Datenanalyse

Abgabe 2

Pascal Poremba (Matr. Nr. 11809911)

21.05.2020

Erste Frage

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

```
## Warning: package 'astsa' was built under R version 3.6.2
```

Gewählter Datensatz: astsa.flu Die erstellten Grafiken:

```
## Warning in `+.default`(p2, resplus): Länge des längeren Objektes
##       ist kein Vielfaches der Länge des kürzeren Objektes
```

```
## Warning in `+.default`(p2, resminus): Länge des längeren Objektes
##       ist kein Vielfaches der Länge des kürzeren Objektes
```

```
plot(res, main="Loess Smoothing and Prediction", xlab="YEARS", ylab="flu-deaths")
lines(p1, col="red")
lines(p2,  col="blue")

lines(smoothedPlus,  col="blue", lty="dashed")
lines(smoothedMinus,  col="blue", lty="dashed")
```

Fragenstellungen: Wie fließt die Spannweite in die LOWESS Schätzung ein?

Bei der LOWESS Schätzung wird eine Gewichtungsfunktion verwendet, welche bewirkt, dass der Einfluss, den benachbarte Werte auf die Glättung an einer Position haben, sich mit zunehmender Entfernung dieser Position verringert. Ausreißer werden hierbei geringer gewertet als bei anderen Verfahren. Die Wahl der Glättungsbreite, welche die Anzahl der in der Berechnung für einen Punkt eingehenden Werte angibt, ist von großer Bedeutung. Die Spannweite ist der Parameter alpha welcher den Grad der Glättung beschreibt. In

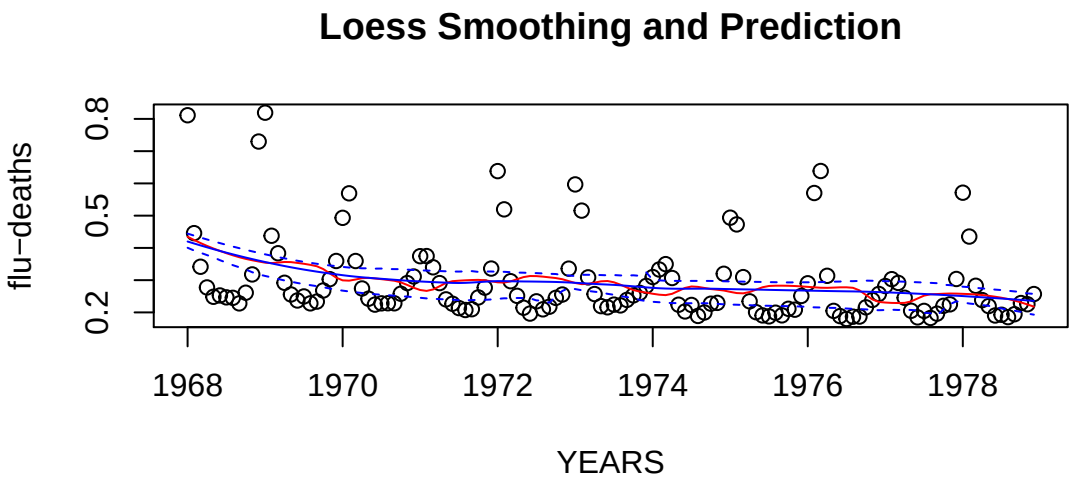


Figure 1: LOWESS Curves

Abbildung **??** ist die rote Kurve mit einer Spannweite von 0.3 gezeichnet, bei der blauen mit einer Spannweite von 0.6.

Wie kann der mit LOWESS ermittelte Trend inhaltlich interpretiert werden?

Bei der Interpretation der LOWESS Schätzung wird erstmals geschaut, ob die vorhandene Linie gerade scheint. Falls die Linie Biegungen aufweist, kann darauf geschlossen werden, dass in diesen Regionen nur wenige Punkte vorhanden sind. Je weniger Punkte sich in einem Bereich befinden, desto stärker beeinflussen diese die Biegung der Linie. Aufgrund der Art der Berechnung kann es sein, dass die Linie nur in den Randregionen links und rechts stark von anderen Punkten abweicht Wenn in genau diesen Regionen wenig Punkte vorhanden sind, geht man trotzdem von einer Linearität aus.

Falls definitiv keine Linearität vorhanden ist kann sich das problematisch auf die Validität der Ergebnisse auswirken.

Wozu dient upper bzw. lower smoothing?

Beim upper bzw. lower Smoothing werden zwei weitere Linien hinzugefügt. Diese Smooths (zu deutsch: Glättungen) zeigen ebenso wie die normale Linie, wie die Verteilung zwischen den Variablen variiert. Sie sind sowohl an den positiven sowie an den negativen Residuen separat anzuwenden. Dies dient zur besseren Veranschaulichung der Punkte, welche weit von der LOWESS Linie entfernt liegen.

Zweite Frage

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.

```
plot(ts.rf1, type="l")
lines(ts.rf2, col="blue",)
lines(ts.rf3, col="blue", lty="dashed")
lines(ts.rf4, col="blue", lty="dashed")
```

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

Die Schätzer können mit den Argumenten trend und scale definiert werden. Hier gibt es für die robust approximation (trend): Median, Repeated Median regression (wird Standardmäßig verwendet), Least Trimmed

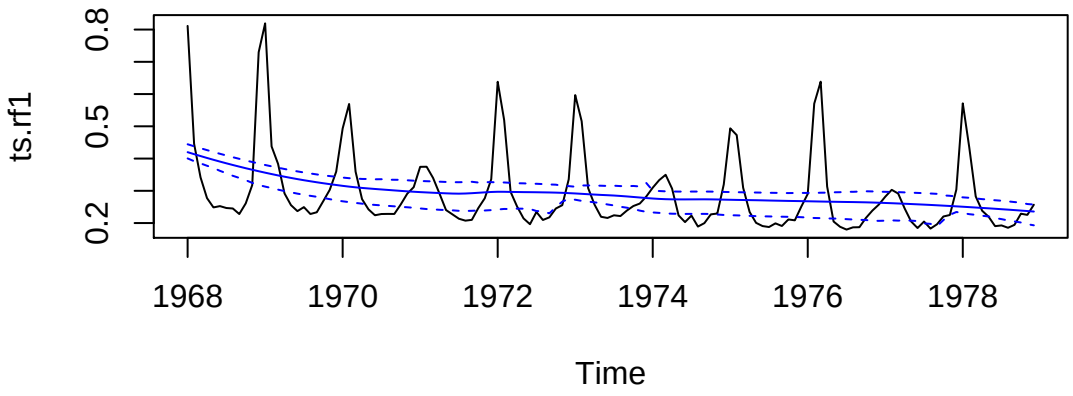


Figure 2: Robustes Filtern

Squares regression und Least Median of Squares regression. Für die robust estimation (scale) gibt es folgende Schätzfunktionen: Median absolute deviation about the median, Rousseeuw's and Croux' (1993) Q_n scale estimator (wird Standardmäßig verwendet), Rousseeuw's and Croux' (1993) S_n scale estimator, Length of the shortest half. Weiters verwendet man Schätzer, die eine hohe Bruchpunktresistenz aufweisen.

Nach welchem Prinzip werden Ausreißer ermittelt?

Die Ausreißer können mir dem Argument outliner definiert werden. Hier sind folgende Prinzipien wählbar: Replace ('trim') large outliers detected by a 3sigma-rule (d=3) by the current level estimate (k=0) (wird Standardmäßig verwendet), Shrink large outliers (d=3) strongly towards the current level estimate (k=1), Shrink large and moderatly sized outliers (d=2) strongly towards the current level estimate (k=1), Shrink large and moderatly sized outliers (d=2) towards the current level estimate (k=2).

Können die Ausreißer inhaltlich interpretiert werden?

In diesem speziellen Fall (flu) sind Ausreißer Monate, in denen es vermehrt Grippetote gab, sei es durch einen besonders kalten Winter, änderungen im Gesundheitssystem oder neuartigen Viren. Hier sind ausreißer nicht zu ignorieren, da sie auch wichtige Informationen beinhalten.

Dritte Frage

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.

```
plot(stl(ts.flu, s.window = "per"))
```

```
ts.first8 <- window(ts.flu, end=c(1975, 12))
ts.last3 <- window(ts.flu, start=c(1976, 0))

plot(ts.first8, xlim=c(1968,1979))
```

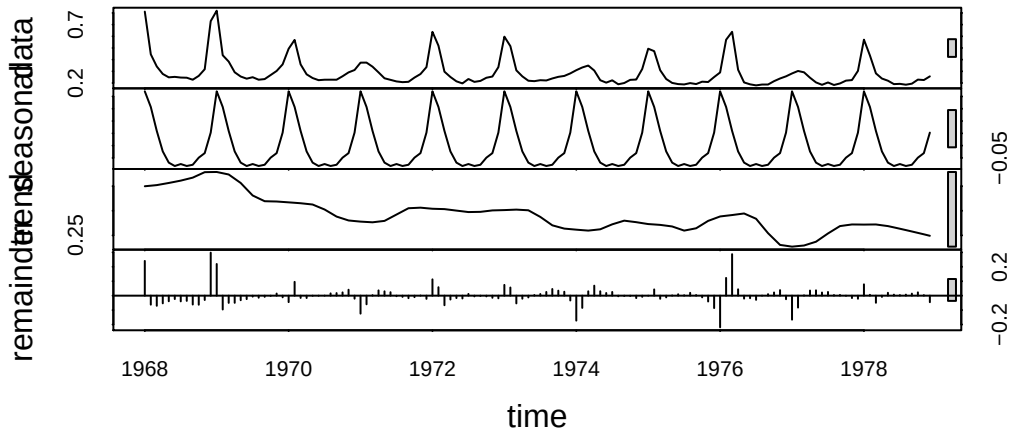


Figure 3: Trend-, Saison-, Restkomponenten

```
lines(ts.last3, col="red")
```

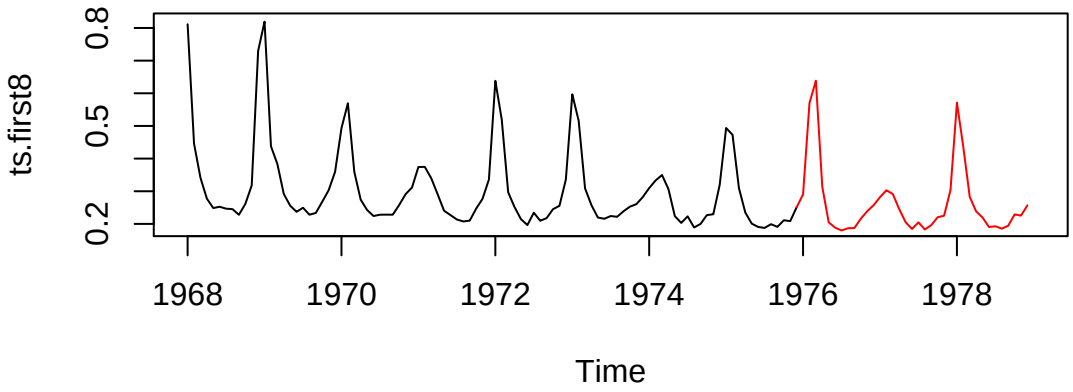


Figure 4: Zeitbereiche

```
hw <- HoltWinters(ts.first8)
smoothedHW <- hw$fitted[, 1]
prog <- predict(hw, n.ahead=36)

plot(hw, xlim=c(1968,1979))
lines(smoothedHW, col="blue")
lines(prog, col="red")
```

Fragestellungen: Sind in der Zeitreihe saisonale Schwankungen und ein Trend erkennbar? Wie sind diese zu interpretieren? Ist die Restkomponente nur noch white noise?

Ja in der Zeitreihe erkennt man bei den gewählten Daten (flu) eine sehr starke saisionale Schwankung (mehr Grippetote im Winter als im Sommer). Ein leichter trend ist auch erkennbar, nämlich eine leichte abnahme von Grippetoten. Das hat vermutlich mit dem Fortschritt der Technik und dem Gesundheitswesens zu tun.

Funktioniert Prognose mit Holt-Winters hier zufriedenstellend?

Nicht zu 100%. Zum Beispiel ist Anfang 1976 noch ein Boom im Vergleich zum Vorjahr, der nicht erkannt wird. 1977 sollten die Zahlen im Vergleich zum Vorjahr halbiert sein, bei der Holt-Winters Prognose ist sie ähnlich groß. Jedoch werden vor allem die Saisionalen schwankungen an sich relativ gut erkannt.

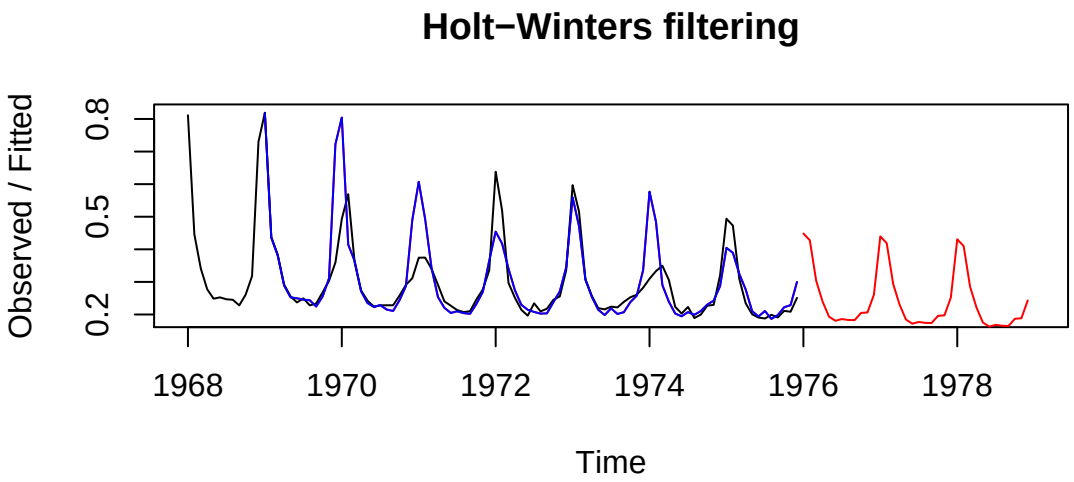


Figure 5: Holt-Winters

Vierte Frage

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 LMS Gerade und die Gerade nach Siegel, 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 image (angewandt auf die Dichte) erstellt werden (versuchen sie, eine gute Perspektive zu finden).

Daten Laden

```
a <- Animals2
body <- a$body
brain <- a$brain
lsrg <- MASS::lmsreg(brain~body)
mb <- mblm::mblm(brain~body, repeated=TRUE)
kqr <- lm(brain~body)
plot(a)
abline(lsrg, col="blue")
abline(mb, col="red")
abline(kqr, col="green")
```

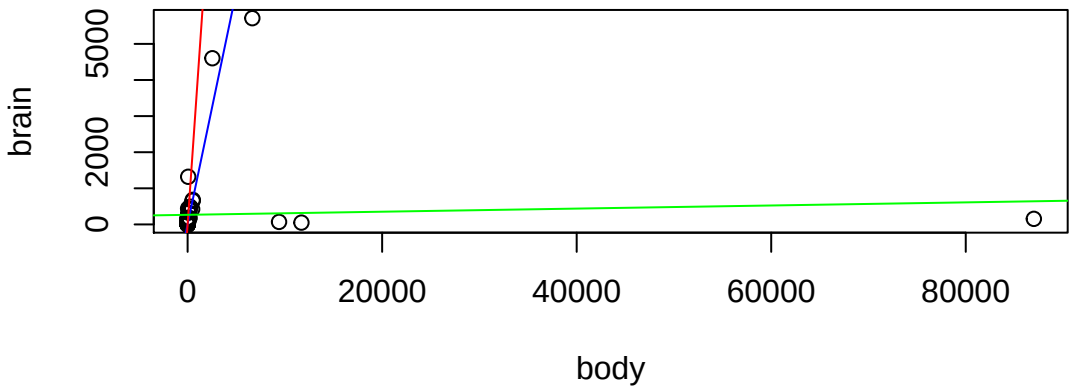


Figure 6: Animals

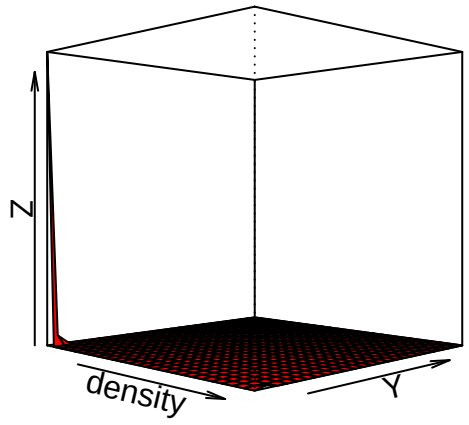
```r
par(mfrow=c(2,1))
density <- MASS::kde2d(body, brain)
dx <- density$x
dy <- density$y
dz <- density$z
persp(density, theta=45, phi=0, r=4, col="red")
image(1:25, 1:25, dz, col = hcl.colors(12, "YlOrRd", rev = TRUE), useRaster=TRUE)
```

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

Die LMS Gerade sieht meiner Meinung nach am plausibelsten aus. Sie geht eher auf die Ausreißer ein als die Gerade von Siegel, folgt jedoch dem eindeutigen Trend.

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

Sie passt die Regression an die 'guten' Punkte im Datensatz an um einen Regressionsschätzer mit hoher Bruchpunktresistenz zu erhalten.



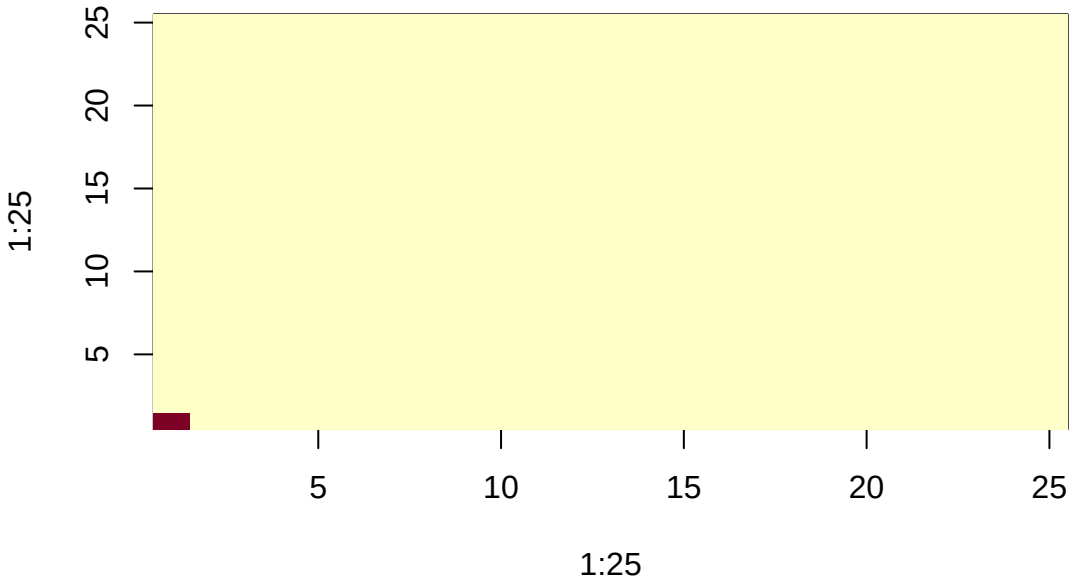


Figure 7: 2D Dichte