Tutorium

Allgemeines Lineares Modell

BSc Psychologie SoSe 2023

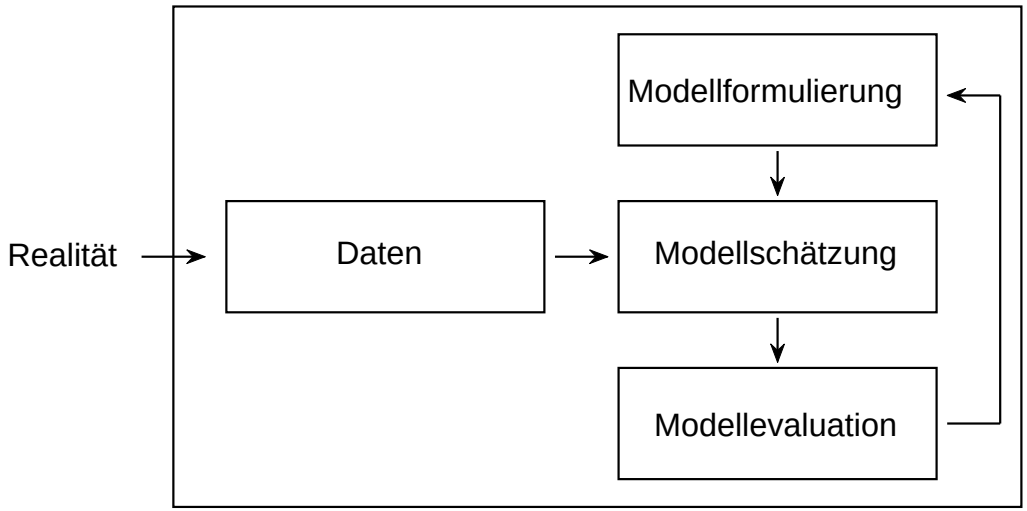
7. Termin: (5) Modellformulierung

Sean Mulready

Selbstkontrollfragen - Modellformulierung

1. Erläutern Sie das naturwissenschaftliche Paradigma.
2. Erläutern Sie die Standardprobleme der Frequentistischen Inferenz.
3. Geben Sie die Definition des Allgemeinen Linearen Modells wieder.
4. Erläutern Sie die deterministischen und probabilistischen Aspekte des ALMs.
5. Wieviele skalare Parameter hat das ALM mit sphärischer Kovarianzmatrix?
6. Warum sind die Komponenten des ALM Zufallsfehlers unabhängig und identisch verteilt?
7. Geben Sie das Theorem zur Datenverteilung des Allgemeinen Linearen Modells wieder.
8. Sind die Komponenten des ALM Datenvektors immer unabhängig und identisch verteilt?
9. Schreiben Sie das Szenario n unabhängig und identisch normalverteilten Zufallsvariablen in ALM Form.
10. Schreiben Sie das Szenario der einfachen linearen Regression in ALM Form.

1. Erläutern Sie das naturwissenschaftliche Paradigma.



- Wir nehmen an, dass eine **Realität** existiert, welche wir idR nur indirekt, teilweise und eingeschränkt beobachten können, indem wir **Daten** erheben (z.B. BDI Fragebogendaten, EEG-Messung).
- Daten $\neq$ Realität. Daten sind eine Beobachtung/Messung der Realität.
- In der (Natur-)wissenschaft bilden wir Theorien und formulieren Modelle über die Realität (**Modellformulierung**). Mithilfe von Modellen treffen wir Vorhersagen über die Realität.
- Wir verwenden Daten, um Modelle zu schätzen (Modellschätzung) und darauf basierend die Güte der Modelle evaluieren (**Modellevaluation**)
- Ergebnisse der Modellevaluation können wiederum dazu verwendet werden die Modellformulierung anzupassen.
- Angepasste/veränderte Modelle können wieder mit Daten geschätzt und deren Güte evaluiert werden.

2. Erläutern Sie die Standardprobleme der Frequentistischen Inferenz.

Standardprobleme Frequentistischer Inferenz

(1) Parameterschätzung

Ziel der Parameterschätzung ist es, einen möglichst guten Tipp für wahre, aber unbekannte, Parameterwerte oderFunktionen dieser abzugeben, typischerweise mithilfe von Daten.

(2) Konfidenzintervalle

Ziel der Bestimmung von Konfidenzintervallen ist es, basierend auf der angenommenen Verteilung der Daten eine quantitative Aussage über die mit Schätzwerten assoziierte Unsicherheit zu treffen.

(3) Hypothesentests

Das Ziel des Hypothesentestens ist es, basierend auf der angenommenen Verteilung der Daten in einer möglichst zuverlässigen Form zu entscheiden, ob ein wahrer, aber unbekannter Parameterwert in einer von zwei sich gegenseitig ausschließenden Untermengen des Parameterraumes liegt.

3. Geben Sie die Definition des Allgemeinen Linearen Modells wieder.

Definition (Allgemeines Lineares Modell)

Es sei

$$\upsilon = X\beta + \varepsilon, \tag{1}$$

wobei

- υ ein n-dimensionaler beobachtbarer Zufallsvektor ist, der *Daten* genannt wird,
- $X \in \mathbb{R}^{n \times p}$ mit $n > p$ eine vorgegebene Matrix ist, die *Designmatrix* genannt wird,
- $\beta \in \mathbb{R}^p$ ein unbekannter Parametervektor ist, der *Betaparametervektor* genannt wird,
- ε ein n-dimensionaler nicht-beobachtbarer Zufallsvektor ist, der *Zufallsfehler* genannt wird und für den angenommen wird, dass mit einem unbekannten Varianzparameter $\sigma^2 > 0$ gilt, dass

$$\varepsilon \sim N\left(0_n, \sigma^2 I_n\right). \tag{2}$$

Dann heißt (1) *Allgemeines Lineares Modell (ALM)*.

Beispiele ALM mit $n = 5$

υ sei ein 5-dimensionalter Zufallsvektor mit Erwartungswertparameter $X\beta \in \mathbb{R}^{n\times p}$ und Kovarianzmatrixparameter $\sigma^2 I_n$. Die Komponenten $\upsilon_1, ..., \upsilon_n$ sind unabhängig aber nicht identisch verteilte Zufallsvariablen der Form $\upsilon_i \sim N(\mu_i, \sigma^2)$ für $i = 1, ..., n$.

Beispiel 1: p = 1 (Wir haben nur einen Betaparameter β)

$$\upsilon \sim N(X\beta, \sigma^2 I_5) \text{ mit } X \in \mathbb{R}^{5\times 1}, \beta \in \mathbb{R}^1, \sigma^2 > 0.$$

$$\upsilon = X\beta + \varepsilon \Leftrightarrow \begin{pmatrix} \upsilon_1 \\ \upsilon_2 \\ \upsilon_3 \\ \upsilon_4 \\ \upsilon_5 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix} = \begin{pmatrix} x_1\beta + \varepsilon_1 \\ x_2\beta + \varepsilon_2 \\ x_3\beta + \varepsilon_3 \\ x_4\beta + \varepsilon_4 \\ x_5\beta + \varepsilon_5 \end{pmatrix}$$

Beispiel 2: p = 2 (Wir haben zwei Betaparameter β_1 und β_2)

$$\upsilon \sim N(X\beta, \sigma^2 I_5) \text{ mit } X \in \mathbb{R}^{5\times 2}, \beta \in \mathbb{R}^2, \sigma^2 > 0.$$

$$\upsilon = X\beta + \varepsilon \Leftrightarrow \begin{pmatrix} \upsilon_1 \\ \upsilon_2 \\ \upsilon_3 \\ \upsilon_4 \\ \upsilon_5 \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \\ x_{51} & x_{52} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix} = \begin{pmatrix} x_{11}\beta_1 + x_{12}\beta_2 + \varepsilon_1 \\ x_{21}\beta_1 + x_{22}\beta_2 + \varepsilon_2 \\ x_{31}\beta_1 + x_{32}\beta_2 + \varepsilon_3 \\ x_{41}\beta_1 + x_{42}\beta_2 + \varepsilon_4 \\ x_{51}\beta_1 + x_{52}\beta_2 + \varepsilon_5 \end{pmatrix}$$

4.Erläutern Sie die deterministischen und probabilistischen Aspekte des ALMs.

- Wir nennen $X\beta \in \mathbb{R}^n$ den *deterministichen Modellaspekt* und ε den *probabilistischen Modellaspekt*.
- *deterministisch* heißt hier, die Komponenten beinhalten keine Zufälligkeit, sondern sind vorgegeben bzw. im Rahmen der Modellformulierung festgelegt.
- *probabilistisch* heißt hier, die Komponenten beinhalten Zufälligkeit. Realisierungen dieser Komponente können aus einer Normalverteilung gezogen werden. Der probabilistische Aspekt modelliert bei Normalverteilungen alle Einflussfaktoren auf υ, die nicht durch den deterministischen Aspekt abgedeckt werden.
- da υ das Ergebnis der Addition deterministischer und probabilistischer Aspekte ist, ist es auch probabilistisch (i.e. zufällig).

5.Wieviele skalare Parameter hat das ALM mit sphärischer Kovarianzmatrix?

Zur Erinnerung: sphärische Koviaranzmatrix

$$\sigma^2 I_n = \sigma^2 \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix} = \begin{pmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{pmatrix}$$

Beispiel $n = 5$

$$\sigma^2 I_5 = \sigma^2 \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 \end{pmatrix}$$

Das ALM mit sphärischer Kovarianzmatrix hat $p + 1$ Parameter (p Betaparameter und 1 Varianzparameter)

6.Warum sind die Komponenten des ALM Zufallsfehlers unabhängig und identisch verteilt?

Wir gehen davon aus, dass alle weiteren Einflüsse, die der deterministische Aspekt des Modells nicht erklärt (auch unbekannte "Störeinflüsse" genannt), *viele* und *unabhängig* sind, und modellieren diese als eine Vielzahl unabhängiger Zufallsvariabeln.

Im Sinne des zentralen Grenzwertsatzes ist die Summe vieler unabhängiger Zufallsvariablen asymptotisch, d.h. für unendlich viele Zufallsvariablen, normalverteilt.

Der Zufallsfehler modelliert also alle nicht durch den deterministischen Aspekt des Modells erklären Einflüsse, die aufaddiert als normalverteilt angenommen werden.

Formal gilt $\varepsilon \sim N(0_n, \sigma^2 I_n)$, wobei der Erwartungswertparameter 0_n bedeutet, dass alle Komponenten $\varepsilon_1, \ldots, \varepsilon_n$ den Erwartungswert 0 haben, und der sphärische Kovarianzmatrixparameter $0_n, \sigma^2 I_n$, bedeutet, dass alle Komponenten die Varianz σ^2 haben und alle Kovarianzen gleich 0 sind.

- $\Rightarrow$ identisch verteilte Komponenten, weil alle Komponenten den Erwartungswert 0 und den Varianzparameter σ^2 haben.
- $\Rightarrow$ unabhängige Komponenten, weil alle nicht-diagonal-Elemente, also alle Kovarianzen zwischen Komponenten gleich 0

7.Geben Sie das Theorem zur Datenverteilung des Allgemeinen Linearen Modells wieder.

Theorem (ALM Datenverteilung)

Es sei

$$\upsilon = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \tag{3}$$

das ALM. Dann gilt

$$\upsilon \sim N(\mu, \sigma^2 I_n) \text{ mit } \mu := X\beta \in \mathbb{R}^n. \tag{4}$$

8.Sind die Komponenten des ALM Datenvektors immer unabhängig und identisch verteilt?

$\upsilon \sim N(X\beta, \sigma^2 I_n)$ mit $\upsilon_i \sim N(\mu_i, \sigma^2)$ für $i = 1, ..., n$.

- Der Kovarianzmatrixparameter ist gegeben gegeben durch $\sigma^2 I_n \in \mathbb{R}^{n \times n}$ $\Rightarrow$ sphärische Kovarianzmatrix $\Rightarrow$ **unabhängige** Komponenten $\upsilon_1, ..., \upsilon_n$
- Der Erwartungswertparameter ist gegeben durch $X\beta \in \mathbb{R}^n$ $\Rightarrow$ Vektor mit n Einträgen, die in Abhängigkeit von der Designmatrix X für jede Komponente υ_i einen anderen Erwartungswert μ_i annehmen kann. $\Rightarrow$ **nicht identisch** verteilte Komponenten υ_i.

9.Schreiben Sie das Szenario n unabhängig und identisch normalverteilten Zufallsvariablen in ALM Form.

Wir betrachten das Szenario von n unabhängigen und identisch normalverteilten Zufallsvariabeln mit Erwartungswertparameter $\mu \in \mathbb{R}$ und Varianzparameter σ^2, mit $\upsilon_i \sim N(\mu, \sigma^2)$ für $i = 1, ..., n$.

Anmerkung: Während wir im Theorem zur Datenverteilung im ALM noch gesehen haben, dass die Komponenten $\upsilon_1, ..., \upsilon_n$ jeweils "individuelle" Verteilungen $\upsilon_i \sim N(\mu_i, \sigma^2)$ mit "individuellen" μ_i haben, und somit nicht identisch verteilt sind, haben wir im Szenario n **unabhängig und *identisch* verteilter Zufallsvariablen** nun nur noch ein μ gegeben, das für alle Komponenten $\upsilon_1, ..., \upsilon_n$, also für alle υ_i gleich ist.

In Matrixschreibweise:

$$\upsilon \sim N(X\beta, \sigma^2 I_n) \text{ mit } X := 1_n \in \mathbb{R}^{n \times 1}, \beta := \mu \in \mathbb{R}^1, \sigma^2 > 0.$$

Unabhängige und identisch normalverteilte ZV - Selbstkontrollfragen

Beispiel $n = 5$ unabhängig und identisch normalverteilte Zufallsvariablen

Wir betrachten das Szenario von $n = 5$ unabhängigen und identisch normalverteilten Zufallsvariablen mit Erwartungswertparameter $\mu \in \mathbb{R}$ und Varianzparameter σ^2, mit $\upsilon_i \sim N(\mu, \sigma^2)$ für $i = 1, ..., 5$.

In Matrixschreibweise:

$$\upsilon \sim N(X\beta, \sigma^2 I_5) \text{ mit } X := 1_5 \in \mathbb{R}^{5\times 1}, \beta := \mu \in \mathbb{R}, \sigma^2 > 0.$$

$$\upsilon = X\beta + \varepsilon = X\mu + \varepsilon \Leftrightarrow \begin{pmatrix} \upsilon_1 \\ \upsilon_2 \\ \upsilon_3 \\ \upsilon_4 \\ \upsilon_5 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix} = \begin{pmatrix} \mu + \varepsilon_1 \\ \mu + \varepsilon_2 \\ \mu + \varepsilon_3 \\ \mu + \varepsilon_4 \\ \mu + \varepsilon_5 \end{pmatrix}$$

10.Schreiben Sie das Szenario der einfachen linearen Regression in ALM Form.

$$\upsilon \sim N(X\beta, \sigma^2 I_n) \text{ mit } X := \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix} \in \mathbb{R}^{n\times 2}, \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \in \mathbb{R}^2, \sigma^2 > 0.$$

Beispiel mit $n = 5$

$$\upsilon = X\beta + \varepsilon \Leftrightarrow \begin{pmatrix} \upsilon_1 \\ \upsilon_2 \\ \upsilon_3 \\ \upsilon_4 \\ \upsilon_5 \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & x_5 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix} = \begin{pmatrix} \beta_0 + x_1\beta_1 + \varepsilon_1 \\ \beta_0 + x_2\beta_1 + \varepsilon_2 \\ \beta_0 + x_3\beta_1 + \varepsilon_3 \\ \beta_0 + x_4\beta_1 + \varepsilon_4 \\ \beta_0 + x_5\beta_1 + \varepsilon_5 \end{pmatrix}$$