Allgemeines Lineares Modell

BSc Psychologie SoSe 2023

Prof. Dr. Dirk Ostwald

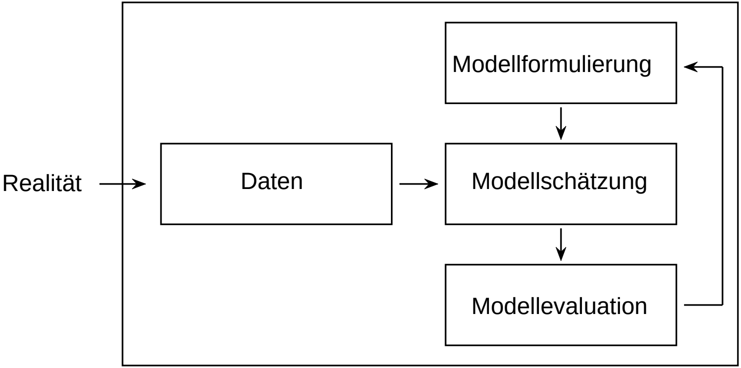
(5) Modellformulierung

Überblick

Modul B2 Inferenzstatistik | Allgemeines Lineares Modell

Datum	Einheit	Thema
13.04.2023	Grundlagen	(1) Regression
20.04.2023	Grundlagen	(2) Korrelation
27.04.2023	Grundlagen	(3) Matrizen
04.05.2023	Grundlagen	(4) Normalverteilungen
11.05.2023	Theorie	(5) Modellformulierung
18.05.2023	Feiertag	
25.05.2023	Theorie	(6) Modellschätzung
01.06.2023	Theorie	(7) T-Statistiken
08.06.2021	Theorie	(8) F-Statistiken
15.06.2021	Anwendung	(9) T-Tests
22.06.2021	Anwendung	(10) Einfaktorielle Varianzanalyse
29.06.2023	Anwendung	(11) Zweifaktorielle Varianzanalyse
06.07.2023	Anwendung	(12) Multiple Regression
13.07.2023	Anwendung	(13) Kovarianzanalyse
20.07.2023	Klausurtermin	
März 2024	Klausurwiederholungstermin	

Naturwissenschaft



Modellformulierung

$$\upsilon = X\beta + \varepsilon, \varepsilon \sim N(0_n, \sigma^2 I_n) \tag{1}$$

Modellschätzung

$$\hat{\beta} = (X^T X)^{-1} X^T \upsilon, \hat{\sigma}^2 = \frac{(\upsilon - X\hat{\beta})^T(\upsilon - X\hat{\beta})}{n-p} \tag{2}$$

Modellevaluation

$$T = \frac{c^T\hat{\beta} - c^T\beta_0}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}}, F = \frac{(\hat{\varepsilon}_0^T\hat{\varepsilon}_0 - \hat{\varepsilon}^T\hat{\varepsilon})/p_1}{\hat{\varepsilon}^T\hat{\varepsilon}/(n-p)} \tag{3}$$

Standardprobleme Frequentistischer Inferenz

(1) Parameterschätzung

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

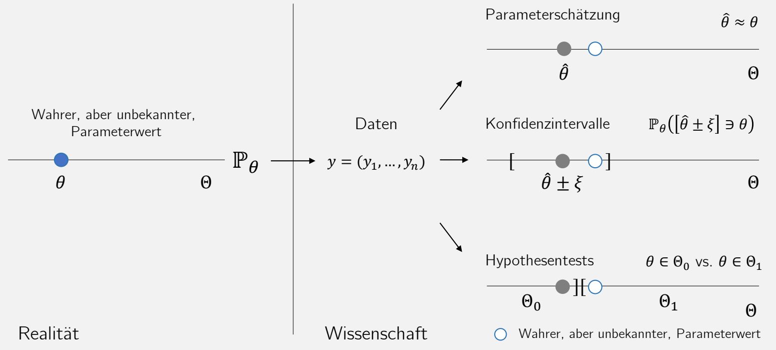
(2) Konfidenzintervalle

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

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.

Modell und Standardprobleme Frequentistischer Inferenz



$\theta := (\beta, \sigma^2)$, $\Theta := \mathbb{R}^p \times \mathbb{R}_{>0}$ $\mathbb{P}_\theta(\upsilon) := \mathbb{P}_{\beta,\sigma^2}(\upsilon)$ mit WDF $p_{\beta,\sigma^2}(y) := N(y; X\beta, \sigma^2 I_n)$

Standardannahmen Frequentistischer Inferenz

Gegeben sei das Allgemeine Lineare Modell. Es wird angenommen, dass ein vorliegender Datensatz eine der möglichen Realisierungen der Daten des Modells ist. Aus Frequentistischer Sicht kann man unendlich oft Datensätze basierend auf einem Modell generieren und zu jedem Datensatz Schätzer oder Statistiken auswerten, z.B. den Betaparameterschätzer

Datensatz (1) : $y^{(1)} = \left(y_1^{(1)}, y_2^{(1)}, ..., y_n^{(1)}\right)^T$ mit $\hat{\beta}^{(1)} = (X^TX)^{-1}X^Ty^{(1)}$

Datensatz (2) : $y^{(2)} = \left(y_1^{(2)}, y_2^{(2)}, ..., y_n^{(2)}\right)^T$ mit $\hat{\beta}^{(2)} = (X^TX)^{-1}X^Ty^{(2)}$

Datensatz (3) : $y^{(3)} = \left(y_1^{(3)}, y_2^{(3)}, ..., y_n^{(3)}\right)^T$ mit $\hat{\beta}^{(3)} = (X^TX)^{-1}X^Ty^{(3)}$

Datensatz (4) : $y^{(4)} = \left(y_1^{(4)}, y_2^{(4)}, ..., y_n^{(4)}\right)^T$ mit $\hat{\beta}^{(4)} = (X^TX)^{-1}X^Ty^{(4)}$

Datensatz (5) : $y^{(5)} = ...$

Um die Qualität statistischer Methoden zu beurteilen betrachtet die Frequentistische Statistik die Wahrscheinlichkeitsverteilungen von Schätzern und Statistiken unter Annahme der Datenverteilung. Was zum Beispiel ist die Verteilung von $\hat{\beta}^{(1)}$, $\hat{\beta}^{(2)}$, $\hat{\beta}^{(3)}$, $\hat{\beta}^{(4)}$, ... also die Verteilung der Zufallsvariable $\hat{\beta} := (X^TX)^{-1}X^T\upsilon$? Wenn eine statistische Methode im Sinne der Frequentistischen Standardannahmen "gut" ist, dann heißt das also, dass sie bei häufiger Anwendung "im Mittel gut" ist. Im Einzelfall, also im Normalfall nur eines vorliegenden Datensatzes, kann sie auch "schlecht" sein.

Überblick

Anwendungsbeispiele Einheiten (5) - (8)

- Unabhängig und identisch normalverteilte Zufallsvariablen | Einstichproben-T-Test
- Einfache lineare Regression

Anwendungsbeispiele Einheiten (9) - (13)

- Zweistichproben-T-Tests
- Einfaktorielle Varianzanalyse
- Zweifaktorielle Varianzanalyse
- Multiple Regression
- Kovarianzanalyse

Allgemeine Theorie

Unabhängige und identisch normalverteilte Zufallsvariablen

Einfache lineare Regression

Selbstkontrollfragen

Allgemeine Theorie

Unabhängige und identisch normalverteilte Zufallsvariablen

Einfache lineare Regression

Selbstkontrollfragen

Definition (Allgemeines Lineares Modell)

Es sei

$$\upsilon = X\beta + \varepsilon, \tag{4}$$

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

$$\varepsilon \sim N\left(0_n, \sigma^2 I_n\right). \tag{5}$$

Dann heißt (4) *Allgemeines Lineares Modell (ALM)*.

Allgemeine Theorie

Bemerkungen

- Wir nehmen durchgängig an, dass $X \in \mathbb{R}^{n \times p}$ vollen Spaltenrang hat, also dass $\operatorname{rg}(X) = p$.
- υ ist ein Zufallsvektor, weil er aus der Addition des Zufallsvektors ε zu dem Vektor $X\beta \in \mathbb{R}^n$ resultiert.
- Wir nennen $X\beta \in \mathbb{R}^n$ den *deterministischen Modellaspekt* und ε den *probabilistischen Modellaspekt*.
- $n \in \mathbb{N}$ bezeichnet durchgängig die Anzahl an Datenpunkten.
- $p \in \mathbb{N}$ bezeichnet durchgängig die Anzahl an Betaparametern.
- Die Gesamtzahl an Parametern des ALMs ist $p + 1$ (p Betaparameterkomponenten und 1 Varianzparameter).
- Der Betaparametervektor wird auch *Gewichtsvektor* oder *Effektvektor* genannt.
- Weil der Kovarianzmatrixparameter von ε als sphärisch angenommen wird, sind die $\varepsilon_1, ..., \varepsilon_n$ unabhängige normalverteilte Zufallsvariablen mit identischem Varianzparameter; weil zusätzlich der Erwartungswertparameter von ε als 0_n angenommen wird, sind die $\varepsilon_1, ..., \varepsilon_n$ auch identisch normalverteilte Zufallsvariablen.
- Für jede Komponente $\upsilon_i, i = 1, ..., n$ von υ impliziert (4) nach Definition des Matrixprodukts, dass

$$\upsilon_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \varepsilon_i \text{ mit } \varepsilon_i \sim N\left(0, \sigma^2\right), \tag{6}$$

  wobei $x_{ij} \in \mathbb{R}$ das ijte Element der Designmatrix X bezeichnet.

Theorem (Datenverteilung des Allgemeinen Linearen Modells)

Es sei

$$\upsilon = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \tag{7}$$

das ALM. Dann gilt

$$\upsilon \sim N(\mu, \sigma^2 I_n) \text{ mit } \mu := X\beta \in \mathbb{R}^n. \tag{8}$$

<u>Beweis</u>

Mit dem Theorem zur linear-affinen Transformation multivariater Normalverteilungen gilt für $\varepsilon \sim N(0_n, \sigma^2 I_n)$ und $\upsilon := I_n\varepsilon + X\beta$, dass

$$\upsilon \sim N\left(I_n 0_n + X\beta, I_n(\sigma^2 I_n)I_n^T\right) = N(X\beta, \sigma^2 I_n) = N(\mu, \sigma^2 I_n) \text{ mit } \mu := X\beta \in \mathbb{R}^n. \tag{9}$$

Bemerkungen

- Im ALM sind die Daten υ also ein n-dimensionaler normalverteilter Zufallsvektor mit Erwartungswertparameter $\mu = X\beta \in \mathbb{R}^n$ und Kovarianzmatrixparameter $\sigma^2 I_n \in \mathbb{R}^{n\times n}$.
- Die Komponenten $\upsilon_1, ..., \upsilon_n$ von υ, also die Datenpunkte, sind damit unabhängige, aber im Allgemeinen nicht identisch verteilte, normalverteilte Zufallsvariablen der Form $\upsilon_i \sim N\left(\mu_i, \sigma^2\right)$ für $i = 1, ..., n$.

Allgemeine Theorie

Unabhängige und identisch normalverteilte Zufallsvariablen

Einfache lineare Regression

Selbstkontrollfragen

Unabhängige und identisch normalverteilte Zufallsvariablen

Wir betrachten das Szenario von n unabhängigen und identisch normalverteilten Zufallsvariablen mit Erwartungswertparameter $\mu \in \mathbb{R}$ und Varianzparameter σ^2,

$$\upsilon_i \sim N(\mu, \sigma^2) \text{ für } i = 1, ..., n. \tag{10}$$

Dann gilt, dass (10) äquivalent ist zu

$$\upsilon_i = \mu + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2) \text{ für } i = 1, ..., n \text{ mit unabhängigen } \varepsilon_i. \tag{11}$$

In Matrixschreibweise ist dies wiederum äquivalent zu

$$\upsilon \sim N(X\beta, \sigma^2 I_n) \text{ mit } X := 1_n \in \mathbb{R}^{n \times 1}, \beta := \mu \in \mathbb{R}^1, \sigma^2 > 0. \tag{12}$$

Bemerkungen

- Wir kennen dieses Modell bereits aus Einheit (9) Grundbegriffe Frequentistischer Inferenz in Wahrscheinlichkeitstheorie und Frequentistsche Inferenz, dort haben wir es geschrieben als

$$\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2) \text{ mit } (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}. \tag{13}$$

Unabhängige und identisch normalverteilte Zufallsvariablen

```r
# Modellformulierung
library(MASS)                                   # Multivariate Normalverteilung
n      = 12                                     # Anzahl von Datenpunkten
p      = 1                                      # Anzahl von Betaparameter
X      = matrix(rep(1,n), nrow = n)             # Designmatrix
I_n    = diag(n)                                # n x n Einheitsmatrix
beta   = 2                                      # wahrer, aber unbekannter, Betaparameter
sigsqr = 1                                      # wahrer, aber unbekannter, Varianzparameter

# Datenrealisierung
y      = mvrnorm(1, X %*% beta, sigsqr*I_n)  # eine Realisierung eines n-dimensionalen ZVs
print(y)
```

```
>  [1]  2.629  1.446  1.717  1.756  1.753  0.178  3.148  2.622  1.994
> [10] -0.437  2.255  0.600
```

Allgemeine Theorie

Unabhängige und identisch normalverteilte Zufallsvariablen

Einfache lineare Regression

Selbstkontrollfragen

Einfache lineare Regression

Wir betrachten das Modell der einfachen linearen Regression aus Einheit (1) Regression,

$$\upsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2) \text{ für } i = 1, ..., n, \tag{14}$$

Wir haben bereits gesehen, dass dieses Modell folgende Datenverteilung besitzt:

$$\upsilon_i \sim N(\mu_i, \sigma^2) \text{ mit } \mu_i := \beta_0 + \beta_1 x_i \text{ für } i = 1, ..., n. \tag{15}$$

In Matrixschreibweise ist dies wiederrum äquivalent zu

$$\upsilon \sim N(X\beta, \sigma^2 I_n) \text{ mit } X := \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \in \mathbb{R}^{n \times 2}, \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \in \mathbb{R}^2, \sigma^2 > 0. \tag{16}$$

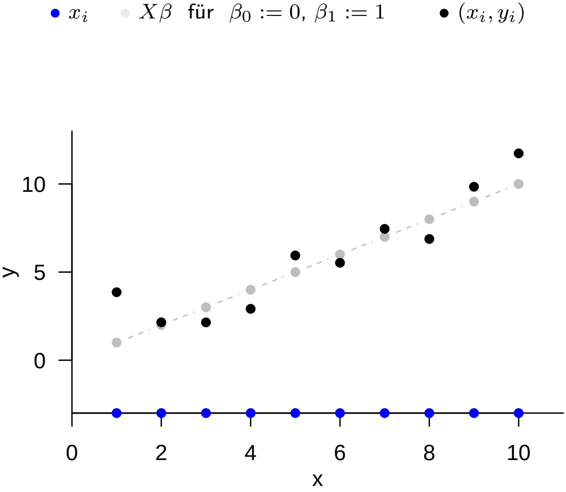
Einfache lineare Regression

```r
# Modellformulierung
library(MASS)                                # Multivariate Normalverteilung
n      = 10                                  # Anzahl von Datenpunkten
p      = 2                                   # Anzahl von Betaparametern
x      = 1:n                                 # Prädiktorwerte
X      = matrix(c(rep(1,n),x), nrow = n)     # Designmatrix
I_n    = diag(n)                             # n x n Einheitsmatrix
beta   = matrix(c(0,1), nrow = p)            # wahrer, aber unbekannter, Betaparameter
sigsqr = 1                                   # wahrer, aber unbekannter, Varianzparameter

# Datenrealisierung
y      = mvrnorm(1, X %*% beta, sigsqr*I_n)  # eine Realisierung eines n-dimensionalen ZVs
print(y)
```

```
>  [1]  1.36  2.47  2.09  4.54  4.95  5.48  5.14  8.51  7.37 12.07
```

Einfache lineare Regression



Allgemeine Theorie

Unabhängige und identisch normalverteilte Zufallsvariablen

Einfache lineare Regression

Selbstkontrollfragen

Selbstkontrollfragen

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 von n unabhängig und identisch normalverteilten Zufallsvariablen in ALM Form.
10. Schreiben Sie das Szenario der einfachen linearen Regression in ALM Form.