Test von Linearen Hypothesen

Exkurs: Non-lineare Hypothesen lassen sich relativ einfach bayesianisch überprüfen

Lineare Hypothese im multiplen Modell: $\overset{\text{Art LGS}}{A\beta = c}$ mit $A \in \mathbb{R}^{q \times (p+1)}$, $c \in \mathbb{R}^q$ und $\text{rank}(A) = q$ ($q \hat{=}$ Anzahl linear unabh. Zeilen in A), q: Anzahl der überprüften Hypothesen

<u>Herleitung der Teststatistik</u> Lösen Minimierungsproblem $\min_\beta \|y - X\beta\|^2$ unter Nebenbedingung $A\beta = c$ mithilfe von Lagrange-Methode

$$\Rightarrow \hat{\hat{\beta}} = \hat{\beta} - \underbrace{(X^TX)^{-1}A^T(A^T(X^TX)^{-1}A^T)^{-1}(A\hat{\beta} - c)}_{:= w} = \hat{\beta} - w$$

Fehlerquadratsumme unter linearer Hypothese $\overset{(H_0)}{A\beta = c}$: $\hat{\hat{\varepsilon}} = y - \hat{\hat{y}} = y - X\hat{\hat{\beta}} \overset{!}{=} y - (X\hat{\beta} - Xw) = (y - X\hat{\beta}) + Xw = \hat{\varepsilon} + Xw$ ($\hat{\hat{\beta}}$ einsetzen)

$$SSE(H_0) = \hat{\hat{\varepsilon}}^T\hat{\hat{\varepsilon}} = (\hat{\varepsilon} + Xw)^T(\hat{\varepsilon} + Xw) = \hat{\varepsilon}^T\hat{\varepsilon} + 2w^TX^T\hat{\varepsilon} + w^TX^TXw = \ldots\ \hat{\varepsilon}^T\hat{\varepsilon} + (A\hat{\beta} - c)^T(A(X^T)^{-1}A^T)^{-1}(A\hat{\beta} - c)$$

$\hat{\varepsilon}^TXw = (Xw)^T\hat{\varepsilon}$

Differenz der Fehlerquadratsumme von Originalmodell und Modell unter linearer Hypothese H_0:

$$SSH = SSE(H_0) - SSE = \hat{\hat{\varepsilon}}^T\hat{\hat{\varepsilon}} - \hat{\varepsilon}^T\hat{\varepsilon} = (A\hat{\beta} - c)^T(A(X^T)^{-1}A^T)^{-1}(A\hat{\beta} - c)$$

Wobei $Var(A\hat{\beta} - c) = \sigma^2 A(X^TX)^{-1}A^T$

Verteilung der normierten Quadratsummen: $\frac{SSH}{\sigma^2} \sim \chi^2_q$, $\frac{SSE}{\sigma^2} = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{\sigma^2} \overset{!}{=} \frac{y^TQy}{\sigma^2} \overset{!}{=} \frac{\hat{\sigma}^2(n-p')}{\sigma^2} \sim \chi^2_{n-p'}$ $(\hat{\varepsilon} = Qy)$ $(\hat{\sigma}^2 = \hat{\varepsilon}^T\hat{\varepsilon}/(n-p'))$

also Hypothese wird wahrscheinlicher abg. → F-Verteilung mit schmaleren Rändern
mehrere Hypothesen zugleich ⇒ $F \in q \uparrow$

Da $\hat{\sigma}^2$ und $\hat{\beta}$ gemäß Modellannahmen voneinander unabh. sind, gilt $TF = \frac{SSH/q}{SSE(n-p')} = \frac{MSH}{MSE} \sim F(q, n-p')$ zentrale F-Verteilung mit FG $q, \underline{n-p'}$ (immer konst.)

⇒ <u>Testentscheidung</u>: Lineare Hypothese H_0 wird abgelehnt, wenn $TF > F_{1-\alpha}(q, n-p')$ <u>rechtsseitiger F-Test</u>

Nullhypothese wird abgelehnt, wenn relativer Unterschied zwischen SSE und SSE(H_0) ausreichend groß ist — in R schnell: anova(modell_restr, modell_full)

↳ Falls ja, dann führt unrestringiertes Modell zu signifikanter Reduktion des SSE ⇒ Alternativ-Hypothese signifikant

($\beta_0 = 0$ testen witzlos, Intercept ist "Verschieber", sodass Modell durch Datenp. läuft)

<u>Spezialfall Overall-Test</u> $H_0: \beta_1 = \ldots = \beta_p = 0$ in allgemeiner Form: $A\beta = \vec{0}$ mit $A = ([0, I])$ (1. Spalte Nuller wegen Intercept)

Testet, ob Modell irgendwas taugt oder komplett random (Nullhypothese) ⇔ Test auf $R^2 = 0$

Wenn $F_0 = \frac{MSM}{MSE} > F_{1-\alpha}(p'-1, n-p')$ Lehnen wir H_0 ab ⇒ $H_1: \exists\, \beta_k \neq 0$ wird auf diesem Niveau angenommen $(t\text{-Wert})^2 = F\text{-Wert}$

(→ Nullhypoth. $\beta_k = 0$) isolierter Test auf $\beta_k = 0$: rechtsseitiger F-Hypothesentest äquiv. zu linksseit. t-Test

Unterschied zu einzelnen t-Tests auf $\beta_k = 0$ für alle k nacheinander: $\overset{\text{Overall}}{F\text{-Test}}$ berücksichtigt mögliche Korrelation zwischen KQ-Koeffizienten $\beta_1, \ldots, \beta_p$

Es kann sein, dass alle individuell nicht sign. $\neq 0$ aber zusammengenommen schon, z.B. bei starker Multikollinearität

Oder einzeln sign. $\neq 0$ aber gemeinsam mit anderer Kovariable (F-Test auf $\beta_k = \beta_j = 0$) nicht signifikant

<u>Fingerübung</u> $p' = 6$ Bestimme Matrix A für Hypothese: $\beta_1 = 0, \beta_2 = 0, \beta_1 = \beta_3, \beta_1 - \beta_2 = 1, \beta_4 = 2$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 2 \end{pmatrix}$$

<u>Übung 3 (Aufgabe 7)</u> Lineares Modell, $p = 6$ (eine Factor-Variable geschlecht)

a) Überprüfung, ob mindestens eine der Kovariable Zusammenhang mit Zielvariable ⇒ Overall F-Test $H_0: \beta_1 = \ldots \beta_6 = 0$

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_6 \end{pmatrix}, c = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$TF \sim F(6, n-7)$

b) Teste ob zweiter Kovariable Zusammenhang mit Zielgröße aufweist ⇒ $H_0: \beta_2 = 0$, $A = (0\ 0\ 1\ 0\ 0\ 0\ 0)$, β: siehe oben, $c = 0$

$TF \sim F(1, n-7)$

c) Teste, ob Zielgröße von erster oder vierter Kovariable linear abhängt ⇒ $H_0: \beta_1 = \beta_4 = 0$, $A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$, $c = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$TF \sim F(2, n-7)$

d) Teste, ob fünfte und sechste Kovariable selben Zusammenhang mit Zielvar. haben ⇒ $H_0: \beta_5 = \beta_6 \Leftrightarrow \beta_5 - \beta_6 = 0$, $A = (0\ 0\ 0\ 0\ 0\ 1\ -1)$, $c = 0$

$TF \sim F(1, n-7)$ <u>Achtung</u>: restringiertes Modell definiert als $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 (x_{i5} + x_{i6})$ In R: lm(y ~ v1 + ... + I(v5+v6))

e) Teste, ob sich Z-Score (y) mit einem Anstieg des Kindesalters (dritte Kov.) um eine Einh. [Monat] um zwei Punkte verringert ⇒ $H_0: \beta_3 = -2$

$TF \sim F(1, n-7)$ Restringiertes Modell: $\ldots - 2x_{i3} \ldots$ aber in R: lm(y ~ v1 + v2 + offset(-2v3) + ...) z.B. für anova $A = (0\ 0\ 0\ 1\ 0\ 0\ 0)$, $c = -2$