Datenanalyse

Abgabe 2

Maximilian Hagn (Matr. Nr. 11808237)

20.05.2020

1. Beispiel

Praxisbeispiel

Gehen sie zur Seite https://robjhyndman.com/hyndsight/time-series-data-in-r/ . Hier werden einige R Pakete gelistet, die Zeitreihendaten beinhalten. Nach Installation eines Paketes kann man mit data(package="packagename") sehen, welche Daten in dem Paket vorhanden sind. Wählen sie einen Datensatz aus, bei dem monatliche Zeitreihen vorliegen (nehmen sie potentiell einen anderen Datensatz als ihre Kolleginnen und Kollegen). Sie können alternativ auch einen Datensatz aus dem Internet wählen, diesen z.B. als csv-Datei speichern, und mit read.csv() in R einlesen. Das resultierende Objekt (z.B. "obj") soll dann als Zeitreihen-Objekt dargestellt werden, was mit dem Befehl ts.obj <- ts(obj,start=...,frequency=12) geschieht - für start muss der Beginn der Zeitreihe angegeben werden. Falls sie keine monatlichen Werte haben, müsste frequency entsprechend geändert werden.

Stellen sie die Zeitreihe grafisch dar. Achten sie dabei auf eine korrekte Achsenbeschriftung. In diese Grafik sollen nun zwei mit LOWESS geglättete Kurven mit unterschiedlicher Farbe eingezeichnet werden. Verwenden sie dazu den Befehl loess und wählen sie einen kleinen und einen großen Wert für den Parameter span. Zeichnen sie für eine der Schätzungen die Glättung der Residuen ("upper und lower smoothing") in der gleichen Farbe, aber mit anderem Strichtyp ein.

Daten laden

```
library(astsa)
data = polio
ts.obj = ts(data, start=c(1970,1), end=c(1984,12), frequency=12)
tm = seq(1970, 1984.999, by=1/12)
```

Lowess Schätzungen erzeugen

```
res <- loess(ts.obj ~ tm)
res02 <- loess(ts.obj ~ tm, span=0.2)
res06 <- loess(ts.obj ~ tm, span=0.6)
```

Lowess Glättungen erzeugen

```
pred02 <- predict(res02, tm)
pred06 <- predict(res06, tm)
ts.obj1 <- ts(pred02,start=c(1970,1), end=c(1984,12),frequency=12)
ts.obj2 <- ts(pred06,start=c(1970,1), end=c(1984),frequency=12)
```

Upper / Lower Smoothing

```
r <- res$residuals-res06$residuals
r <- r*10

resupper <- lowess(ts.obj[sign(r)==1], r[sign(r)==1], f=0.6)$y
rp <- ts(resupper,start=c(1970,1), end=c(1984,12),frequency=12)

reslower <- lowess(ts.obj[sign(r)==-1], r[sign(r)==-1], f=0.6)$y
rm <- ts(reslower,start=c(1970,1), end=c(1984,12),frequency=12)

upperSmoothing <- ts.obj2+resupper
lowerSmoothing <- ts.obj2+reslower
```

Abbildung zeichnen

```
plot(res, main="Poliomyelitis cases in US", xlab="Years",
     ylab="Cases in the US", ylim=c(-1,5))
lines(ts.obj1, col="red")
lines(ts.obj2, col="green")
lines(upperSmoothing, col="blue", lty="dashed")
lines(lowerSmoothing, col="blue", lty="dashed")
```

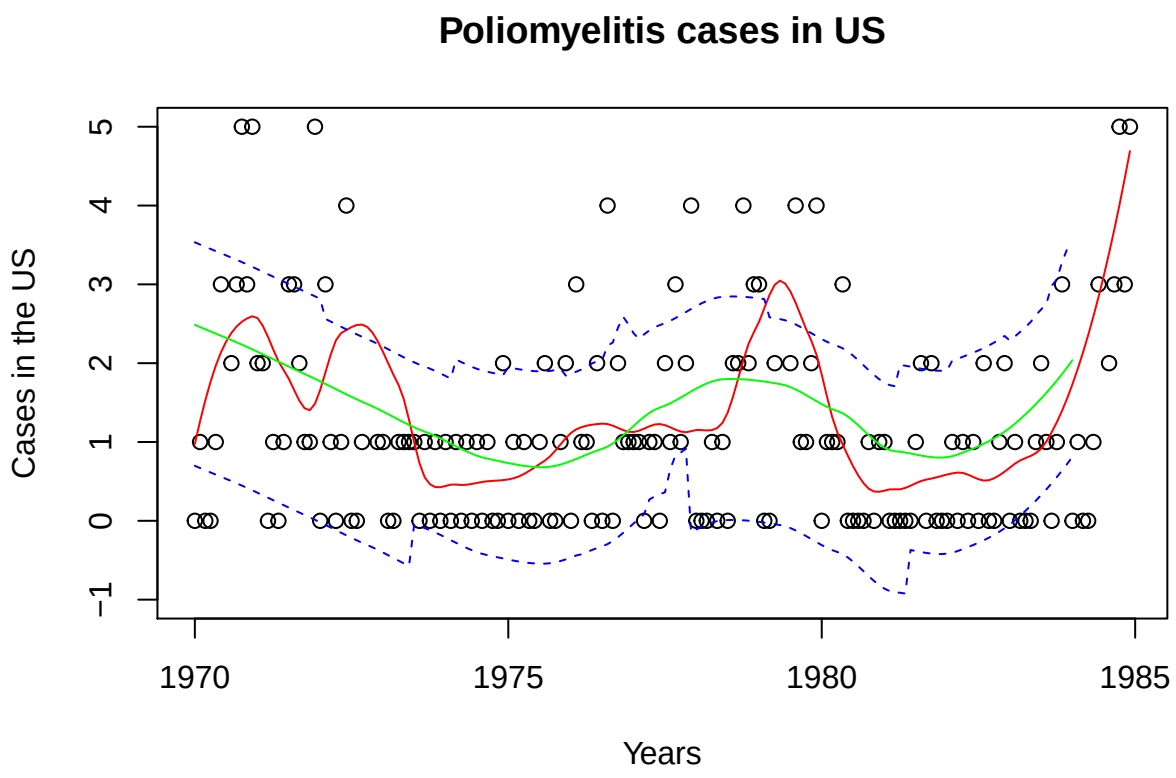

Figure 1: Lowess Verteilung der Polio Fälle in den US

Fragestellungen

Wie fließt die Spannweite in die LOWESS Schätzung ein?

Die Spannweite beschreibt den Grad der Glättung. In Abbildung 1 sieht man deutlich, dass die Rote Kurve mit einer Spannweite von 0.2 mehr Ausreisser umfasst als die grüne Kurve mit Spannweite 0.6. Sprich werden weniger benachbarte Werte mit einbezogen und so zusagen übersprungen wird die Kurve dadurch glatter.

Wie kann der mit LOWESS ermittelte Trend inhaltlich interpretiert werden?

Mit der Lowess Kurve können Verbindungen zwischen den Variablen und Trends erkannt werden. Bei den Punkten, an denen die Kurve starke Biegungen aufweist sind weniger Daten vorhanden, als an den geraden Stücken. Diese Biegungen entstehen unter anderem durch das Auftreten von Ausreissern. Diese können durch eine höhere Skalierung der Spannweite elemeniniert werden.

Wozu dient upper bzw. lower smoothing?

Das Upper and Lower Smoothing bietet zusätzliche Streuungsinformationen. Es werden zwei weitere Linien (hier blau strichliert) zum Diagramm hinzugefügt, die ebenfalls Informationen zu den Realtionen zwischen den Daten aufweisen. Hierfür werden die Daten nach oben und unten hin versetzt und wieder geglättet. Somit kann besser veranschaulicht werden wie weit die Ausreisser von der eigentlichen Glättung entfernt liegen.

2. Beispiel

Praisbeispiel

Nehmen sie die gleichen Daten wie in Beispiel 1 und wenden sie robustes Filtern an. Nehmen sie dazu aus dem Paket robfilter die Funktion robust.filter(), und wählen sie einen geeigneten Parameter für width. Visualisieren sie die Zeitreihe gemeinsam mit der geglätteten Zeitreihe. Werden Ausreißer erkannt? Wenn ja, zeichnen sie die Ausreißer in der Grafik ein.

Daten laden

```
library(robfilter)
```

Robustes Filtern anwenden

```
robf1  <- robfilter::robust.filter(ts.obj, width=7)
robf2  <- robfilter::robust.filter(ts.obj2, width=7)
robupper <- robfilter::robust.filter(upperSmoothing, width=7)
roblower <- robfilter::robust.filter(lowerSmoothing, width=7)

ts.rf1 <- ts(robf1$y, start=c(1970,1), end=c(1984,12),frequency=12)
ts.rf2 <- ts(robf2$y, start=c(1970,1), end=c(1984,12),frequency=12)
ts.rf3 <- ts(robupper$y, start=c(1970,1), end=c(1984,12),frequency=12)
ts.rf4 <- ts(roblower$y, start=c(1970,1), end=c(1984,12),frequency=12)
```

Abbildung zeichnen

```
plot(ts.rf1, type="l", main="Robustes Filtern", xlab="Years", ylab="Cases in the US")
lines(ts.rf2, col="blue",)
lines(ts.rf3, col="blue", lty="dashed")
lines(ts.rf4, col="blue", lty="dashed")
```

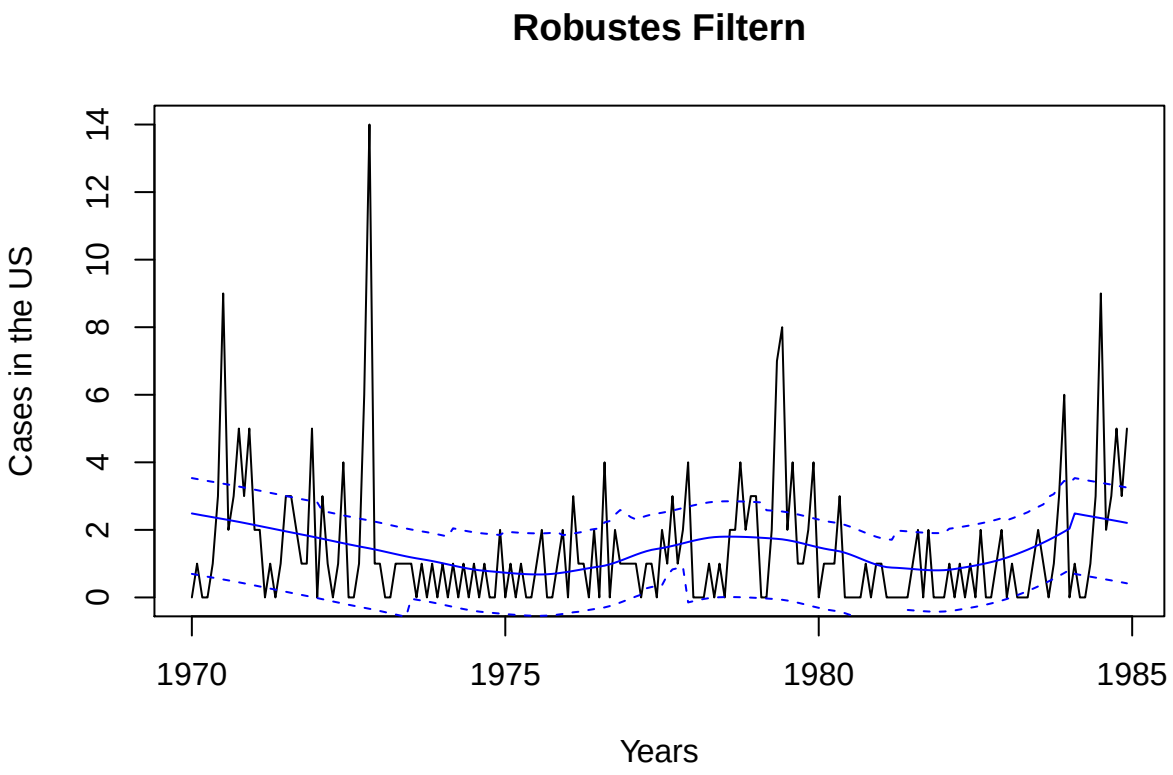

Figure 2: Robustes Filtern auf Poliodaten in den US anwenden

Fragestellungen

robust.filter() berechnet intern lokale Mittel und Streuungen. Welche Schätzer werden hierfür verwendet?

Die Schätzer können über die Parameter trend und scale definiert werden. Bei trend wird standardmäßig Repeated Median Regression verwendet. Bei scale wird normalerweise der Q_n scale estimator benutzt.

Nach welchem Prinzip werden Ausreißer ermittelt?

Der Umgang mit Ausreißern kann über den Parameter "outliner" definiert werden. Standardmäßig werden alle großen Ausreisser, die über die 3-Sigma-Regel identifiziert werden in durchschnittliche Werte umgewandelt.

Können die Ausreißer inhaltlich interpretiert werden?

Bei den Poliodaten können die Ausreisser als Monate in denen es besonders viele oder wenige Poliofälle in den US gegeben hat interpretiert werden.

3. Beispiel

Nehmen sie die gleichen Daten wie in Beispiel 1. Erstellen sie zunächst einen Plot der Trend-, Saison- und Restkomponenten.

Unterteilen sie nun die Daten in zwei aufeinanderfolgende Zeitbereiche, wobei der letzte Bereich etwa 1-4 Jahre sein soll (je nachdem, wie lang ihre Zeitreihe ist). Der erste Bereich wird dazu dienen, ein Modell zu schätzen, und mit dem zweiten Bereich kann mit dem Modell eine Prognose erstellt werden, die dann mit den gemessenen Daten verglichen werden kann. Dazu müssen sie den ersten Bereich mittels plot() zeichnen, den Bereich der x-Achse vergrößern, und danach mit lines den zweiten Bereich einzeichnen.

Berechnen sie nun anhand des ersten Bereiches eine Holt-Winters Schätzung für die Parameter. Zeichnen sie die erhaltenen geglätteten Werte in die Grafik mit anderer Farbe ein. Verwenden sie das berechnete Modell zur Prognose der restlichen Werte, und zeichnen sie die prognostizierten Werte mit einer auffallenden Farbe zusätzlich in die Grafik ein.

Trend-, Saison-, Restkomponenten Abbildung zeichnen

```
plot(stl(ts.obj, s.window = "per"), main="Trend-, Saison-, Restkomponenten Diagramm")
```

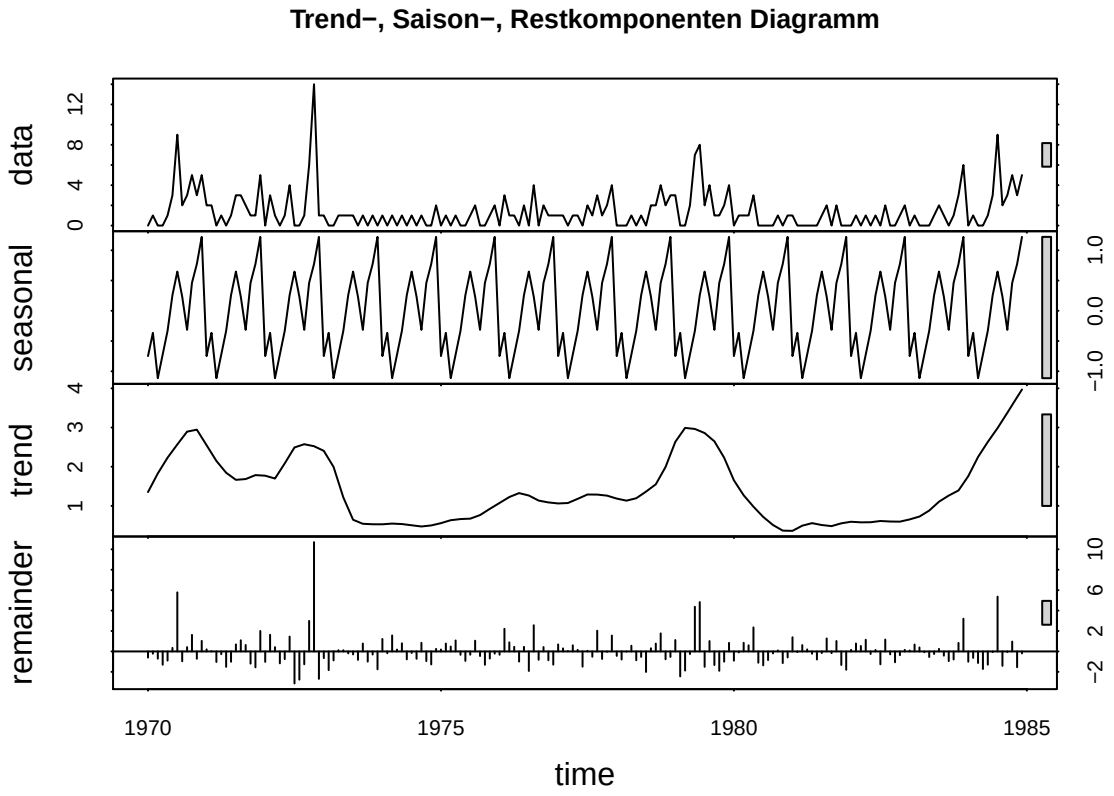

Figure 3: Trend-, Saison-, Restkomponenten Diagramm

Daten unterteilen

```
ts.first <- window(ts.obj, end=c(1979, 12))
ts.last4 <- window(ts.obj, start=c(1980, 0))
```

Unterteilungen zeichnen

```
plot(ts.first, xlim=c(1970,1984), xlab="Years", ylab="Cases in the US",
     main="Unterteilung in Bereiche")
lines(ts.first, col="black")
lines(ts.last4, col="red")
```

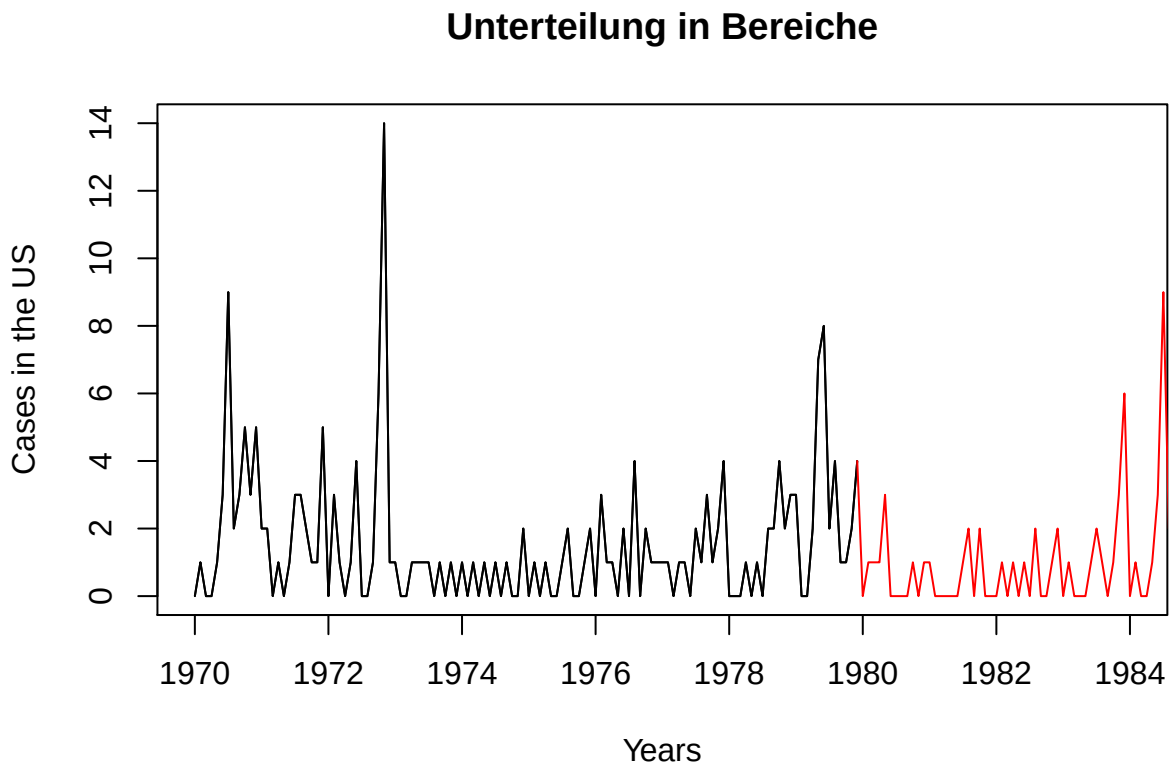

Figure 4: Unterteilung in Bereiche

Holt-Winters Schätzung anhand des ersten Bereichs

```
hw <- HoltWinters(ts.first)
holtWinterObj <- hw$fitted[, 1]
prog <- predict(hw, n.ahead=60)
```

Holt-Winters Schätzung einzeichnen

```
plot(hw, xlim=c(1970,1984), ylab="Cases in the US", xlab="Years", main="Holt-Winters Schätzung")
lines(holtWinterObj, col="brown")
lines(prog, col="red")
```

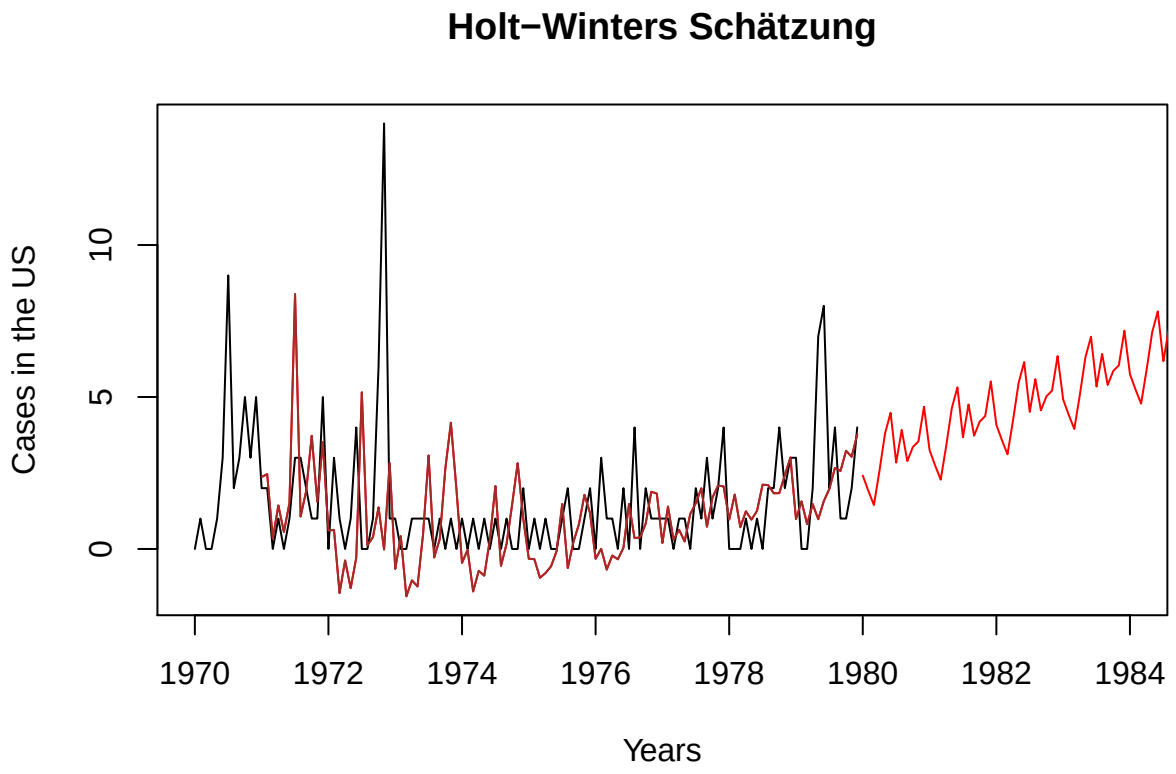

Figure 5: Holt-Winters Schätzung

Fragestellungen

Sind in der Zeitreihe saisonale Schwankungen und ein Trend erkennbar? Wie sind diese zu interpretieren?

Ja es gibt starke saisonale Schwankungen, so steigen die Fälle der Poliopatienten zum Ende jeden Jahres startk an. Im generellen Trend ist zu erkennen, dass die Fälle weiterhin stark ansteigen. Die saisonalen Schwankungen könnten durch die wechselden Wetterbedinungen begründet werden.Der generelle Trend könnte durch falsche Maßnahmen erklärt werden.

Ist die Restkomponente nur noch white noise?

Ja im Grunde genommen schon, jedoch sind noch einige starke Ausreißer zu erkennen.

Funktioniert Prognose mit Holt-Winters hier zufriedenstellend?

Die Hold-Winters Prognose sucht nach Mustern im angegebenen Zeitraum. Dieser wird dann für die restlichen Jahre in zyklischen Mustern eingefügt. Umso länger die Zeitspanne die prognostiziert wird, desto verfälschter wird das Ergebnis. Vergleicht man díe Trendanalyse mit der prognostizierten Kurve sieht man jedoch, dass der Trend übereinstimmt.

4. Beispiel

Praxisbeispiel

Laden sie nun die Daten Animals2 aus dem Paket library(robustbase) und logarithmieren sie die Daten.

Zeichnen sie nun das Gehirngewicht gegen das Körpergewicht in eine Grafik. Zeichnen sie die übliche kleinste Quadrate Regressionsgerade sowie die Gerade nach Siegel und die LTS Gerade, in jeweils unterschiedlicher Farbe ein. Erstellen sie abschließend noch 2 Grafiken der geschätzten 2-dimensionalen Dichte der Daten. Die erste Grafik soll die 3D Darstellung der Dichte zeigen (Befehl persp) und die 2. Grafik soll mit dem Befehl contour (angewandt auf die Dichte) erstellt werden (versuchen sie, eine gute Perspektive zu finden).

Daten Laden

```
library(robustbase)
animals <- Animals2
brain <- animals$brain
body <- animals$body
```

Werte berechnen

```
ltsline <- MASS::lmsreg(brain~body)
line <- mblm::mblm(brain~body, repeated=TRUE)
lsregression <- lm(brain~body)
```

Abbildung zeichnen

```
plot(animals, main="Gehirngewicht gegen Körpergewicht",
     xlim=c(0,20000), xlab="Body-Weight", ylab="Brain-Weight")
abline(ltsline, col="green")
abline(line, col="blue")
abline(lsregression, col="brown")
```

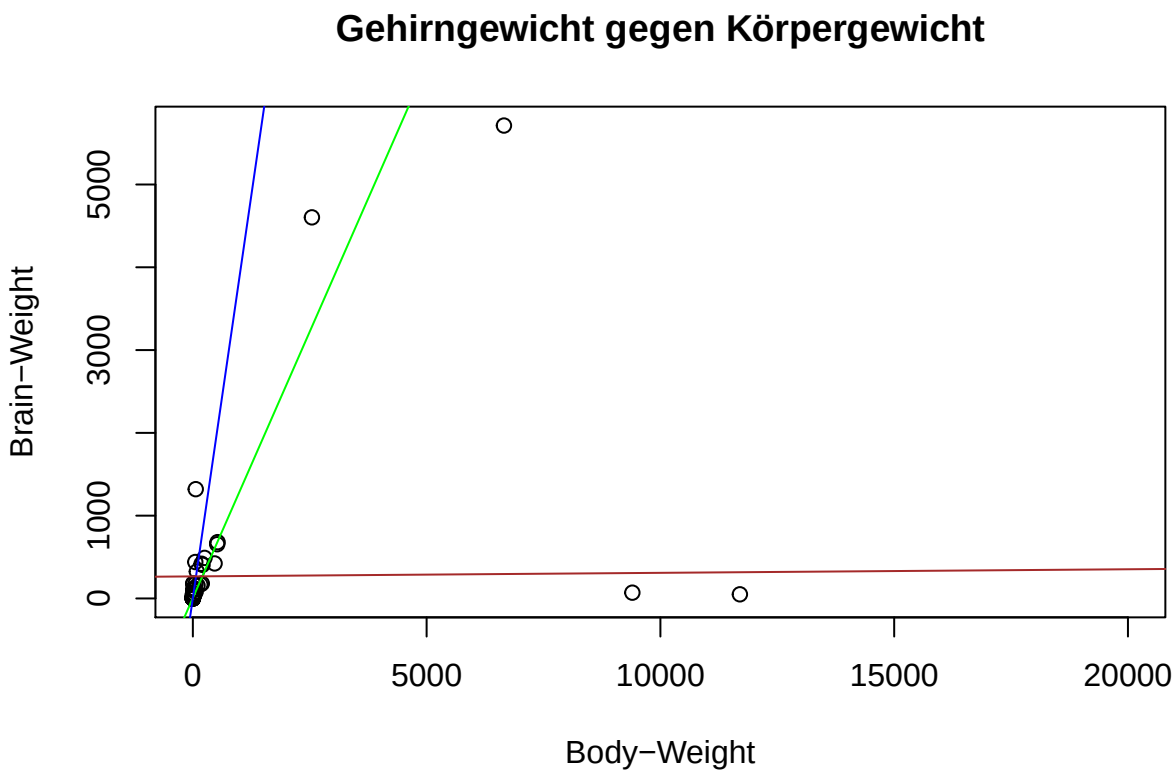

Figure 6: Gehirngewicht gegen Körpergewicht

Daten berechnen

```
par(mfrow=c(2,1))
density <- MASS::kde2d(body, brain)
```

Abbildung zeichnen

```
persp(density, theta=45, phi=0, r=4, col="blue",
      main="3D Darstellung der Dichte")
```

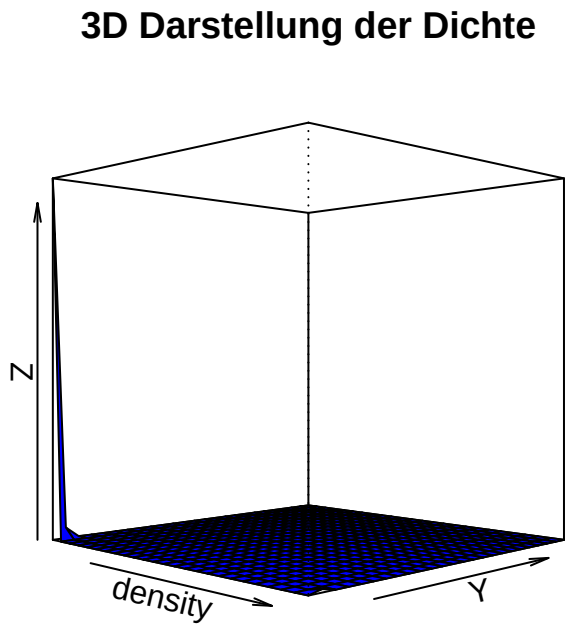

Figure 7: 3D Darstellung der Dichte mit persp

Daten berechnen

```
densityx <- density$x
densityy <- density$y
densityz <- density$z
```

Abbildung zeichnen

```
contour(densityx,densityy, densityz, ylim=c(0,250), xlim=c(0,3200),
        main="2D Darstellung der Dichte")
```

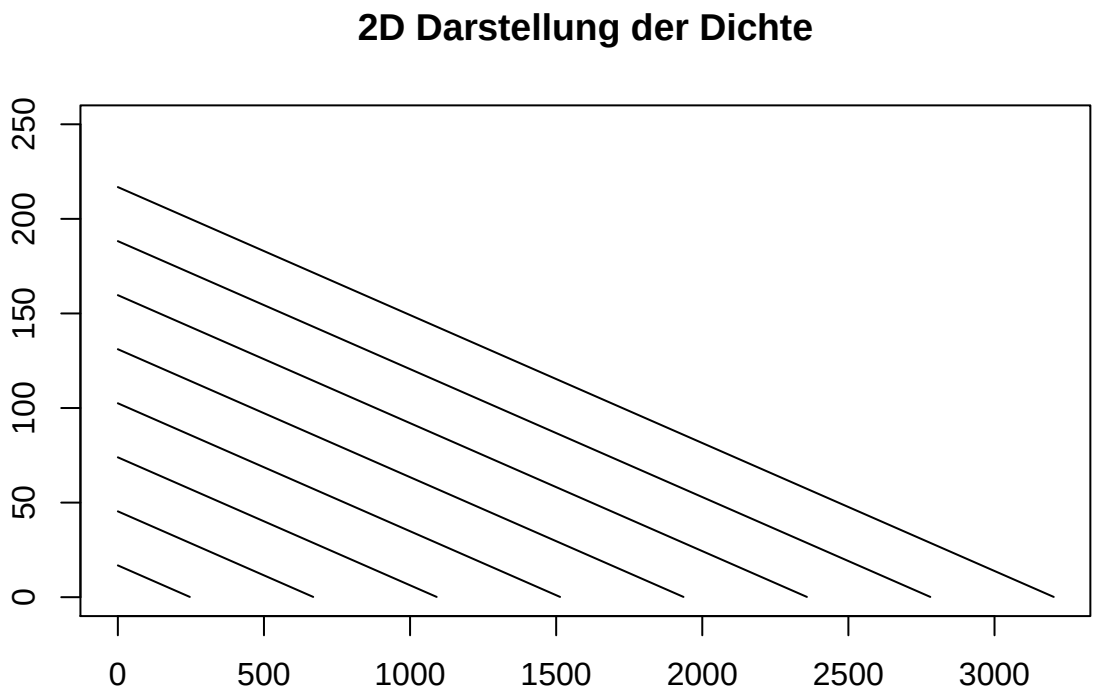

Figure 8: Darstellung der Dichte mit Contour

Fragestellungen

Welche der Geraden eignet sich ihrer Meinung nach für diesen Datensatz am besten (Begründung!)

Meiner Ansicht nach eigenet sich die LTS Gerate am besten, da sie alle Werte gleich einbezieht und somit den geringsten Abstand zu allen Werten hat. Die beiden Anderen schließen besonders hoch bzw. niedrige Ausreißer aus und neigen sich somit stark auf eine Seite.

Beschreiben sie kurz, wie diese (von ihnen gewählte) Gerade berechnet wird?

Die LTS Regression minimiert die Summe der quadratischen Residuen. Alle Residuen werden mit beta hoch zwei multipliziert und summiert.