h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

Antje Jahn
Multivariate Statistik
Sommersemester 2019

Arbeitsblatt 6

A 1 In einer Studie wird die Korrelation zwischen dem Prostata-spezifischen Antigen (PSA) mit verschiedenen klinischen Parametern untersucht. Die Stichprobe besteht aus Daten von 97 Männern vor radikaler Prostatektomie, d.h. Entfernung der Prostata. Die Daten finden Sie auf moodle. Die folgenden Variablen sind erfasst:

- psa: PSA-Wert (Zielgröße)
- cavol: Volumen des Tumors
- svi: Tumorausbreitung in die Bläschendrüse (ja=1/nein=0)
- age: Alter des Patienten
- gleason: Gleason Score (Score aus zwei Gleason-Graden des Tumorgewebes, der die Ausdifferenzierung des Tumors beschreibt; höhere Werte sind mit einer schlechteren Prognose assoziiert)
- pgg45: % der untersuchten Tumorgewebe mit Grad 4 oder 5 (höhere Werte sind mit einer schlechteren Prognose assoziiert)

a) Führen Sie eine lineare Regression mit psa als abhängiger Variable durch, betrachten Sie die diagnostischen Plots und entscheiden Sie sich ggf. anhand dieser Plots für Transformationen von Ziel- und/oder Einflussgrößen. Verbessert Ihr ausgewähltes Modell die Modellanpassung? Woran machen Sie dies fest?

b) Führen Sie mit den ggf. transformierten Variablen verschiedene Variablenselektionsverfahren (stepwise forward (`MASS::stepAIC()`), stepwise backward(`MASS::stepAIC()`), subset (`leaps::regsubset()`)) durch und entscheiden Sie sich für ein abschliessendes Modell.

c) Vergleichen Sie die Residuenanalysen zwischen dem Basismodell (ohne Variablentransformationen und -selektionen) und ihrem final selektierten Modell. Haben sich die Ergebnisse verbessert? Identifizieren Sie in Ihren Residuenplots auffällige Beobachtungen im Basismodell. Sind diese im finalen Modell weiterhin auffällig? Falls nicht, durch welchen Schritt in der Modellwahl wurde dies vermutlich behoben?

A 2 Es sei X eine p-dimensionale Zufallsvariable mit Kovarianzmatrix Σ. Als symmetrische Matrix ist Σ diagonalisierbar, d.h. es gibt eine orthogonale $p \times p$ -Matrix A und eine Diagonalmatrix Λ, so dass $\Sigma = A\Lambda A^T$. Betrachten Sie die transformierten Zufallsvariablen

$$Y_1 := (A^T X)_1, \qquad Y_2 := (A^T X)_2, \qquad \dots$$

Zeigen Sie, dass die Y_i unkorreliert sind.

Hinweis: Nutzen Sie die Rechenregeln für Kovarianzmatrizen und die Orthogonalität von A.

A 3 Lesen Sie den Datensatz *framingham* in `R` ein (Moodle). Hierin sind für 1352 Probanden u.a. der diastolische Blutdruck DBP (X_1) und der systolische Blutdruck SBP (X_2) erfasst.

a) Erstellen Sie einen Scatterplot und berechnen Sie die Korrelation zwischen DBP und SBP.

b) Wählen Sie einen beliebigen **normierten** Vektor $a^T = (a_1, a_2)$, definieren Sie

$$Y := a^T X = a_1 X_1 + a_2 X_2$$

und berechnen Sie für jeden Probanden i die Variable Y_i. Wieviel Prozent der Gesamtvarianz erklärt die Variable Y, d.h. berechnen Sie den Anteil erklärter Varianz $\frac{Var(Y)}{Var(X_1)+Var(X_2)}$.

c) Finden Sie einen Vektor a, bei dem der Anteil erklärter Varianz besonders groß wird. Programmieren Sie dazu eine R-Funktion, die als Funktion in a_1 den Anteil erklärter Varianz berechnet und ermitteln Sie z.B. mit der R-Funktion optimise die Maximalstelle dieser Funktion.

d) Die R-Funktion `princomp(X)` schlägt Ihnen einen Vektor a vor, den Sie durch `loadings(princomp(X))[,1]` erhalten. Vergleichen Sie diesen mit Ihrem Ergebnis aus Teil c).

e) Nun sollen Sie die beiden Variablen X_1 und X_2 zunächst standardisieren (Funktion `scale`) und a)-d) mit den standardisierten Variablen ausführen. Wie ändert sich das Ergebnis? Erklären Sie diese Veränderung.