

Übungen zu Angewandte Statistik

(Abgabe am Montag, den 23.05.2016, 12:15h)

1. Es sei $\vec{Z} \sim N_{\mu, K}$ ein n -dimensionaler Zufallsvektor und $A \in \mathbb{R}^{r_1 \times n}$, $B \in \mathbb{R}^{r_2 \times n}$ seien beliebige Matrizen mit $r_1 \leq r_2$. Weiter Sei C eine symmetrische, nicht negativ-definite $n \times n$ Matrix.

(a) Beweise die folgenden Aussagen:

- Falls $AKB^T = 0$, dann sind $A\vec{Z}$ und $B\vec{Z}$ unabhängig.
- Falls $AKC = 0$, dann sind $A\vec{Z}$ und $\vec{Z}^T C \vec{Z}$ unabhängig.

(b) Betrachte das lineare Modell $\vec{X} = A\vec{\beta} + \vec{\varepsilon}$, $\vec{\varepsilon} \sim N_{0, \sigma^2 I}$. Zeige, dass sich der ML-Schätzer

$$\hat{\sigma}^2 = \frac{1}{n-r} \|\vec{X} - A\vec{\beta}\|^2$$

als Bilinearform

$$\hat{\sigma}^2 = \frac{1}{n-r} \varepsilon^T \tilde{P} \varepsilon, \tilde{P} = I - A(A^T A)^{-1} A^T$$

schreiben lässt.

(c) Zeige, dass die ML-Schätzer $\hat{\sigma}$, $\hat{\beta}$ unabhängig sind.

(6 Punkte)

2. Wir betrachten die Daten über Weltrekorde von Laufstrecken aus der Vorlesung. Lade ihn von der Vorlesungshomepage herunter.

(a) Erzeuge einen `data.frame` `weltrek.ma`, welcher nur die Ergebnisse von Männern enthält. Entferne zudem die Einträge, welche `NA`-Einträge aufweisen. (Hinweis: der Befehl `is.na()` ist hier nützlich).

(b) Erzeuge ein Streudiagramm (X_i, Y_i) , wobei X_i die Laufzeit in Sekunden und Y_i der Strecke in Metern entspricht.

(c) Mache dich mit den R-Befehlen `lm()`, `lm()$coeff`, sowie `summary(lm())` vertraut. Passe folgende lineare Modelle an den Datensatz an:

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.
- $Y_i = \beta_0 + \beta_1 \log X_i + \varepsilon_i$.
- $\log Y_i = \beta_0 + \beta_1 \log X_i + \varepsilon_i$,

wobei $\varepsilon_i \sim N_{0,1}$.

(d) Füge dem Streudiagramm aus a) die drei Regressionskurven hinzu (**Hinweis:** Es ist möglicherweise nötig die Ergebnisse der Regression aus c) zu sortieren, damit der `lines`-Befehl nur die direkt benachbarten Punkte miteinander verbindet). Welches Modell sollte verwendet werden und warum? Argumentiere über die Informationen, die über `summary(lm())` erhalten werden.

Hinweis: Der nötige Stoff zu dieser Aufgabe wird in der Vorlesung vom 18.05 behandelt.

(6 Punkte)