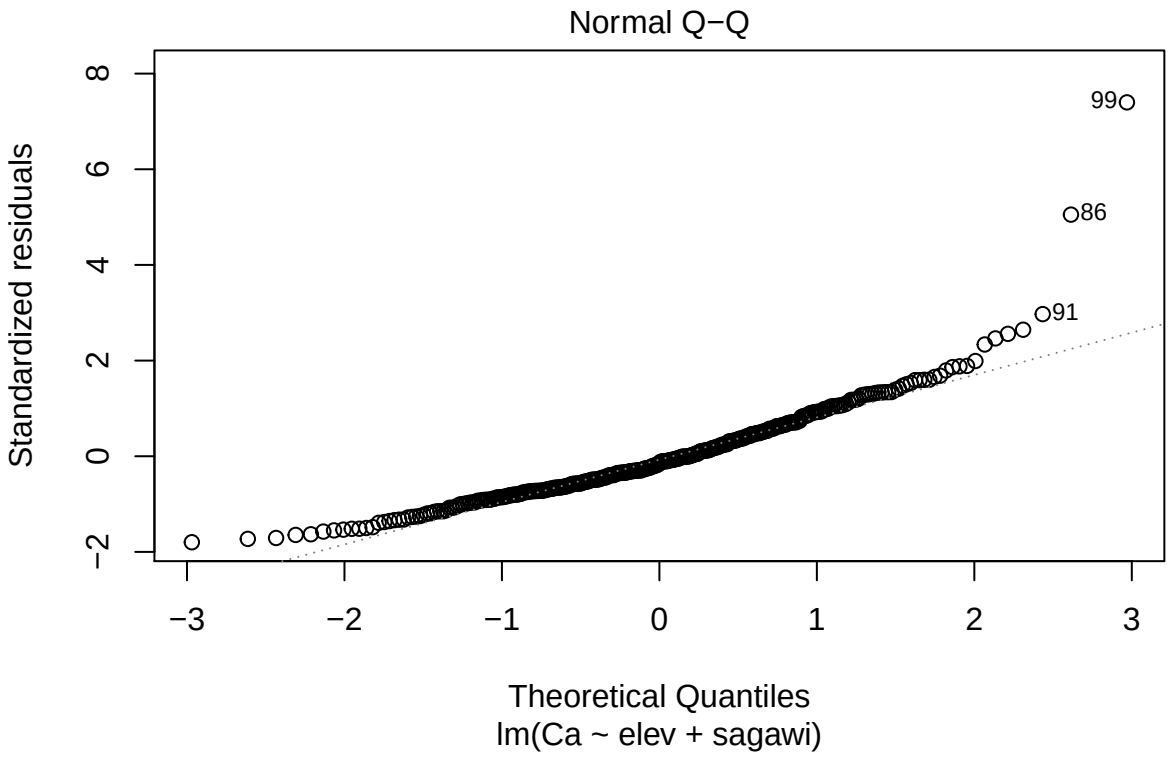
Bestimmtheitsmaß: R-squared, R^2 = 0,1315 Durch das Höhenmodell und den SAGA-Wetness-Index können etwa 13% der Werte der austauschbaren Ca-Ionen erklärt werden.

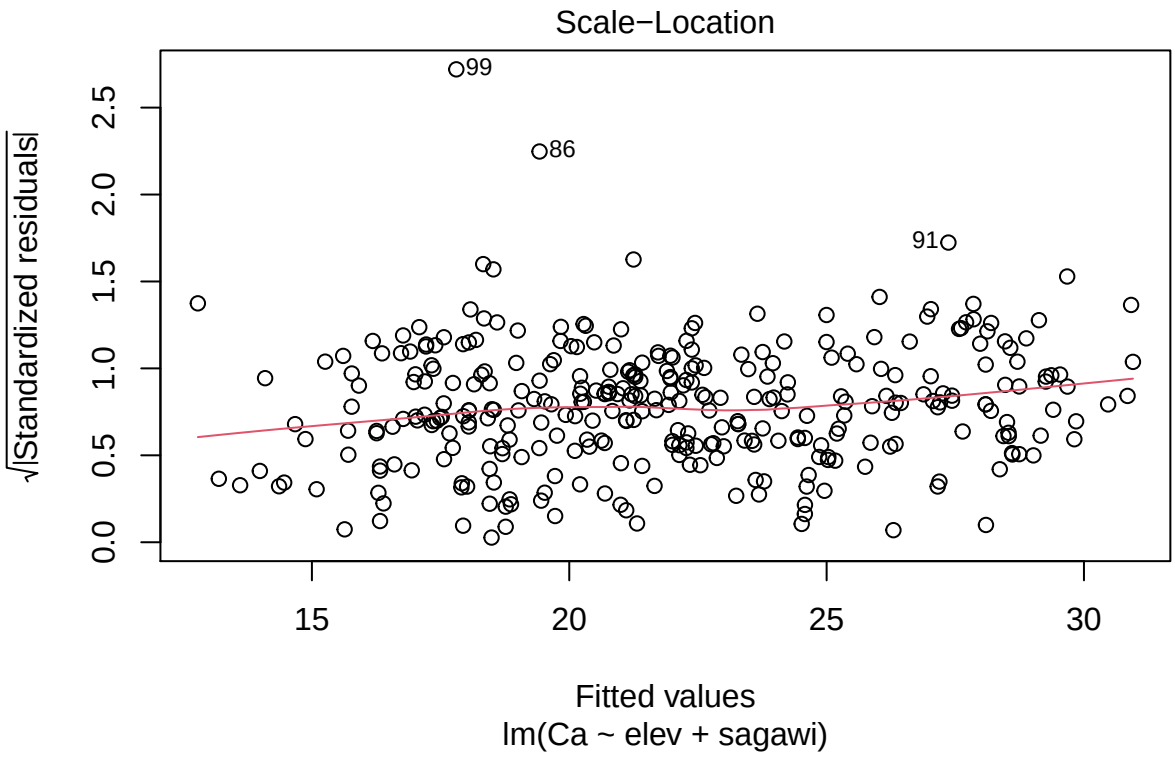
Kovariablen des abschließende Regressionsmodell: Signifaganz: Ja, da P-Wert < 0.05. Es besteht ein Zusammenhang zwischen den Ca_exch-Werten und den ausgewählten Kovariablen.

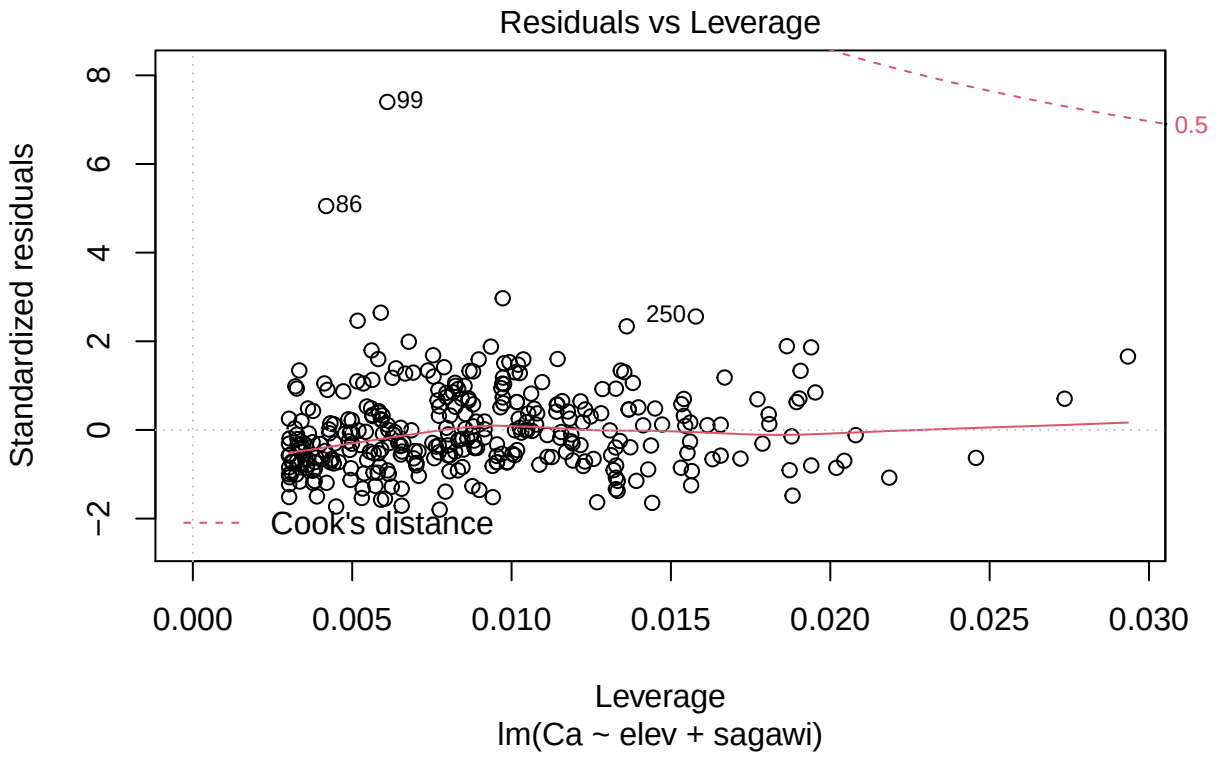
c) Untersuchen Sie ihren Fit aus a) hinsichtlich • fehlender Normalverteilung der Residuen • Ausreißern und high-leverage points • Heteroskedastizität • nicht-linearer Regression. Nennen Sie alle Annahmeverletzungen, die Sie finden können. BegründenSie ihre Auswahl. (4 Punkte)

```
plot(model)
```


Residuals vs Fitted
Residuals
80
60
40
20
0
−20
99
86
91
15
20
25
30
Fitted values
lm(Ca ~ elev + sagawi)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Ca ~ elev + sagawi)
99
86
91
−3
−2
−1
0
1
2
3
−2
0
2
4
6
8


Scale−Location
99
86
91
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
15
20
25
30
Fitted values
lm(Ca ~ elev + sagawi)



- fehlender Normalverteilung der Residuen

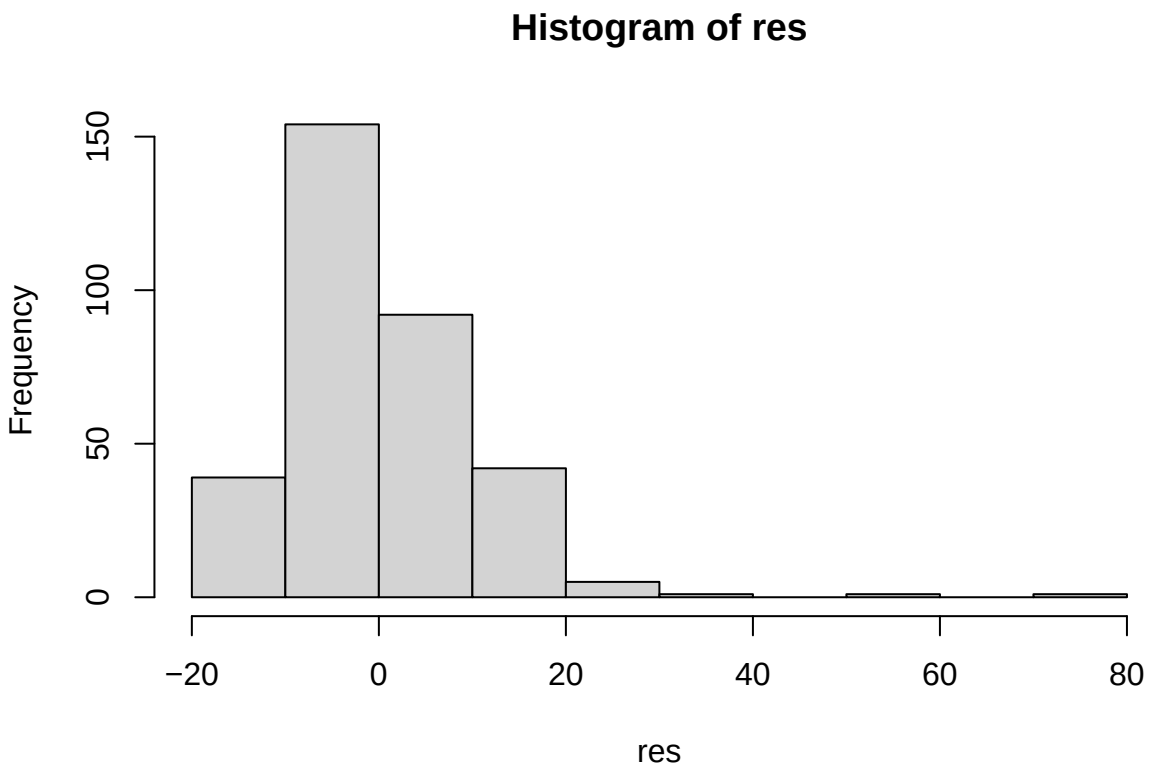
```
# Residuen
res <- resid(model)

# Normalverteilung der Residuen
shapiro.test(res)
```

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

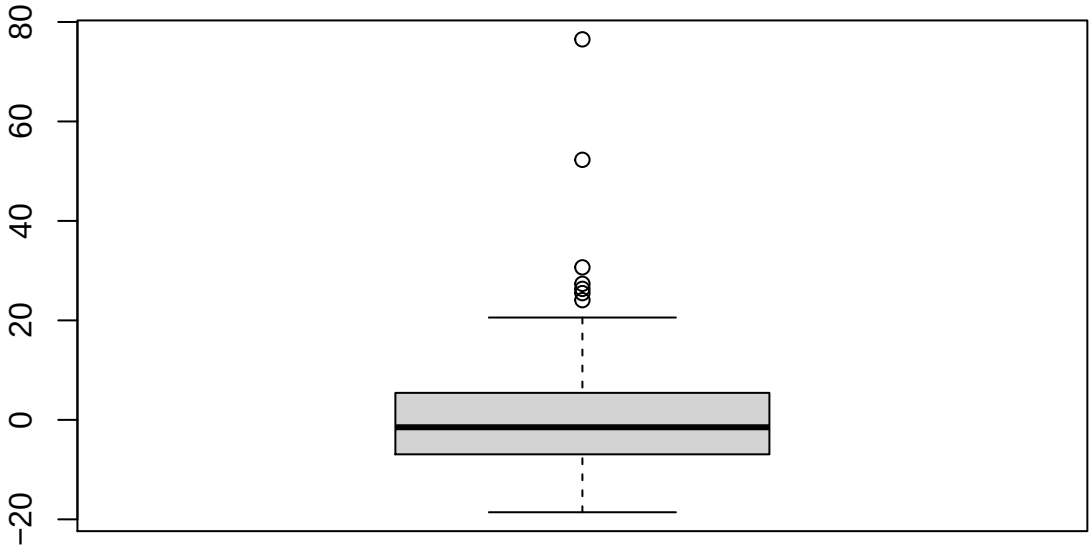
Der W-Wert ist mit 0,89 zwar dicht an 1, aber der p-Wert ist wesentlich kleiner 0,05. Die Residuen sind nicht normalverteilt.

- Ausreißern und high-leverage points

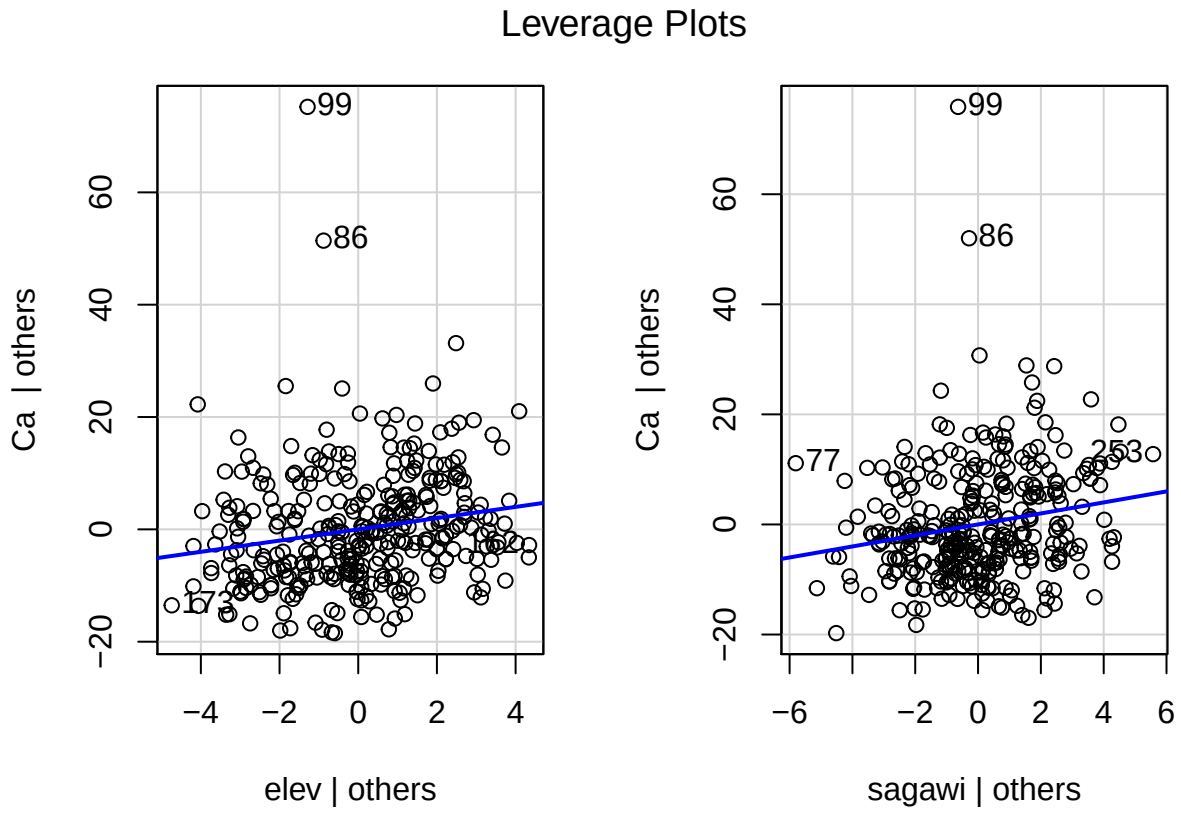
```
# Verteilung der Residuen
hist(res)
```




```
boxplot(res)
```




```
# high-leverage points
leveragePlots(model = model)
```




```
outlierTest(model)
```

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

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

- Heteroskedastizität

```
# Breusch-Pragan Test
library(AER)
```

```
## Warning: package 'AER' was built under R version 4.0.2
```

```
## Loading required package: lmtest
```

```
## Warning: package 'lmtest' was built under R version 4.0.2
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Warning: package 'sandwich' was built under R version 4.0.2
```

```
## Loading required package: survival
```

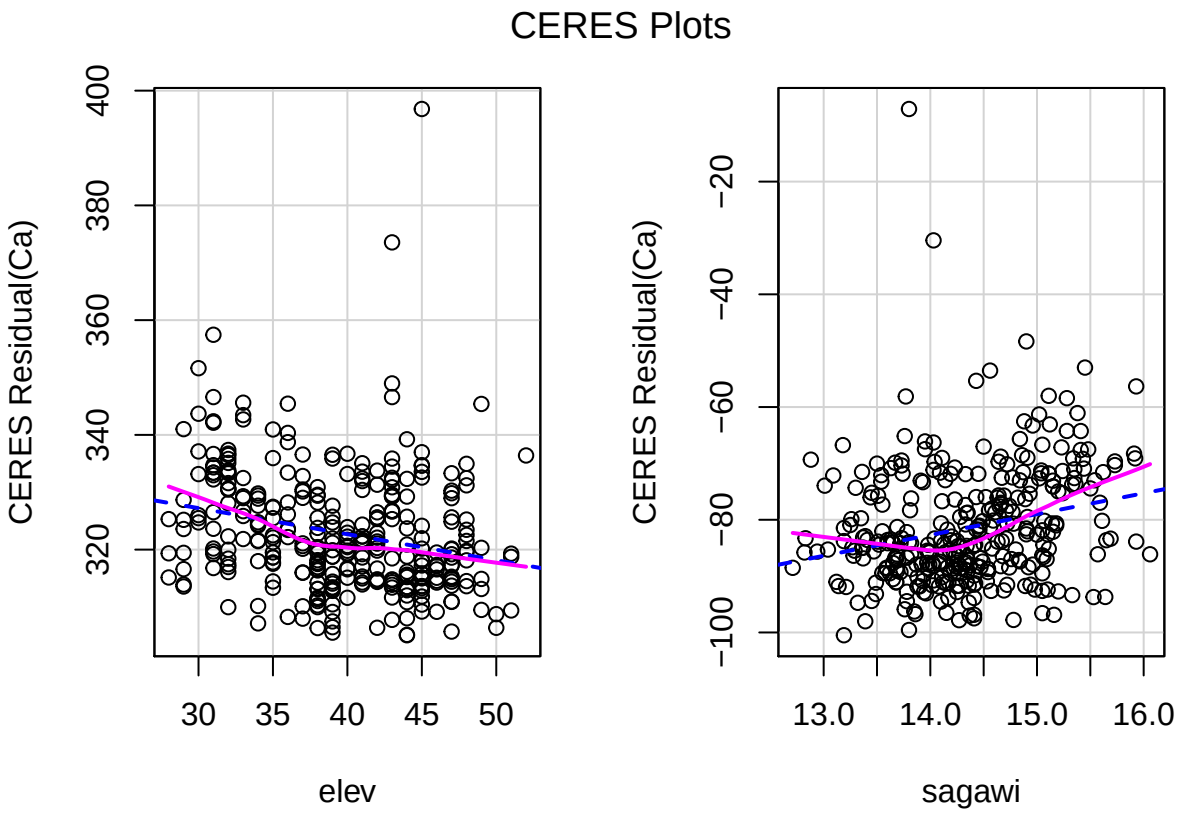
```
bptest(model)
```

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

Der p-Wert ist $> 0{,}05$, es liegt also eine Heteroskedastizität vor.

- nicht-linearer Regression.

```
# nicht-lineare Regression
ceresPlots(model)
```



Die Funktionen der tatsächliche Werte und die der Residuen unterscheiden sich nur leicht. Außer im oberen Wertebereich des SAGAWI unterscheiden sich die Werte kaum. 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 (s. summary).

d) Führen Sie nun eine Vorhersage mit dem generierten Regressionsmodell durch. Nutzen Sie das Objekt „terrain“ des geladenen Workspace als Ziel-Grid. Ermitteln Sie anschließend den RMSE dieser Methode, indem Sie eine LOO-Kreuzvalidierung durchführen. (2 Punkte)

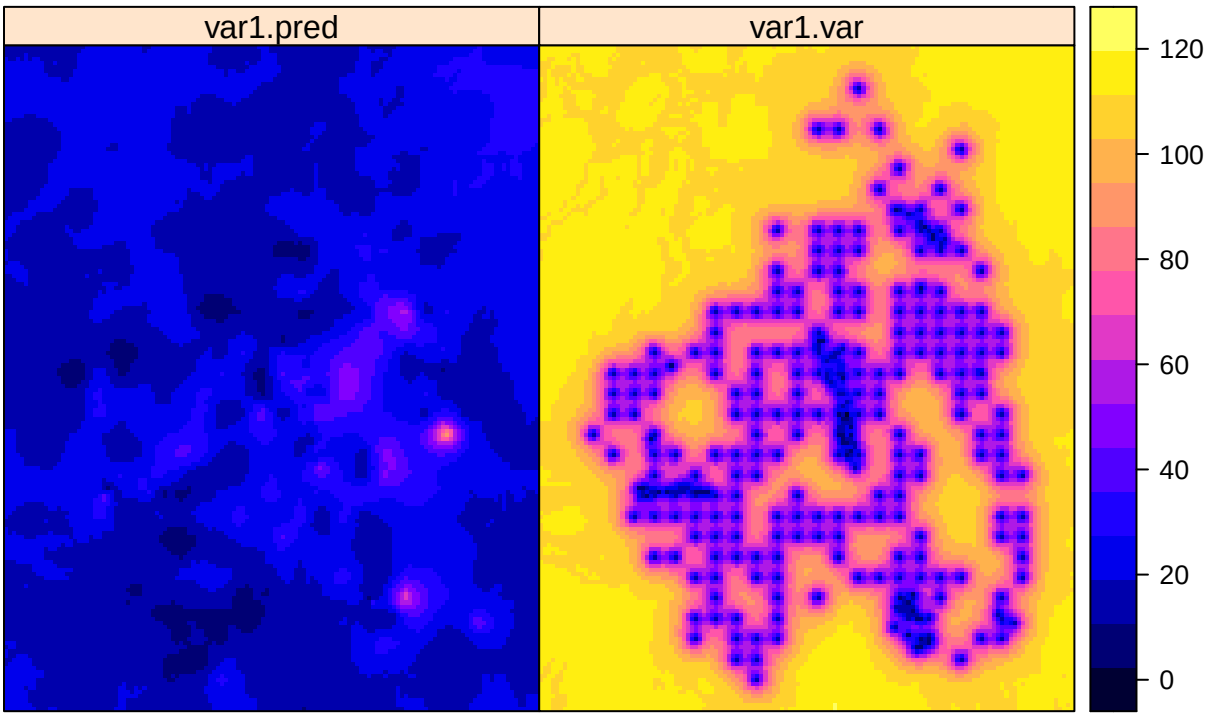
```
library(gstat)
variomodel <- variogram(Ca ~ elev + sagawi,
                        data = ljz,
                        cutoff = 2202,
```

```
                          width = 150)

m <- vgm(model = "Exp",
         cutoff = 2202)

fvariomodel <- fit.variogram(variomodel,
                             model = m,
                             fit.method = 7)

# Ordinary Kriging
interpol_Ca <- krige(Ca_exch ~ yingtan_elevation + SAGAWI,
                     ljz,
                     terrain,
                     model = fvariomodel)
```

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

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

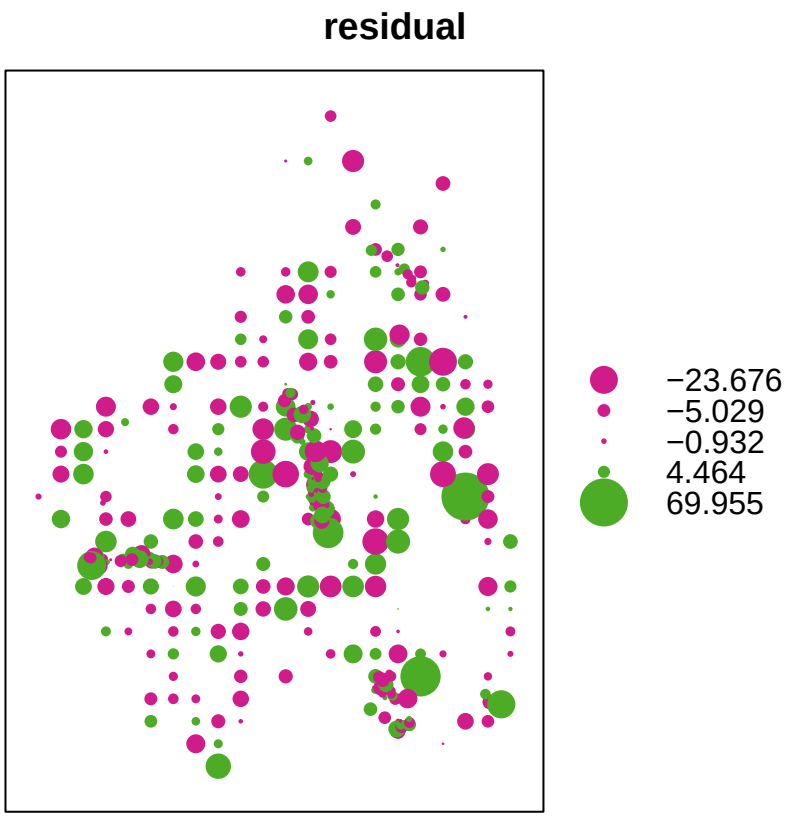
```
spplot(interpol_Ca)
```




```
# Koordinatensystem anpassen
proj4string(terrain) <- CRS("+proj=utm +zone=50 +ellps=WGS84 +datum=WGS84")
```

```
## Warning in `proj4string<-`(`*tmp*`, value = new("CRS", projargs = "+proj=utm +zone=50 +datum=WGS84 +
## +init=epsg:32650 +proj=utm +zone=50 +datum=WGS84 +units=m +no_defs +ellps=WGS84 +towgs84=0,0,0
## without reprojecting.
```

```
## For reprojection, use function spTransform
proj4string(ljz) <- proj4string(terrain)

## Warning in ReplProj4string(obj, CRS(value)): A new CRS was assigned to an object with an existing CRS
## +init=epsg:32650 +proj=utm +zone=50 +datum=WGS84 +units=m +no_defs +ellps=WGS84 +towgs84=0,0,0
## without reprojecting.
## For reprojection, use function spTransform
#Leave-one-out-Cross-Validation
LOOCV <- gstat::krige.cv(Ca_exch ~ yingtan_elevation + SAGAWI,
                  ljz,
                  model = fvariomodel)

bubble(LOOCV, "residual")
```




```
# RMSE
rmse <- function(x,y) {
  sqrt(mean((x-y)^2))
}

rmse(x = LOOCV$var1.pred,
     y = LOOCV$observed)

## [1] 9.29913
```