StatistischeModellierung

Zlabinger Christof

2024-01-23

Aufgabe 1

Lade den Datensatz 'state.x77' in R. Beschreibe die Daten anhand der internen Hilfe.

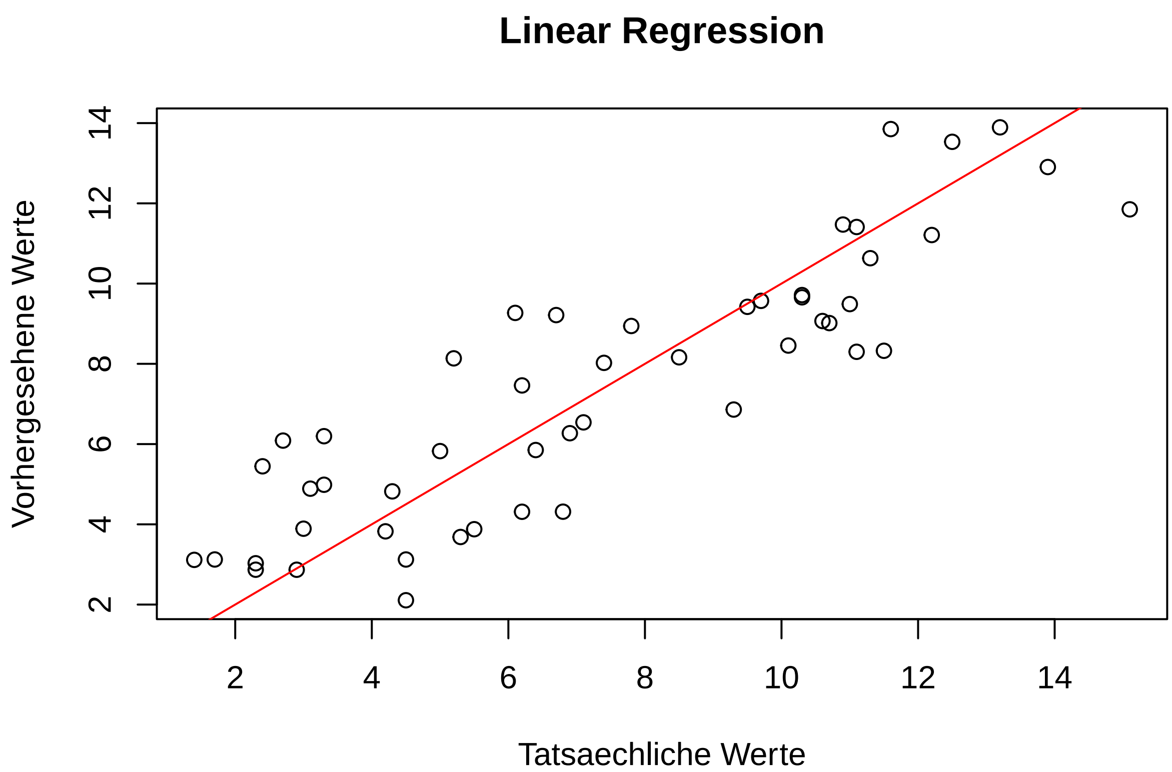
```
##    Population        Income       Illiteracy       Life Exp    
##  Min.   :  365   Min.   :3098   Min.   :0.500   Min.   :67.96  
##  1st Qu.: 1080   1st Qu.:3993   1st Qu.:0.625   1st Qu.:70.12  
##  Median : 2838   Median :4519   Median :0.950   Median :70.67  
##  Mean   : 4246   Mean   :4436   Mean   :1.170   Mean   :70.88  
##  3rd Qu.: 4968   3rd Qu.:4814   3rd Qu.:1.575   3rd Qu.:71.89  
##  Max.   :21198   Max.   :6315   Max.   :2.800   Max.   :73.60  
##      Murder          HS Grad          Frost             Area       
##  Min.   : 1.400   Min.   :37.80   Min.   :  0.00   Min.   :  1049  
##  1st Qu.: 4.350   1st Qu.:48.05   1st Qu.: 66.25   1st Qu.: 36985  
##  Median : 6.850   Median :53.25   Median :114.50   Median : 54277  
##  Mean   : 7.378   Mean   :53.11   Mean   :104.46   Mean   : 70736  
##  3rd Qu.:10.675   3rd Qu.:59.15   3rd Qu.:139.75   3rd Qu.: 81162  
##  Max.   :15.100   Max.   :67.30   Max.   :188.00   Max.   :566432
```

Es handelt sich um eine Matrix mit 50 Zeilen und 8 Reihen welche Dated der US Staaten beinhalten. Jede Zeile entspricht einem Staat. Diese Reihen beinhalten die:

- Population im Jahr 1975 in 100 Einwohnern
- Das Einkommen pro Person in 1974
- Die Prozent an Analphabeten in 1970
- Die Lebenserwartung von 1969-1971
- Die Mordrate pro 100,000 Einwohnern
- Die Prozent der Highschool Absolventen
- Die Mittlere Anzahl an Tagen an denen es in der Hauptstadt oder in einer grossen Stadt, in den Jahren 1931-1960, es unter 0°c hatte.
- Die Flaeche der Laender

Ermittle ein lineares Regressionsmodell, dass die Mordrate ('Murder') durch die unabhängigen Variablen Population, Income, Illiteracy,und Life Exp(ectancy) erklärt. Schreibe die Modellgleichung an und interpretiere die Werte der Koeffizienten im Kontext.

```r
model <-  lm(state.x77[,"Murder"] ~ state.x77[,"Population"] + state.x77[,"Income"] + state.x77[,"Illite

plot(state.x77[,"Murder"], predict(model), main = "Linear Regression", xlab = "Tatsaechliche Werte", yla
abline(0,1, col = "red")
```




```r
coef <- coef(model)
model_equation <- paste("Y=",round(coef[1],2), "+",
                        round(coef[2],4),"* X1 +",
                        round(coef[3],4),"* X2 +",
                        round(coef[4],2),"* X3 +",
                        round(coef[5],2),"* X4")

print(model_equation)
```

```
## [1] "Y= 112.84 + 2e-04 * X1 + 5e-04 * X2 + 2.27 * X3 + -1.57 * X4"
```

Aus der resultierenden Regressionsformel laest sich erkennen, dass die Illiteracy den groessten Einfluss auf die Mordrate hat. Den kleinsten Einfluss hat die Population. Eine niedrige Murder rate bringt eine hoehere Life Exp.

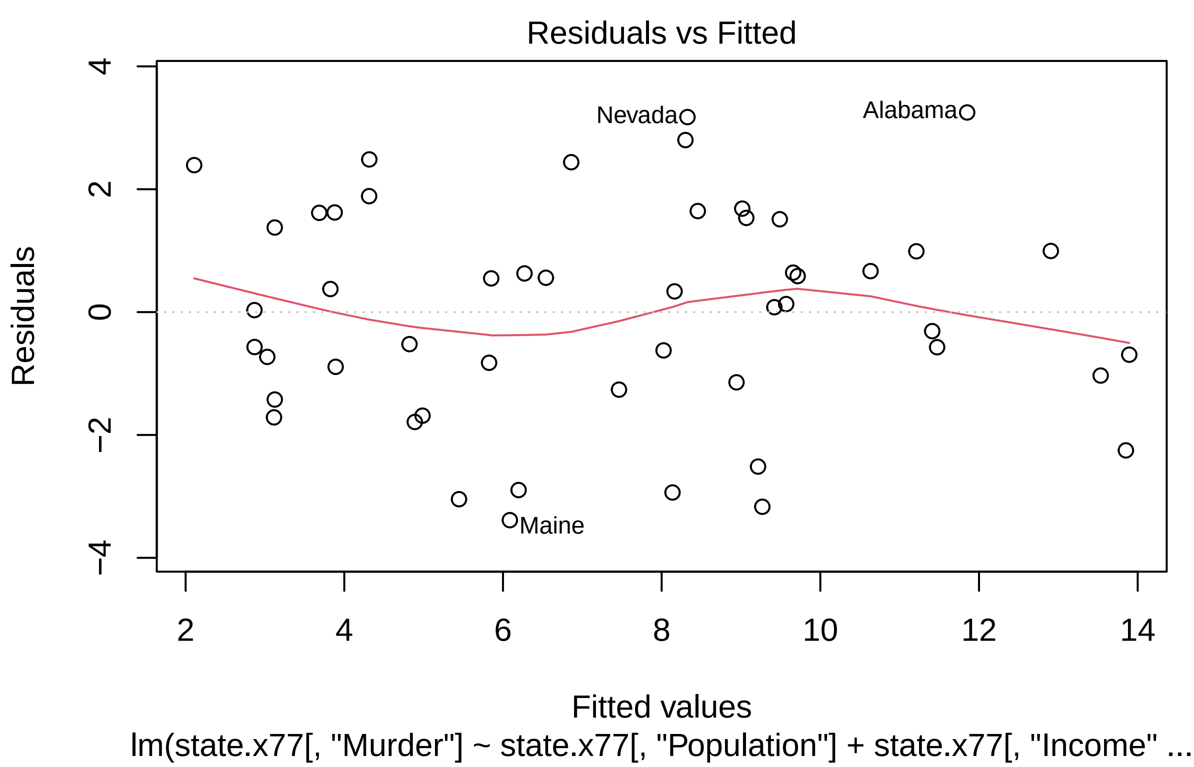
Führe alle fünf für dieses Regressionsmodell geltenden Modellvoraussetzungen an und überprüfe diese Voraussetzungen nachweislich anhand der Zusammenfassung (summary), Quality Plots der Regression und der pairwise Scatterplot Matrix. Erkläre, ob diese Modell überhaupt gültig ist. Falls es gültig ist, gib die Qualität der Erklärung durch das Modell an.

```r
model <-  lm(state.x77[,"Murder"] ~ state.x77[,"Population"] + state.x77[,"Income"] + state.x77[,"Illit
# Korrelation
summary(model)
```

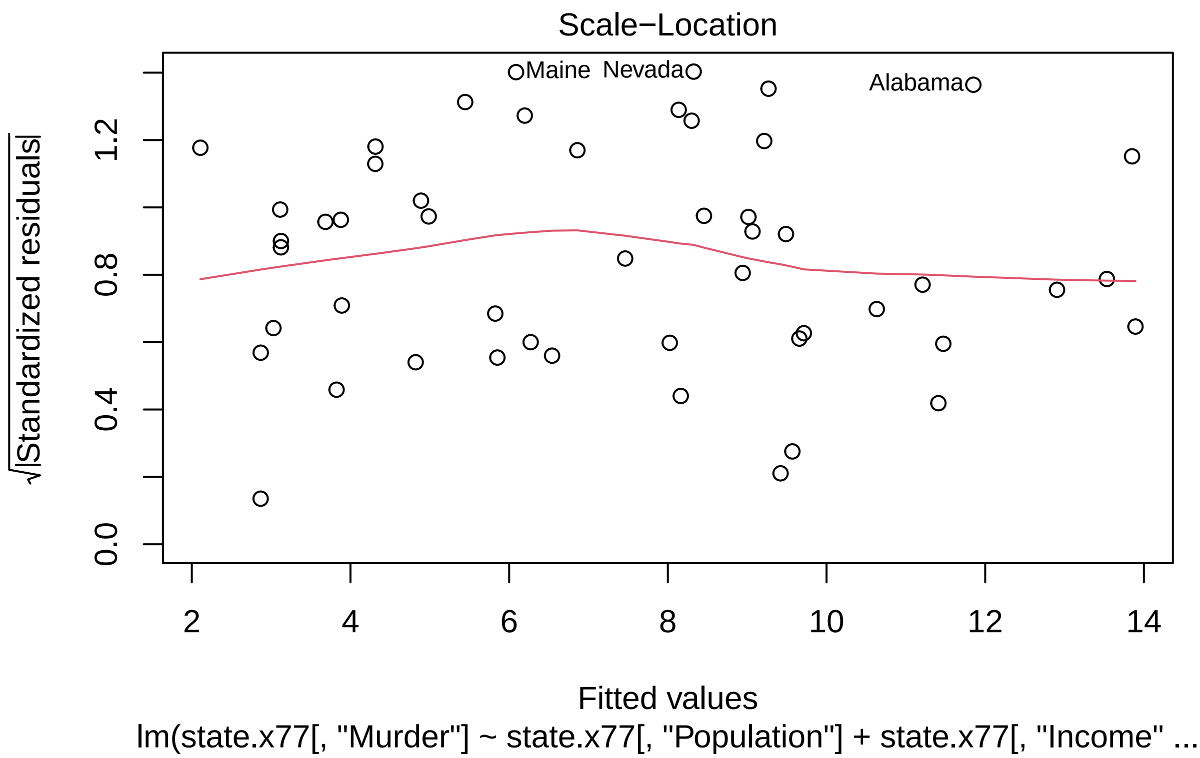
```
##
## Call:
## lm(formula = state.x77[, "Murder"] ~ state.x77[, "Population"] +
##     state.x77[, "Income"] + state.x77[, "Illiteracy"] + state.x77[,
##     "Life Exp"])
##
```

```
## Residuals:
##     Min      1Q Median      3Q     Max
## -3.387 -1.116  0.105  1.478  3.249
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.128e+02  1.740e+01   6.487 5.90e-08 ***
## state.x77[, "Population"]  2.059e-04  6.131e-05   3.358 0.001606 **
## state.x77[, "Income"]      4.524e-04  4.956e-04   0.913 0.366230
## state.x77[, "Illiteracy"]  2.265e+00  5.651e-01   4.008 0.000227 ***
## state.x77[, "Life Exp"]   -1.566e+00  2.419e-01  -6.474 6.15e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.824 on 45 degrees of freedom
## Multiple R-squared:  0.7758, Adjusted R-squared:  0.7558
## F-statistic: 38.92 on 4 and 45 DF,  p-value: 4.532e-14
```

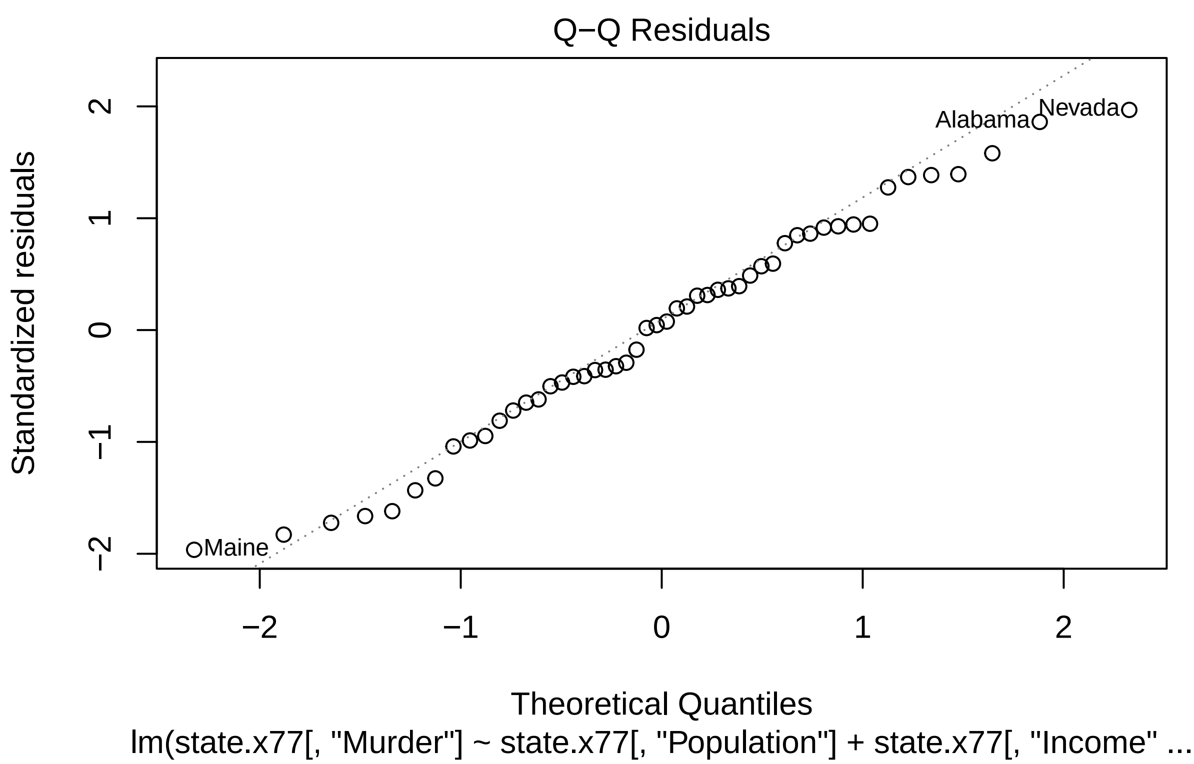
```r
# Systematische Fehler
plot(model, which = 1)
```




```r
# homoskedastitzitaet
plot(model, which = 3)
```




```
# Modellfehler normalverteilt
plot(model, which = 2)
```



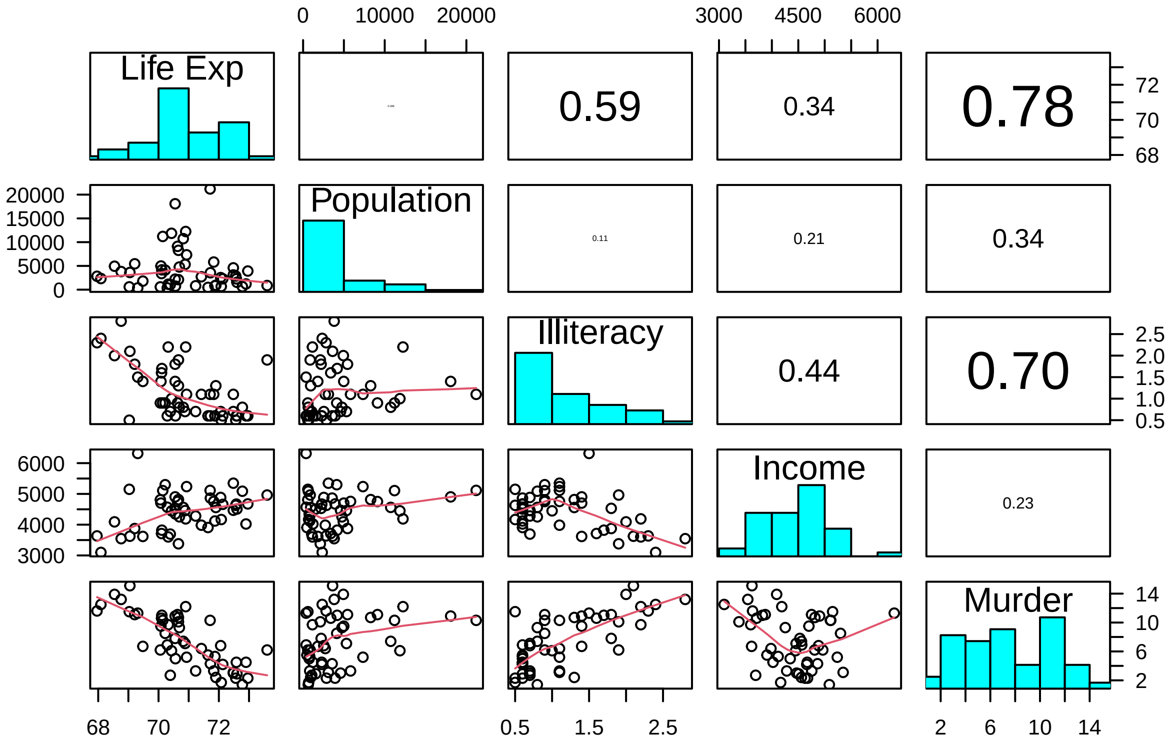

```
# multikollinearitaet
p <- state.x77[,c("Population","Income", "Illiteracy", "Life Exp", "Murder")]
print(cor(p))
```

```
##              Population     Income Illiteracy    Life Exp     Murder
## Population  1.00000000  0.2082276  0.1076224 -0.06805195  0.3436428
## Income      0.20822756  1.0000000 -0.4370752  0.34025534 -0.2300776
## Illiteracy  0.10762237 -0.4370752  1.0000000 -0.58847793  0.7029752
## Life Exp   -0.06805195  0.3402553 -0.5884779  1.00000000 -0.7808458
## Murder      0.34364275 -0.2300776  0.7029752 -0.78084575  1.0000000
```

```
panel.hist <- function(x, ...) {
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5))
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}

panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...) {
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste0(prefix, txt)
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
}

# Create a matrix of scatterplots
pairs(state.x77[, c("Life Exp" , "Population" , "Illiteracy", "Income", "Murder")],
      lower.panel = panel.smooth,
      upper.panel = panel.cor,
      diag.panel = panel.hist,
      las=1)
```

```
## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter
```

```
## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter
```



Es liegen keine systematischen Fehler vor da die Fehlervarianz konstant ist, es eine kollinearitaet zwischen der Illiteracy und der Life Exp, Murder rate, Income, alle Werte im QQ-Pot liegen nahe an der Geraden somit sind die Modellfehler Nomalverteilt. -> Modell sinvoll

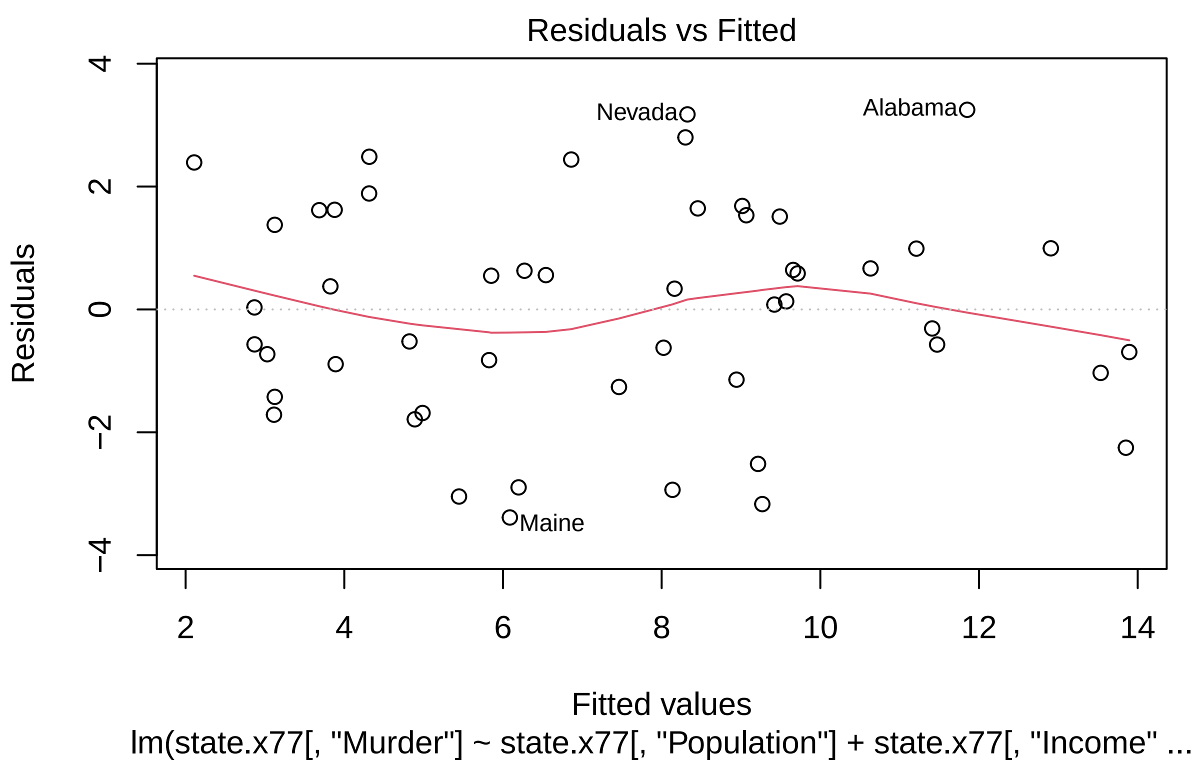
Führe eine Modellselektion der relevanten erklärenden Variablen durch.

```
model <-  lm(state.x77[,"Murder"] ~ state.x77[,"Population"] + state.x77[,"Income"] + state.x77[,"Illit
# Korrelation
summary(model)
```

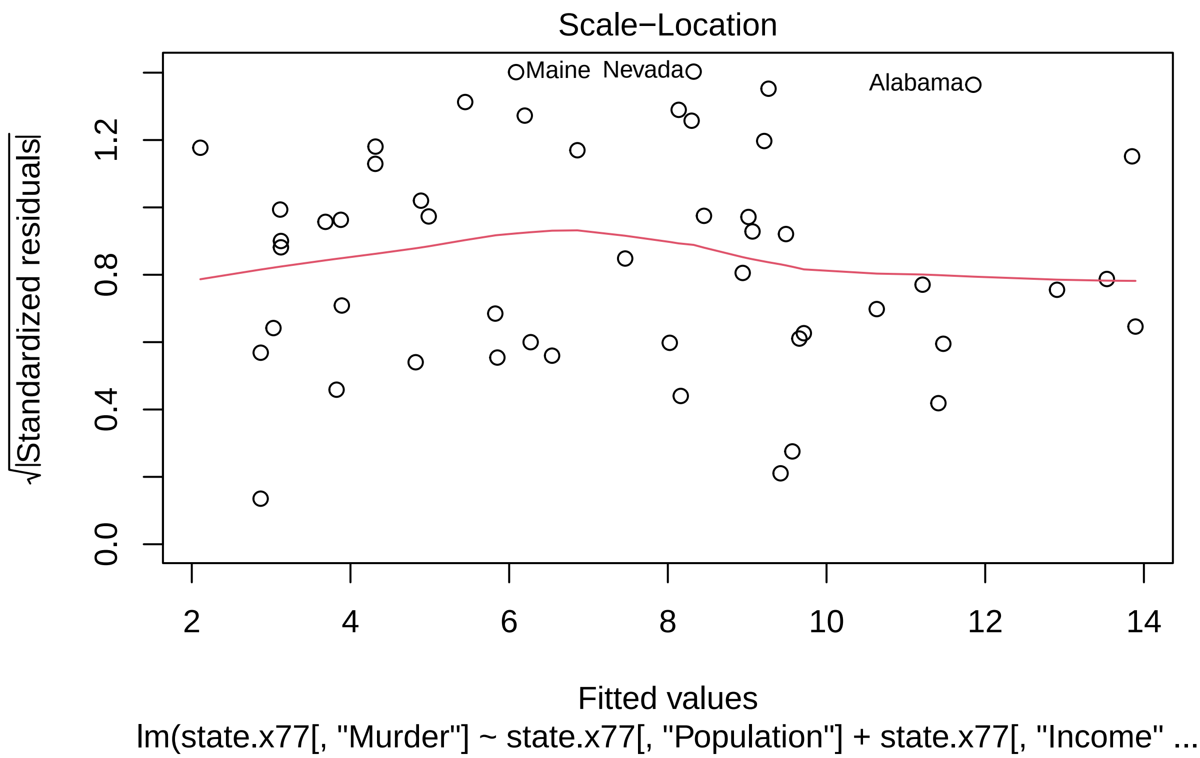
```
##
## Call:
## lm(formula = state.x77[, "Murder"] ~ state.x77[, "Population"] +
##     state.x77[, "Income"] + state.x77[, "Illiteracy"] + state.x77[,
##     "Life Exp"])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.387 -1.116  0.105  1.478  3.249
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.128e+02  1.740e+01   6.487 5.90e-08 ***
```

```
## state.x77[, "Population"]  2.059e-04  6.131e-05   3.358 0.001606 **
## state.x77[, "Income"]      4.524e-04  4.956e-04   0.913 0.366230
## state.x77[, "Illiteracy"]  2.265e+00  5.651e-01   4.008 0.000227 ***
## state.x77[, "Life Exp"]   -1.566e+00  2.419e-01  -6.474 6.15e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.824 on 45 degrees of freedom
## Multiple R-squared:  0.7758, Adjusted R-squared:  0.7558
## F-statistic: 38.92 on 4 and 45 DF,  p-value: 4.532e-14
```

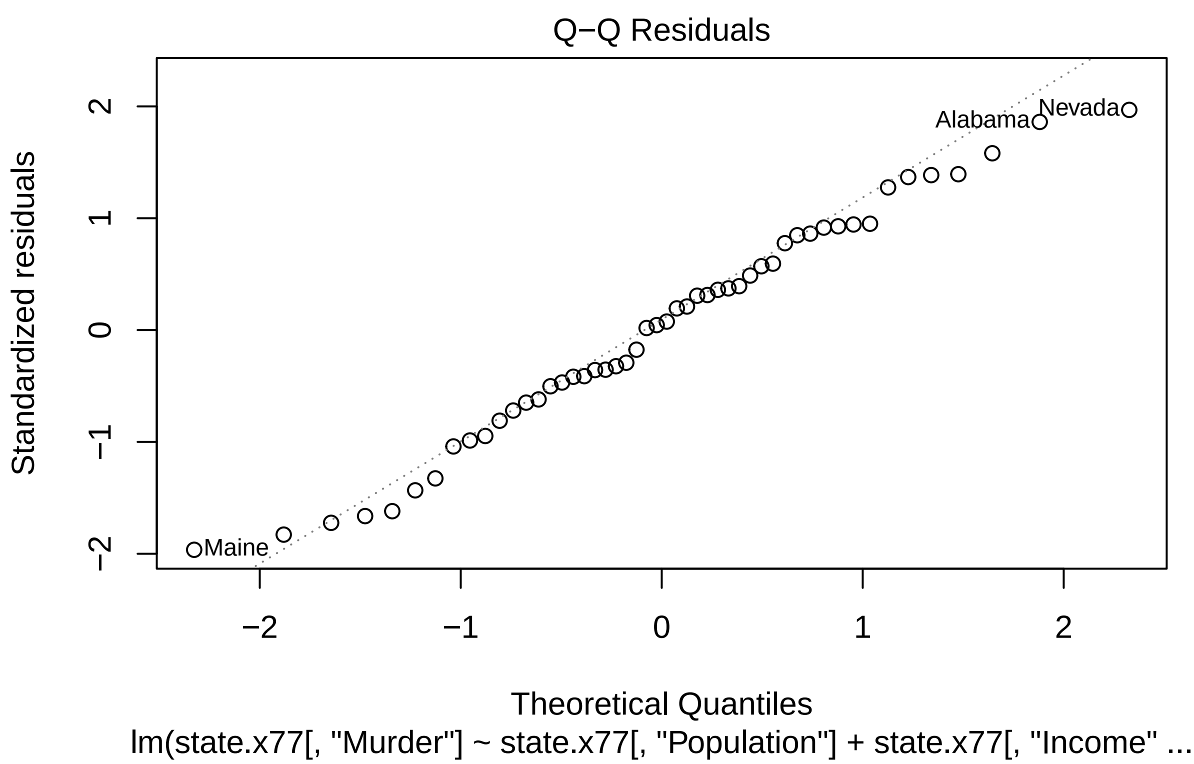
```r
# Systematische Fehler
plot(model, which = 1)
```




```r
# homoskedastitzitaet
plot(model, which = 3)
```




```r
# Modellfehler normalverteilt
plot(model, which = 2)
```

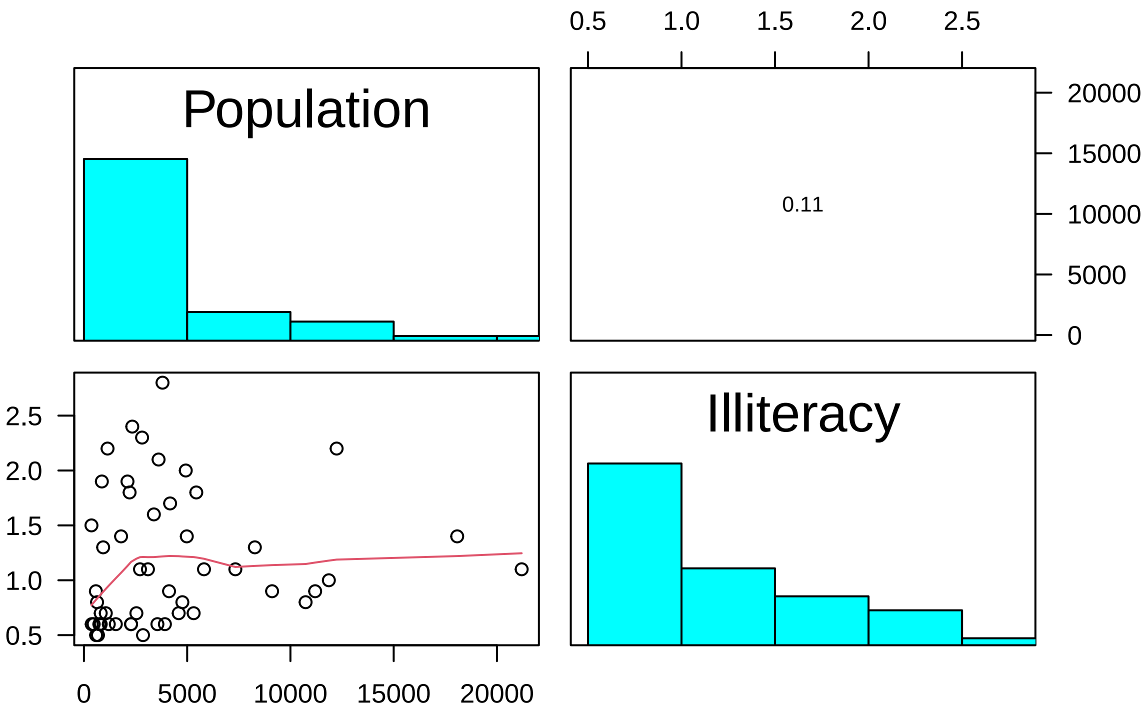



```r
# multikollinearitaet
p <- state.x77[,c("Population", "Illiteracy")]
print(cor(p))
```

```
##             Population Illiteracy
## Population  1.0000000  0.1076224
## Illiteracy  0.1076224  1.0000000
```

```
panel.hist <- function(x, ...) {
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5))
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}

panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...) {
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste0(prefix, txt)
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
}

# Create a matrix of scatterplots
pairs(state.x77[, c("Population", "Illiteracy")],
      lower.panel = panel.smooth,
      upper.panel = panel.cor,
      diag.panel = panel.hist,
      las=1)
```

```
## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter
```



Aufgabe 2

Installiere das Package 'MASS' mithilfe der Funktion install.packages. Lade den Datensatz 'Pima.tr' in R. Beschreibe die Daten anhand der internen Hilfe.

```
library(MASS)
data(Pima.tr)
?Pima.tr
```

Pima.tr ist ein Datensatz welcher Daten ueber Indische Frauen ueber 21 welche in der naehe von Phoenix Arizona wohnen beinhaltet. Die beinhalteten Daten sind:

- npreg

```
-    Die anzahl an Schwangerschaften
```

- glu
  - Plasma glucose konzentration
- bp
  - Blutdruck
- skin
  - Die dicke der Haut am triceps
- bmi
  - Body mass index
- ped
  - Diabetes-Stammbaumfunktion
- age
  - Alter

- type
  - Ob die Person von der WHO gesehen diabetis hat.

Ermittle ein logistisches Regressionsmodell, dass das Auftreten von Diabetes ('type') durch die übrigen unabhängigen Variablen Alter (age), Anzahl der Schwangerschaften (npreg), BMI, Glukosespiegel (glu), Blutdruck (bp), familiäre Häufung von Diabetesfällen (ped) und Hautfaltendickemessung am Oberarm (skin) erklärt. Schreibe die Modellgleichung an und interpretiere die Werte der Koeffizienten im Kontext.

```r
library(MASS)
library(ggplot2)

data(Pima.tr)

model <- glm(Pima.tr$type ~ Pima.tr$npreg + Pima.tr$glu + Pima.tr$bp + Pima.tr$skin + Pima.tr$bmi + Pima

coef <- coef(model)
model_equation <- paste("Y = ",round(coef[1],2), " + ",
                        round(coef[2],4)," * skin + ",
                        round(coef[3],2)," * glu + ",
                        round(coef[4],2)," * bmi + ",
                        round(coef[5],2)," * ped + ",
                        round(coef[6],4)," * bp + ",
                        round(coef[7],2)," * age + ",
                        round(coef[8],2)," * npreg")

summary(model)
```

```
##
## Call:
## glm(formula = Pima.tr$type ~ Pima.tr$npreg + Pima.tr$glu + Pima.tr$bp +
##     Pima.tr$skin + Pima.tr$bmi + Pima.tr$ped + Pima.tr$age, family = binomial,
##     data = Pima.tr)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -9.773062   1.770386  -5.520 3.38e-08 ***
## Pima.tr$npreg  0.103183   0.064694   1.595  0.11073
## Pima.tr$glu    0.032117   0.006787   4.732 2.22e-06 ***
## Pima.tr$bp    -0.004768   0.018541  -0.257  0.79707
## Pima.tr$skin  -0.001917   0.022500  -0.085  0.93211
## Pima.tr$bmi    0.083624   0.042827   1.953  0.05087 .
## Pima.tr$ped    1.820410   0.665514   2.735  0.00623 **
## Pima.tr$age    0.041184   0.022091   1.864  0.06228 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 256.41  on 199  degrees of freedom
## Residual deviance: 178.39  on 192  degrees of freedom
```

```
## AIC: 194.39
##
## Number of Fisher Scoring iterations: 5
```

```
print(model_equation)
```

```
## [1] "Y =  -9.77  +  0.1032  * skin +  0.03  * glu +  0  * bmi +  0  * ped +  0.0836  * bp +  1.82  *
```

Begriffe

- Scatterplot Matrix
- Zeigt mehrere Streudiagramme. Bei mehreren variabeln kann die korrelation darstellen.
- lineare Regression
- Stellt eine Gleichung auf welche moeglichst genau durch alle Datenpunkte geht.
- Quality Plots
- Hilft die Qualitaet eines Modells zu ueberpruefen.
- Residuen
- Die Differenz zwischen vorhergesagten Werten und tatsaechlichen Werten.
- Regressionskoeffizienten
- Parameter einer Regressionsgleichung
- Regressionsmodell
- Zusammenhand zwischen anhaengiger und unabhaengigen variabeln.
- Modellgleichung
- Gleich wie Regressionsmodell
- logistische Regression
- Modell fuer binaere abhaengige variabeln.