Tutorium Allgemeines Lineares Modell

BSc Psychologie SoSe 2022

7. Termin: Normalverteilungen

Belinda Fleischmann

Follow-up

Matrizen

- Eine quadratische Matrix $A \in \mathbb{R}^{n \times n}$ ist nur dann invertierbar, wenn gilt, dass $\det(A) \neq 0$.

Normalverteilungen

- Die WDF der Verteilung des Zufallsvektors y entspricht der WDF der gemeinsamen Verteilung der Zufallsvariablen $y_1, ..., y_n$, die aufgrund der Unabhängigkeit der Zufallsvariablen $y_1, ..., y_n$ wiederum dem Produkt der WDFen aller y_i entspricht.

$$p_y(v) = p_{y_1,...,y_n}(v_1, ..., v_n) = \prod_{i=1}^{n} p_{y_i}(v_i)$$

- Realisierungen aus multivariater Normalverteilung generieren (Daten simulieren) mit `MASS::mvrnorm`, eine Alternative für `mvtnorm::rmvnorm`

```
library(MASS)

# Parameterdefinition
mu = c(1,1) # \mu \in \mathbb{R}^2
Sigma = matrix(c(0.2, 0.15, 0.15, 0.2), 2) # \Sigma in \mathbb{R}^{2 \times 2}

Realisierungen = mvrnorm(n = 100, mu = mu, Sigma = Sigma)
```

Normalverteilungen (Fortführung)

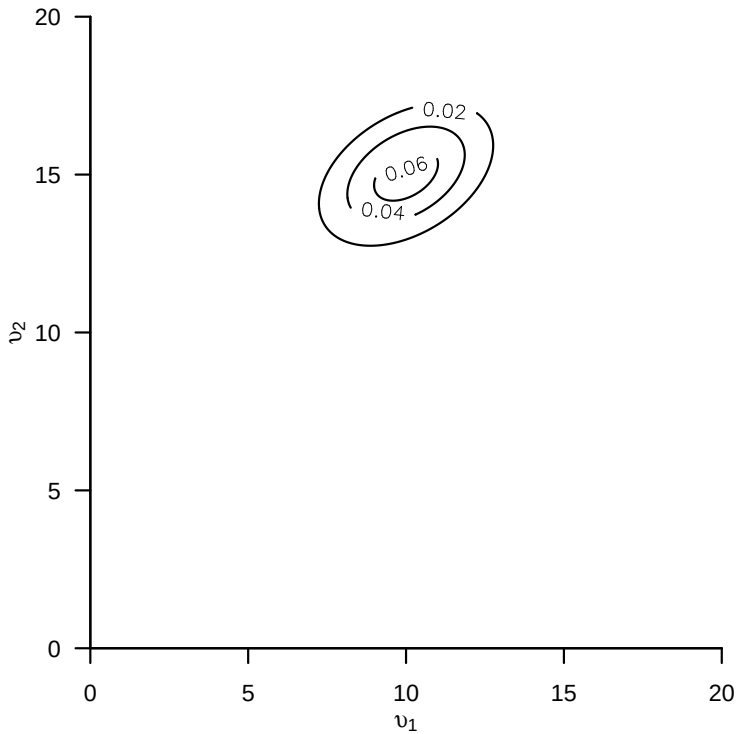
- Visualisierung der WDF (SKF Nr. 6)

```r
# R Paket für multivariate Normalverteilungen
library(mvtnorm)

# Parameterdefinition
mu     = c(10,15)                                  # Erwartungswertparameter
Sigma  = matrix(c(3,1,1,2), 2)                     # Kovarianzmatrixparameter

# Ergebnisraumdefintion
y_min  = 0                                         # y_i Minimum
y_max  = 20                                        # y_i Maximum
y_res  = 1e3                                       # y_i Auflösung (1e3 --> 1000 Werte)
y_1    = seq(y_nin, y_nax, length.out = y_res)     # y_1 Raum
y_2    = seq(y_nin, y_nax, length.out = y_res)     # y_2 Raum
y      = expand.grid(y_1,y_2)                      # y = (y_1,y_2)^T Raum

# Wahrscheinlichkeitsdichtefunktionauswertung
WDF = dmvnorm(as.matrix(y), mu, Sigma)             # Multivariate WDF (als Vektor)
p      = matrix(WDF, nrow = y_res)                 # Matrixkonversion der WDF
```

```r
# Visualisierung
contour(
  y_1,
  y_2,
  p,
  xlim      =  c(y_nin,y_nax),
  ylim      =  c(y_nin,y_nax),
  xlab      = TeX("$\\upsilon_1$"),
  ylab      = TeX("$\\upsilon_2$"),
  nlevels   = 5)
```

Normalverteilungen (Fortführung)

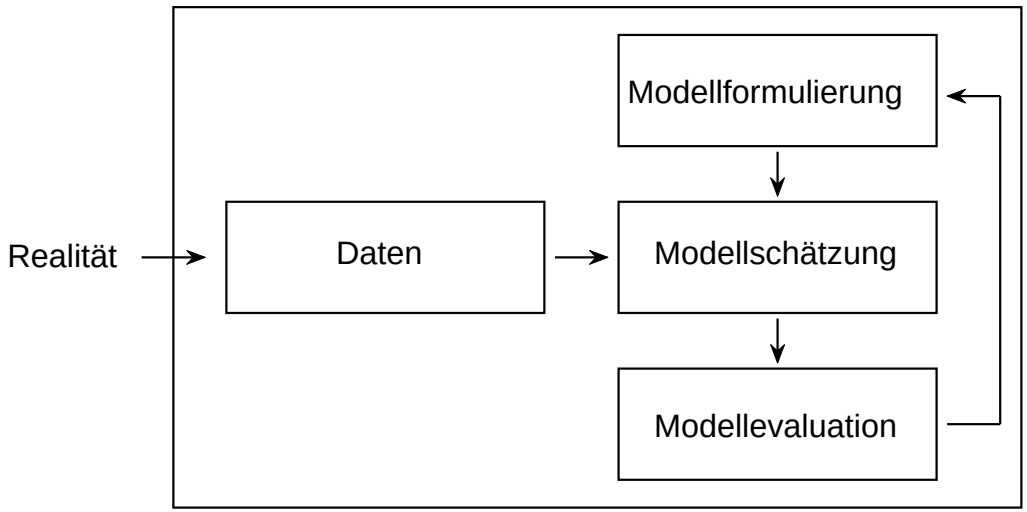
- Visualisierung der WDF (SKF Nr. 6)



Selbstkontrollfragen - Modellformulierung

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.

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 die wahren, aber unbekannten, Parameterwerte (oder eine Funktion derer) abzugeben, typischerweise basierend auf der Beobachtung einer Datenrealisierung.

(2) Konfidenzintervalle

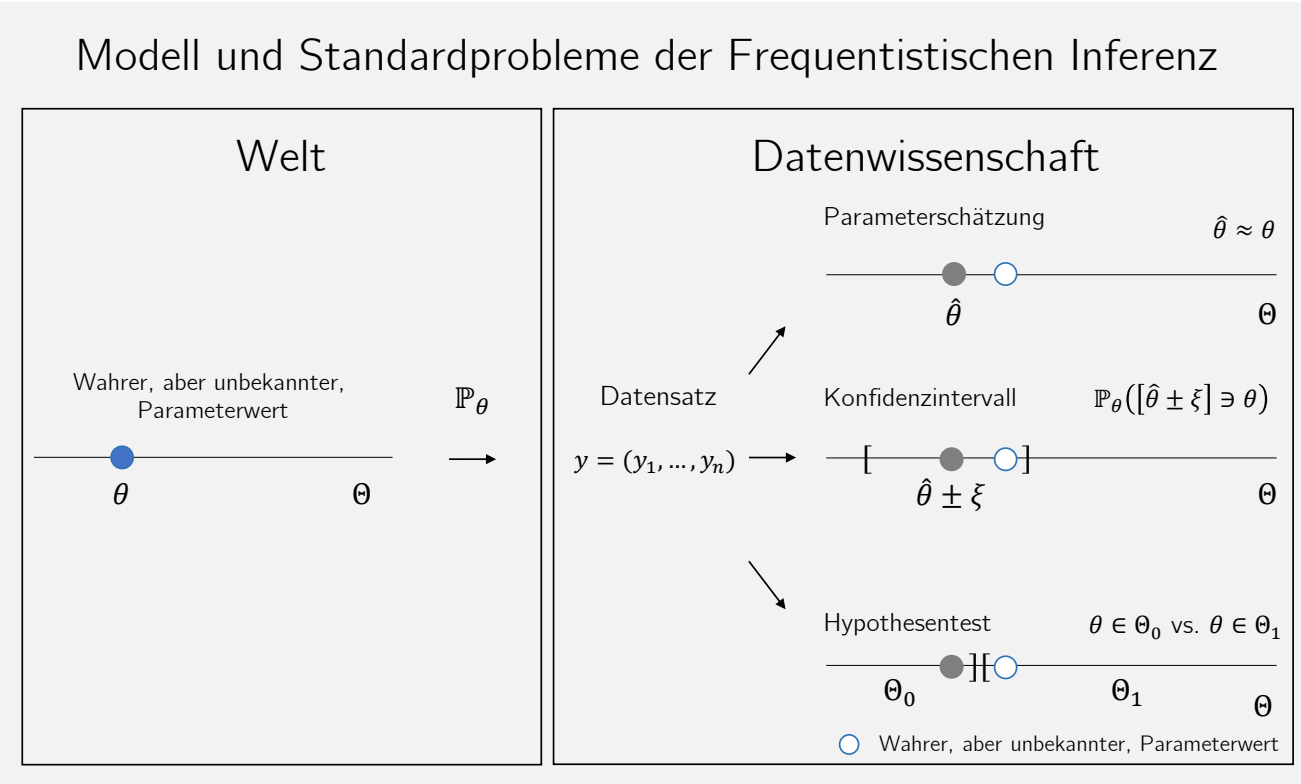
Das Ziel der Bestimmung von Konfidenzintervallen ist es, basierend auf der Verteilung möglicher Parameterschätzwerte eine quantitative Aussage über die mit dem Schätzwert assoziierte Unsicherheit zu treffen.

(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.

3. Setzen Sie das naturwissenschaftliche Paradigma und die Frequentistische Inferenz in Beziehung.

Zur Veranschaulichung



3. Setzen Sie das naturwissenschaftliche Paradigma und die Frequentistische Inferenz in Beziehung.

- Im Rahmen der Frquentistischen Inferenz gehen wir davon aus, dass es in der **Realität** (Welt) wahre, aber unbekannte Parameterwerte gibt (z.B. ein Wert, der den Zusammenhang zwischen zwei Variablen beschreibt).
- Eine **Modellformulierung** drückt eine Theorie über die Realität (Welt) formal aus. So ist beispielsweise $y = X\beta + \epsilon, \epsilon \sim N(0_n, \sigma^2 I_n)$ eine Modellformulierung, die einen linearer Zusammenhang zwischen x und y behauptet.
- Im Rahmen der **Modellschätzung** werden Parameterwerte basierend auf erhobenen **Daten** geschätzt.
- 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.
- Im Rahmen der **Modellevaluation** werden z.B. Konfidenzintervalle bestimmt und Hypothentests durchgeführt, um die Güte der Modelle zu bewerten.
- Um die Qualität statistischer Methoden zu beurteilen, betrachtet die frequentistische Statistik die Wahrscheinlichkeitsverteilungen von Schätzern und Statistiken unter Annahme der Datenverteilung.

4. Geben Sie die Definition des ALMs in generativer Form wieder.

Definition (Allgemeines Lineares Modell)

Es sei

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

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

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

Dann wird (1) *Allgemeines Lineares Modell (ALM) in generativer Form* genannt.

Beispiele ALM mit $n = 5$

y sei ein 5-dimensionalter Zufallsvektor mit Erwartungswertparameter $X\beta \in \mathbb{R}^{n \times p}$ und Kovarianzmatrixparameter $\sigma^2 I_n$. Die Komponenten $y_1, ..., y_n$ sind unabhängig aber nicht identisch verteilte Zufallsvariablen der Form $y_i \sim N(\mu_i, \sigma^2)$ für $i = 1, ..., n$.

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

(im Tutorium)

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

(im Tutorium)

5. 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 Normalverteilungne gezogen werden. Der probabilistische Aspekt modelliert bei Normalverteilungen alle Einflussfaktoren auf y, die nicht durch den deterministischen Askept abgedeckt werden.
- y das Ergebnis der Addition deterministischer und probabilistischer Aspekte ist, ist es auch probabilistisch (i.e. zufällig).

6. Wieviele 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)

7. Warum sind die Komponenten des ALM Zufallsfehler 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 $\epsilon \sim N(0_n, \sigma^2 I_n)$, wobei der Erwartungswertparameter 0_n bedeutet, dass alle Komponenten $\epsilon_1, \ldots, \epsilon_n$ 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.

- $\Rightarrow$ identische verteilte Komponenten, weil alle Komponenten den Erwartungswert 0 und Varianzparameter σ^2 haben.
- $\Rightarrow$ unabhängige Komponenten, weil alle nicht-digonal-Elemente, also alle Kovarianzen zwischen Komponenten gleich 0 (vgl. Tutorium (4) Normalverteilungen - SKF 10)

8. Geben Sie das Theorem zur ALM Datenverteilung wieder.

Theorem (ALM Datenverteilung)

Es sei

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

das ALM in generativer Form. Dann gilt

$$y \sim N(X\beta, \sigma^2 I_n). \tag{4}$$

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

$y \sim N(X\beta, \sigma^2 I_n)$ mit $y_i \sim N(\mu_i, \sigma^2)$ 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 $y_1, ..., y_n$
- 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 y_i einen anderen Erwartungswert μ_i annehmen kann. $\Rightarrow$ **nicht** identisch verteilte Komponenten y_i.

10. Schreiben Sie das Szenario n unabhängig und identisch verteilter Zufallsvariablen als ALM in Matrixschreibweise.

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

Anmerkung: Während wir im Theorem zur Datenverteilung im ALM noch gesehen haben, dass die Komponenten $y_1, ..., y_n$ jeweils "individuelle" Verteilungen $y_i \sim N(\mu_i, \sigma^2)$ 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 $y_1, ..., y_n$, also für alle y_i gleich ist.

In Matrixschreibweise:

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

Beispiel $n = 5$ unabhängig und identisch normalverteilte Zufallsvariable

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

In Matrixschreibweise:

(im Tutorium)

11. Schreiben Sie das Szenario der einfachen linearen Regression als ALM in Matrixschreibweise.

(im Tutorium)

Beispiel mit $n = 5$

(im Tutorium)

12. Generieren Sie 100 Datensätze von 12 unabhängig und identisch verteilten Zufallsvariablen.

```
library(MASS)                                  # package für Funktion MASS::mvrnorm laden
options(width = 300)                           # Ausgabe layout anpassen

# Modellformulierung
n      = 12                                    # Anzahl von Datenpunkten
p      = 1                                     # Anzahl von Betaparametern
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(100, X %*% beta, sigsqr*I_n) # 100 Realisierung eines n-dimensionalen ZVs
print(y[1:10,])                                # Ausgabe der ersten 10 Datensätze
```

```
>        [,1] [,2]   [,3]  [,4] [,5]  [,6] [,7] [,8]  [,9] [,10] [,11]   [,12]
>  [1,] 0.850 1.09 0.2944 1.663 3.38 3.535 2.89 2.24 1.288 3.736 1.644  1.5706
>  [2,] 1.728 5.11 1.1446 1.784 3.68 1.584 1.62 2.63 2.662 1.155 0.936  3.3605
>  [3,] 2.457 0.93 1.8551 2.621 3.18 1.479 2.61 2.42 2.291 1.038 3.077  1.9291
>  [4,] 1.983 1.30 1.6756 0.716 1.85 2.851 2.00 3.98 2.198 3.017 3.182  1.7278
>  [5,] 1.458 1.80 1.8274 0.700 1.82 2.334 1.48 1.49 0.796 0.504 2.198 -0.4467
>  [6,] 2.876 2.74 0.7639 1.623 2.80 1.171 1.36 0.89 1.960 0.815 1.600  2.0655
>  [7,] 2.741 1.95 0.0977 2.104 1.58 1.781 1.36 1.05 2.687 2.630 2.616  0.9015
>  [8,] 1.835 3.31 1.9055 1.296 3.21 0.455 2.11 2.48 2.705 4.101 3.974  1.3668
>  [9,] 1.249 2.09 2.0326 3.497 3.24 2.233 3.18 1.20 2.991 1.386 3.885 -0.0637
> [10,] 0.747 1.37 2.4613 1.697 2.69 2.031 2.45 2.23 3.144 0.365 0.411  4.6489
```

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.

Teil 1/2 - Ausformulierung

(im Tutorium)

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.

Teil 2/2 - in R

```
library(MASS)                                  # package für Funktion MASS::mvrnorm laden
options(width = 300)                           # Ausgabe layout anpassen

# Modellformulierung
n      = 12                                    # Anzahl von Datenpunkten
p      = 2                                     # Anzahl von Betaparametern
x      = seq(1,2,1/11)                         # 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)              # wahre, aber unbekannte, Betaparameter
sigsqr = 1                                     # wahrer, aber unbekannter, Varianzparameter

# Datenrealisierung
y      = mvrnorm(100, X %*% beta, sigsqr*I_n) # 100 Realisierungen des n-dimensionalen ZVs
print(y[1:10,])                                # Ausgabe der ersten 10 Datensätze
```

```
>          [,1]   [,2]  [,3]    [,4]  [,5]  [,6]    [,7]  [,8] [,9] [,10]   [,11] [,12]
>  [1,]  1.386  2.235 1.072  1.8511 2.206 1.288  0.8333 2.637 2.21  1.11  1.3306 0.863
>  [2,]  0.161 -0.291 0.130  1.3859 2.628 1.575  0.1828 2.164 3.40  3.29  1.7400 1.720
>  [3,]  0.470  0.247 0.290  1.7129 0.653 0.793  1.0448 2.800 1.69  2.66 -0.0101 1.106
>  [4,] -1.293 -0.746 0.563 -0.7296 1.010 0.924  2.9839 1.810 1.29  3.11  0.3748 2.137
>  [5,]  2.855  2.135 1.437  2.3144 0.280 1.153  1.0338 2.214 2.46  2.12  0.7943 1.251
>  [6,]  0.375  1.306 0.743  0.2590 2.330 0.852  1.5435 1.520 1.15  1.41  3.5069 2.518
>  [7,]  1.176  1.891 2.650  0.7679 0.883 1.136  3.3286 2.541 1.80  1.68  1.2693 1.808
>  [8,] -0.366  1.806 2.014  1.3627 0.323 1.763  2.5768 1.620 1.79  1.60  3.4758 2.029
>  [9,] -0.346  0.774 0.268 -0.0193 1.563 2.253  0.9647 0.714 2.58  3.57  0.4599 2.359
> [10,]  1.947 -1.161 0.909 -0.4578 3.077 3.205 -0.0476 0.576 4.30  1.74  1.1176 1.971
```