



Statistik II

Übungsblatt 5

Abgabe: Do. 16.11.2006, 10.00 Uhr, Briefkasten: Statistik II.

Bei jeder Aufgabe können maximal 5 Punkte erreicht werden.

1. "COOP"-Datensatz:

Bei einer chemischen Analyse wurden 7 verschiedene Präparate (S_{PC}) in 6 verschiedenen Laboratorien (Lab) untersucht und dabei die Konzentration ($Conc$) eines bestimmten Stoffes gemessen. Nach jeweils drei Monaten wurden Wiederholungsmessungen (Bat) durchgeführt.

- Erstellen Sie ein sinnvolles lineares Modell für $Conc$ in Abhängigkeit der anderen Variablen für den Teildatensatz des Präparats $S1$.
- Begründen Sie Ihre Modellwahl und interpretieren Sie Ihre Ergebnisse.
- In welcher Form würden Sie die Variable S_{PC} in das Modell einfließen lassen?
- Führen Sie eine lineare Modellierung des kompletten Datensatzes durch.

2. "Bank"-Datensatz:

Analysieren Sie den Bank-Datensatz mit linearen Modellen. Es soll analysiert werden, ob die Variable "Minority" (Rasse) einen Einfluss auf das aktuelle Gehalt der Personen ($Salnow$) hat.

- Betrachten Sie den Einfluss der anderen Variablen auf "Salnow"!
- Schauen Sie sich nun an, ob Abhängigkeiten zwischen "Minority" und den anderen Variablen bestehen.
- Unter Hinzunahme der Informationen aus (a) und (b) erstellen Sie ein geeignetes Modell zur Beantwortung der Frage nach dem Einfluss von "Minority" (Rasse) auf das aktuelle Gehalt, welches alle notwendigen Faktoren und Interaktionen beinhaltet.

3. "SMSA-Datensatz":

- Führen Sie eine lineare Regression für die Mortalitätsrate durch.

- Starten Sie mit dem saturierten Modell.
- Eliminieren Sie alle nicht signifikanten Variablen.
- Nehmen Sie nun alle Ausreißer heraus, sie Ihrer Meinung nach das Modell zu stark beeinflussen.
- Vergleichen Sie Ihr bestes Modell mit dem aus (a). Was hat sich geändert?

4. Verständnisfragen

Sie wollen die Körpergröße von Personen linear durch das Gewicht und das Geschlecht modellieren. Man könnte entweder getrennte Regressionen für die Frauen und für die Männer berechnen, also zwei Modelle

$$E[y_{if}] = \alpha_f + \beta_f \cdot x, \quad E[y_{im}] = \alpha_m + \beta_m \cdot x,$$

und die Steigungen der Geraden vergleichen oder ein gemeinsames lineares Modell der Form

$$E[y_{ij}] = \mu + \alpha_j + \beta_j \cdot x, \quad j = m, f$$

aufstellen und die Schätzungen für die Parameter β_m und β_f vergleichen. Erläutern Sie die Unterschiede der zwei Vorgehensweisen.

5. "Cats"-Datensatz:

Im Cats-Datensatz wurde jeweils für 144 Katzen das Geschlecht (Sex), das Körpergewicht (Bwt) und das Gewicht des Herzens (Hwt) erhoben. Der Datensatz ist in der R Library "MASS" zu finden. Bearbeiten Sie jeweils mit R:

- (a) Wenden Sie die Situation aus Aufgabe 4 auf die Schätzung des Herzgewichts an, und bestimmen sie die jeweiligen Koeffizienten!
- (b) Sei y eine stetige Variable und A, B, C und D vier Faktoren. Notieren Sie jeweils die Modellformel in R-Notation für das Modell
 - i. mit allen Faktoren, aber ohne Interaktionen
 - ii. mit allen 2-fach Interaktionen
 - iii. mit allen Interaktionen von Grad kleiner 4 und ohne die Interaktion zwischen A, C und D