Hausübung 04 - Lineare Regression

Baier Sebastian, Figlmüller Magdalena, Schwarzböck Alice

13.06.2024

Contents

Allgemeine Hinweise zur Übung

Für alle Beispiele gelten folgende Aufgabenstellungen:

- Überprüfen Sie alle erforderlichen statistischen Voraussetzungen für die Gültigkeit dieses Modells mithilfe 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

1.1 Aufgabenstellung

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.

1.2 Beschreibung des Datensatzes

Es folgt eine kurze Beschreibung des Datensatzes, ehe die beiden Zielvariablen (Säuglinssterblichkeit/Infant Mortality und Bruttoinlandsprodukt/GDP) genauer untersucht und anschließend für die Erstellung eines Regressionsmodells herangezogen werden.

- **region:** umfasst die Weltregionen Afrika, Asien, Karibik, Europa, Lateinamerika, Nordamerika, Nordatlantik und Ozeanien.
- **group:** ist ein beschreibender Faktor, demzufolge Länder der OECD (Organization for Economic Co-operation and Development), Afrika oder anderen Ländern (other) zugeordnet sind.
- **fertility:** beschreibt die Fruchtbarkeitsrate als Anzahl der Kinder pro Frau.
- **ppgdp:** das per capita gross domestic product (GDP) beschreibt das Bruttoinlandsprodukt in US-Dollar.
- **lifeExpF:** beschreibt die Lebenserwartung von Frauen in Jahren.
- **pctUrban:** beschreibt den Stadtanteil in Prozent.
- **infantMortality:** beschreibt die Zahl der Todesfälle bei Säuglingen vor dem 1. Geburtstag je 1.000 Lebendgeburten.

Als erster Schritt wurde der Datensatz mittels na.omit gefiltert um Zeilen mit fehlenden Werten zu entfernen. Glimpse liefert einen Einblick in die Datenstruktur und Summary gibt eine Übersicht über Minima, Maxima, Median und die 1.+3. Quantile der beiden Zielvariablen "Infant Mortality" und "GDP":

```
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~
```

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

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

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

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

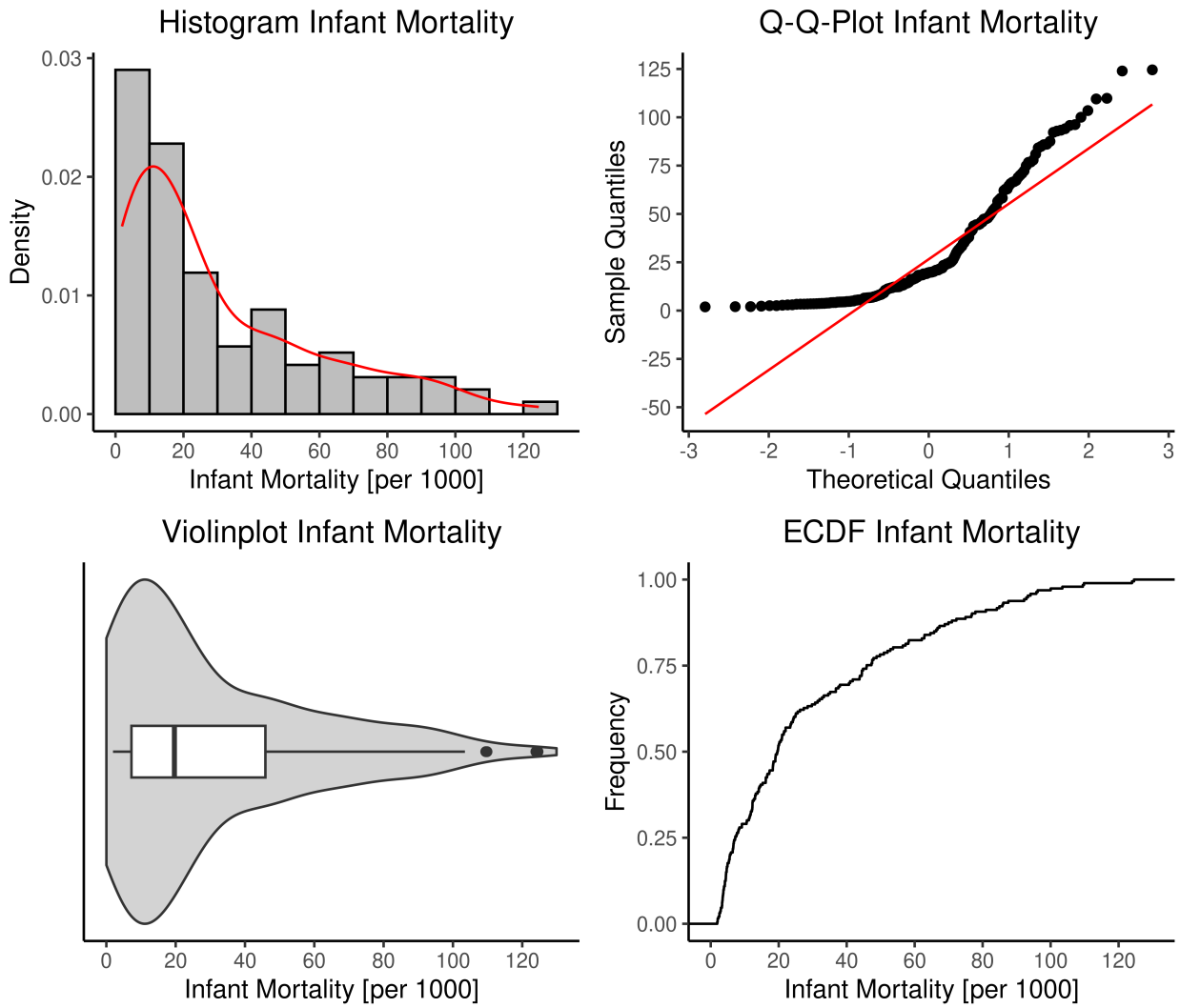
Mit na.omit wurden aus den Originaldaten (213 Zeilen) insgesamt 20 Zeilen herausgefiltert.

Hinsichtlich der Variable **Infant Mortality** fällt auf, dass der **arithmetische Mittelwert (30.739)** deutlich vom **Median (19.637)** abweicht.

Dies macht sich bei der Variable **GDP** noch deutlicher bemerkbar: Hier beträgt der **arithmetische Mittelwert** 1.229×10^4 und der **Median** liegt bei **4495.8**.

In beiden Fällen deutet der Unterschied zwischen Mittelwert und Median auf eine möglicherweise schiefe Datenlage hin, was im Anschluss mittels explorativer Datenanalyse verdeutlicht wird.

1.3 Explorative Analyse: Infant Mortality



Anhand der vier Plots lässt sich feststellen, dass es sich hinsichtlich der Variable Infant Mortality um **unimodale, rechtsschiefe Daten** handelt:

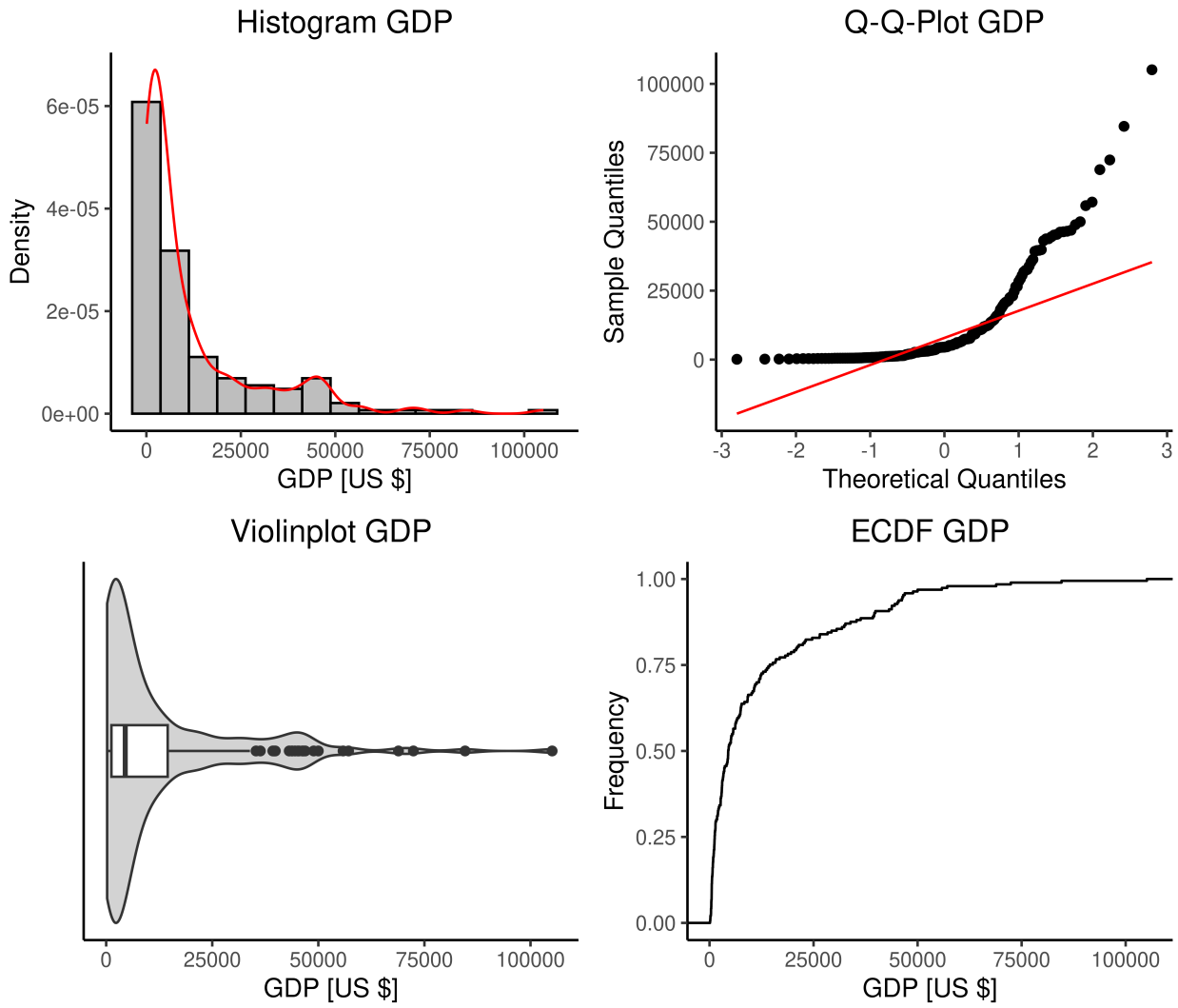
- das **Histogramm** deutet auf einen leichten (rasch ansteigenden) linken Rand und einen schweren

(langsam auslaufenden) rechten Rand hin. Dies ist plausibel, da die Daten mit 0 ein striktes unteres Limit haben, was nach oben hin nicht der Fall ist.

- der **Q-Q-Plot** bestätigt dies, da der linke Rand zur imaginären Referenzmitte hin und der rechte Rand von der Referenzmitte weg "gebogen" ist.
- der **Violin-/Boxplot** spiegelt diesen Trend ebenfalls wider. Zudem ist der im Vergleich zum Mittelwert (30.739) nach links verzerrte Median (19.637) im Boxplot deutlich sichtbar. Bei den beiden rechts außerhalb des 3. Quantils liegenden Datenpunkten handelt es sich aufgrund der schiefen Datenlage nicht zwingend um Ausreißer.
- auch der **ECDF Plot** weist mit der anfänglich starken und später abflachenden Steigung ebenfalls auf eine rechtsschiefe Datenlage hin. Eindeutige Ausreißer sind nicht erkennbar.

Die schiefe Datenlage wird auch durch die Skewness von 1.198 bestätigt.

1.4 Explorative Analyse: Gross Domestic Product (GDP)



Anhand der vier Plots lässt sich feststellen, dass es sich hinsichtlich der Variable GDP ebenfalls um unimodale, **rechtsschiefe Daten** handelt:

- das **Histogramm** deutet auch hier einen leichten linken Rand und einen schweren rechten Rand an.

Wieder haben die Daten mit 0 ein striktes unteres Limit, was nach oben hin nicht der Fall ist. Es zeigt sich, dass die Daten zwar sehr weit verteilt sind, jedoch ab etwa 60.000 US-Dollar gegen Null tendieren.

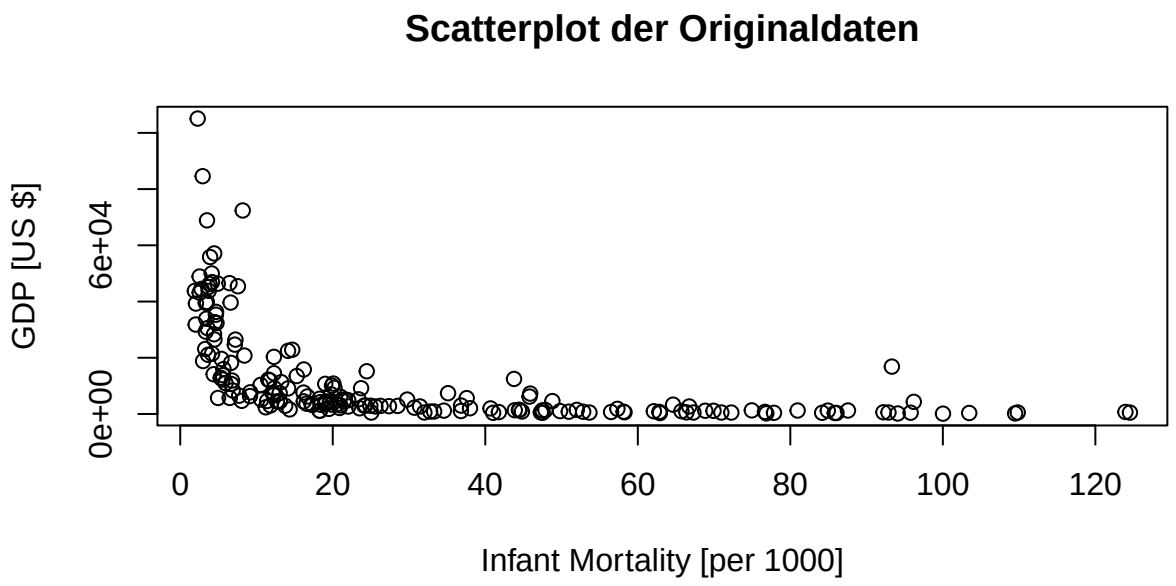
- der **Q-Q-Plot** bestätigt den leichten linken und schweren rechten Rand, da der linke Rand zur imaginären Referenzmitte hin und der rechte Rand von der Referenzmitte weg "gebogen" ist. Am rechten Rand ist ein möglicher Ausreißer erkennbar.
- der **Violin-/Boxplot** spiegelt die rechtsschiefe Datenlage ebenfalls wider. Der im Vergleich zum Mittelwert (1.229×10^4) nach links verzerrte Median (4495.8) ist auch hier im Boxplot deutlich sichtbar. Bei den außerhalb des 3. Quantils liegenden Datenpunkten handelt es sich aufgrund der schiefen Datenlage nicht zwingend um Ausreißer.
- auch der **ECDF Plot** weist mit der anfänglich starken und später abflachenden Steigung auf eine rechtsschiefe Datenlage hin. Aufgrund des langen Abstandes zwischen dem vorletzten und letzten Datenpunkt kann der letzte Datenpunkt möglicherweise als Ausreißer eingestuft werden.

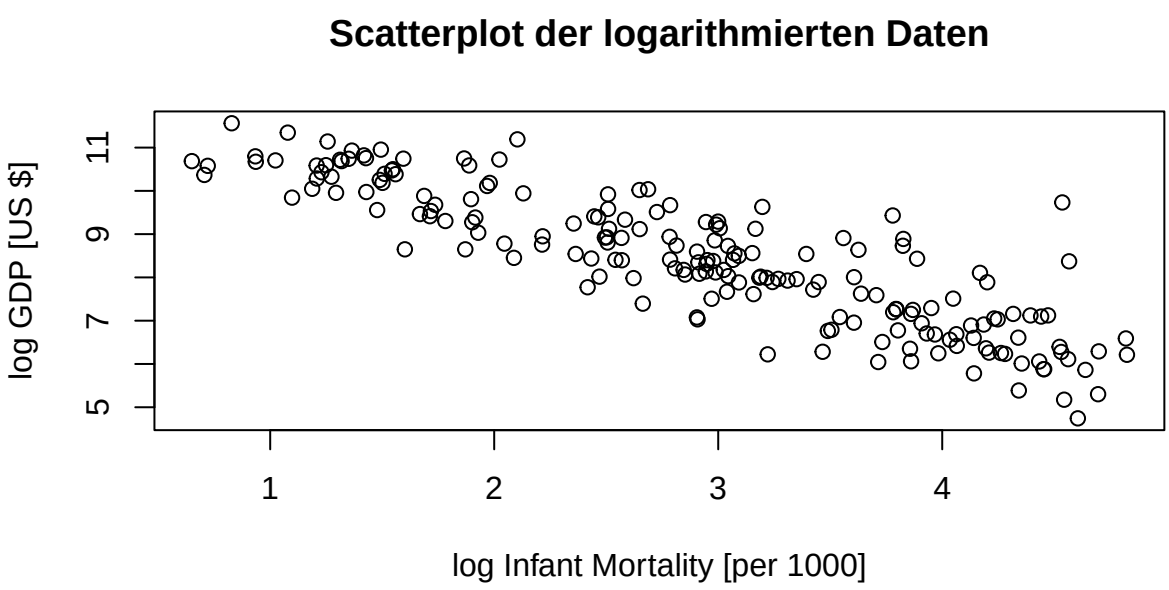
Die schiefe Datenlage wird auch durch die Skewness von 2.212 bestätigt.

1.5 Korrelations- und Lineritätscheck zwischen Infant Mortality und GDP

Datentransformation durch Logarithmieren

Die o.g. rechtschiefe Datenlage deutet darauf hin, dass die Daten vor der Modellierung durch Logarithmieren transformiert werden müssen. Um den Unterschied zu veranschaulichen, sind in Folge die beiden Variablen in ihrer originalen Form bzw. in logarithmierter Form gegeneinander geplottet:





Wie im ersten Scatterplot ersichtlich, sind die Originaldaten aufgrund der schweren Ränder der x- und y-Variablen kurvenförmig angeordnet. Bei dieser Datenverteilung ergibt es wenig Sinn, ohne vorhergehende Transformation eine Regressionsgerade durch die Datenwolke zu legen. Die Kurvenform ließ sich allerdings durch logarithmieren der beiden Variablen beheben. Die Anordnung der logarithmierten Datenpunkte deutet einen negativen linearen Zusammenhang zwischen Infant Mortality und GDP an, worauf im Anschluss mittels Pearson-Korrelation getestet wurde:

```
>
>   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 ergibt einen Wert von -0.875. Dies bestätigt obige Annahme eines Zusammenhangs zwischen niedriger Infant Mortality bei hohem GDP.

Auch der p-Wert $< 2.2e\text{-}16$ zeigt, dass die Nullhypothese (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
```

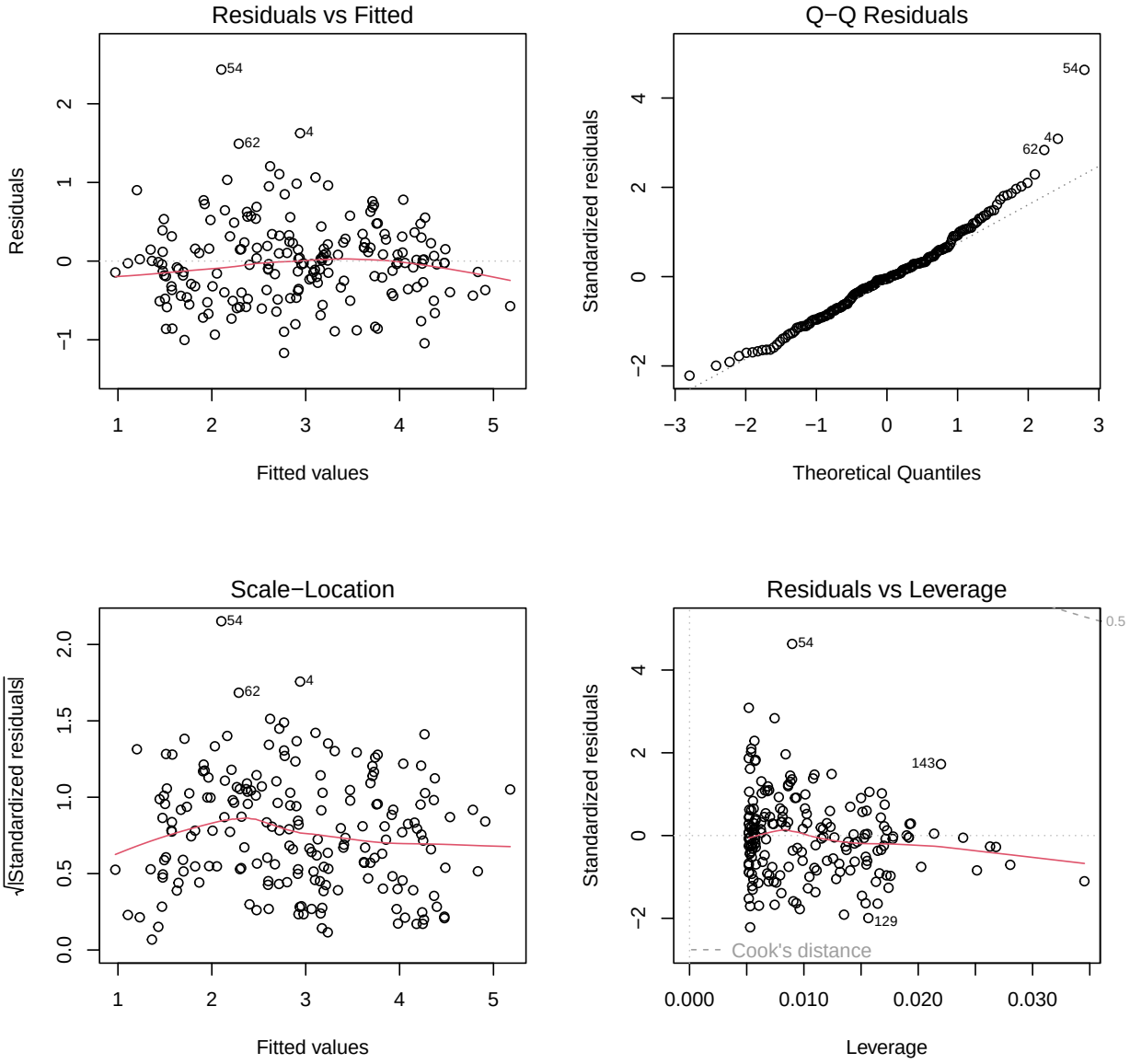
Hier gilt:

- H0: kein linearer Zusammenhang
- H1: linearer Zusammenhang

Die Summary der Residuen sagt aus, dass der Fehler mit einem Median von -0.0235 im Mittel sogut wie bei Null liegt - somit kann ein systematischer Fehler ausgeschlossen werden. Die Modellkoeffizienten a (Intercept) und b (Steigung) sind ungleich 0. Dies deutet darauf hin, dass tatsächlich ein sinnvoller linearer Zusammenhang existiert.

Die F-Statistik (625.9) und der p-Wert von $< 2.2e\text{-}16$ lassen ein Verwerfen der Nullhypothese zu. Die Korrelation (R^2) zwischen den tatsächlichen Datenpunkten und den zugehörigen Werten auf der Regressionsgeraden liegt bei 0.765, was aussagt, dass von den abhängigen Variablen ca. 76% der Varianz durch dieses lineare Modell erklärt werden können. Der Standardfehler der Residuen beträgt 0.5281 bei 191 Freiheitsgraden (=bei 193 Beobachtungen).

Überprüfung der Residuen



- **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. **Die Voraussetzungen für Regression sind daher erfüllt.**
- **Normal Q-Q:** Die Residuen liegen zum Großteil auf der Linie. Da sie sich am rechten Rand eindeutig nach oben abheben (schwerer Rand), sind die **Residuen nicht normalverteilt**. Das Regressionsmodell bleibt somit zwar gültig, jedoch darf kein t-Test durchgeführt werden. Auf der rechten Seite zeigt sich ein Ausreißer (54) - dieser sollte bei dem Datenumfang allerdings kein Problem darstellen. Der Großteil der Daten (min. 95%) liegt innerhalb des angemessenen Intervalls von [-2,2] Standardabweichungen.
- **Scale-Location:** Es sind keine systematischen Verläufe erkennbar und die Residuals wirken homoskedastisch.

- **Residuals vs Leverage:** Dieser Plot ermöglicht die Bewertung von Ausreißern bei der Regression (zweidimensionale Deklaration). Die Datenpunkte 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). Dieser Plot zeigt auch, dass der im Q-Q Plot ersichtliche Ausreißer (54) nicht entfernt werden muss, da er den Verlauf der Geraden nicht maßgeblich beeinflusst bzw. die Gerade nicht aushebelt.

Erstellung der Modell-Gleichung

- **Allgemeine Regressionsgleichung:**

$$yreg(i) = a + b * xreg(i) + \epsilon(i)$$

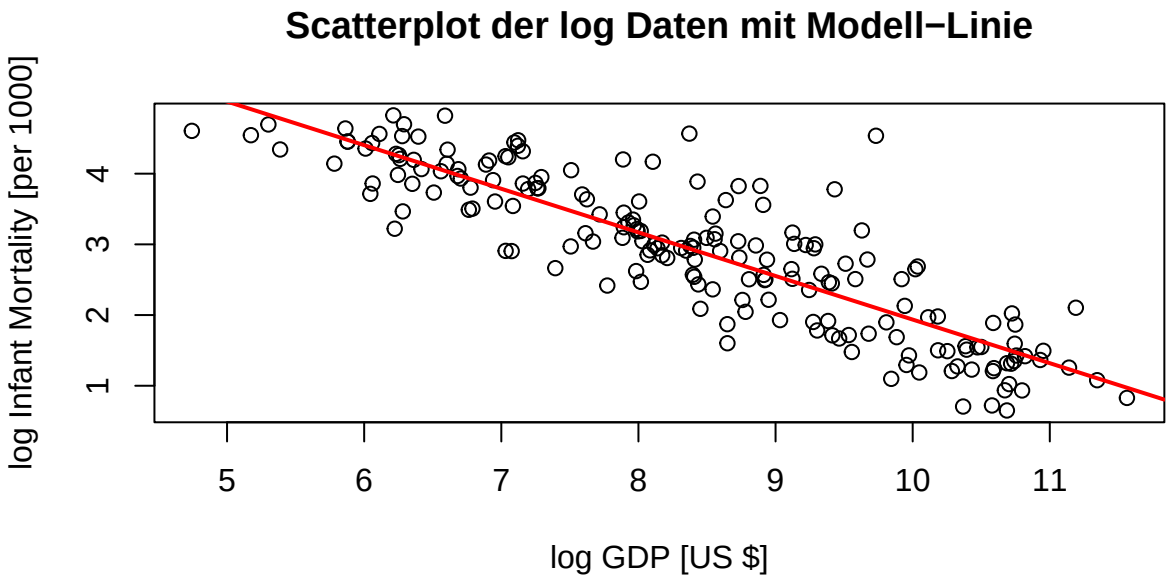
- **Modell-Gleichung:**

$$log(yreg) = 8.10377 - 0.61680 \cdot log(xreg)$$

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

* **Dabei gilt:** 8.10377 ist die Intercept und -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:

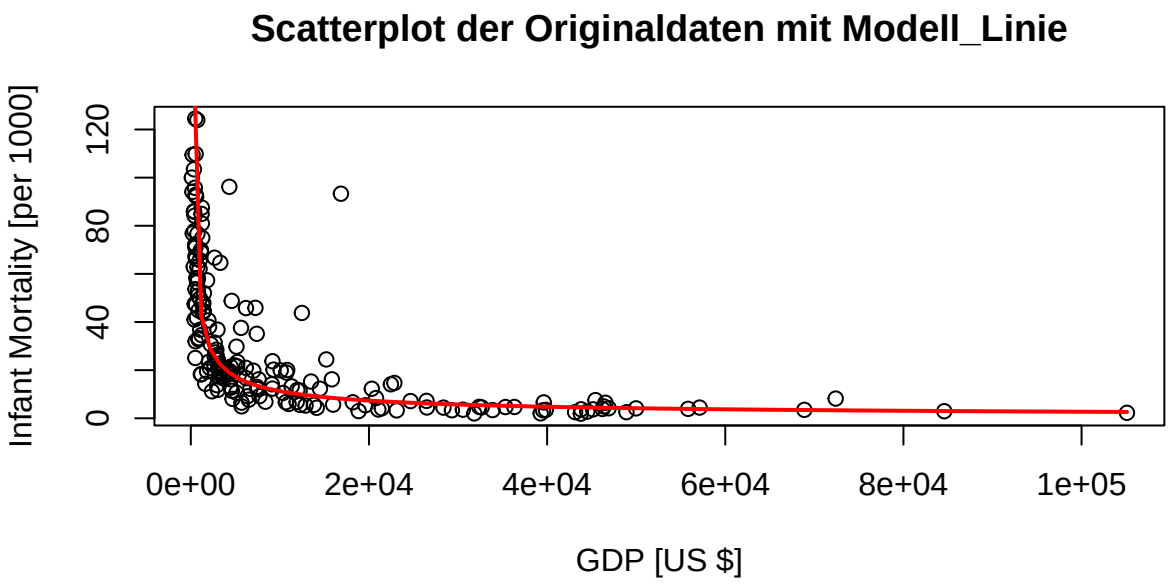
$$exp(log(infantMortality)) = exp(-0.61680 * log(ppgdp) + 8.10377)$$

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

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, das Bruttoinlandsprodukt (y-Achse) unendlich gross wäre. Zudem bedeutet eine Erhöhung der Säuglingssterblichkeitsrate um 1, dass das Bruttoinlandsprodukt um das 3306.912-fache sinkt.



Aufgabe 2: Schweiz

2.1 Aufgabenstellung

- 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 ...
```

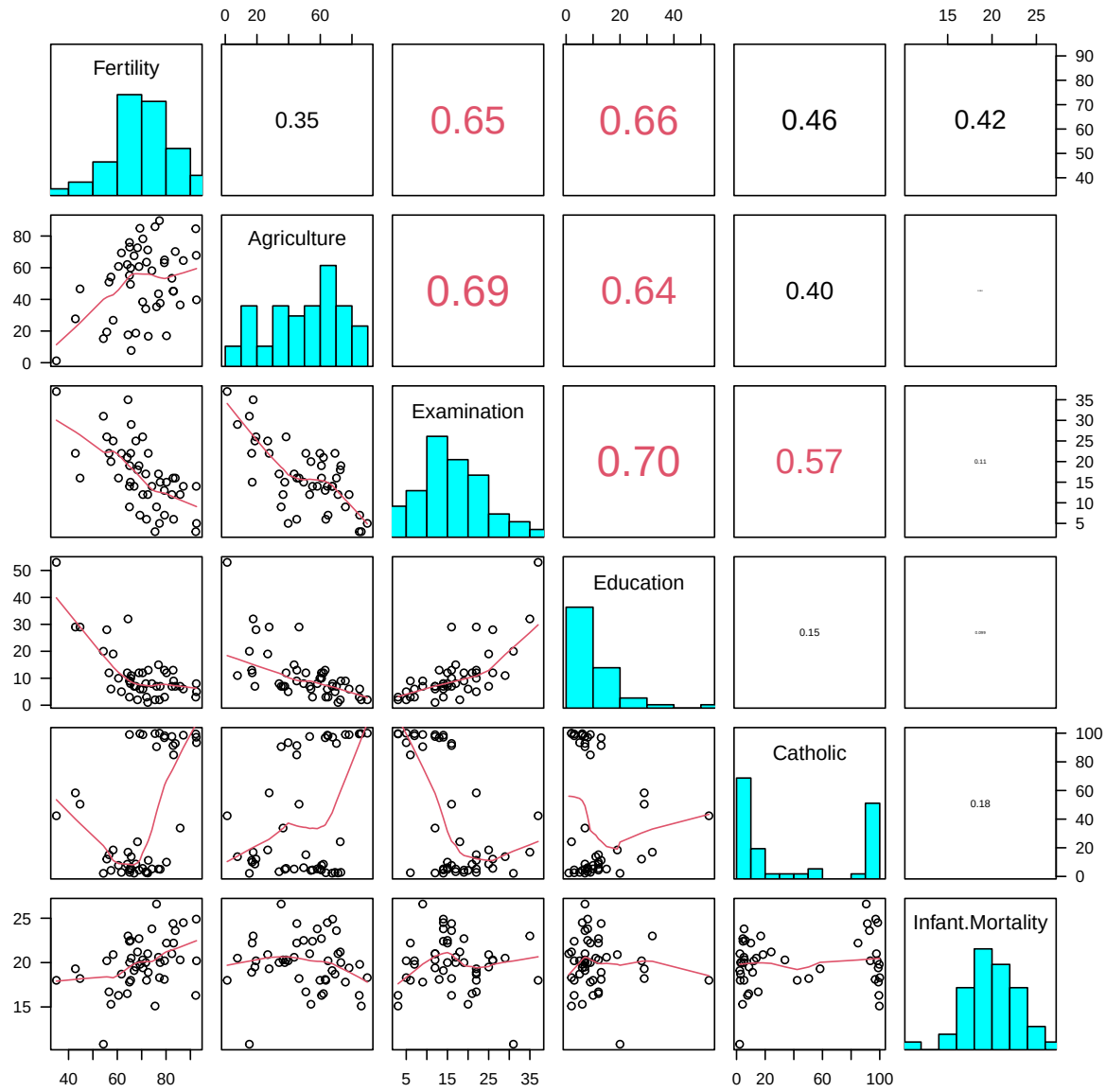
```
summary(swiss)
```

```
##    Fertility     Agriculture    Examination     Education     Catholic
##  Min.   :35.0   Min.   : 1.2   Min.   : 3.0   Min.   : 1   Min.   :  2.1
##  1st Qu.:64.7   1st Qu.:35.9   1st Qu.:12.0   1st Qu.: 6   1st Qu.:  5.2
##  Median :70.4   Median :54.1   Median :16.0   Median : 8   Median : 15.1
##  Mean   :70.1   Mean   :50.7   Mean   :16.5   Mean   :11   Mean   : 41.1
##  3rd Qu.:78.4   3rd Qu.:67.7   3rd Qu.:22.0   3rd Qu.:12   3rd Qu.: 93.1
##  Max.   :92.5   Max.   :89.7   Max.   :37.0   Max.   :53   Max.   :100.0
##  Infant.Mortality
##  Min.   :10.8
##  1st Qu.:18.1
##  Median :20.0
##  Mean   :19.9
##  3rd Qu.:21.7
##  Max.   :26.6
```

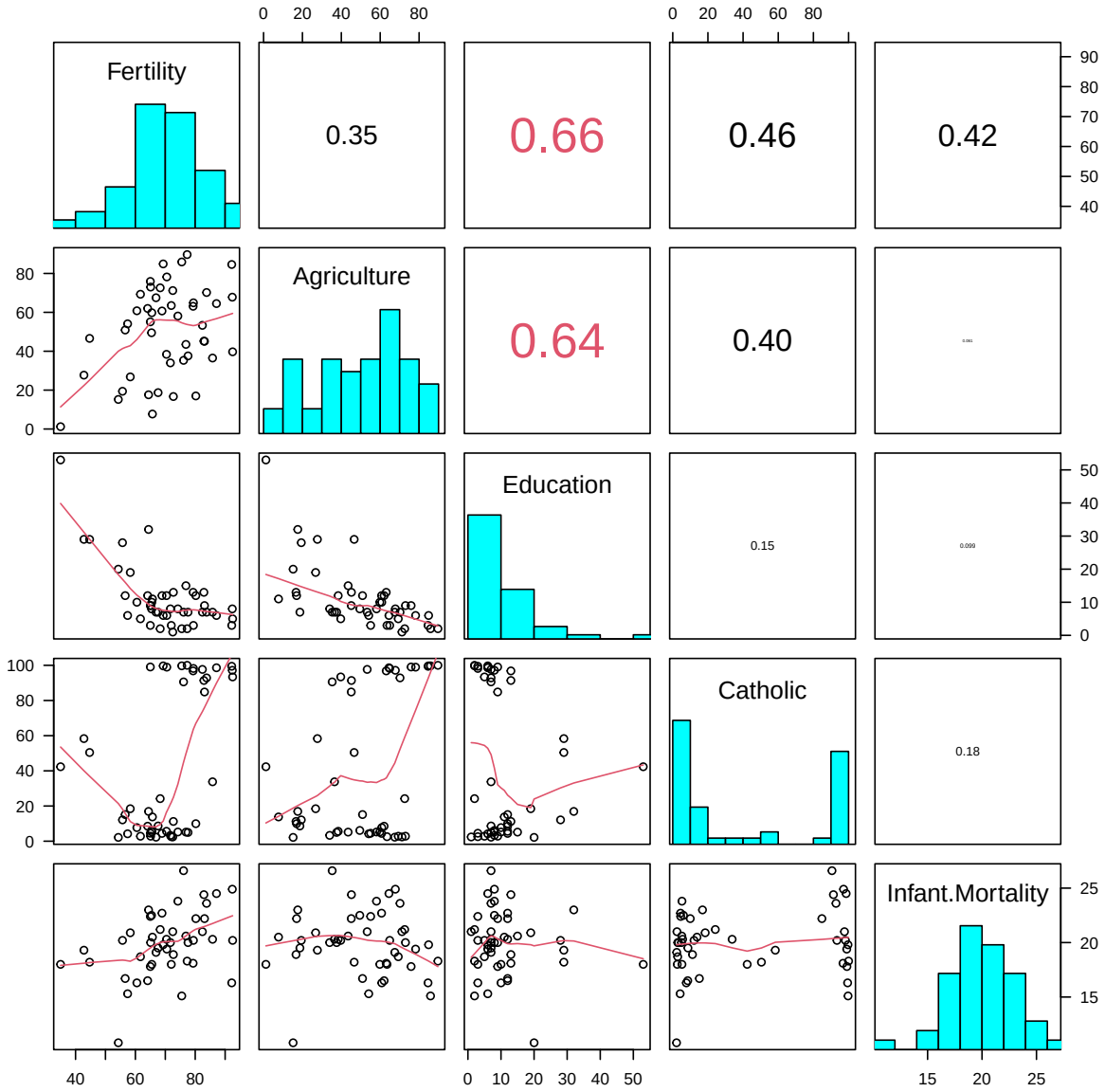
In dieser Aufgabenstellung geht es darum, das optimale lineare Modell zu finden um die Variable Education durch die übrigen Variablen zu beschreiben. Hierfür müssen allerdings zuerst die statistischen Voraussetzungen überprüft bzw. evaluiert werden.

2.2 Überprüfung der statistischen Voraussetzungen

Überprüfung auf Unabhängigkeit der Kovariablen



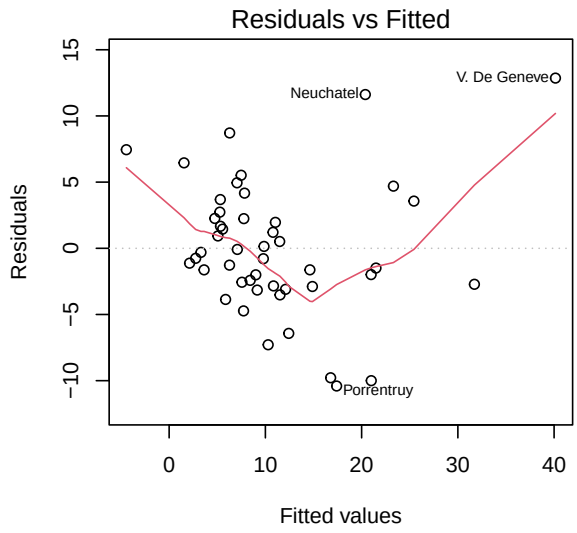
Aus dem Scatterplot sehen wir, dass die Variable **Examination** mit fast allen Variablen hoch (R >= 0.5) korreliert. Da dies für eine Modellbildung als erklärende Variable nicht zulässig ist, wird diese aus dem Datensatz entfernt und die übrigen Daten via Scatterplot nochmalig geprüft.

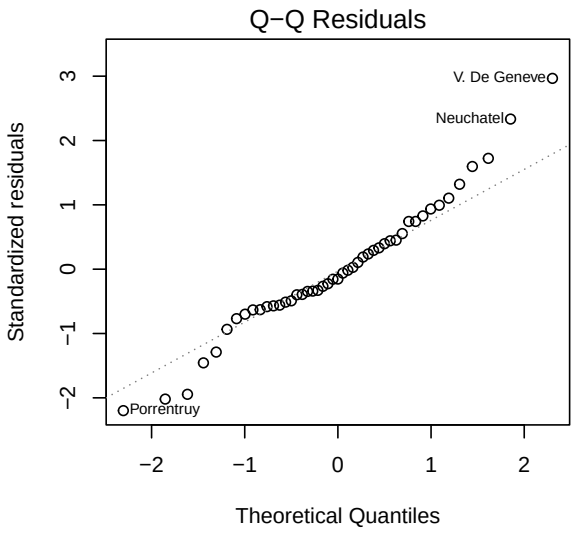


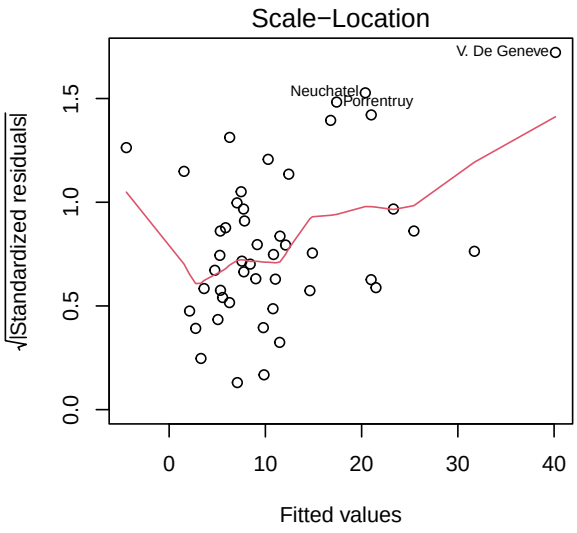
Zwischen den erklärenden Variablen (ausgenommen Education, welche erklärt werden soll) ist nach Entfernung der Variable Examination keine Korrelation mehr erkennbar. Da die Grundvoraussetzung, dass Kovariablen nicht miteinander korrelieren dürfen ($R < 0.5$) nun erfüllt ist, kann mit diesem Datensatz weitergearbeitet werden.

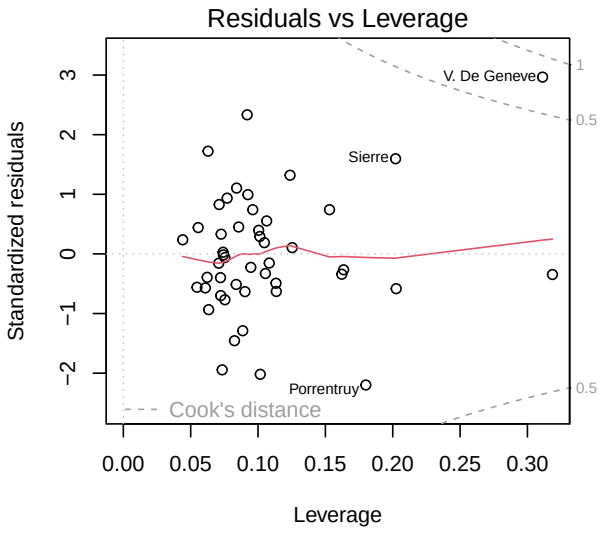
Überprüfung der Residuen

```
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
```

Anhand der oben angeführten Summary und Plots kann überprüft werden, ob nachfolgende Anforderungen für eine multiple lineare Regression erfüllt sind:

- Das Modell besitzt keinen systematischen Fehler: **Ja**, die Lage des Medians (-0.7571), welcher Nähe Null liegt, deutet darauf hin, dass die Residuen um Null herum zentriert sind. Diese Bedingung scheint daher erfüllt.
- Die Fehlervarianz ist für alle Beobachtungen in etwa gleich groß (homoskedastisch): **Ja**, der Residual vs. Fitted und der Scale-Location Plot deuten dies an, wobei die Datenpunkte nicht so gleichmäßig um die 0-Linie angesiedelt sind wie das bei Beispiel 1 der Fall war.
- Die Residuen sind unabhängig, bzw. gleichartig verteilt: **Ja**, im Residual vs. Fitted Plot lässt sich kein eindeutiger Trend in der Verteilung erkennen.
- Die Residuen sind normalverteilt: **Nein**, im Q-Q Plot ist erkennbar, dass die Residuen schwere Ränder aufweisen und somit streng genommen **nicht normalverteilt** sind. Der Großteil der Werte liegt jedoch auf der Geraden, weshalb möglicherweise dennoch einige Tests für Normalverteilung in Frage kommen (?)
- Es gibt keine lineare Abhängigkeit zwischen den Regressoren: **Ja**, dies wurde anhand der Korrelations-Scatterplots und der Korrelationskoeffizienten (< 0.5) aufgezeigt.

Es sind daher alle notwendigen Bedingungen für das Erstellen eines Regressionsmodells erfüllt.

Im **Residuals vs. Leverage Plot**, der Ausreißer bei Regressionsmodellen aufzeigt, 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 möglicherweise entfernt werden muss.

2.3 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:

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

2.4 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.

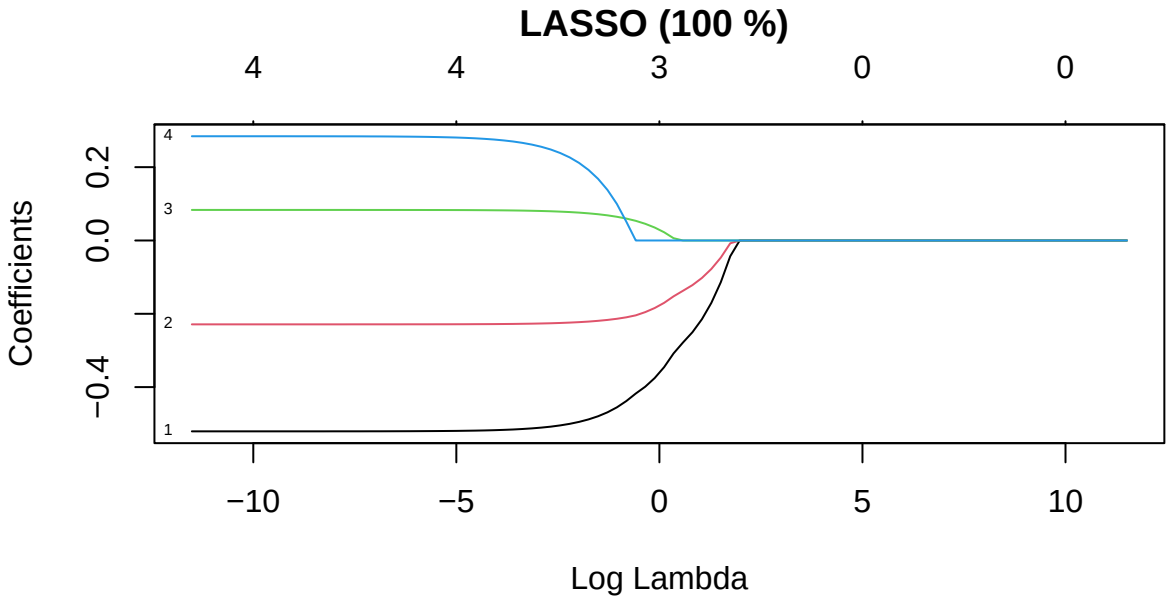
Das bedeutet für 1% mehr in der "common standardized fertility measure" sind es 0.48883% weniger, für 1% mehr Männer in der Landwirtschaft 0.23799% weniger und für 1% mehr Katholiken sind es 0.08440% mehr Menschen mit einer höheren Ausbildung als Grundschulausbldung.

2.5 Erweiterung: LASSO Regression

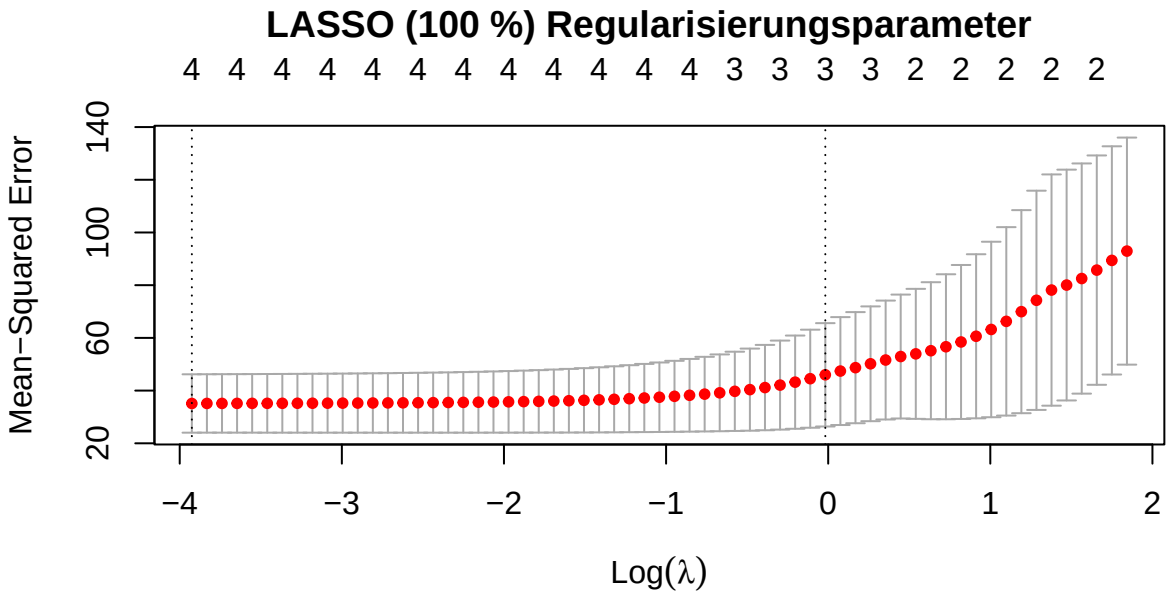
LASSO-Regression (Least Absolute Shrinkage and Selection Operator) ist eine Methode der linearen Regression, die durch L1-Regularisierung eine Strafe für die Summe der absoluten Werte der Regressionskoeffizienten hinzufügt, wodurch einige Koeffizienten auf Null gesetzt werden können. Das führt zu einer Schrumpfung der Koeffizienten und ermöglicht die Auswahl der wichtigsten Variablen, wodurch das Modell vereinfacht und die Überanpassung reduziert werden soll.

Diese Methode wurde verwendet, um die Vorhersage-Genauigkeit des Modells zu erhöhen bzw. zu automatisieren. Um die Datenvoraussetzung zu erfüllen, wurde vorab die mit anderen Kovariablen korrelierende Examination-Variable entfernt.

```
library(glmnet)
library(caret)

data <- swiss %>% select(-Examination)

# Unabhängige Kovariablen
x <- as.matrix(swiss[, c("Fertility", "Agriculture", "Catholic", "Infant.Mortality")])

# Zielvariable
y <- swiss$Education

# Plot 1: Koeffizienten aller inkludierten Kovariablen vs. Log Lambda
lambda.grid <- 10^seq(-5,5, length=100)
fitL <- glmnet(x=x, y=y, alpha=1, lambda= lambda.grid)
plot(fitL,xvar="lambda",label=TRUE, main = "LASSO (100 %) \n")
```




```
# LASSO mit Kreuzvalidierung
lasso_model <- cv.glmnet(x, y, alpha = 1)

# Plot 2: Kreuzvalidierungsfehler
plot(lasso_model, main = "LASSO (100 %) Regularisierungsparameter  \n")
```




```
# Bestes Lambda
best_lambda <- lasso_model$lambda.min
best_lambda
```

```
> [1] 0.0197
```

```
# Endgültiges LASSO-Modell
final_lasso_model <- glmnet(x, y, alpha = 1, lambda = best_lambda)

# Koeffizienten des endgültigen Modells
coef(final_lasso_model)
```

```
> 5 x 1 sparse Matrix of class "dgCMatrix"
>                       s0
> (Intercept)      49.9422
> Fertility        -0.5170
> Agriculture      -0.2279
> Catholic          0.0823
> Infant.Mortality  0.2740
```

Interpretation der Plots

Die **erste Grafik (Koeffizienten vs. Log Lambda)** zeigt, dass die den verschiedenen Kovariablen zugehörigen Koeffizienten an unterschiedlichen Stellen rapide nach Null abfallen oder aufsteigen. Wenn die Null-Linie erreicht ist, wird die Variable aus dem Modell ausgeschlossen.

Die x-Achse der **zweiten Grafik (Plot des cv.glmnet-Objekts)** zeigt die Werte des Regularisierungsparameters log(lambda), der die Stärke der Regularisierung steuert. Ein höherer Wert von lambda führt zu einer stärkeren Regularisierung, wodurch mehr Koeffizienten auf null gesetzt werden. Wenn ein Koeffizient auf null gesetzt wird, heißt das er wird als nicht signifikant befunden und aus dem Modell ausgeschlossen.

"cv" steht hier für Kreuzvalidierung, was in Kombination mit LASSO, Elastic Net oder Ridge dazu verwendet wird, den optimalen Wert für den Regularisierungsparameter lambda zu finden. Der Punkt auf der x-Achse, der den niedrigsten mittleren Kreuzvalidierungsfehler (= Mean-Squared Error) hat, entspricht dem besten lambda-Wert, hier: 0.0197. Dies ist der Wert, der das Modell mit der besten Vorhersagegenauigkeit auf den Validierungsdatensatz liefert. Im Zuge der LASSO-Analyse wurden keine der Kovariablen auf null gesetzt, da diese vom Modell als signifikant identifiziert wurden.

Interpretation der Koeffizienten

Die resultierenden Koeffizienten des Modells können wie folgt interpretiert werden:

- die **Intercept (49.9422)** ist der geschätzte Wert von "Education", wenn alle anderen unabhängigen Variablen (Fertility, Agriculture, Catholic, Infant Mortality) gleich null sind. In diesem Kontext ist es der Basiswert der Bildung, wenn keine der erklärenden Variablen vorhanden ist.
- **Fertility** und **Agriculture** haben negative Koeffizienten, was darauf hindeutet, dass höhere Werte dieser Variablen mit niedrigeren Bildungsniveaus verbunden sind.
- **Catholic** und **Infant Mortality** haben positive Koeffizienten, was bedeutet, dass höhere Werte dieser Variablen mit höheren Bildungsniveaus verbunden sind.

Diese Trends entsprechen dem unter Punkt 2.3 genannte Modell #1 - zusammenfassend trägt LASSO daher hier keine nennenswerten Mehrwert zur Analyse bei.

Die angepasste LASSO-Modellgleichung lautet:

$$Education_i = 49.9422 - 0.5170 \times x_{Fertility,i} + -0.2279 \times x_{Agriculture,i} + 0.0823 \times x_{Catholic,i} + 0.2740 \times x_{Infant.Mortality,i}$$

Zuletzt wurde noch überprüft, wie sich die Koeffizienten verändern, wenn statt des Lambda-Minimalwerts ein strengerer Lambda-Wert gewählt wird - nämlich der größtmögliche Lambda-Wert, bei dem der Fehler

noch innerhalb eines Standardfehlers des Minimalwerts liegt. Dieser Wert (hier: 0.982) und der Minimalwert für Lambda sind im LASSO-Plot oberhalb durch zwei strichlierte Linien markiert.

Wie nachfolgende Koeffizientenausgabe zeigt, wird bei Verwendung des größeren Lambda-Werts die Variable **Infant Mortality auf 0 gesetzt** und fällt dadurch aus der Modellierung heraus. Die Koeffizienten der übrigen Variablen (Fertility, Agriculture, Catholic) sind etwas kleiner, zeigen aber dieselben Trends wie zuvor.

```
# Wahl des zweiten Lambda-Schwellenwerts (strenger)
sbest_lambda <- lasso_model$lambda.1se
sbest_lambda
```

```
> [1] 0.982
```

```
# Strengeres LASSO-Modell
final_lasso_model2 <- glmnet(x, y, alpha = 1, lambda = sbest_lambda)

# Koeffizienten des strengeren Modells
coef(final_lasso_model2)
```

```
> 5 x 1 sparse Matrix of class "dgCMatrix"
>                       s0
> (Intercept)      44.2885
> Fertility        -0.3636
> Agriculture      -0.1787
> Catholic          0.0302
> Infant.Mortality  .
```

Aufgabe 3: USA

3.1 Aufgabenstellung

- 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.

3.2 Beschreibung des Datensatzes

- Population (Schätzung der Bevölkerungsanzahl zum 1. Juli 1975)
- Income (Pro-Kopf-Einkommen in 1974, in $)
- Illiteracy (Analphabetismus in 1970, in %)
- Life Exp (Lebenserwartung in Jahren von 1969-71)
- Murder (Mord und nicht-fahrlässige Tötungsrate pro 100.000 Einwohner in 1976)
- HS Grad (Personen mit High School Abschluss in 1970, in %)
- Frost (durchschnittliche Anzahl von Tagen an denen die Mindesttemperatur von 1931-1960 in einer Haupt- oder Großstadt unter dem Gefrierpunkt lag)
- Area (Landfläche in Quadratmeilen, mi²; nicht Teil der Aufgabenstellung)

```
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" ...
```

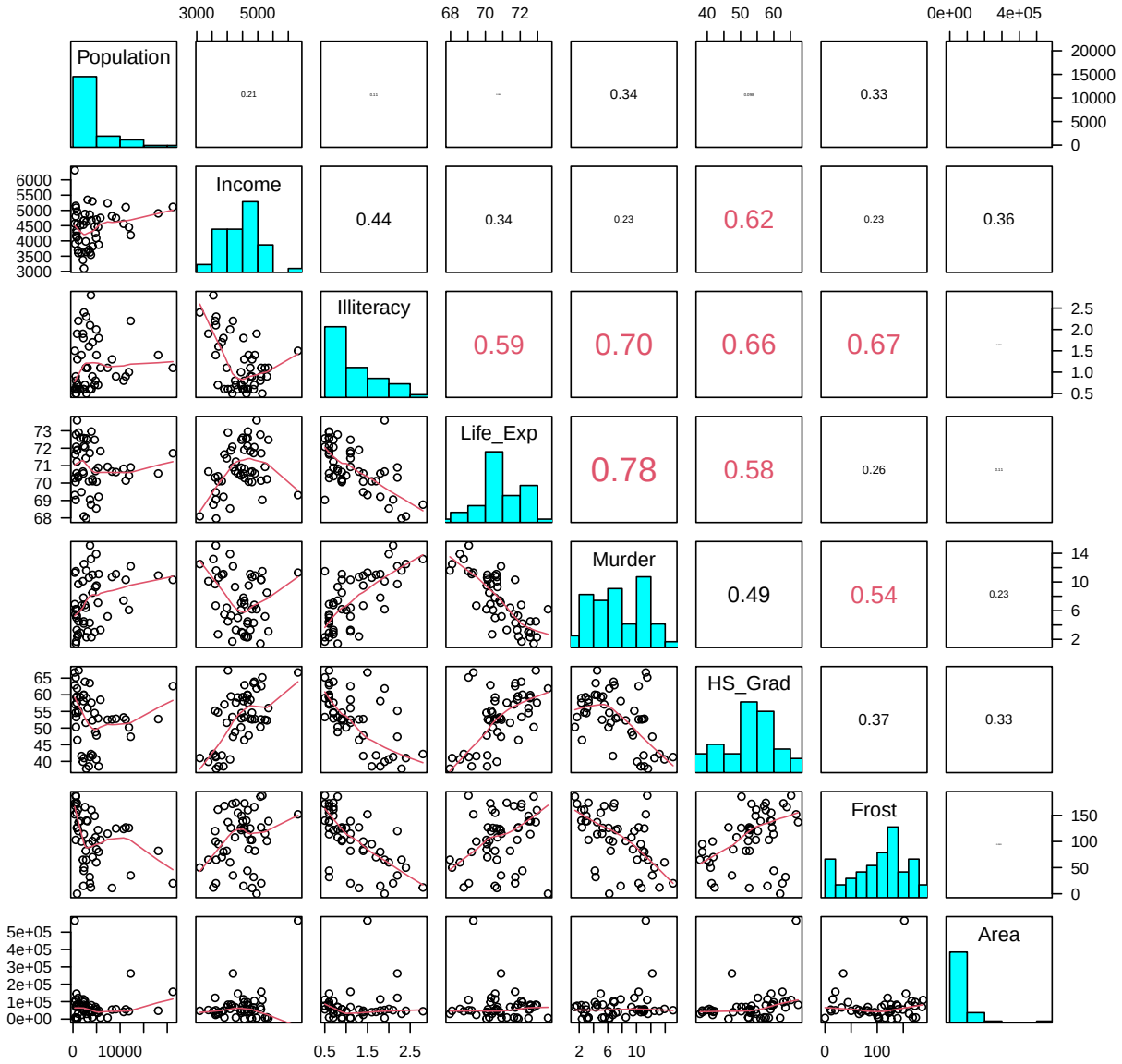
```
class(state.x77)
```

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

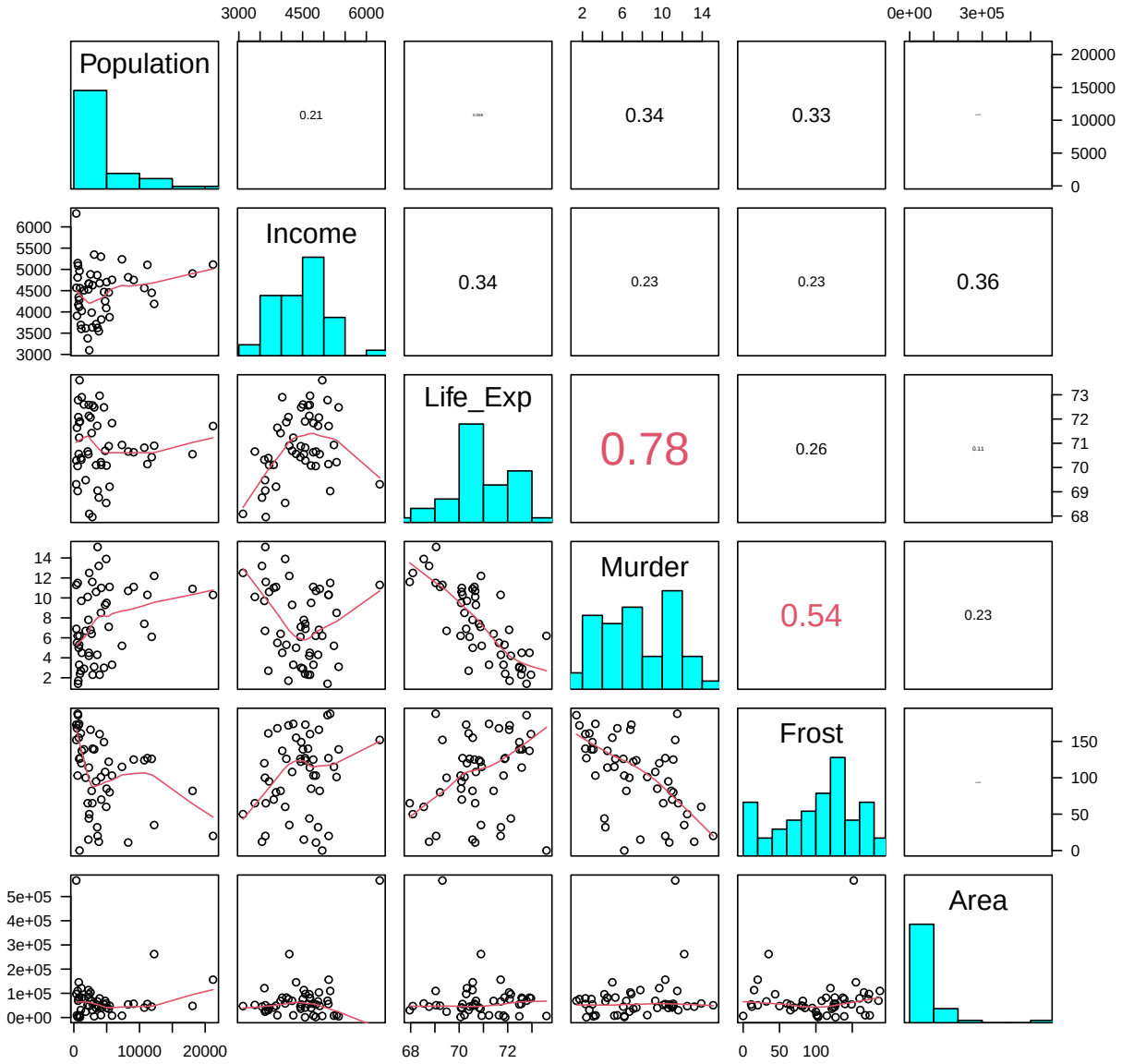
3.3 Scatterplots

Zunächst wurden Scatterplots erstellt, um einen Überblick darüber zu erhalten, welche Variablen mit "Murder" in einem linearen Zusammenhang stehen könnten:

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



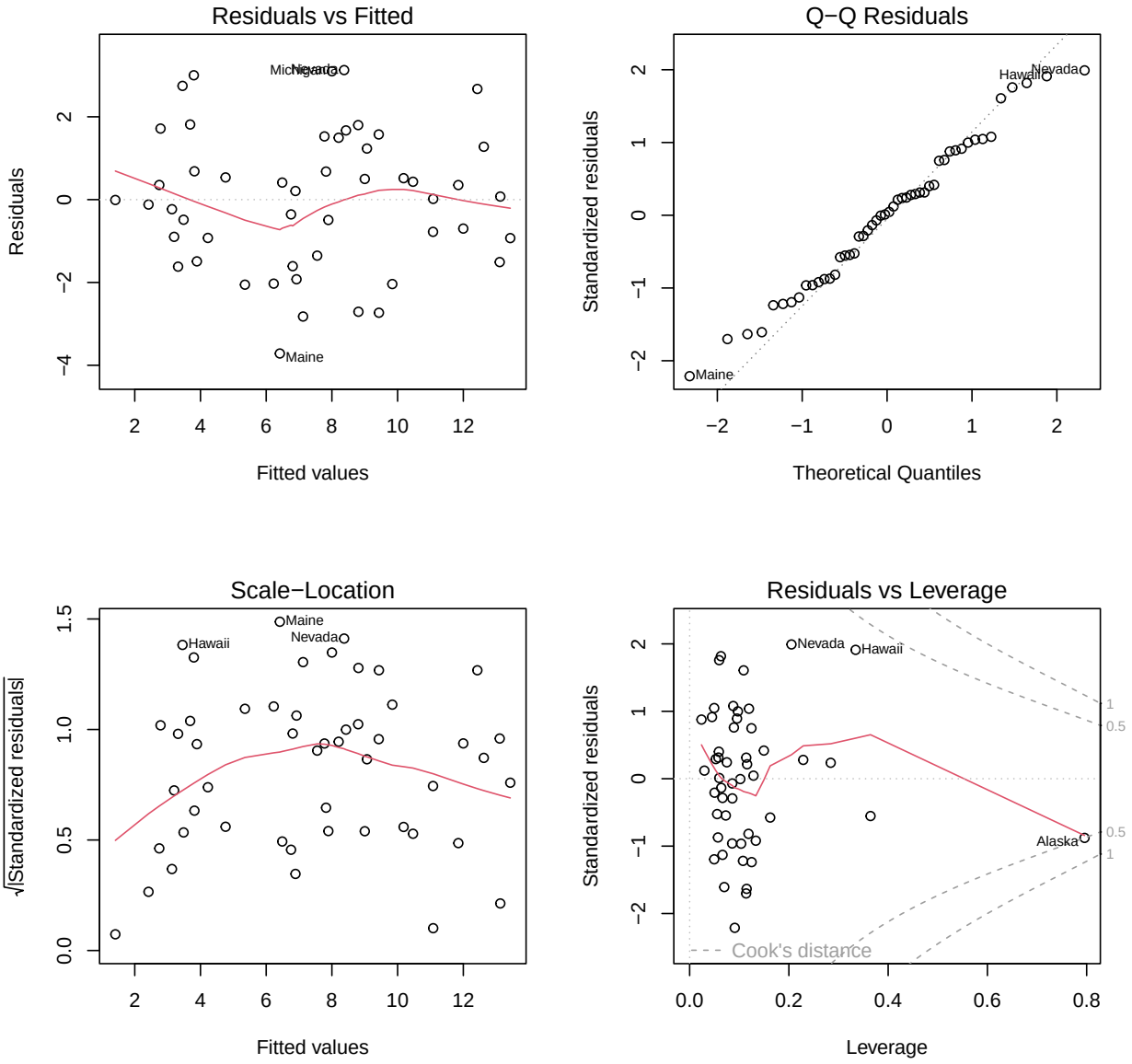
Da wir sehen, dass die Variable “HS_Grad” mit nahezu allen anderen Variablen sehr gut korreliert, ist diese als erklärende Variable auszuschließen. Gleiches gilt für “Illiteracy”.



Nun sind nurmehr Variablen übrig, welche als erklärende Variablen nicht ausreichend (R >= 0.5) miteinander korrelieren.

3.4 Überprüfung der statistischen Voraussetzungen

Residuen und Summary




```
>
> 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
```

In unserem ersten erstellten Modell sehen wir, dass nicht alle Variablen zur Erklärung herangezogen werden können. Income spielt mit einem p-Wert von 0.530396 keine signifikante Rolle. Die Area spielt im Vergleich zu den übrigen Variablen auch nur eine untergeordnete Rolle, wurde aber mittels t-Test (dem aufgrund der normalverteilten Residuen lt. Q-Q Plot vertraut werden darf) mit einem p-Wert von 0.01584 als schwach signifikant bewertet.

- Im Plot **Residuals vs. Fitted** ist eine gleichmäßig verteilte Punktewolke zu sehen, welche 1) auf keinen systematischen Fehler hinweist (gemeinsam mit dem Median, der nahezu bei 0 liegt), 2) zeigt, dass die Residuen homoskedastisch sind und 3) auch keine Korrelationen erkennen lässt. Die Bedingungen für ein lineares Regressionsmodell sind somit erfüllt.
- Im Plot **Residuals vs. Leverage** sehen wir eine Beobachtung "Alaska", welche exakt auf der Cooks Distance liegt und daher vermutlich eine große Hebelwirkung erzielt. Diese ist daher als Ausreißer einzustufen und zu entfernen. Es existieren noch zwei weitere Beobachtungen "Nevada" und "Hawaii", welche eine relativ große Hebelwirkung aufweisen und nahe an der Cooks Distance liegen - jedoch noch innerhalb des akzeptablen Bereichs, weshalb sie nicht entfernt wurden.
- Der Plot **Scale Location** weist eine leichte Kurve auf, vermutlich verursacht durch die Variablen Hawaii, Maine und Nevada. Diese erscheint jedoch akzeptabel.
- Der **Q-Q Residuals** Plot zeigt, dass die Daten zum größten Teil an oder sehr Nahe bei der Geraden liegen. Die Residuen können somit als normalverteilt bewertet werden. Das Durchführen von t-Tests ist daher hier grundsätzlich erlaubt.

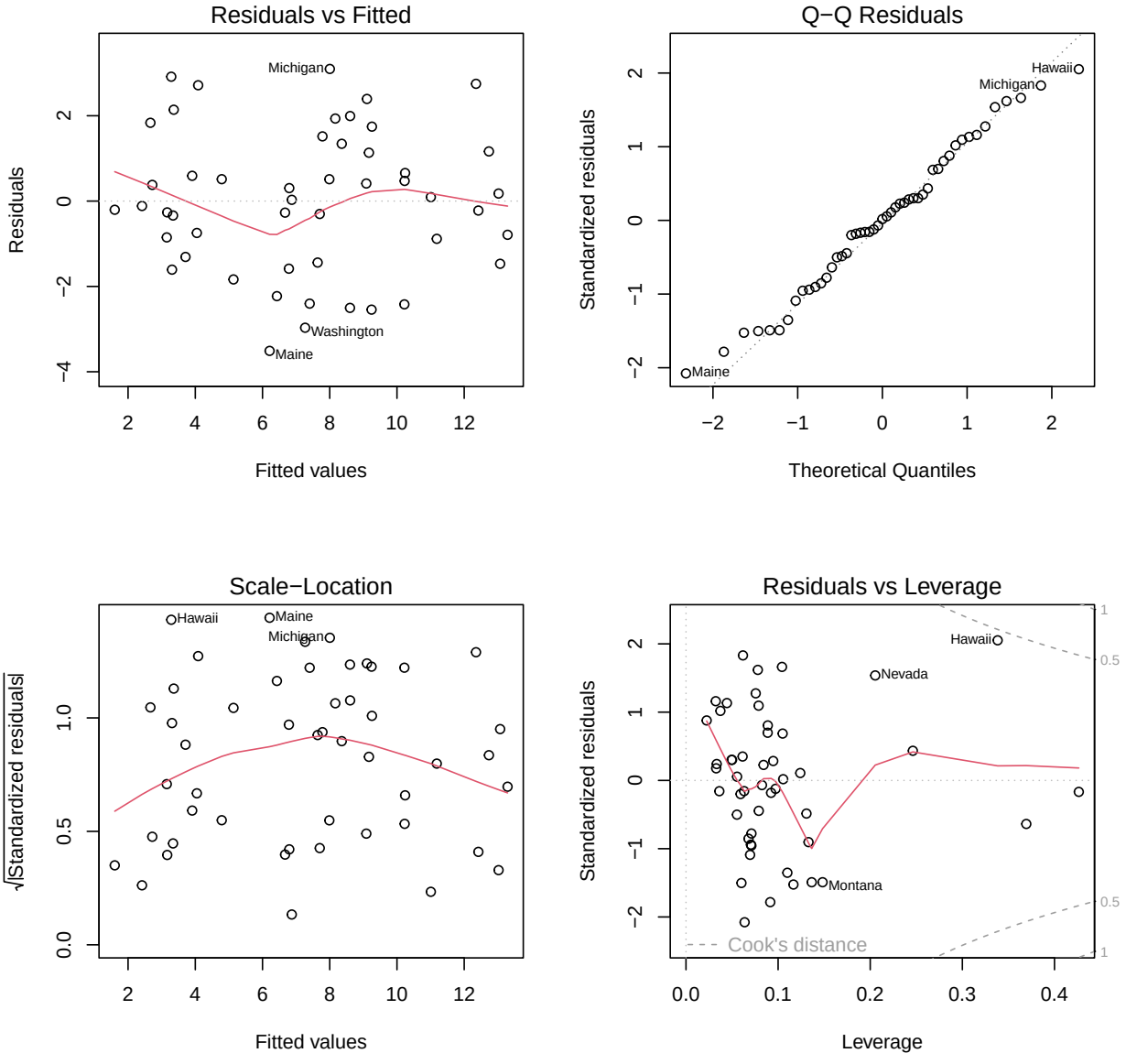
Im folgenden Schritt wurden die oben genannte Variable Income (als nicht ausreichend erklärende Variable) sowie die Beobachtung "Alaska" als Ausreißer entfernt:

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

```
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
```



```
##            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 nurmehr 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 erklärenden Koariablen keine Zusammenhänge / Korrelationen untereinander ($<$ 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 mit den errechneten Koeffizienten angepasste Modellgleichung lautet demnach:

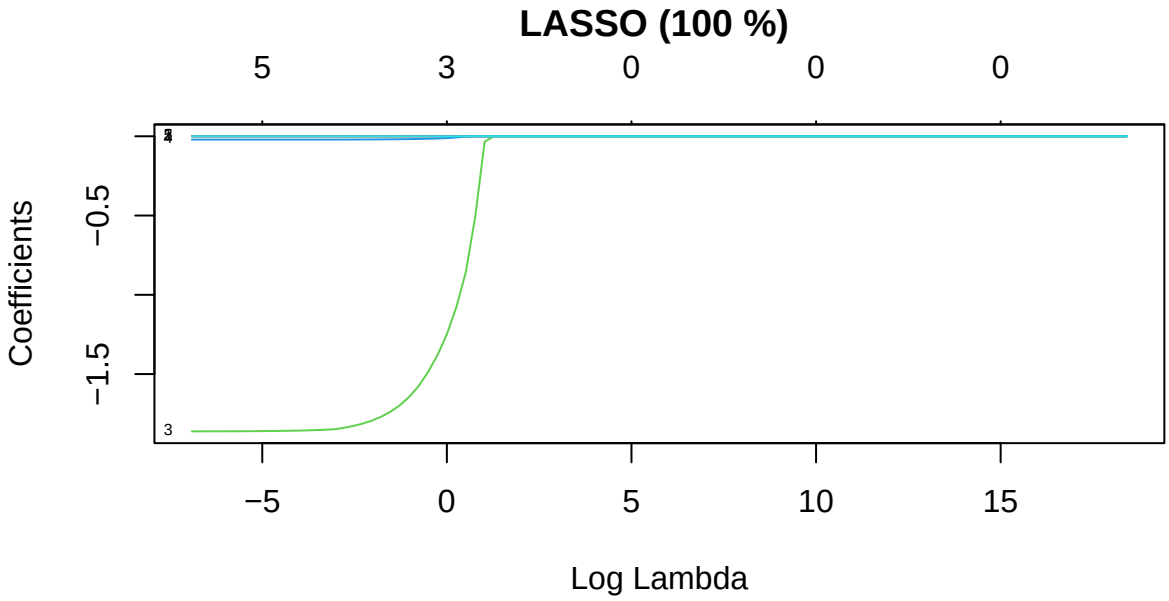
$$Murder_i = (1.414e{+}02){+}(1.430e{-}04){\times}x_{Population,i}{+}{-}1.879e{+}00{\times}x_{Life_Exp,i}{+}(-2.135e{-}02){\times}x_{Frost,i}{+}(1.246e{-}05){\times}x_{Area,i}$$

3.5 Erweiterung: Elastic Net

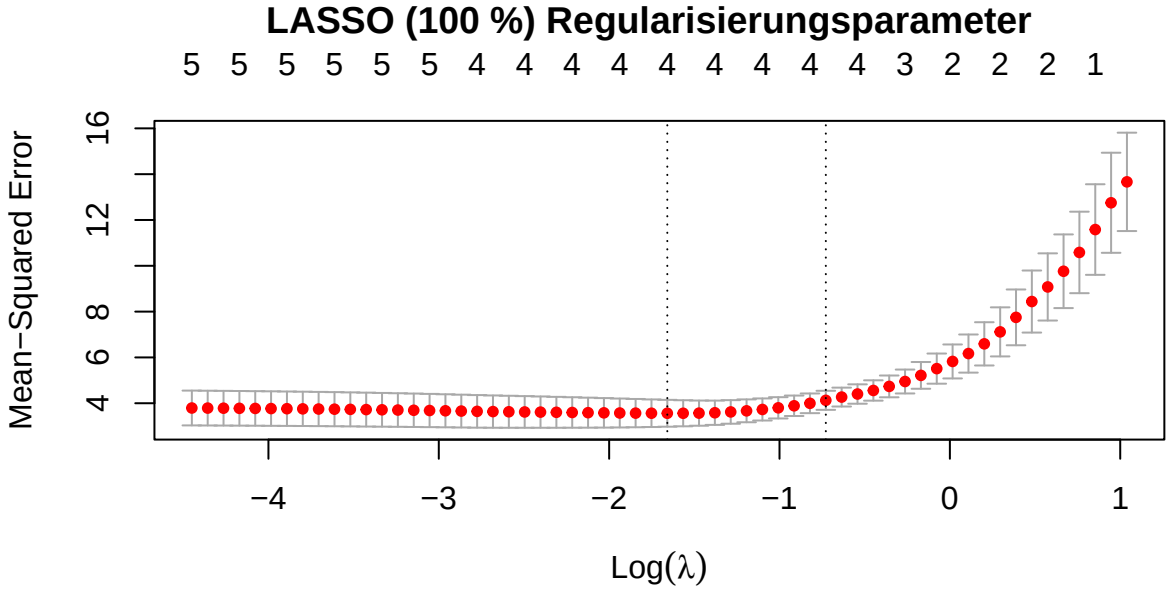
Um zu veranschaulichen, welchen Einfluss LASSO, Ridge und Elastic Net Regressionsalgorithmen auf diesen (bereinigten) Datensatz haben, sind im Anschluss folgende Szenarien dargestellt:

1) 100% LASSO (alpha = 1)
2) 100% Ridge (alpha = 0)
3) 50%-50% Elastic Net (alpha = 0.5)

```
# Plot 1a: LASSO - Verlauf der Koeffizienten/Variablen entlang Log Lambda
fitL <- glmnet(x=X, y=y, alpha=1, lambda= lambda.grid)
plot(fitL,xvar="lambda",label=TRUE, main = "LASSO (100 %) \n")
```




```
# Plot 1b: LASSO - Kreuzvalidierung, um den besten Lambda-Wert zu finden
cv_modelL <- cv.glmnet(X, y, alpha = 1)
plot(cv_modelL, main = "LASSO (100 %) Regularisierungsparameter  \n")
```



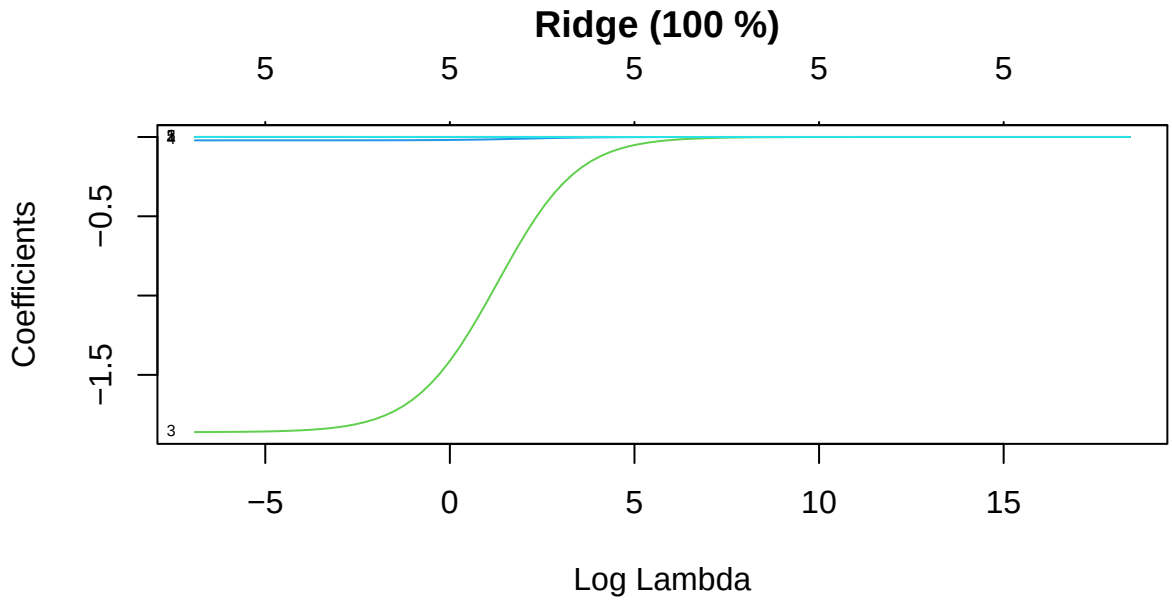

```
# Extrahiere den besten (niedrigsten) Lambda-Wert, Lasso
best_lambdaL <- cv_modelL$lambda.min
best_lambdaL
```

```
> [1] 0.19
```

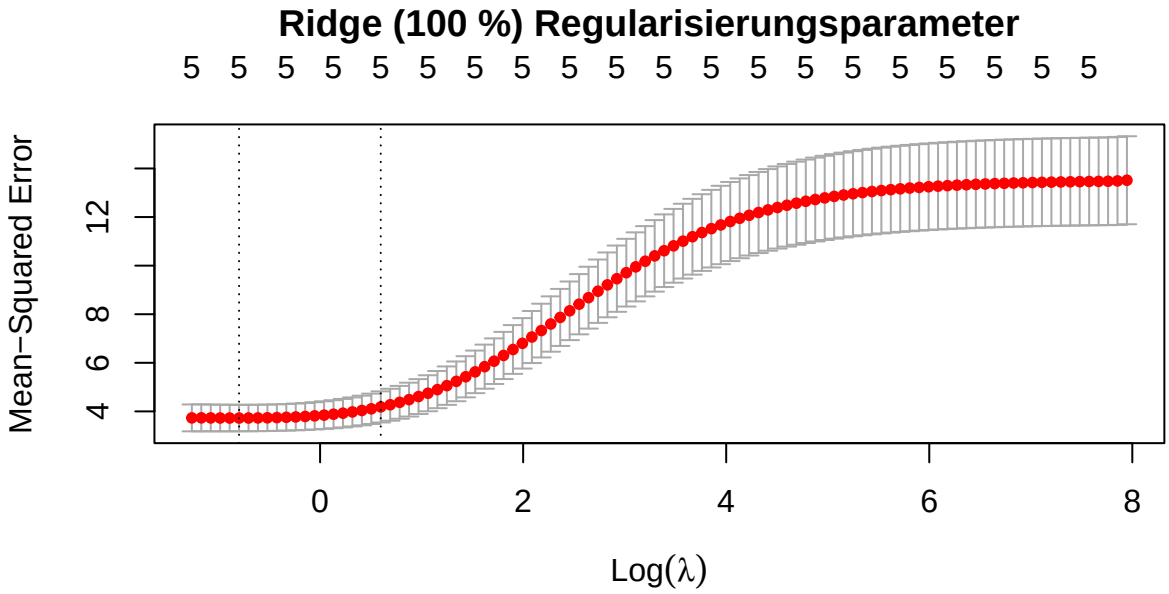
```r
# Passe das finale LASSO Modell mit dem besten Lambda-Wert an
LASSO_fit <- glmnet(X, y, alpha = 1, lambda = best_lambdaL)

# Koeffizienten des finalen LASSO Modells
print(coef(LASSO_fit))
```

```
> 6 x 1 sparse Matrix of class "dgCMatrix"
>                     s0
> (Intercept)  1.33e+02
> Population   1.19e-04
> Income       .
> Life_Exp    -1.75e+00
> Frost       -1.96e-02
> Area         8.87e-06
```

```r
# Plot 2a: Ridge - Verlauf der Koeffizienten/Variablen entlang Log Lambda
fitR <- glmnet(x=X, y=y, alpha=0, lambda= lambda.grid)
plot(fitR,xvar="lambda",label=TRUE, main = "Ridge (100 %) \n")
```




```r
# Plot 2b: Ridge - Kreuzvalidierung, um den besten Lambda-Wert zu finden
cv_modelR <- cv.glmnet(X, y, alpha = 0)
plot(cv_modelR, main = "Ridge (100 %) Regularisierungsparameter  \n")
```



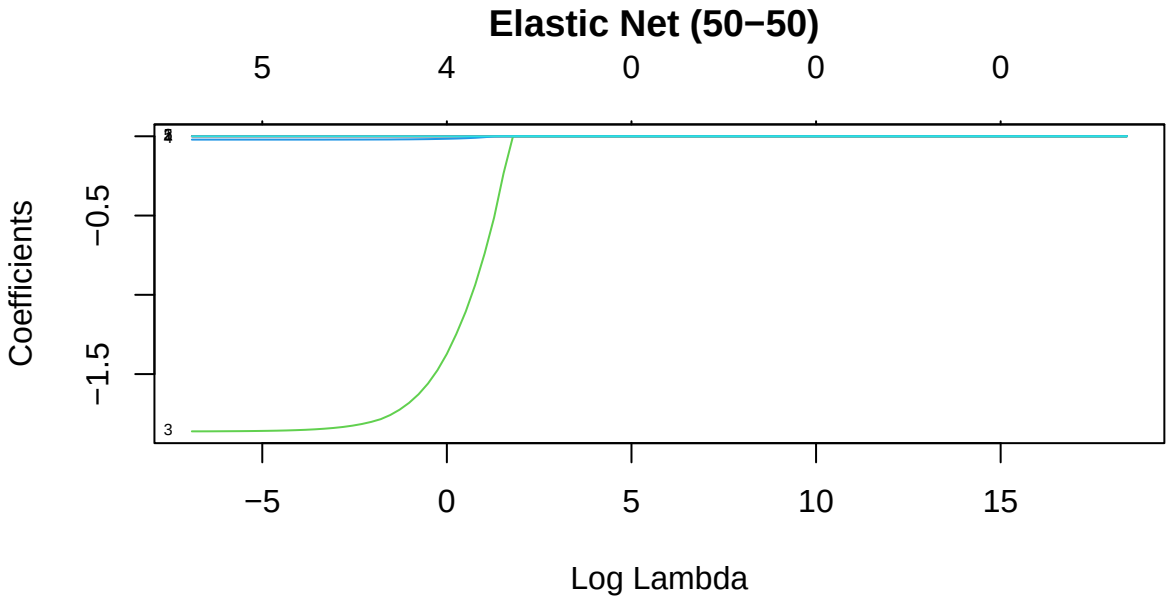

```r
# Extrahiere den besten (niedrigsten) Lambda-Wert, Lasso
best_lambdaR <- cv_modelR$lambda.min
best_lambdaR
```

```
> [1] 0.45
```

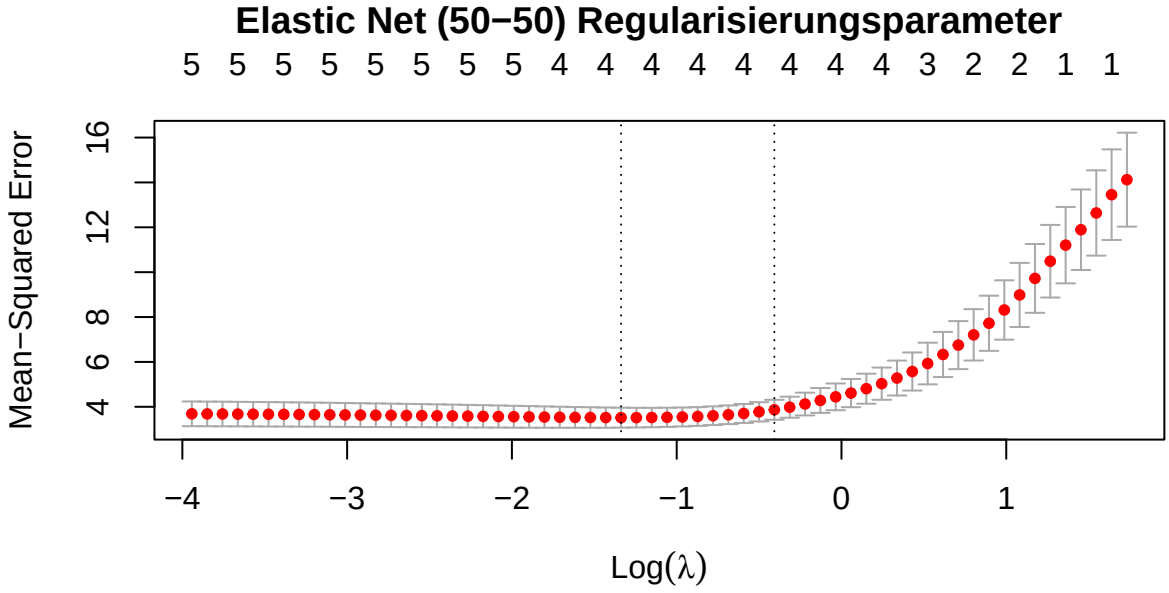
```r
# Passe das finale Ridge Modell mit dem besten Lambda-Wert an
Ridge_fit <- glmnet(X, y, alpha = 0, lambda = best_lambdaR)

# Koeffizienten des finalen Ridge Modells
print(coef(Ridge_fit))
```

```
> 6 x 1 sparse Matrix of class "dgCMatrix"
>                      s0
> (Intercept)  1.24e+02
> Population   1.54e-04
> Income      -3.60e-04
> Life_Exp    -1.62e+00
> Frost       -2.00e-02
> Area         1.05e-05
```

```r
# Plot 3a: Elastic Net - 50% - 50%
fitEN <- glmnet(x=X, y=y, alpha=0.5, lambda= lambda.grid)
plot(fitEN,xvar="lambda",label=TRUE, main = "Elastic Net (50-50) \n")
```




```
# Plot 3b: Elastic Net - Kreuzvalidierung, um den besten Lambda-Wert zu finden
cv_modelEN <- cv.glmnet(X, y, alpha = 0.5)
plot(cv_modelEN, main = "Elastic Net (50-50) Regularisierungsparameter  \n")
```




```
# Extrahiere den besten (niedrigsten) Lambda-Wert, Elastic Net
best_lambdaEN <- cv_modelEN$lambda.min
best_lambdaEN
```

```
> [1] 0.262
```

```r
# Passe das finale Elastic Net Modell mit dem besten Lambda-Wert an
elastic_net_fit <- glmnet(X, y, alpha = 0.5, lambda = best_lambdaEN)

# Koeffizienten des finalen 50-50 Elastic Net Modells
print(coef(elastic_net_fit))
```

```
> 6 x 1 sparse Matrix of class "dgCMatrix"
>                     s0
> (Intercept)  1.31e+02
> Population   1.25e-04
> Income       .
> Life_Exp    -1.73e+00
> Frost       -1.99e-02
> Area         9.57e-06
```

Im Vergleich der LASSO und Ridge **Coefficients vs. Log Lambda Plots** zeigt sich, dass bei LASSO die Koeffizienten an einem gewissen Punkt rapide fallen/steigen, 0 erreichen und somit aus dem Modell ausscheiden. Bei der Ridge Regression hingegen nähern sich die Werte zwar 0 an, da es sich um eine Exponentialfunktion handelt, wird der Nullwert jedoch nie ganz erreicht.

In allen drei Fällen (1 - LASSO, 2 - Ridge, 3 - Elastic Net 50/50) zeigt sich hinsichtlich der Zielvariable Murder ein ähnlicher Trend:

- die **Income** Variable wurde von LASSO und Elastic Net eliminiert (nicht signifikant, wie schon zuvor mittels t-Test ermittelt), nicht jedoch von Ridge.
- die Koeffizienten der **Population** und **Area** Variablen haben eine Steigung > 0. Dies bedeutet, dass ein **Anstieg der Bevölkerung und der Fläche eines Gebiets** im analysierten Datensatz mit einer **Zunahme der Mordrate** einhergeht (schwach signifikant lt. t-Test).
- die Koeffizienten der **Life Expectancy** und **Frost** Variablen haben eine Steigung < 0. Die negativen Steigungen für Lebenserwartung und Frost deuten auf eine umgekehrte Beziehung zur Mordrate hin. Dies bedeutet, dass **Verbesserungen der Lebenserwartung und eine Zunahme der Frosttage** mit einer **Verringerung der Mordrate** verbunden sind (hochsignifikant lt. t-Test).

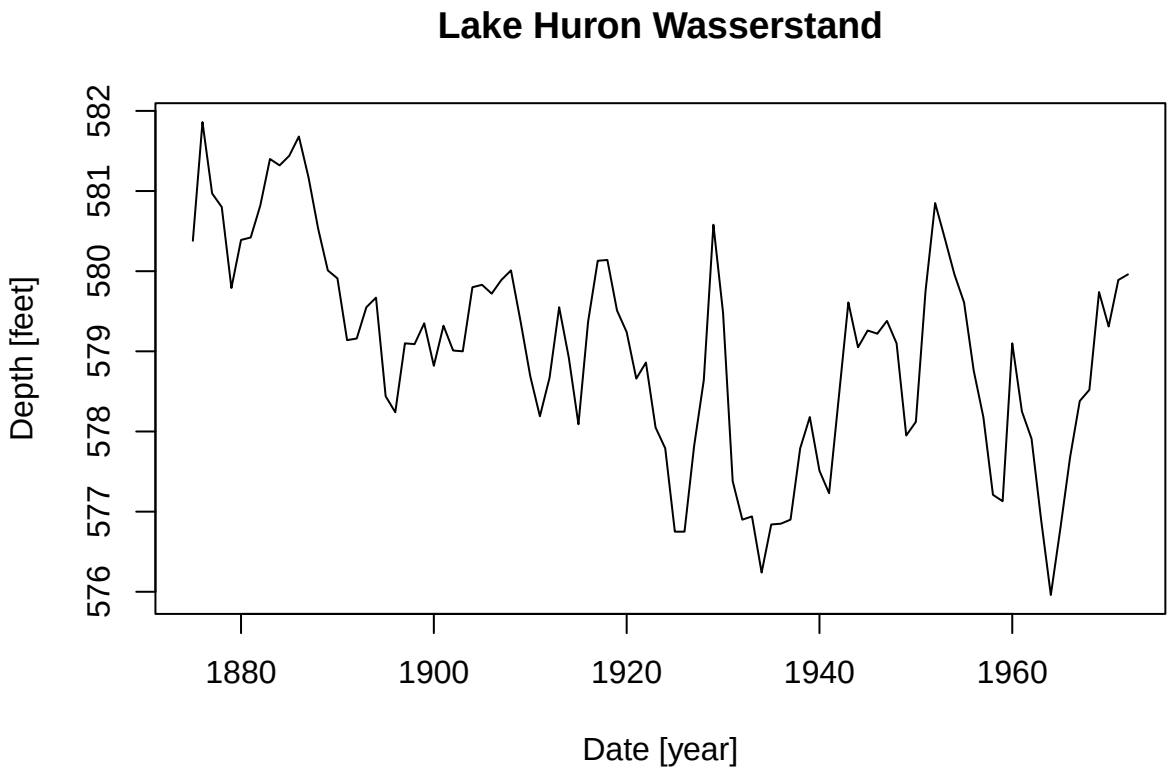
Aufgabe 4: Lake Huron

4.1 Aufgabenstellung

- 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.

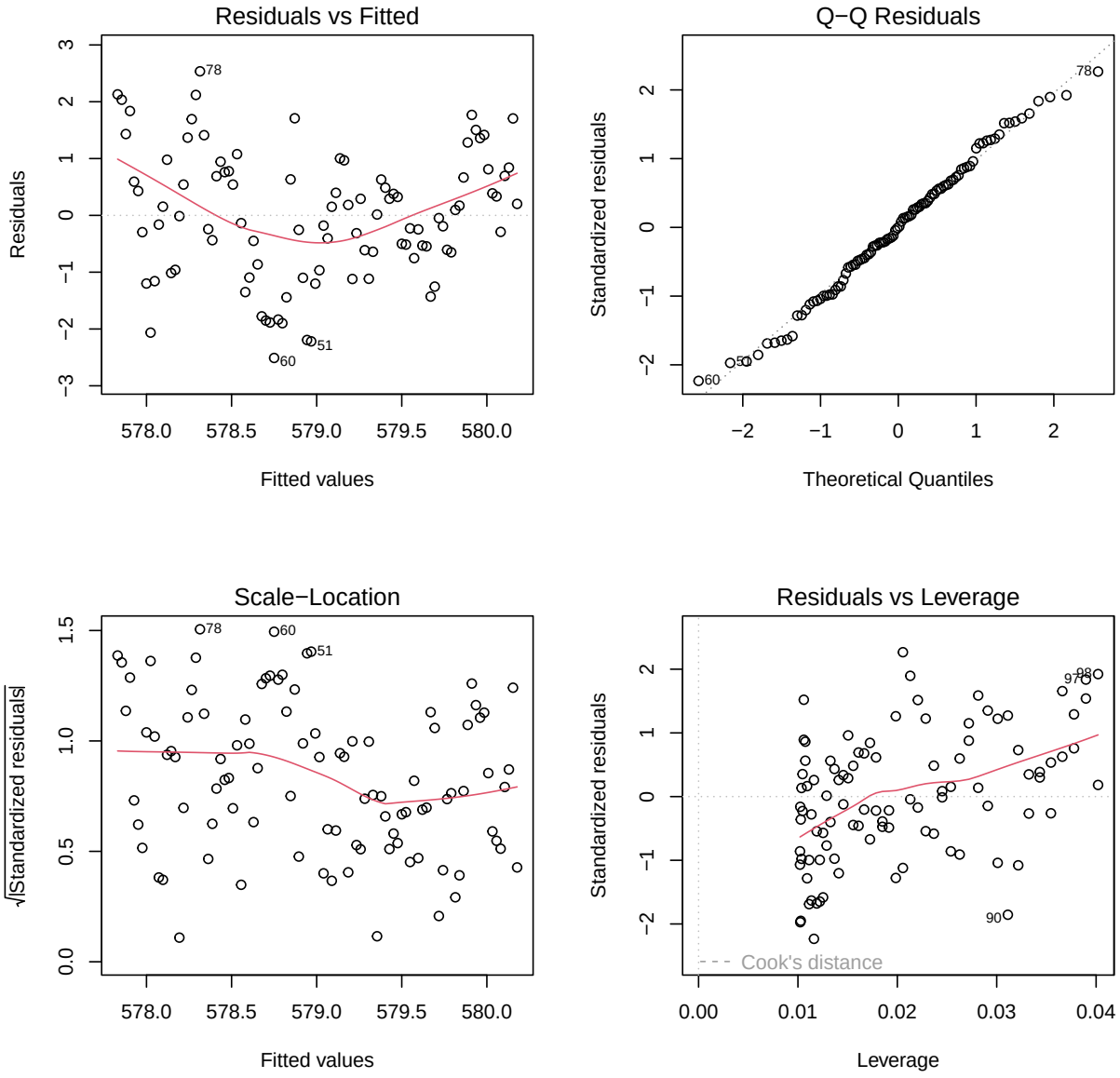
4.2 Beschreibung des Datensatzes und erste Untersuchung der Daten

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




```
Mean:              579.00

Median:            579.12

Min:               575.96

Max:               581.86

Variance:          1.74

Standardeviation:  1.32
```

4.3 Modellanpassung



Anhand der Modellanpassung 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.

4.4 Interpretation der Ergebnisse

Der Datensatz Lake Huron kann unter anderem deswegen nicht weiter bearbeitet werden (im Sinne eines linearen Modells), da wir hier nur eine tatsächliche Variable vorliegen haben. Zwar können wir die Zeit ablesen, allerdings eignet sich ein lineares Modell nicht unbedingt als vorhersage der weiteren Entwicklung des Sees. So wäre hier eine Zeitreihenanalyse deutlich sinnvoller, da es sich nicht um ein gewöhnliches Element handelt, was sich so gut linear berechnen lässt sondern auch deswegen, weil hier noch Umweltfaktoren vorhanden sind, die sich in den Messdaten zwar widerspiegeln (als Visualisierung), so aber nicht hinterlegt sind. Dazu gehören unter anderem Starkwetterereignisse oder auch Jahreszeiten im Allgemeinen. Dies deutet im Übrigen auch auf korrelierte Daten hin, was wiederrum eine Regel für verletzt, welche besagt, dass die Daten keinen systematischen Fehler aufweisen dürfen. Exakt das ist aber hier der Fall.

Aufgabe 5: Pima Indians

5.1 Aufgabenstellung

- 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.

5.2 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. Dieser Datensatz stellt einen Teildatensatz eines größeren Datensatzes, bestehend aus Pima.tr (n=200), Pima.tr2 (n=132) und Pima.te (Pima.tr sowie weiteren Daten, welche allerdings fehlende Daten beinhalten) dar. Im Folgenden wird nun Pima.tr weiter betrachtet.

- **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.)

```
suppressMessages(library(MASS))
suppressMessages(library(corrplot))
suppressMessages(library(pROC))
suppressMessages(library(magrittr))
suppressMessages(library(tidyr))
suppressMessages(library(dplyr))

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
```

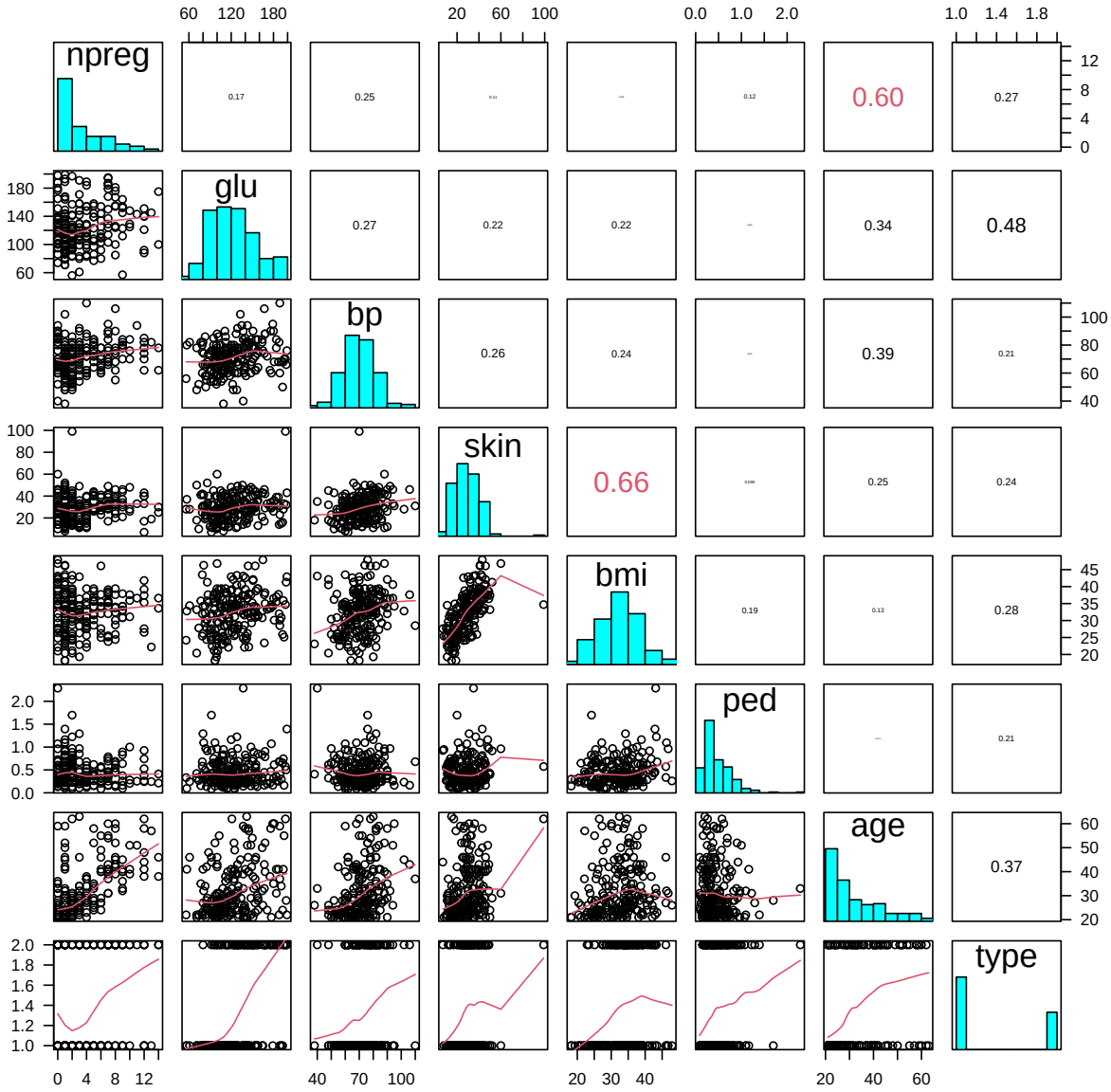
```r
str(Pima.tr)
```

```
## 'data.frame':    200 obs. of  8 variables:
##  $ npreg: int  5 7 5 0 0 5 3 1 3 2 ...
##  $ glu  : int  86 195 77 165 107 97 83 193 142 128 ...
##  $ bp   : int  68 70 82 76 60 76 58 50 80 78 ...
##  $ skin : int  28 33 41 43 25 27 31 16 15 37 ...
##  $ bmi  : num  30.2 25.1 35.8 47.9 26.4 35.6 34.3 25.9 32.4 43.3 ...
##  $ ped  : num  0.364 0.163 0.156 0.259 0.133 ...
##  $ age  : int  24 55 35 26 23 52 25 24 63 31 ...
##  $ type : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 2 1 1 1 2 ...
```

```r
data0_Pimatr <- Pima.tr
```

5.3 Entfernen korrelierender Variablen

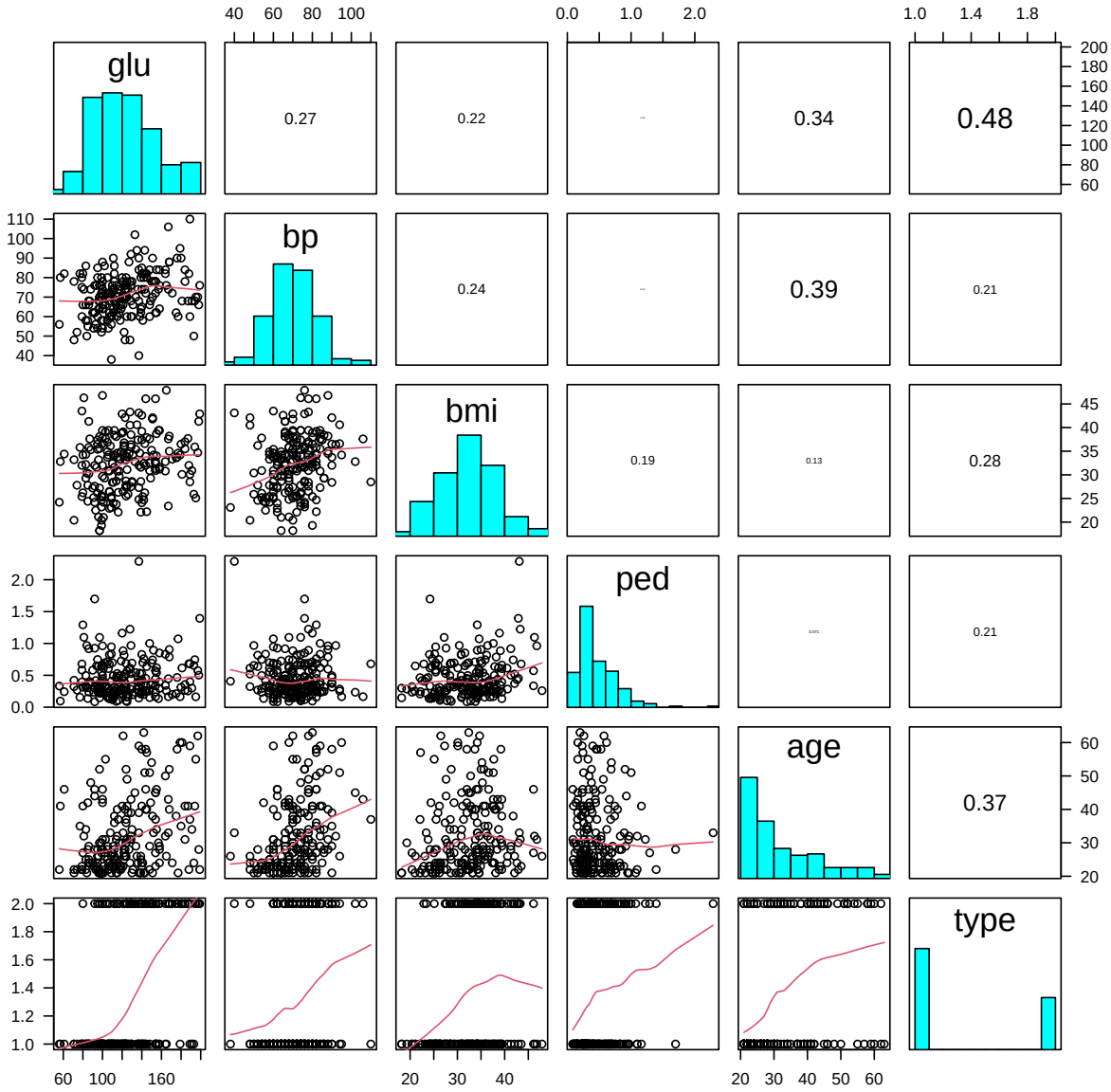
Da wir das Auftreten von Diabetes (Variable 'type' durch die anderen Variablen erklären wollen, müssen wir zuerst einmal schauen, ob Variablen gibt, die - von type abgesehen - miteinander korrelieren und diese entfernen.



Wir sehen, dass die Variablen ‘skin’ und ‘bmi’ stark miteinander korrelieren. Daher müsste eine dieser Variablen entfernt werden. Weiters korrelieren die Variablen ‘npreg’ und ‘age’ stark miteinander.

Die Variable ‘bmi’ ist ein berechneter Wert, welcher sich aus zwei gemessenen Werten zusammen setzt. Zudem ist Korrelation mit type etwas höher verglichen zu skin. Daher wird die Variable ‘skin’ aus dem Datensatz entfernt.

Die Variable ‘npreg’ steht für die Anzahl der Schwangerschaften und ist damit rein für Frauen relevant, während ‘age’ ein Faktor ist, welche für Frauen und Männer relevant ist. Es wird daher die Variable ‘npreg’ entfernt um die Korrelation eher allgemein gültig zu halten.



Wir haben hier prinzipiell die Situation, dass die Variable type nominal-skaliert ist mit zwei Ausprägungen "ja" und "nein" (=kategorial). Dadurch ergibt sich eine Binomial-Verteilung. Als Folge ist hier eine logistische Regression anzuwenden.

Zwischen den erklärenden Variablen (ausgenommen type, welche erklärt werden soll) ist keine Korrelation mehr erkennbar. Da die Grundvoraussetzung, dass Kovariablen nicht miteinander korrelieren dürfen ($R < 0.5$) nun erfüllt ist, kann mit diesem Datensatz weitergearbeitet werden.

5.4 Aufstellen des generalisierten linearen Modells

```
(modell_glm <- glm(type~.,data=data1_Pimatr,family=binomial(link = "logit")))
```

```
>
```

```
> Call:  glm(formula = type ~ ., family = binomial(link = "logit"), data = data1_Pimatr)
>
> Coefficients:
> (Intercept)          glu           bp          bmi          ped          age
>    -9.76294      0.03158     -0.00517      0.07872      1.72920      0.06053
>
> Degrees of Freedom: 199 Total (i.e. Null);  194 Residual
> Null Deviance:        256
> Residual Deviance: 181    AIC: 193
```

```
summary(modell_glm)
```

```
>
> Call:
> glm(formula = type ~ ., family = binomial(link = "logit"), data = data1_Pimatr)
>
> Coefficients:
>             Estimate Std. Error z value Pr(>|z|)
> (Intercept) -9.76294    1.68999   -5.78  7.6e-09 ***
> glu          0.03158    0.00675    4.68  2.9e-06 ***
> bp          -0.00517    0.01824   -0.28   0.7767
> bmi          0.07872    0.03281    2.40   0.0164 *
> ped          1.72920    0.66009    2.62   0.0088 **
> age          0.06053    0.01890    3.20   0.0014 **
> ---
> 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: 181.00  on 194  degrees of freedom
> AIC: 193
>
> Number of Fisher Scoring iterations: 5
```

Die Variable bp trägt nichts zum Regresionsmodell bei und werden dabei entfernt.

```
data2_Pimatr <- data1_Pimatr
data2_Pimatr$bp <- NULL
(modell_glm_2 <- glm(type~.,data=data2_Pimatr,family=binomial(link = "logit")))
```

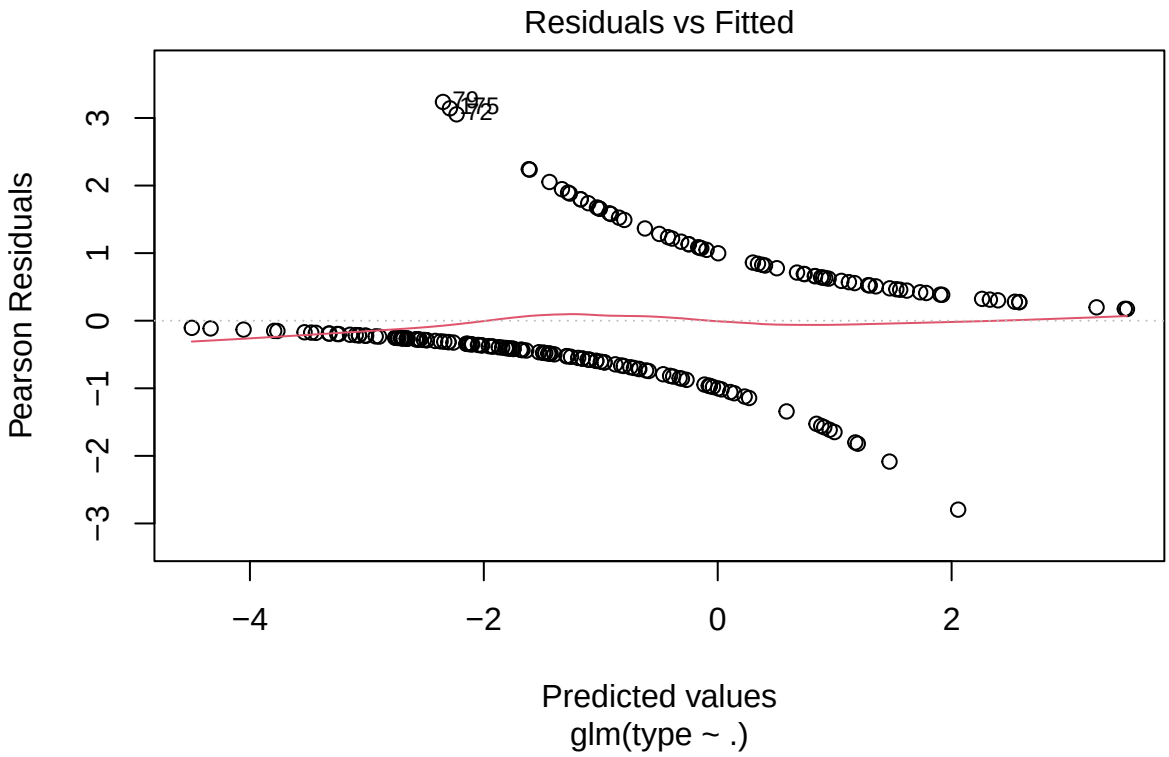
```
##
## Call:  glm(formula = type ~ ., family = binomial(link = "logit"), data = data2_Pimatr)
##
## Coefficients:
## (Intercept)          glu          bmi          ped          age
##     -9.9714       0.0313       0.0770       1.7198       0.0586
##
## Degrees of Freedom: 199 Total (i.e. Null);  195 Residual
## Null Deviance:       256
## Residual Deviance: 181   AIC: 191
```

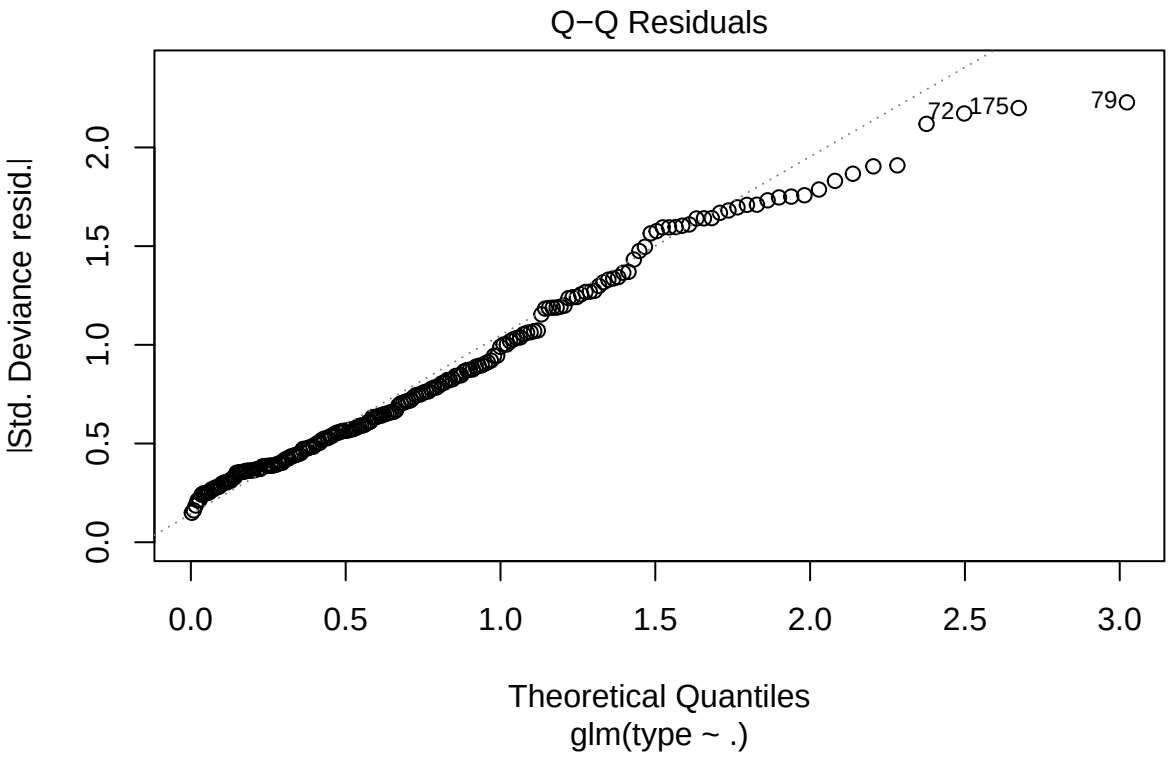
```r
summary(modell_glm_2)
```

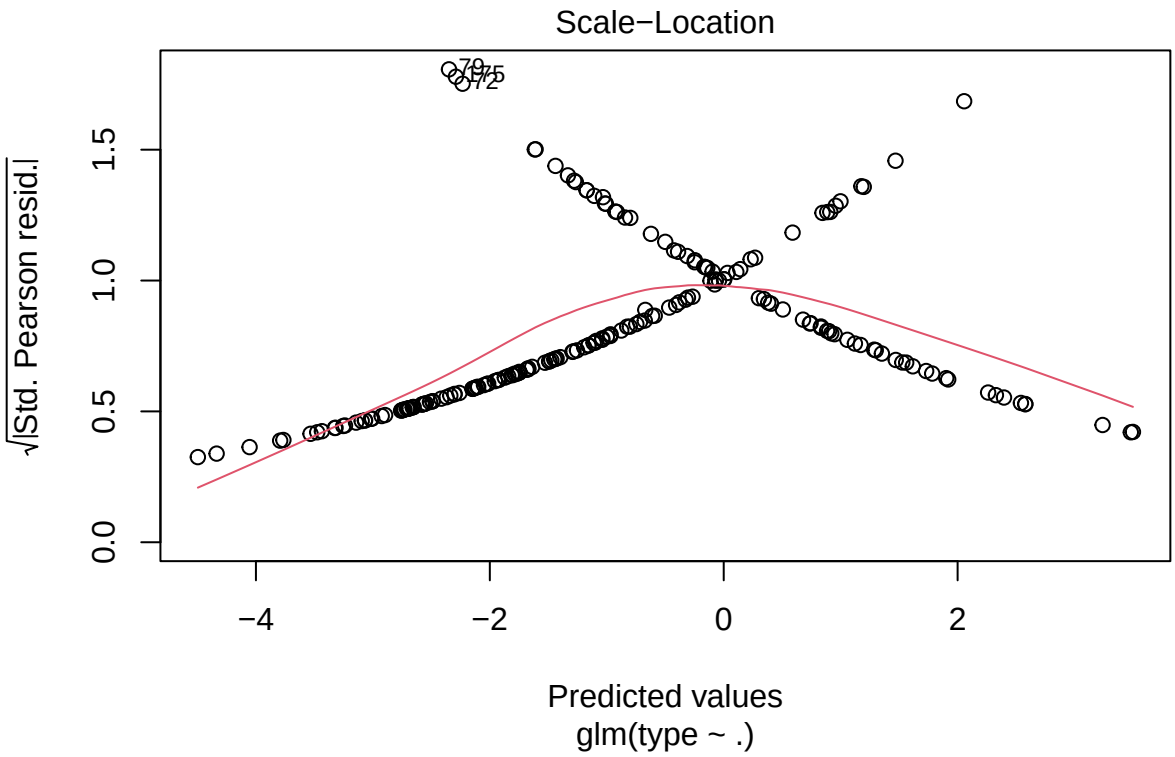
```
##
## Call:
## glm(formula = type ~ ., family = binomial(link = "logit"), data = data2_Pimatr)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.97139    1.52759   -6.53  6.7e-11 ***
## glu          0.03126    0.00663    4.72  2.4e-06 ***
## bmi          0.07703    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
```

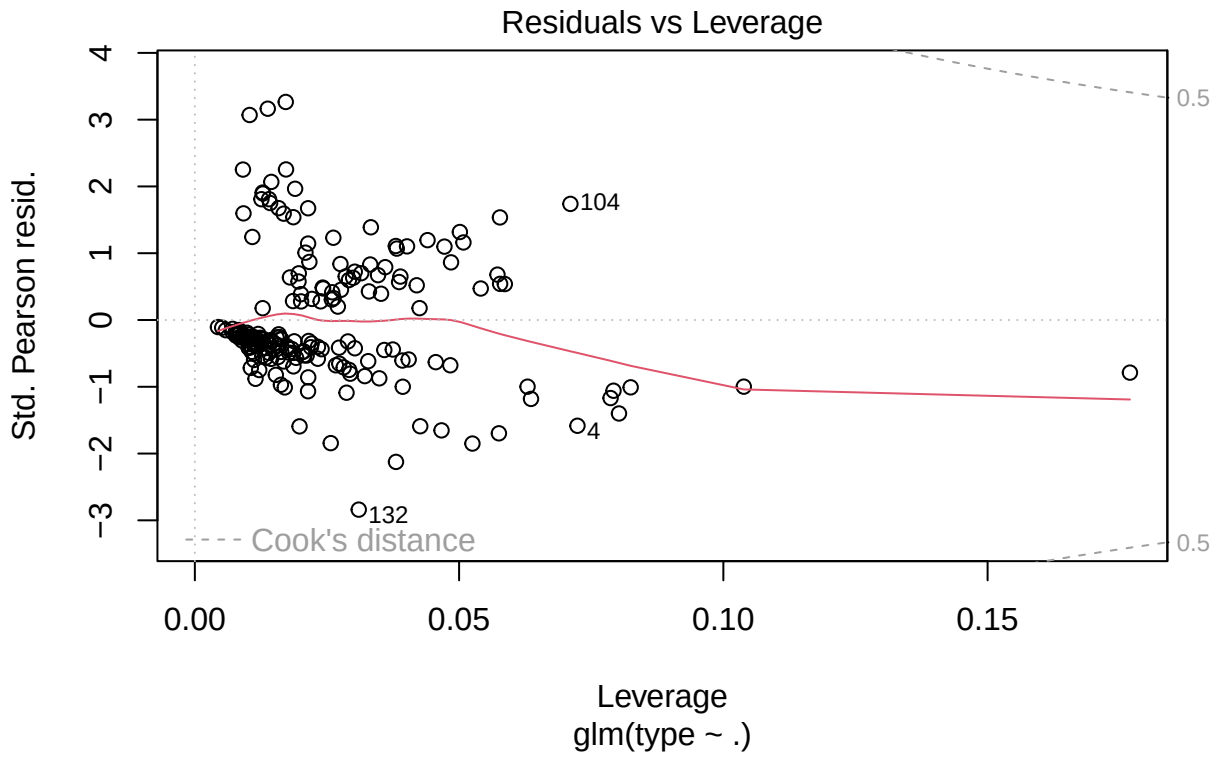
Wir haben nun nur mehr Variablen, welche zum Modell etwas beitragen. Wir überprüfen nun die Residuen Plots.

```r
plot(modell_glm_2)
```


Residuals vs Fitted
Pearson Residuals
Predicted values
glm(type ~ .)
79
75
72
3
2
1
0
−1
−2
−3
−4
−2
0
2


Q−Q Residuals
|Std. Deviance resid.|
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
72
175
79
Theoretical Quantiles
glm(type ~ .)


Scale−Location
79
175
172
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
−4
−2
0
2
Predicted values
glm(type ~ .)



5.5 Erste Interpretation der Ergebnisse

Mithilfe der oben angeführten Plots (ausgenommen der Scatterplots) werden nun die Anforderungen für ein generalisiertes lineares Modell geprüft.

Auf Basis folgender Quelle: https://schmidtpaul.github.io/crashcouRse/intro_glm_carrot.html Punkt: "Lösung 2: Generalisierte Modelle"

... fallen zwei Annahmen für die Residuen weg, welche bei linearen Modellen wichtig sind: * Varianzhomogenität * Normalverteilung der Fehler

Nicht zu vergessen ist hierbei, dass wir durch die link-funktion "logit" eine logistische Transformation durchgeführt haben.

Gemäß Recherche scheint der Plot "Residuals vs. fitted" eine übliche Verteilung der Residuen bei einer binomial Verteilung darzustellen und wäre daher in Ordnung.

Jedenfalls scheint hier die bimodalität nicht unbedingt zu stören und auch die beiden leichten Ränder, welche im Q-Q Plot zu sehen sind, dürften das Ergebnis kaum beeinflussen.

Der Plot Residuals vs. Leverage zeigt einen Datenpunkt (48) an, welcher weitab der anderen Datenpunkt liegt. Jedoch befindet sich dieser nach wie vor innerhalb der Cook's Distance und dürfte daher keinen zu starken Einfluss haben, so dass er eventuell das Modell aushebeln könnte. Wir können diesen daher im Datensatz belassen.

Auf Basis folgender Quelle: http://giscience.courses-pages.gistools.geog.uni-heidelberg.de/einfuerung_statistik/generalisierte-lineare-modelle-f%C3%BCr-z%C3%A4hldaten.html

gilt es zu beachten, dass es seitens der generalisierten linearen Modelle keinen R^2 Wert mehr gibt, die Confidence aber anders abschätzbar ist, nämlich:

```
modell_glm_2$deviance/modell_glm_2$null.deviance
```

```
## [1] 0.706
```

Demnach scheint das Modell nicht so schlecht zu sein, da deutlicher größer als 0.5.

5.6 Modellgleichung (logistisch)

```
modell_glm_2$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:

$$logit_(type_i) = -9.97138818 + 0.03125508 \times x_{glu,i} + 0.07703027 \times x_{bmi,i} + 1.71979415 \times x_{ped,i} + 0.05860297 \times x_{age,i} + \varepsilon_i$$

5.7 Umkehrung der linearen Transformation

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

Hierfür gilt:

Für die Umkehrung des Terms sowie dessen Interpretation wurde ChatGPT zu Hilfe genommen!

$$p_i = 1 - \frac{1}{1 + e^{-9.97138818 + 0.03125508 \times x_{glu,i} + 0.07703027 \times x_{bmi,i} + 1.71979415 \times x_{ped,i} + 0.05860297 \times x_{age,i}}}$$

Daraus folgt:

$$p_i = 1 - \frac{1}{1 + 4.662204 \times 10^{-5} \times (1.031777)^{x_{glu,i}} \times (1.080150)^{x_{bmi,i}} \times (5.585137)^{x_{ped,i}} \times (1.060360)^{x_{age,i}}}$$

5.8 Interpretation des Ergebnisses

Der Interzept ist mit 4,6*-10^5 sehr (!) gering, daher ist die Basiswahrscheinlichkeit NICHT an Diabetis zu erkranken recht gering. Die Frage ist hier, ob das Ergebnis so stimmen kann. Ein Anstieg der Plasmaglukosekonzentration um eine Einheit ist mit einer Erhöhung des Logits für Diabetes um 0.031 verbunden. Dies bedeutet, dass höhere Glukosewerte die Wahrscheinlichkeit für Diabetes erhöhen. Ein Anstieg des BMI um eine Einheit ist mit einer Erhöhung des Logits für Diabetes um 0.077 verbunden. Höherer BMI erhöht die Wahrscheinlichkeit für Diabetes. Ein Anstieg der Diabetes-Stammbaumfunktion um eine Einheit ist mit einer Erhöhung des Logits für Diabetes um 1.72 verbunden. Dies deutet darauf hin, dass eine stärkere genetische Prädisposition (gemessen durch die Stammbaumfunktion) die Wahrscheinlichkeit für Diabetes stark erhöht. Ein Anstieg des Alters um ein Jahr ist mit einer Erhöhung des Logits für Diabetes um 0.0586 verbunden. Ältere Menschen haben eine höhere Wahrscheinlichkeit, an Diabetes zu erkranken.

Ende der Zuhilfenahme von ChatGPT

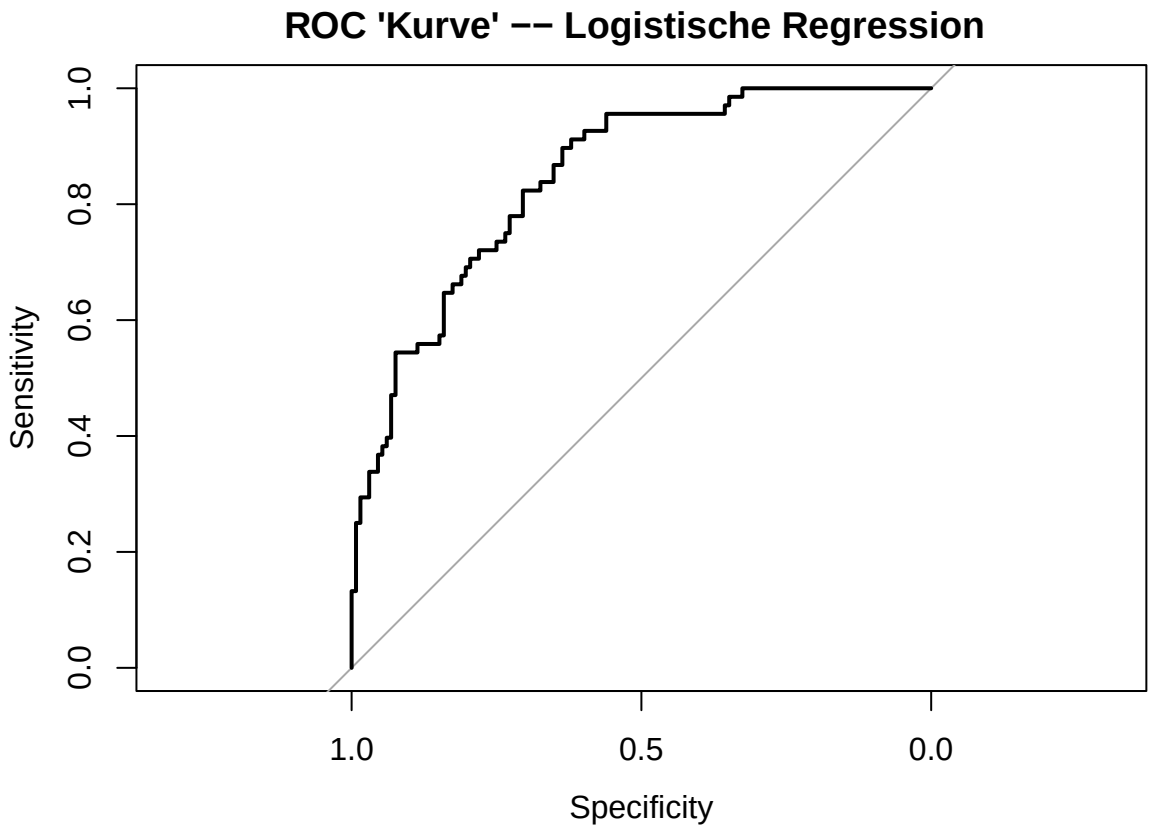
5.9 ROC Kurve

```
predictions_filtered <- predict(modell_glm_2, data2_Pimatr, type="response")
roc_curve_filtered <- roc(data2_Pimatr$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 0.8385695, somit hat die Entfernung der Koeffizient npreg, bp und skin, die Qualität des Modells nicht signifikant beeinflusst.

5.10 Erstellung der Confusion-Matrix

```r
Pima_Original <- Pima.tr
Pima_Original$truefalse <- ifelse(predict(modell_glm_2, type = "response") > 0.5, "Yes", "No")
```

```r
data2_Pimatr$truefalse <- ifelse(predict(modell_glm_2, type = "response") > 0.5, "Yes", "No")
```

```r
library(caret)
model2 = table(predicted = data2_Pimatr$truefalse, actual = Pima_Original$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
```

```r
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
##
```

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 vorhergesagte 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%.

5.11 Cross Validation

Für die Cross-Validation wurde das Jackknifing verwendet. Hier werden zuerst die Daten "zerschnitten" und anschließend nach und nach durchgetestet.

```
library(glmnet)
jack_data<-as.data.frame(data1_Pimatr)
X <- as.matrix(jack_data[, c("glu", "bp", "bmi", "ped", "age")])
#y <- jack_data$`type`
y <- ifelse(jack_data$type == "Yes", 1, 0)

lambda.grid <- 10^seq(-3, 8, length=100)

n <- nrow(jack_data)
errors <- numeric(n)

for (i in 1:n) {
  X_train <- X[-i,]
  y_train <- y[-i]
  X_test <- X[i,,drop=FALSE]
  y_test <- y[i]
  cv_fit <- cv.glmnet(x=X_train, y=y_train, alpha=1, lambda=lambda.grid)
  best_lambda <- cv_fit$lambda.min
  fitL <- glmnet(x=X_train, y=y_train, alpha=1, lambda=best_lambda)
  y_pred <- predict(fitL, newx=X_test, s=best_lambda)
  errors[i] <- (y_test - y_pred)^2
}

mse <- mean(errors)
cat("Mean Squared Error: ", mse, "\n")
```

```
## Mean Squared Error:  0.158
```

Der Mean Squared Error ist mit 0.158 recht niedrig, was für eine gute Qualität des Modells spricht.

5.12 Elastic Net

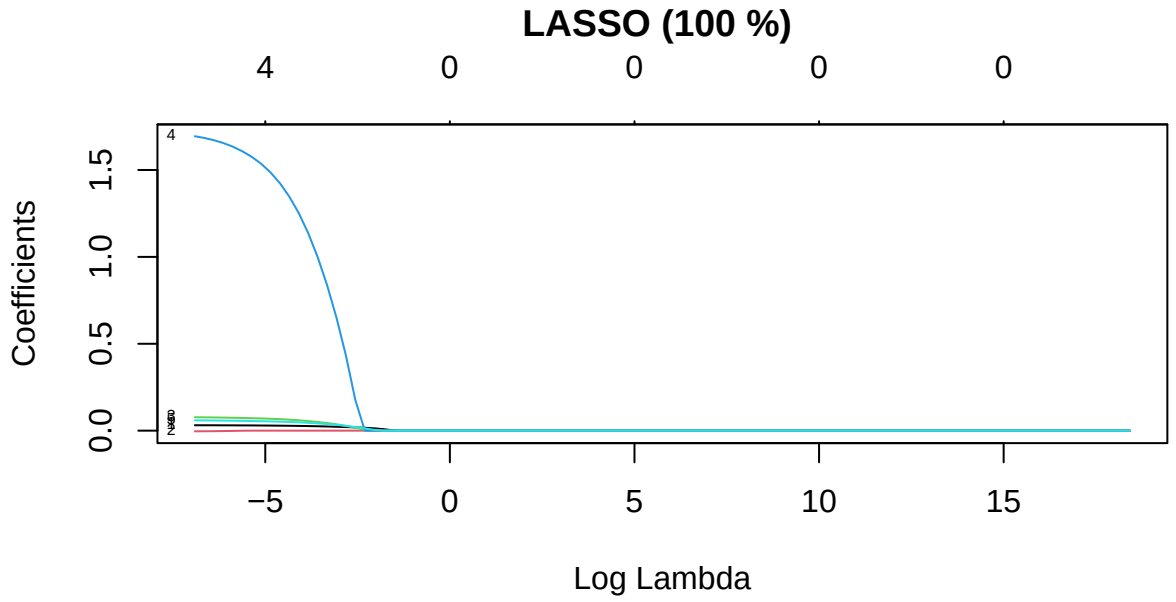
LASSO-Regression (Least Absolute Shrinkage and Selection Operator) ist eine Methode der linearen Regression, die durch L1-Regularisierung eine Strafe für die Summe der absoluten Werte der Regressionskoeffizienten hinzufügt, wodurch einige Koeffizienten auf Null gesetzt werden können. Das führt zu einer Schrumpfung der Koeffizienten und ermöglicht die Auswahl der wichtigsten Variablen, wodurch das Modell vereinfacht und die Überanpassung reduziert werden soll.

Ridge hingegeben geht von der Annahme aus, dass alle Variablen irgendwie wichtig sind und setzt diese ggf. nahe Null, wirft sie aber eher nicht raus.

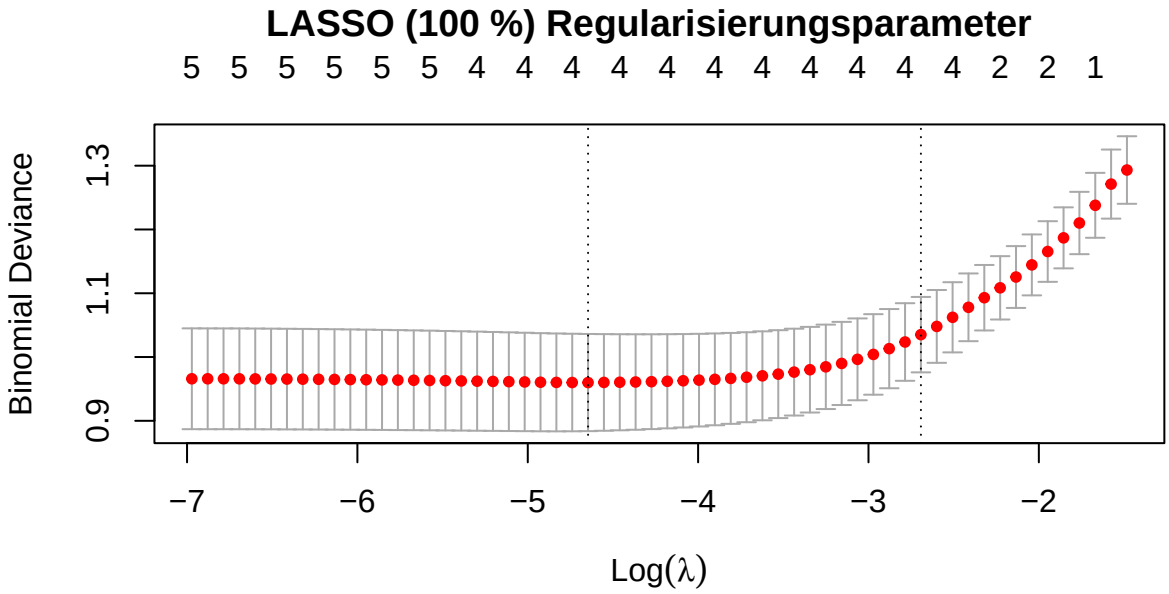
Elastische Netze bzw. Elastic Nets sind eine Möglichkeit beide Technologien miteinander zu verbinden um auf ein möglichst ideales Ergebnis zu kommen.

Um die Datenvoraussetzung zu erfüllen, wurde vorab die mit anderen Kovariablen korrelierende Examination-Variable entfernt.

```
# Plot 1a: LASSO - Verlauf der Koeffizienten/Variablen entlang Log Lambda
#y_binary <- ifelse(elnet_data$type == "Yes", 1, 0)
fitL <- glmnet(x=X, y=y, family="binomial", alpha=1, lambda= lambda.grid)
plot(fitL,xvar="lambda",label=TRUE, main = "LASSO (100 %) \n")
```




```
# Plot 1b: LASSO - Kreuzvalidierung, um den besten Lambda-Wert zu finden
cv_modelL <- cv.glmnet(X, y, family="binomial", alpha = 1)
plot(cv_modelL, main = "LASSO (100 %) Regularisierungsparameter  \n")
```



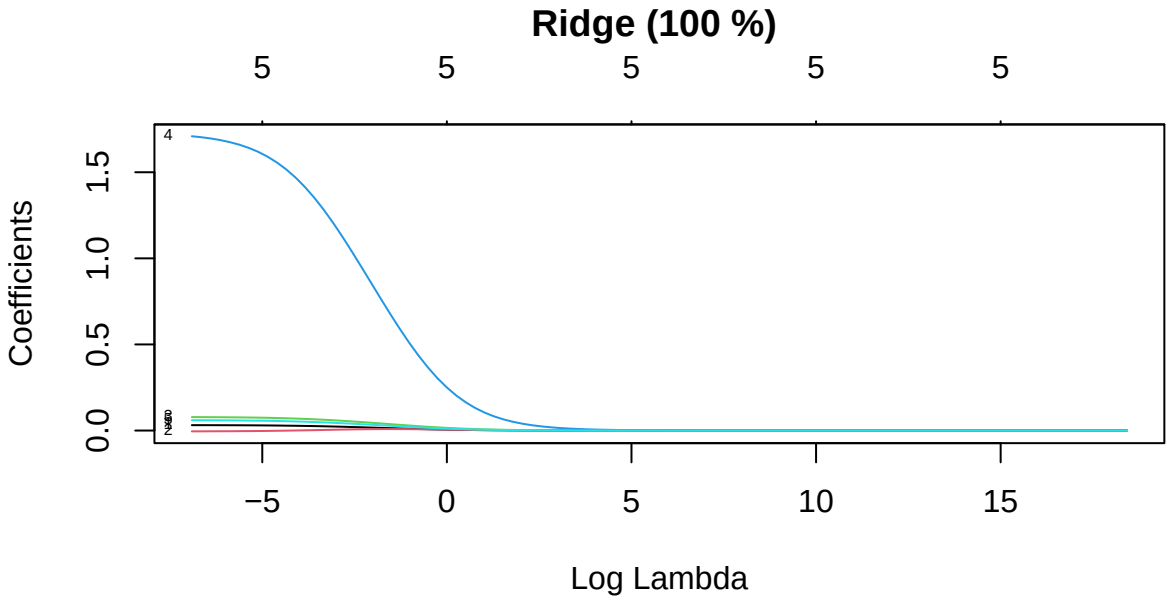

```
# Extrahiere den besten (niedrigsten) Lambda-Wert, Lasso
best_lambdaL <- cv_modellL$lambda.min
best_lambdaL
```

```
> [1] 0.0096
```

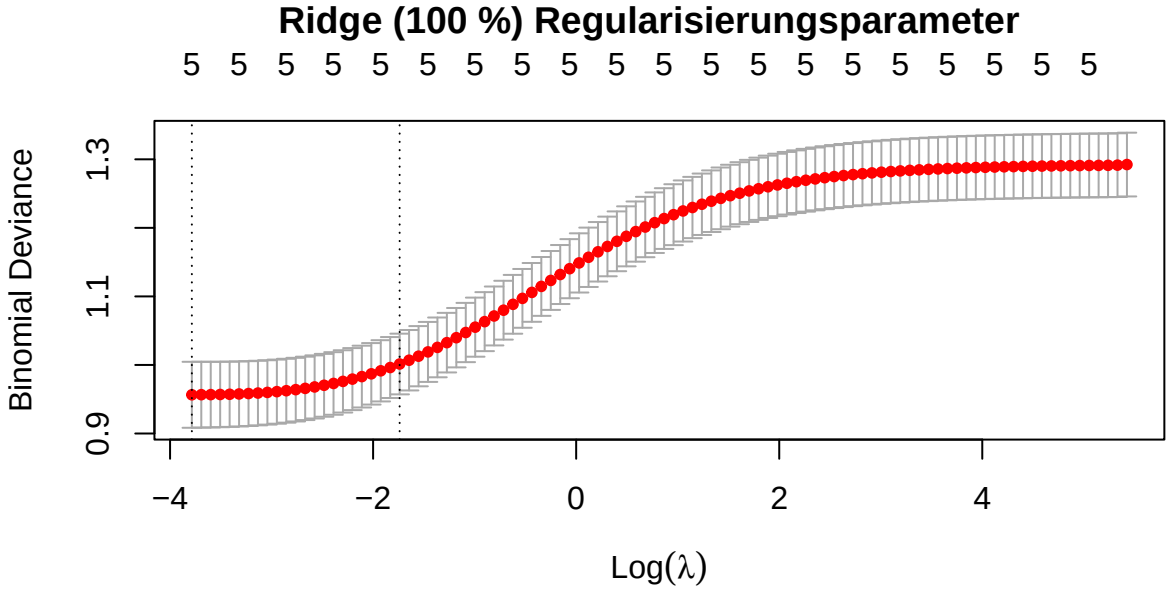
```
# Passe das finale LASSO Modell mit dem besten Lambda-Wert an
LASSO_fit <- glmnet(X, y, family="binomial", alpha = 1, lambda = best_lambdaL)

# Koeffizienten des finalen LASSO Modells
print(coef(LASSO_fit))
```

```
> 6 x 1 sparse Matrix of class "dgCMatrix"
>                    s0
> (Intercept) -8.9600
> glu          0.0289
> bp           .
> bmi          0.0665
> ped          1.4354
> age          0.0521
```

```
# Plot 2a: Ridge - Verlauf der Koeffizienten/Variablen entlang Log Lambda
fitR <- glmnet(x=X, y=y, family="binomial", alpha=0, lambda= lambda.grid)
plot(fitR,xvar="lambda",label=TRUE, main = "Ridge (100 %) \n")
```




```
# Plot 2b: Ridge - Kreuzvalidierung, um den besten Lambda-Wert zu finden
cv_modelR <- cv.glmnet(X, y=y, family="binomial", alpha = 0)
plot(cv_modelR, main = "Ridge (100 %) Regularisierungsparameter  \n")
```



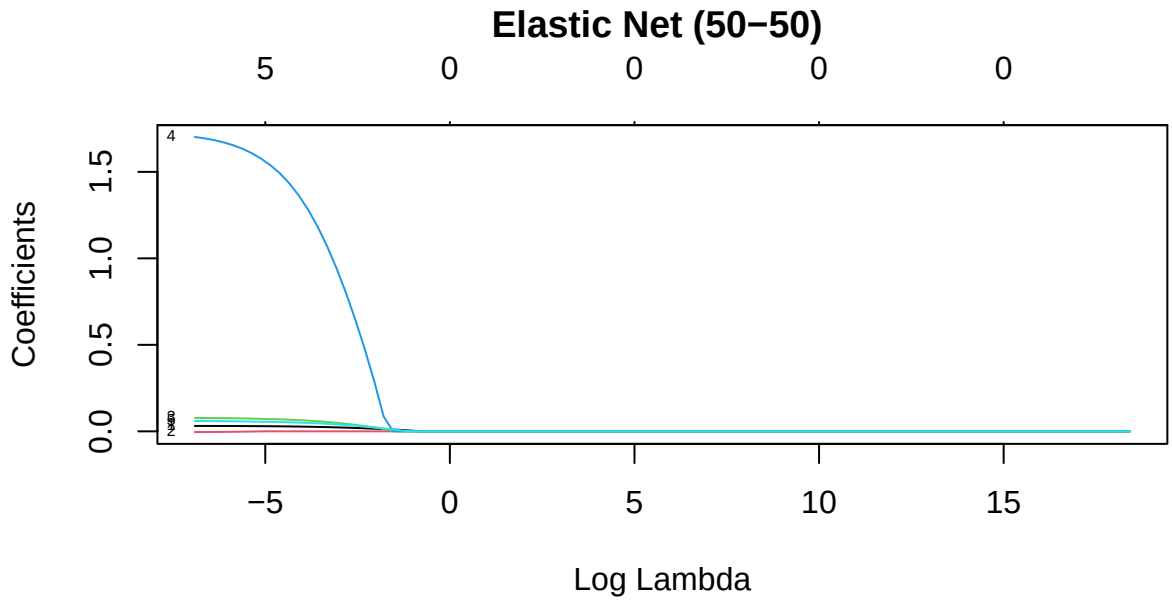

```
# Extrahiere den besten (niedrigsten) Lambda-Wert, Lasso
best_lambdaR <- cv_modelR$lambda.min
best_lambdaR
```

```
> [1] 0.0227
```

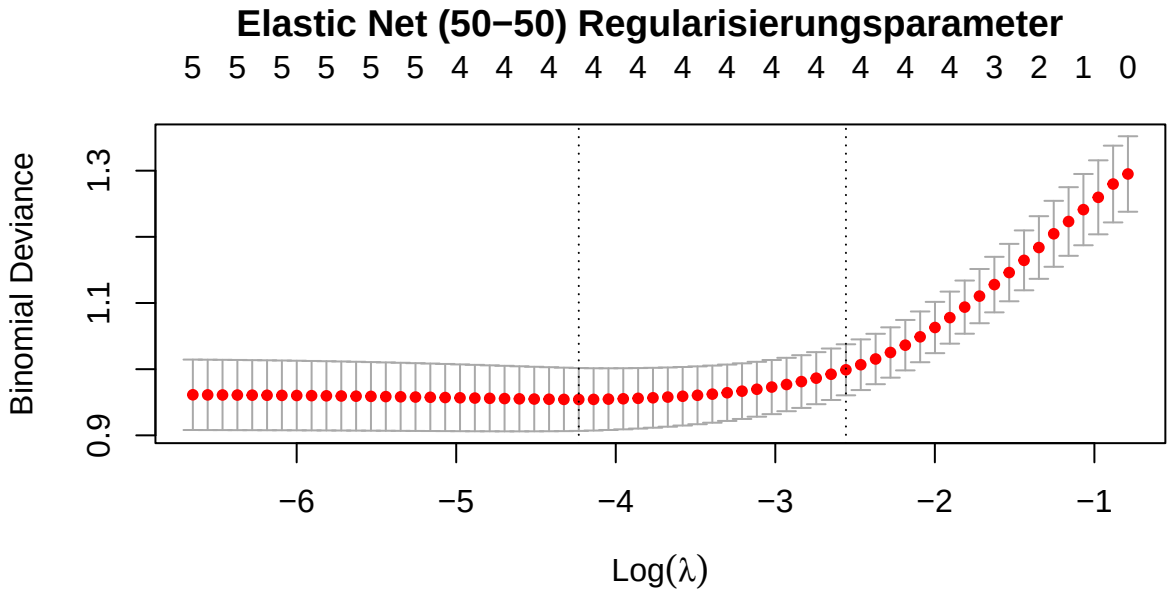
```r
# Passe das finale Ridge Modell mit dem besten Lambda-Wert an
Ridge_fit <- glmnet(X, y=y, family="binomial", alpha = 0, lambda = best_lambdaR)

# Koeffizienten des finalen Ridge Modells
print(coef(Ridge_fit))
```

```
> 6 x 1 sparse Matrix of class "dgCMatrix"
>                    s0
> (Intercept) -8.72086
> glu          0.02657
> bp           0.00149
> bmi          0.06703
> ped          1.40049
> age          0.05079
```

```r
# Plot 3a: Elastic Net - 50% - 50%
fitEN <- glmnet(x=X, y=y, family="binomial", alpha=0.5, lambda= lambda.grid)
plot(fitEN,xvar="lambda",label=TRUE, main = "Elastic Net (50-50) \n")
```




```r
# Plot 3b: Elastic Net - Kreuzvalidierung, um den besten Lambda-Wert zu finden
cv_modelEN <- cv.glmnet(X, y=y, family="binomial", alpha = 0.5)
plot(cv_modelEN, main = "Elastic Net (50-50) Regularisierungsparameter  \n")
```




```
# Extrahiere den besten (niedrigsten) Lambda-Wert, Elastic Net
best_lambdaEN <- cv_modelEN$lambda.min
best_lambdaEN
```

```
> [1] 0.0145
```

```
# Passe das finale Elastic Net Modell mit dem besten Lambda-Wert an
elastic_net_fit <- glmnet(X, y=y, family="binomial", alpha = 0.5, lambda = best_lambdaEN)

# Koeffizienten des finalen 50-50 Elastic Net Modells
print(coef(elastic_net_fit))
```

```
> 6 x 1 sparse Matrix of class "dgCMatrix"
>                    s0
> (Intercept) -8.7820
> glu          0.0279
> bp           .
> bmi          0.0662
> ped          1.4032
> age          0.0514
```

In allen drei Fällen (1 - LASSO, 2 - Ridge, 3 - Elastic Net 50/50) zeigt sich hinsichtlich der Zielvariable ein ähnlicher Trend:

- die bp Variable wurde von LASSO und Elastic Net eliminiert (nicht signifikant, wie schon zuvor ermittelt), nicht jedoch von Ridge
- sämtliche anderen Koeffizienten haben eine Steigung > 0. Dies bedeutet, dass ein Anstieg bspw. des Glukosegehaltes 'glu' mit einem höheren Risiko für Diabetis behaftet ist, was ebenfalls für die Variablen BMI 'bmi', die Diabetes Pedigree Funktion 'ped', sowie für das Alter 'age' der Fall ist.