



## Statistik II

### Übungsblatt 2

**Abgabe:** Do. 2.11.2006, 10.00 Uhr, Briefkasten: Statistik II.

Bei jeder Aufgabe können maximal 5 Punkte erreicht werden.

1. Folgende Daten stammen von einem bei der Papierherstellung durchgeführten Experiment. Dabei wurde in der Phase des Blattpressens mit fünf verschiedenen Druckstärken gearbeitet. Ziel des Experiments war es, den Einfluß der Druckstärke auf die Reißfestigkeit des Papiers zu untersuchen. Für jede der fünf Druckstärken  $A, B, C, D$  und  $E$  wurden vier Blatt Papier dem Reißtest unterzogen. Die angegebenen Werte geben den sogenannten Reißfaktor an, eine Maßzahl, die den Prozentsatz einer Standardkraft widerspiegelt, der benötigt wird, das Blatt zu zerreißen.

Druckstärke beim Blattpressen	Reißfaktor			
A	112	119	117	113
B	108	99	112	118
C	120	106	102	109
D	110	101	99	104
E	100	102	96	101

Führen Sie eine Varianzanalyse für dieses Experiment durch! Erläutern Sie Ihre Vorgehensweise und beschreiben Sie Ihre Ergebnisse und Interpretationen.

2. Tatsächlich stellt sich jedoch heraus, daß die in Aufgabe 1 verwendeten Kategorien für Druckstärken stehen, die auf logarithmischer Skala äquidistant gewählt sind, und zwar  $A = 35.0$ ,  $B = 49.5$ ,  $C = 70.0$ ,  $D = 99.0$  und  $E = 140.0$ . Führen Sie unter diesen geänderten Gegebenheiten eine lineare Regression durch! Wie unterscheiden sich die Ergebnisse zu denen aus Aufgabe 1?
3. "SMSA-Datensatz":
  - (a) Führen Sie eine lineare Regression für die Mortalitätsrate in Abhängigkeit von Bevölkerungsdichte, Anteil Nichtweißer, Einkommen, Stickoxidbelastung und jährlicher Niederschlagsmenge durch.  
Gehen Sie dabei schrittweise vor, einmal vorwärts (d.h. schrittweise Hinzunahme von Variablen) und einmal rückwärts (d.h. beginnend mit dem vollen Modell, schrittweises Entfernen von Variablen).
  - (b) Vergleichen Sie die resultierenden Modelle und verteidigen Sie die jeweilige Vorgehensweise gegenüber Ihrem Auftraggeber.
  - (c) Welche anderen Methoden zur Variablenselektion wären denkbar?
  - (d) Berechnen Sie in Ihrem Lieblingsmodell Hebelwirkungen und Cook's  $D$ . Erläutern und interpretieren Sie Ihr Ergebnis. Welche Punkte würden Sie als Ausreißer deklarieren? Lassen sich diese Punkte alle über die Hebelwirkung bzw. Cook's  $D$  finden?
  - (e) Betrachten Sie einen Parallelkoordinatenplot für `mortality` und die Variablen Ihres Modells. Finden Sie, daß das Konzept der Hebelwirkung bzw. Cook's  $D$  dem optischen Eindruck von den Daten gerecht wird? Ergibt sich ein stimmiges Bild oder fallen Ihnen auch Punkte auf, die Sie nicht als Ausreißer deklarieren würden?

4. Gegeben sei das ANOVA-Modell

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + \epsilon_{ij}, \\ \sum_i n_i \alpha_i &= 0, \quad i = 1, \dots, k, j = 1, \dots, n_i, \epsilon_{ij} \sim N(0, \sigma^2), \sigma^2 \text{ unbekannt}, \end{aligned} \quad (1)$$

also  $k$  Gruppen mit je  $n_i, i = 1, \dots, k$  Beobachtungen.

(a) Zeigen Sie: Das ANOVA-Modell (1) kann als lineares Modell der Form

$$Y = X\beta + \epsilon, \quad Y, \epsilon \in \mathbb{R}^{p+1}, X \in \mathbb{R}^{n \times p+1}, E(\epsilon) = 0 \quad (2)$$

geschrieben werden, wobei  $X$  nicht notwendigerweise vollen Rang besitzt.

(b) Definition: Ein *Kontrast*  $C$  ist eine lineare Funktion der (wahren) Mittelwerte  $\alpha_i$  mit

$$C = \sum_i c_i \alpha_i \text{ und } \sum_i c_i = 0, c_i \in \mathbb{R}, i = 1, \dots, k.$$

Zeigen Sie, daß im Modell (2) nur Kontraste schätzbar sind.

5. Verständnisfragen

- Erläutern Sie die Konzepte von Hebelwirkung (leverage) und Cook's  $D$ . Was sagen diese Größen aus (keine Formeln!) und was ist der Hauptunterschied zwischen ihnen?
- Beide Maßzahlen beschreiben den Einfluß einzelner Punkte. Wie wird das realisiert? Was ist die Schwäche dieses Ansatzes?
- Zur Modellfindung wird unter anderem auch die sogenannte Schrittweise Regression vorgeschlagen. Erläutern Sie diese Idee. Müssen Vorwärts- und Rückwärtsselektion immer zum selben Modell kommen? Begründen Sie Ihre Antwort.
- Was ist Ihrer Meinung nach die Schwäche dieser Verfahren?