Copenhagen Häuser Datensatz
(3 verschiedene Reihenfolgen)

Model	Df	Deviance	Resid. Df	Resid. Dev
NULL			71	833.66
H	3	376.30	68	457.36
C	1	38.83	67	418.52
I	2	78.52	65	340.01
S	2	44.66	63	295.35
H:C	3	39.06	60	256.29
I:S	4	106.37	56	149.92
H:S	6	60.67	50	89.25
H:I	6	12.59	44	76.67
C:I	2	16.70	42	59.97
C:S	2	16.02	40	43.95
C:S	2	5.13	61	290.23
C:I	2	17.54	59	272.69
H:I	6	16.89	53	255.80
H:S	6	60.81	47	194.99
I:S	4	107.00	43	87.99
H:C	3	44.04	40	43.95
I:S	4	106.37	59	188.98
H:C	3	39.06	56	149.92
H:S	6	60.67	50	89.25
C:I	2	15.86	48	73.40
H:I	6	13.43	42	59.97
C:S	2	16.02	40	43.95

Balance

Gegeben sei der Datensatz

$$\{y_i; x_{i1}, x_{i2}, \dots x_{ip}\} \qquad i = 1, \dots n$$

Die $\{y_i\}$ betrachten wir als unabhängige Stichproben der Zufallsvariablen $\{Y_i\}$. Die $\{x_{.j}\}$ können stetige oder kategorielle Faktoren darstellen.

Wir passen das lineare Modell mit $(p+1)$ Parametern $\{\beta_j\}(j = 0, \dots p)$ den Daten an.

$$Y_i = X_i\beta + \epsilon_i \qquad \epsilon_i \text{ u.i.v. } \sim N(0, \sigma^2)$$

Das Modell kann auch als

$$Y = X\beta + \epsilon \qquad \epsilon \sim N(0, \sigma^2 I)$$

oder

$$Y_i \sim N(X_i\beta, \sigma^2) \qquad \{Y_i\} \text{ unabhängig}$$

geschrieben werden.

Schätzer von β

Für lineare Modelle ist der ML-Schätzer ein KQ-Schätzer und $\widehat{\beta}$ minimiert $(y - X\beta)'(y - X\beta)$

$$\widehat{\beta} = (X'X)^{-1}X'y$$

Dann gilt

$$\begin{aligned} RQS &= y'y - y'X\widehat{\beta} - \widehat{\beta}X'y + \widehat{\beta}X'X\widehat{\beta} \\ &= y'y - y'X(X'X)^{-1}X'y \end{aligned}$$

Aber

$$\begin{aligned} \widehat{y} &= X\widehat{\beta} \\ &= X(X'X)^{-1}X'y \end{aligned}$$

$$\begin{aligned} \widehat{y}'\widehat{y} &= y'X(X'X)^{-1}X'X(X'X)^{-1}X'y \\ &= y'X(X'X)^{-1}X'y \end{aligned}$$

Um die Wichtigkeit der einzelnen Faktoren zu messen, zerlegen wir diese Modell Quadratsumme.

Die Zerlegung der Modell Quadratsumme

Schreiben wir den Parametervektor um:

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

Dann haben wir, dass

$$\begin{aligned} Y &= X\beta + \epsilon \\ &= X_1\beta_1 + X_2\beta_2 + \epsilon \end{aligned}$$

Und die Normalgleichungen

$$X'X\widehat{\beta} = X'y$$

werden so geschrieben:

$$\begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix}$$

Die Orthogonalität von X_1 und X_2 bedeutet, dass

$$X_1'X_2 = X_2'X_1 = 0$$

was dazu führt, dass wir β_1 und β_2 getrennt betrachten können:

$$X_1'X_1\widehat{\beta}_1 = X_1'y$$

und

$$X_2'X_2\widehat{\beta}_2 = X_2'y$$

$$\widehat{\beta}_i = (X_i'X_i)^{-1}X_i'y$$

Und die vom mit $\{\beta_i\}$ verbundenen Faktor erklärten Quadratsumme ist

$$\begin{aligned} QS(\widehat{\beta}_i) &= y'X_i(X_i'X_i)^{-1}X_i'y \\ &= \widehat{\beta}_i'X_i'y \end{aligned}$$