Modeldiagnostik

VB

Institut für Veterinär-Epidemiologie und Biometrie, FU Berlin

11/17/2020

Wiederholung

Modelldiagnostik

Unregelmäßigkeiten in den Daten oder falsche Angaben zu den Beziehungen zwischen den Prädiktoren und der Antwortvariablen können dazu führen, dass Sie sich für ein Modell entscheiden, das äußerst ungenau ist.

Einerseits können Sie daraus schließen, dass ein Prädiktor und eine Antwortvariable keine Beziehung zueinander haben, obwohl dies tatsächlich der Fall ist.

Auf der anderen Seite können Sie den Schluss ziehen, dass ein Prädiktor und eine Antwortvariable zusammenhängen, obwohl dies nicht der Fall ist!

Es kann auch vorkommen, dass Sie ein Modell erhalten, das in realen Umgebungen schlechte Vorhersagen mit erheblichen und unnötigen Fehlern macht.

Modelldiagnostik

Eine Reihe von Techniken, die als Regressionsdiagnose bezeichnet werden, bieten die erforderlichen Tools zur Bewertung der Angemessenheit des Regressionsmodells und können Ihnen dabei helfen, Probleme aufzudecken und zu beheben.

Modelldiagnostik (1)

- Normalverteilung von Residuen
- Unabhängigkeit
- Linearität
- Homoskedastizität
- Ausreißer
- Kollinearität

Normalverteilung von Residuen

Wenn die abhängige Variable für einen festen Satz von Prädiktorwerten normalverteilt ist, sollten die Residuenwerte normalverteilt sein (mit einem Mittelwert von 0).

Das Q-Q-Plot ist ein Wahrscheinlichkeitsdiagramm der standardisierten Residuen gegen die Werte die für die normalverteilten Residuen zu erwarten sind. Wenn Sie die Normalitätsannahme erfüllt haben, sollten die Punkte in diesem Diagramm auf der geraden 45-Grad-Linie liegen.

z-score

$$z = \frac{x - \bar{x}}{\hat{\sigma}}$$

Q-Q-Plot

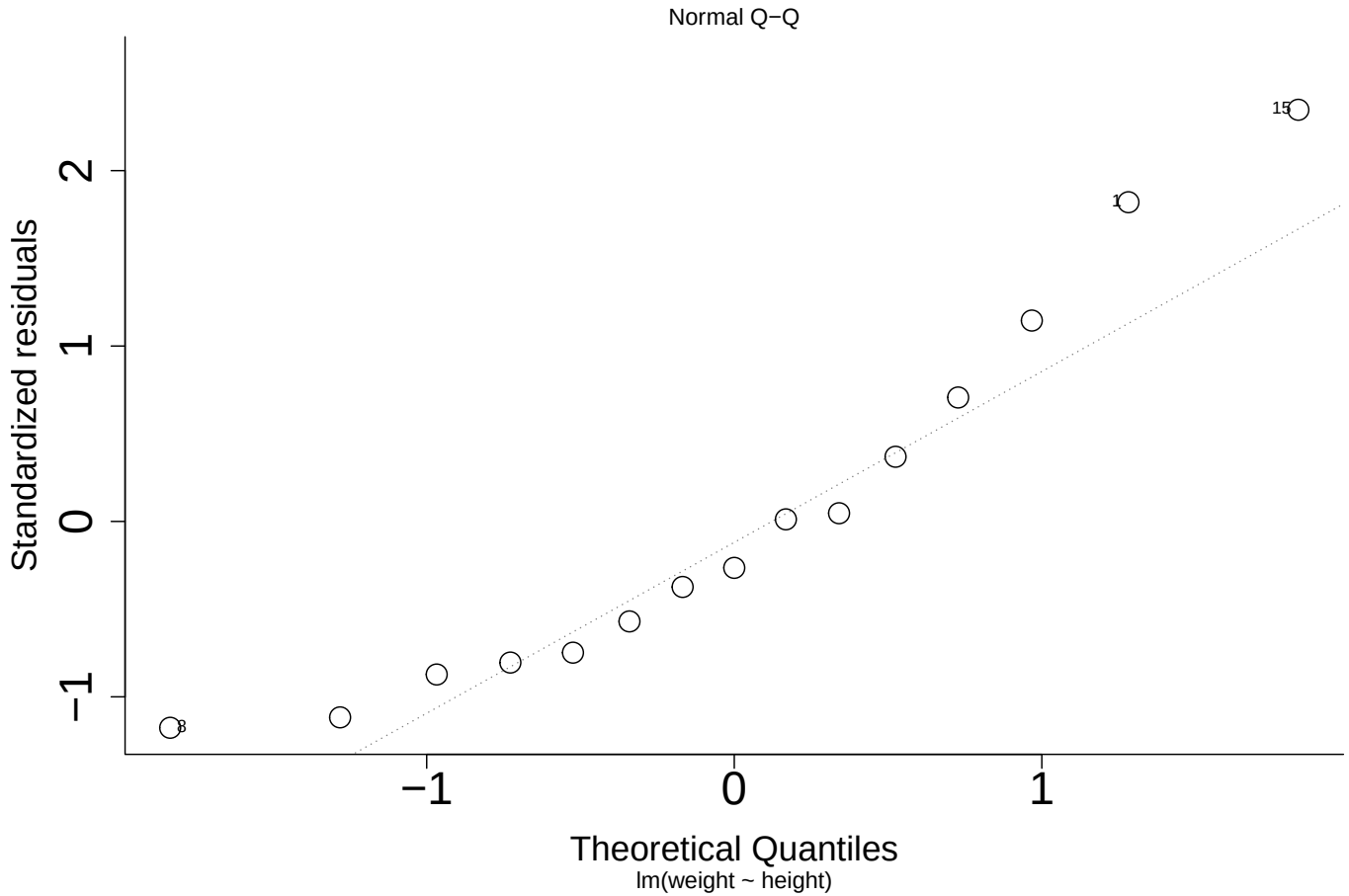

Figure 1: Modelldiagnostik für Gewicht vs Größe für Frauen

Quantil-Quantil-Diagramm (Q-Q plot)

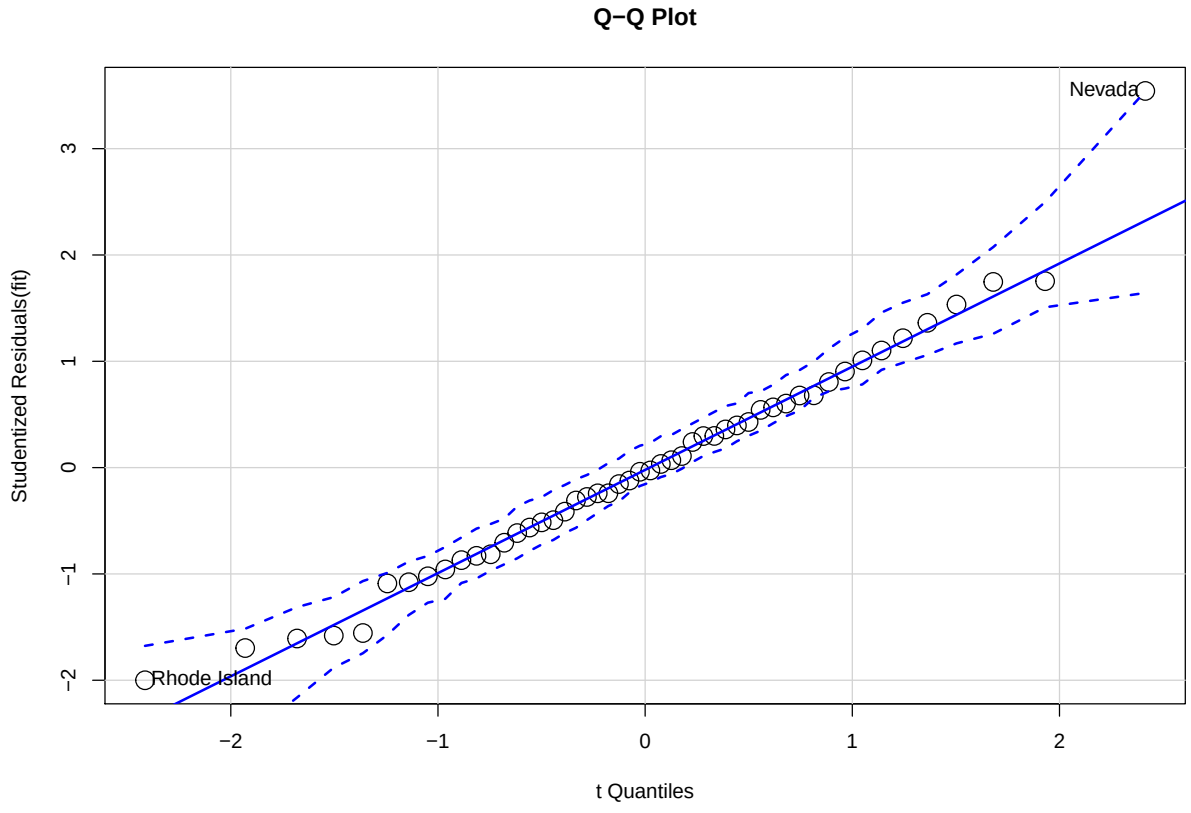

Figure 2: Modelldiagnostik auf Normalverteilung der Residuen

```
states["Nevada",]
```

```
##        Murder Population Illiteracy Income Frost
## Nevada   11.5        590        0.5   5149   188
```

```
fitted(fit)["Nevada"]
```

```
##   Nevada
## 3.878958
```

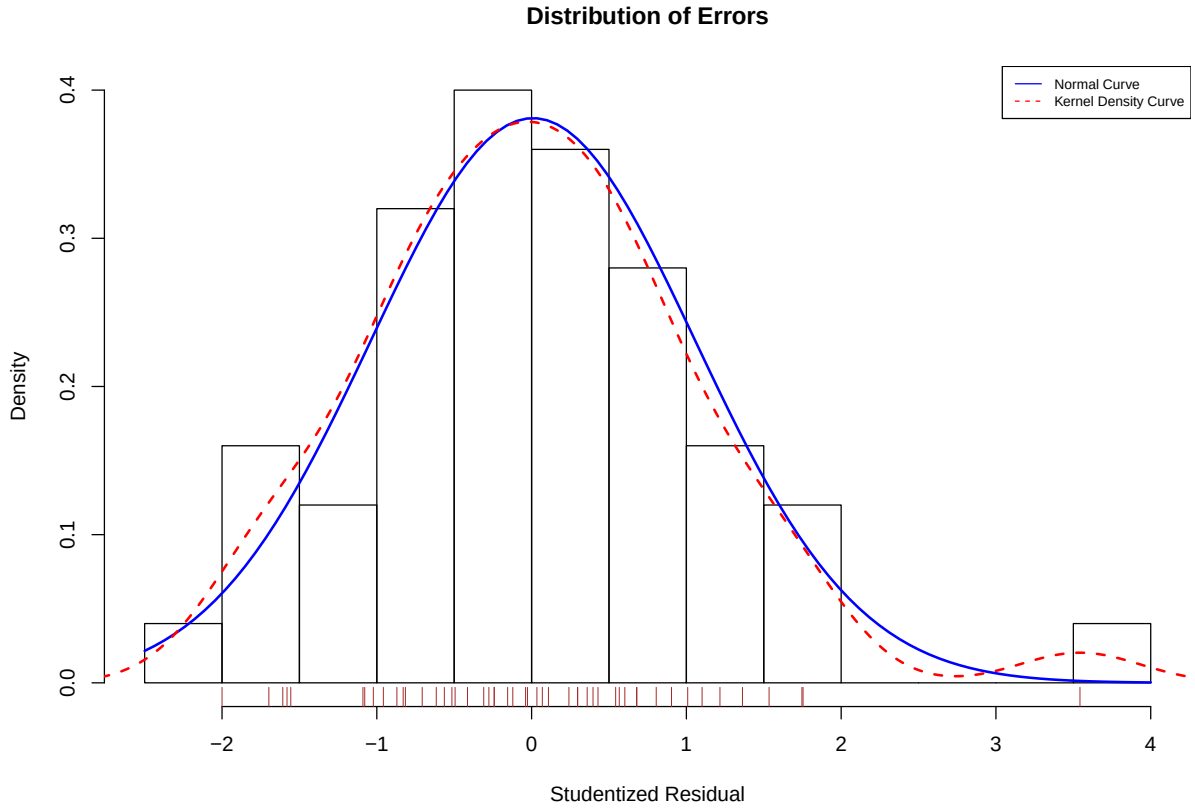

Figure 3: Modelldiagnostik auf Normalverteilung der Residuen

Quantil-Quantil-Diagramm (Q-Q plot) (1)

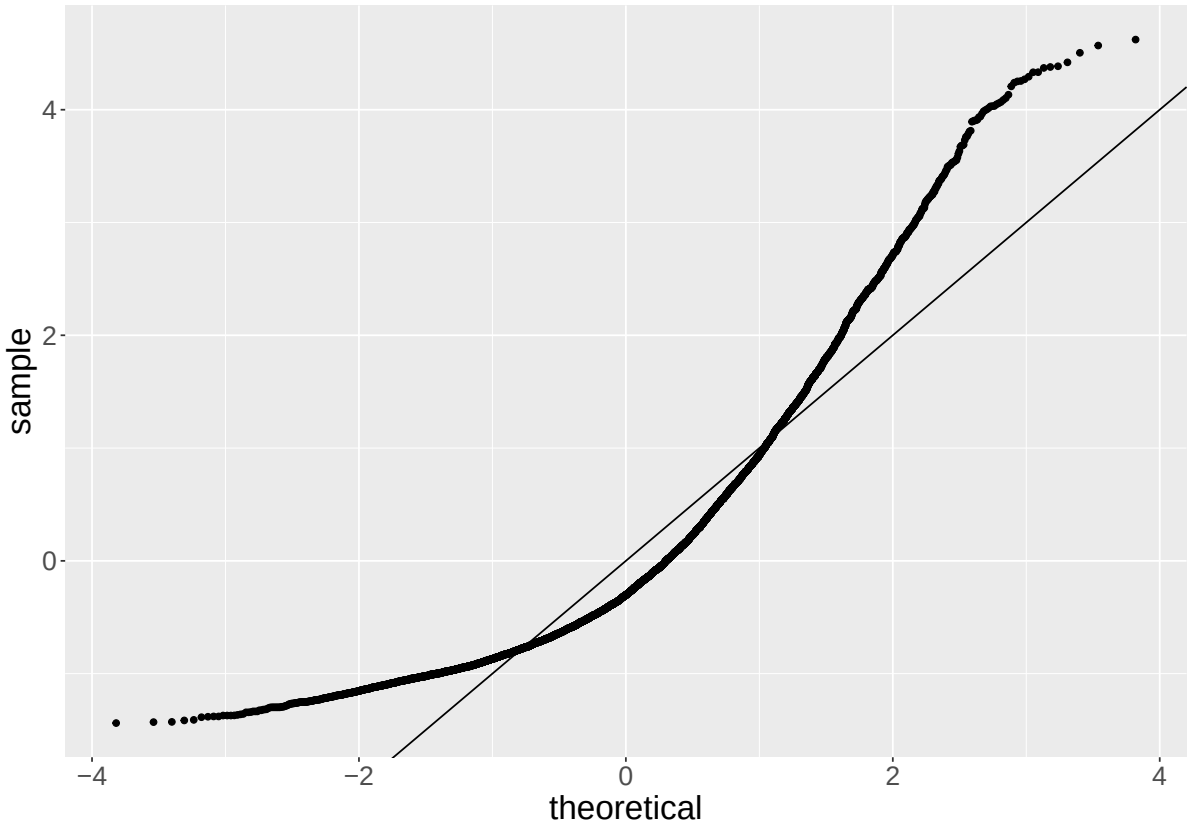

Figure 4: Evaluation of the normality of residuals assumption for interval from waiting period to conception (wpc) regression.

Normalverteilung von Residuen (1)

```
ggplot(model1, aes(sample = .stdresid)) +
  stat_qq() +
  theme(text = element_text(size = 20)) +
  geom_abline()
```

Shapiro-Wilk's test

p $<$ 0.05 indicates *non-normality*

```
states <- as.data.frame(state.x77[,c("Murder", "Population",
                        "Illiteracy", "Income", "Frost")])
fit <- lm(Murder ~ Population + Illiteracy + Income + Frost, data=states)
shapiro.test(resid(fit))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit)
## W = 0.98264, p-value = 0.6672
```

Unabhängigkeit

- ergibt sich in der Regel aus der Art der Daten
- mangelnde Unabhängigkeit, z. bei mehreren Beobachtungen eines einzelnen Tieres oder derselben Herde
- mangelnde Unabhängigkeit, wenn serielle Korrelationen wie Messungen während einer bestimmten Jahreszeit vorliegen

Unabhängigkeit (1)

Sie können nicht feststellen, ob die abhängigen Variablenwerte von diesen Darstellungen unabhängig sind. Sie müssen Ihr Verständnis dafür verwenden, wie die Daten gesammelt wurden. Es gibt keinen Grund zu der Annahme, dass das Gewicht einer Frau das Gewicht einer anderen Frau beeinflusst. Wenn Sie herausfinden, dass die Daten aus Familien stammen, müssen Sie möglicherweise Ihre Vermutung der Unabhängigkeit anpassen.

Durbin–Watson test

```
states <- as.data.frame(state.x77[,c("Murder", "Population",
                        "Illiteracy", "Income", "Frost")])
fit <- lm(Murder ~ Population + Illiteracy + Income + Frost, data=states)
durbinWatsonTest(fit)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.2006929      2.317691   0.222
##  Alternative hypothesis: rho != 0
```

Der nicht signifikante p-Wert (p = 0,288) deutet auf eine fehlende Autokorrelation und umgekehrt auf eine Unabhängigkeit von Fehlern hin. Der Verzögerungswert (in diesem Fall 1) gibt an, dass jede Beobachtung mit der im Datensatz daneben liegenden Beobachtung verglichen wird. Obwohl der Test für zeitabhängige Daten geeignet ist, gilt er weniger für Daten, die nicht auf diese Weise gruppiert wurden.

Linearität

Wenn die abhängige Variable linear mit den unabhängigen Variablen zusammenhängt, sollte es keine systematische Beziehung zwischen den Residuen und den vorhergesagten (dh angepassten) Werten geben. Mit anderen Worten, das Modell sollte alle in den Daten vorhandenen systematischen Abweichungen erfassen und nur zufälliges Rauschen zurücklassen.

Linearität (1)

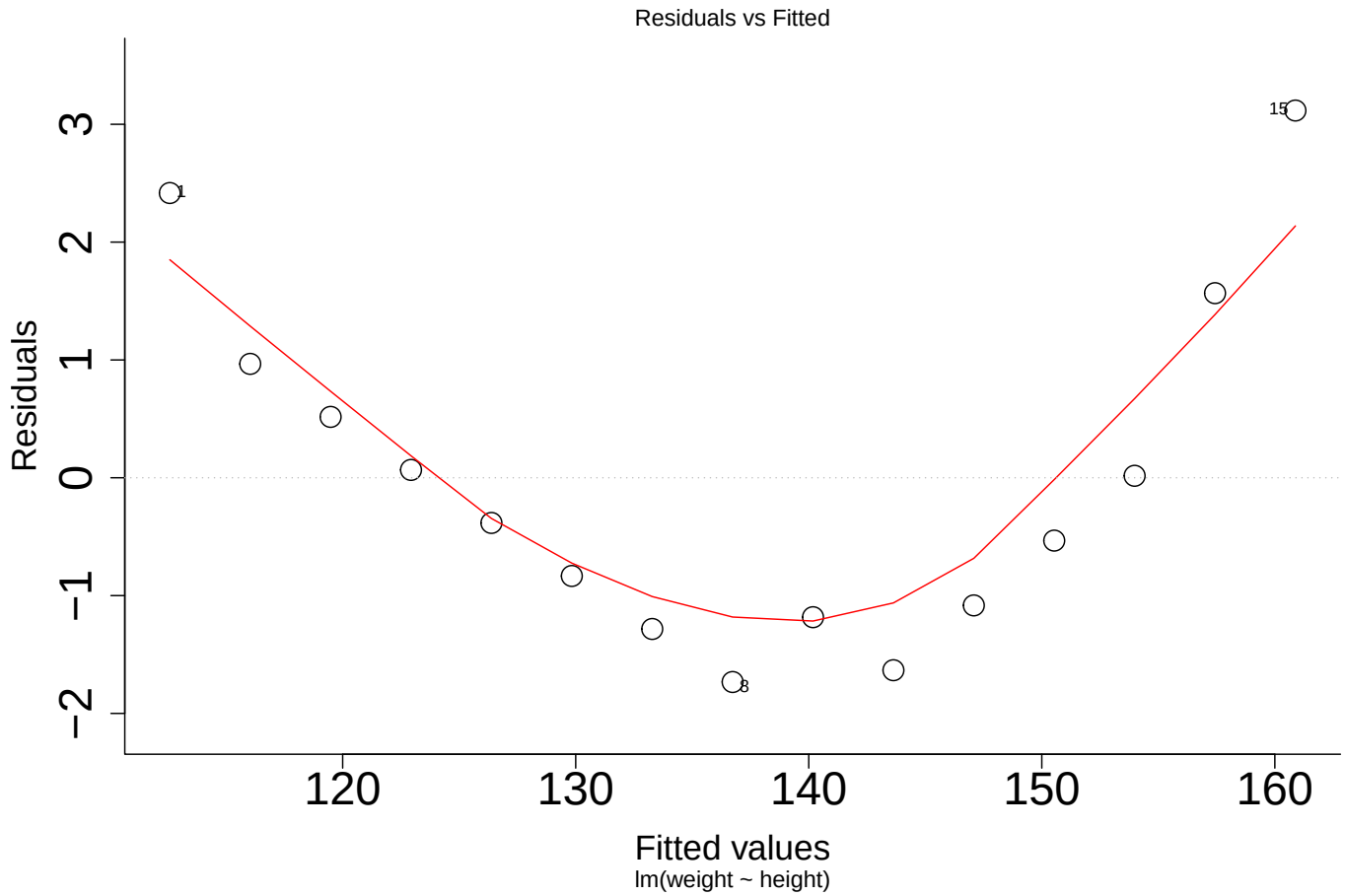

Figure 5: Modelldiagnostik auf Linearität für Gewicht vs Größe für Frauen

Linearität (2)

Im Diagramm Residuen vs. angepasste Werte sehen Sie deutliche Hinweise auf eine gekrümmte Beziehung, was darauf hindeutet, dass Sie der Regression möglicherweise einen quadratischen Term hinzufügen möchten.

Component plus residual plots

$$\text{Residuals} + \beta_i X_i \text{ versus } X_i$$

Die Nichtlinearität in einer dieser Darstellungen deutet darauf hin, dass Sie die funktionale Form dieses Prädiktors in der Regression möglicherweise nicht angemessen modelliert haben. In diesem Fall müssen Sie möglicherweise krummlinige Komponenten wie Polynome hinzufügen, eine oder mehrere Variablen transformieren (z. B. $\log(X)$ anstelle von X verwenden) oder die lineare Regression zugunsten einer anderen Regressionsvariante aufgeben.

Component plus residual plots (1)

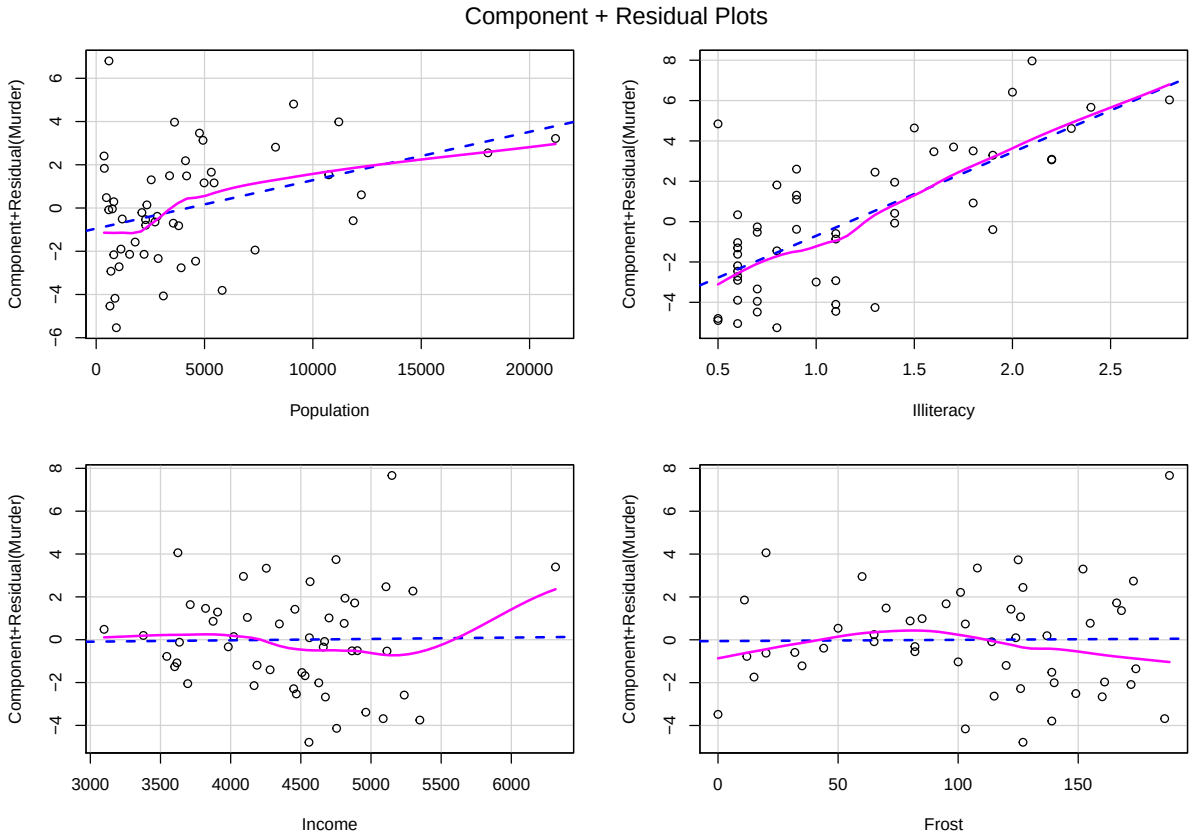

Figure 6: Modelldiagnostik auf Linearität

Homoskedastizität

Wenn Sie die Annahme einer konstanten Varianz erfüllt haben, sollten die Punkte im Diagramm "Skalenposition" (unten links) ein zufälliges Band um eine horizontale Linie sein. Sie scheinen diese Annahme zu erfüllen.

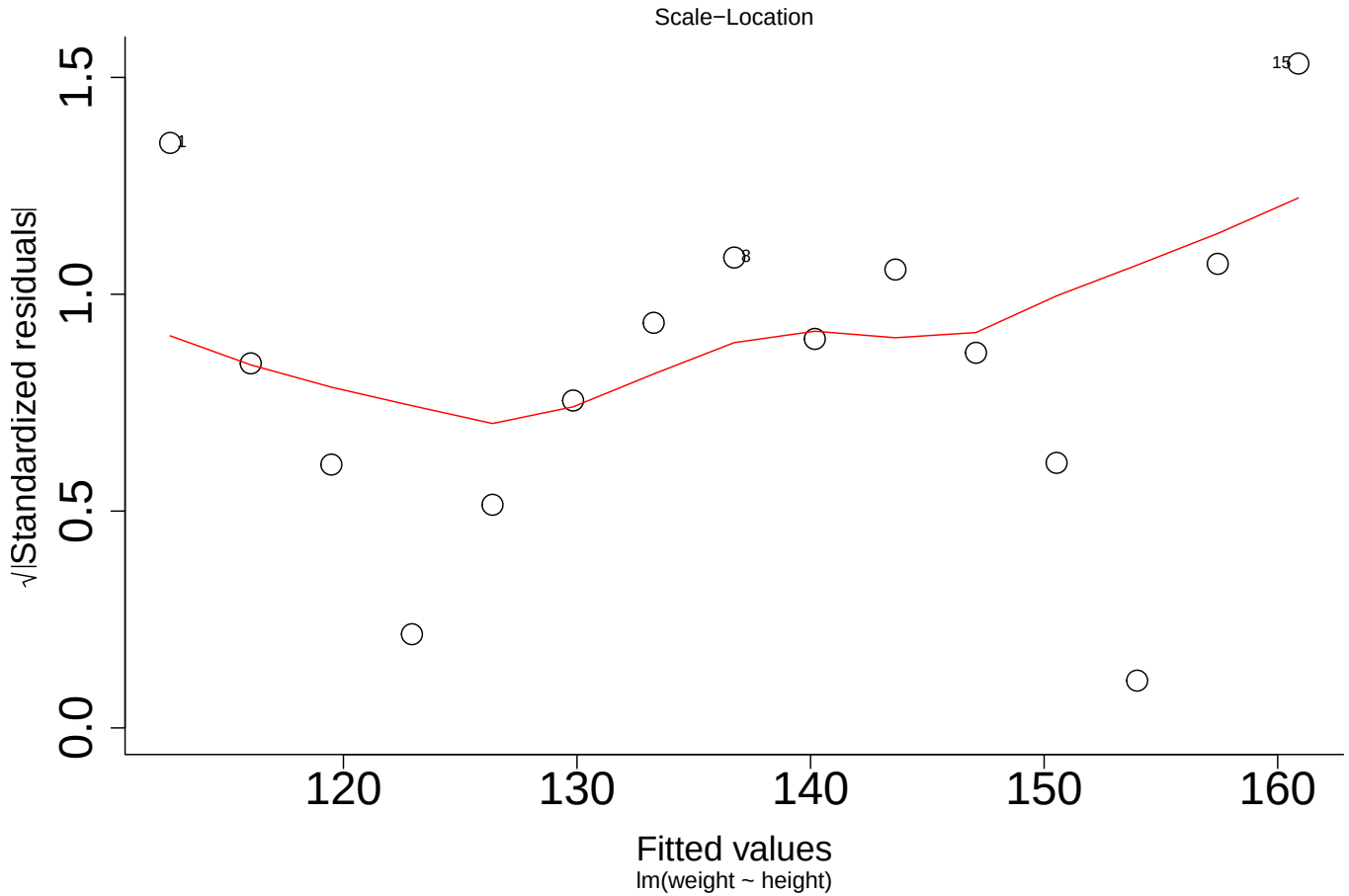

Figure 7: Modelldiagnostik auf Homoskedastizität für Gewicht vs Größe für Frauen

ncvTest()

Die Funktion ncvTest () testet die Hypothese der konstanten Fehlervarianz gegenüber der Alternative, dass sich die Fehlervarianz mit der Höhe der angepassten Werte ändert.

Ein signifikantes Ergebnis deutet auf eine Heteroskedastizität (nicht konstante Fehlervarianz) hin.

```
ncvTest(fit)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.8052115, Df = 1, p = 0.36954
```

Gesamtevaluation von den Annahmen des linearen Modells

```
library(gvlma)
gvmodel <- gvlma(fit)
summary(gvmodel)
```

```
##
## Call:
## lm(formula = weight ~ height, data = women)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7333 -1.1333 -0.3833  0.7417  3.1167
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
## height        3.45000    0.09114   37.85 1.09e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.525 on 13 degrees of freedom
## Multiple R-squared:  0.991,  Adjusted R-squared:  0.9903
## F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = fit)
##
##                       Value   p-value                   Decision
## Global Stat        16.5866 0.0023251 Assumptions NOT satisfied!
## Skewness            1.5577 0.2119999    Assumptions acceptable.
## Kurtosis            0.1019 0.7496131    Assumptions acceptable.
## Link Function      14.1218 0.0001713 Assumptions NOT satisfied!
## Heteroscedasticity  0.8052 0.3695398    Assumptions acceptable.
```

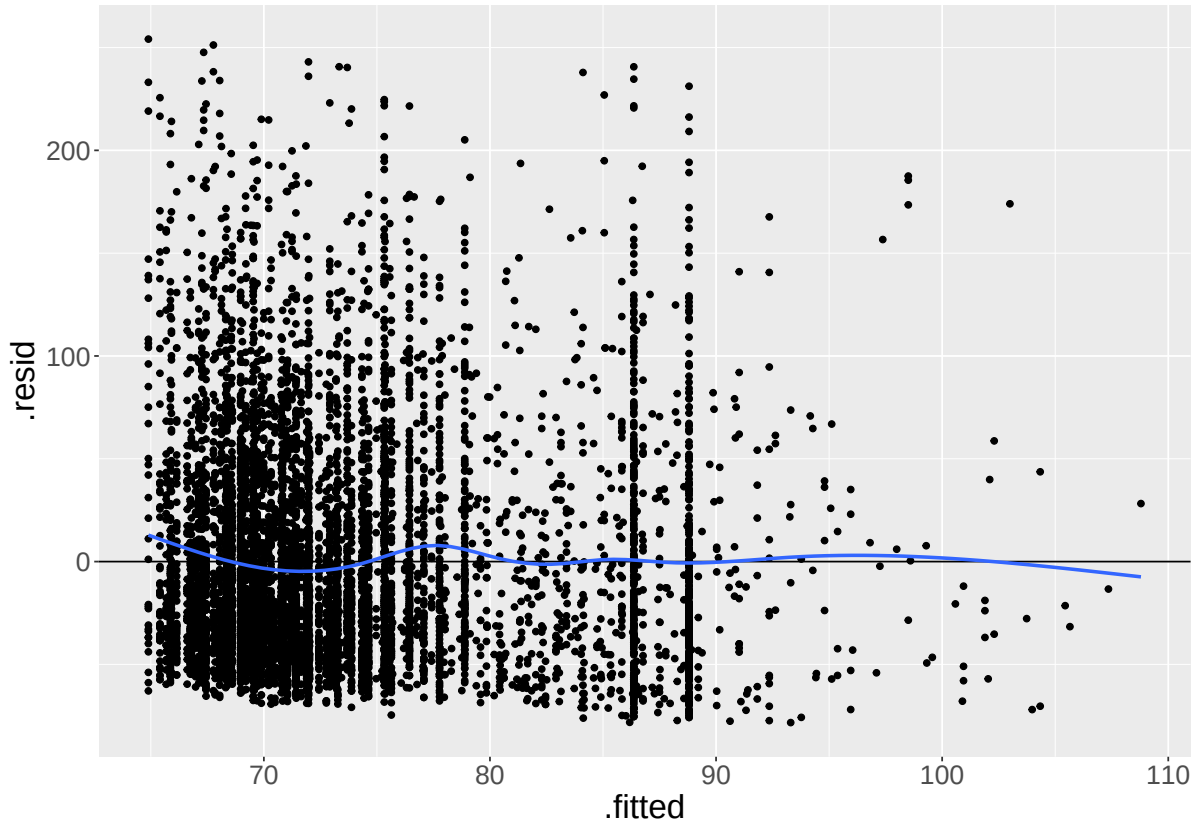

Figure 8: Evaluation of the homoscedasticity assumption for interval from waiting period to conception (wpc) regression.

Homoskedastizität (1)

```
library(car)
ncvTest(model1)
library(lmtest)
bptest(model1)
ggplot(model1, aes(.fitted, .resid)) +
  geom_hline(yintercept = 0) +
  geom_point() +
  geom_smooth(se = FALSE)
```

```
library(car)
ncvTest(model1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 18.21011, Df = 1, p = 1.9783e-05
```

```
library(lmtest)
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 19.601, df = 7, p-value = 0.006499
```

Multikollinearität

Multikollinearität besteht, wenn zwei Variable korreliert sind (z.B. Geburdatum und Alter)

Dies führt zu großen Konfidenzintervallen für Modellparameter und erschwert die Interpretation einzelner Koeffizienten.

Multikollinearität kann mithilfe einer Statistik namens *Varianzinflationsfaktor* (VIF) erfasst werden.

Für jede Prädiktorvariable gibt die Quadratwurzel des VIF den Grad an, in dem das Konfidenzintervall für den Regressionsparameter dieser Variablen relativ zu einem Modell mit unkorrelierten Prädiktoren (daher der Name) erweitert wird.

Multikollinearität - variance inflation factor

```
states <- as.data.frame(state.x77[,c("Murder", "Population",
                        "Illiteracy", "Income", "Frost")])
fit <- lm(Murder ~ Population + Illiteracy + Income + Frost, data=states)
vif(fit)
```

```
## Population Illiteracy     Income      Frost
##   1.245282   2.165848   1.345822   2.082547
```

```
sqrt(vif(fit)) > 2
```

```
## Population Illiteracy     Income      Frost
##      FALSE      FALSE      FALSE      FALSE
```

Für jede Prädiktorvariable gibt die Quadratwurzel des VIF den Grad an, in dem das Konfidenzintervall für den Regressionsparameter dieser Variablen relativ zu einem Modell mit unkorrelierten Prädiktoren (daher der Name) erweitert wird.

$\sqrt{\text{VIF}} > 2$ indicates a multicollinearity problem.

Ausreißer

- Ein *Ausreißer* ist eine Beobachtung, die vom angepassten Regressionsmodell nicht gut vorhergesagt wird (d.h. ein großes positives oder negatives Residuum aufweist). Faustregel: standardisierte Residuen, die größer als 2 oder kleiner als –2 sind.
- Eine Beobachtung mit einem hohen *Hebelwert* (*leverage*) weist eine ungewöhnliche Kombination von Prädikatorwerten auf. Das heißt, es ist ein Ausreißer im Prädiktorbereich. Der Wert der abhängigen Variablen wird nicht zur Berechnung des Hebels einer Beobachtung verwendet.
- Eine *einflussreiche* (*influential*) Beobachtung ist eine Beobachtung, die einen unverhältnismäßigen Einfluss auf die Bestimmung der Modellparameter hat. Einflussreiche Beobachtungen werden anhand einer Statistik identifiziert, die als *Cooks-Distanz* oder *Cooks D* bezeichnet wird.

Ausreißer (1)

Ausreißer (2)

```
fit1 <- lm(weight ~ height, data=women)
outlierTest(fit1)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 15 2.970125           0.011698      0.17548
```

Beispiel: ist polynomiale Regression besser?

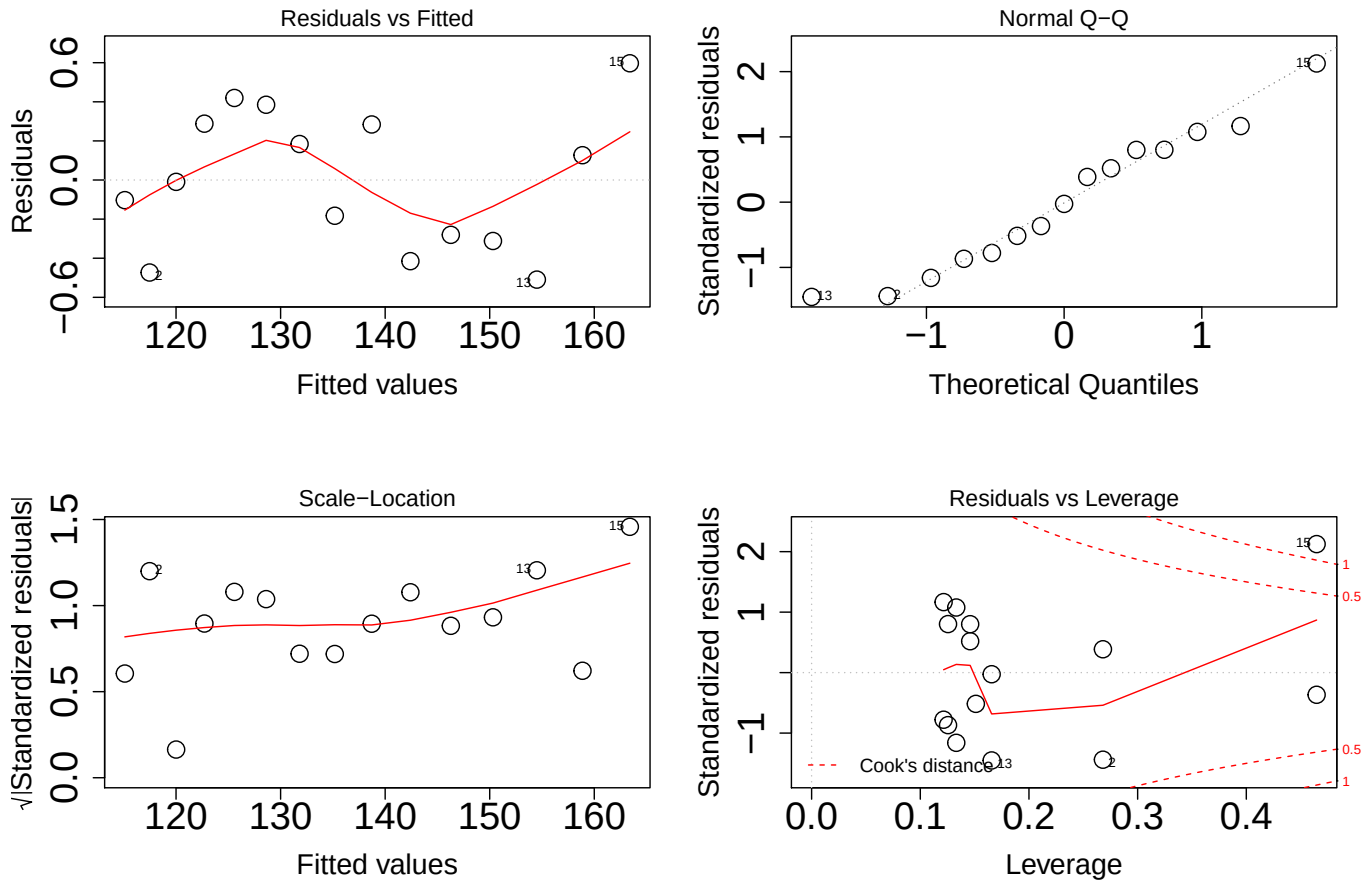

Figure 9: Modelldiagnostik für polynomiale Regression für Gewicht vs Größe für Frauen

Beispiel: Ausreißer rausnehmen.

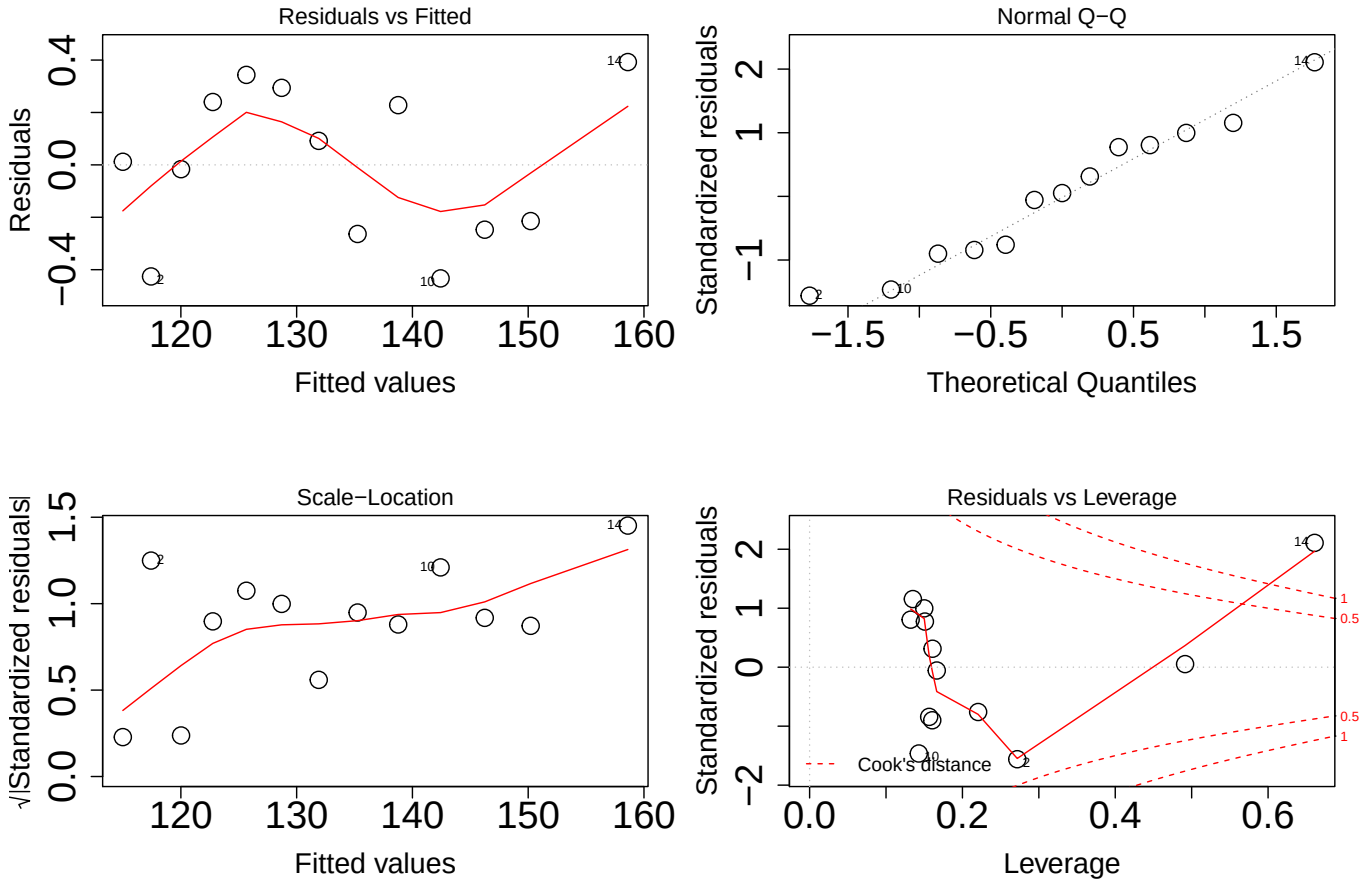

Figure 10: Modelldiagnostik für polynomiale Regression ohne 13. und 15. Messwerte für Gewicht vs Größe für Frauen

Man soll sehr vorsichtig beim Rausnehmen der Messwerte vorgehen. Das Modell soll an die Daten angepasst werden, nicht umgekehrt!

Beispiel 2

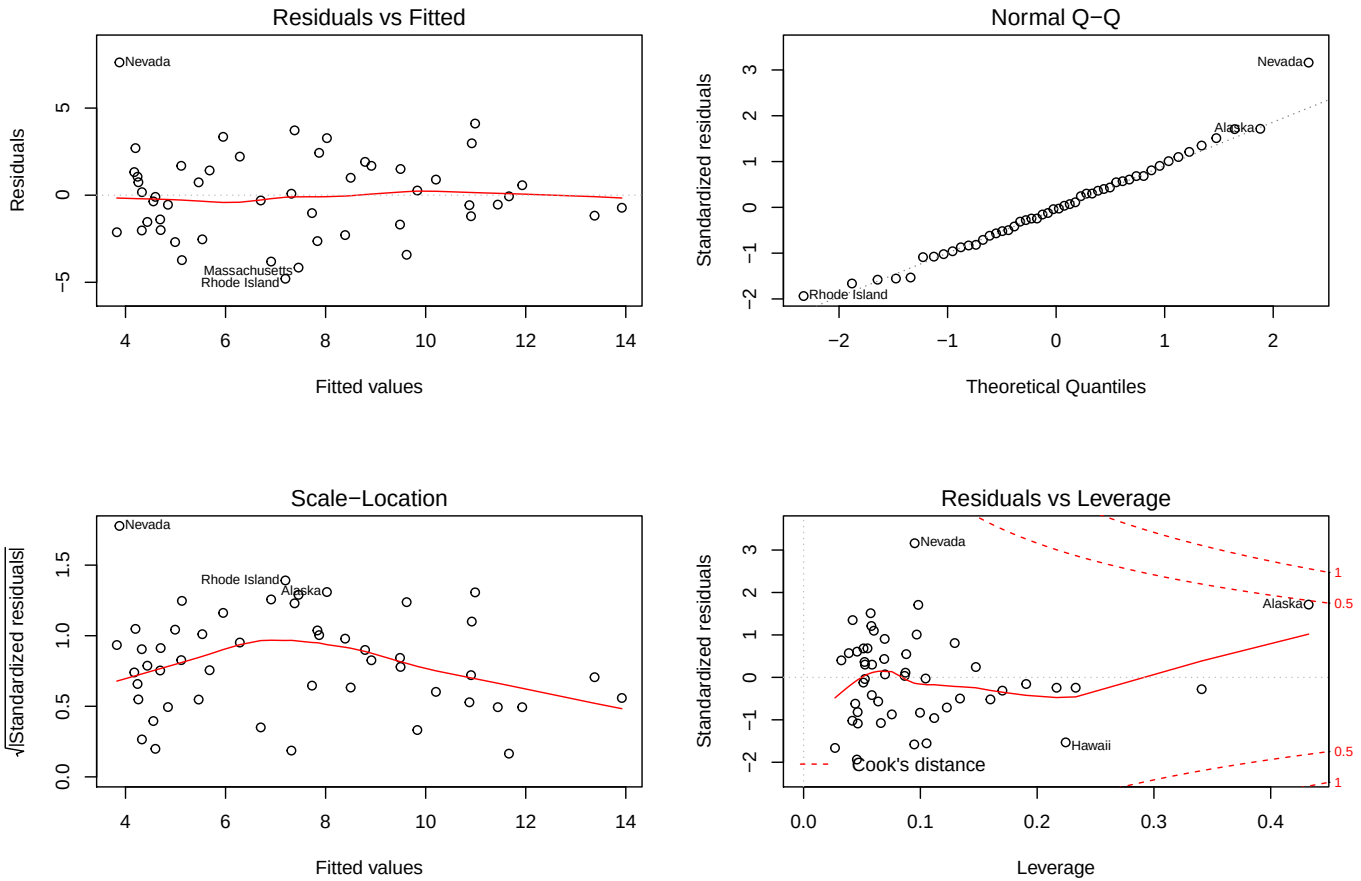

Figure 11: Modelldiagnostik für Tötungsrate

Hebelwert (Leverage)

Beobachtungen mit hoher Hebelwirkung werden durch die *hat-Statistik* identifiziert.

Für einen gegebenen Datensatz beträgt der durchschnittliche *hat*-Wert p/n, wobei p die Anzahl der im Modell geschätzten Parameter (einschließlich des Abschnitts) und n die Stichprobengröße ist.

Grob gesagt sollte eine Beobachtung mit einem Hutwert, der das Zwei- oder Dreifache des durchschnittlichen *hat*-Werts übersteigt, untersucht werden.

Beobachtungen mit hohem Hebel können einflussreiche Beobachtungen sein oder auch nicht. Das hängt davon ab, ob sie auch Ausreißer sind.

Hebelwert (Leverage) (1)

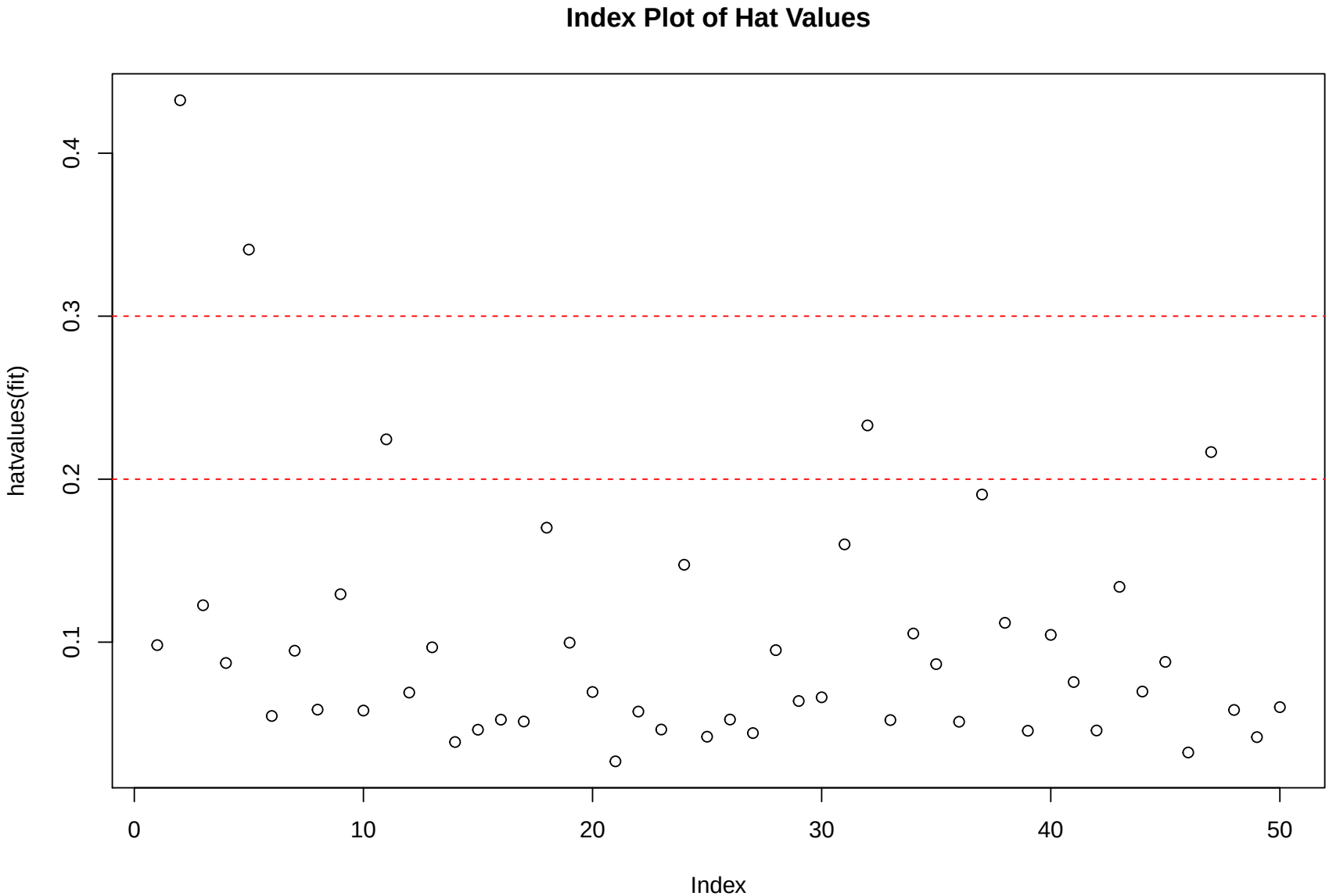

Figure 12: hat-Statistik.

Einflüßreicher Messwerte: Cooks D-Wert

Cooks D-Werte von mehr als $4/(n-k-1)$, wobei n die Stichprobengröße und k die Anzahl der Prädiktorvariablen ist, weisen auf einflussreiche Beobachtungen hin.

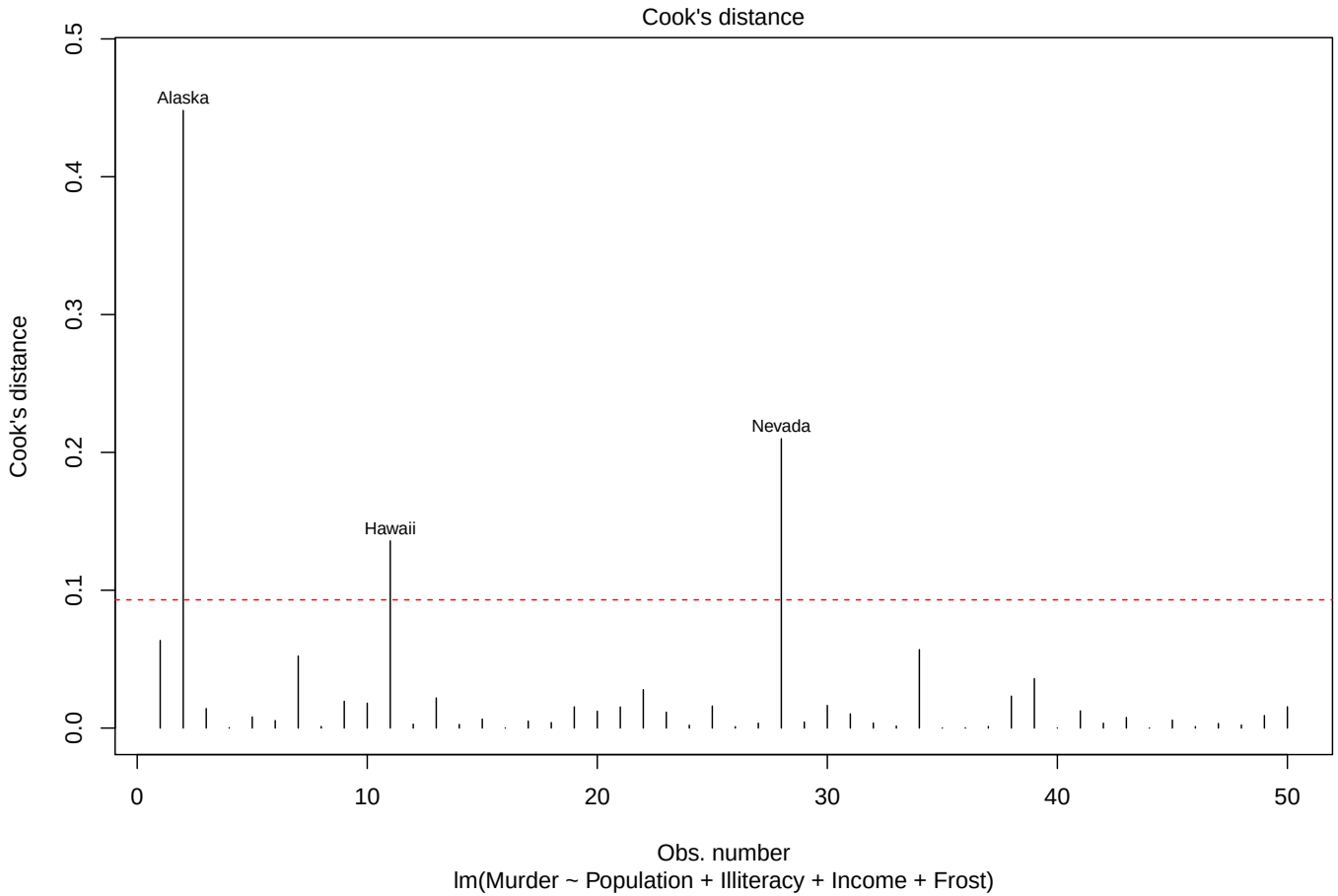

Figure 13: Cooks D-Werte.

Einflüßreicher Messwerte: Added-variable plots

Cooks D-Diagramme können helfen, einflussreiche Beobachtungen zu identifizieren, liefern jedoch keine Informationen darüber, wie sich diese Beobachtungen auf das Modell auswirken.

Für eine Antwortvariable und k Prädiktorvariablen erstellen man k *Added-variable plots* wie folgt.

Für jeden Prädiktor X_k die Residuen aus der Regression der Antwortvariablen auf den anderen $k-1$ Prädiktoren werden gegen die Residuen aus der Regression von X_k auf den anderen $k-1$ Prädiktoren dargestellt.

Einflüßreicher Messwerte: Added-variable plots (1)

```
avPlots(fit, ask=FALSE, id.method="identify")
```

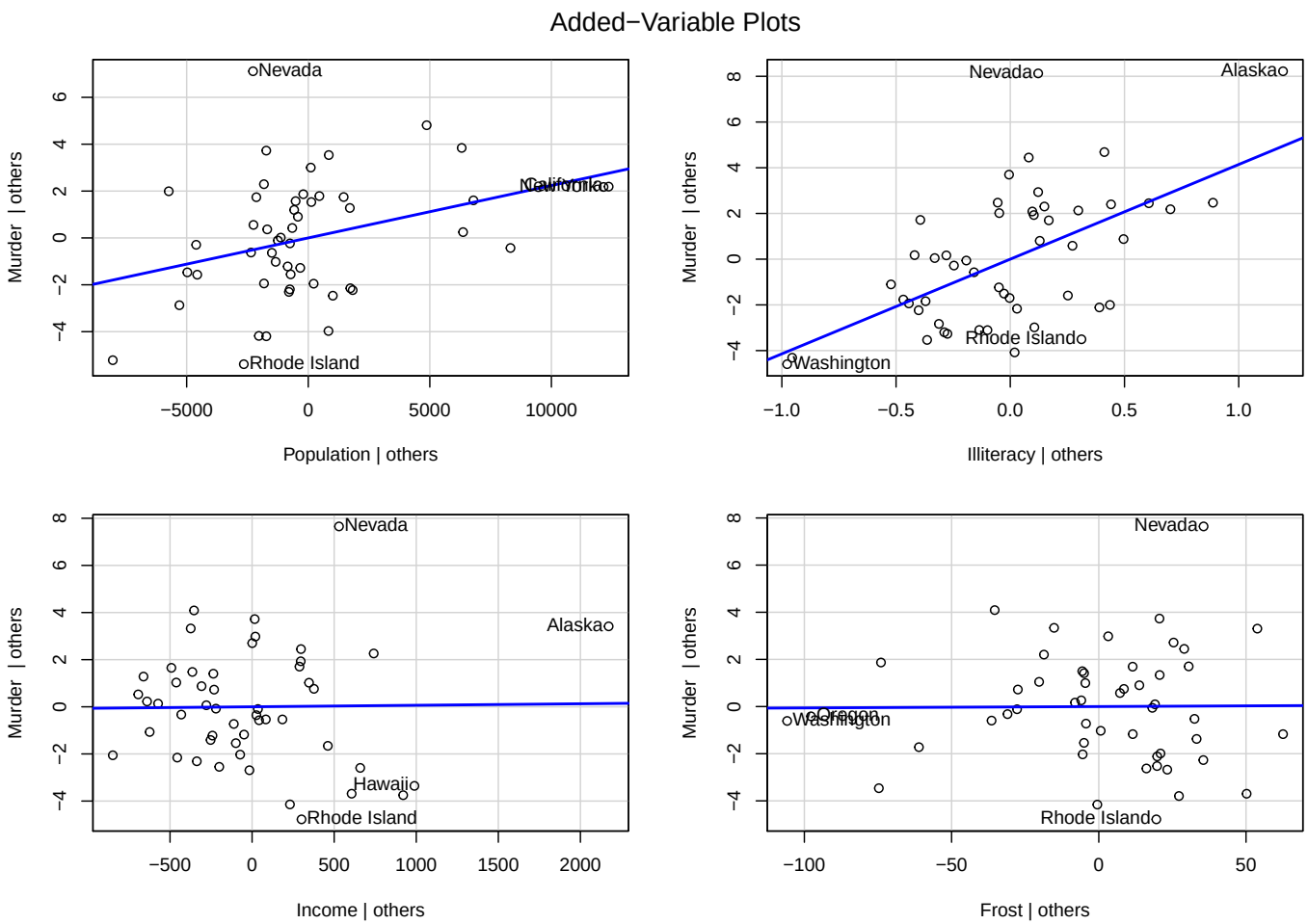

Figure 14: Added-variable plots zur Bewertung der Auswirkung einflussreicher Beobachtungen.

influencePlot()

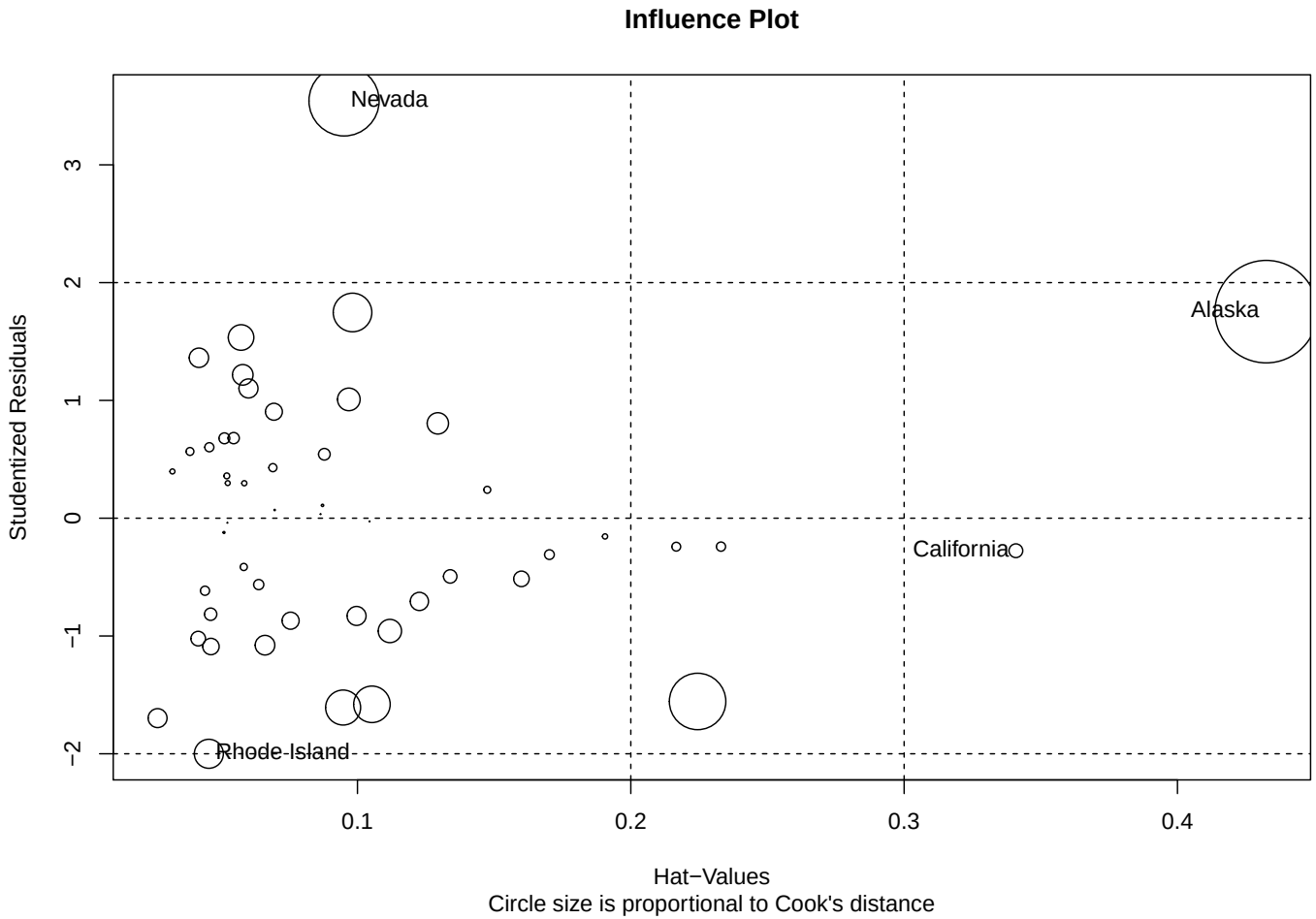

Figure 15: Outliers, leverage and influence zusammen.

influencePlot() (1)

```
library(car)
influencePlot(fit, id.method="identify", main="Influence Plot",
              sub="Circle size is proportional to Cook's distance")
```

```
##                  StudRes        Hat       CookD
## Alaska         1.7536917 0.43247319 0.448050997
## California    -0.2761492 0.34087628 0.008052956
## Nevada         3.5429286 0.09508977 0.209915743
## Rhode Island -2.0001631 0.04562377 0.035858963
#
```