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

Antje Jahn
Multivariate Statistik
Sommersemester 2019

Arbeitsblatt 5

A 1 Zeigen Sie, dass im multiplen linearen Regressionsmodell gilt

$$Cov(Y_i, \hat{Y}_i) = h_{ii} \quad Var(Y_i) = \sigma^2 \quad Var(\hat{Y}_i) = h_{ii}\sigma^2 \quad Cor(Y_i, \hat{Y}_i) = \sqrt{h_{ii}}$$

wobei h_{ii} das i-te Diagonalelement von $H = X(X^T X)^{-1} X^T$ Hinweis: Schreiben Sie $\hat{Y}_i$ als $(X\hat{b})_i$ und setzen Sie für $\hat{b}$ den KQ-Schätzer ein. Rechnen Sie dann mit den bekannten Rechenregeln für Kovarianzmatrizen.

A 2 Untersuchen Sie nochmals im Datensatz Auto (ISLR) den Zusammenhang zwischen Meilen per Gallon und verschiedenen Einflussvariablen.

a) Führen Sie jeweils für das Modell mit mpg und log(mpg) als abhängiger Variable und PS, Jahr und Ursprungsland als unabhängigen Variablen eine Modelldiagnostik durch, indem Sie sich geeignete Residuenplots anschauen. Welches Modell erscheint Ihnen anhand der Plots passender?

b) Weisen die diagnostischen Plots auf auffällige Beobachtungen hin? Schauen Sie sich diese im Datensatz an: In welcher Hinsicht sind diese Beobachtungen auffällig?

c) Führen Sie ein Subset-Selektionsverfahren mit allen verfügbaren Variablen (ausser Fahrzeugname) durch. Plotten Sie für das subset-Verfahren die Modellkomplexität (Anzahl an Kovariaten) gegen das minimale BIC, das minimale AIC (äquivalent zu Cp) und das maximale R^2 innerhalb der jeweiligen Komplexität. Für welches Modell würden Sie sich jeweils entscheiden? Begründen Sie den Unterschied.

d) Berechnen Sie Varianzinflationsfaktoren in einem Modell mit allen Kovariablen und in Ihrem selektierten Modell. Würden Sie auf Basis der Varianzinflationsfaktoren weitere Kovariablen aus dem selektierten Modell entfernen?

A 3

M1 und M2 seien zwei verschachtelte lineare Regressionsmodelle, d.h. M1 ist ein Modell mit p Kovariaten $X_1, \ldots, X_p$ und M2 ein Modell mit $q > p$ Kovariaten $X_1, \ldots, X_p, X_{p+1}, \ldots X_q$. Der Likelihood-Ratio-Test prüft die Nullhypothese

$$H_0 = \{b_{p+1} = b_{p+2} = \cdots = b_q = 0\}$$

mit der Teststatistik

$$T = 2 \cdot (LL_{M2} - LL_{M1}) \underset{H_0}{\sim} \chi_q^2$$

Dabei sind LL_{M1} und LL_{M2} die Werte der Log-Likelihood-Funktion an der Stelle des Maximum-Likelihood-Schätzers aus Modell M1 bzw. M2.

Zur Entscheidung innerhalb eines Variablenselektionsverfahrens, ob zu einem linearen Regressionsmodell (M1) mit p Kovariaten eine weitere Kovariate X_{p+1} hinzugefügt wird (M2), können folgende Entscheidungskriterien herangezogen werden

i) X_{p+1} wird in das Modell aufgenommen wenn sich dadurch das AIC verringert, d.h. $AIC_{M2} < AIC_{M1}$

ii) X_{p+1} wird in das Modell aufgenommen wenn der LR-Test ein zum Signifikanzniveau $\alpha = 0.1572992$ signifikantes Ergebnis liefert.

a) Zeigen Sie, dass beide Vorgehensweisen i) und ii) näherungsweise zur selben Entscheidung kommen. Nutzen Sie dabei dass das AIC-Kriterium auch als $AIC = -2LL + 2k$ formuliert werden kann mit k=Anzahl Modellparameter

b) Überprüfen Sie die Äquivalenz exemplarisch, d.h. für einige beliebig ausgewählte Modelle und Variablen, an dem Auto-Datensatz (Die Funktion `lmtest::lrtest()` berechnet Likelihood-Ratio-Tests)