Vorlesung: Lineare Modelle

Prof. Dr. Helmut Küchenhoff

Institut für Statistik, LMU München

SoSe 2015

0 Einführung und Beispiele

1 Das einfache lineare Regressionsmodell

2 Das multiple lineare Regressionsmodell

3 Quadratsummenzerlegung und statistische Inferenz im multiplen linearen Regressionsmodell

5 Diskrete Einflußgrößen: Dummy- und Effektkodierung, Mehrfaktorielle Varianzanalyse

Quadratsummenzerlegung

Gegeben sei das Modell (2.1) mit Design-Matrix X und

$$rg(X) = p + 1 =: p'. \tag{3.1}$$

Dann gilt:

$$\underbrace{(Y - \bar{Y})'(Y - \bar{Y})}_{\text{SST}} = \underbrace{(Y - \hat{Y})'(Y - \hat{Y})}_{\text{SSE}} + \underbrace{(\hat{Y} - \bar{Y})'(\hat{Y} - \bar{Y})}_{\text{SSM}} \tag{3.2}$$

Interpretation:

SST : Gesamt-Streuung, (korrigierte) Gesamt-Quadratsumme, "Total"
SSE : Fehler-Quadratsumme, "Error"
SSM : Modell-Quadratsumme, "Model"

Zerlegung ohne Absolutglied

Die Zerlegung (3.2) setzt ein Absolutglied in der Regression voraus.

Weitere Zerlegung, die nicht notwendig ein Absolutglied in der Regression voraussetzt:

$$\underbrace{Y'Y}_{\text{SST*}} = \underbrace{(Y - \hat{Y})'(Y - \hat{Y})}_{\text{SSE}} + \underbrace{\hat{Y}'\hat{Y}}_{\text{SSM*}} \tag{3.3}$$

SST^*	:	nicht korrigierte Gesamt-Quadratsumme Erfasst auch Abweichungen von $Y = 0$ und nicht nur von $\bar{Y}$.
SSE	:	Fehler-Quadratsumme, wie bei (3.2)
SSM^*	:	nicht korrigierte Modell-Quadratsumme

Nachweis der Quadratsummenzerlegung (3.3)

$$\begin{aligned} Y'Y &= Y'IY \\ &= Y'(Q+P)Y \\ &= Y'QY + Y'PY \\ &= Y'Q'QY + Y'P'PY \\ &= \hat{\varepsilon}'\hat{\varepsilon} + \hat{Y}'\hat{Y} \qquad\qquad \text{qed} \end{aligned}$$

P und Q Projektionsmatrizen.

Nachweis der Quadratsummenzerlegung (3.2) I

Wir definieren:

$$P_e = e(e'e)^{-1}e', \quad Q_e = I - P_e \text{ mit } e = (1, 1, \dots, 1)'. \tag{3.4}$$

P_e entspricht der Regression auf eine Konstante ($E(Y) = \beta_0$)

$$\begin{aligned} P_e Y &= \quad \bar{Y} && \text{(„Mittelwert bilden“)} \\ Q_e Y &= \quad Y - \bar{Y} && \text{(„Mittelwertsbereinigung“)} \end{aligned}$$

$$\begin{aligned} (Y - \bar{Y})'(Y - \bar{Y}) &= Y'Q_e'Q_eY = Y'Q_eY \\ &= Y'Q_e(Q + P)Y = Y'Q_eQY + Y'Q_ePY \\ &\overset{(*)}{=} \dots \end{aligned}$$

Nachweis der Quadratsummenzerlegung (3.2) II

$$(*) \quad Q_e QY = Q_e\hat{\varepsilon} \underset{\uparrow}{=} \hat{\varepsilon} = QY$$

„Residuen haben Mittelwert 0“

$$(*) \quad Q_e P Q_e P Y = Q_e P Q_e \hat{Y} = Q_e P(\hat{Y} - \bar{Y})$$
$$= Q_e P\hat{Y} - Q_e P\bar{Y} \underset{\uparrow}{=} Q_e\hat{Y} - Q_e\bar{Y} = Q_e\hat{Y} - 0$$

„Regression auf $\hat{Y}$ und $\bar{Y}$ liefert Identität“

$$\begin{aligned}
(Y - \bar{Y})'(Y - \bar{Y}) &= Y'Q_e'Q_eY = Y'Q_eY \\
&= Y'Q_e(Q + P)Y = Y'Q_eQY + Y'Q_ePY \\
&\overset{(*)}{=} Y'QY + Y'Q_ePQ_ePY \\
&= \hat{\varepsilon}'\hat{\varepsilon} + Y'P'Q_ePY \\
&= \hat{\varepsilon}'\hat{\varepsilon} + (\hat{Y} - \bar{Y})'(\hat{Y} - \bar{Y})
\end{aligned}$$

Erwartungswerte der Quadratsummen

Wir betrachten das multiple Regressionsmodell (2.1) mit (2.2) bis (2.4) und

$$P_e = e(e'e)^{-1}e', \quad Q_e = I - P_e \text{ mit } e = (1, 1, \dots, 1)'.$$

Dann gilt für die Erwartungswerte der Quadratsummen:

$$\begin{aligned}
E(SST^*) = E(Y'Y) &= \sigma^2 n + \beta' X' X \beta && (3.5)\\
E(SST) = E(Y - \bar{Y})'(Y - \bar{Y}) &= \sigma^2 (n-1) + \beta'(Q_e X)'(Q_e X)\beta && (3.6)\\
E(SSE) = E(\hat{\varepsilon}'\hat{\varepsilon}) &= \sigma^2 (n - p') && (3.7)\\
E(SSM^*) = E(\hat{Y}'\hat{Y}) &= \sigma^2 p' + \beta' X' X \beta && (3.8)\\
E(SSM) = E(\hat{Y} - \bar{Y})'(\hat{Y} - \bar{Y}) &= \sigma^2 p + \beta'(Q_e X)'(Q_e X)\beta && (3.9)
\end{aligned}$$

Bemerkungen

1. Die stochastischen Eigenschaften der Quadratsummen werden zur Konstruktion von Tests bezüglich β genutzt
2. $\beta = 0 \Longrightarrow E(Y'Y) = \sigma^2 n$
3. $\beta_1, ...\beta_p = 0 \Longrightarrow E(SSM) = \sigma^2 p$

Zum Nachweis von 3. benutzt man, dass Q_e der „Mittelwertsbereinigungs- Operator“ ist:

$$Q_e x = (I - P_e)x = x - \bar{x}$$

Die erste Spalte von $Q_e X$ ist also der Nullvektor.

Beweisidee

Allgemein berechnet man den Erwartungswert von quadratischen Formen wie folgt:

$$
\begin{aligned}
E(Y'AY) &= E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} Y_i Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} E(Y_i Y_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} E(Y_i) E(Y_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} Cov(Y_i, Y_j) \\
&= E(Y)' A E(Y) + Sp(AV(Y)) \\
V(Y) &:= \text{Varianz-Kovarianzmatrix von Y}
\end{aligned}
$$

Unter Benutzung von $Sp(AV(Y)) = sp(A) * \sigma^2 = rang(A) * \sigma^2$ für Projektionsmatrizen erhält man obige Identitäten.

Mittlere Quadratsummen

Wir definieren entsprechend der Zahl der Freiheitsgrade die **mittleren Quadratsummen**:

$$
\begin{aligned}
MST^* &:= \frac{SST^*}{n} && (3.10)\\
MST &:= \frac{SST}{n-1} && (3.11)\\
MSE &:= \frac{SSE}{n-p'} && (3.12)\\
MSM^* &:= \frac{SSM^*}{p'} && (3.13)\\
MSM &:= \frac{SSM}{p} && (3.14)
\end{aligned}
$$

Verteilungsdefinitionen

Normalverteilung

Ein n-dimensionaler Zufallsvektor Z heißt multivariat normalverteilt, falls für seine Dichtefunktion gilt:

$$f_z(z) = \frac{1}{\sqrt{\det \Sigma}\sqrt{(2\pi)^n}} \exp\left[-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\right] \qquad (3.15)$$

mit positiv definiter, symmetrischer Matrix Σ.

Bezeichnung: $Z \sim N_n(\mu, \Sigma)$

Eigenschaften der Normalverteilung

Ist $Z \sim N_n(\mu, \Sigma)$, so gilt:

1. **Momente:**
$$\begin{aligned} E(Z) &= \mu \\ V(Z) &= \Sigma \end{aligned}$$

2. **Lineare Transformationen:**
Ist $A : R^n \longrightarrow R^m$ eine lineare Tranformation mit $rg(A) = m$
$$\Longrightarrow AZ \sim N_m(A\mu, A\Sigma A') \tag{3.16}$$

3. **Orthogonale Transformation in unabhängige Komponenten (Spektralzerlegung)**
Es existiert eine Matrix $T \in R^{n \times n}$ mit $T'T = I$ und $T\Sigma T' = \text{ diag } (\lambda_1, \ldots, \lambda_n)$, so dass
$$TZ \sim N_n(T\mu, \text{ diag } (\lambda_1, \ldots, \lambda_n)) \text{ gilt.} \tag{3.17}$$

Chi-Quadrat-Verteilung I

Ist $Z \sim N_n(\mu, I)$ so heißt $X = Z'Z$ (nicht-zentral) Chi-Quadrat-verteilt.

Bezeichnung: $X \sim \chi^2(n, \delta)$

n:	Zahl der Freiheitsgrade
$\delta := \mu'\mu$:	Nichtzentralitätsparameter

Im Fall $\delta = 0$ erhält man die zentrale $\chi^2(n)$-Verteilung.

Chi-Quadrat-Verteilung II

Eigenschaften: Ist $X \sim \chi^2(n, \delta)$, so gilt:

1. **Momente:**

$$\begin{aligned} E(X) &= n + \delta \\ V(X) &= 2n + 4\delta \end{aligned}$$

2. **Allgemeiner Bezug zur Normalverteilung**

$$Z \sim N_n(\mu, \Sigma) \quad \Longrightarrow Z'\Sigma^{-1}Z \sim \chi^2(n, \mu'\Sigma^{-1}\mu) \tag{3.18}$$

t-Verteilung

Seien Z und W voneinander unabhängige Zufallsgrößen mit

$$\begin{aligned} Z &\sim N(\delta, 1), \\ W &\sim \chi^2(n). \end{aligned}$$

Dann heißt $X = \frac{Z}{\sqrt{\frac{W}{n}}}$ (nicht-zentral) t-verteilt.

Bezeichnung: $X \sim t(n, \delta)$

n: Zahl der Freiheitsgrade,

δ: Nicht-Zentralitätsparameter

Erwartungswert: Ist $X \sim t(n, \delta)$,so gilt:

$$E(X) = \delta \sqrt{\frac{n}{2}} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} \quad \text{für } n > 1. \tag{3.19}$$

F-Verteilung

Seien W_1 und W_2 voneinander unabhängige Zufallsgrößen mit

$$\begin{array}{rcl} W_1 & \sim & \chi^2(n_1, \delta) \\ W_2 & \sim & \chi^2(n_2) \end{array}$$

Dann heißt $X = \frac{W_1/n_1}{W_2/n_2}$ (nicht-zentral) F-verteilt.

Bezeichnung: $X \sim F(n_1, n_2, \delta)$

n_1: Zählerfreiheitsgrade
n_2: Nennerfreiheitsgrade
δ: Nichtzentralitätsparameter

Erwartungswert: Ist $X \sim F(n_1, n_2, \delta)$, so gilt:

$$E(X) = \frac{n_2}{n_2 - 2}(1 + \delta/n_1) \text{ für } n_2 > 2. \tag{3.20}$$

Unabhängigkeit von Quadratsummen (Satz von Cochran)

Sei $Z \sim N(\mu, I)$, dim $Z =$ n,
A $\in R^{n \times n}$, $A' = A$, $rg(A) = r$, $A^2 = A$,
$B \in R^{n \times n}, B^2 = B,\ B = B'$,
$C \in R^{m \times n}$.

Dann gilt:

$$Z'AZ \sim \chi^2(r, \mu' A\mu) \tag{3.21}$$

$$CA = 0 \implies CZ \text{ und } Z'AZ \text{ sind unabh.} \tag{3.22}$$

$$AB = 0 \implies Z'AZ \text{ und } Z'BZ \text{ sind unabh.} \tag{3.23}$$

Beweis von (3.21)

Spektralzerlegung von A

$A = T\Sigma T'$

$$\Sigma = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & 0 \end{pmatrix} \quad \text{da } A^2 = A$$

$$\underbrace{}_{r}$$

$\Rightarrow Z'AZ = Z'T\Sigma T'Z$

Für $\widetilde{Z} := T'Z$ gilt: $Var(\widetilde{Z}) = T'IT = I$

$\Rightarrow Z'AZ = \widetilde{Z}'\Sigma\widetilde{Z} = \sum_{i=1}^{r} \widetilde{z}_i^2 \sim \chi^2(r, \mu' A\mu)$

Beweis von (3.22) und (3.23)

$$Cov(CZ, Z'A) = CCov(Z, Z')A = CA = 0$$

Damit sind CZ und $Z'A$ unabhängig (NV-Annahme!)
$\Rightarrow CZ$ und $(Z'A)(Z'A)' = Z'AZ$ unabh.

qed

$$Cov(AZ, Z'B) = ACov(Z, Z')B = AB = 0$$

Damit sind AZ und $Z'B$ unabhängig (NV-Annahme!)
$\Rightarrow (AZ)'(AZ) = Z'AZ$ und $(Z'B)(Z'B)' = Z'BZ$ unabh.

qed

Verteilung des KQ-Schätzers unter Normalverteilung

Sei multiples Regressionsmodell (2.1) mit (2.5) und $rg(X) = p'$ gegeben. Für den KQ-Schätzer $\hat{\beta}$ gilt:

$$
\begin{aligned}
\hat{\beta} &\sim N(\beta, \sigma^2 (X'X)^{-1}) && (3.24)\\
\hat{\Sigma}_{\hat{\beta}} &:= \hat{\sigma}^2 (X'X)^{-1} && (3.25)\\
(n-p')\frac{\hat{\sigma}^2}{\sigma^2} &\sim \chi^2(n-p') && (3.26)\\
\hat{\sigma} \text{ und } \hat{\beta} &\ \text{sind}\ \text{unabhängig} && (3.27)\\
\hat{\sigma}_{\hat{\beta}_k} &:= \sqrt{c_{kk}}\hat{\sigma}, && (3.28)\\
& \quad (c_{kk} \text{ entspr. Diagonalelement von } (X'X)^{-1})\\
\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}} &\sim t(n-p', 0) && (3.29)
\end{aligned}
$$

Beweis

(3.24) (Normalverteilung von β) folgt aus Eigenschaften der NV (3.16)

(3.26) folgt aus (3.21), da $\hat{\sigma}^2 = \frac{1}{n-p'}\hat{\varepsilon}'\hat{\varepsilon} = \frac{1}{n-p'} Y'QY$
$rg(Q) = sp(Q)$

(3.27) folgt aus (3.22):

$$C \in R^{p \times n}, CA = 0 \Longrightarrow CZ \text{ und } Z'AZ \text{ sind unabh.}$$

mit

$$C := (X'X)^{-1}X', \quad A := Q, \quad Z := Y$$
$$CA = (X'X)^{-1}X'Q = (X'X)^{-1}X'(I - P) =$$
$$(X'X)^{-1}X' - (X'X)^{-1}X'X(X'X)^{-1}X' = 0.$$

Overall-Tests

Sei Modell (2.1) mit (2.5) und $rg(X) = p'$ gegeben. Dann gilt für die mittlere Quadratsummen:

$$F_O = \frac{MSM}{MSE} \quad \sim \quad F(p, n - p', \sigma^{-2}\beta'(Q_e X)'(Q_e X)\beta) \qquad (3.30)$$

(3.30) folgt aus (3.21) und (3.23) mit

$$\begin{aligned} MSM &= SSM/p = Y'(P - P_e)Y/p \\ MSE &= SSE/(n - p') = Y'QY/(n - p') \\ Q(P - P_e) &= QP - QP_e = 0 \end{aligned}$$

Die Verteilung wird zur Konstruktion des folgenden Tests benutzt:
$H_0^O : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
Lehne H_0^O ab, falls

$$F_O > F_{1-\alpha}(p, n - p') \qquad (3.31)$$

Overall Test mit $\beta_0 = 0$

$$F_O^* = \frac{MSM^*}{MSE} \quad \sim \quad F(p', n - p', \sigma^{-2}\beta'(X'X)\beta) \tag{3.32}$$

$H_0^{O*} : \beta_0 = \beta_1 = \cdots = \beta_p = 0$

Lehne H_0^{O*} ab, falls

$$F_0^* > F_{1-\alpha}(p', n - p') \tag{3.33}$$

$F_{1-\alpha}(p, n - p')$: $(1 - \alpha)$-Quantil der zentralen $F(p, n - p')$–Verteilung.

Dieser Test wird nur in Ausnahmefällen angewendet.

Lineare Hypothesen

Es sollen Hypothesen, die sich mit Hilfe linearer Transformationen von β darstellen lassen, betrachtet werden:

$$A \in R^{a \times (p+1)} \ , c \in R^a,$$

$$A\beta = c \ \text{ mit } rg(A) = a$$

Beispiele:

$$p = 2 \ \ \beta_1 = \beta_2 \quad \leftrightarrow \quad A = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}, c = 0$$

$$p = 4 \ \ \beta_3 = \beta_4 = 0 \quad \leftrightarrow \quad A = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, c = 0$$

$$p = 4 \ \ \beta_2 = 3 \ \ \beta_3 + \beta_4 = 1 \quad \leftrightarrow \quad A = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}, c = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$$

Allgemeine lineare Hypothese, Wald-Test

Sei das Modell (2.1) mit (2.5) und $A \in R^{a \times p'}$, $rg(A) = a, c \in R^a$ gegeben.

$$\begin{aligned}
V(A\hat{\beta} - c) &= \sigma^2 A(X'X)^{-1}A' && (3.34)\\
SSH &:= (A\hat{\beta} - c)'(A(X'X)^{-1}A')^{-1}(A\hat{\beta} - c) && (3.35)\\
\sigma^{-2} SSH &\sim \chi^2(a, \sigma^{-2}(A\beta - c)'(A(X'X)^{-1}A')^{-1}(A\beta - c)) && (3.36)\\
MSH &:= \frac{SSH}{a} && (3.37)
\end{aligned}$$

$$\frac{MSH}{MSE} \sim F(a, n - p', \sigma^{-2}(A\beta - c)'(A(X'X)^{-1}A')^{-1}(A\beta - c)) \quad (3.38)$$

SSH: Quadratsumme, die die Abweichung von der Hypothese $A\beta = c$ beschreibt.

Test nach Wald: $H_0 : A\beta = c$. Lehne H_0 ab, falls:

$$TF = MSH/MSE > F_{1-\alpha}(a, n - p') \quad (3.39)$$

Overall-Test und zweiseitiger t-Test

Die Overall-Tests aus (3.31) und (3.33) sind spezielle Wald-Tests.

Nachweis von (3.33) erfolgt direkt, da hier A $=$ I ist.

Der Nachweis von (3.31) erfolgt später.

Der mit Hilfe von (3.29) konstruierte zweiseitige Test auf $\beta_k = \beta_k^0$ ist ebenfalls ein Wald-Test.

Der Nachweis erfolgt direkt unter Benutzung von $A = (0 \cdots 1 \cdots 0)$

Lineare Hypothese und Modell mit Restriktion

Wir betrachten das lineare Modell mit der linearen Restriktion $A\beta = c$.
Alle Schätzungen aus diesem Modell werden mit $\hat{}$ gekennzeichnet.

Wir lösen das Minimierungsproblem:

$$(y - X\beta)'(y - X\beta) \to \min \quad \text{unter} \quad A\beta = c$$

mit der Lagrange-Methode:

$$S(\beta, \lambda) = (y - X\beta)'(Y - X\beta) + 2\lambda'(A\beta - c)$$

λ ist der Vektor der Lagrange-Multiplikationen.

Berechnung der Lösung $\hat{\hat{\beta}}$

$$
\begin{aligned}
\frac{\partial S}{\partial \beta} &= 0 \Leftrightarrow -2X'Y + 2X'X\hat{\hat{\beta}} + 2A'\lambda = 0 \\
\frac{\partial S}{\partial \lambda} &= 0 \Leftrightarrow A\hat{\hat{\beta}} = c \\
A'\lambda &= X'Y - X'X\hat{\hat{\beta}} \quad | \cdot (X'X)^{-1} \\
(X'X)^{-1}A'\lambda &= \hat{\beta} - \hat{\hat{\beta}} \quad | \cdot A \ (*) \\
\Rightarrow A(X'X)^{-1}A'\lambda &= A\hat{\beta} - c \quad (A\hat{\hat{\beta}} = c) \\
\Rightarrow \lambda &= (A(X'X)^{-1}A')^{-1}(A\hat{\beta} - c) \\
(*) \Rightarrow \hat{\hat{\beta}} &= \hat{\beta} - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - c) \\
w &:= (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - c)
\end{aligned}
$$

Darstellung von SSH

$$
\begin{aligned}
\hat{\hat{Y}} &= X\hat{\hat{\beta}} = \hat{Y} - Xw \\
\hat{\hat{\varepsilon}} &= Y - \hat{\hat{Y}} = \hat{\varepsilon} + Xw \\
\Rightarrow \hat{\hat{\varepsilon}}'\hat{\hat{\varepsilon}} &= \hat{\varepsilon}'\hat{\varepsilon} + w'X'Xw \;\; (\text{da} \;\; X'\hat{\varepsilon} = 0) \\
&= \hat{\varepsilon}'\hat{\varepsilon} + (A\hat{\beta} - c)'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - c) \\
\Rightarrow SSH &= \hat{\hat{\varepsilon}}'\hat{\hat{\varepsilon}} - \hat{\varepsilon}'\hat{\varepsilon}
\end{aligned}
$$

Damit haben wir eine andere Möglichkeit, SSH aus den Residuenquadratsummen zu berechnen.

Lineare Hypothese (alternative Durchführung)

Sei das Modell (2.1) mit (2.5) und $A \in R^{a \times p'}$, $rg(A) = a, c \in R^a$ gegeben.

$$H_0 : A\beta = c$$

Berechne SSE des Modells und des Modells unter H_0:

$$
\begin{aligned}
SSE &= \hat{\varepsilon}'\hat{\varepsilon}' \\
SSE(H_0) &:= \hat{\hat{\varepsilon}}'\hat{\hat{\varepsilon}} && (3.40) \\
SSH &= SSE(H_0) - SSE && (3.41) \\
MSH &= SSH/a \\
TF &:= MSH/MSE = \frac{(SSE(H_0) - SSE)/a}{SSE/(n-p')} && (3.42) \\
TF &\sim F(a, n-p') \text{ unter } H_0 && (3.43)
\end{aligned}
$$

Das Vorgehen entspricht dem Wald- Test (3.39).

Spezialfall: Overall - Test

Overall- Test

$$H_0^O : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

Das Modell unter H_0 ist $y = \beta_0$

$$\begin{aligned} SSE(H_0) &= (Y - \hat{\hat{\beta}}_0)'(Y - \hat{\hat{\beta}}_0) = SST \\ SSH &= SSE(H_0) - SSE = SSM \end{aligned}$$

Damit entspricht hier der Wald -Test dem Test (3.31)

Weglassen von Einflussgrößen

Lineare Hypothese:

$$H_0 : \beta_k = \beta_{k+1} = \ldots = \beta_p = 0$$

Das Modell unter H_0 ist M1:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1}$$

$$\begin{aligned} SSE(H_0) &= SSE(M1) \\ SSH &= SSE(M1) - SSE \end{aligned}$$

Vergleich der SSE der beiden Modelle.

Beispiel: Welche Faktoren stehen in Zusammenhang mit den Fehlern bei einem Lesetest

Daten von Christa Kieferle (Pädagogik, LMU)
Daten von 180 Kindern aus den 8 Klassen (3. und 4. Klassen Grundschule)

Zielgröße:
Anzahl der Fehler bei einem Lesetest

potentielle Einflussgrößen:
Geschlecht, Jahrgang, Leseförderzeit, sonstiges lesen (1= oft,.. 5= fast nie), Gameboy (1= oft,.. 5= fast nie), Jahrgang.

Modell

$$Y = \beta_0 + \beta_1 * GE + \beta_2 * JG + \beta_3 * LZ + \beta_4 * WOL + \beta_5 * WOG + \beta_6 * WOTV$$

Y:	Anzahl der Fehler
GE:	Indikator für männlich (= 1 für männlich, 0 sonst)
JG:	Indikator für Klassenstufe (=1 für 3. Klasse, 0 für 4.Klasse)
LZ:	Lesezeit in der Schule
WOL:	Wie oft wird sonst gelesen
WOG:	Wie oft wird Gameboy gespielt
WOTV:	Wie oft TV

Interpretation der Parameter, z.B β_1 : Geschlechtsunterschied

Annahme: Bei WOL ist der durchschnittliche Unterschied in der Fehlerzahl zwischen den Stufen 1-5 gleich groß (Änderung um 1 entspricht β_4

Auswertung

1. Schätzung des Gesamtmodells
2. Gibt es einen Zusammenhang zwischen den Variablen WOL, WOG, WOTV und Y ?
3. Ist der Einfluss von WOG und WOTV auf die Fehlerzahl gleich ?
4. Gibt es einen Zusammenhang von Lesezeit und WOL und der Fehlerzahl ?
5. Wie lässt die Assoziation zur Lesezeit quantifizieren?

Fragen 2-4 lassen sich als lineare Hypothesen in dem Modell formulieren. Fragen 1 und 5 betreffen die Schätzung des Modells.