Statistik IV für Nebenfachstudierende

1. Grundlagen

Andreas Mayr

Institut für Statistik, LMU München

Sommersemester 2017

Organisatorisches

Vorlesungen

- **Montag** (ganzes Semester)

  12:15 – 13:45 , A120

- **Mittwoch** (bis Mitte Juni)

  14:15 – 15:45, A213

Übung (Leitung: M. Schneider)

- **Donnerstag**

  12:15 – 13:45, A120

Klausur

- **Mittwoch, 26.07.2017 (geplant)**

  14:00 – 16:00

Übersicht

1. Grundlagen
2. Schätzen und Testen
3. Multivariate lineare Regression
4. Mehrkategorielle Regression
5. Diskriminanzanalyse, Klassifikation
6. Clusteranalyse

1. Grundlagen

1. Datenmatrix

$$\underline{\underline{X}} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

i-ter Beobachtungsvektor:

$\underline{x}_i =$ $\underline{\underline{X}} =$

j-ter Vektor, zur j-ten Variablen:

$\underline{x}_{\cdot j} =$ $\underline{\underline{X}} =$

Zusammen:

$\underline{\underline{X}} =$ $=$

Zufallsvariablen

Meist: Man beobachtet zufällig, unabhängig die **Zufallsvariablen**

$X_1, ..., X_p$. Dadurch werden die Spalten der Matrix generiert.

2. Mehrdimensionale Zufallsvariablen

Eindimensional: Wie wird der Zufall charakterisiert?

- Zugrunde liegende Verteilung, unbekannt aber vorhanden.
- Dichte auf Träger
  - diskret
  - stetig
- Alternative: Verteilungsfunktion

Mehrdimensional: ▶ Zufallsvektor

$\underline{x} \; =$

- Dichte auf Träger
  - diskret:

    $f(\underline{x}) = f(x_1, \ldots, x_p) =$

- stetig: Gemeinsame Wahrscheinlichkeit

$$\mathbb{P}(X_1 \in [a_1, b_1], X_2 \in [a_2, b_2], \ldots) =$$

- Alternative: Verteilungsfunktion:

$$F(x_1, \ldots, x_p) = \mathbb{P}(X_1 \leq x_1, \ldots X_p \leq x_p) =$$

$$f(\underline{x}) =$$

Erwartungswertvektor:

$$\underline{\mu} \;=\; E(\underline{X}) \;=$$

Erwartungswert

$E(X)$ bzw. $E(\underline{X})$ sind **feste** Größen (fixe Parameter), keine Zufallsvariablen mehr.

3. Varianz - Kovarianz - Struktur

$$\sigma_{ij} = \mathsf{Cov}(X_i, X_j) =$$

$$\sigma_{ii} = \mathsf{Cov}(X_i, X_i) =$$

$$\sigma_{ii} = \sigma_i^2$$

Kovarianzmatrix

$$\underline{\underline{\Sigma}} \quad =$$

Schreibweise

$$\underline{\underline{\Sigma}} \;=\; \mathsf{Cov}(\underline{X}) \;= \mathbb{E}\left((\underline{X}-\underline{\mu})(\underline{X}-\underline{\mu})^{\top}\right) \;=$$

Inhaltlich interessanter: Korrelation

$$r_{ij} = \frac{\mathsf{Cov}(X_i, X_j)}{\sqrt{\mathsf{Var}(X_i)}\sqrt{\mathsf{Var}(X_j)}}$$

Korrelationsmatrix

$$\mathsf{R} = (r_{ij}) =$$

Kovarianz zwischen zwei Vektoren

$$\underline{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \qquad \underline{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_q \end{pmatrix}$$

$$\underline{\underline{\Sigma}}_{X,Y} = \mathsf{Cov}(X, Y) =$$

Regeln:

Zufallsvektoren $\underline{X}$ und $\underline{Y}$.

- $\mathbb{E}(\underline{X} + \underline{Y}) =$
- $\mathbb{E}(\underline{\underline{A}}\ \underline{X} + \underline{b}) =$
- $\mathrm{Cov}(\underline{\underline{A}}\ \underline{X} + \underline{b}) =$

Spezialfälle:

- $\mathrm{Cov}(\underline{X}, \underline{X}) = \mathrm{Cov}(\underline{X})$
- $\mathrm{Cov}(\underline{X}, \underline{Y}) = \mathrm{Cov}(\underline{X}, \underline{Y})^\top$
- $\mathrm{Cov}(\underline{a}^\top \underline{X}) =$

Eigenschaften:

- $\underline{\underline{\Sigma}} = \text{Cov}(\underline{X}) =$
  1. symmetrische Matrix
  2. nicht negativ definit

     Matrix $\underline{\underline{\Sigma}}$ ist nicht negativ definit g.d.w.:

     - $\underline{a}^\top \underline{\underline{\Sigma}}\ \underline{a} \geq 0 \quad \forall \underline{a}$

4. Normalverteilung

$p = 1$:

Dichte $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

$p > 1$: $\underline{x} = (x_1, \ldots, x_p)^\top$

Dichte $f(\underline{x}) = \frac{1}{(2\pi)^{\frac{1}{2}} \underline{\underline{\Sigma}}^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\underline{x} - \underline{\mu})^\top \underline{\underline{\Sigma}}^{-1}(\underline{x} - \underline{\mu})\right)$

5. Standardisierung eines Zufallsvektors

mit $\mu = \mathbb{E}(\underline{x})$ und $\underline{\underline{\Sigma}} = \mathsf{Cov}(x)$, beliebig verteilt.

- $p = 1$:

  $z = \frac{x-\mu}{\sigma} \Rightarrow \mathbb{E}(z) = 0, \quad \mathsf{Var}(z) = 1$

- $p > 1$:

  "$\frac{\underline{x}-\underline{\mu}}{\underline{\underline{\Sigma}}}$" $\longrightarrow$ Richtiger Gedanke, **falsche** Notation!

  Notwendig ist Wurzel aus $\underline{\underline{\Sigma}}$:

  - Voraussetzung $\underline{\underline{\Sigma}}$ pos.definit, d.h

$$\underline{a}^\top \underline{\underline{\Sigma}}\ \underline{a} > 0 \quad \forall \underline{a} \neq \underline{0}$$

  - Dann existiert eine symmetrische Matrix $\underline{\underline{\Sigma}}^{\frac{1}{2}}$, so dass

$$\underline{\underline{\Sigma}} = \underline{\underline{\Sigma}}^{\frac{1}{2}}\ \underline{\underline{\Sigma}}^{\frac{1}{2}}$$

  - Und für $\underline{\underline{\Sigma}}^{-1}$ gilt:

$$\underline{\underline{\Sigma}}^{-1} = \underline{\underline{\Sigma}}^{-\frac{1}{2}}\ \underline{\underline{\Sigma}}^{-\frac{1}{2}}$$

Standardisieren von multiv. ZV

$$\underline{Z} = \underline{\underline{\Sigma}}^{-\frac{1}{2}}(\underline{x} - \underline{\mu}),$$

weil

$$\begin{aligned}
\mathbb{E}(\underline{Z}) &= \underline{\underline{\Sigma}}^{-\frac{1}{2}}(\mathbb{E}(\underline{x}) - \underline{\mu}) = 0 \\
\mathsf{Cov}(\underline{Z}) &= \underline{\underline{\Sigma}}^{-\frac{1}{2}}\ \underline{\underline{\Sigma}}\ \underline{\underline{\Sigma}}^{-\frac{1}{2}} \\
&= \underline{\underline{\Sigma}}^{-\frac{1}{2}}\ \underline{\underline{\Sigma}}^{\frac{1}{2}}\ \underline{\underline{\Sigma}}^{\frac{1}{2}}\ \underline{\underline{\Sigma}}^{-\frac{1}{2}} = \underline{\underline{\mathbf{I}}}
\end{aligned}$$

Graphisch:

Aussagen

Wenn $X \sim N(\underline{\mu}, \underline{\underline{\Sigma}})$:

(1) $\underline{Y} = \underline{\underline{A}}\ \underline{x} + \underline{a}$

- Es gilt:

$$\underline{Y} \sim N(\underline{\underline{A}}\ \underline{\mu} + \underline{a},\ \underline{\underline{A}}\ \underline{\underline{\Sigma}}\ \underline{\underline{A}}^{\top})$$

- Daraus folgt:

(2) Bedingte Verteilung

- Zwei Variablengruppen

$\underline{x} =$ $\underline{y} =$

Gesucht ist Verteilung von $\underline{Y}|\underline{X}$

Wenn $(\underline{Y}, \underline{X})$ gemeinsam normalverteilt, dh.

- Bedingte Dichte für $\underline{Y}|\underline{X}$, d.h

$$f_{Y|X}(\underline{y}|\underline{x}) = \frac{f(\underline{y}, \underline{x})}{f_x(\underline{x})}$$

folgt

$$\underline{Y}|\underline{X} \sim N(\underline{\mu}_{Y|X},\ \underline{\underline{\Sigma}}_{Y|X})$$

wieder normalverteilt mit

$$\underline{\mu}_{Y|X} = \underline{\mu}_y + \underline{\underline{\Sigma}}_{yx}\ \underline{\underline{\Sigma}}_x^{-1}(\underline{x} - \underline{\mu})$$

$$\underline{\underline{\Sigma}}_{Y|X} = \underline{\underline{\Sigma}}_y - \underline{\underline{\Sigma}}_{yx}\underline{\underline{\Sigma}}_x^{-1}\underline{\underline{\Sigma}}_{xy}$$

Beispiel: $p = q = 1$