Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering
IDP Institut für Datenanalyse
und Prozessdesign

2018-12

CAS Datenanalyse

Probepruefung Zeitreihen und Prognosen

Name:	Punkte:	Note:

Bewertung

Geben sie bei quantitativen Fragen immer alle benutzten Formeln an, bevor sie die Zahlen einsetzen und Schlussresultate berechnen. Erfragte Begründungen müssen in ganzen Sätzen verständlich ausformuliert sein. Fehlt eine Begründung, bzw. ein klar ersichtlicher Lösungsweg, so wird auch bei korrekten Zahlenresultaten höchstens die Hälfte des Punktemaximums vergeben.

Erlaubte Hilfsmittel:

open book", d.h. beliebige schriftliche Hilfsmittel sind erlaubt. Taschenrechner sind ebenfalls zugelassen. Ein Laptop darf zur Bearbeitung der Aufgaben eingesetzt werden. Auf dem Computer darf R-Studio und ein PDF-Reader betrieben werden. Bereits bestehende R-Files dürfen eingesetzt werden.

Viel Erfolg

frequency=12

Wir betrachten in dieser Aufgabe die mittlere Milchleistung pro Kuh in den USA zwischen 1994 und 2005. Die Daten befinden sich im File `milk_data.rda`. Es handelt sich um Monatsdaten die in einem Data Frame vorliegen, vorhanden sind 12 komplette Jahre ab Januar 1994. Im Data Frame sind die die folgenden Variablen enthalten:

start=c(1994, 1)

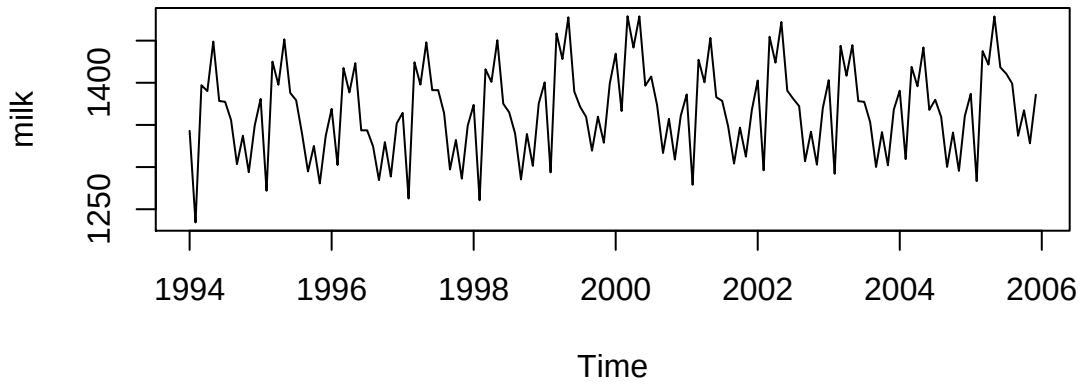
`milk`: mittlere Milchleistung pro Kuh
`zeitschritte`: Zeitschritte
`monat`: Monate

a) (1 Punkt) Definieren Sie die Spalte `milk` gemäss den Angaben von oben in R sinnvoll und korrekt als Zeitreihe. Geben Sie den verwendeten R-Befehl an.

```
milk <- ts(milk_data$milk, start = c(1994,1), frequency = 12)
```

b) (2 Punkte) Schauen Sie sich die Zeitreihe in R an. Ist die Zeitreihe stationär? Begründen Sie ihre Antwort.

```
plot(milk)
```

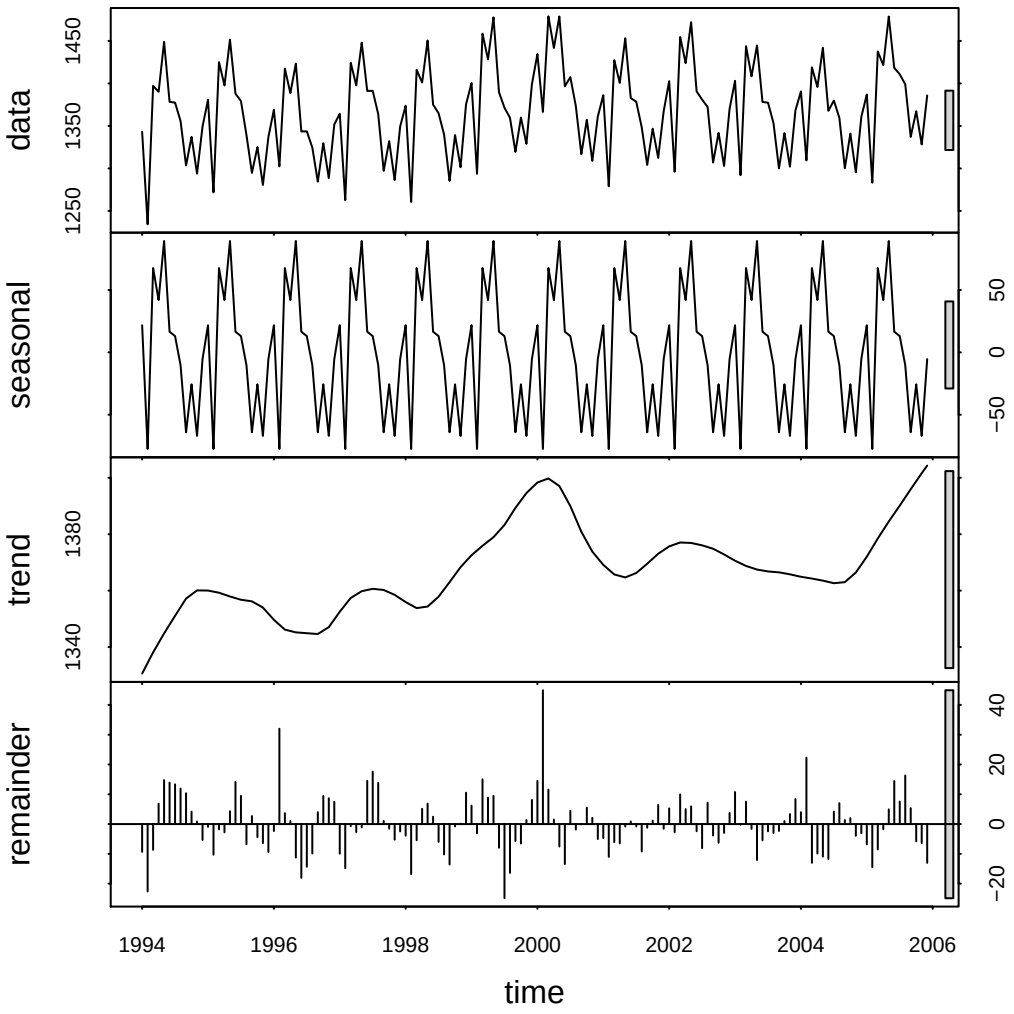



```
# Nein die Zeitreihe ist nicht Sationär.
# Sie weist eine Saisonalität auf und allenfalls einen leichten positiven Trend.
```

c) Führen Sie eine STL-Zerlegung der Zeitreihe durch, wobei Sie annehmen können, dass die Saisonkomponente konstant ist.

i) (1 Punkt) Welche der drei Komponeten der STL-Zerlegung dominiert die Struktur der Zeitreihe?
ii) (1 Punkt) Ist eine log-Transformation der Daten angebracht?

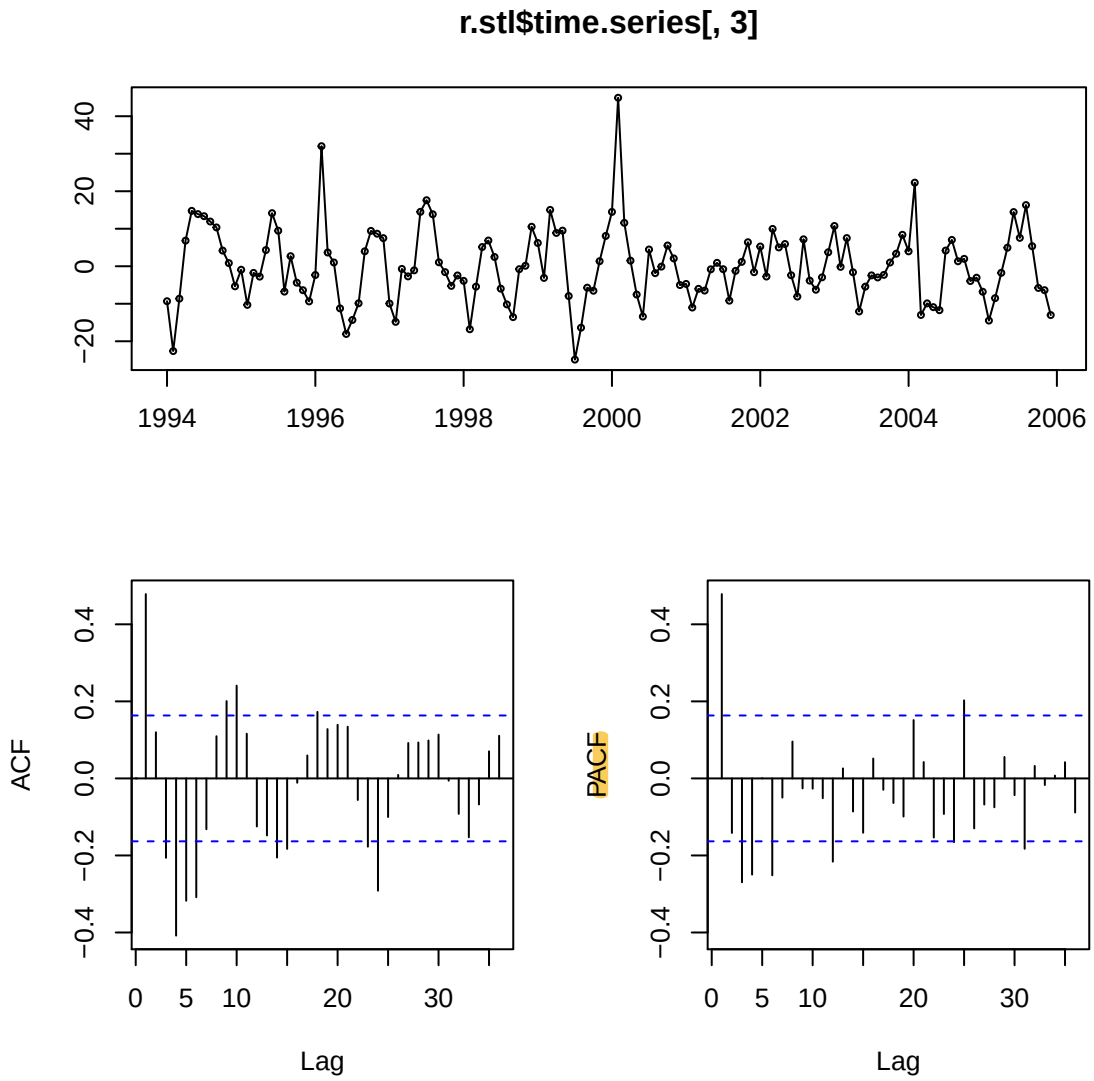
```
plot(r.stl <- stl(milk, s.window = "periodic"))
```




```
# i) Die Saisonkomponente dominiert die ZR ii) Nein, eine log-Transformation
# ist nicht nötig (es ändert sich nichts am Zerlegungsmuster). Varianz ist
# bereits mehr oder weniger konstant.
```

d) (1 Punkt) Nehmen Sie nun den stationären Restterm aus der STL-Zerlegung. Ist das Anpassen eines AR(p)-Modells angebracht? Argumentieren Sie mit Hilfe des ACF und PACF-Plot.

```
tsdisplay(r.stl$time.series[,3])
```




```
# ACF-Plot zeigt noch deutlich saisonales Muster
# evtl. leichte exponetielle Abnahem der Korrelation zu erkennen
# PACF zeigt bei lag 1 einen deutlichen cut-off , aber
# einige partielle Korrelation scheinen nicht Null zu sein
# es ist also nicht so klar, Saison kann nicht ganz entfernt werden
```

e) (2 Punkte) Unabhängig von ihren Argumenten in d) sollen Sie nun aus der PACF die zwei plausibelsten Ordnungen p für ein AR(p)-Modell ablesen und angeben. Erklären Sie ihre Wahl.

144 Beobachtungen

```
# Interpretaiton bis max. Lag 21 = 10*log10(144)
# p = 12 oder p = 6
# Nach dem Lag 12 oder 6 ist ein klarer drop-off zu erkennen
```

f) (2 Punkte) Passen Sie nun ein AR(p)-Modell an die Daten des Restterms an, welches zum kleinsten AIC-Wert führt. Geben Sie dessen Ordnung p an. Geben Sie den verwendeten R-Code an.

```
r.burg <- ar.burg(r.stl$time.series[,3])
# Order selected 6
```

g) Passen Sie nun ein additives Holt-Winters Modell an die in a) generierte Zeitreihe an.

i) (2 Punkte) Welche Bedeutung und welchen Wert hat der Parameter β? Welchen Schluss ziehen Sie aus diesem Resultat?

ii) (1 Punkt) Wie gross ist die mit dem Holt-Winters Modell prognostizierte Milchleistung für den Juni 2006.

```
r.hw <- hw(milk)
r.hw$model$par[2]
```

```
##          beta
## 0.0001149179
```

```
# i)
# Glättungsparameter der Steigung ist klein, d.h. der Steigungsparameter b wird aus dem Mittelwert a
#ii)
## Aufgabe 1: Prognostizierte Milchleistung pro Kuh im Juni 2006 ist
hw(milk)$mean[6]
```

```
## [1] 1407.483
```

h) Berechnen Sie das Zeitreihen-Regressionsmodell `milk ~ zeitschritte + monat` für die gegeben Daten (eine Transformation ist nicht nötig).

i) (0.5 Punkte) Geben Sie den dazu verwendeten R-Befehl an.

ii) (0.5 Punkt) Ändert sich die durchschnittliche Milchleistung der Kühe zwischen 1994 und 2005 signifikant ($\alpha = 1\%$) aufgrund der Resultate des Regressionsmodells?

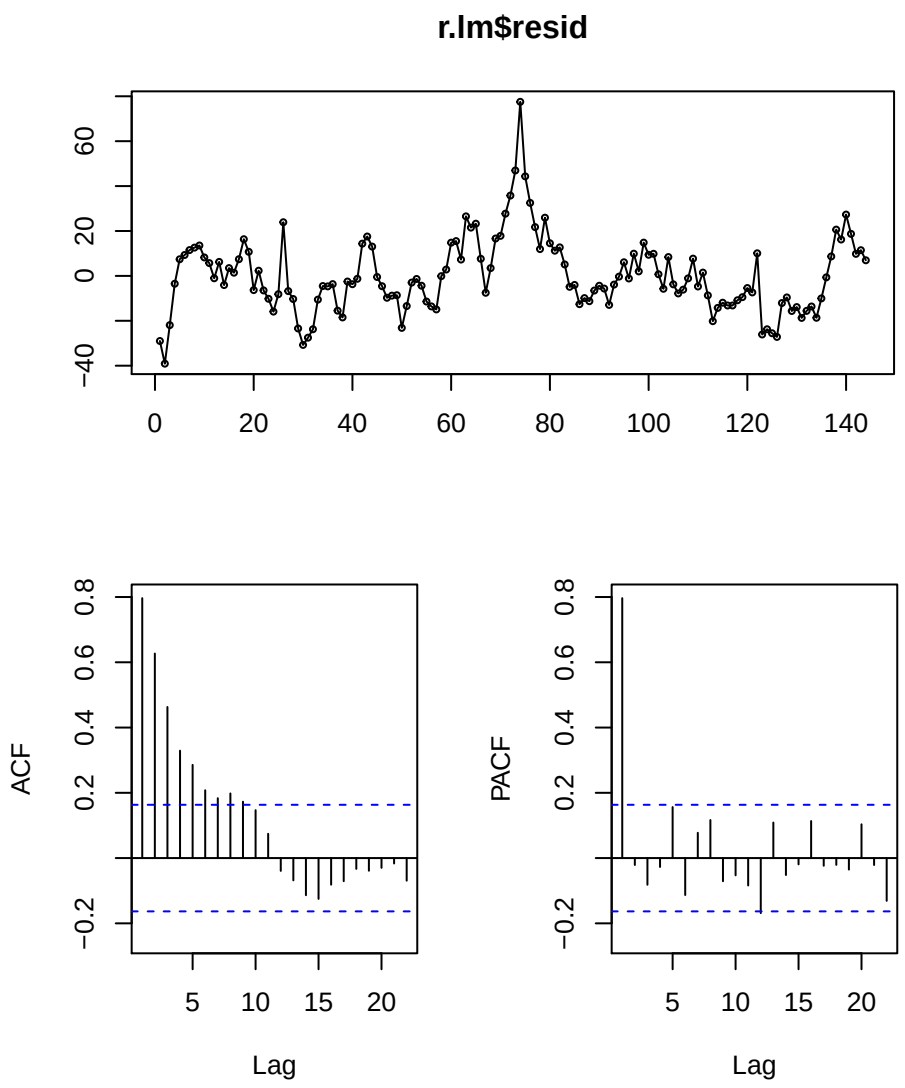
iii) (1 Punkte) Können Sie dem p-Wert von `milk` vertrauen? Bitte begründen Sie ihre Antwort.

```
# i)
r.lm <- lm(milk ~ zeitschritte + monat, data = milk_data)
# ii) Ja, gemäss dem summary output ist der Koeffizient von zeitschritte
# hoch signifikant << 1%
summary(r.lm)
```

```
##
## Call:
## lm(formula = milk ~ zeitschritte + monat, data = milk_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.090 -10.360  -3.242   9.850  77.532
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1392.7956     5.5402 251.396  < 2e-16 ***
## zeitschritte      0.2165     0.0347   6.239 5.63e-09 ***
## monatAugust     -51.3968     7.0433  -7.297 2.52e-11 ***
## monatDecember   -45.0949     7.0474  -6.399 2.56e-09 ***
## monatFebruary  -119.8645     7.0423 -17.021  < 2e-16 ***
## monatJanuary    -21.0162     7.0427  -2.984 0.003394 **
## monatJuly       -28.3650     7.0427  -4.028 9.49e-05 ***
## monatJune       -25.0367     7.0423  -3.555 0.000526 ***
## monatMarch       25.5721     7.0420   3.631 0.000403 ***
```

```
## monatMay          47.6108     7.0420   6.761 4.10e-10 ***
## monatNovember  -106.9855     7.0461 -15.184  < 2e-16 ***
## monatOctober    -66.0054     7.0450  -9.369 2.80e-16 ***
## monatSeptember -104.8684     7.0441 -14.888  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.25 on 131 degrees of freedom
## Multiple R-squared:  0.9046, Adjusted R-squared:  0.8959
## F-statistic: 103.5 on 12 and 131 DF,  p-value: < 2.2e-16
```

```
# iii)
tsdisplay(r.lm$resid)
```




```
# Nein, dem p-Wert kann man nicht trauen, Schätzung der sd ist nicht korrekt
# -> Teststatistik ist falsch Grund Residuen sind autokorreliert
```

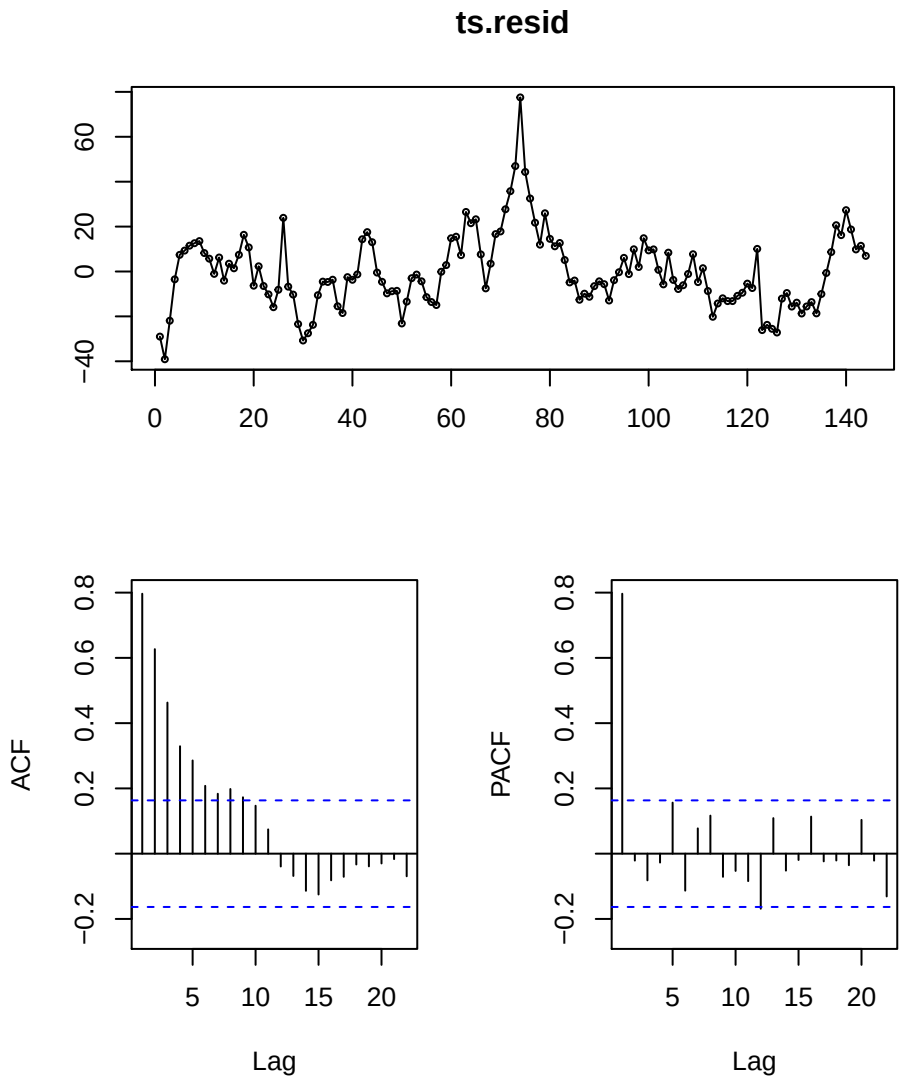
i) (2 Punkt) Welches AR(p)-Modell beschreibt die Residuen des Regressionsmodell aus h) gut?

```
# i)
ts.resid <- ts(r.lm$residuals)
```

```
ar.burg(ts.resid)
```

```
##
## Call:
## ar.burg.default(x = ts.resid)
##
## Coefficients:
##      1
## 0.8057
##
## Order selected 1  sigma^2 estimated as  94.95
```

```
tsdisplay(ts.resid)
```




```
## Aufgabe 2: AR(1) aus ACF und PACF oder geschätzt mit burg-Algorithmus
```

j) Verwenden sie nun die Funktion `gls()`, um im Zeitreihen-Regressionsmodell von Frage h) auch noch die Korrelation der Residuen zu berücksichtigen

i) (2 Punkte) Geben Sie die dafür die notwendigen R-Befehle an.

```r
corStruct <- corARMA(form = ~ zeitschritte, p = 1, q = 0)
r.gls <- gls(milk ~ zeitschritte + monat, data= milk_data, correlation = corStruct)
```

ii) (1 Punkt) Wie beantworten Sie nun die Frage h-ii?

```r
summary(r.gls)
```

```
## Generalized least squares fit by REML
##   Model: milk ~ zeitschritte + monat
##   Data: milk_data
##        AIC     BIC    logLik
##   1049.632 1092.76 -509.8158
##
## Correlation Structure: AR(1)
##  Formula: ~zeitschritte
##  Parameter estimate(s):
##       Phi
## 0.8293887
##
## Coefficients:
##                    Value Std.Error   t-value p-value
## (Intercept)    1389.1432  9.696251 143.26601  0.0000
## zeitschritte      0.2658  0.108648   2.44671  0.0157
## monatAugust     -52.2249  5.039041 -10.36405  0.0000
## monatDecember   -47.2380  5.043997  -9.36520  0.0000
## monatFebruary  -119.5128  4.037777 -29.59866  0.0000
## monatJanuary    -20.4880  4.636179  -4.41916  0.0000
## monatJuly       -28.9577  4.657178  -6.21787  0.0000
## monatJune       -25.4167  4.044673  -6.28399  0.0000
## monatMarch       25.7492  3.034596   8.48521  0.0000
## monatMay         47.4267  3.035718  15.62290  0.0000
## monatNovember  -108.7259  5.258019 -20.67812  0.0000
## monatOctober    -67.3991  5.323612 -12.66041  0.0000
## monatSeptember -105.9609  5.252523 -20.17334  0.0000
##
##  Correlation:
##                (Intr) ztschr mntAgs mntDcm mntFbr mntJnr mntJly montJn
## zeitschritte   -0.806
## monatAugust    -0.241 -0.021
## monatDecember  -0.249 -0.029  0.506
## monatFebruary  -0.220  0.005  0.331  0.611
## monatJanuary   -0.254  0.003  0.416  0.792  0.764
## monatJuly      -0.224 -0.017  0.807  0.423  0.280  0.351
## monatJune      -0.196 -0.013  0.623  0.337  0.227  0.282  0.765
## monatMarch     -0.165  0.005  0.230  0.414  0.666  0.514  0.196  0.159
## monatMay       -0.148 -0.008  0.422  0.236  0.160  0.199  0.514  0.666
## monatNovember  -0.253 -0.029  0.593  0.828  0.515  0.662  0.492  0.389
## monatOctober   -0.254 -0.028  0.697  0.699  0.442  0.565  0.572  0.449
## monatSeptember -0.251 -0.025  0.827  0.595  0.383  0.485  0.673  0.524
##                mntMrc montMy mntNvm mntOct
## zeitschritte
## monatAugust
## monatDecember
## monatFebruary
```

```
## monatJanuary
## monatJuly
## monatJune
## monatMarch
## monatMay        0.113
## monatNovember   0.352  0.270
## monatOctober    0.304  0.309  0.836
## monatSeptember  0.265  0.358  0.705  0.835
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -1.96904398 -0.59509967 -0.05377022  0.58543553  4.23470679
##
## Residual standard error: 18.22628
## Degrees of freedom: 144 total; 131 residual
# p-Wert = 0.0157  > 0.01 knapp nicht signifikant
```