Statistik IV für Nebenfachstudierende

3.2. Grundkonzept der multivariaten linearen Regression

Prof. Dr. Andreas Mayr

Institut für Statistik, LMU München

Sommersemester 2017

Grundidee

- Bis jetzt hatten wir als Zielgröße Beobachtungen $\underline{y} = (y_1, ..., y_n)^\top$.
  - Dazu p Einflussgrößen (multiple lineare Regression).
  - Führt zu p verschiedenen Koeffizienten $\underline{\beta} = (\beta_0, \beta_1, ..., \beta_p)^\top$.
- Jetzt multivariate Zielgröße der Dimension q: $\underline{y}_i = (y_{i1}, ..., y_{iq})^\top$
  - Führt zu $p \cdot q$ verschiedenen Koeffizienten:
  - Für jeden der q Zielgrößen ein Koeffizientenvektor, $\underline{\beta}_{(1)}, ..., \underline{\beta}_{(q)}$.
  - ▶ Ziel: Gemeinsame Analyse, gemeinsame Hypothesentests.

Daten:

$$(\underline{y}_i, \underline{x}_i)\ ,\ \ i = 1, \ldots, n \ \ \text{mit}$$

$$\underline{y}_i^\top = (y_{i1}, \ldots, y_{iq})$$
$$\underline{x}_i^\top = (1, x_{i1}, \ldots, x_{ip})$$

- q abhängige Variablen
- p konstante und fixe Kovariablen (ohne Intercept)

Klassisches lineares Modell

$$y_{ij} = \mathbf{x}_i^\top \underline{\beta}_{(j)} + \varepsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots q,$$

Matrixdarstellungen:

Vektorielle Form

$$\begin{pmatrix} y_{i1} \\ \vdots \\ y_{iq} \end{pmatrix} = \begin{pmatrix} \underline{x}_i^\top & & & \\ & \underline{x}_i^\top & & \\ & & \ddots & \\ & & & \underline{x}_i^\top \end{pmatrix} \begin{pmatrix} \underline{\beta}_{(1)} \\ \vdots \\ \underline{\beta}_{(q)} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iq} \end{pmatrix}$$

bzw.

$$\underline{y}_i = \underline{\underline{X}}_i \, \underline{\beta} + \boldsymbol{\epsilon}_i$$

und als Gesamtmodell

$$\underline{y} = \underline{\underline{X}} \, \underline{\beta} + \boldsymbol{\epsilon}.$$

wobei

$$\underline{\beta}_{(j)}^{\top} = (\beta_{0j}, \beta_{1j}, \ldots \beta_{pj}), \quad \underline{\beta}^{\top} = (\underline{\beta}_{(1)}^{\top}, \ldots, \underline{\beta}_{(q)}^{\top}).$$

Multivariate Formulierung I

$$\begin{pmatrix} \underline{y}_1^\top \\ \vdots \\ \underline{y}_n^\top \end{pmatrix} = \begin{pmatrix} \underline{x}_1^\top \\ \vdots \\ \underline{x}_n^\top \end{pmatrix} (\underline{\beta}_{(1)}, \ldots, \underline{\beta}_{(q)}) + \begin{pmatrix} \boldsymbol{\epsilon}_1^\top \\ \vdots \\ \boldsymbol{\epsilon}_n^\top \end{pmatrix}$$

Multivariate Formulierung II

$$(\underline{y}_{(1)}, ..., \underline{y}_{(q)}) = \begin{pmatrix} \underline{x}_1^\top \\ \vdots \\ \underline{x}_n^\top \end{pmatrix} (\underline{\beta}_{(1)}, \dots, \underline{\beta}_{(q)}) + (\boldsymbol{\epsilon}_{(1)}, ..., \boldsymbol{\epsilon}_{(q)})$$

Als Gesamtmodell

$$\mathbf{Y} = \underline{\underline{X}}_0 \mathbf{B} + \mathbb{E}$$

mit $\underline{\underline{X}}_0 = \begin{pmatrix} \underline{x}_1^\top \\ \vdots \\ \underline{x}_n^\top \end{pmatrix}, \underline{\beta}_{(j)} = \begin{pmatrix} \beta_{0j} \\ \vdots \\ \beta_{pj} \end{pmatrix}, \underline{y}_{(j)} = \begin{pmatrix} y_{1j} \\ \vdots \\ y_{nj} \end{pmatrix}, \boldsymbol{\epsilon}_{(j)} = \begin{pmatrix} \epsilon_{1j} \\ \vdots \\ \epsilon_{nj} \end{pmatrix}.$

Beispiele ?

Verteilungsannahme

- Cov$(\underline{y}_i) = \underline{\underline{\Sigma}}$
- Normalverteilung in Struktur
- Unkorreliertheit der Beobachtungen
- Zusammen:

Annahmen Störterm

$$(1) \quad \mathbb{E}(\boldsymbol{\epsilon}_i) = \mathbf{0}$$

$$(2) \quad \mathrm{cov}(\boldsymbol{\epsilon}_i) = \underline{\underline{\Sigma}} \;,\quad i = 1, \ldots, n$$

$$\mathrm{cov}(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_j) = \mathbf{0} \;,\quad i \neq j$$

$$\Rightarrow \mathrm{cov}\left(\boldsymbol{\epsilon}_{(s)}, \boldsymbol{\epsilon}_{(t)}\right) = \sigma_{st}\,\mathbf{I},\; s \neq t$$

(3) Normalverteilung: $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \underline{\underline{\Sigma}})$

Kleinste–Quadrate–Schätzung

$$\sum_{j=1}^{q} (\underline{y}_{(j)} - \underline{\underline{X}}_0 \, \underline{\beta}_{(j)})^\top (\underline{y}_{(j)} - \underline{\underline{X}}_0 \, \underline{\beta}_{(j)}) \to \min$$

Lösung (voller Rang von $\underline{\underline{X}}$):

$$\hat{\underline{\beta}}_{(j)} = (\underline{\underline{X}}_0^\top \underline{\underline{X}}_0)^{-1} \underline{\underline{X}}_0^\top \underline{y}_{(j)} \quad \text{bzw.} \quad \hat{\mathbf{B}} = (\underline{\underline{X}}_0^\top \underline{\underline{X}}_0)^{-1} \underline{\underline{X}}_0^\top \mathbf{Y}$$

Als Vektor

$$\hat{\underline{\beta}} = \text{BlockDiag}\{(\underline{\underline{X}}_0^\top \underline{\underline{X}}_0)^{-1} \underline{\underline{X}}_0^\top\} \underline{y}_0 \quad \text{oder} \quad \hat{\underline{\beta}} = (\underline{\underline{X}}^\top \underline{\underline{X}})^{-1} \underline{\underline{X}}^\top \underline{y}.$$

mit

$$\underline{y}_0^\top = (\underline{y}_{(1)}^\top, \ldots, \underline{y}_{(q)}^\top), \; \underline{y}^\top = (\underline{y}_1^\top, \ldots, \underline{y}_n^\top).$$

Kleinste–Quadrate–Schätzung

Anmerkung:

- In KQ-Schätzung fließt $\underline{\underline{\Sigma}}$ nicht ein.
- $\underline{\hat{\beta}}_{(q)}$ hängt nicht von $\underline{\underline{\Sigma}}$ ab.
- KQ-Schätzung identisch mit separater Optimierung!

Warum dann das Ganze?

- Simultane multivariate Tests (z.B., MANOVA)

Schätzung der Kovarianz

$$\hat{\underline{\underline{\Sigma}}} = \frac{1}{n-p-1} \sum_{i=1}^{n} (\underline{y}_i - \underline{\underline{X}}_i \hat{\underline{\beta}})(\underline{y}_i - \underline{\underline{X}}_i \hat{\underline{\beta}})^\top$$

$$\hat{\underline{\underline{\Sigma}}} = \frac{1}{n-p-1} \mathbb{E}^\top \mathbb{E}$$

wobei

$$\mathbb{E} = \mathbf{Y} - \underline{\underline{X}}_0 \hat{\mathbf{B}} \text{ Residuenmatrix.}$$

Gauß-Markov-Theorem

(1) $\mathbb{E}(\hat{\mathbf{B}}) = \mathbf{B}$, $\mathbb{E}(\hat{\underline{\underline{\Sigma}}}) = \underline{\underline{\Sigma}}$

(2) $\operatorname{cov}(\underline{\hat{\beta}}_{(i)}, \underline{\hat{\beta}}_{(j)}) = \sigma_{ij}(\underline{\underline{X}}_0^\top \underline{\underline{X}}_0)^{-1}$

(3) $\hat{\mathbf{B}}$ ist BLUE

Weiter gilt mit Normalverteilungsannahme

(4) $\hat{\mathbf{B}}$ ist normalverteilt

(5) $\mathbf{Q}_e = \hat{\mathbb{E}}^\top \hat{\mathbb{E}} \sim W_p(\underline{\underline{\Sigma}}, n - p - 1)$

(6) $\hat{\mathbf{B}}$ und $\hat{\underline{\underline{\Sigma}}}$ sind unabhängig.