Hausübung 04 - Lineare Regression

Baier Sebastian, Figlmüller Magdalena, Schwarzböck Alice

29.11.2023

Contents

Allgemeine Hinweise zur Übung

Für alle Beispiele gelten folgende Aufgabenstellungen:

- Überprüfen Sie alle erforderlichen statistischen Voraussetzungen für die Gültigkeit dieses Modells mtihilfe der quality plots der Residuen und gegebenenfalls Scatterplots.
- Führen Sie eine Modellselektion durch und wählen anhand statistischer Kriterien ein optimales Modell aus. Argumentieren Sie anhand Kriterien für die Signifkanz von Koeffizienten und gegebenenfalls zusätzlich von Modellen.
- Schreiben Sie das Regressionsmodell und die angepasste Modellgleichung des optimalen Modells explizit an.
- Interpretieren Sie die Werte die Koeffizienten im Sachzusammenhang.

Aufgabe 1: Datentransformation

- Wählen Sie den Datensatz UN aus der library 'car'.
- Filtern Sie erst 'NA' mit der Funktion na.omit.
- Erklären Sie dann infant mortality durch gross domestic product.
- Explorieren Sie die Daten, bevor Sie ein Modell anpassen.

In diesem Datensatz sollen nun, wie oben beschrieben, die Infant Mortality (Säuglingssterblichkeit) und das Gross Domestic Product (Bruttoinlandsprodukt) erklärt werden. Hierfür schauen wir uns zuerst die Daten etwas genauer an um anschließend ein Modell dafür zu erstellen.

Beschreibung des Datensatzes: * **region:** Größere Regionen in der Welt, jedoch nicht zwingend Kontinente (Original [en]: Region of the world: Africa, Asia, Caribbean, Europe, Latin Amer, North America, NorthAtlantic, Oceania) * **group:** Beschreibender Faktor, welcher beinhaltet, ob Länder der OECD oder anden Organisationen angehören (Original [en]: A factor with levels oecd for countries that are members of the OECD, the Organization for Economic Co-operation and Development, as of May 2012, africa for countries on the African continent, and other for all other countries. No OECD countries are located in Africa.) * **fertility:** Fruchtbarkeitsrate als Zahl der Kinder/Frau (Original [en]: Total fertility rate, number of children per woman.) * **ppgdp:** Bruttoinlandsprodukt in US-Dollar (Original [en]: Per capita gross domestic product in US dollars.) * **pctUrban:** Prozent des Stadtanteils (Original [en]: Percent urban.) * **infantMortality:** Todesfällen bei Säuglingen im Alter von einem Jahr je 1.000 Lebendgeburten (Original [en]: Infant deaths by age 1 year per 1000 live births)

Zuerst filtern wir den Datensatz mittels na.omit, was wir wie folgt durchführen und uns anschließend die Datenstruktur via glimpse ausgeben lassen.

```
UN_new <- UN %>% na.omit()
glimpse(UN_new)
```

```
## Rows: 193
## Columns: 7
## $ region          <fct> Asia, Europe, Africa, Africa, Latin Amer, Asia, Caribb~
## $ group           <fct> other, other, africa, africa, other, other, other, oec~
## $ fertility       <dbl> 5.968, 1.525, 2.142, 5.135, 2.172, 1.735, 1.671, 1.949~
## $ ppgdp           <dbl> 499.0, 3677.2, 4473.0, 4321.9, 9162.1, 3030.7, 22851.5~
## $ lifeExpF        <dbl> 49.49, 80.40, 75.00, 53.17, 79.89, 77.33, 77.75, 84.27~
## $ pctUrban        <dbl> 23, 53, 67, 59, 93, 64, 47, 89, 68, 52, 84, 89, 29, 45~
## $ infantMortality <dbl> 124.535, 16.561, 21.458, 96.191, 12.337, 24.272, 14.68~
```

```r
summary(UN_new$infantMortality)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.916   7.243  19.637  30.739  45.892 124.535
```

```r
summary(UN_new$ppgdp)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    114.8   1239.8   4495.8  12291.1  14497.3 105095.4
```

Die Originaldaten enthielten 213 Zeilen, während die gefilterten Daten nur mehr 193 Zeilen enthalten. Es wurden daher 20 Zeilen herausgefiltert.

Was weiterhin auffällt, ist, dass der arithmetische Mittelwert von infantmortality mit 30.739 gegenüber dem Median mit 19.637 deutlich abweicht.

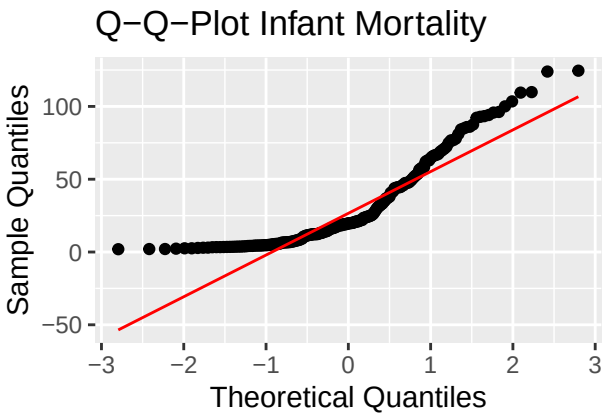
Noch deutlich stärker wird dieser Effekt beim Bruttoinlandsprodukt 'ppdgdp' sichtbar. Hier ist der arithmetische Mittelwert mit 1.229×10^4 gegenüber dem Median mit 4495.8 noch unterschiedlicher.

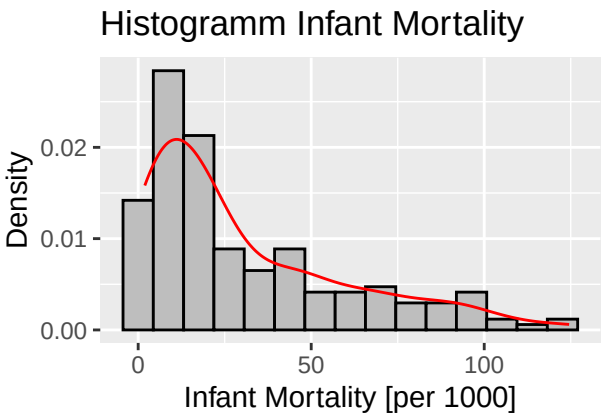
In beiden Fällen ist hier möglicherweise eine Schiefe der Daten zu erwarten.

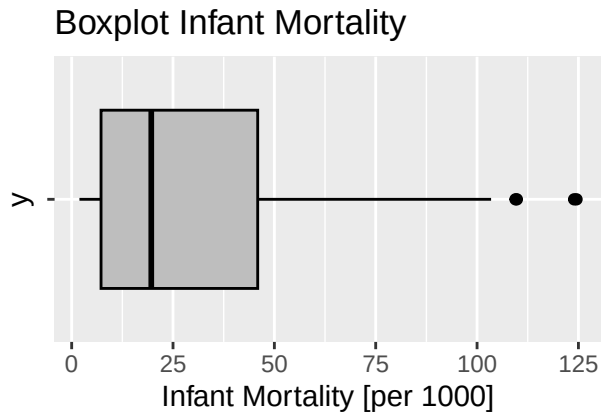
In Folge wird der Datensatz exploratorisch weiter untersucht.

Explorative Analyse - Infant Mortality

```
## $y
## [1] ""
##
## attr(,"class")
## [1] "labels"
```







Auffällig im QQ-Plot ist der schwere Rand auf der linken Seite des Plots. Dies wird auch die Skweness von 1.198 bestätigt.

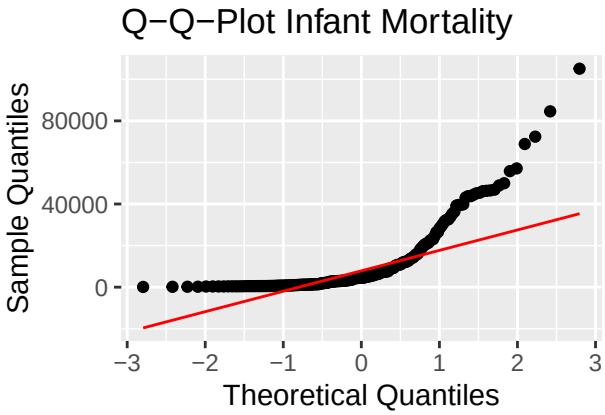
Dies wird auch durch das Histogramm belegt, welches eine Spitze bei etwa 12 von 1000 Säuglingen in der Säuglingssterblichkeit zeigt. Dies weicht deutlich vom arithmetischen Mittelwert (30.739) und dem Median (19.637) ab.

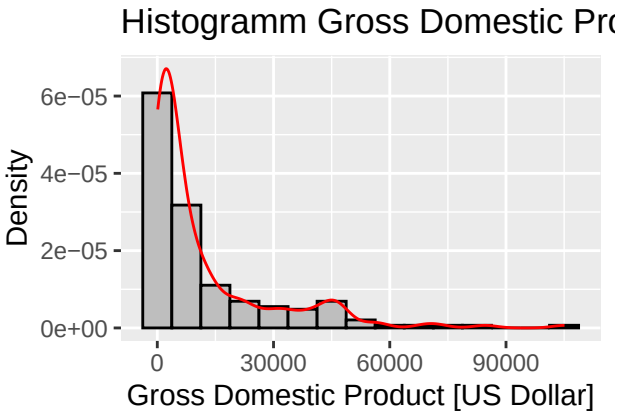
Diese Schiefe wird im Boxplot ebenfalls nochmal sehr gut sichtbar. Anzumerken ist hierbei, dass Boxplots für normalverteilte optimal sind. Die hier dargestellten Ausreißer, müssen demnach keine Ausreißer sein sondern werden als solche dargsetellt, da diese aufgrund der Schiefe außerhalb des Quantiles des Boxplots liegen.

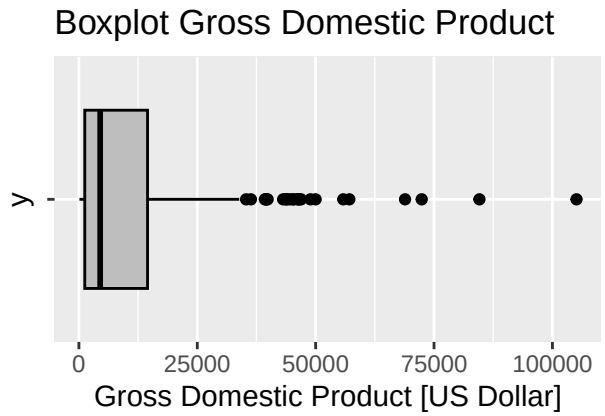
Zusammenfassend zeigt sich jedoch die Unimodalität und Rechtsschiefe der Daten.

Explorative Analyse - Gross Domestic Product

```
## $y
## [1] ""
##
## attr(,"class")
## [1] "labels"
```







Auffällig im QQ-Plot ist der extrem schwere Rand auf der rechten Seite des Plots. Dies wird auch die Skweness 2.212 bestätigt. Im Histogramm zeigt sich, dass die Daten zwar sehr weit verteilt sind, jedoch ab etwa 60.000 US-Dollar gegen Null tendieren. Weiterhin weichen der arithmetische Mittelwert (1.229×10^4) und der Median (4495.8) deutlich voneinander ab, was die Grafiken ebenfalls nochmalig unterstreichen.

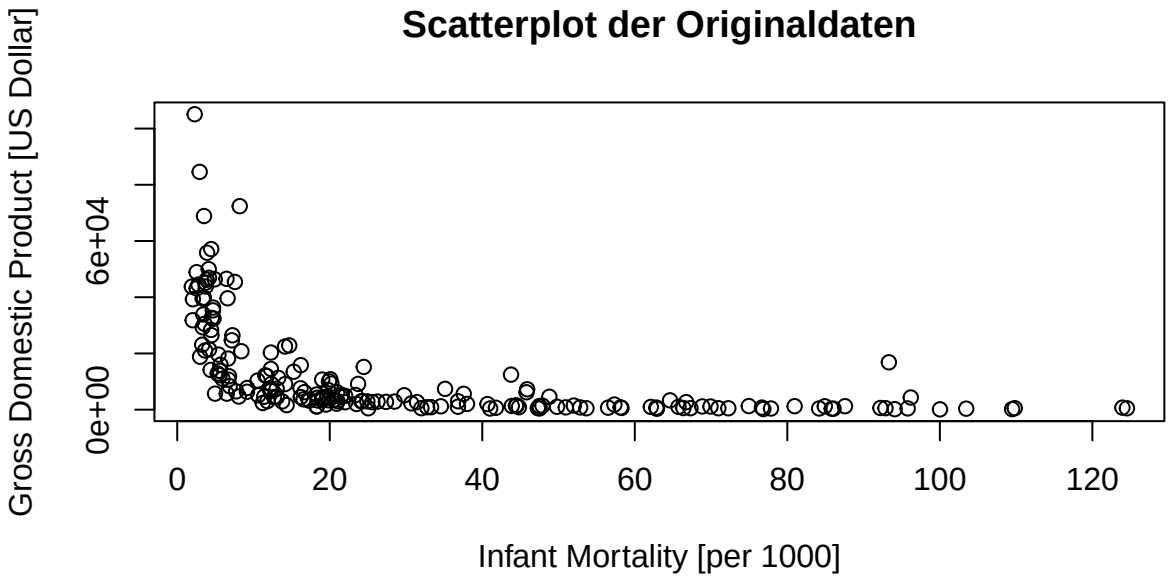
Diese Schiefe wird im Boxplot ebenfalls nochmal sehr gut sichtbar. Anzumerken ist hierbei, dass Boxplots für normalverteilte optimal sind. Die hier dargestellten Ausreißer, müssen demnach keine Ausreißer sein sondern werden als solche dargsetellt, da diese aufgrund der Schiefe außerhalb des Quantiles des Boxplots liegen.

Zusammenfassend zeigt sich jedoch die Unimodalität und rechtsscheife der Daten.

Korrelations- und Lineritätscheck zwischen der Infant Mortality sowie dem Gross Domestic Product

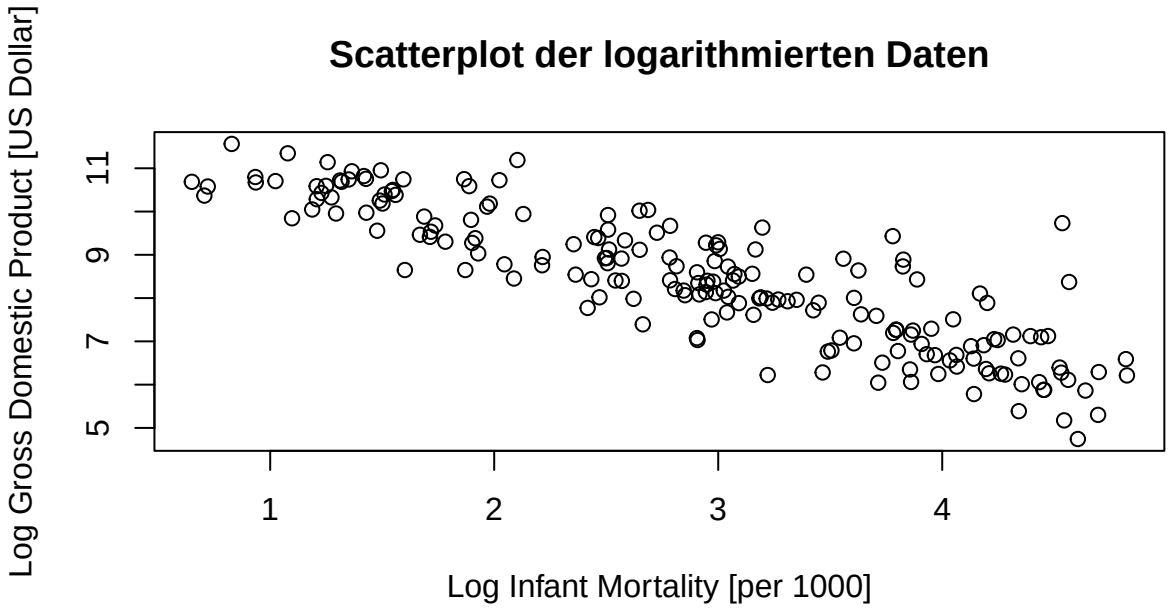
Originaldaten

Zuerst schauen wir uns einen Scatterplot der Originaldaten an.



Wir sehen an der Stelle, dass diese Daten eine Kurve beschreiben und es hier wenig Sinn macht eine Regressionsgerade durchzulegen. Allerdings liegen die Daten so, dass man eventuell durch logarithmieren eine gerichtete Punktewolke hinbekommt, womit eine Korrelation wieder deutlich mehr Sinn machen würde.

Logarithmierte Daten



Hier lässt sich bereits eine negativ gerichtete Gerade erahnen. Eine negative Korrelation würde hier bedeuten, dass ein linearer Zusammenhang beobachtet werden kann. So scheint es - aufgrund dieser Daten - einen Zusammenhang zwischen einer niedrigen Säuglingssterblichkeit bei hohem Bruttoinlandsprodukt oder umgekehrt einer hohen Säuglingssterblichkeit bei einem niedrigen Bruttoinlandsprodukt zu geben. Dies wird nun mit der Pearson-Korrelation getestet.

```
>
>   Pearson's product-moment correlation
>
> data:  log(UN_new$infantMortality) and log(UN_new$ppgdp)
> t = -25, df = 191, p-value <2e-16
> alternative hypothesis: true correlation is not equal to 0
> 95 percent confidence interval:
>  -0.905 -0.838
> sample estimates:
>    cor
> -0.875
```

Die Pearson-Korrelation zeigt einen Wert von -0.875. Dies bestätigt obige Annahme eines Zusammenhangs zwischen niedriger Säuglingssterblichkeit bei hohem Bruttoinlandsprodukt.

Auch der p-value mit < 2.2e-16 zeigt, dass H0 (keine Korrelation) verworfen werden kann.

Für die Modellanpassung werden die logarithmierten Daten verwendet.

Modellanpassung mit logarithmierten Daten

```
log.lm <- lm(log(UN_new$infantMortality)~log(UN_new$ppgdp))
summary(log.lm)
```

```
>
> Call:
> lm(formula = log(UN_new$infantMortality) ~ log(UN_new$ppgdp))
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -1.1679 -0.3674 -0.0235  0.2454  2.4350
>
> Coefficients:
>                   Estimate Std. Error t value Pr(>|t|)
> (Intercept)         8.1038     0.2109    38.4   <2e-16 ***
> log(UN_new$ppgdp)  -0.6168     0.0247   -25.0   <2e-16 ***
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 0.528 on 191 degrees of freedom
> Multiple R-squared:  0.766,   Adjusted R-squared:  0.765
> F-statistic:  626 on 1 and 191 DF,  p-value: <2e-16
```

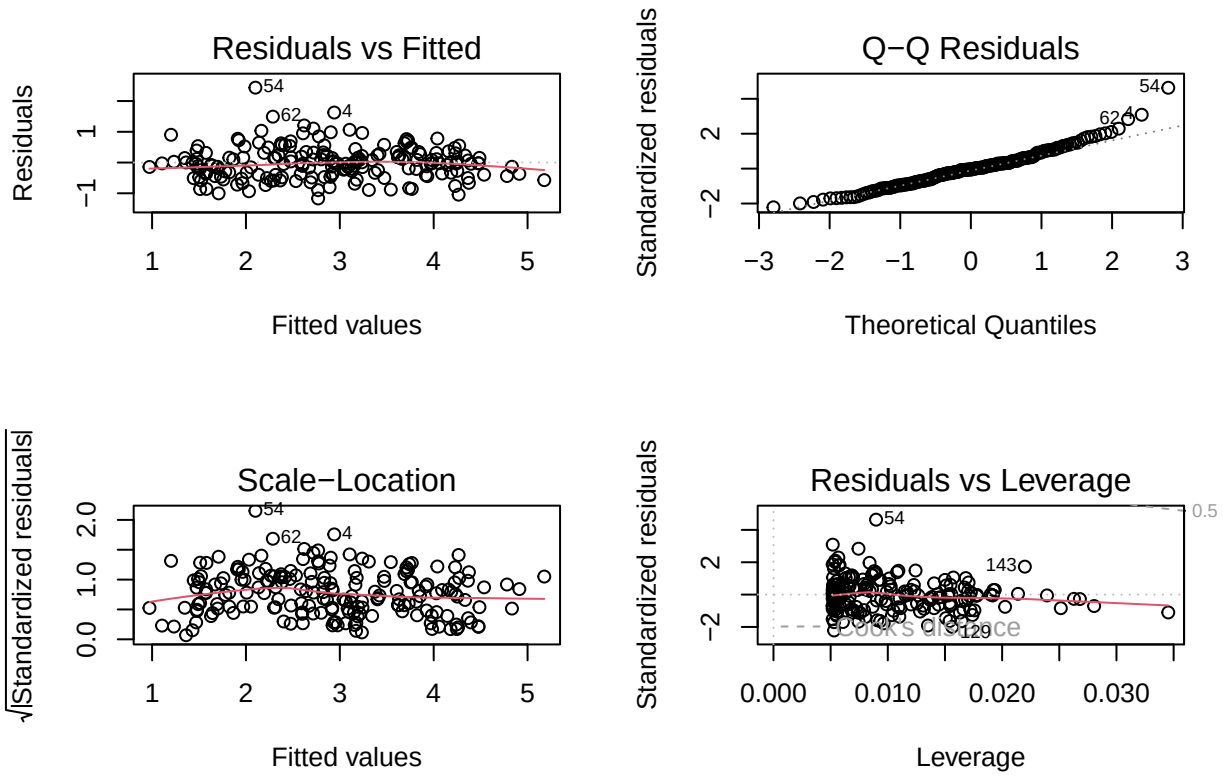
Für hier gilt:

- H0: kein linearer Zusammenhang
- H1: linearer Zusammenhang

Die F-Statistik (625.9) und der p-value von < 2.2e-16 lässt uns die H0 verwerfen. Durch den Adjusted R-squared von 0.765 kann ausgesagt werden, dass von den abhängigen Variablen ca. 76% der Varianz durch das dieses lineare Modell erklärt werden können. Der Standardfehler der Residuen beträt 0.5281 bei 191 Freiheitsgraden.

Überprüfung der Residuen

```
par(mfrow = c(2, 2))
plot(log.lm)
```



- **Residuals vs Fitted:** Die Residuen sind homoskedastisch. Sie sind auch um den 0-Wert zentriert, was auf keinen systematischen Fehler hinweist. Zudem sind die Residuen nicht korreliert und beeinflussen sich nicht gegenseitig.
- **Normal Q-Q:** Die Residuen liegen zum allergrößten Teil auf der Linie und heben sich nur am Rechten Rand etwas ab (schwerer Rand). Die Residuen wirken daher normalverteilt. Es gibt zwar auf der rechten Seite zusätzlich noch einen Ausreißer (54) - jedoch sollte dieser bei diesem Datenumfang kein Problem darstellen.
- **Scale-Location:** Es sind keine Verläufe erkennbar und die Residuals wirken nicht heteroskedastisch.
- **Residuals vs Leverage:** Die Punkte liegen innerhalb der Cook's distance, wodurch sich kein starker (negativer) Hebeleffekt erkennen lässt. Zwar ist ganz rechts ein Punkt zu sehen, dieser liegt allerdings nur knapp neben der Linie und hätte daher eher einen positiven Hebeleffekt (im Gegensatz zu solchen Residuals, welche außerhalb der Cooks-Distance liegen würden).

Erstellung der Modell-Gleichung

- **Allgemeine Regressionsgleichung:**

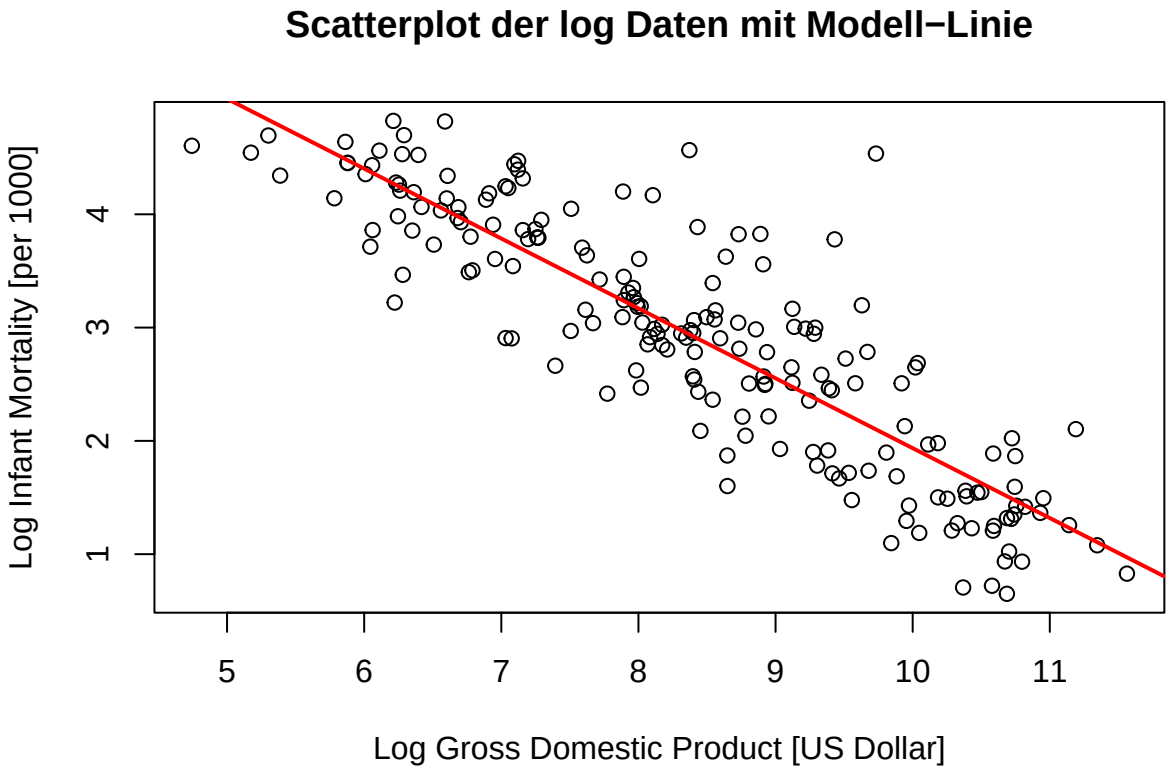
$$y(i) = alpha + beta * x$$

- **Modell-Gleichung:**

$$log(infantMortality) = 8.10377 - 0.61680 \cdot log(ppgdp)$$

* **Dabei gilt:** Die 8.10377 ist das Intercept und die -0.61680 ist die Steigung.

Scatterplot mit Regressionsgeraden



Umkehrung der linearen Transformation

Um wieder zu den Originaldaten zurückzugelangen, muss das durch Logarithmierung erstellte Modell wieder umgeformt werden.

Hierfür gilt:

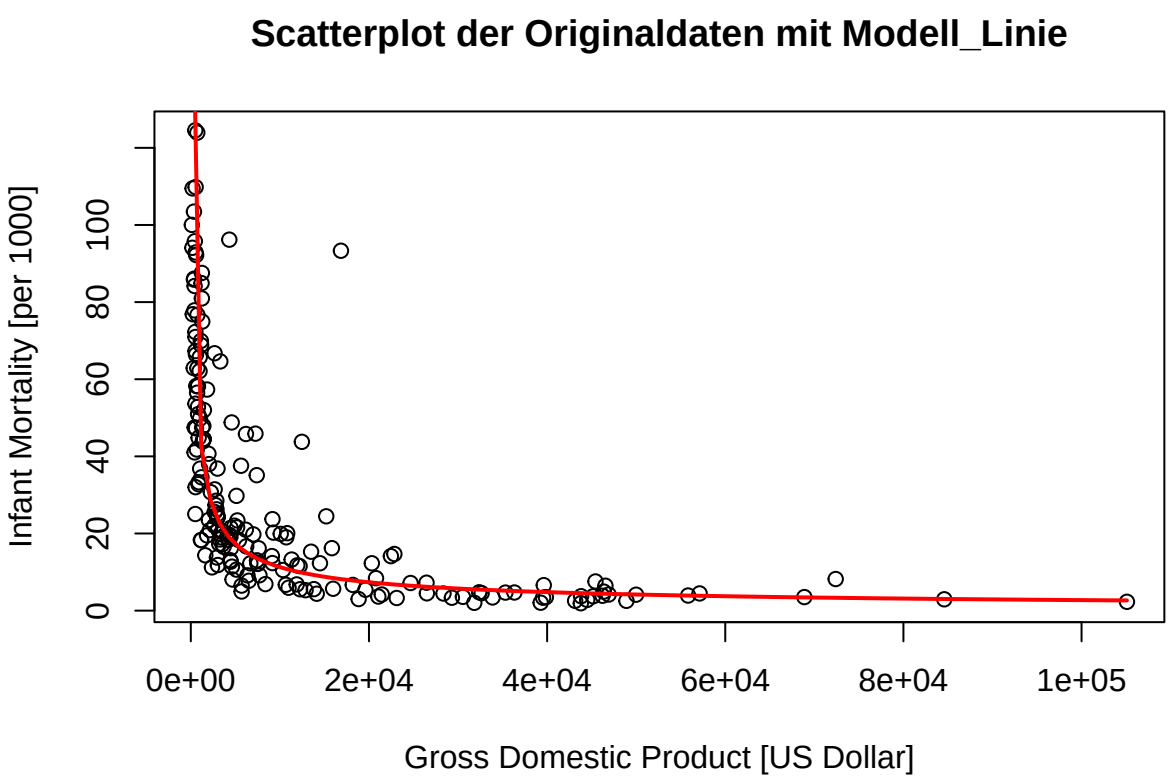
$$exp(log(infantMortality)) = exp(-0.61680 * log(ppgdp) + 8.10377)$$

$$infantMortality = ppgdp^{-0.61680} * exp(8.10377)$$

Die Funktion für das Modell der Original-Daten lautet somit:

$$ppgdp = 3306.912 \cdot (infantMortality)^{-0.61680}$$

Durch betrachten der Gleichung wird ersichtlich, dass wenn die Säuglingssterblichkeit (x-Achse) auf Null gehen würde, wäre das Bruttoinlandsprodukt (y-Achse) unendlich gross. Zudem bedeutet eine Erhöhung der Säuglingssterblichkeitsrate um 1, dass das Bruttoinlandsprodukt um das 3306.912 fache sinkt.


Scatterplot der Originaldaten mit Modell_Linie
Infant Mortality [per 1000]
0
20
40
60
80
100
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Gross Domestic Product [US Dollar]

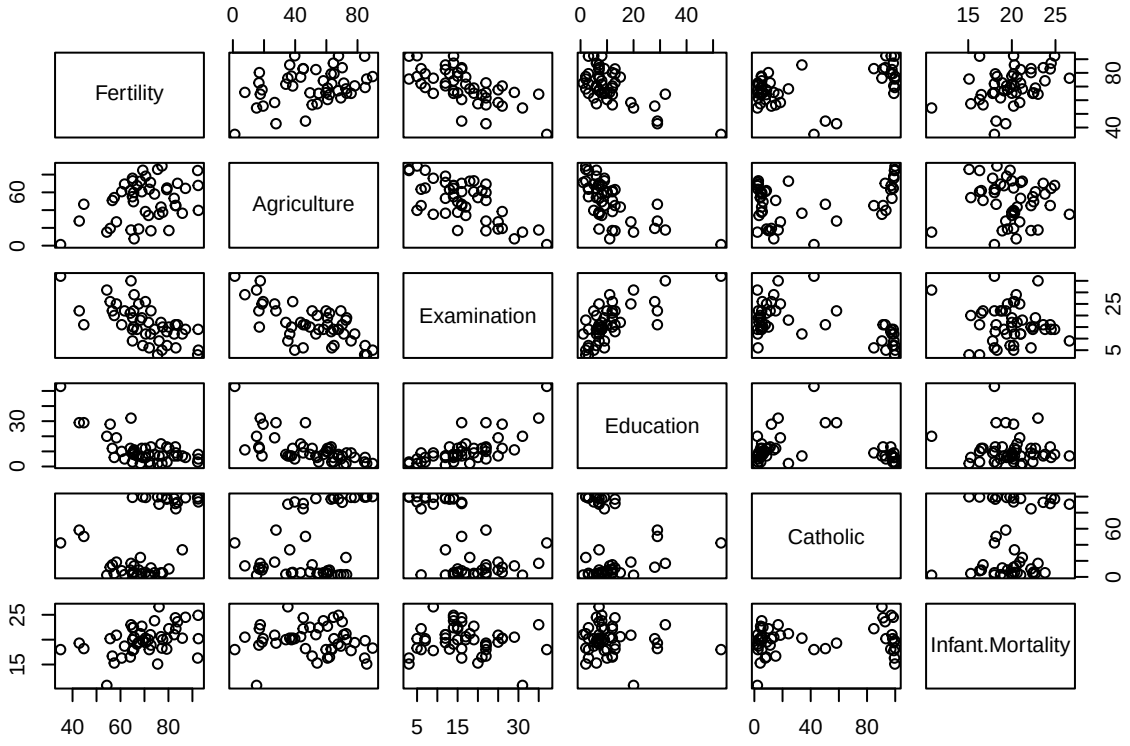
Aufgabe 2: Schweiz

- Wir kehren zurück zu den Variablen “Fertility”, “Agriculture”, “Education”, “Catholic” und “Infant. Mortality” aus dem R Datensatz swiss des R package utils.
- Passen Sie für die oben genannten Variablen ein Modell an, das Education durch die übrigen Variablen erklärt, soweit dies zulässig ist.

```
library(utils)
str(swiss)
```

```
## 'data.frame':    47 obs. of  6 variables:
##  $ Fertility       : num  80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 ...
##  $ Agriculture     : num  17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ...
##  $ Examination     : int  15 6 5 12 17 9 16 14 12 16 ...
##  $ Education       : int  12 9 5 7 15 7 7 8 7 13 ...
##  $ Catholic        : num  9.96 84.84 93.4 33.77 5.16 ...
##  $ Infant.Mortality: num  22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ...
```

```
plot(swiss)
```



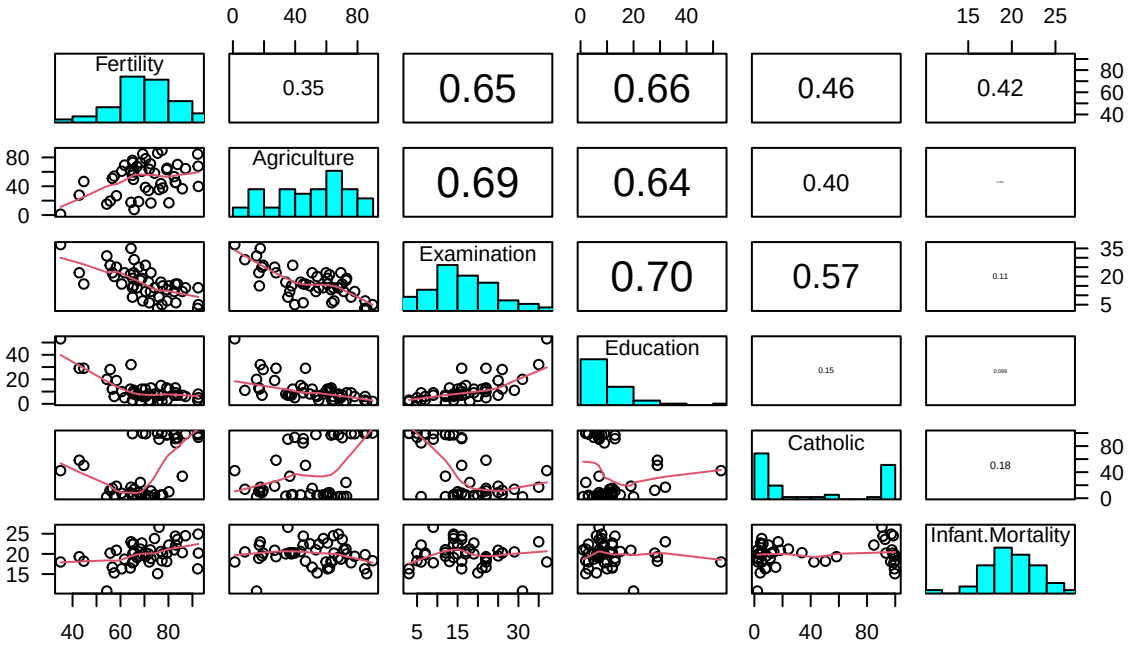
In dieser Aufgabenstellung geht es darum, das optimale lineare Modell zu finden um die Variable Education durch die übrigen Variablen zu beschreiben. Hierfür müssen allerdings zuerst die statistischen Voraussetzungen hierfür überprüft bzw. evaluiert werden.

Überprüfung der statistischen Voraussetzungen

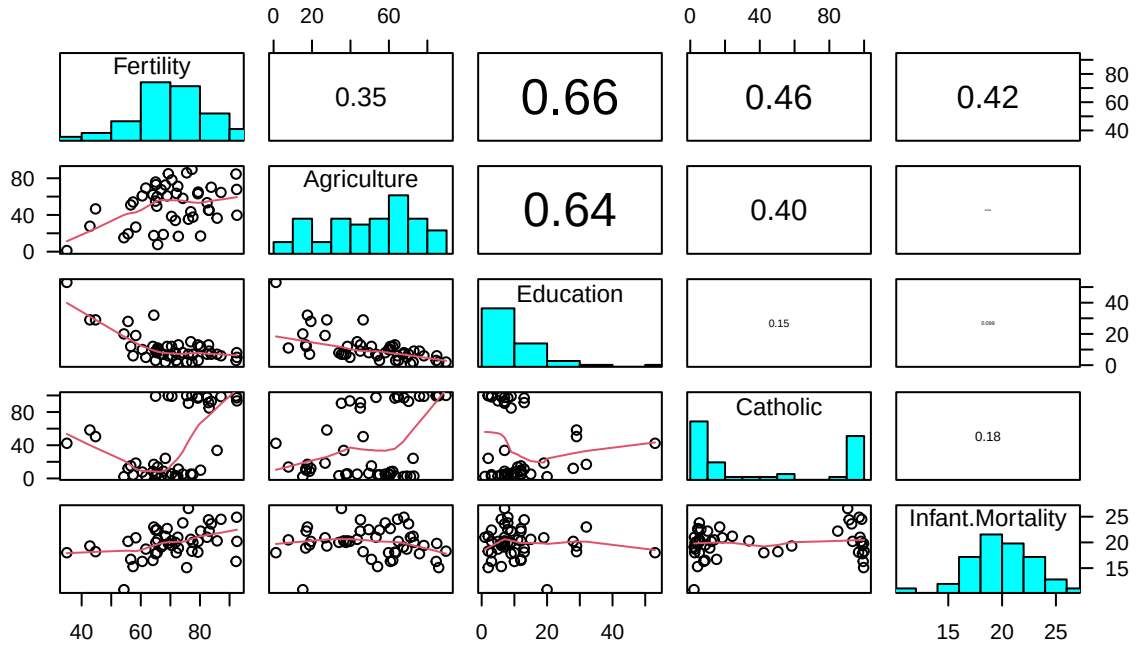
Zusammenfassung und Residuenplots

```
>                 Fertility Agriculture Examination Education Catholic
```

```
> Fertility            1.000      0.3531      -0.646   -0.6638    0.464
> Agriculture          0.353      1.0000      -0.687   -0.6395    0.401
> Examination         -0.646     -0.6865       1.000    0.6984   -0.573
> Education           -0.664     -0.6395       0.698    1.0000   -0.154
> Catholic             0.464      0.4011      -0.573   -0.1539    1.000
> Infant.Mortality     0.417     -0.0609      -0.114   -0.0993    0.175
>                  Infant.Mortality
> Fertility                  0.4166
> Agriculture               -0.0609
> Examination               -0.1140
> Education                 -0.0993
> Catholic                   0.1755
> Infant.Mortality           1.0000
```



Aus dem Scatterplot sehen wir, dass die Variable Examination mit fast allen anderen Variablen hoch korreliert. Da dies für eine Modellbildung als erklärende Variable nicht zulässig ist, wird diese aus dem Datensatz entfernt und die Daten via Scatterplot nochmalig geprüft.

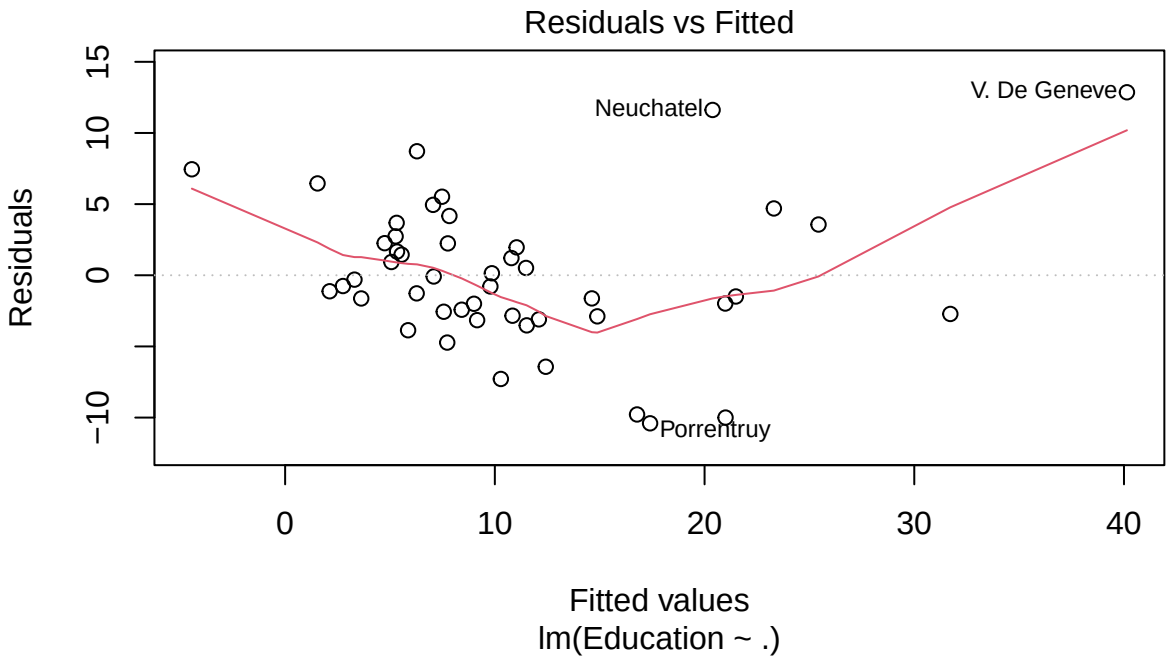


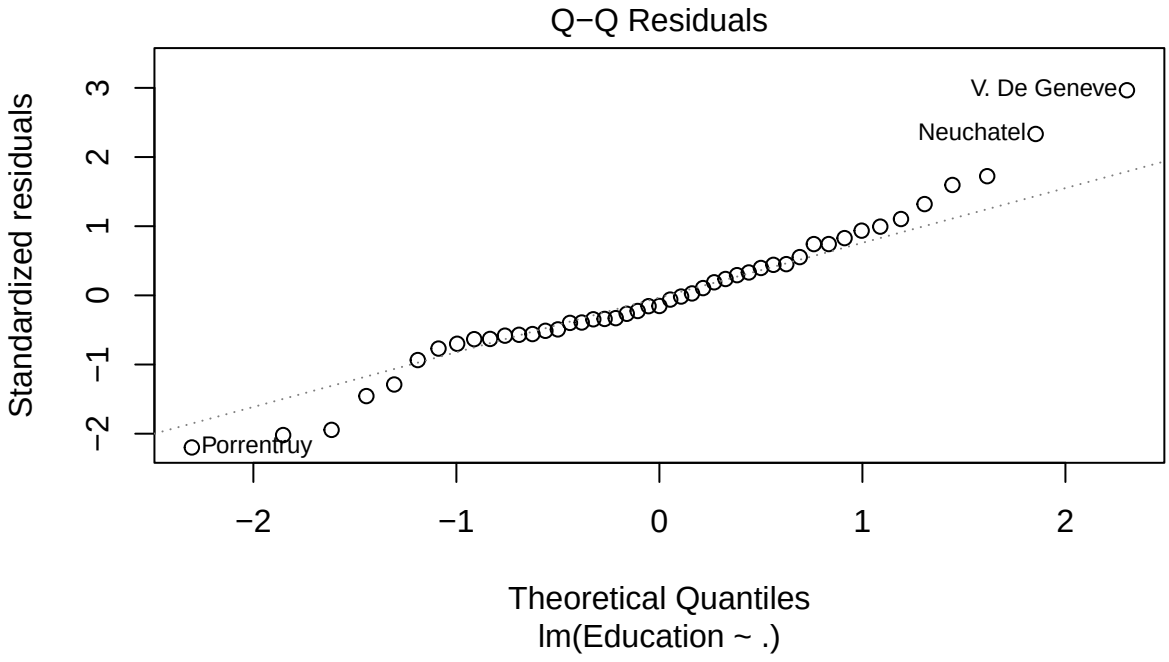
Zwischen den erklärenden Variablen (ausgenommen Education, welche erklärt werden soll) ist nun keine Korrelation mehr erkennbar. Daher kann nun mit diesem Datensatz weitergearbeitet werden.

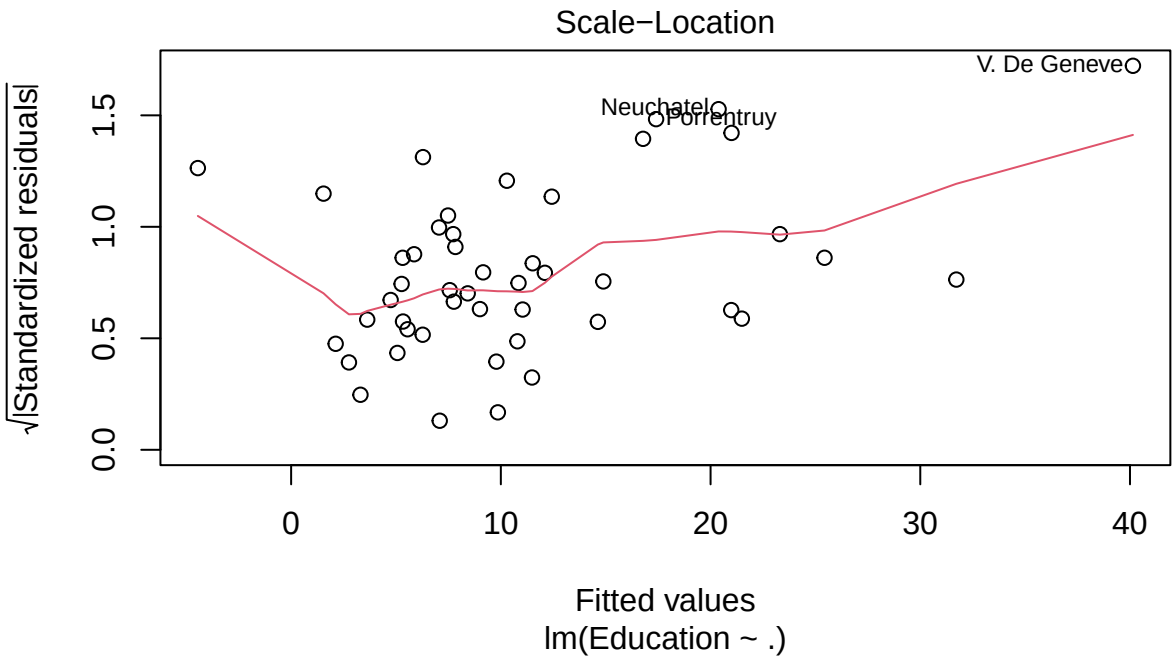
```
modell <- lm(Education ~ ., data)
summary(modell)
```

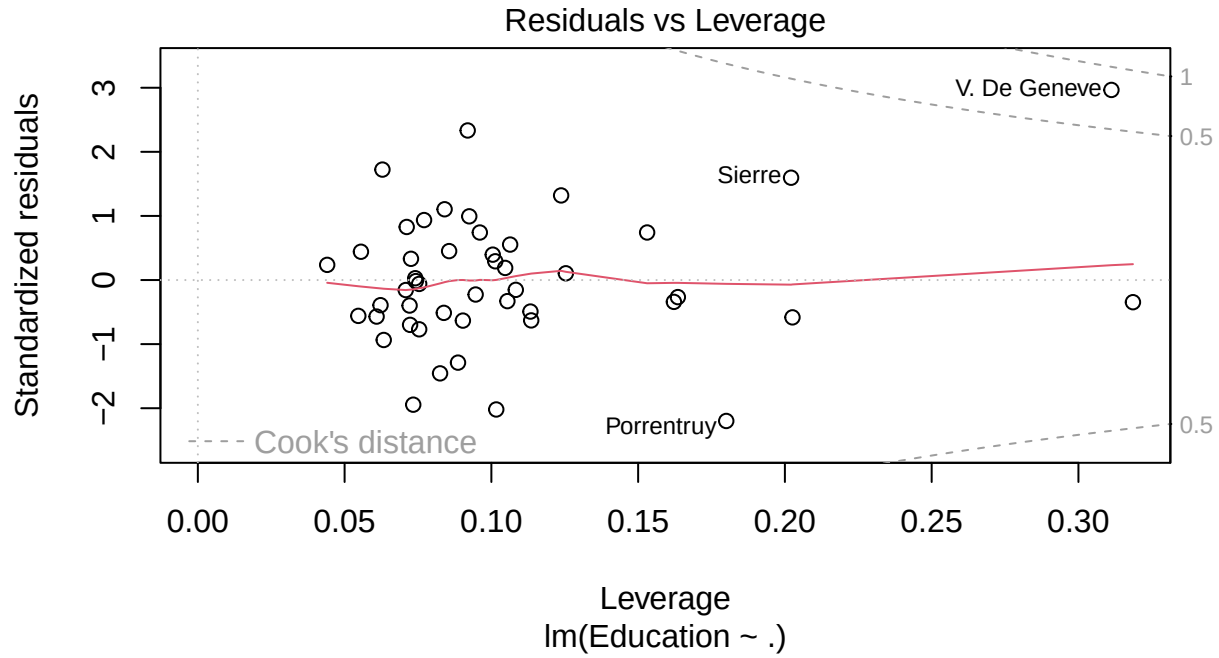
```
>
> Call:
> lm(formula = Education ~ ., data = data)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -10.403  -2.780  -0.757   2.493  12.859
>
> Coefficients:
>                  Estimate Std. Error t value Pr(>|t|)
> (Intercept)       49.9930     6.1864    8.08  4.3e-10 ***
> Fertility         -0.5207     0.0787   -6.62  5.1e-08 ***
> Agriculture       -0.2288     0.0391   -5.86  6.4e-07 ***
> Catholic           0.0833     0.0218    3.82  0.00043 ***
> Infant.Mortality   0.2844     0.3004    0.95  0.34924
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 5.22 on 42 degrees of freedom
> Multiple R-squared:  0.73,    Adjusted R-squared:  0.705
> F-statistic: 28.5 on 4 and 42 DF,  p-value: 1.8e-11
```

```
plot(modell)
```

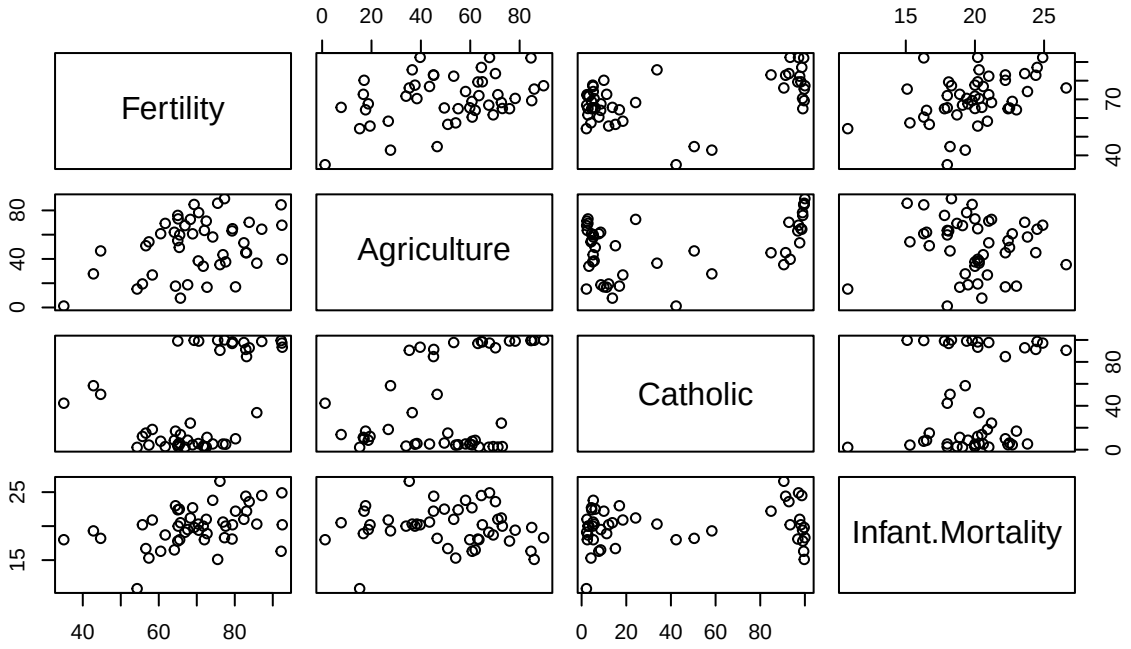

Residuals vs Fitted
Residuals
15
10
5
0
−10
Neuchatel
V. De Geneve
Porrentruy
0
10
20
30
40
Fitted values
lm(Education ~ .)


Q−Q Residuals
Standardized residuals
3
2
1
0
−1
−2
V. De Geneve
Neuchatel
Porrentruy
−2
−1
0
1
2
Theoretical Quantiles
lm(Education ~ .)






```
explanatory<-data %>% select(-Education)
plot(explanatory)
```




```
cor(explanatory)
```

```
>                  Fertility Agriculture Catholic Infant.Mortality
> Fertility            1.000      0.3531    0.464           0.4166
> Agriculture          0.353      1.0000    0.401          -0.0609
> Catholic             0.464      0.4011    1.000           0.1755
> Infant.Mortality     0.417     -0.0609    0.175           1.0000
```

Interpretation der weiteren Plots

Mithilfe der oben angeführten Plots (ausgenommen der ersten beiden Scatterplots) werden nun die Anforderungen für eine multiple lineare Regression geprüft. Aus dem Skript lassen sich folgende Annahmen und Voraussetzungen für eine multiple Regression ableiten: * Das Modell besitzt keinen systematischen Fehler * Die Fehlervarianz ist für alle Beobachtungen gleich groß (homoskedastisch) * Die Komponenten des FEhlerterms sind nicht korreliert * Der Modellfehler sei normalverteilt * Es gibt keine lineare Abhängigkeit zwischen den Regressoren

Aufgrund des Medians (-0.7571), welcher in der Nähe der 'Nulllinie' liegt, kann man schließen, dass die Residuen um Null herum zentriert sind. Diese Bedingung ist daher erfüllt. Am Residual vs. Fitted sowie am Scale-Location Plot ist zu sehen dass die Residuen sowohl unkorreliert als auch homoskedastisch sind.
In der Scatterplot-Matrix ist visuell keine Korrelation zwischen erklärenden Variablen erkennbar. Auch die Korrelationskoeffizienten betragen alle unter 0.5. Daher sind alle Bedingungen für eine multiple lineare Regression erfüllt.

Im Residuals vs. Leverage Plot fallen jedoch diverse Dinge auf. So existiert ein Punkt ziemlich am Ende mit einer hohen Hebelwirkung, welche allerdings nicht negativ ist. Die Punkte Sierre und Porrentuy liegen etwas symmetrisch und innerhalb der Hooks-Distance und müssen daher auch nicht entfernt werden. Kritisch ist allerdings der Punkt V. De Geneve, welcher außerhalb der Hooks-Distance liegt, daher eine große (negative) Hebelwirkung zeigt und daher in Folge entfernt werden sollte nach bisherigem Kenntnissstand.

Eine weitere Bedingung, welche insbesondere für das Testen der Parameter benötigt wird, weniger jedoch für

die lineare Regression selbst, ist die Normalverteilung der Daten. Diese kann über den Q-Q Plot evaluiert werden. In dieser ist zu erkennen, dass es ein paar Punkte gibt, welche sich von der Geraden wegbewegen in Form eines schweren Randes. Dies betrifft nur wenige Punkte, wobei einer davon sowieso entfernt wird aufgrund der Hooks-Distance. Die Daten sind daher nicht komplett normalverteilt, allerdings sind es nur sehr wenige Werte die von der Norm abweichen. Man kann daher von einer ausreichenden Normalverteilung der Daten ausgehen.

Erstellen des Modells

Wir erstellen daher nun ein Modell, welches den Punkt "V. De Geneve" nicht mehr enthält und schauen, wie gut unser Modell funktioniert.

Modell #1: Punkt "V. De Geneve" entfernt, alle Spalten bis auf Examination enthalten

```
data <- swiss %>% select(-Examination)
neudata=rbind(data[1:44,],data[46:47,])
neumodell <- lm(Education ~ ., neudata)
summary(neumodell)
```

```
>
> Call:
> lm(formula = Education ~ ., data = neudata)
>
> Residuals:
>    Min     1Q Median     3Q    Max
> -8.846 -2.313 -0.268  1.918 13.299
>
> Coefficients:
>                  Estimate Std. Error t value Pr(>|t|)
> (Intercept)       41.8071     6.0962    6.86  2.6e-08 ***
> Fertility         -0.4147     0.0778   -5.33  3.8e-06 ***
> Agriculture       -0.1943     0.0367   -5.30  4.3e-06 ***
> Catholic           0.0611     0.0207    2.95   0.0053 **
> Infant.Mortality   0.2602     0.2704    0.96   0.3417
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 4.7 on 41 degrees of freedom
> Multiple R-squared:  0.63,    Adjusted R-squared:  0.594
> F-statistic: 17.4 on 4 and 41 DF,  p-value: 1.97e-08
```

Wir stellen hier zum einen fest, dass unser Modell nur relativ unzureichend ist mit einem Multiple R-squared: 0.6299 sowie Adjusted R-squared: 0.5938. Weiters stellen wir fest, dass die Spalte "Infant.Mortality" zum Modell nichts beiträgt (p=0.34170), und daher in Folge entfernt wird.

Modell #2: Punkt "V. De Geneve" entfernt, alle Spalten bis auf Examination und Infant Mortality enthalten

```
data2 <- swiss %>% select(-Examination, -Infant.Mortality)
neudata2=rbind(data2[1:44,],data2[46:47,])
neumodell2 <- lm(Education ~ ., neudata2)
summary(neumodell2)
```

```
>
> Call:
> lm(formula = Education ~ ., data = neudata2)
>
```

```
> Residuals:
>    Min     1Q Median     3Q    Max
> -9.014 -2.441 -0.769  2.641 14.020
>
> Coefficients:
>             Estimate Std. Error t value Pr(>|t|)
> (Intercept)  45.2686     4.9166    9.21  1.2e-11 ***
> Fertility    -0.3847     0.0712   -5.40  2.9e-06 ***
> Agriculture  -0.2024     0.0357   -5.68  1.2e-06 ***
> Catholic      0.0619     0.0207    2.99   0.0047 **
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 4.7 on 42 degrees of freedom
> Multiple R-squared:  0.622,   Adjusted R-squared:  0.595
> F-statistic:   23 on 3 and 42 DF,  p-value: 5.78e-09
```

Interessanterweise wurde unser Modell etwas schlechter (Multiple R-squared: 0.6216, Adjusted R-squared: 0.5945; wobei wir uns erstmal auf den R2-Wert beziehen), obwohl wir die Spalte Infant Mortality entfernt hatten. Dies erscheint uns etwas paradox, da wir uns eigentlich ein deutlich besser angepasstes Modell erwartet hatten. Unsere Überlegung ist daher, den Hebelpunkt "V. De Geneve" doch in unseren Daten zu lassen in der Hoffnung auf ein besser angepasstes Ergebnis. Dies wird nun getestet.

Modell #3: Punkt "V. De Geneve" behalten, alle Spalten bis auf Examination und Infant Mortality enthalten

```
data2 <- swiss %>% select(-Examination, -Infant.Mortality)
neumodell3 <- lm(Education ~ ., data2)
summary(neumodell3)
```

```
>
> Call:
> lm(formula = Education ~ ., data = data2)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -10.085  -2.952  -0.668   3.252  12.971
>
> Coefficients:
>             Estimate Std. Error t value Pr(>|t|)
> (Intercept)  53.8505     4.6491   11.58  8.3e-15 ***
> Fertility    -0.4888     0.0710   -6.88  1.9e-08 ***
> Agriculture  -0.2380     0.0378   -6.30  1.3e-07 ***
> Catholic      0.0844     0.0217    3.88  0.00035 ***
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 5.22 on 43 degrees of freedom
> Multiple R-squared:  0.725,   Adjusted R-squared:  0.706
> F-statistic: 37.7 on 3 and 43 DF,  p-value: 4.12e-12
```

Aus den R2 sowie den R2-adjusted Werten ist zu erkennen, dass es sich bei dem Modell ohne der Variable "Infant Mortality" sowie inklusive dem Wert "V. De Geneve" um das Beste Modell handelt, welches wir erstellt hatten.

Wir zeigen nun daher die Formel für das Regressionsmodell.

Regressionsmodell

Das daraus reslultierende Regressionsmodell lautet:

$$Education_i = \alpha + \beta_{Fertility} \times x_{Fertility,i} + \beta_{Agriculture} \times x_{Agriculture,i} + \beta_{Catholic} \times x_{Catholic,i} + \varepsilon_i$$

Modellgleichung

Die angepasste Modellgleichung ist:

$$Education_i = 53.8505 + -0.48883 \times x_{Fertility,i} + -0.23799 \times x_{Agriculture,i} + 0.08440 \times x_{Catholic,i}$$

Interpretation der Koeffizienten

Der intercept alpha bedeutet, dass 45.2686% der Bevölkerung eine Ausbildung höher als die der Grundschule hätte wenn... * die Bevölkerung komplett unfruchtbar wäre * kein Mann mehr in der Landwirtschaft tätig wäre und * niemand katholisch wäre.

Die einzelnen Beta-Koeffizienten stellen dar wie stark (prozentual) die Bevölkerung mit einer Ausbildung höher als die der Grundschule steigen würde, vorausgesetzt das dazugehörige Maß steigt um 1% während die anderen gleich bleiben.

Das bedeutet für 1% mehr in der “common standardized fertility measure” sind es 0.48883% weniger, für 1% mehr Männer in der Landwirtschaft 0.23799% weniger und für 1% mehr Katholiken sind es 0.08440% mehr Menschen mit einer höheren Ausbildung als Grundschulausbldung.

Aufgabe 3: USA

- Wir kehren zurück zu den Variablen “Population”, “Income”, “Illiteracy”, “Life.Exp”, “Murder”, “HS Grade” und “Frost” aus dem R Datensatz state.x77.
- Passen Sie für die oben genannten Variablen ein lineares Modell (lm) an, das “Murder” durch die übrigen Variablen erklärt, soweit dies zulässig ist.

```
glimpse(state.x77)
```

```
##  num [1:50, 1:8] 3615 365 2212 2110 21198 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ...
##   ..$ : chr [1:8] "Population" "Income" "Illiteracy" "Life Exp" ...
```

```
class(state.x77)
```

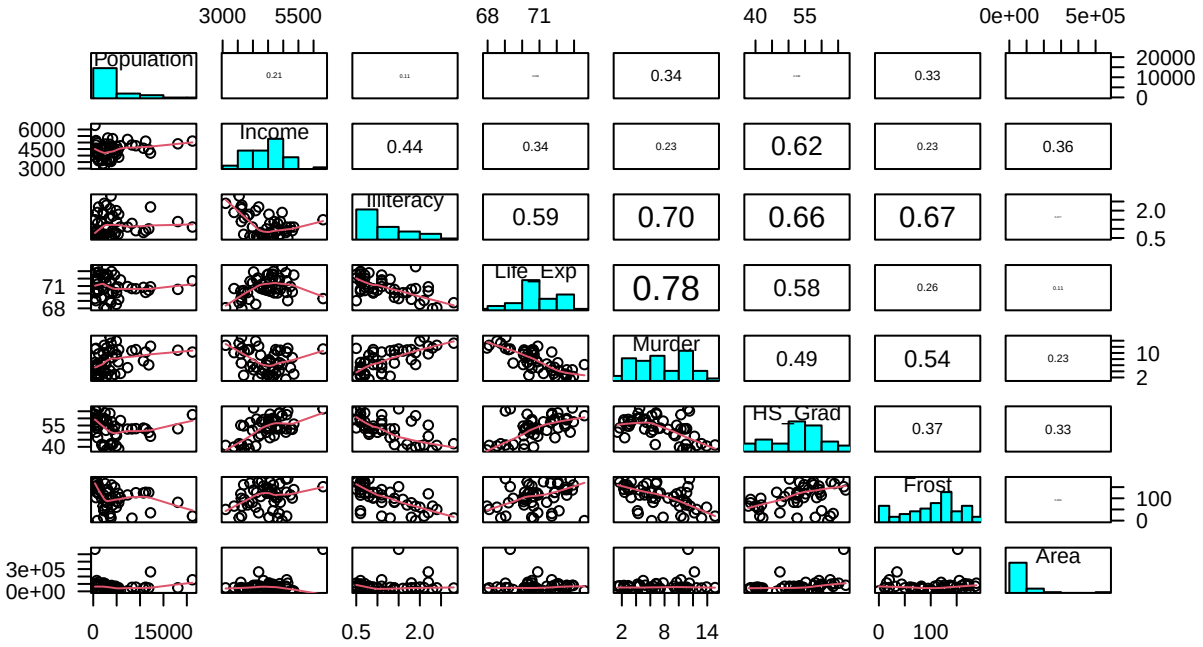
```
## [1] "matrix" "array"
```

Beschreibung des Datensatzes: * **Population:** Bevölkerungsanzahl am 01. Juli 1975 (Original [en]: population estimate as of July 1, 1975) * **Income:** Einkommen pro Kopf (Stand: 1974) (Original [en]: per capita income (1974)) * **Illiteracy:** Analphabetismus (Stand: 1970; % der Bevölkerung) (Original [en]: illiteracy (1970, percent of population)) * **Life Exp:** Lebenserwartung in Jahren (Original [en]: ife expectancy in years (1969–71)) * **Murder:** Anzahl an Mördern und nicht fahrlässige Tötungen je 100.000 Einwohnern (Original [en]: murder and non-negligent manslaughter rate per 100,000 population (1976)) * **HS Grad:** Highschool Abschlüsse in Prozent (Original [en]: percent high-school graduates (1970)) * **Frost:** Durchschnittstage mit Frost (Temperaturen unterhalb des Gefrierpunktes) in den Haupt- oder Großstädten (Original [en]: mean number of days with minimum temperature below freezing (1931–1960) in capital or large city) * **Area:** Fläche der Region in Quadratmeilen (Original [en]: land area in square miles)

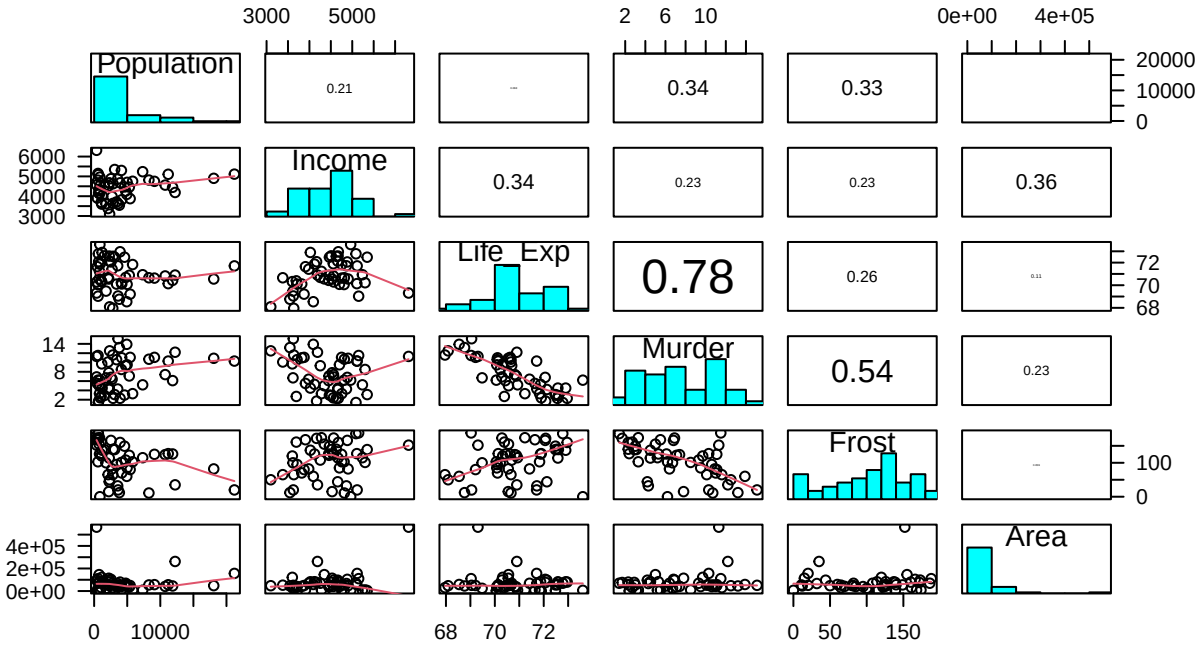
Scatterplots

Zunächst wurden Scatterplots erstellt, um einen Überblick darüber zu erhalten welche Variablen mit “Murder” in einem linearen zusammenhang zu stehen scheinen.

```
rm(data)
data<-as.data.frame(state.x77) %>%
  rename(Life_Exp = "Life Exp",
         HS_Grad = "HS Grad")
```



Da wir sehen, dass die Variable “HS_Grad” mit nahezu allen anderen Variablen sehr gut korreliert, ist diese als erklärende Variable auszuschließen. Gleiches gilt für Illiteracy.

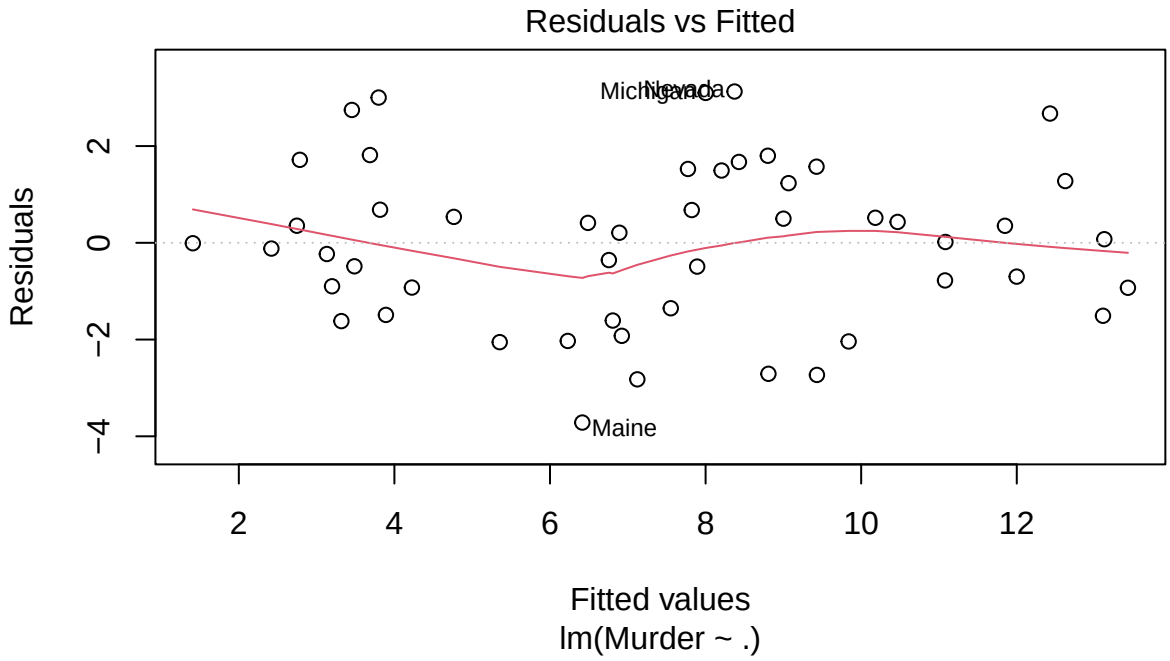


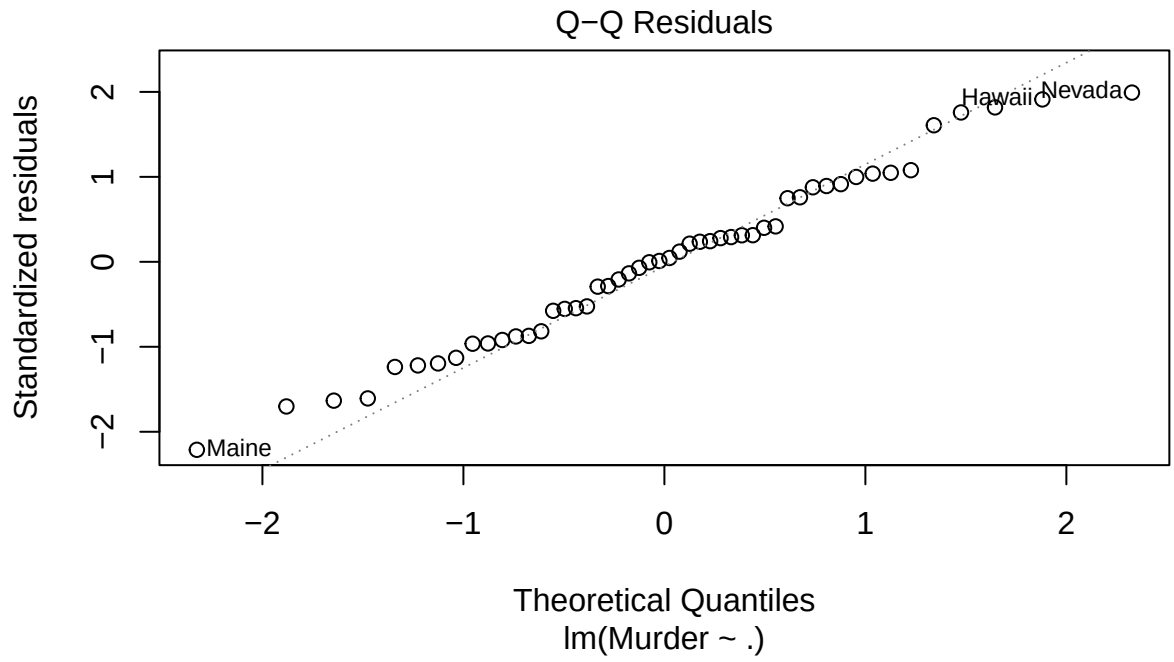
Nun sind nurmehr Variablen übrig, welche als erklärende Variablen nicht hoch miteinander korrelieren.

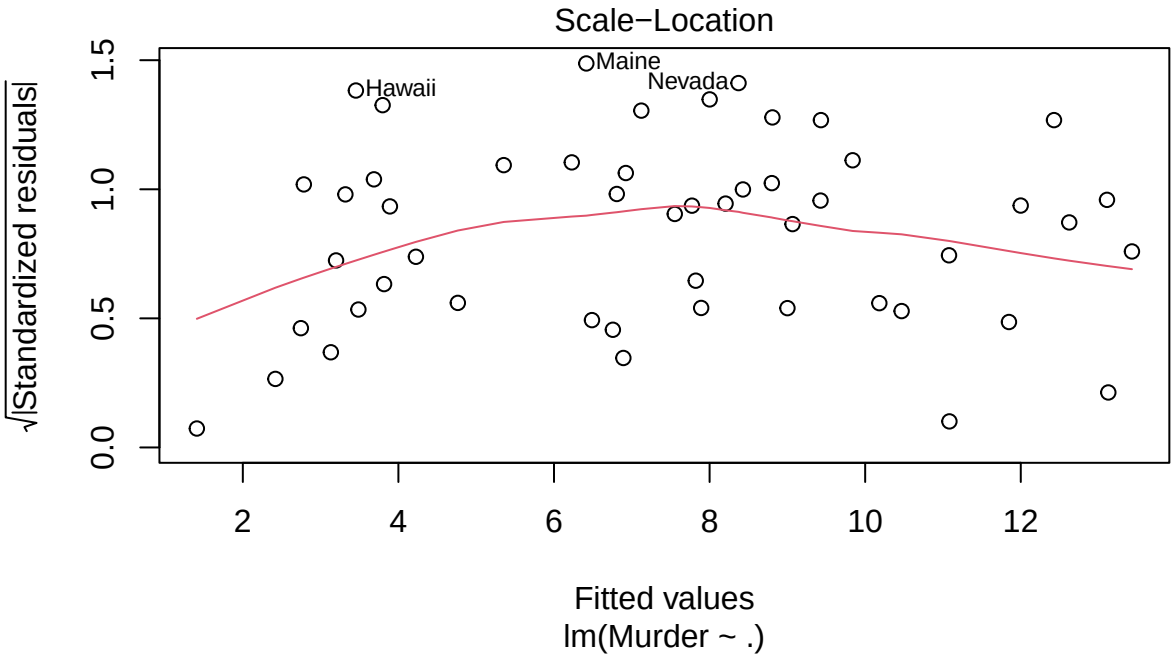
Überprüfung der statistischen Voraussetzungen

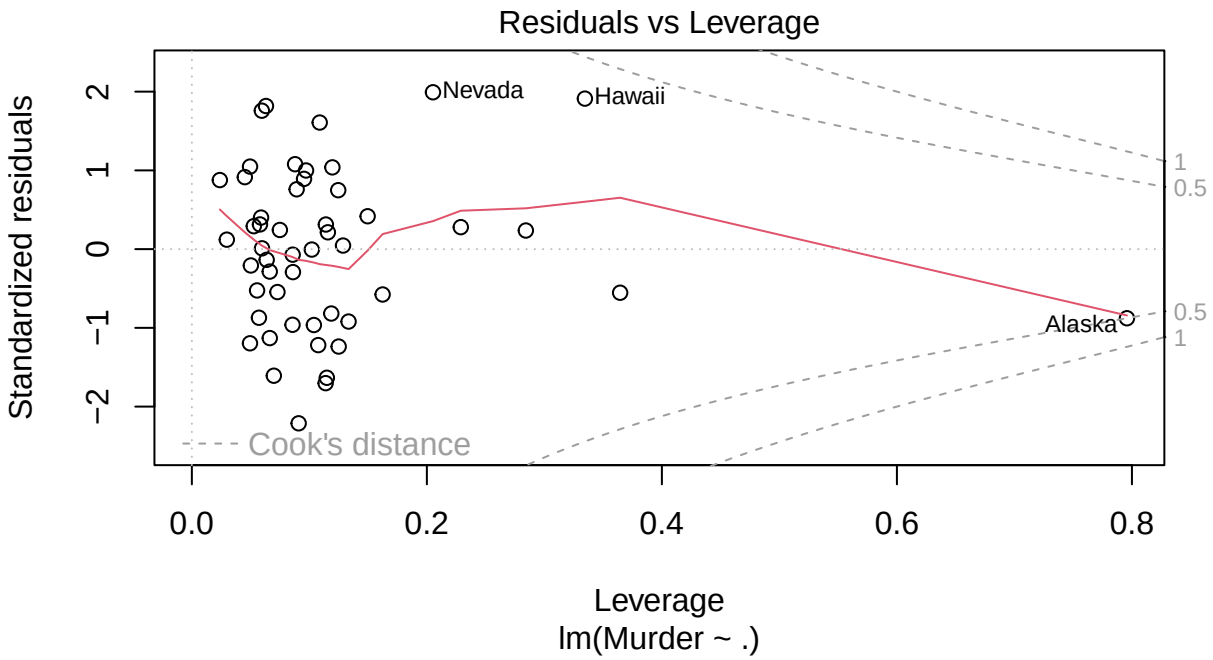
Summary und Residuenplots

```
>
> Call:
> lm(formula = Murder ~ ., data = data2)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -3.716 -1.247   0.046   1.266   3.128
>
> Coefficients:
>              Estimate Std. Error t value Pr(>|t|)
> (Intercept)  1.36e+02   1.43e+01    9.54  2.8e-12 ***
> Population   1.71e-04   6.33e-05    2.70  0.00970 **
> Income      -3.25e-04   5.14e-04   -0.63  0.53040
> Life_Exp    -1.79e+00   2.12e-01   -8.42  1.0e-10 ***
> Frost       -2.12e-02   5.49e-03   -3.87  0.00036 ***
> Area         8.28e-06   3.30e-06    2.51  0.01584 *
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 1.76 on 44 degrees of freedom
> Multiple R-squared:  0.796,   Adjusted R-squared:  0.772
> F-statistic: 34.2 on 5 and 44 DF,  p-value: 4.14e-14
```




Q−Q Residuals
Standardized residuals
Hawaii
Nevada
Maine
−2
−1
0
1
2
Theoretical Quantiles
lm(Murder ~ .)


Scale−Location
√|Standardized residuals|
Maine
Hawaii
Nevada
0.0
0.5
1.0
1.5
2
4
6
8
10
12
Fitted values
lm(Murder ~ .)



In unserem ersten erstellen Modell sehen wir, dass dass nicht alle Variablen zur Erklärung herangezogen werden können. Income spielt keine Rolle mit einem P-Value von 0.530396. Die Region spielt hier auch nur eine untergeordnete Rolle im Vergleich zu den anderen Variablen, kann aber durchaus als erklärende Variable hinzugezogen werden.

Im Plot **Residuals vs. Fitted** haben wir eine gleichmäßig verteilte Punktewolke, welche keinen systematischen Fehler aufweist und auch keine Korrelationen erkennen lassen. Die Bedingung wäre daher erfüllt.

Im Plot **Residuals vs. Leverage** sehen wir eine Beobachtung "Alaska", welche exakt auf der Cooks Distance liegt und daher eine große Hebelwirkung erzielt. Diese ist daher zu entfernen. Es existieren noch zwei weitere Beobachtungen "Nevada" und "Hawaii", welche eine relativ große Hebelwirkung aufweisen und sehr Nahe der Cooks Distance liegen - wir aber noch nicht entfernen müssen, da sie nach wie vor innerhalb dieses Bereiches liegen.

Beim Plot **Scale Location** sehen wir eine leichte Drehung der Daten, vermutlich verursacht durch die Variablen Hawaii, Maine und Nevada. Jedoch scheint der Plot noch in Ordnung zu sein.

Die **Q-Q Residuals** zeigen nahezu normalverteilte Daten, welche für unsere Regression ausreichen sein sollte.

Im Folgenden Schritt entfernen wir nun die Variable Income als nicht ausreichend erklärende Variable sowie die Beobachtung "Alaska".

```
data3 <- data2 %>% select(-Income)
data3 <- data3[-2,]
```

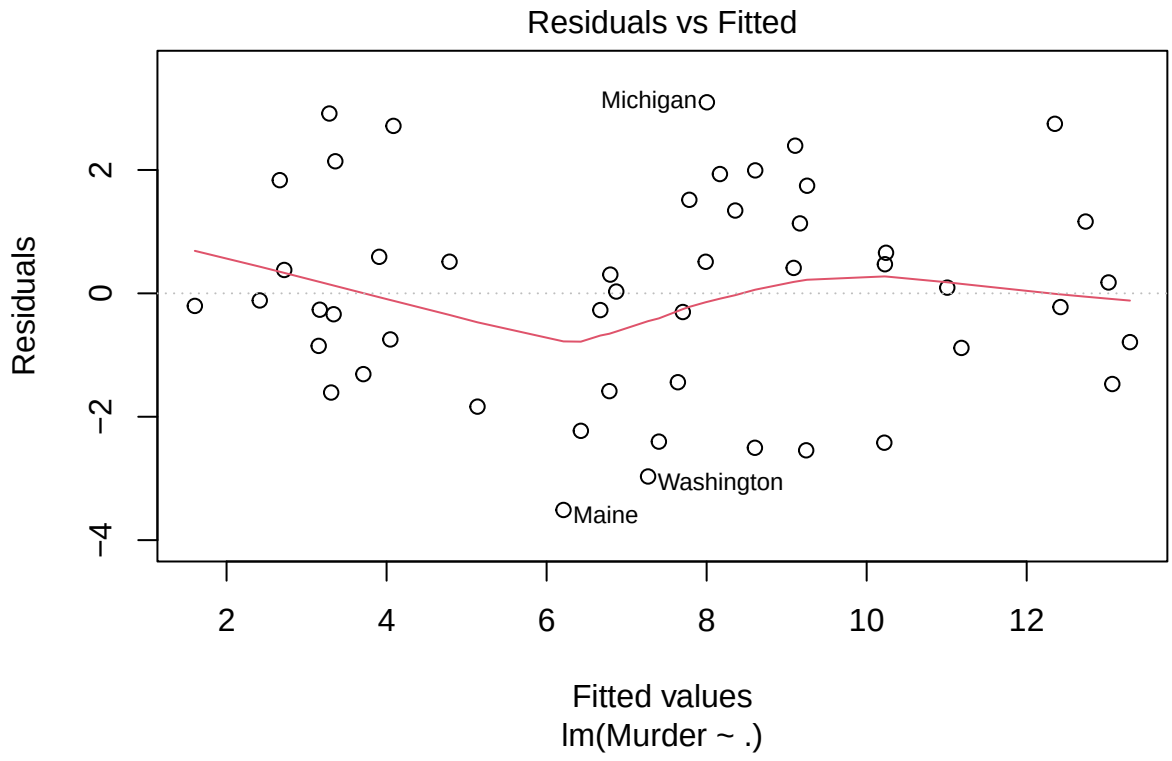
Im folgenden schauen wir uns das Modell nochmal an:

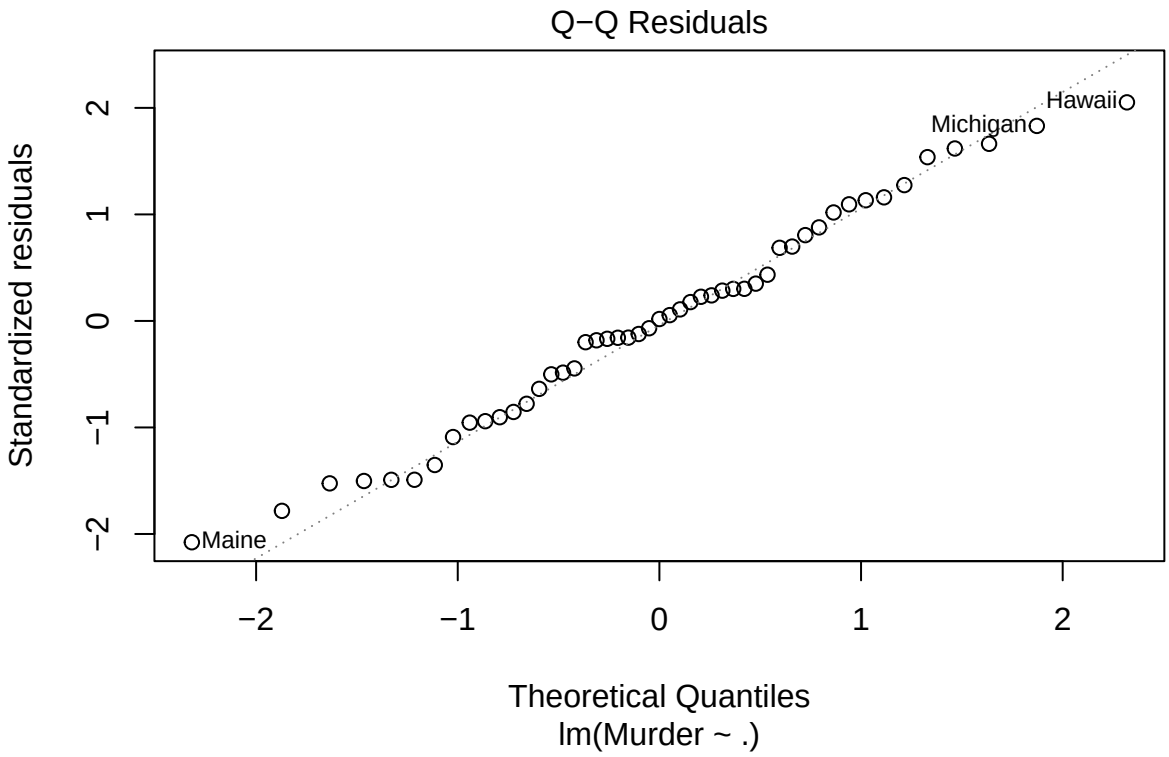
```
modell_new <- lm(Murder ~ ., data3)
summary(modell_new)
```

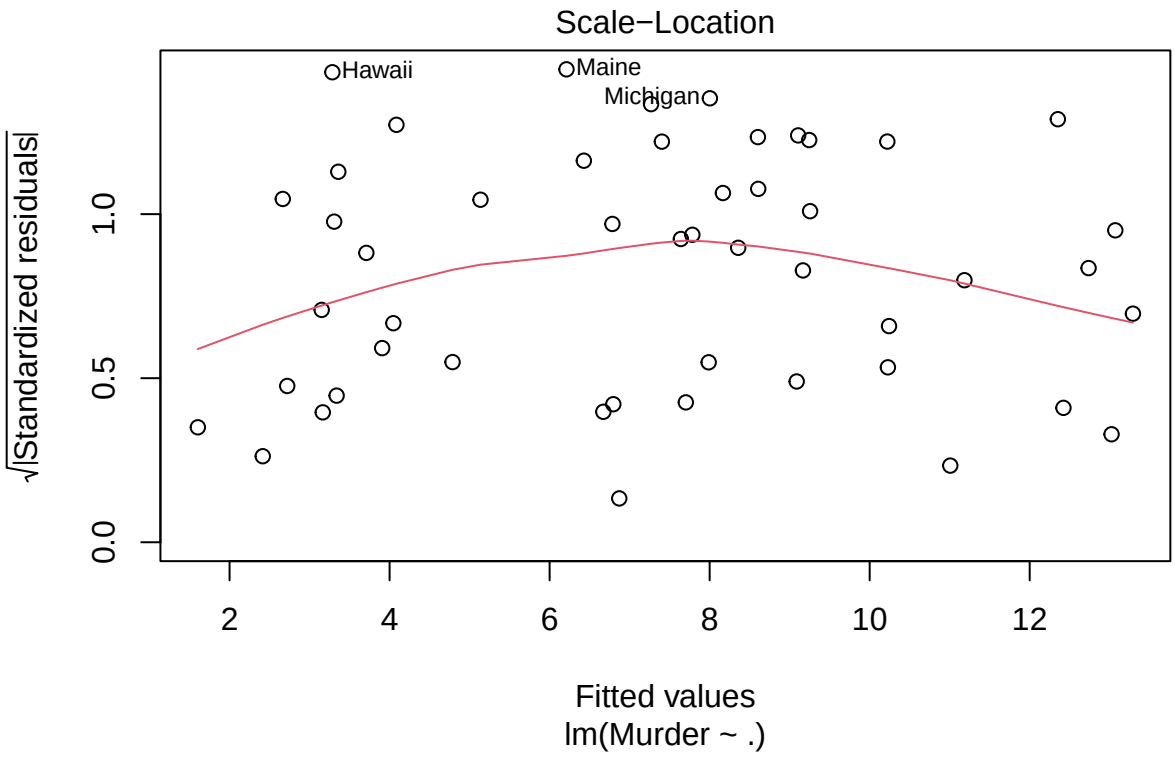
```
##
## Call:
## lm(formula = Murder ~ ., data = data3)
##
```

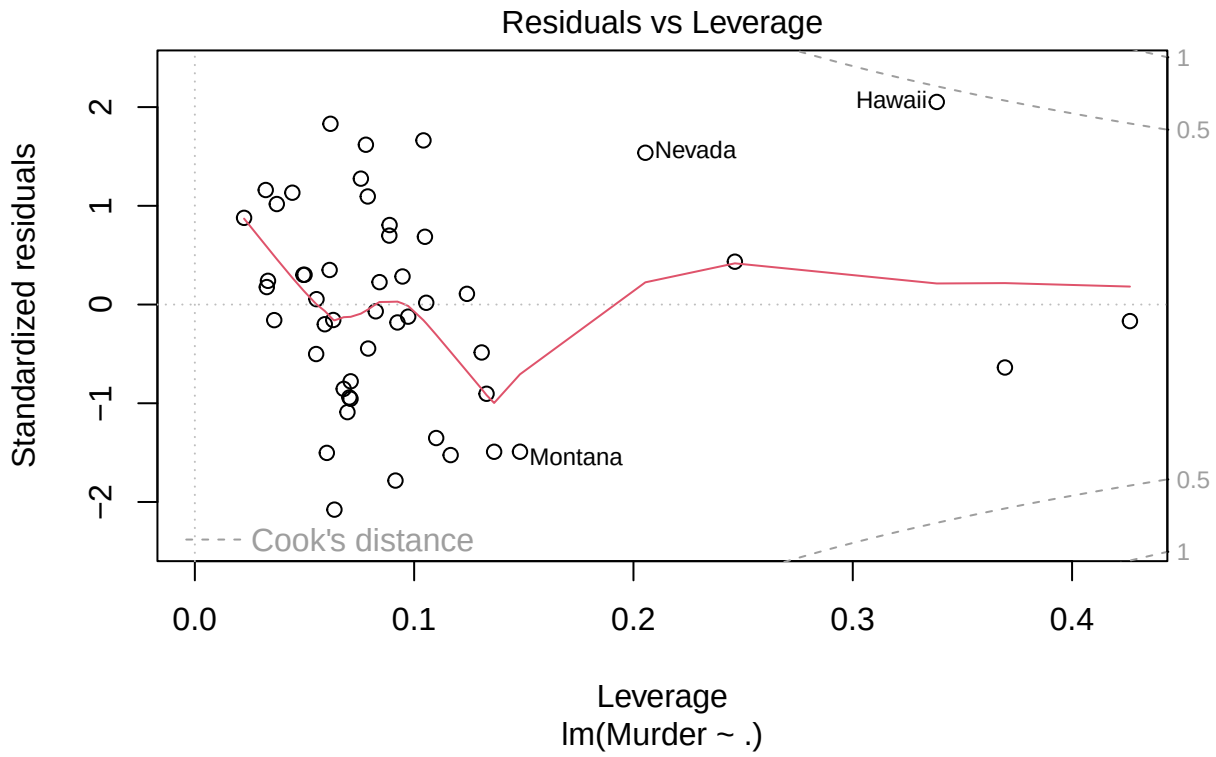
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.511 -1.309  0.029  1.164  3.097
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.41e+02   1.39e+01   10.19  3.8e-13 ***
## Population   1.43e-04   6.09e-05    2.35  0.02351 *
## Life_Exp    -1.88e+00   1.98e-01   -9.49  3.3e-12 ***
## Frost       -2.14e-02   5.33e-03   -4.01  0.00023 ***
## Area         1.25e-05   5.55e-06    2.25  0.02982 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.75 on 44 degrees of freedom
## Multiple R-squared:  0.794,  Adjusted R-squared:  0.775
## F-statistic: 42.4 on 4 and 44 DF,  p-value: 1.45e-14
```

```
plot(modell_new)
```




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Murder ~ .)
Maine
Michigan
Hawaii
−2
−1
0
1
2


Scale−Location
Hawaii
Maine
Michigan
√|Standardized residuals|
0.0
0.5
1.0
2
4
6
8
10
12
Fitted values
lm(Murder ~ .)




```
data4 <- data3 %>% select(-Murder)
cor(data4)
```

```
##            Population Life_Exp   Frost    Area
## Population     1.0000  -0.0912 -0.3209  0.2361
## Life_Exp      -0.0912   1.0000  0.2910  0.0633
## Frost         -0.3209   0.2910  1.0000 -0.0951
## Area           0.2361   0.0633 -0.0951  1.0000
```

Unser neues Modell zeigt nun nur mehr statistisch erklärende Daten, wobei Area (p-Value von 0.029819) und Population (p-Value von 0.023509) im Vergleich einen deutlich geringeren Einfluss haben. Der Median (0.0295) zeigt außerdem, dass die Daten nahe der Null-Linie verteilt sind und das Modell daher gültig sein sollte. Ebenso zeigen die ausgerechneten Korrelationen der erklärenden Variablen keine Zusammenhänge / Korrelationen zueinander (< 0.5).

Regressionsmodell

Das daraus reslultierende Regressionsmodell lautet:

$$Murder_i = \alpha + \beta_{Population} \times x_{Population,i} + \beta_{Life_Exp} \times x_{Life_Exp,i} + \beta_{Frost} \times x_{Frost,i} + \beta_{Area} \times x_{Area,i} + \varepsilon_i$$

Modellgleichung

Die angepasste Modellgleichung ist:

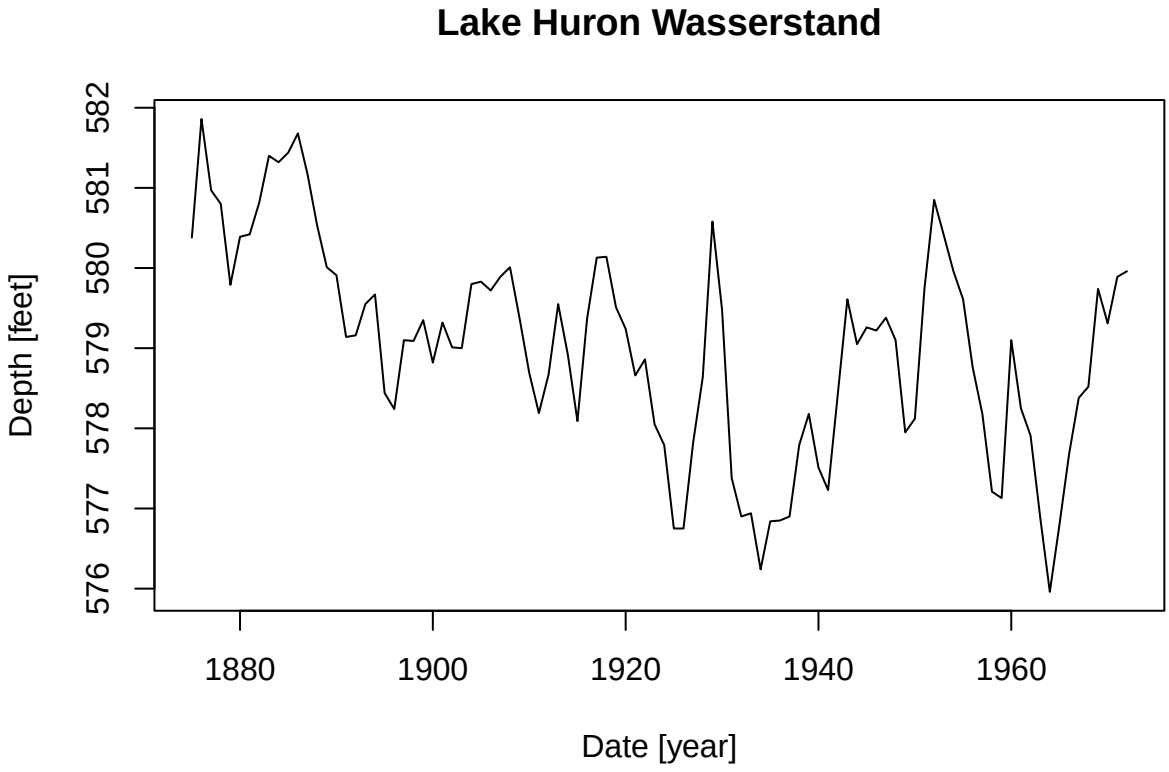
$$Murder_i = 1.414e{+}02 + 1.430e{-}04 \times x_{Population,i} + -1.879e{+}00 \times x_{Life_Exp,i} + -2.135e{-}02 \times x_{Frost,i} + 1.246e{-}05 \times x_{Area,i}$$

Aufgabe 4: Lake Huron

- Wir kehren zurück zum Datensatz "LakeHuron".
- Passen Sie ein Modell an, das den Zeittrend modelliert.
- Überprüfen Sie alle erforderlichen statistischen Voraussetzungen für die Gültigkeit dieses Modells mtihilfe der quality plots der Residuen.

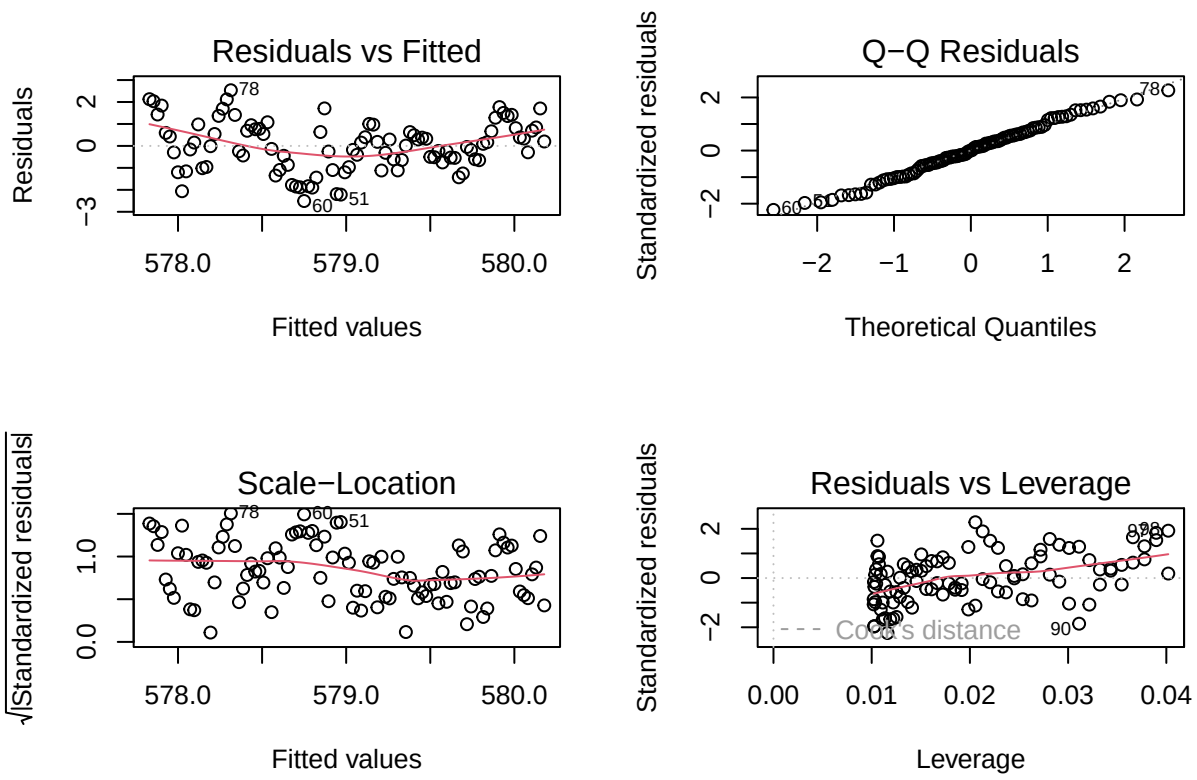
Beschreibung des Datensatzes: Jährliche Messungen des Pegels des Huron-Sees in Fuß, 1875-1972.

Erste Untersuchung der Daten




```
Mean:               579.00
Median:             579.12
Min:                575.96
Max:                581.86
Variance:           1.74
Standardeviation:   1.32
```

Modellanpassung



Anhand der Modelanpassung können folgende Aussagen getroffen werden:

- Im Plot *_**Residuals vs Fitted:** Die Residuen schwanken relativ gleichmäßig um Null und die angepasste Rote Linie ist leicht gebogen, **was einen systematischen Fehler hindeutet bzw. eben korrelierte Daten**. Damit ist eine **Voraussetzung für eine multiple Regression nicht mehr gegeben**, weshalb die Analyse an dieser Stelle abgebrochen wird. Die Alternative wäre hier, das Regressionsmodell zu erstellen und anschließend zu sagen, dass dieses Modell aufgrund vorher genannten systematischen Fehlers nicht gültig wäre.
- Im Plot ***Scale-Location:** Die Standardabweichung der Residuen scheinen relativ konstant und es kann von einer Homoskedastizität ausgegangen werden.
- Im Plot **Normal Q-Q:** Die Punkte liegen großteils auf beziehungsweise sehr nah an der Referenzlinie, weshalb man von einer Normalverteilung ausgehen kann.
- Im Plot **Residuals vs Leverage:** Da alle Werte innerhalb der Cook's Distance liegen kann kein Hebelpunkt indentifiziert werden. Denoch ist auch hier noch der Kurvenverlauf wie in Residuals vs Fitted in Teilen zu erkennen.

Aufgabe 5: Pima Indians

- Laden Sie den Datensatz 'Pima.tr' aus der library 'MASS'.
- 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.
- Ermitteln Sie die prädiktive Qualität des Modells mithilfe einer Receiver Operating Characteristic (ROC) Kurve.
- Führen Sie auch die False Positive, False Negative, True Positive und True Negative Raten in einer Tabelle (Konfusionsmatrix) an.

Beschreibung des Datensatzes: Eine Population von Frauen, die mindestens 21 Jahre alt waren, von den Pima-Indianern abstammten und in der Nähe von Phoenix, Arizona, lebten, wurde nach den Kriterien der Weltgesundheitsorganisation auf Diabetes getestet. Die Daten wurden vom US National Institute of Diabetes and Digestive and Kidney Diseases erhoben. Wir haben die 532 vollständigen Datensätze verwendet, nachdem wir die (größtenteils fehlenden) Daten zum Seruminsulin herausgenommen haben.

- **npreg:** Anzahl der Schwangerschaften (Original [en]: number of pregnancies.)
- **glu:** Plasmaglukosekonzentration bei einem oralen Glukosetoleranztest. (Original [en]: plasma glucose concentration in an oral glucose tolerance test.)
- **bp:** diastolischer Blutdruck (mm Hg). (Original [en]: diastolic blood pressure (mm Hg).)
- **skin:** Dicke der Trizepshautfalte (mm). (Original [en]: triceps skin fold thickness (mm).)
- **bmi:** BMI (Original [en]: body mass index (weight in kg/(height in m)
- **ped:** Diabetes-Stammbaumfunktion. (Original [en]: diabetes pedigree function.)
- **age:** Alter in Jahren (Original [en]: age in years.)
- **type:** Diabetis ja/nein (Original [en]: Yes or No, for diabetic according to WHO criteria.)

```
library(MASS)
library(corrplot)
library(pROC)
head(Pima.tr)
```

```
##   npreg glu bp skin  bmi   ped age type
## 1     5  86 68   28 30.2 0.364  24   No
## 2     7 195 70   33 25.1 0.163  55  Yes
## 3     5  77 82   41 35.8 0.156  35   No
## 4     0 165 76   43 47.9 0.259  26   No
## 5     0 107 60   25 26.4 0.133  23   No
## 6     5  97 76   27 35.6 0.378  52  Yes
```

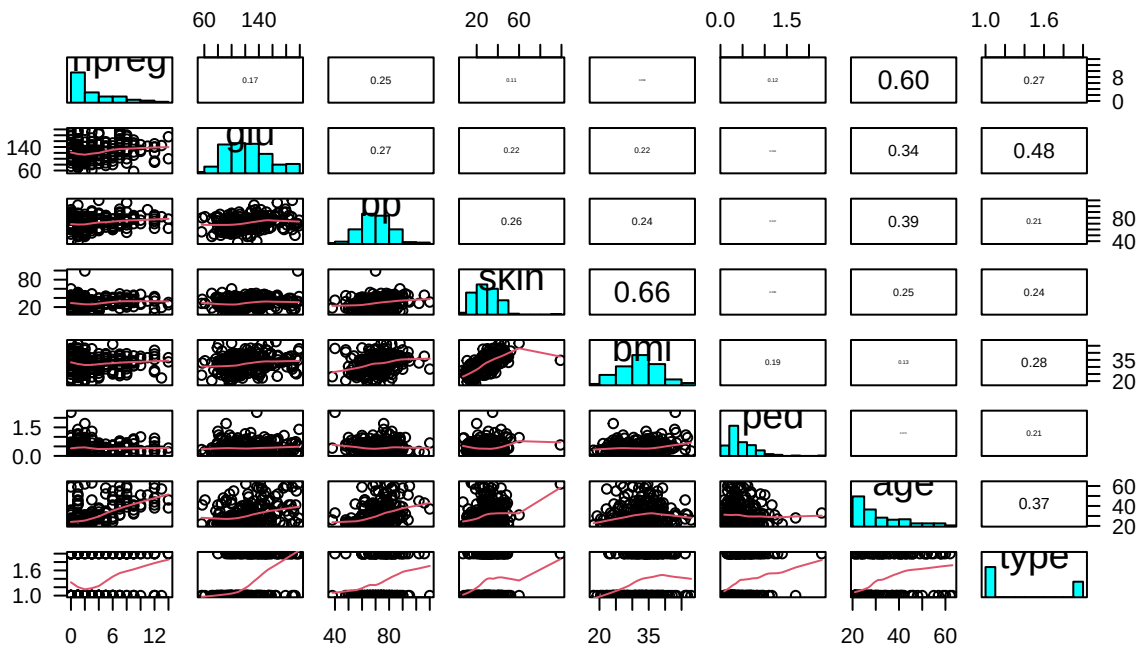
```
str(data)
```

```
## 'data.frame':    50 obs. of  8 variables:
##  $ Population: num  3615 365 2212 2110 21198 ...
##  $ Income    : num  3624 6315 4530 3378 5114 ...
##  $ Illiteracy: num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
##  $ Life_Exp  : num  69 69.3 70.5 70.7 71.7 ...
##  $ Murder    : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...
##  $ HS_Grad   : num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40.6 ...
##  $ Frost     : num  20 152 15 65 20 166 139 103 11 60 ...
##  $ Area      : num  50708 566432 113417 51945 156361 ...
```

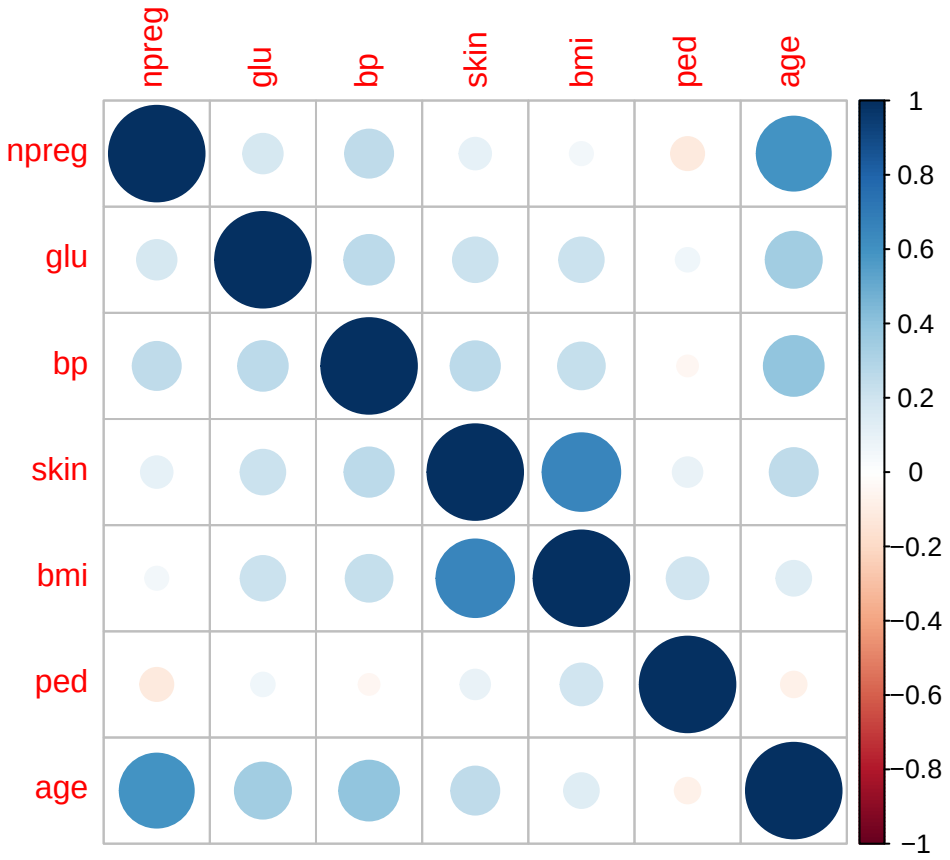
```
PrimaOriginal <- Pima.tr
```

Wir haben hier prinzipiell die Situation, dass die Variable type nominal-skaliert ist mit zwei Ausprägungen "ja" und "nein". Wir können weiters hier von einer Binominalverteilung ausgehen, da es nur zwei Ergebnisse

gibt diesbezüglich. Als Folge wäre hier eine logistische Regression anzuwenden.




```
Prima_test <- PrimaOriginal
Prima_test$type <- NULL
cor(Prima_test) %>% corrplot()
```




```
cor(Prima_test)
```

```
##          npreg    glu      bp   skin    bmi      ped     age
## npreg   1.0000 0.1705  0.2521 0.1090 0.0583 -0.1195  0.5989
## glu     0.1705 1.0000  0.2694 0.2176 0.2168  0.0607  0.3434
## bp      0.2521 0.2694  1.0000 0.2650 0.2388 -0.0474  0.3911
## skin    0.1090 0.2176  0.2650 1.0000 0.6590  0.0954  0.2519
## bmi     0.0583 0.2168  0.2388 0.6590 1.0000  0.1906  0.1319
## ped    -0.1195 0.0607 -0.0474 0.0954 0.1906  1.0000 -0.0714
## age     0.5989 0.3434  0.3911 0.2519 0.1319 -0.0714  1.0000
```

Wir sehen, dass hier BMI und skin miteinander korrelieren sowie npreg und age. Da der Datensatz recht groß ist, haben wir uns dazu entschieden erstmal zu schauen, ob diese Koeffizienten überhaupt etwas zum Modell beitragen und erstmal ein generalisiertes lineares Modell zu erstellen.

Modell 1 - alle Koeffizienten vorhanden

```
(modell_glm <- glm(type~.,data=PrimaOriginal,family=binomial(link = "logit")))
```

```
##
## Call:  glm(formula = type ~ ., family = binomial(link = "logit"), data = PrimaOriginal)
##
## Coefficients:
## (Intercept)        npreg          glu           bp         skin          bmi
##    -9.77306      0.10318      0.03212     -0.00477     -0.00192      0.08362
##         ped          age
##     1.82041      0.04118
```

```
##
## Degrees of Freedom: 199 Total (i.e. Null);  192 Residual
## Null Deviance:       256
## Residual Deviance: 178   AIC: 194
```

```
summary(modell_glm)
```

```
##
## Call:
## glm(formula = type ~ ., family = binomial(link = "logit"), data = PrimaOriginal)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.77306    1.77039   -5.52  3.4e-08 ***
## npreg        0.10318    0.06469    1.59   0.1107
## glu          0.03212    0.00679    4.73  2.2e-06 ***
## bp          -0.00477    0.01854   -0.26   0.7971
## skin        -0.00192    0.02250   -0.09   0.9321
## bmi          0.08362    0.04283    1.95   0.0509 .
## ped          1.82041    0.66551    2.74   0.0062 **
## age          0.04118    0.02209    1.86   0.0623 .
## ---
## 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: 194.4
##
## Number of Fisher Scoring iterations: 5
```

Wie wir in obigem Modell sehen, tragen die Koeffizienten npreg, bp und skin ohnehin nichts zum Modell bei und werden daher entfernt. Damit erübrigen sich auch die Korrelationen der ersten explorativen Analyse.

Modell 2 - Koeffizienten npreg, bp und skin entfernt.

```
Prima_filtered <- PrimaOriginal
Prima_filtered$npreg <- NULL
Prima_filtered$bp <- NULL
Prima_filtered$skin <- NULL

(modell_glm2 <- glm(type~.,data=Prima_filtered,family=binomial(link = "logit")))
```

```
##
## Call:  glm(formula = type ~ ., family = binomial(link = "logit"), data = Prima_filtered)
##
## Coefficients:
## (Intercept)          glu          bmi          ped          age
##     -9.9714       0.0313       0.0770       1.7198       0.0586
##
## Degrees of Freedom: 199 Total (i.e. Null);  195 Residual
## Null Deviance:       256
## Residual Deviance: 181   AIC: 191
```

```
summary(modell_glm2)
```

```
##
## Call:
## glm(formula = type ~ ., family = binomial(link = "logit"), data = Prima_filtered)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.97139    1.52759   -6.53  6.7e-11 ***
## glu          0.03126    0.00663    4.72  2.4e-06 ***
## bmi          0.07703    0.03225    2.39  0.01692 *
## ped          1.71979    0.65609    2.62  0.00876 **
## age          0.05860    0.01757    3.33  0.00085 ***
## ---
## 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: 181.08  on 195  degrees of freedom
## AIC: 191.1
##
## Number of Fisher Scoring iterations: 5
```

Wir sehen, dass hier nur noch Koeffizienten dabei sind, welche zum Modell etwas beitragen. Hinzu kommt, dass sich die p-Values der Koeffizienten age und bmi deutlich 'gebessert' haben bzw. niedriger sind.

ROC Kurve

Wir machen hier zu beiden Modellen eine ROC-Kurve um die prädiktive Qualität des Modells zu beurteilen.

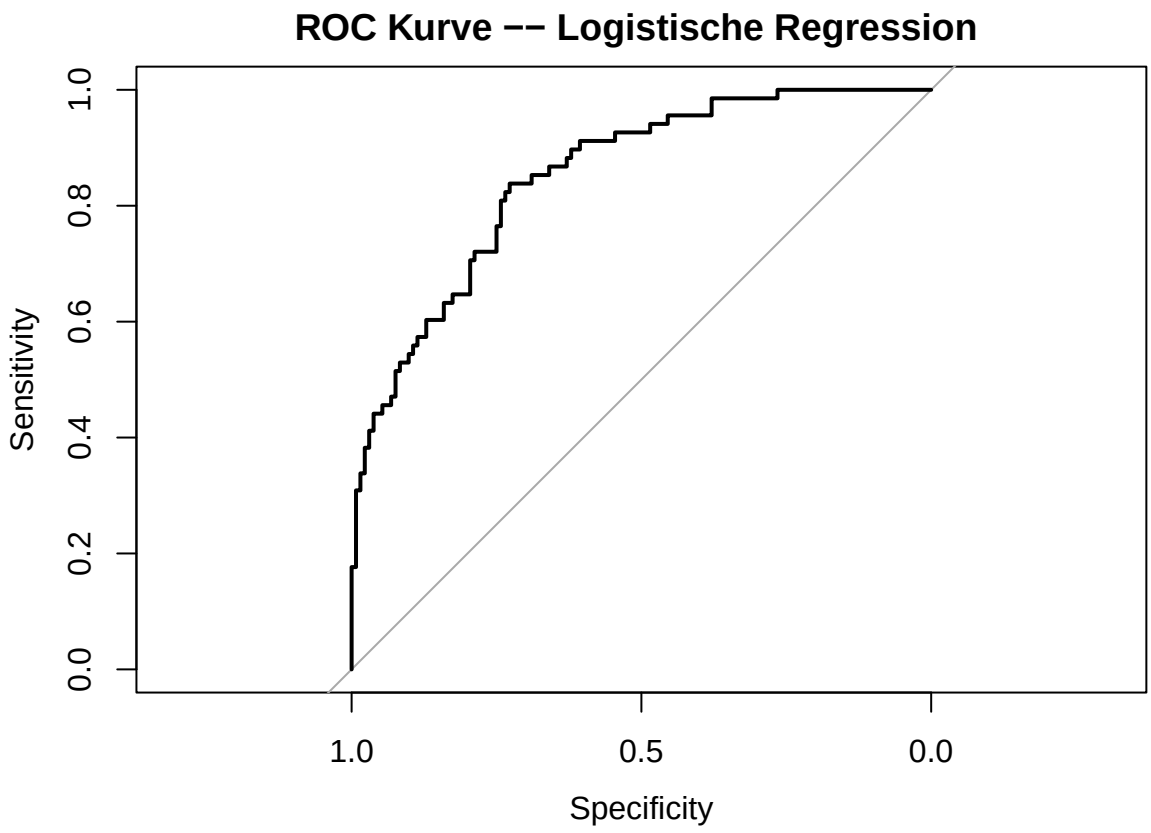
Modell 1 - alle Koeffizienten vorhanden

```
predictions <- predict(modell_glm, PrimaOriginal, type="response")
roc_curve <- roc(Pima.tr$type, predictions)
```

```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

```
plot(roc_curve,main ="ROC Kurve -- Logistische Regression ")
```




```
as.numeric(roc_curve$auc)
```

```
## [1] 0.85
```

Wir sehen hier mit einem ROC-Wert von 0.8502674 ein relativ gut angepasstes Modell.

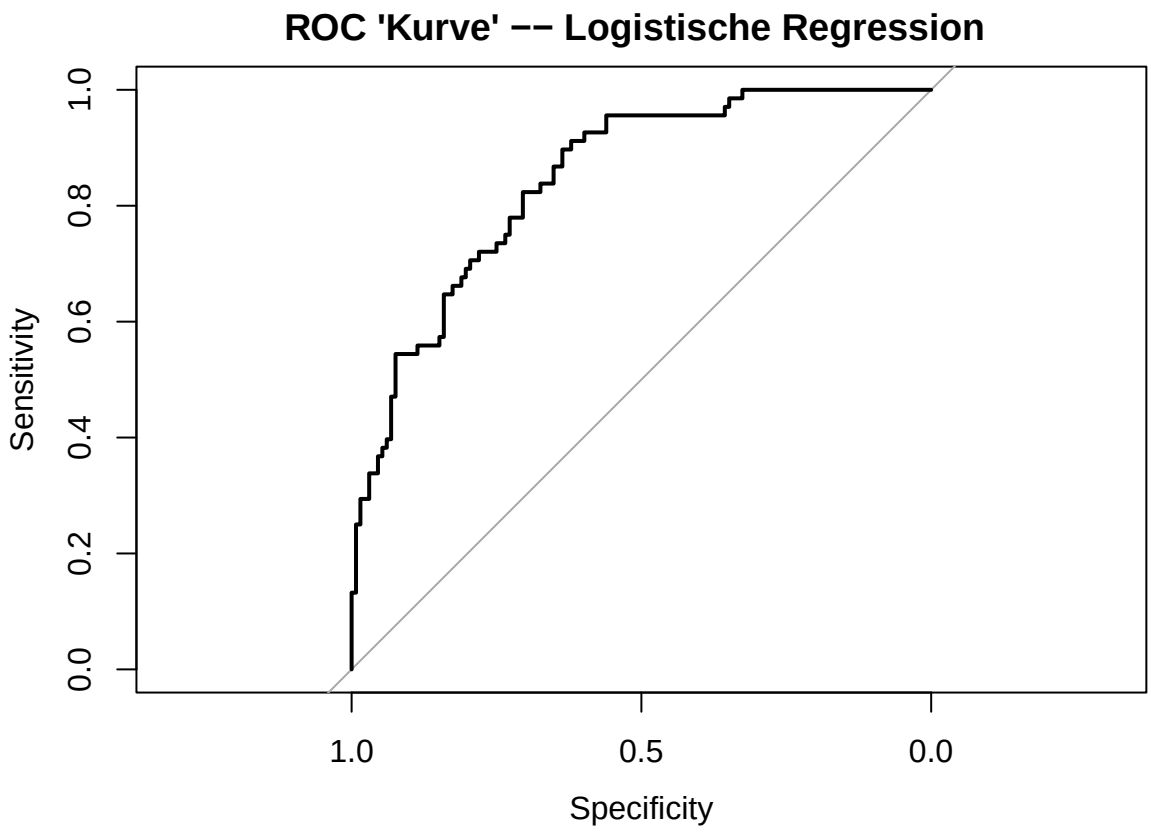
Modell 2 - Koeffizienten npreg, bp und skin entfernt.

```
modell_glm2 <- glm(type~.,data=Prima_filtered,family=binomial(link = "logit"))
predictions_filtered <- predict(modell_glm2, Prima_filtered, type="response")
roc_curve_filtered <- roc(Prima_filtered$type, predictions_filtered)
```

```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

```
plot(roc_curve_filtered,main ="ROC 'Kurve' -- Logistische Regression ")
```




```
as.numeric(roc_curve_filtered$auc)
```

```
## [1] 0.846
```

Für Modell 2 ergibt sich ein ROC-Wert von 0.846, somit hat die Entfernung der Koeffizient npreg, bp und skin, die Qualität des Modells nicht signifikant beeinflusst.

Modellgleichung

Modell 1 - alle Koeffizienten vorhanden

Die Grundgleichung lautet wie folgt

$$logit_(type_i) = \alpha + \beta_{npreg} \times x_{npreg,i} + \beta_{glu} \times x_{glu,i} + \beta_{bp} \times x_{bp,i} + \beta_{skin} \times x_{skin,i} + \beta_{bmi} \times x_{bmi,i} + \beta_{ped} \times x_{ped,i} + \beta_{age} \times x_{age,i} + \varepsilon_i$$

Das Modell im Detail daher:

$$logit_(type_i) = -9.773061533 + 0.103183427 \times x_{npreg,i} + 0.032116823 \times x_{glu,i} + -0.004767542 \times x_{bp,i} + -0.001916632 \times x_{skin,i} + 0.083623912 \times x_{bmi,i} + 1.820410367 \times x_{ped,i} + 0.041183529 \times x_{age,i} + \varepsilon_i$$

Modell 2 - nicht beitragende Koeffizienten verworfen

```
modell_glm2$coefficients
```

```
## (Intercept)         glu         bmi         ped         age
##     -9.9714      0.0313      0.0770      1.7198      0.0586
```

Die Grundgleichung lautet wie folgt

$$logit_(type_i) = \alpha + \beta_{glu} \times x_{glu,i} + \beta_{bmi} \times x_{bmi,i} + \beta_{ped} \times x_{ped,i} + \beta_{age} \times x_{age,i} + \varepsilon_i$$

Das Modell daher:

$$logit_(type_i) = -10.01125925 + 0.03143671 \times x_{glu,i} + 0.07726048 \times x_{bmi,i} + 1.72763912 \times x_{ped,i} + 0.05891378 \times x_{age,i} + \varepsilon_i$$

Erstellung der Confusion-Matrix

```
PrimaOriginal$truefalse <- ifelse(predict(modell_glm2, type = "response") > 0.5, "Yes", "No")
view(PrimaOriginal)
```

```
Prima_filtered$truefalse <- ifelse(predict(modell_glm2, type = "response") > 0.5, "Yes", "No")
```

```
model1 = table(predicted = PrimaOriginal$truefalse, actual = PrimaOriginal$type)
library(caret)
model1_con_mat = confusionMatrix(model1, positive = "Yes")
c(model1_con_mat$overall["Accuracy"],
  model1_con_mat$byClass["Sensitivity"],
  model1_con_mat$byClass["Specificity"])
model1_con_mat
```

```
library(caret)
model2 = table(predicted = Prima_filtered$truefalse, actual = Prima_filtered$type)
model2_con_mat = confusionMatrix(model2, positive = "Yes")
c(model2_con_mat$overall["Accuracy"],
  model2_con_mat$byClass["Sensitivity"],
  model2_con_mat$byClass["Specificity"])
```

```
##    Accuracy Sensitivity Specificity
##       0.770       0.559       0.879
```

```
model2_con_mat
```

```
## Confusion Matrix and Statistics
##
##          actual
## predicted  No Yes
##       No  116  30
##       Yes  16  38
##
##                Accuracy : 0.77
##                  95% CI : (0.705, 0.826)
##     No Information Rate : 0.66
##     P-Value [Acc > NIR] : 0.000477
##
##                   Kappa : 0.461
##
##  Mcnemar's Test P-Value : 0.055270
##
```

```
##             Sensitivity : 0.559
##             Specificity : 0.879
##          Pos Pred Value : 0.704
##          Neg Pred Value : 0.795
##              Prevalence : 0.340
##          Detection Rate : 0.190
##    Detection Prevalence : 0.270
##       Balanced Accuracy : 0.719
##
##        'Positive' Class : Yes
##
```

Nach dem Vergleich beider Konfusion-Matrizen stellen wir fest, dass es keinen Unterschied in der Accuracy, Sensitivity sowie Specificity gibt. Aufgrund der geringeren Komplexität des zweiten generalisierten linearen Modells (weniger Koeffizienten) entscheiden wir uns für dieses.

Bei der Konfusionsmatrix handelt es sich um ein binäres Zweiklassenmodell, welches die Verteilung der vorhergesagten und tatsächlichen Werte widerspiegelt.

In diesem Fall:

True Positives: 38 True Negatives: 116 False Positives: 16 False Negatives: 30

Die Genauigkeit umschreibt die insgesamt richtig klassifizierten Werte im Verhältnis zu allen klassifzierten. In unserem sind das 154 korrekt vorhergesasgte Werte von 200, was einer Genauigkeit von 77% entspricht.

Die Sensitivität, oder auch True-Positive-Rate, umschreibt die Fälle in denen positiv klassifizierte Datenpunkte auch tatsächlich positiv waren. Diese beträgt in unserem Fall 55,88%

Die Spezifizität, oder auch True Negative Rate, misst alle Fälle in denen negativ klassifizierte Datenpunkte auch tatsächlich negativ waren. Diese beträgt in unserem Fall 87,88%.