

5.Tutorium Multivariate Verfahren

- Multivariate Regression -

Hannah Busen:

15.06.2015 und 22.06.2015

Nicole Schüller:

16.06.2015 und 23.06.2015

Institut für Statistik, LMU München

Gliederung

- 1 Problemstellung
- 2 Multivariates Regressionsmodell
- 3 Lineare Hypothesen

Gliederung

- 1 Problemstellung
- 2 Multivariates Regressionsmodell
- 3 Lineare Hypothesen

Zur Erinnerung: Univariates Regressionsmodell

- Modellierung des Zusammenhangs von p Einflussgrößen auf eine **skalare** Zielgröße:

$$y_i = \beta_0 + x_{i1}\beta_1 + \dots + x_{ip}\beta_p + \epsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$$

- In Matrixschreibweise: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{n \times 1} = \underbrace{\begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & & \vdots & \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}}_{n \times (p+1)} \underbrace{\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}}_{(p+1) \times 1} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{n \times 1}$$

Multivariates Regressionsmodell

- Modellierung des Zusammenhangs von p Einflussgrößen auf q Zielgrößen. Die abhängige Variable ist jetzt ein **Vektor**:

$$\mathbf{y}_i = (y_{i1}, \dots, y_{iq})^\top$$

- Modell für das i -te Individuum und die j -te Zielgröße:

$$y_{ij} = \beta_{0j} + x_{i1}\beta_{1j} + \dots + x_{ip}\beta_{pj} + \epsilon_{ij} = \mathbf{x}_i^\top \boldsymbol{\beta}_{(j)} + \epsilon_{ij}$$

β_{rj} : Einfluss der r -ten x -Variable auf die j -te y -Variable

- Für jede Zielgröße $j = 1, \dots, q$ erhält man einen eigenen Regressionskoeffizientenvektor $\boldsymbol{\beta}_{(j)}$

Gliederung

- 1 Problemstellung
- 2 Multivariates Regressionsmodell**
- 3 Lineare Hypothesen

Darstellung mit Matrizen

- Modellgleichung: $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$
- Matrix der Zielgrößen:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1^\top \\ \vdots \\ \mathbf{y}_n^\top \end{pmatrix} = \begin{pmatrix} y_{11} & \dots & y_{1q} \\ \vdots & & \vdots \\ y_{n1} & \dots & y_{nq} \end{pmatrix}$$

- Matrix der Regressionskoeffizienten:

$$\mathbf{B} = (\boldsymbol{\beta}_{(1)} \quad \dots \quad \boldsymbol{\beta}_{(q)}) = \begin{pmatrix} \beta_{01} & \dots & \beta_{0q} \\ \vdots & & \vdots \\ \beta_{p1} & \dots & \beta_{pq} \end{pmatrix}$$

- Matrix der Störgrößen:

$$\mathbf{E} = \begin{pmatrix} \epsilon_{11} & \dots & \epsilon_{1q} \\ \vdots & & \vdots \\ \epsilon_{n1} & \dots & \epsilon_{nq} \end{pmatrix}$$

- Designmatrix \mathbf{X} wie im univariaten Regressionsmodell!

Annahmen

- $E(\boldsymbol{\epsilon}_i) = \mathbf{0}$
- $\text{cov}(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_k) = \mathbf{0}, i \neq k$
→ Individuen sind untereinander unkorreliert!
- $\text{cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}$ mit $\text{cov}(\epsilon_{il}, \epsilon_{ij}) = \sigma_{lj}$
→ Beobachtungen eines Individuums können korreliert sein!
- Verteilungsannahme:

$$\boldsymbol{\epsilon}_i \sim N_q(\mathbf{0}, \boldsymbol{\Sigma}) \quad \text{bzw.} \quad \mathbf{y}_i \sim N_q(\mathbf{x}_i^\top \mathbf{B}, \boldsymbol{\Sigma})$$

Schätzung

- zu schätzen ist die Regressionskoeffizientenmatrix \mathbf{B} und die Kovarianzmatrix $\mathbf{\Sigma}$
- die KQ-Schätzung entspricht der ML-Schätzung unter Normalverteilungsannahme
- $\hat{\beta}_{(j)} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}_{(j)}$, $j = 1, \dots, q$
⇒ die Schätzungen für die Regressionskoeffizientenvektoren entsprechen denen des zugehörigen univariaten Regressionsmodells!
- $\hat{\mathbf{B}} = \begin{pmatrix} \hat{\beta}_{(1)} & \dots & \hat{\beta}_{(q)} \end{pmatrix} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$
- $\hat{\mathbf{\Sigma}} = \frac{1}{n-p-1} \hat{\mathbf{E}}^T \hat{\mathbf{E}}$
- $\hat{\mathbf{E}} = \mathbf{Y} - \mathbf{X} \hat{\mathbf{B}} = \mathbf{Y} - \hat{\mathbf{Y}}$ (Residuenmatrix)

Gliederung

- 1 Problemstellung
- 2 Multivariates Regressionsmodell
- 3 Lineare Hypothesen**

Allgemeines Konzept

- Alle linearen Hypothesen lassen sich durch

$$H_0 : \mathbf{CBD} = \mathbf{\Gamma} \quad \text{vs.} \quad H_1 : \mathbf{CBD} \neq \mathbf{\Gamma}$$

formalisieren!

- Die Matrizen \mathbf{C} , \mathbf{D} und $\mathbf{\Gamma}$ sind durch die jeweilige Hypothese festgelegt.
- **Merke:**
 - $\text{rg}(\mathbf{C})$ entspricht der Anzahl an Hypothesen
 - die Matrix \mathbf{C} wählt die Zeilen (Kovariablen) von \mathbf{B} aus
 - die Matrix \mathbf{D} wählt die Spalten (Responsevariablen) von \mathbf{B} aus
 - mit $H_0 : \mathbf{CB} = \mathbf{\Gamma}$ können nur Hypothesen untersucht werden, die alle Responsevariablen gleichzeitig betreffen

Overall-Test

- Hat **mindestens** eine der Einflussgrößen einen Einfluss auf eine der Zielgrößen?
- Hypothesen:

$$H_0 : \beta_{rj} = 0 \quad \forall r \in \{1, \dots, p\} \text{ und } j \in \{1, \dots, q\} \quad \text{vs.}$$

$$H_1 : \beta_{rj} \neq 0 \quad \text{für mind. ein } r \text{ und } j$$

- Verwendung der linearen Hypothese $CBD = \Gamma$ mit

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots \end{pmatrix}, \quad \mathbf{\Gamma} = \mathbf{0} \text{ und } \mathbf{D} = \mathbf{I}_q$$

Test auf Signifikanz einer Einflussgröße

- Hat die r -te Variable einen Einfluss auf eine der Zielgrößen?
- Hypothesen:

$$H_0 : \beta_r = \begin{pmatrix} \beta_{r1} \\ \vdots \\ \beta_{rq} \end{pmatrix} = \mathbf{0} \quad \text{vs.}$$

$$H_1 : \beta_r \neq \mathbf{0}$$

- Wähle $\mathbf{C} = (0 \dots 0 \ 1 \ 0 \dots 0)$ mit 1 an $(r + 1)$ -ter Stelle
- Testgröße:

$$F = \frac{n - p - q}{q(n - p - 1)a_{rr}} \hat{\beta}_r^\top \hat{\Sigma}^{-1} \hat{\beta}_r \sim F(q, n - p - q)$$

a_{rr} stellt das r -te Diagonalelement von $(\mathbf{X}^\top \mathbf{X})^{-1}$ dar