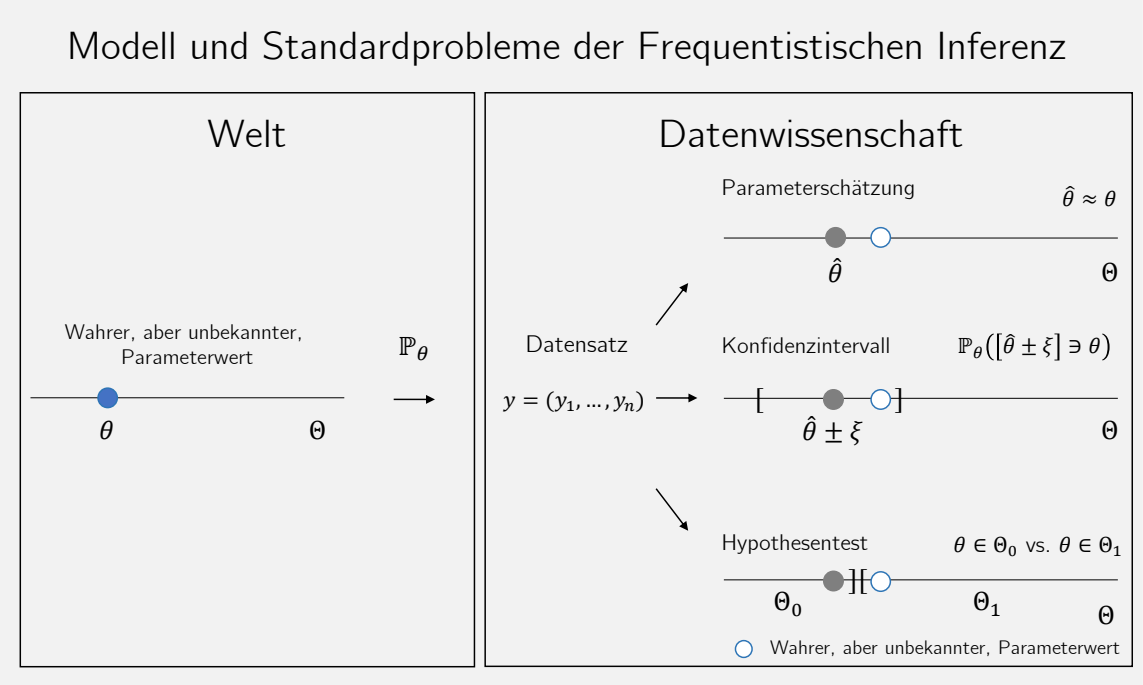
Tutorium Allgemeines Lineares Modell

BSc Psychologie SoSe 2022

8. Termin: (6) Modellschätzungen

Belinda Fleischmann

Wiederholung - Frequentistisches Weltbild



- Wir nehmen an, dass wahre, aber unbekannte Parameter existieren.
- Wir nehmen weiterhin an, dass probabilistische Prozesse existieren, die - gegeben dieser wahren, aber unbekannten Parameter - Datensätze generieren können.
- Für diese probabilistischen Prozesse nehmen wir an, dass ihnen bestimmte Verteilungen bzw. Wahrscheinlichkeitsdichten zugrundeliegen (z.B. Normalverteilung)
- Wir verwenden erhobene Daten dafür, Parameterwerte zu schätzen.
- Dabei bilden die angenommenen Verteilungen bzw. Warhscheinlichkeitsdichten der probabilistischen Prozesse (die, wie wir annehmen, die Daten generiert haben), die Grundlage für Parameterschätzung und Modellevaluation.

Selbstkontrollfragen - Modellschätzungen

1. Geben Sie das Betaparameterschätzer Theorem wieder.
2. Geben Sie das Varianzparameterschätzer Theorem wieder.
3. Geben Sie die Beta- und Varianzparameterschätzer des ALM Szenarios von n unabhängigen und identisch normalverteilten Zufallsvariablen wieder.
4. Geben Sie die Beta- und Varianzparameterschätzer des ALM Szenarios der einfachen linearen Regression wieder.
5. Wie unterscheiden sich die Betaparameterschätzer des ALM Szenarios der einfachen linearen Regression und die Parameter der Ausgleichsgerade aus Einheit (1) Regression?
6. Simulieren Sie die Unverzerrtheit des Varianzparameterschätzers im ALM Szenario von n unabhängigen und identisch normalverteilten Zufallsvariablen in einem R Skript.
7. Simulieren Sie die Unverzerrtheit des Varianzparameterschätzers im ALM Szenario der einfachen linearen Regression in einem R Skript.

1. Geben Sie das Betaparameterschätzer Theorem wieder.

Theorem (Betaparameterschätzer)

Es sei

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

das ALM in generativer Form und es sei

$$\hat{\beta} := (X^T X)^{-1} X^T y. \tag{2}$$

der *Betaparameterschätzer*. Dann gilt für festes $y \in \mathbb{R}^n$, dass

$$\hat{\beta} = \arg\min_{\tilde{\beta}} (y - X\tilde{\beta})^T (y - X\tilde{\beta}) \tag{3}$$

und dass $\hat{\beta}$ ein unverzerrter Maximum Likelihood Schätzer von $\beta \in \mathbb{R}^p$ ist.

Beispiel für $\hat{\beta}$ bei ALM mit $n = 5$ und $p = 2$

(vgl. Beispiel (2) für ALM aus Tut. 5, nach Aufg. 4)

Wir betrachten das ALM

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

$$y = X\beta + \varepsilon \Leftrightarrow \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \\ x_{51} & x_{52} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix} = \begin{pmatrix} x_{11}\beta_1 + x_{12}\beta_2 + \varepsilon_1 \\ x_{21}\beta_1 + x_{22}\beta_2 + \varepsilon_2 \\ x_{31}\beta_1 + x_{32}\beta_2 + \varepsilon_3 \\ x_{41}\beta_1 + x_{42}\beta_2 + \varepsilon_4 \\ x_{51}\beta_1 + x_{52}\beta_2 + \varepsilon_5 \end{pmatrix}$$

Dann sieht die Betaparameterschätzerformel ausgeschrieben wie folgt aus

$$\hat{\beta} := (X^T X)^{-1} X^T y = \left(\begin{pmatrix} x_{11} & x_{21} & x_{31} & x_{41} & x_{51} \\ x_{12} & x_{22} & x_{32} & x_{42} & x_{52} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \\ x_{51} & x_{52} \end{pmatrix} \right)^{-1}$$
$$\times \begin{pmatrix} x_{11} & x_{21} & x_{31} & x_{41} & x_{51} \\ x_{12} & x_{22} & x_{32} & x_{42} & x_{52} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix}$$

Beispiel für $\hat{\beta}$ bei ALM mit $n = 5$ und $p = 2$ (fortgeführt)

$$= \begin{pmatrix} x_{11}^2 + x_{21}^2 + x_{31}^2 + x_{41}^2 + x_{51}^2 & x_{11}x_{12} + x_{21}x_{22} + x_{31}x_{32} + x_{41}x_{42} + x_{51}x_{52} \\ x_{12}x_{11} + x_{22}x_{21} + x_{32}x_{31} + x_{42}x_{41} + x_{52}x_{51} & x_{12}^2 + x_{22}^2 + x_{32}^2 + x_{42}^2 + x_{52}^2 \end{pmatrix}^{-1}$$

$$\times \begin{pmatrix} x_{11} & x_{21} & x_{31} & x_{41} & x_{51} \\ x_{12} & x_{22} & x_{32} & x_{42} & x_{52} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix}$$

$$= \quad \dots$$

$$= \quad \begin{pmatrix} \hat{\beta_1} \\ \hat{\beta_2} \end{pmatrix} = \hat{\beta} \in \mathbb{R}^p$$

Wenn wir uns die Dimensionen der einzelnen Terme anschauen, wird klar, dass das Ergebnis am Ende 2×1-dimensional ist. Im Detail:

- Für den ersten Term $(X^T X)$ halten wir fest, dass die Inverse A^{-1} einer Matrix $A \in \mathbb{R}^{n \times n}$ die gleiche Größe hat wie die Matrix, also $A^{-1} \in \mathbb{R}^{n \times n}$. Entsprechend hat das Ergebnis des ersten Terms 2 Zeilen und 2 Spalten, i.e. (2×2)
- Für den zweiten Term $(X^T y)$ gilt $(2 \times 5)(5 \times 1) = (2 \times 1)$.
- Somit ergibt sich insgesamt für $\hat{\beta}$ die Dimension $(2 \times 2)(2 \times 1) = (2 \times 1)$.

2. Geben Sie das Varianzparameterschätzer Theorem wieder.

Theorem (Varianzparameterschätzer)

Es sei

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \tag{4}$$

das ALM in generativer Form. Dann ist

$$\hat{\sigma}^2 := \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n - p} \tag{5}$$

ein unverzerrter Schätzer von $\sigma^2 > 0$.

3. Geben Sie die Beta- und Varianzparameterschätzer des ALM Szenarios von n unabhängigen und identisch normalverteilten Zufallsvariablen wieder.

Wir betrachten das Szenario von n Unabhängige und identisch normalverteilte Zufallsvariablen mit Erwartungswertparameter $\mu \in \mathbb{R}$ und Varianzparameter σ^2. D.h., für jede Komponente des Datenvektors gilt $y_i \sim N(\mu, \sigma^2)$ für $i = 1, \ldots, n$. Äquivalent dazu können wir das generative Modell schreiben als

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

In Matrixschreibweise:

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

$$y = X\beta + \varepsilon = X\mu + \varepsilon \Leftrightarrow \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} \mu + \varepsilon_1 \\ \vdots \\ \mu + \varepsilon_n \end{pmatrix}$$

Dann gilt für die Beta- und Varianzparameterschätzer

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} y_i =: \bar{y} \text{ und } \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 =: s_y^2.$$

In diesem Szenario ist der Betaparameterschätzer mit dem Stichprobenmittel $\bar{y}$ der $y_1, \ldots, y_n$ und der Varianzparameterschätzer mit der Stichprobenvarianz s_y^2 der $y_1, \ldots, y_n$ identisch.

3. Geben Sie die Beta- und Varianzparameterschätzer des ALM Szenarios von n unabhängigen und identisch normalverteilten Zufallsvariablen wieder.

Anmerkungen:

- $\hat{\beta} := (X^T X)^{-1} X^T y$ ist wie wir im Theorem (Betaparameterschätzer) gelernt haben, ein erwartungstreuer Schätzer für $\beta \in \mathbb{R}^p$ des ALM.
- Weiterhin haben wir gelernt, dass im Szenario von n u.i.(normal-)v. Zufallsvariablen gilt, dass $X := I_n \in \mathbb{R}^n$, und somit $\hat{\beta} =: \bar{y}$ (siehe Herleitung auf VO-Folie 18).
- Entsprechend ist $\bar{y}$ ein erwartungstreuer Schätzer für β im ALM Szenario von n unabh. u.i.(normal-)v. Zufallsvariablen.
- Analog ist s_y^2 ein erwartungstreuer Schätzer für σ^2 im ALM Szenario von n u.i.(normal-)v. Zufallsvariablen.

4. Geben Sie die Beta- und Varianzparameterschätzer des ALM Szenarios der einfachen linearen Regression wieder.

Wir betrachten das generative Modell der einfachen linearen Regression

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

In Matrixschreibweise

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

$$y = X\beta + \varepsilon \Leftrightarrow \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} \beta_0 + x_1\beta_1 + \varepsilon_1 \\ \vdots \\ \beta_0 + x_n\beta_1 + \varepsilon_n \end{pmatrix}$$

Dann gilt für die Beta- und Varianzparameterschätzer

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \bar{y} - \frac{c_{xy}}{s_x^2}\bar{x} \\ \frac{c_{xy}}{s_x^2} \end{pmatrix} \text{ and } \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2,$$

wobei $\bar{x}$ und $\bar{y}$ die Stichprobenmittel der $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$, respektive, bezeichnen, c_{xy} die Stichprobenkovarianz der $x_1, \ldots, x_n$ und s_x^2 die Stichprobenvarianz der $x_1, \ldots, x_n$ bezeichnen.

5. Wie unterscheiden sich die Betaparameterschätzer des ALM Szenarios der einfachen linearen Regression und die Parameter der Ausgleichsgerade aus Einheit (1) Regression?

- Sowohl die Betaparameter im ALM Szenario der einfachen linearen Regression, als auch die Parameter der Ausgleichsgerade aus Einheit (1) minimieren die Summe der quadrierten Abweichungen
- Der Unterschied beim Bestimmen der Parameter zwischen zwischen Ausgleichsgerade und ALM Szenario der einfachen linearen Regression ist, dass bei der Ausgleichsgerade keine Annahmen über die Verteilung der Zufallsfehler getroffen werden, im ALM schon.

6. Simulieren Sie die Unverzerrtheit des Varianzparameterschätzers im ALM Szenario von n unabhängigen und identisch normalverteilten Zufallsvariablen in einem R Skript.

Wdhl.: Simulation der Schätzerunverzerrtheit von $\hat{\beta}$

```
library(MASS)
# Modellformulierung
n        = 12                                    # Anzahl von Datenpunkten
p        = 1                                     # Anzahl von Betparametern
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

# Frequentistische Simulation
n_sims     = 100                                 # Anzahl Realisierungen des n-dimensionalen ZVs (Anzahl Datensätze)
beta_hats = rep(NaN,n_sims)                      # array für beta_hats nach allg. ALM Schätzerformel
beta_hats_uiv = rep(NaN, n_sims)                 # array für beta_hats nach Schätzerformel für Szenario uiv. ZVen
for(i in 1:n_sims){
  y               = mvrnorm(1, X %*% beta, sigsqr*I_n) # eine Realisierung eines n-dimensionalen ZVs (1Datensatz)
  beta_hats[i]    = solve(t(X) %*% X) %*% t(X) %*% y   # \hat{\beta} = (X^T)X^{-1}X^Ty
  beta_hats_uiv[i] = mean(y)}                          # \hat{\beta} = \bar{\beta}

# Ausgabe
cat("Wahrer, aber unbekannter, Betaparameter                                    : ", beta,
    "\nGeschätzter Erwartungswert des Betaparameterschätzers nach allg. ALM Formel   : ", mean(beta_hats),
    "\nGeschätzter Erwartungswert des Betaparameterschätzers nach Formel für uiv. ZV : ", mean(beta_hats_uiv))
```

```
> Wahrer, aber unbekannter, Betaparameter                                    :  2
> Geschätzter Erwartungswert des Betaparameterschätzers nach allg. ALM Formel   :  1.97
> Geschätzter Erwartungswert des Betaparameterschätzers nach Formel für uiv. ZV :  1.97
```

6. Simulieren Sie die Unverzerrtheit des Varianzparameterschätzers im ALM Szenario von n unabhängigen und identisch normalverteilten Zufallsvariablen in einem R Skript.

Simulation der Schätzerunverzerrtheit von $\hat{\sigma}^2$

```
library(MASS)
# 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

# Frequentistische Simulation
n_sims     = 100                               # Anzahl Realisierungen des n-dimensionalen ZVs (Anzahl Datensätze)
beta_hats = rep(NaN,n_sims)                    # array für beta_hats
sigsqu_hats = rep(NaN,n_sims)                  # array für sigma^2_hats nach allg. ALM-Schätzformel
sigsqu_hats_uiv = rep(NaN, n_sims)             # array für sigma^2_hats nach Schätzerformel für Szenario uiv. ZVen
for(i in 1:n_sims){
  y           = mvrnorm(1, X %*% beta, sigsqr*I_n)   # eine Realisierung eines n-dimensionalen ZVs (1 Datensatz)
  beta_hats[i] = solve(t(X) %*% X) %*% t(X) %*% y    # \hat{\beta} = (X^T)X^{-1}X^Ty
  sigsqu_hats[i] =  (t(y - X %*% beta_hats[i]) %*%   # \frac{(y - X\hat{\beta})^T(y - X\hat{\beta})}{n - p}
                     (y - X %*% beta_hats[i]))/(n-p)
  sigsqu_hats_uiv[i] = var(y)}                       # \hat{\simga}^2 = s^2_y

# Ausgabe
cat("Wahrer, aber unbekannter, Varianzparameter                                    : ", sigsqr,
    "\nGeschätzter Erwartungswert des Varianzparameterschätzers nach allg. ALM Formel   : ", mean(sigsqu_hats),
    "\nGeschätzter Erwartungswert des Varianzparameterschätzers nach Formel für uiv. ZV : ", mean(sigsqu_hats_uiv))
```

```
> Wahrer, aber unbekannter, Varianzparameter                                    :  1
> Geschätzter Erwartungswert des Varianzparameterschätzers nach allg. ALM Formel   :  1.08
> Geschätzter Erwartungswert des Varianzparameterschätzers nach Formel für uiv. ZV :  1.08
```

6. Simulieren Sie die Unverzerrtheit des Varianzparameterschätzers im ALM Szenario von n unabhängigen und identisch normalverteilten Zufallsvariablen in einem R Skript.

Achtung:

- Das Bestimmen des Betaparameterschätzers $\hat{\beta}$ mit `mean(y)` und des Varianzparameterschätzers $\hat{\sigma}^2$ mit `var(y)` wird hier nur angeführt, um zu zeigen, dass es in diesem Szenario das gleiche ergibt wie über die ALM-Schätzerformeln $\hat{\beta} = (X^T)X^{-1}X^T y$ und $\hat{\sigma}^2 = \frac{(y-X\hat{\beta})^T(y-X\hat{\beta})}{n-p}$.
- Wenn wir multidimensionale Datensätze und Modelle haben, wird das nicht mehr funktionieren. Daher empfiehlt es sich, immer die allgemeinen ALM-Formeln für Parameterschätzungen zu verwenden.

7. Simulieren Sie die Unverzerrtheit des Varianzparameterschätzers im ALM Szenario der einfachen linearen Regression in einem R Skript.

Wdhl.: Simulation der Schätzerunverzerrtheit von $\hat{\beta}$

```
library(MASS)

# Modellformulierung
n         = 12                                  # 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)            # wahre, aber unbekannte, Betaparameter
sigsqr    = 1                                   # wahrer, aber unbekannter, Varianzparameter

# Frequentistische Simulation
n_sims     = 1e3                                # Anzahl Realisierungen des n-dimensionalen ZVs (Anzahl Datensätze)
beta_hats = matrix(rep(NaN,p*n_sims), nrow = p) # array für beta_hats nach allg. ALM Schätzerformel
for(i in 1:n_sims){
  y               = mvrnorm(1, X %*% beta, sigsqr*I_n) # eine Realisierung eines n-dimensionalen ZVs (1 Datensatz)
  beta_hats[,i]    = solve(t(X) %*% X) %*% t(X) %*% y  # \hat{\beta} = (X^T)X^{-1}X^Ty
}

# Ausgabe
cat("Wahrer, aber unbekannter, Betaparameter                                    : ", beta,
    "\nGeschätzter Erwartungswert des Betaparameterschätzers nach allg. ALM Formel  : ", rowMeans(beta_hats))
```

```
> Wahrer, aber unbekannter, Betaparameter                                    :  0 1
> Geschätzter Erwartungswert des Betaparameterschätzers nach allg. ALM Formel  :  0.00666 0.998
```

7. Simulieren Sie die Unverzerrtheit des Varianzparameterschätzers im ALM Szenario der einfachen linearen Regression in einem R Skript.

Simulation der Schätzerunverzerrtheit von $\hat{\sigma}^2$

```
library(MASS)
# Modellformulierung
n         = 12                                  # 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)            # wahre, aber unbekannte, Betaparameter
sigsqr    = 1                                   # wahrer, aber unbekannter, Varianzparameter

# Frequentistische Simulation
n_sims      = 1e3                               # Anzahl Realisierungen des n-dimensionalen ZVs (Anzahl Datensätze)
beta_hats = matrix(rep(NaN,p*n_sims), nrow = p) # array für beta_hats nach allg. ALM Schätzerformel
sigsqu_hats = rep(NaN,n_sims)                   # array für sigma^2_hats nach allg. ALM-Schätzformel
for(i in 1:n_sims){
  y               = mvrnorm(1, X %*% beta, sigsqr*I_n) # eine Realisierung eines n-dimensionalen ZVs (1 Datensatz)
  beta_hats[,i]    = solve(t(X) %*% X) %*% t(X) %*% y  # \hat{\beta} = (X^T)X^{-1}X^Ty
  sigsqu_hats[i] =  (t(y - X %*% beta_hats[,i]) %*%     # \frac{(y - X\hat{\beta})^T(y - X\hat{\beta})}{n - p}
                     (y - X %*% beta_hats[,i]))/(n-p)}

# Ausgabe
cat("Wahrer, aber unbekannter, Betaparameter                                    : ", sigsqr,
    "\nGeschätzter Erwartungswert des Betaparameterschätzers nach allg. ALM Formel  : ", mean(sigsqu_hats))
```

```
> Wahrer, aber unbekannter, Betaparameter                                    :  1
> Geschätzter Erwartungswert des Betaparameterschätzers nach allg. ALM Formel  :  1.01
```