Analyse des „Old Faithful Geyser“ Datensatzes

Knowledgedump.org — E. Keller

Inhaltsverzeichnis

1 Laden der Daten **3**

2 Old Faithful Geysir Daten **3**

3 Vorhersage der nächsten Eruption **7**

4 Modellalternativen **16**

5 Fazit **21**

6 Referenzen **21**

1 Laden der Daten

```
> rm(list=ls())                  #Loeschen von ggf. definierten Objekten
> library("datasets")          #Laden des Pakets mit "faithful" Datensatz
```

2 Old Faithful Geysir Daten

Die Daten zu den Ausbrüchen (Eruptionsdauer und Zeit zwischen den Eruptionen) des „Old Faithful Geysirs“ im Yellowstone National Park (Wyoming, USA) sind bei Statistikern beliebt, da eine Korrelation zwischen Ausbruchsdauer und Zeit zwischen den Eruptionen besteht, welche weder zu groß, noch zu klein ist. Mit dem Wissen über die vorherigen Ausbrüche können mehr oder weniger genaue Vorhersagen getroffen werden, wann der nächste Ausbruch stattfinden wird.
Infolgedessen existiert eine Vielzahl von Datensätzen, welche sich mit dem Geysir befassen.
In dieser Analyse betrachten wir ausschließlich den in dem R-Paket „datasets“ enthaltenen Datensatz „faithful“.

```
> faithful[1:10,]
```

```
   eruptions waiting
1      3.600      79
2      1.800      54
3      3.333      74
4      2.283      62
5      4.533      85
6      2.883      55
7      4.700      88
8      3.600      85
9      1.950      51
10     4.350      85
```

```
> str(faithful)
```

```
'data.frame':        272 obs. of  2 variables:
 $ eruptions: num  3.6 1.8 3.33 2.28 4.53 ...
 $ waiting  : num  79 54 74 62 85 55 88 85 51 85 ...
```

```
> summary(faithful)
```

```
   eruptions        waiting
 Min.   :1.600   Min.   :43.0
 1st Qu.:2.163   1st Qu.:58.0
 Median :4.000   Median :76.0
```

```
 Mean   :3.488   Mean   :70.9
 3rd Qu.:4.454   3rd Qu.:82.0
 Max.   :5.100   Max.   :96.0
```

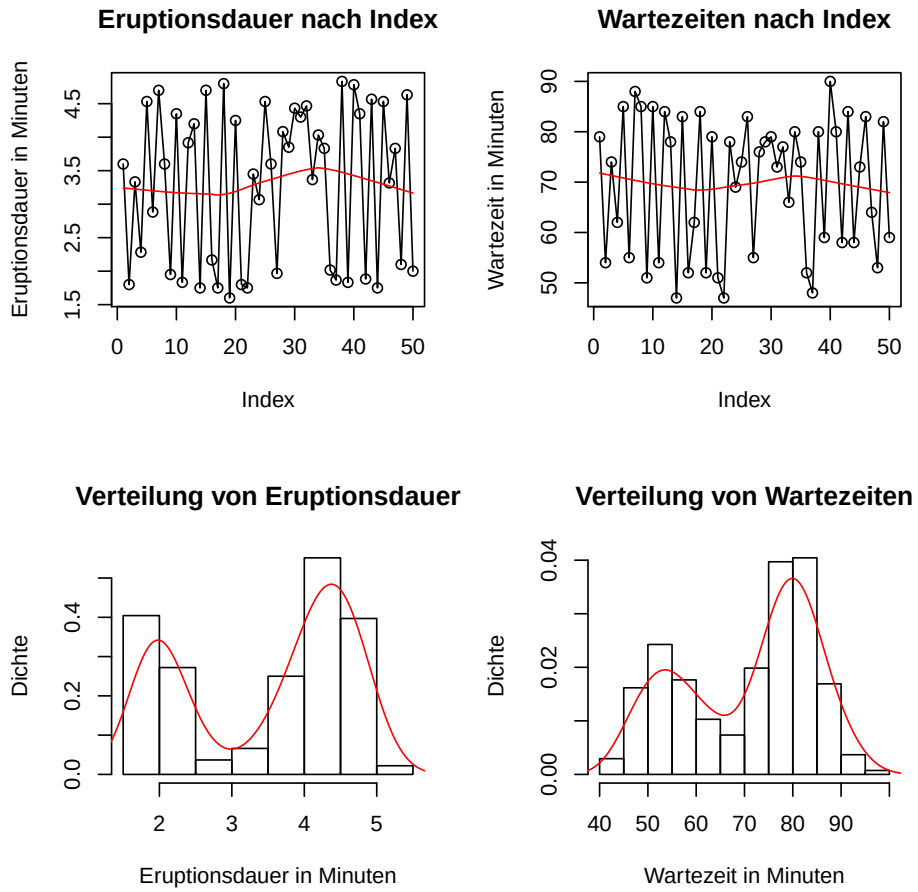
```
> help(faithful)
```

Man sieht, dass der Datensatz 2 Variablen mit numerischen Werten aufweist und 272 Objekte umfasst. Außerdem liegen die Eruptionszeiten zwischen 1,6 Minuten und 5,1 Minuten, mit einem Mittelwert von ca. 3,5 Minuten. Die Wartezeiten liegen zwischen 43 und 96 Minuten, wobei diese im Mittel ca. 71 Minuten betragen. Dies legt nahe, dass eine symmetrische Verteilung der Eruptions- und Wartezeiten besteht (vorausgesetzt die Daten sind tatsächlich chronologisch geordnet, siehe unten).
Durch Aufrufen der Hilfefunktion erhalten wir zusätzliche wichtige Informationen zu den Attributen des Datensatzes:

Variable	Name	Typ	Erklärung
[,1]	eruptions	numeric	Eruptionsdauer in Minuten
[,2]	waiting	numeric	Zeit zwischen dem Beginn zweier Eruptionen

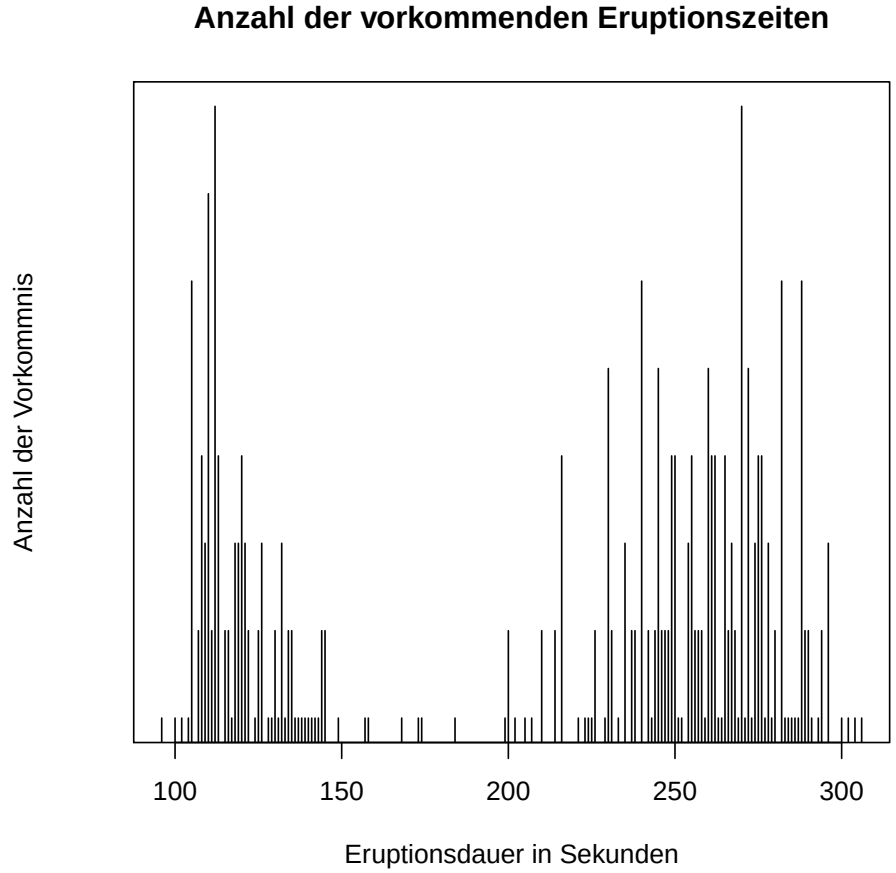
Nicht ersichtlich ist hierbei, ob die Werte konsekutive Eruptionen beschreiben, oder ob Lücken bestehen. Auch bei der Quelle der Daten ([1] W. Härdle, 1991, App. 3, S.201), wird dies nicht erläutert.
Aus diesem Grund kann der Einfluss von vorhergegangenen Eruptionen auf eine kommende Eruption nicht sicher untersucht werden, auch wenn es eine Korrelation zwischen den verschiedenen Eruptionsdauern und Wartezeiten zu geben scheint - welche auch logisch sinnvoll wäre, da meist auf kurze Eruptionen längere kommen und umgekehrt. Dies kann man sich anhand der folgenden plots und Dichteapproximationen vor Augen führen, welche auch die Bimodalität der Wartezeit/Eruptions-Verteilungen verdeutlichen.

```
> par(mfrow=c(2,2))
> plot(faithful[1:50,1], type="l", main="Eruptionsdauer nach Index",
+ xlab="Index", ylab="Eruptionsdauer in Minuten")
> panel.smooth(1:50,faithful[1:50,1])
> plot(faithful[1:50,2], type="l", main="Wartezeiten nach Index",
+ xlab="Index", ylab="Wartezeit in Minuten")
> panel.smooth(1:50,faithful[1:50,2])
> hist(faithful$eruptions, main = "Verteilung von Eruptionsdauer",
+ xlab = "Eruptionsdauer in Minuten", ylab = "Dichte", freq=F)
> lines(density(faithful$eruptions), col = "red")
> hist(faithful$waiting, main = "Verteilung von Wartezeiten",
+ xlab = "Wartezeit in Minuten", ylab = "Dichte", freq=F)
> lines(density(faithful$waiting), col = "red")
```



Die Hilfefunktion erwähnt zudem, dass vermutlich bei mehreren „eruptions“ Werten eine Rundung auf 5 Sekunden erfolgte.

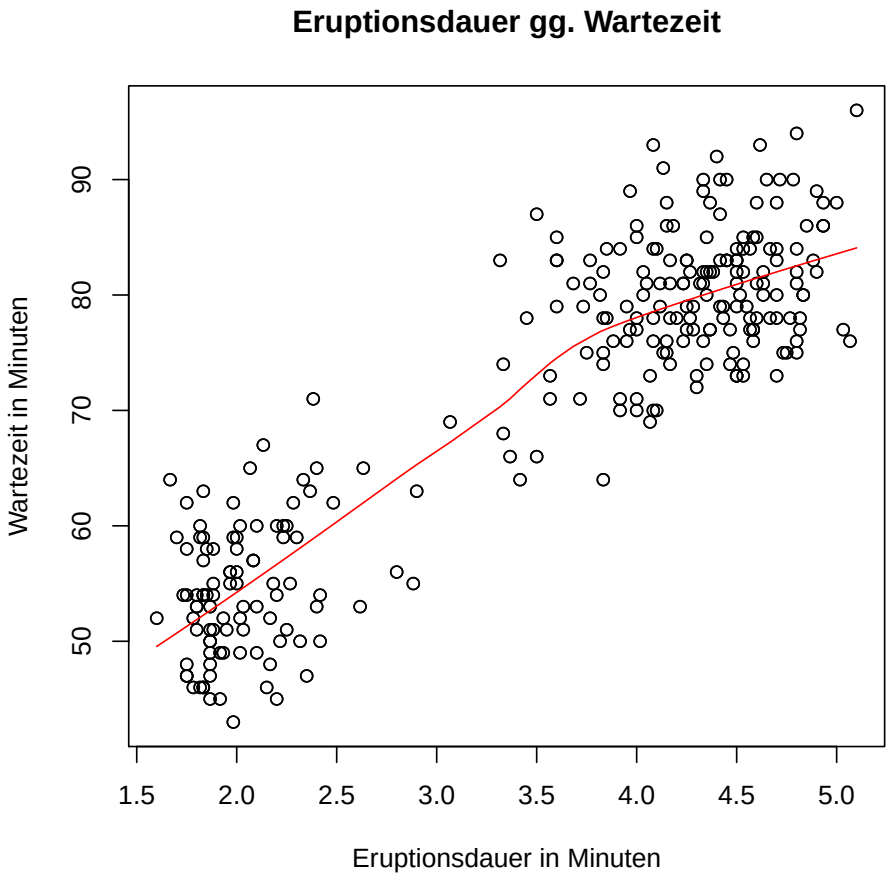
```
> par(mfrow=c(1,1))
> f_sek<-table(round(faithful[,1]*60))
> plot(names(f_sek), f_sek, type="h",
+ main="Anzahl der vorkommenden Eruptionszeiten",
+ xlab="Eruptionsdauer in Sekunden", ylab="Anzahl der Vorkommnis")
> f_sek[f_sek>=4]

105 108 110 112 113 120 216 230 240 245 249 250 255 260 261 262
  6   4   7   8   4   4   4   5   6   5   4   4   4   5   4   4
265 270 272 275 276 282 288
  4   8   5   4   4   6   6
```



Es scheint tatsächlich eine Rundung auf 5 Sekunden erfolgt zu sein. Die Wirkung auf den Datensatz, bzw. unsere kommende Schätzung ist bei der Stichprobengröße des Datensatzes und einer derart kleinen Rundung jedoch vernachlässigbar.

Um ein besseres Gefühl für die Daten zu bekommen, plotten wir die Eruptions- gegen die Wartezeiten und sehen, was man bereits in den Plots nach Datensatz-Index ahnen konnte: Die Warte-/Eruptionszeiten bilden zwei Cluster, welche eine merkliche Abgrenzung aufweisen - es handelt sich somit um eine bimodale Verteilung.
Die durchgeführte lokal gewichtete Polynom-Regression legt nahe, dass die Bereiche in etwa um eine Gerade, mit leichtem Knick beim Anfang des zweiten Bereichs, angeordnet sind.

```
> plot(faithful, main="Eruptionsdauer gg. Wartezeit",
+ xlab="Eruptionsdauer in Minuten", ylab="Wartezeit in Minuten")
> panel.smooth(faithful[,1],faithful[,2])  #lowess() mit default Werten
```



3 Vorhersage der nächsten Eruption

Anhand der vorhandenen Daten kann man nun versuchen, die nächste Eruption vorherzusagen. Früher geschah dies mit Hilfe einer linearen Gleichung durch die Parkwächter (vgl. [2] Cook & Weisberg, 1982, S.40 Beispiel 2.3.2):

$$\widehat{\text{Wartezeit}} = 30 + 10 * \text{Eruptionsdauer in Minuten}$$

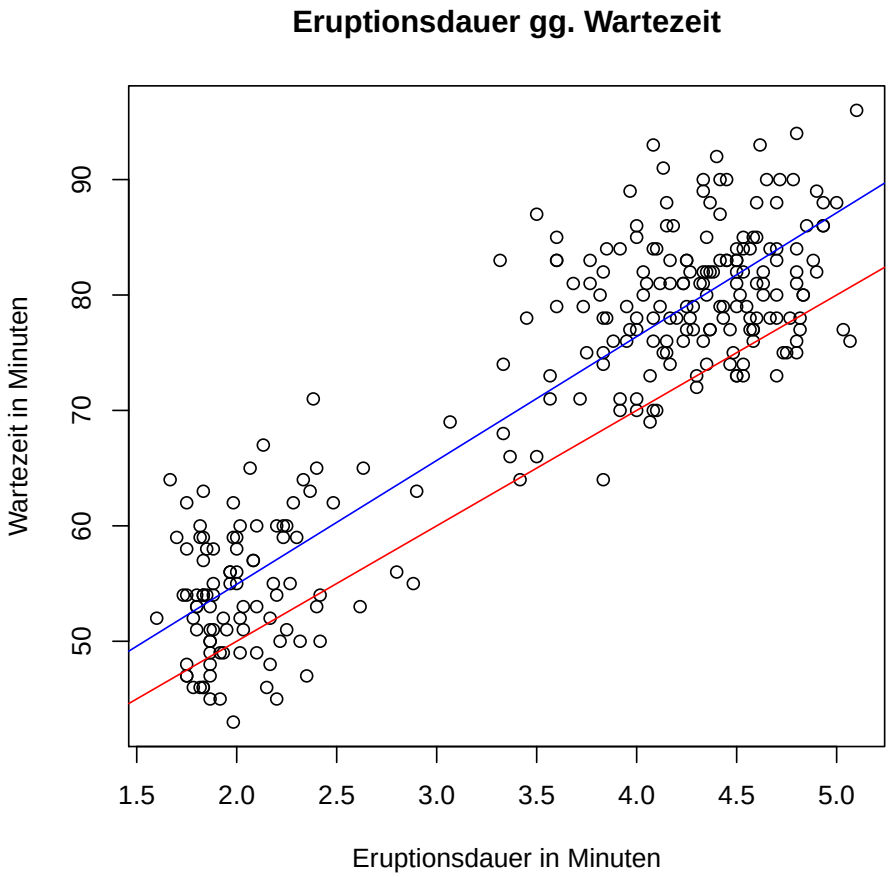
Diese musste aber angepasst werden, da sich aus geologischen Gründen die mittlere Wartezeit über die Jahre immer mehr vergrößert hat - es bestehen Vermutungen, dass diese Veränderungen durch Erdbeben hervorgerufen wurden.
Aber auch künftige und aktuelle Schätzmodelle orientieren sich an demselben Prinzip: Für eine kurze Eruption wird eine kurze Wartezeit vorhergesagt, während nach langen Eruptionen eine lange Wartezeit vorhergesagt wird. So findet sich auf der Internetseite des Nationalparks folgende einfache Vorhersageformel:

C Length of Eruption	D Interval Until Next Eruption
less than 3 minutes	68 minutes
more than 3 minutes	94 minutes

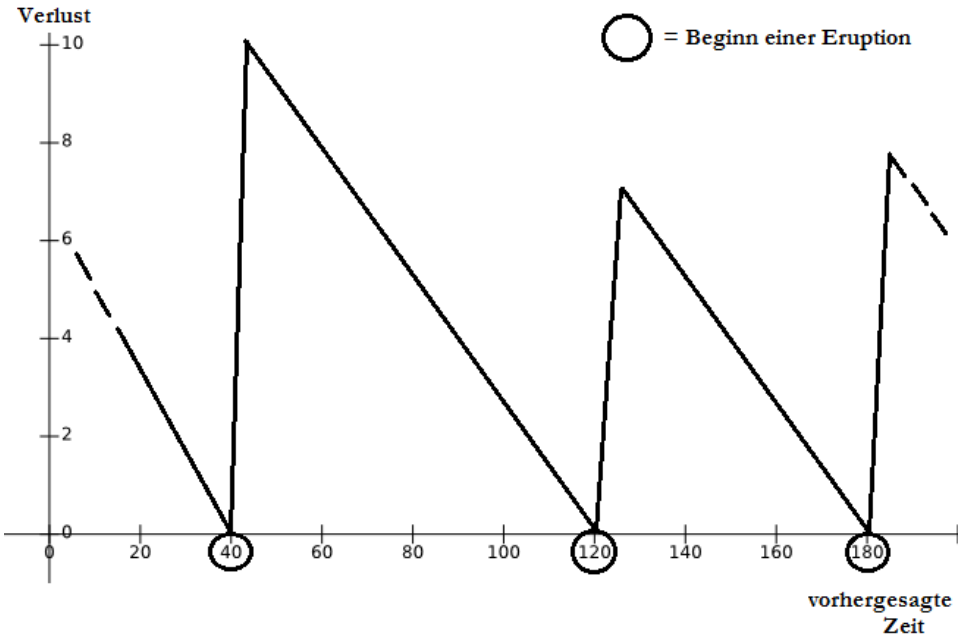
Quelle: http://www.nps.gov/yell/learn/kidsyouth/predict-old-faithful.htm

Wir wollen nun mit Hilfe des Datensatzes ein alternatives Vorhersagemodell für die alte lineare Gleichung aufstellen, wobei wir ein lineares Regressionsmodell mit der Wilkinson-Rodgers Notation $waiting \sim eruptions$ anwenden. Im Anschluss vergleichen wir diese Modelle.

```
> lmfaithful<-lm(faithful$waiting ~ faithful$eruptions, data=faithful)
> lmfaithful

Call:
lm(formula = faithful$waiting ~ faithful$eruptions, data = faithful)

Coefficients:
       (Intercept)  faithful$eruptions
             33.47               10.73

> plot(faithful, main="Eruptionsdauer gg. Wartezeit",
+ xlab="Eruptionsdauer in Minuten", ylab="Wartezeit in Minuten")
> abline(lmfaithful, col="blue")
> abline(30,10 ,col="red")
> pred1<-faithful$eruptions*10.73+33.47        # =lmfaithful$fitted
> pred2<-faithful$eruptions*10+30
> summary(pred1)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  50.64   56.68   76.39   70.89   81.26   88.19

> summary(pred2)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  46.00   51.63   70.00   64.88   74.54   81.00

> summary(faithful$waiting)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   43.0    58.0    76.0    70.9    82.0    96.0
```



Schon eine oberflächliche Betrachtung zeigt, dass das sehr einfache, alte Schätzmodell für unseren Datensatz keine zufriedenstellenden Vorhersagen liefert und im Mittel zu niedrige Wartezeiten vorhersagt. Dies bedeutet jedoch nicht, dass es zwangsweise eine schlechtere Modellwahl für die Parkwächter ist, da für sie die Besucherzufriedenheit im Mittelpunkt steht. Die Verlustfunktion ist abhängig von dem Ziel der Schätzung - will man die Eruption möglichst genau vorhersagen, bietet sich beispielsweise eine quadratische Verlustfunktion an. Will man jedoch die Wartezeit der Besucher gering halten, so könnte die Verlustfunktion beispielsweise wie folgt aussehen:



In diesem Fall wäre das Modell der Parkwächter überlegen, da das lineare Regressionsmodell viel zu oft zu lange Wartezeiten vorhersagt, weswegen in etwa die Hälfte der Eruptionen verpasst werden würden. Wird der Eruptionsbeginn um 1 Minute verpasst, sehen wir dies noch als akzeptabel an.

```
> spaet1<-subset(pred1-faithful$waiting, pred1-faithful$waiting>1)
> spaet2<-subset(pred2-faithful$waiting, pred2-faithful$waiting>1)
> spaet1

  [1]  9.40459  3.39350  5.24750  4.72191  1.78400  5.24750
  [7]  8.10909  1.28059  2.03609  6.60900  4.40091  3.59791
 [13]  3.11241  5.50291  5.32809  9.10909 10.59809  3.00300
 [19]  1.18209  9.97400  5.32809  2.86459  2.35341  5.47391
 [25]  4.24750  2.97400  5.90100 10.90100  2.75500  5.39000
 [31] 11.83891  4.47391  2.81709  7.46300  4.10891  1.93741
 [37]  8.72191  3.50291  7.15641  7.60900  2.50291  4.04700
 [43]  7.00300  6.50291  9.25509  1.86441  2.18209  8.33141
 [49]  8.55041  8.10891  2.07250  9.03941  2.79491  8.50291
 [55]  7.51400  7.13809  2.67459  2.28409  8.15641  3.96309
 [61]  3.18209  6.11241 10.47409  1.82800  6.39000 12.07600
 [67]  5.02500  6.64559  2.21109  5.03941  5.64559  1.23309
 [73]  8.75500  9.40441  4.18191  5.64559  4.28409  3.03609
 [79]  7.13809  1.89359  3.97400  8.97400  3.31718  6.61250
 [85]  5.54691  4.82800  6.60159  3.32791  5.21109  3.90100
 [91]  4.50291  6.13441  2.89009  6.22200  1.99950  3.50291
 [97]  6.57259  1.25491  5.49941  3.29150  7.28059  5.40441
[103]  7.25841  2.42659  2.99950 11.68550  1.58700  6.14550
```

```
[109]  3.07600  8.75500  4.49941  6.61991 11.74759  9.43750
[115] 10.53950  6.96641  7.40091
```

```
> spaet2
```

```
 [1] 3.83 1.33 2.33 4.33 3.00 4.00 4.67 3.67 2.00 1.67 2.33 3.17
[13] 3.17 1.67 4.17 3.67 2.00 2.33 1.17 1.17 3.33 7.00 2.00 4.17
[25] 2.33 2.00 1.50 1.83 2.17 6.50 2.00 6.83 2.50 5.50 2.17
```

Wir sehen, dass die Besucher mit dem Parkwächter-Modell viel seltener eine Eruption verpassen - nur in etwa 13% der Eruptionsanfänge werden um mehr als eine Minute verpasst. Mit dem linearen Modell läge dieser Wert bei ca. 43%. Somit liegt die Vermutung nahe, dass das Modell absichtlich zu niedrige Wartezeiten vorhersagt, damit weniger Besucher die Eruptionen verpassen. Für das genaue Vorhersagen der Eruptionen ist es jedoch nicht gut geeignet.
Im Folgenden untersuchen wir die Genauigkeit des linearen Regressionsmodells.

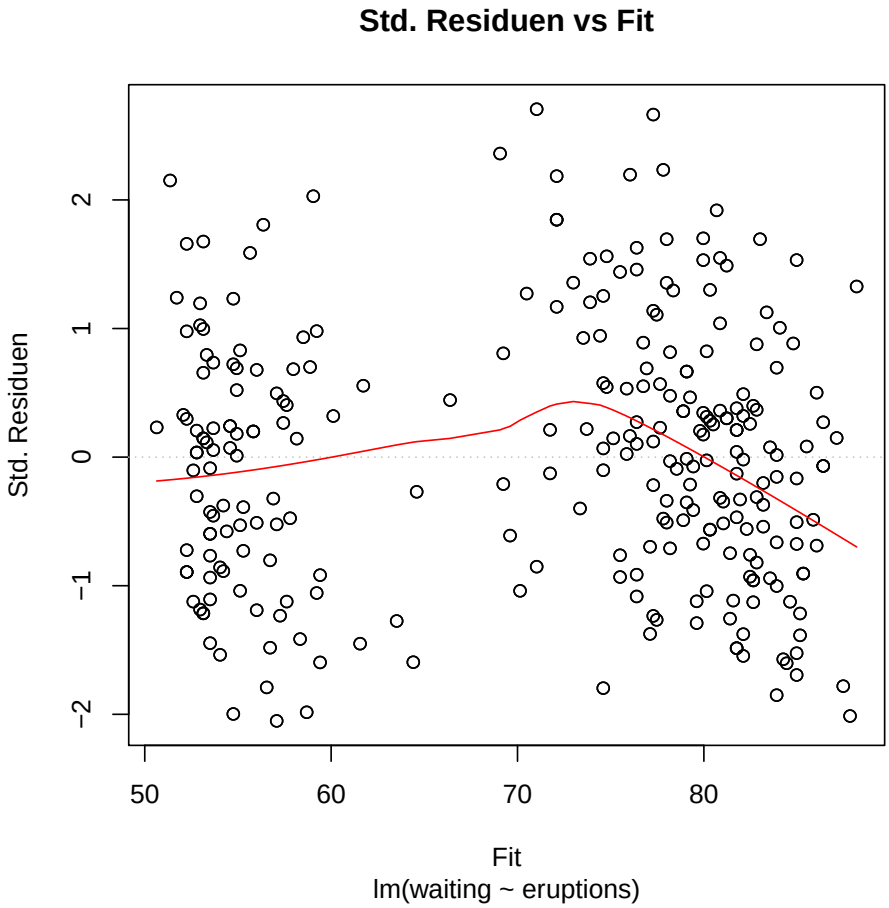
```
> summary(lmfaithful)
```

```
Call:
lm(formula = faithful$waiting ~ faithful$eruptions, data = faithful)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0796  -4.4831   0.2122   3.9246  15.9719

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         33.4744     1.1549   28.98   <2e-16 ***
faithful$eruptions  10.7296     0.3148   34.09   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.914 on 270 degrees of freedom
Multiple R-squared:  0.8115,        Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

```
> plot(lmfaithful$fitted, stdres(lmfaithful),
+ sub = "lm(waiting ~ eruptions)", main = "Std. Residuen vs Fit",
+ xlab = "Fit", ylab = "Std. Residuen")
> abline(h = 0, lty = 3, col = "gray")
> panel.smooth(lmfaithful$fitted,stdres(lmfaithful))
> str(subset(faithful$eruptions, faithful$eruptions>=3.5))
```

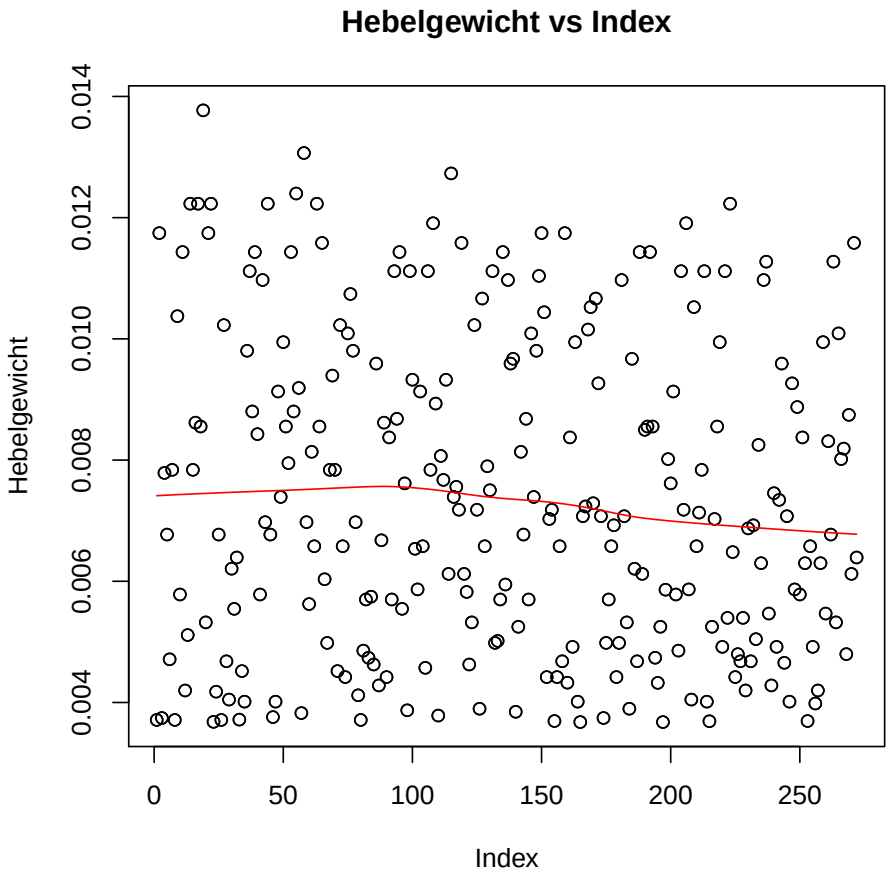
```
 num [1:168] 3.6 4.53 4.7 3.6 4.35 ...
```

```
> str(subset(faithful$eruptions, faithful$eruptions<3.5))

 num [1:104] 1.8 3.33 2.28 2.88 1.95 ...
```



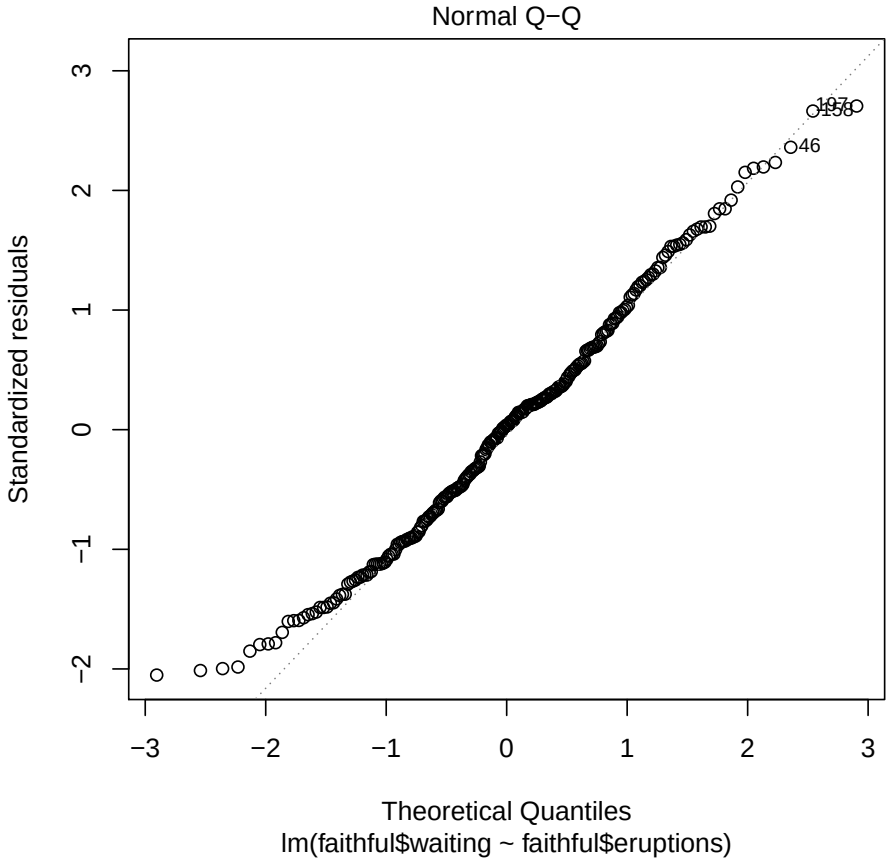
Durch die Zusammenfassung erhalten wir zunächst einige allgemeine Informationen. So ist ersichtlich, dass die Residuen fast symmetrisch um 0 verteilt sind und keine zu großen Abweichungen, gegeben der geschätzten Standardabweichung von ca. 6 (Minuten), aufweisen.
Der angepasste R^2-Wert drückt aus, wie viel Prozent von der beobachteten Varianz der Messdaten durch das lineare Modell erklärt wird. Bei unserem Modell liegt er bei 81%, was ein hinreichend guter Wert ist.
Anhand des plots sehen wir, dass die standardisierten Residuen relativ symmetrisch um den Nullpunkt liegen und keine offensichtlichen Ausreißer existieren. Der Abwärtstrend der lokal gewichteten Polynom-Regressionskurve könnte dabei darin begründet sein, dass sich in der zweiten Häufung der Eruptionslängen mehr Datenpunkte befinden.

```
> plot(hatvalues(lmfaithful), main = "Hebelgewicht vs Index",
+ ylab = "Hebelgewicht", xlab = "Index")
> panel.smooth(1:272,hatvalues(lmfaithful))
```



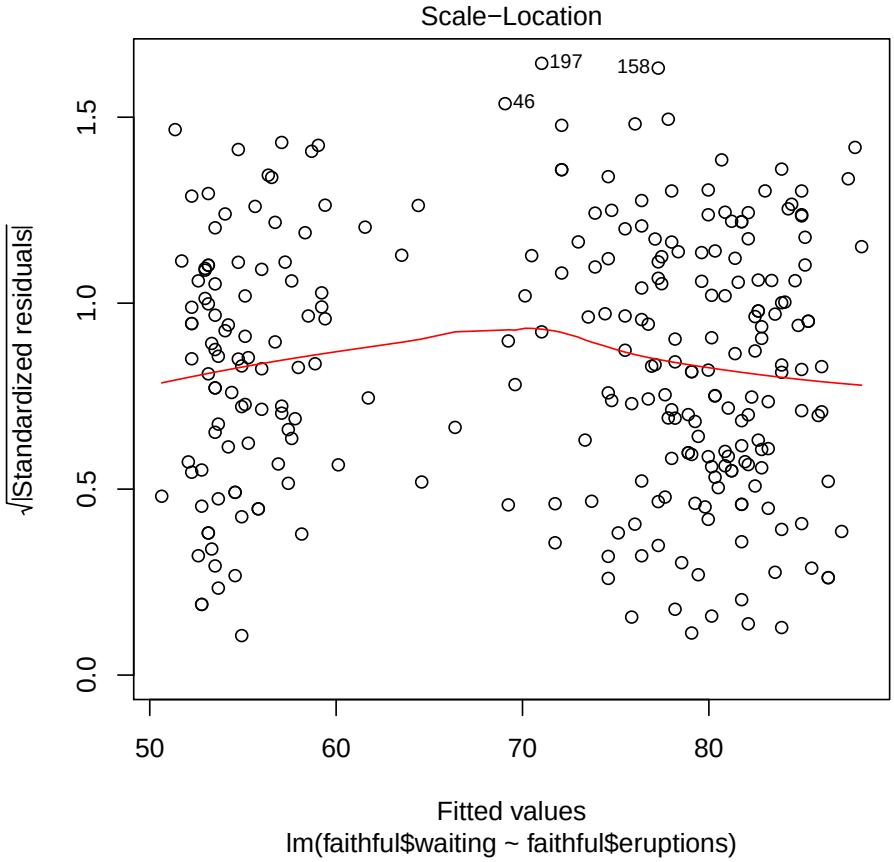
Da keiner der Datenwerte ein überdurchschnittliches Residuum oder Hebelgewicht besitzt, wird das Modell nicht durch Ausreißer beeinflusst. Dies ließ sich bereits an den ersten Scatterplots erahnen, auf denen auch keine besonderen Punkte erkennbar waren. Um zu überprüfen ob der Fehlerterm ϵ normalverteilt ist, betrachten wir nun den QQ-Plot.

```
> plot(lmfaithful, which=2)
```



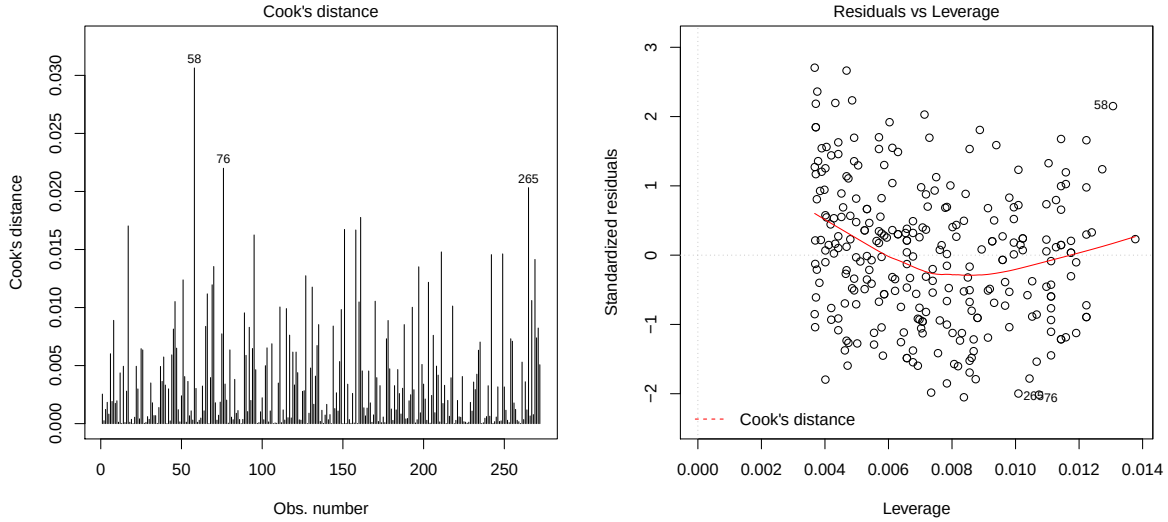
Der QQ-Plot zeichnet die Quantile der Normalverteilung gegen die der standardisierten Residuen. Da diese für unser Modell in etwa auf der gewünschten QQ-Linie (qqline(stdres(lmfaithful))) liegen und nur im unteren „Tail“ von der Geraden abweichen, kann man davon ausgehen, dass die Normalverteilungshypothese wahrscheinlich zutrifft oder wenigstens nicht grob verletzt wird.

```
> plot(lmfaithful, which=3)
```



Der Scale-Location plot liefert uns Auskunft über die beobachtete Varianz der Residuen bzgl. der entsprechenden Fit-Werte. Es wird ersichtlich, dass die Standardabweichung zur Mitte hin ansteigt und dann wieder absinkt. Da es sich hierbei aber nur um eine Fluktuation von $> 0,5$ bzgl. den Std. Residuen handelt und dies auch an der mangelnden Anzahl von Datenwerten mit einem Fit von ~ 65 liegen kann, können wir davon ausgehen, dass die Daten zumindest annähernd homoskedastisch verteilt sind - d.h., dass die Varianz annähernd konstant bleibt.

```
> par(mfrow=c(1,2))
> plot(lmfaithful, which=4:5)
```



In den letzten beiden Analyseplots werden wir nochmals bestätigt, dass der Datensatz nicht übermäßig durch irgendwelche Ausreißer beeinflusst wird. Die maximale beobachtete Cooks-Distanz liegt bei knapp über $0,03$ - wobei ein Datenpunkt in der Regel erst bei Werten um 1 oder höher als Ausreißer in Betracht gezogen wird.

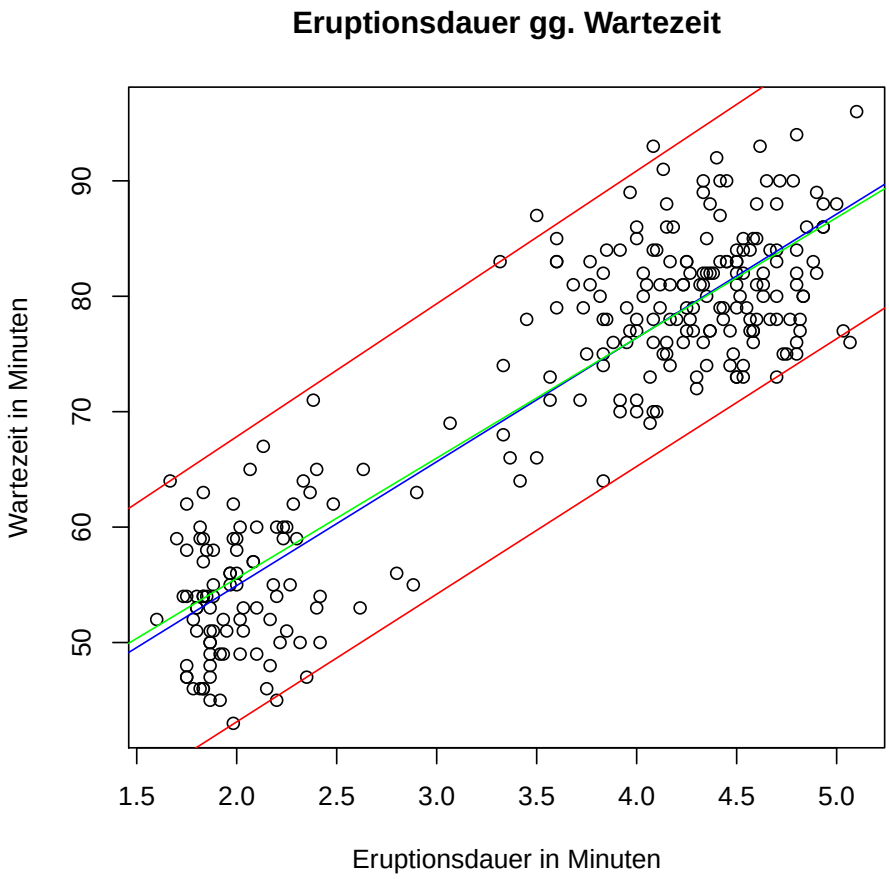
Wir sehen, dass das lineare Regressionsmodell in unserem Fall eine zufriedenstellende Lösung zum Vorhersageproblem darstellt. Im Folgenden werden wir noch einige alternativen Herangehensweisen prüfen, um zu sehen, ob sich ein besseres Modell finden lässt.

4 Modellalternativen

Eine Alternative zum OLS-Schätzer („ordinary least squares“) stellt der Quantilregressions-Schätzer dar. Während ersterer Ergebnisse liefert, welche den bedingten Erwartungswert der Response unter gegebenen Daten approximieren, versucht zweiterer ein bedingtes Quantil (z.B. den Median) anzunähern.
Als Schätzer ist für uns somit hauptsächlich ein Modell interessant, welches sich am Median orientiert, während die anderen Quantil-Schätzer zur Analyse von gegebenen Daten verwendet werden können (z.B. Erkennung von Ausreißern).

```
> if (!require("quantreg"))
+ install.packages("quantreg",
+ repos="http://cran.us.r-project.org", dependencies=TRUE)
> library(quantreg)
> rq1<-rq(waiting~eruptions, data=faithful, tau=0.01)
> rq50<-rq(waiting~eruptions, data=faithful, tau=0.5)
> rq99<-rq(waiting~eruptions, data=faithful, tau=0.99)
> plot(faithful, main="Eruptionsdauer gg. Wartezeit",
+ xlab="Eruptionsdauer in Minuten", ylab="Wartezeit in Minuten")
> abline(lmfaithful, col="blue")
```

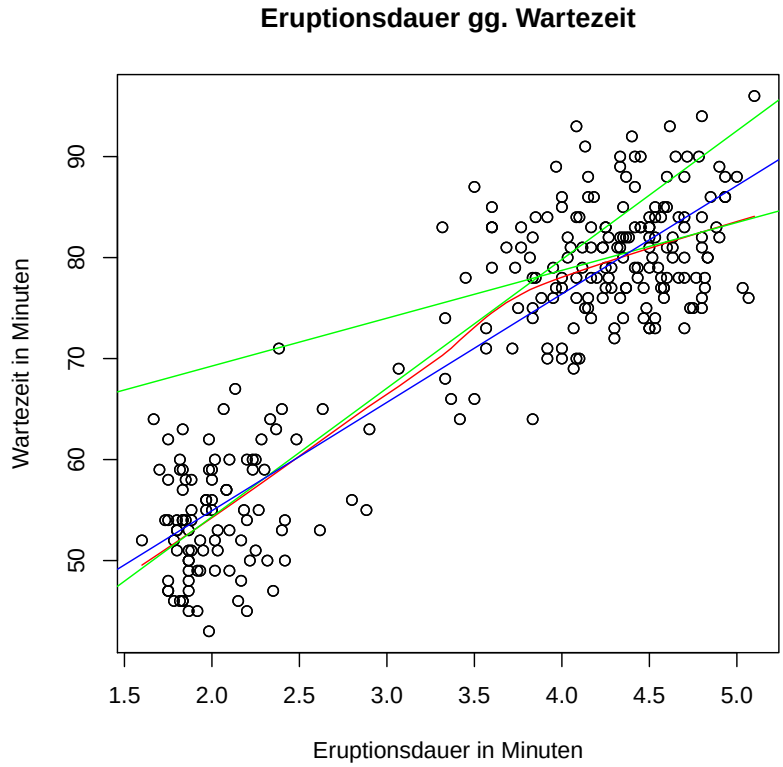
```
> abline(rq1, col="red")
> abline(rq99, col="red")
> abline(rq50, col="green")
```



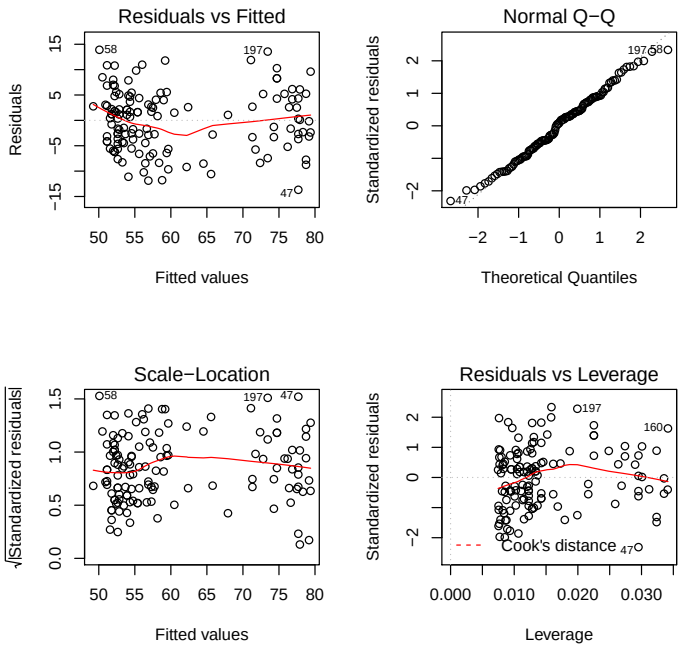
Wir sehen, dass der Medianschätzer kein signifikant verschiedenes Ergebnis zum linearen Regressionsschätzer liefert.
Zuletzt möchten wir untersuchen, ob ein Modell bestehend aus 2 linearen Regressionsgeraden - jeweils eines für jeden Cluster - auch gute Ergebnisse hervorbringt.
Der Nutzen dieser Methode hängt maßgeblich von den verwendeten Untergruppen ab. So liefert eine Aufteilung in Eruptionslängen > 3 und ≤ 3 ein komplett unbrauchbares Ergebnis, während eine Aufteilung in > 4 und ≤ 4 der lokal gewichteten Polynomregression sehr nahe kommt. Der Schnitt liegt in etwa bei einer Eruptionslänge von 3,85. Somit müsste das erste lineare Modell bei Eruptionslängen kleiner 3,85 und das zweite für Werte darüber verwendet werden.

```
> plot(faithful, main="Eruptionsdauer gg. Wartezeit",
+ xlab="Eruptionsdauer in Minuten", ylab="Wartezeit in Minuten")
> panel.smooth(faithful[,1],faithful[,2])#lowess() mit default Werten
> f1<-subset(faithful,faithful[,1]<4)
```

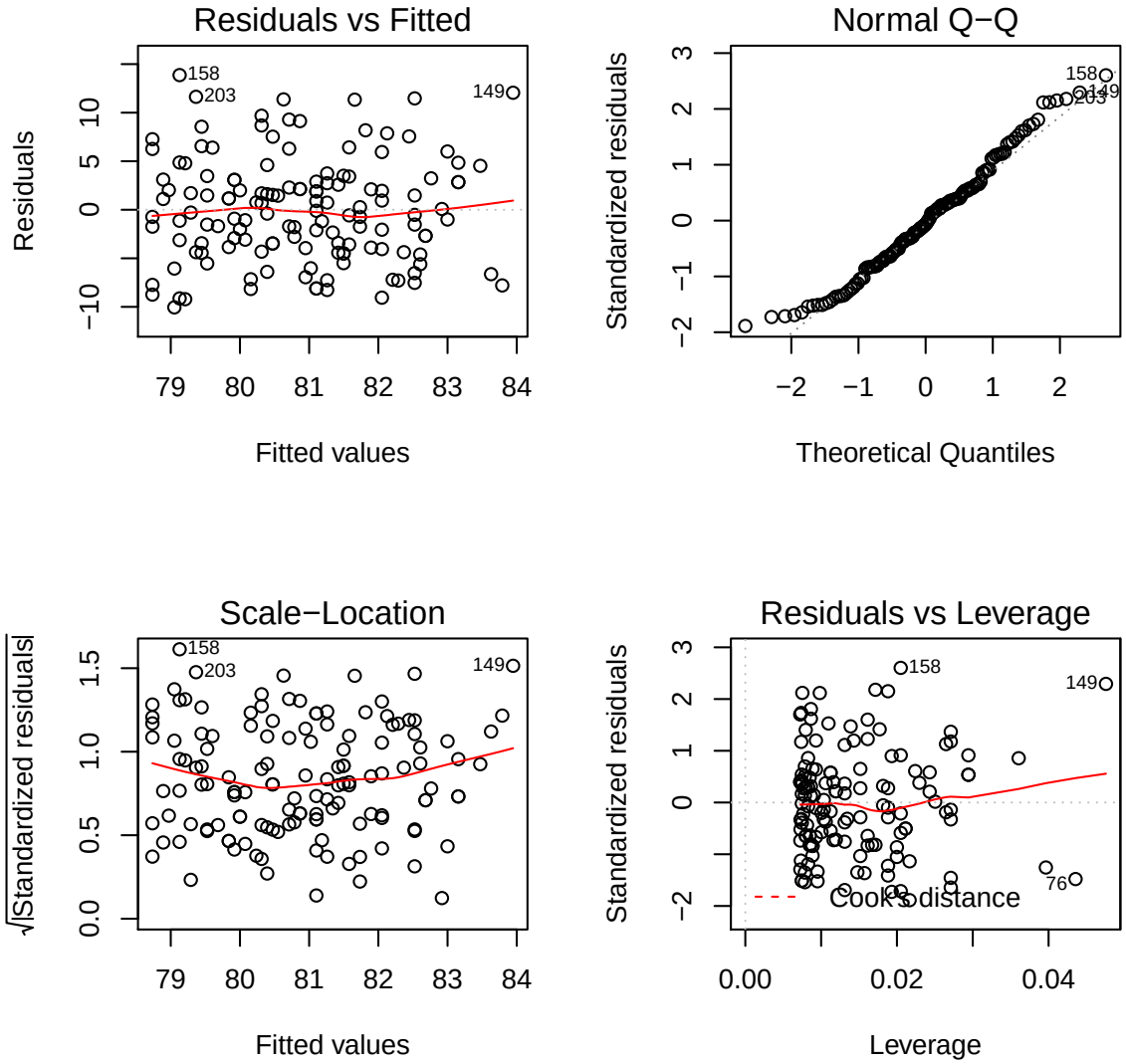
```
> f2<-subset(faithful,faithful[,1]>=4)
> lmf1<-lm(waiting~eruptions, data=f1)
> lmf2<-lm(waiting~eruptions, data=f2)
> abline(lmf1, col="green")
> abline(lmf2, col="green")
> abline(lmfaithful, col="blue")
> summary(lmf2)

Call:
lm(formula = waiting ~ eruptions, data = f2)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0525  -3.9345  -0.3414   3.0804  13.8716

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   59.784      7.464   8.010  4.6e-13 ***
eruptions      4.738      1.673   2.832  0.00533 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.385 on 136 degrees of freedom
Multiple R-squared:  0.0557,        Adjusted R-squared:  0.04875
F-statistic: 8.021 on 1 and 136 DF,  p-value: 0.005326

> summary(lmf1)

Call:
lm(formula = waiting ~ eruptions, data = f1)

Residuals:
     Min       1Q   Median       3Q      Max
-13.6831  -4.2391   0.5538   4.2537  13.9142

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.8463     1.7385   16.59   <2e-16 ***
eruptions    12.7411     0.6655   19.14   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.004 on 132 degrees of freedom
Multiple R-squared:  0.7352,        Adjusted R-squared:  0.7332
F-statistic: 366.5 on 1 and 132 DF,  p-value: < 2.2e-16
```




```
> par(mfrow=c(2,2))
> plot(lmf1)
```



```
> plot(lmf2)
```



Man sieht, dass die resultierenden Geraden (grün) der „lowess“ Kurve (rot) sehr ähnlich sind. Eine oberflächliche Analyse zeigt, dass vor allem bei der zweiten Regressionsgerade der Adjusted R^2 Wert sehr klein ist. Dieser ist bei einer Aufteilung der Datenpunkte jedoch kein zweckmäßiges Mittel zur Bestimmung der Regressionsgüte. Im Allgemeinen scheinen die Analyseplots darauf hinzudeuten, dass auch dieses Modell eine gute Schätzung liefert, wobei höchstens der zweite QQ-Plot auf mögliche Unstimmigkeiten bei der Fehlerverteilung hinweist.
Somit wäre dieses Regressionsmodell eine Alternative zum einfachen linearen Modell, welches außerdem den „Knick“ in der Verteilung der Datenpunkte ab dem zweiten Cluster miteinbezieht.

5 Fazit

Der faithful-Datensatz weist eine bimodale Verteilung mit relativ großer Varianz auf. Will man bei gegebener Eruptionslänge die Wartezeit bis zur nächsten Eruption schätzen, stellt die lineare Regression ein verlässliches und relativ einfaches Mittel dar. Bei einer quadratischen Verlustfunktion ist diese eine gute Wahl - bei anderen Verlustfunktionen, bzw. praktischer Anwendung, könnte hingegen ein alternatives Modell bessere Resultate erzielen.
Eine weitere Untersuchung des Datensatzes könnte hier ansetzen und versuchen, beispielsweise die beste Schätzung für den Fall zu finden, dass die Wahrscheinlichkeit die Wartezeit zu überschätzen gering gehalten werden soll.
Ließe sich in Erfahrung bringen, ob die Datenwerte konsekutive Eruptionen beschreiben, könnte man außerdem weitere Variablen in das Modell einbauen, indem man auch vorangegangene Eruptionen bei der Schätzung einbezieht.

6 Referenzen

[1] W. Härdle (1991); Smoothing Techniques with Implementation in S; New York: Springer Verlag
[2] Cook, R. D., & Weisberg, S. (1982); Residuals and Influence in Regression; Chapman and Hall.