5.Tutorium Multivariate Verfahren

- Multivariate Regression -

Andreas Hölzl:

16.06.2014

Shuai Shao:

12.06.2014 und 20.06.2014

Institut für Statistik, LMU München

Gliederung

Gliederung

Zur Erinnerung: Univariates Regressionsmodell

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

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

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

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

Multivariates Regressionsmodell

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

$$\mathbf{y_i} = (y_{i1}, \ldots, y_{iq})^\top$$

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

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

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

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

Gliederung

Darstellung mit Matrizen

- Modellgleichung: $\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E}$
- Matrix der Zielgrößen:

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

- Matrix der Regressionskoeffizienten:

$$\boldsymbol{B} = \begin{pmatrix} \beta_{(1)} & \cdots & \beta_{(q)} \end{pmatrix} = \begin{pmatrix} \beta_{01} & \cdots & \beta_{0q} \\ \vdots & & \vdots \\ \beta_{p1} & \cdots & \beta_{pq} \end{pmatrix}$$

- Matrix der Störgrößen:

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

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

Annahmen

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

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

Schätzung

- zu schätzen ist die Regressionskoeffizientenmatrix $\mathbf{B}$ und die Kovarianzmatrix $\mathbf{\Sigma}$
- die KQ-Schätzung entspricht der ML-Schätzung unter Normalverteilungsannahme
- $\widehat{\beta}_{(j)} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}_{(j)}\ ,\ j = 1, \ldots, q$

  $\Rightarrow$ die Schätzungen für die Regressionskoeffizientenvektoren entsprechen denen des zugehörigen univariaten Regressionsmodells!
- $\widehat{\boldsymbol{B}} = \begin{pmatrix} \widehat{\beta}_{(1)} & \ldots & \widehat{\beta}_{(q)} \end{pmatrix} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ bzw.
- $\hat{\mathbf{\Sigma}} = \frac{1}{n-p-1} \widehat{\boldsymbol{E}}^\top \widehat{\boldsymbol{E}}$
- $\widehat{\boldsymbol{E}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ (Residuenmatrix)

Gliederung

Allgemeines Konzept

- Alle lineare Hypothesen lassen sich durch

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

  formalisieren!
- Die Matrizen $\mathbf{C}$,$\mathbf{D}$ und $\mathbf{\Gamma}$ sind durch die jeweilige Hypothese festgelegt.
- **Merke:**
  - $s = \text{rg}(\mathbf{C})$ 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 $H_0 : \mathbf{CB} = \mathbf{\Gamma}$ könnnen nur Hypothesen untersucht werden, die alle Responsevariablen gleichzeitig betreffen

Overall-Test

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

$$H_0 : \beta_{sj} = 0 \quad \forall\, s \in \{1, \ldots, p\} \text{ und } j \in \{1, \ldots, q\} \qquad \text{vs.}$$
$$H_1 : \beta_{sj} \neq 0 \quad \text{für mind. ein s und j}$$

- Verwendung der linearen Hypothese $\boldsymbol{CBD} = \boldsymbol{\Gamma}$ mit

$$\boldsymbol{C} = \begin{pmatrix} 0 & 1 & 0 & 0 & \ldots \\ 0 & 0 & 1 & 0 & \ldots \\ \vdots & \vdots & \vdots & \ddots & \vdots \end{pmatrix}, \; \boldsymbol{\Gamma} = \mathbf{0} \text{ und } \mathbf{D} = \mathbf{I}_p$$

Test auf Signifikanz einer Einflussgröße

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

$$H_0 : \boldsymbol{\beta_s} = \begin{pmatrix} \beta_{s1} \\ \vdots \\ \beta_{sq} \end{pmatrix} = 0 \qquad \text{vs.}$$

$$H_1 : \boldsymbol{\beta_s} \neq 0$$

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

$$F = \frac{n-p-q}{q(n-p-1)a_{ss}} \widehat{\boldsymbol{\beta}}_{\boldsymbol{s}}^{\top} \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\beta}}_{\boldsymbol{s}} \ \sim \ F(q, n-p-q)$$

a_{ss} stellt das s-te Diagonalelement von $(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}$ dar

Wilk's Lambda

- M das multivariate lineare Modell
- M_0 das multivariate lineare Modell mit $\mathbf{CBD} = \mathbf{\Gamma}$
- Test auf Gültigkeit von $H_0(M_0)$:

$$\mathbf{\Lambda} = \frac{|SSP(M)|}{|SSP(M_0)|} = \frac{|SSP(M)|}{|SSP(M) + SSP(M_0|M)|}$$

mit

$$SSP(M) = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}})^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}) \quad \text{(volles Modell)}$$

$$SSP(M_0) = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}_{\boldsymbol{0}})^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}_{\boldsymbol{0}}) \quad \text{(restringiertes Modell)}$$

$$SSP(M_0|M) = SSP(M_0) - SSP(M)$$

- Unter H_0 gilt: $\mathbf{\Lambda} \sim \mathbf{\Lambda}(q, n-p-1, rang(\boldsymbol{C}))$
  $\Rightarrow$ H_0 wird abgelehnt, falls $\mathbf{\Lambda} < \Lambda(q, n-p-1, rang(\boldsymbol{C}); \alpha)$