Multiple lineare Regression

Das Modell

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \qquad i = 1, \ldots, n$$

y_i : Zielvariable
$x_{i1}, \ldots, x_{ip}$: erklärende Variablen, fest.
$\beta_0, \ldots, \beta_p$: unbekannte Parameter, Regressionskoeffizienten.
ϵ_i : zufälliger Rest oder Fehler. $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$ und $Cov(\epsilon_i, \epsilon_j) = 0$ für $i \neq j$.
Für ϵ_i normalverteilt gilt $y_i \sim N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \sigma^2)$ und $Cov(y_i, y_j) = 0$ für $i \neq j$.

in Matrixschreibweise:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$\mathbf{y}$: Zielvariablenvektor der Länge n.
$\mathbf{X}$: Designmatrix der Dimension $n \times (p+1)$.
$\boldsymbol{\beta}$: Parametervektor der Länge $p+1$.
$\boldsymbol{\epsilon}$: Fehlervektor, $E(\boldsymbol{\epsilon}) = \mathbf{0}$ und $Cov(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$.
Für ϵ_i normalverteilt gilt $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

Methode der kleinsten Quadrate

Gesucht sind $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ so, dass
$Q = \sum_{i=1}^n r_i^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}))^2$ minimal wird.

Normalgleichungen:

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_0} &= -2\sum_{i=1}^n (y_i - (\beta_0 + \ldots + \beta_p x_{ip})) = 0 \\
\frac{\partial Q}{\partial \beta_1} &= -2\sum_{i=1}^n (y_i - (\beta_0 + \ldots + \beta_p x_{ip}))x_{i1} = 0 \\
\vdots \quad & \quad \vdots \\
\frac{\partial Q}{\partial \beta_p} &= -2\sum_{i=1}^n (y_i - (\beta_0 + \ldots + \beta_p x_{ip}))x_{ip} = 0
\end{aligned}$$

$$\mathbf{X}^t(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}$$

oder

$$\mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^t\mathbf{y}$$

Least-Squares-Schätzungen:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ und $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$
mit Normalverteilung: $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1})$

Geschätzte Werte und Residuen:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} = \mathbf{H}\mathbf{y} \quad \mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y} \qquad \mathbf{H} \text{ heisst } \textbf{Hat-Matrix}$$

Tests und Vertrauensintervalle

ANOVA-Tabelle:

Source	Sum of squares	df	Mean square	F^*
Regres	$SSR = \sum(\hat{y}_i - \bar{y})^2$	p	MSR	MSR/MSE
Resid	$SSE = \sum(y_i - \hat{y}_i)^2$	$n-1-p$	MSE	
Total	$SST = \sum(y_i - \bar{y})^2$	$n-1$		

Globaler F-Test:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad \text{gegen} \quad H_A : \text{ mindestens ein } \beta_j \neq 0$$

Testgrösse $F^* = MSR/MSE$ hat unter H_0 eine F-Verteilung mit p und n–p–1 Freiheitsgraden. Verwerfe H_0, wenn $F^* > F_{95\%,p,n-p-1}$.

Bestimmtheitsmass:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad adjR^2 = 1 - \left(\frac{n-1}{n-p-1}\right)\frac{SSE}{SST}$$

Tests von individuellen Parametern:

$$H_0 : \beta_j = 0 \quad \text{gegen} \quad H_A : \beta_j \neq 0$$

Teststatistik

$$t^* = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{(\mathbf{X}^t\mathbf{X})^{-1}_{jj}}}$$

hat unter H_0 eine t-Verteilung mit n–p–1 Freiheitsgraden. Verwerfe H_0, wenn $|t^*| > t_{97.5\%,n-p-1}$.

95%-Vertrauensintervall für β_j : $\quad \hat{\beta}_j \pm t_{97.5\%,n-p-1} \cdot \hat{\sigma}\sqrt{(\mathbf{X}^t\mathbf{X})^{-1}_{jj}}$

95%-Vertrauensintervall für $E(y_0)$: $\quad \hat{y}_0 \pm t_{97.5\%,n-p-1} \cdot \hat{\sigma}\sqrt{\mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0}$ wobei $\mathbf{x}_0^t = (1 x_{01} x_{02} \cdots x_{0p})^t$

95%-Prognoseintervall für eine zukünftige Beobachtung: $\quad \hat{y}_0 \pm t_{97.5\%,n-p-1} \cdot \hat{\sigma}\sqrt{1 + \mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0}$

Multicollinearität:

Sind x_1 und x_2 korreliert, dann ändern sich die geschätzten Koeffizienten, je nachdem welche Variablen im Modell sind. Es ist möglich, dass der globale F-Test signifikant ist und alle einzelnen t-Tests sind nicht signifikant.

Partielle F-Tests:

Effekt von $p-q$ Variablen gemeinsam testen. Partitioniere Parametervektor und Designmatrix :

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \\ \beta_{q+1} \\ \vdots \\ \beta_p \end{pmatrix} \quad \text{und} \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix},$$

Modell: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$
Teste auf $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ gegen $H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$
Testgrösse

$$F^* = \frac{(SSR_{H_1} - SSR_{H_0})/(p-q)}{SSE_{H_1}/(n-p-1)}$$

hat unter H_0 eine F-Verteilung. Verwerfe H_0, wenn $F^* > F_{95\%,p-q,n-p-1}$.

Modelldiagnostik

Residuenplot:
Normalplot der Residuen r_i
Residuen r_i gegen geschätzte y-Werte $\hat{y}_i$
Residuen r_i gegen eine erklärende Variable x_i des Modells
Residuen r_i gegen eine neue Variable x_i', die nicht im Modell ist
Residuen r_i gegen den Index i

Ausreisser und einflussreiche Beobachtungen:
Ausreisser: Beobachtung mit grossem Residuum
einflussreiche Beobachtung: Beobachtung mit grossem Einfluss auf Parameterschätzungen
Hebelpunkt: Beobachtung mit extremen x-Werten

Leverages: Diagonalelemente h_{ii} der Hat-Matrix $\mathbf{H}$. Messen, wie extrem Beobachtungen bezüglich der x-Variablen sind. Es gilt $0 \leq h_{ii} \leq 1$. Hebelpunkte: $h_{ii} > 2(p+1)/n$.
Gefährlich, wenn r_i und h_{ii} gross. Betrachte Plot r_i gegen h_{ii} .

Cook's Distanz:

$$D_i = \frac{\sum(\hat{y}_j - \hat{y_{j(i)}})^2}{(p+1)\hat{\sigma}^2} = \frac{h_{ii}}{1-h_{ii}} \cdot \frac{r_i^{*2}}{p+1}$$

$$r_i^* = \frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \qquad \text{heisst studentisiertes Residuum}$$

Punkte mit $D_i > 1$ sollten genauer untersucht werden.