Allgemeines Lineares Modell

BSc Psychologie SoSe 2022

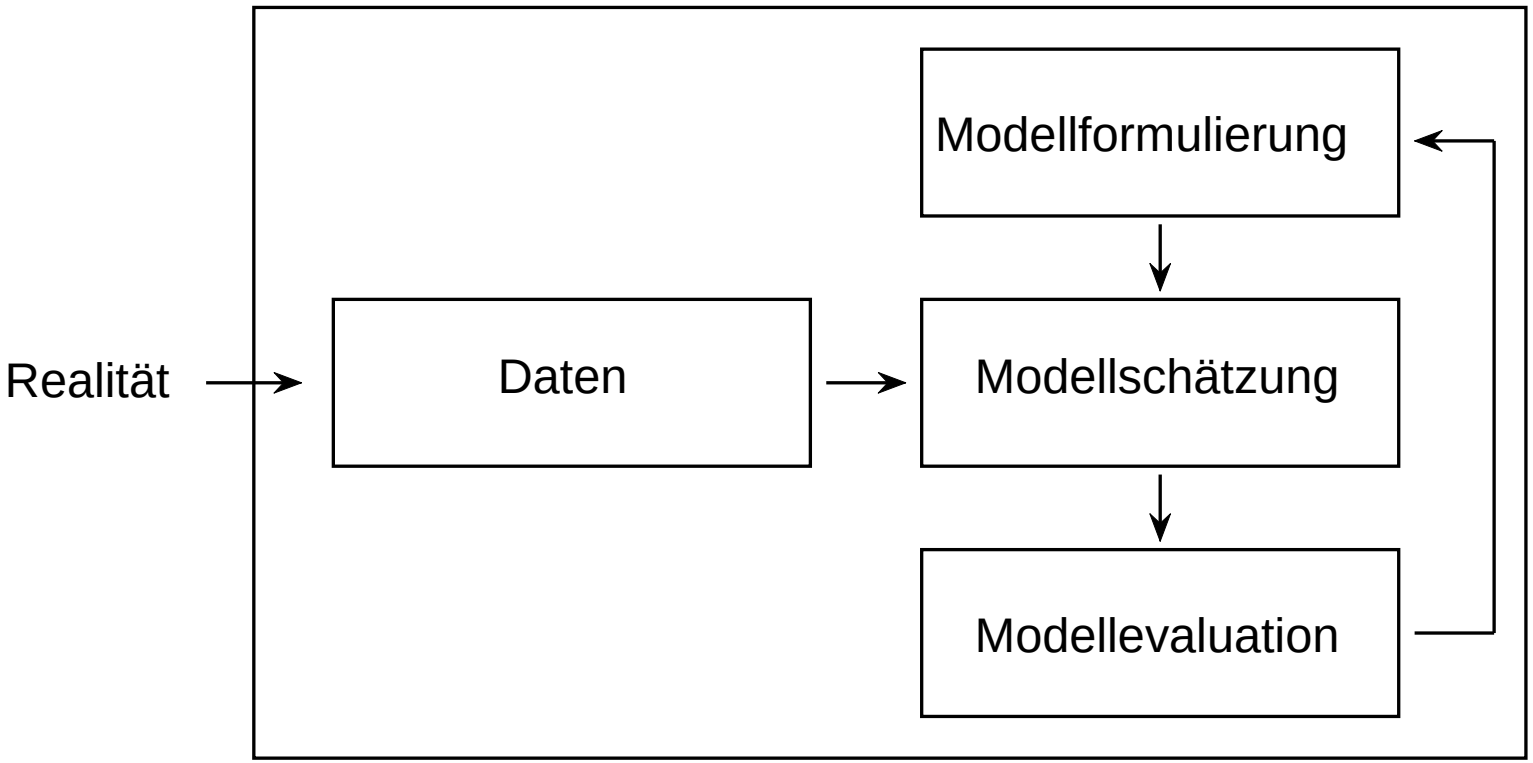
Prof. Dr. Dirk Ostwald

(5) Modellformulierung

Überblick

Datum	Einheit	Thema
08.04.2022	Grundlagen	(1) Regression
	Osterpause	
22.04.2022	Grundlagen	(2) Korrelation
29.04.2022	Grundlagen	(3) Matrizen
06.05.2022	Grundlagen	(4) Normalverteilungen
13.05.2022	Theorie	(5) Modellformulierung
20.05.2022	Theorie	(6) Modellschätzung
27.05.2022	Theorie	(7) Modellevaluation
03.06.2021	Anwendung	(8) Studiendesign
10.06.2021	Anwendung	(9) T-Tests
17.06.2021	Anwendung	(10) Einfaktorielle Varianzanalyse
24.06.2022	Anwendung	(11) Zweifaktorielle Varianzanalyse
01.07.2022	Anwendung	(12) Multiple Regression
08.07.2022	Anwendung	(13) Kovarianzanalyse
Juli 2022	Klausurtermin	
März 2023	Klausurwiederholungstermin	

Naturwissenschaft



Überblick

Modellformulierung

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

Modellschätzung

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

Modellevaluation

$$T = \frac{c^T \hat{\beta} - c^T \beta_0}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}}, F = \frac{(\hat{\varepsilon}_1^T \hat{\varepsilon}_1 - \hat{\varepsilon}^T \hat{\varepsilon})/p_2}{\hat{\varepsilon}^T \hat{\varepsilon}/(n - p)} \tag{3}$$

Standardprobleme Frequentistischer Inferenz

(1) Parameterschätzung

Ziel der Parameterschätzung ist es, einen möglichst guten Tipp für die wahren, aber unbekannten, Parameterwerte (oder eine Funktion derer) abzugeben, typischerweise basierend auf der Beobachtung einer Datenrealisierung.

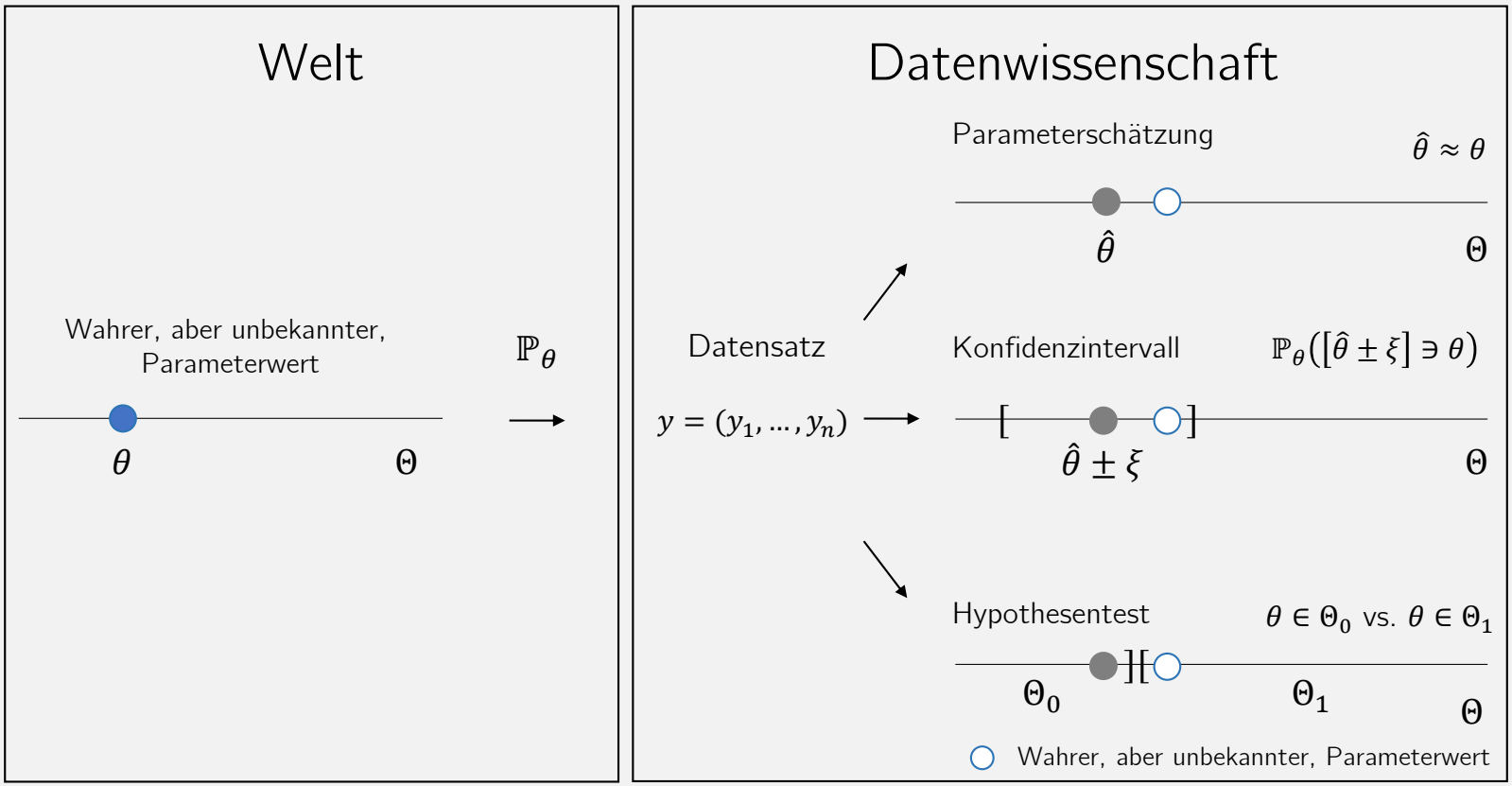
(2) Konfidenzintervalle

Das Ziel der Bestimmung von Konfidenzintervallen ist es, basierend auf der Verteilung möglicher Parameterschätzwerte eine quantitative Aussage über die mit dem Schätzwert assoziierte Unsicherheit zu treffen.

(3) Hypothesentests

Das Ziel der Auswertung von Hypothesentests ist es, basierend auf der angenommenen Verteilung der Daten in einer möglichst sinnvollen Form zu entscheiden, ob ein wahrer, aber unbekannter Parameterwert, sich in einer von zwei sich gegenseitig ausschließenden Untermengen des Parameterraumes, welche man als Hypothesen bezeichnet, liegt.

Modell und Standardprobleme der Frequentistischen Inferenz



Überblick

Standardannahmen Frequentistischer Inferenz

Gegeben sei ein statistisches Modell mit. 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

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

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

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

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

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

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 $\hat{\beta} := (X^T X)^{-1} X^T y$? 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 Einheit (5) - (7)

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

Anwendungsbeispiele Einheit (8) - (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

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

wobei

- y ein n-dimensionaler beobachtbarer Zufallsvektor ist, der *Daten* genannt wird,
- $X \in \mathbb{R}^{n \times p}$ eine vorgegebene Matrix ist, die *Designmatrix* genannt wird,
- $\beta \in \mathbb{R}^p$ ein unbekannter Parametervektor ist, der *Betaparametervektor* genannt wird und
- 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 wird (4) *Allgemeines Lineares Modell (ALM) in generativer Form* genannt.

Bemerkungen

- y 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 Betaparameter und 1 Varianzparameter).
- Der Betaparametervektor wird auch *Gewichtsvektor* oder *Effektvektor* genannt.
- Weil der Kovarianzmatrixparameter von ε als sphärisch angenommen wird, sind die $\varepsilon_1, \ldots, \varepsilon_n$ unabhängige normalverteilte Zufallsvariablen mit identischem Varianzparameter; weil zusätzlich der Erwartungswertparameter von ε als 0_n angenommen wird, sind die $\varepsilon_1, \ldots, \varepsilon_n$ auch identisch normalverteilte Zufallsvariablen.
- Für jede Komponente $y_i, i = 1, \ldots, n$ von y impliziert (4) nach Definition des Matrixprodukts, dass

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

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

Theorem (ALM Datenverteilung)

Es sei

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

das ALM in generativer Form. Dann gilt

$$y \sim N(X\beta, \sigma^2 I_n). \tag{8}$$

Beweis

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

$$y \sim N\left(I_n 0_n + X\beta, I_n(\sigma^2 I_n)I_n^T\right) = N(X\beta, \sigma^2 I_n). \tag{9}$$

Bemerkungen

- Im ALM sind die Daten y also ein n-dimensionaler normalverteilter Zufallsvektor mit Erwartungswertparameter $X\beta \in \mathbb{R}^n$ und Kovarianzmatrixparameter $\sigma^2 I_n \in \mathbb{R}^{n\times n}$.
- Die Komponenten $y_1, \ldots, y_n$ von y, also die Datenpunkte, sind damit unabhängige, aber im Allgemeinen nicht identisch verteilte, normalverteilte Zufallsvariablen der Form $y_i \sim N\left((X\beta)_i, \sigma^2\right)$ für $i = 1, \ldots, n$.

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,

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

Dann gilt, dass (10) äquivalent ist zu

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

In Matrixschreibweise ist dies wiederum äquivalent zu

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

Bemerkungen

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

$$X_1, \ldots, X_n \sim N(\mu, \sigma^2) \text{ mit } (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}. \tag{13}$$

- Bitte verwechseln Sie nicht die Designmatrix $X \in \mathbb{R}^{n \times p}$ des ALMs und die Zufallsvariablen $X_1, \ldots, X_n$ der Szenarien in Wahrscheinlichkeitstheorie und Frequentistische Inferenz.

Unabhängige und identisch normalverteilte Zufallsvariablen

```r
# Libraries
library(MASS)                                   # Multivariate Normalverteilung

# Modellformulierung
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 generative Modell der einfachen linearen Regression aus Einheit (1) Regression,

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

wobei wir hier nun die Zufallsvariablen $y_1, \ldots, y_n$ mit kleinen Buchstaben bezeichnen.

Wir haben bereits gesehen, dass dieses Modell äquivalent ist zu dem Normalverteilungsmodell der Regession

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

In Matrixschreibweise ist dies wiederrum äquivalent zu

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

Einfache lineare Regression

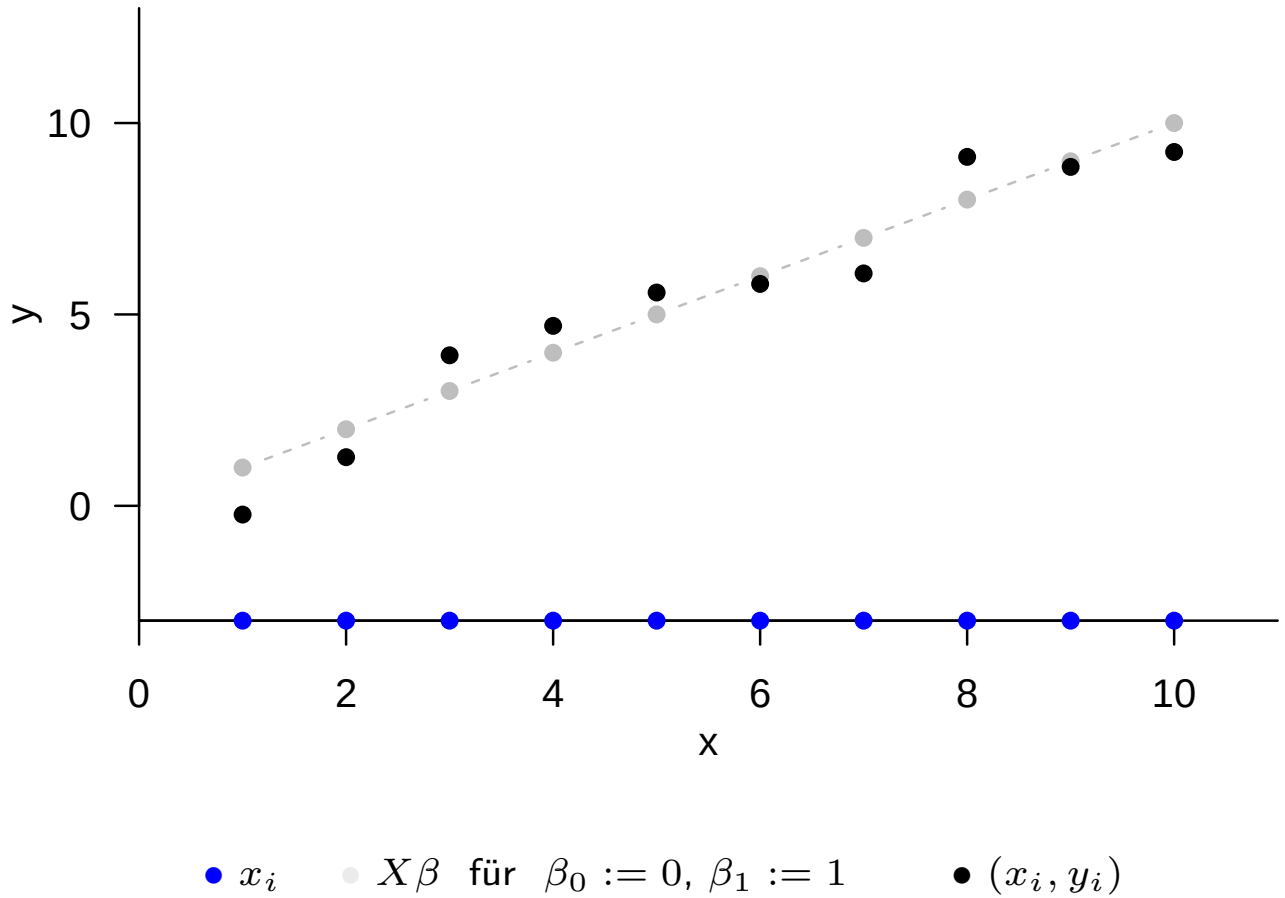
```r
# Libraries
library(MASS)                                   # Multivariate Normalverteilung

# Modellformulierung
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. Setzen Sie das naturwissenschaftliche Paradigma und die Frequentistische Inferenz in Beziehung.
4. Geben Sie die Definition des ALMs in generativer Form wieder.
5. Erläutern Sie die deterministischen und probabilistischen Aspekte des ALMs.
6. Wieviele Parameter hat das ALM mit sphärischer Kovarianzmatrix?
7. Warum sind die Komponenten des ALM Zufallsfehler unabhängig und identisch verteilt?
8. Geben Sie das Theorem zur ALM Datenverteilung wieder.
9. Sind die Komponenten des ALM Datenvektors unabhängig und identisch verteilt?
10. Schreiben Sie das Szenario n unabhängig und identisch verteilter Zufallsvariablen als ALM in Matrixschreibweise.
11. Schreiben Sie das Szenario der einfachen linearen Regression als ALM in Matrixschreibweise.
12. Generieren Sie 100 Datensätze von 12 unabhängig und identisch verteilten Zufallsvariablen.
13. Generieren Sie 100 Datensätze von eines einfachen linearen Regressionsmodels mit 12 äquidistanten Werten der unabhängigen Variable im Intervall $[1, 2]$, wobei $x_1 := 1$ und $x_{12} := 2$ sein sollen.