Praktische Übung zur Zeitreihenanalyse
Sommersemester 2023

Lennart Oelschläger ✉
14-tägig, Mo, 16:15 – 17:45, T2-213

Aufgabenblatt 3 – Lösungen

Aufgabe 1 (Multiple lineare Regression)

a) Wie wird das multiple lineare Regressionsmodell definiert, und wie können die Modellparameter geschätzt werden?

> Eine Variable y wird mit k Regressoren x_k durch $y = \sum_{j=1}^{k} x_j\beta_j + u$ erklärt. Der u Term beinhaltet den Modellfehler, die β_j's sind Modellparameter. Es sollen konkrete Werte für die Parameter gefunden werden, die am besten zu gegebenen Daten passen. Die populärste Methode ist der Kleinste-Quadrate-Schätzer: Er bestimmt solche Schätzwerte, sodass die Summe der quadrierten Residuen minimal ist.

b) Welche Eigenschaften hat der Kleinste-Quadrate Schätzer, und welche Voraussetzungen müssen dafür erfüllt sein?

> Unter den Annahmen MLR.1 (lineares Modell), MLR.2 (Zufallsstichprobe), MLR.3 (Information in den Regressoren, keine Multikollinearität) und MLR.4 (bedingte Fehlererwartung ist Null) ist der KQ-Schätzer erwartungstreu. Gilt zusätzlich MLR.5 (Homoskedastie), so ist die Varianz des KQ-Schätzers die minimale Varianz unter allen linearen und erwartungstreuen Schätzern (Gauss-Markov Theorem). Gilt auch MLR.6 (Fehler sind normalverteilt), so ist der KQ-Schätzer normalverteilt. Unter einer technischen Annahme an die Daten ist er konsistent.

c) Bitte prognostizieren Sie den y Wert für $x_1 = 6$ und $x_2 = -1$, gegeben die Daten

$$
\begin{aligned}
x_1 &= \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 1 & 2 & 3 & 4 & 5 \end{pmatrix}, \\
x_2 &= \begin{pmatrix} 1 & 1 & 2 & 2 & 3 & 3 & 4 & 4 & 5 & 5 \end{pmatrix}, \\
y &= \begin{pmatrix} 1 & 3 & 1 & 5 & 2 & -3 & -3 & -1 & -2 & -2 \end{pmatrix}.
\end{aligned}
$$

> Erstelle die drei Datenvektoren in `R` und erzeuge daraus einen `data.frame` mittels `data <- data.frame(y, x_1, x_2)`. Schätze das lineare Modell mit dem Aufruf `model <- lm(y ~ x_1 + x_2, data)` und prognostiziere mit `model$coefficients %*% c(1, 6, -1)`. Ich erhalte den Wert 12.

Aufgabe 2 (Trendmodellierung durch polynomiale Regression)

a) Sie finden im Lernraum der PÜ den Datensatz `hermannslauf_frauen.csv` mit den Hermannslaufbestzeiten der Frauen. Bitte erstellen Sie eine Grafik der Bestzeiten in Minuten. Können Sie Elemente des Komponentenmodells erkennen und interpretieren?

Siehe R Code für die Erstellung der Grafik. Im klassischen additiven Komponentenmodell wird angenommen, dass sich die Beobachtung x_t zum Zeitpunkt t additiv als $x_t = T_t + S_t + Z_t + R_t$ ergibt:

- T_t ist der Trendwert bei t (langfristige Veränderung des Erwartungswertes),
- S_t ist der saisonale Wert bei t (periodische Schwankungen),
- Z_t ist die zyklische Komponente bei t (nicht-periodische Schwankungen),
- R_t ist die verbleibende Variation bei t (Restkomponente).

Wir erkennen Trend (Zeiten wurden besser, aber scheinen inzwischen gesättigt) und Restkomponente (unbeobachtete Einflüsse, zum Beispiel das Wetter), keine Saisonalitäten oder Zyklen (dafür gäbe es hier auch keine Interpretation).

b) Bitte passen Sie polynomiale Trendmodelle verschiedenen Grades mittels linearer Regression an die Daten an. Wie würden Sie den Polynomgrad wählen und warum?

Wir betrachten polynomiale Trendmodelle der Form $T_t = \beta_0 + \beta_1 t + \cdots + \beta_k t^k$ für verschiedene $k \in \mathbb{N}$. Zur Bestimmung von k können wir F-Tests, Modellselektionskriterien und Kreuzvalidierung betrachten. Siehe R Code für die Anpassungen und Modellselektionen.

c) Im Jahr 2005 wurde die Laufstrecke um 500 Meter verlängert. Modellieren Sie an dieser Stelle einen Strukturbruch in der Zeitreihe und testen Sie auf statistische Signifikanz.

Siehe R Code für die Modellierung des Strukturbruchs und Test auf Signifikanz.

Aufgabe 3 (Trendbereinigung mittels variate-difference Methode)

Betrachten Sie die folgenden drei Zeitreihen für $t = 1, \ldots, T$ mit Restkomponente u_t:

- $a_t = 3 + u_t$
- $b_t = a_t + 0.4t$
- $c_t = b_t + 0.3t^2$

a) Bitte berechnen Sie jeweils die erste und zweite Differenz der drei Zeitreihen.

- $\Delta a_t = \Delta u_t$, $\Delta^2 a_t = \Delta^2 u_t$
- $\Delta b_t = 0.4 + \Delta u_t$, $\Delta^2 b_t = \Delta^2 u_t$
- $\Delta c_t = 0.1 + 0.6t + \Delta u_t$, $\Delta^2 c_t = 0.6 + \Delta^2 u_t$

b) Die Restkomponente u_t sei eine standardnormalverteilte und unabhängige Zufallsvariable. Welchen Erwartungswert und Varianz haben die Zeitreihen sowie ihre erste und zweite Differenz für gegebenes t jeweils?

- $\mathrm{E}(a_t) = 3$, $\mathrm{E}(\Delta a_t) = 0$, $\mathrm{E}(\Delta^2 a_t) = 0$
- $\mathrm{Var}(a_t) = 1$, $\mathrm{Var}(\Delta a_t) = 2$, $\mathrm{Var}(\Delta^2 a_t) = 6$
- $\mathrm{E}(b_t) = 3 + 0.4t$, $\mathrm{E}(\Delta b_t) = 0.4$, $\mathrm{E}(\Delta^2 b_t) = 0$
- $\mathrm{Var}(b_t) = 1$, $\mathrm{Var}(\Delta b_t) = 2$, $\mathrm{Var}(\Delta^2 b_t) = 6$
- $\mathrm{E}(c_t) = 3 + 0.4t + 0.3t^2$, $\mathrm{E}(\Delta c_t) = 0.1 + 0.6t$, $\mathrm{E}(\Delta^2 c_t) = 0.6$
- $\mathrm{Var}(c_t) = 1$, $\mathrm{Var}(\Delta c_t) = 2$, $\mathrm{Var}(\Delta^2 c_t) = 6$

c) Simulieren Sie die Zeitreihe c_t bis $T = 100$ und berechnen Sie die erste und zweite Differenz. Visualisieren Sie anschließend das Ergebnis.

```
T <- 100
t <- 1:T
u <- rnorm(T)
c <- ts(3 + 0.4*t + 0.3*t^2 + u)
plot(c)
plot(diff(c))
plot(diff(c, difference = 2))
```