

4. MULTIPLE LINEARE REGRESSION (Maddala Kap. 4)

Datengenerierender Prozess:

$$Y_i = \underbrace{\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}_{\text{deterministischer Teil}} + \underbrace{u_i}_{\text{Störterm}}$$

Annahmen bezüglich u_i wie bei der einfachen linearen Regression.

Stichprobe ($i = 1, \dots, n$):

$$\begin{array}{l} x_{11}, x_{12}, \dots, x_{1i}, \dots, x_{1n} \\ x_{21}, x_{22}, \dots, x_{2i}, \dots, x_{2n} \\ \vdots \\ x_{j1}, x_{j2}, \dots, x_{ji}, \dots, x_{jn} \\ \vdots \\ x_{k1}, x_{k2}, \dots, x_{ki}, \dots, x_{kn} \end{array} \quad \begin{array}{l} \\ \\ \\ \text{exogene (erklärende)} \\ \text{Variablen} \\ \\ \\ \end{array}$$

$$Y_1, Y_2, \dots, Y_i, \dots, Y_n \quad \begin{array}{l} \\ \\ \\ \text{endogene (abhängige)} \\ \text{Variable} \end{array}$$

Schätzung der Regressionsparameter ($\alpha, \beta_1, \dots, \beta_j, \dots, \beta_k$) mit der Methode der kleinsten Quadrate:

Schätzgleichung

$$y_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_k x_{ki} + \underbrace{\hat{u}_i}_{\text{Residuum}}$$

Wir legen $\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_j, \dots, \hat{\beta}_k$ so fest, dass

$$RSS = \sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \dots - \hat{\beta}_k x_{ki})^2$$

minimal wird (RSS: residual sum of squares). Erste Ableitungen gleich 0 setzen:

$$\partial RSS / \partial \hat{\alpha} = 0, \quad \partial RSS / \partial \hat{\beta}_j = 0 \quad (j = 1, \dots, k)$$

ergibt $k+1$ Gleichungen mit $k+1$ Unbekannten.

Die so erhaltenen Schätzwerte $\hat{\alpha}$, $\hat{\beta}_j$ ($j = 1, \dots, k$) sind unverzerrt, konsistent und effizient.

$\hat{\sigma}^2 = \text{RSS}/(n-k-1)$ ist eine unverzerrte Schätzung für die Störterm-Varianz σ^2 .

Varianzanalyse

Wir definieren:

$$\text{TSS} = S_{yy} = \sum (y_i - \bar{y})^2 \quad \text{mit } \bar{y} = (1/n) \sum y_i$$

$$\text{RSS} = \sum \hat{u}_i^2$$

$$\begin{array}{rcc} \text{TSS} & = & \text{RSS} + \text{ESS} \\ | & & | \\ \text{total} & & \text{residual explained} \\ \hline & & \text{sum of squares} \end{array}$$

$$S_{1y} = \sum (x_{1i} - \bar{x}_1)(y_i - \bar{y}) \quad \text{mit } \bar{x}_1 = (1/n) \sum x_{1i}$$

$$S_{2y} = \sum (x_{2i} - \bar{x}_2)(y_i - \bar{y}) \quad \text{mit } \bar{x}_2 = (1/n) \sum x_{2i}$$

$$\cdot$$

$$S_{ky} = \sum (x_{ki} - \bar{x}_k)(y_i - \bar{y}) \quad \text{mit } \bar{x}_k = (1/n) \sum x_{ki}$$

Für ESS kann man auch schreiben:

$$\text{ESS} = \hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y} + \dots + \hat{\beta}_k S_{ky}$$

Multiples Bestimmtheitsmass:

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = (\hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y} + \dots + \hat{\beta}_k S_{ky}) / \text{TSS}$$

R ist der multiple Korrelationskoeffizient.

Für R^2 wird ausführlicher auch geschrieben $R_{y \cdot 12 \dots k}^2$.

STATISTISCHE AUSSAGEN IM MULTIPLLEN LINEAREN REGRESSIONSMODELL
(Hypothesentest, Berechnung von Vertrauensbereichen)

Stichprobenverteilung von $\hat{\alpha}$, $\hat{\beta}_1$, ... $\hat{\beta}_k$: Unter der Annahme, dass $u_i \sim \text{IN}(0, \sigma^2)$, sind die Schätzungen $\hat{\alpha}$, $\hat{\beta}_1$, ... $\hat{\beta}_k$ normalverteilt mit $E(\hat{\alpha}) = \alpha$, $E(\hat{\beta}_1) = \beta_1$, .. $E(\hat{\beta}_k) = \beta_k$ und Varianzen/Kovarianzen, die proportional zu σ^2 sind.

Für $k = 2$ (d.h. $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$) gilt z.B. (vgl. Maddala S. 96):

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{11}(1 - r_{12}^2)}$$

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{S_{22}(1 - r_{12}^2)}$$

$$\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{-\sigma^2 r_{12}}{S_{12}(1 - r_{12}^2)}$$

worin: r_{12} : einfacher Korrelationskoeff. zwischen x_{1i} und x_{2i}

$$S_{11} = \sum (x_{1i} - \bar{x}_1)^2 \quad S_{22} = \sum (x_{2i} - \bar{x}_2)^2$$

$$S_{12} = \sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$$

Wir wissen nun z.B., dass

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{var}(\hat{\beta}_1)}} \sim N(0,1) \text{ , d.h. standard-normalverteilt ist.}$$

Das im Ausdruck für $\text{var}(\hat{\beta}_1)$ erscheinende σ^2 ist aber unbekannt und muss durch die Schätzung $\hat{\sigma}^2 = \text{RSS}/(n-3)$ ersetzt werden. Die geschätzte Varianz und der entsprechende Standardfehler von $\hat{\beta}_1$ sind:

$$S^2(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{S_{11}(1 - r_{12}^2)} \quad \text{SE}(\hat{\beta}_1) = \sqrt{S^2(\hat{\beta}_1)}$$

Die Stichprobenverteilung von $\hat{\beta}_1$ ist somit gegeben durch

$$\frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\hat{\beta}_1)} \sim t_{n-3} \quad (\text{t-verteilt mit } n-3 \text{ Freiheitsgraden})$$

Allgemein (bei $j = 1, \dots, k$ erklärenden Variablen) gilt:

$$\frac{\hat{\beta}_j - \beta_j}{\text{SE}(\hat{\beta}_j)} \sim t_{n-k-1} \quad (j = 1, \dots, k) \text{ ,} \quad \frac{\hat{\alpha} - \alpha}{\text{SE}(\hat{\alpha})} \sim t_{n-k-1}$$

Auf Basis dieser t-Verteilungen lassen sich für einzelne Regressionsparameter Hypothesentests durchführen (z.B. $H_0: \beta_3 = 0$) und Vertrauensbereiche berechnen.

Beim Test von verbundenen Hypothesen (z.B. $H_0: \beta_2 = 1$ und $\beta_3 = 0$) und bei der Berechnung gemeinsamer Vertrauensbereiche mehrerer Regressionsparameter ist auf die F-Verteilung abzustellen (vgl. Maddala S. 116):

$$F = \frac{(\text{RRSS} - \text{URSS})/r}{\text{URSS}/(n-k-1)} \sim F_{r, n-k-1}$$

URSS: Summe quadrierter Residuen der unrestringierten Regression

RRSS: Summe quadrierter Residuen der durch H_0 restringierten Regression

r: Anzahl Restriktionen (z.B. 2)

EINFACHE, PARTIELLE UND MULTIPLE KORRELATIONSKOEFFIZIENTEN

Wir betrachten als Beispiel $y_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{u}_i$.

multiple Korrelation:

$$R_{y \cdot 12}^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = (\hat{\beta}_1 s_{1y} + \hat{\beta}_2 s_{2y}) / s_{yy}$$

$$RSS = TSS(1 - R_{y \cdot 12}^2)$$

einfache Korrelation:

$$y_i = \hat{\alpha} + \hat{\beta} x_{1i} + \hat{v}_i$$

$$r_{y1}^2 = \frac{ESS'}{TSS} = 1 - \frac{RSS'}{TSS} = \frac{\hat{\beta} s_{1y}}{s_{yy}}$$

$$RSS' = TSS(1 - r_{y1}^2)$$

partielle Korrelation:

$$y_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{u}_i$$

$$r_{y2 \cdot 1}^2 = 1 - \frac{RSS}{RSS'}, \quad (\text{relative Verminderung von } RSS' \text{ durch Einbezug von } x_{2i})$$

$$RSS = RSS'(1 - r_{y2 \cdot 1}^2)$$

Daraus ergibt sich die folgende Beziehung zwischen den drei Korrelationskoeffizienten:

$$TSS(1 - R_{y \cdot 12}^2) = TSS(1 - r_{y1}^2)(1 - r_{y2 \cdot 1}^2)$$

$$(1 - R_{y \cdot 12}^2) = (1 - r_{y1}^2)(1 - r_{y2 \cdot 1}^2)$$

Für $k = 3$:

$$(1 - R_{y \cdot 123}^2) = (1 - r_{y1}^2)(1 - r_{y2 \cdot 1}^2)(1 - r_{y3 \cdot 12}^2)$$

andere Reihenfolge des Einbezugs:

$$(1 - R_{y \cdot 123}^2) = (1 - r_{y3}^2)(1 - r_{y1 \cdot 3}^2)(1 - r_{y2 \cdot 13}^2)$$

KORRIGIERTES \bar{R}^2 (R^2 adjusted for degrees of freedom):

- Durch Hinzufügen weiterer erklärender Variablen vermindert sich RSS zwangsläufig, so dass R^2 steigt.
- Unverzerrte Schätzung der Störterm-Varianz: $\hat{\sigma}^2 = \text{RSS}/(n-k-1)$. Wir teilen durch $n-k-1$ (und nicht durch $n-1$), weil zwischen den n berechneten Residuen in Form der geschätzten Regressionsparameter $(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_k)$ $k+1$ Abhängigkeiten bestehen. Anders formuliert: Nur $n-k-1$ Residuen sind frei ($n-k-1$ Freiheitsgrade). Im Unterschied zu RSS vermindert sich $\hat{\sigma}^2$ bei Hinzufügen weiterer erklärender Variablen nicht zwangsläufig, sondern wird irgendwann (bei Hinzufügen von Variablen, die keinen oder nur einen sehr geringen Erklärungsgehalt haben) wieder grösser.
- Analoge "Korrektur" von R^2 (Maddala S. 125 f.):

$$(1 - \bar{R}^2) = \frac{n-1}{n-k-1} (1 - R^2)$$

Erklärung:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}/(n-1)}{\text{TSS}/(n-1)}$$

$$\bar{R}^2 = 1 - \frac{\text{RSS}/(n-k-1)}{\text{TSS}/(n-1)}$$

Diese Korrektur bewirkt, dass \bar{R}^2 bei Hinzufügen weiterer erklärender Variablen genau dann zu fallen beginnt, wenn $\hat{\sigma}^2$ zunimmt.