Allgemeines Lineares Modell

BSc Psychologie, SoSe 2024

Joram Soch

(7) T-Statistiken

Modellformulierung

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

Modellschätzung

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

Modellevaluation

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

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.

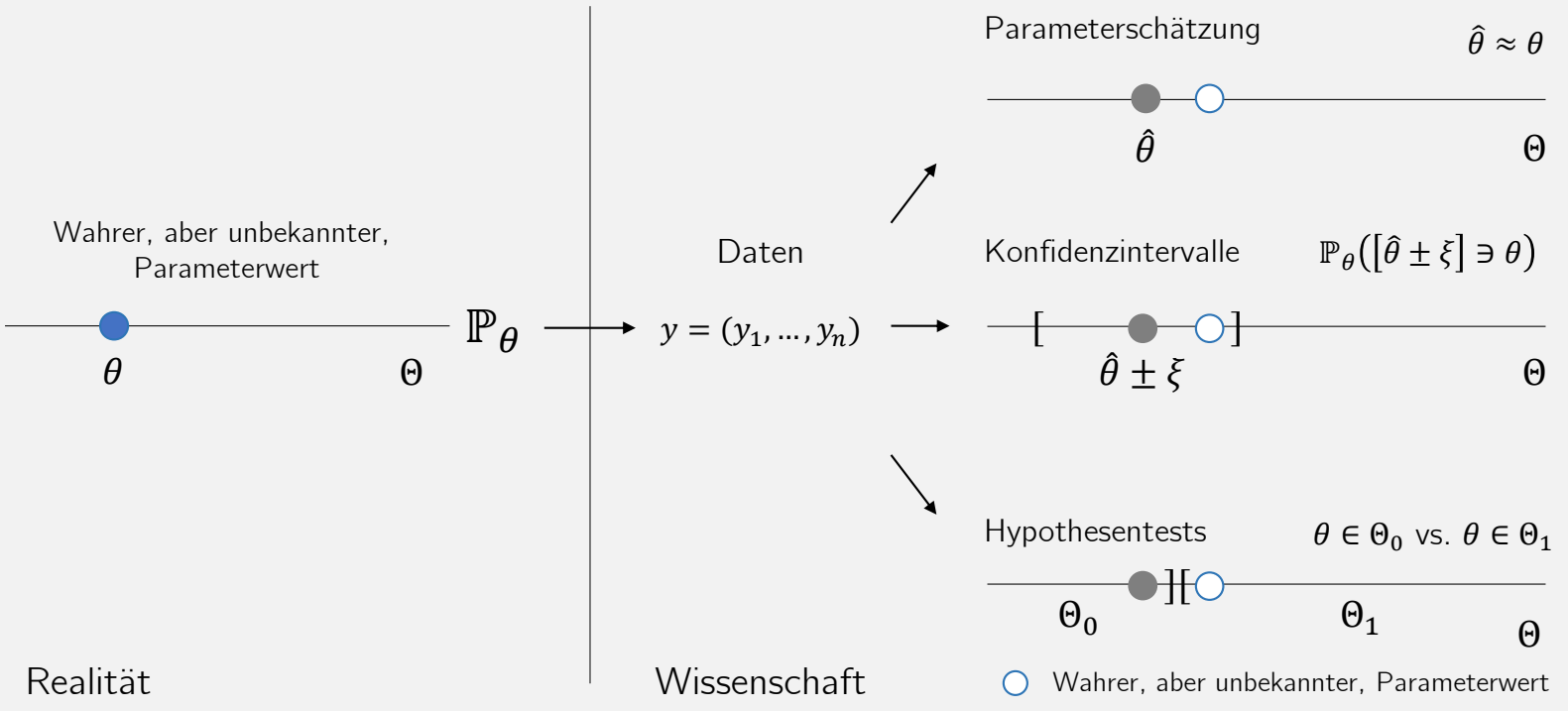
(2) Konfidenzintervalle

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



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

Überblick

- In dieser Einheit führen wir T-Statistiken als Maße zur Evaluation von Betaparameterschätzern im ALM ein. T-Statistiken quantifizieren dabei die geschätzten Effekte des Betaparameterschätzers in Bezug zur durch den Varianzparameterschätzer geschätzten Residualvariabilität. Der Wert einer T-Statistik ist also zunächst einmal einfach als Signal-zu-Rauschen-Verhältnis (*signal-to-noise ratio*) zu verstehen.
- T-Statistiken erlauben weiterhin die Evaluation von Linearkombinationen der Komponenten des Betaparameterschätzers im Sinne Frequentistischer Konfidenzinteralle und Hypothesentests. Wir betrachten hier zunächst nur die funktionale Form von T-Statistiken und ihre Frequentistische Verteilung zum Zwecke der Konfidenzintervallbestimmung. Der Einsatz von T-Teststatistiken zum Zwecke von Einstichproben- und Zweistichproben-T-Tests folgt später (siehe Einheit (9) in *Allgemeines Lineares Modell*).

T-Zufallsvariablen

T-Statistiken

Konfidenzintervalle

Selbstkontrollfragen

T-Zufallsvariablen

T-Statistiken

Konfidenzintervalle

Selbstkontrollfragen

Definition (t-Zufallsvariable)

$Z \sim N(0,1)$ sei eine Z-Zufallsvariable, $U \sim \chi^2(n)$ sei eine χ^2-Zufallsvariable mit Freiheitsgradparameter n. Weiterhin seien Z und U unabhängige Zufallsvariablen. Dann nennen wir die Zufallsvariable

$$T := \frac{Z}{\sqrt{U/n}} \tag{4}$$

eine t-verteilte Zufallsvariable mit Freiheitsgradparameter n. Wir schreiben $T \sim t(n)$. Die Wahrscheinlichkeitsdichtefunktion (WDF) einer t-Zufallsvariable bezeichnen wir mit $t(x;n)$. Die kumulative Verteilungsfunktion (KVF) und inverse KVF einer t-Zufallsvariable bezeichnen wir mit $\Psi(x;n)$ bzw. $\Psi^{-1}(x;n)$.

Bemerkungen

- Teilt man eine standardnormal-verteilte Zufallsvariable durch die Wurzel aus einer Chi-Quadrat-verteilten Zufallsvariable, geteilt durch ihren Freiheitsgradparameter, so erhält man eine t-verteilte Zufallsvariable.
- Die Definition und das folgende Theorem gehen auf Student (1908) zurück.
- Zabell (2008) gibt hierzu einen historischen Überblick.

Theorem (WDF einer t-Zufallsvariable)

T sei eine t-Zufallsvariable mit Ergebnisraum $\mathbb{R}$ und Freiheitsgradparameter n. Dann ist die Wahrscheinlichkeitsdichtefunktion von T gegeben durch

$$t(\cdot; n) : \mathbb{R} \to \mathbb{R}_{>0}, x \mapsto t(x; n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \tag{5}$$

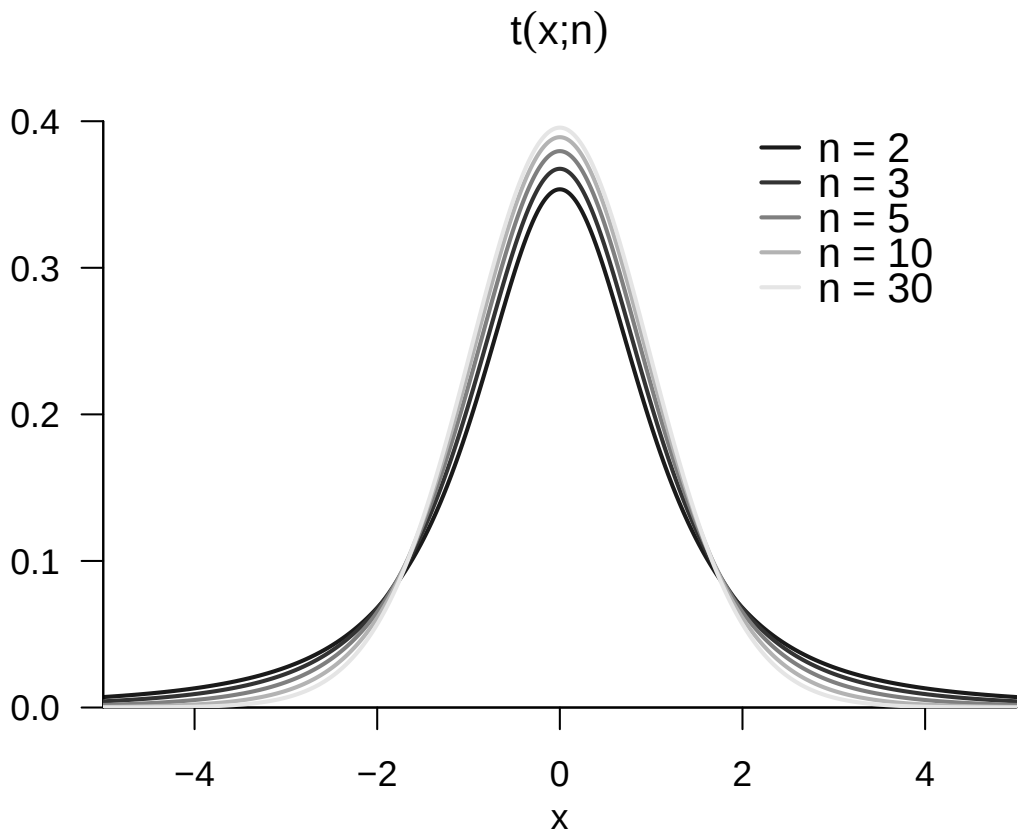
wobei Γ die Gammafunktion bezeichne.

Bemerkungen

- Wir verzichten auf einen Beweis.
- Das Theorem ist eines der zentralen Resultate der Frequentistischen Statistik.
- Die t-Verteilung ist um 0 symmetrisch. Steigendes n verschiebt Wahrscheinlichkeitsmasse von den Ausläufen zum Zentrum hin. Ab $n = 30$ gilt $t(x; n) \approx N(x; 0, 1)$.

T-Zufallsvariablen

Wahrscheinlichkeitsdichtefunktionen von t-Zufallsvariablen



Definition (Nichtzentrale t-Zufallsvariable)

$X \sim N(\delta, 1)$ sei eine normalverteilte Zufallsvariable mit Erwartungswertparameter δ, $U \sim \chi^2(n)$ sei eine χ^2-Zufallsvariable mit Freiheitsgradparameter n. Weiterhin seien X und U unabhängige Zufallsvariablen. Dann nennen wir die Zufallsvariable

$$T := \frac{X}{\sqrt{U/n}} \tag{6}$$

eine nichtzentral t-verteilte Zufallsvariable mit Nichtzentralitätsparameter δ und Freiheitsgradparameter n. Wir schreiben $T \sim t(\delta, n)$. Die WDF einer nichtzentralen t-Zufallsvariable bezeichnen wir mit $t(x; \delta, n)$. Die KVF und inverse KVF einer nichtzentralen t-Zufallsvariable bezeichnen wir mit $\Psi(x; \delta, n)$ bzw. $\Psi^{-1}(x; \delta, n)$.

Bemerkungen

- Teilt man eine normalverteilte Zufallsvariable mit Erwartungswertparameter $\mu = \delta$ und Varianzparameter $\sigma^2 = 1$ durch die Wurzel aus einer Chi-Quadrat-verteilten Zufallsvariable, geteilt durch ihren Freiheitsgradparameter, so erhält man eine nichtzentral t-verteilte Zufallsvariable.
- Eine nichtzentrale t-Zufallsvariable mit $\delta = 0$ ist eine t-Zufallsvariable. Es gilt also $t(x; 0, n) = t(x; n)$.

Theorem (WDF einer nichtzentralen t-Zufallsvariable)

T sei eine nichtzentrale t-Zufallsvariable mit Ergebnisraum $\mathbb{R}$, Nichtzentralitätsparameter δ und Freiheitsgradparameter n. Dann ist die WDF von T gegeben durch

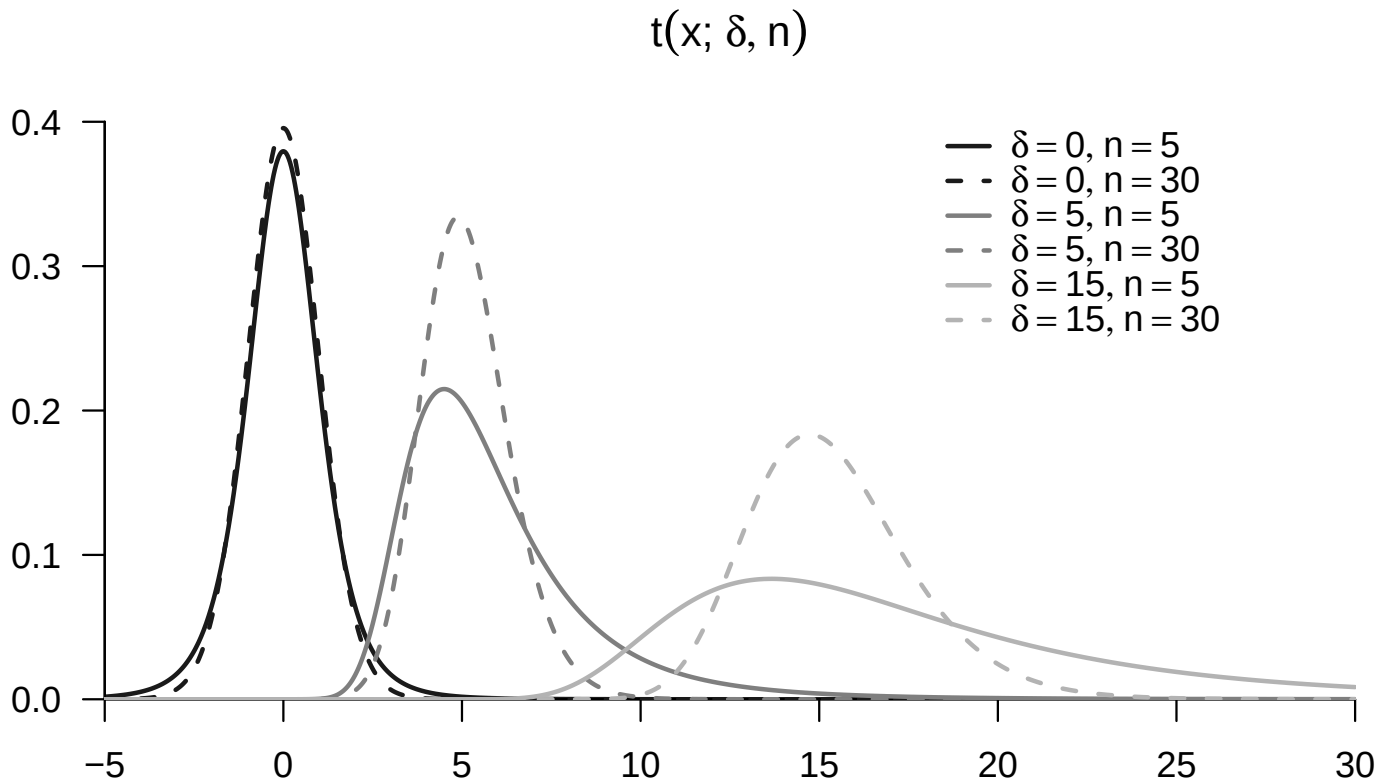
$$\begin{aligned} t(\cdot;\delta,n):\mathbb{R}\to\mathbb{R}_{>0}, x\mapsto t(x;\delta,n) := & \frac{1}{2^{\frac{n-1}{2}}\,\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}} \\ & \times\int_0^\infty \tau^{\frac{n-1}{2}}\,\exp\left(-\frac{\tau}{2}\right)\exp\left(-\frac{1}{2}\left(x\sqrt{\frac{\tau}{n}}-\delta\right)^2\right)\,d\tau\,, \end{aligned} \tag{7}$$

wobei $\exp$ die Exponentialfunktion und Γ die Gammafunktion bezeichne.

Bemerkung

- Wir verzichten auf einen Beweis.
- Die funktionale Form der WDF findet sich zum Beispiel in Lehmann (1986), Seite 254, Gleichung (80).

Wahrscheinlichkeitsdichtefunktionen nichtzentraler t-Zufallsvariablen



T-Zufallsvariablen

T-Statistiken

Konfidenzintervalle

Selbstkontrollfragen

Definition (Kontrastgewichtsvektor)

Gegeben seien das ALM

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N\left(0_n, \sigma^2 I_n\right) \tag{8}$$

mit dem Betaparametervektor $\beta \in \mathbb{R}^p$. Ein Kontrastgewichtsvektor ist dann ein Vektor derselben Dimensionalität $c \in \mathbb{R}^p$, sodass das Skalarprodukt der beiden Vektoren eine lineare Kombination der Betaparameter β mit den Kontrastgewichten c darstellt:

$$\langle c, \beta \rangle = c^{\mathrm{T}}\beta = \sum_{i=1}^{p} c_i \beta_i \ . \tag{9}$$

Bemerkungen

- Der Kontrastgewichtsvektor projiziert β auf einen Skalar $c^{\mathrm{T}}\beta \in \mathbb{R}$.
- Die Wahl p-dimensionaler Einheitsvektoren für c erlaubt die Auswahl einzelner Komponenten von β bzw. $\hat{\beta}$.
- Eine generelle Wahl von c erlaubt die Evaluation beliebiger Linearkombinationen von β bzw. $\hat{\beta}$.

Beispiel

Gegeben sei der Betaparametervektor

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} \tag{10}$$

und die Kontrastgewichtsvektoren

$$c_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad c_2 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \quad c_3 = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} . \tag{11}$$

Dann gilt:

- Der Kontrastgewichtsvektor c_1 wählt die dritte Komponente des Betaparametervektors aus.
- Der Kontrastgewichtsvektor c_2 berechnet die Differenz zwischen der ersten und zweiten Komponente des Betaparametervektors.
- Der Kontrastgewichtsvektor c_3 beschreibt eine Linearkombination der Betaparameterkomponenten $\beta_1, \dots, \beta_4$ mit den Gewichten $1, \dots, 4$.

Definition (T-Statistik für Kontrastgewichtsvektor)

Gegeben seien das ALM

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N\left(0_n, \sigma^2 I_n\right) \tag{12}$$

sowie die Betaparameter- und Varianzparameterschätzer

$$\hat{\beta} := (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y \text{ und } \hat{\sigma}^2 := \frac{(y - X\hat{\beta})^{\mathrm{T}} (y - X\hat{\beta})}{n - p} \ . \tag{13}$$

Dann ist für einen Kontrastgewichtsvektor $c \in \mathbb{R}^p$ und einen Nullparameter $\beta_0 \in \mathbb{R}^p$ die *T-Statistik* definiert als

$$T := \frac{c^{\mathrm{T}} \hat{\beta} - c^{\mathrm{T}} \beta_0}{\sqrt{\hat{\sigma}^2 c^{\mathrm{T}} (X^{\mathrm{T}} X)^{-1} c}} \ . \tag{14}$$

Bemerkungen

- Die T-Statistik hängt via $\hat{\beta}$ und $\hat{\sigma}^2$ von den Daten y ab.
- Der Kontrastgewichtsvektor projiziert $\hat{\beta}$ auf einen Skalar $c^{\mathrm{T}} \hat{\beta} \in \mathbb{R}$.

T-Statistiken

Bemerkungen (fortgeführt)

Die Wahl von $\beta_0 \in \mathbb{R}^p$ erlaubt es, die T-Statistik unterschiedlich einzusetzen:

- Wählt man $\beta_0 := 0_p$, so erhält man mit der T-Statistik eine Deskriptivstatistik, die es erlaubt, geschätzte Regressoreffekte, also Komponenten oder Linearkombinationen von $\hat{\beta}$, im Sinne eines Signal-zu-Rauschen-Verhältnisses in Bezug zu der durch $\hat{\sigma}^2$ quantifizierten Residualdatenvariabilität zu setzen. Der Nenner der T-Statistik stellt dabei sicher, dass insbesondere die adäquate (Ko-)Standardabweichung der entsprechenden Betaparameterkomponentenkombination als Bezugsgröße dient, da es sich bei $\hat{\sigma}^2 \left(X^{\mathrm{T}} X\right)^{-1}$ bekanntlich um die Kovarianz des Betaparameterschätzers handelt. Folgende erste Intuition ist in diesem Kontext hilfreich:

$$T = \frac{\text{geschätzte Effektstärke}}{\text{geschätzte stichprobenumfangskalierte Datenvariabilität}} \tag{15}$$

- Wählt man für $\beta_0 = \beta$, also den wahren, aber unbekannten Betaparameterwert, so eröffnet die T-Statistik die Möglichkeit, für die einzelnen Komponenten des Betaparametervektors Konfidenzintervalle zu bestimmen.
- Deklariert man schließlich $\beta_0 \in \Theta_0$ im Kontext eines Testszenarios als das Element einer Nullhypothese Θ_0, so eröffnet die T-Statistik die Möglichkeit Hypothesentest-basierter Inferenz über Betaparameterkomponenten und ihre Linearkombinationen im Rahmen des des ALMs.

Theorem (Verteilung der T-Statistik)

Gegeben seien das ALM

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \tag{16}$$

sowie die Betaparameter- und Varianzparameterschätzer

$$\hat{\beta} := (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y \text{ und } \hat{\sigma}^2 := \frac{(y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})}{n - p} \ . \tag{17}$$

Schließlich sei für einen Kontrastgewichtsvektor $c \in \mathbb{R}^p$ und einen Nullparameter $\beta_0 \in \mathbb{R}^p$ die T-Statistik

$$T := \frac{c^{\mathrm{T}}\hat{\beta} - c^{\mathrm{T}}\beta_0}{\sqrt{\hat{\sigma}^2 c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}} \ . \tag{18}$$

Dann gilt:

$$T \sim t(\delta, n - p) \text{ mit } \delta = \frac{c^{\mathrm{T}}\beta - c^{\mathrm{T}}\beta_0}{\sqrt{\sigma^2 c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}}. \tag{19}$$

Bemerkungen

- Die T-Statistik folgt einer nichtzentralen t-Verteilung, wobei sich der Nichtzentralitätsparameter aus der Wahl von Kontrastgewichtsvektor sowie Nullparameter und der Freiheitsgradparameter aus den Dimensionen der Designmatrix $X \in \mathbb{R}^{n \times p}$ ergibt.
- T ist eine Funktion der Parameterschätzer, δ ist eine Funktion der wahren, aber unbekannten, Parameter
- Für $c^{\mathrm{T}}\beta = c^{\mathrm{T}}\beta_0$, also bei Zutreffen der Nullhypothese, gilt $\delta = 0$ und damit $T \sim t(n - p)$.
- Für $c^{\mathrm{T}}\beta \neq c^{\mathrm{T}}\beta_0$ kann die Verteilung von T zur Herleitung von Powerfunktionen benutzt werden.

T-Statistiken

Beweis

Wir wissen, dass der Betaparameterschätzer einer multivariaten Normalverteilung folgt (siehe Einheit (6) in *Allgemeines Lineares Modell*):

$$\hat{\beta} \sim N\left(\beta, \sigma^2 (X^{\mathrm{T}} X)^{-1}\right) \ . \tag{20}$$

Mit dem Theorem zur linear-affinen Transformation multivariat normalverteilter Zufallsvektoren (vgl. Einheit (4) in *Allgemeines Lineares Modell*) gilt daher

$$c^{\mathrm{T}} \hat{\beta} - c^{\mathrm{T}} \beta_0 \sim N\left(c^{\mathrm{T}} \beta - c^{\mathrm{T}} \beta_0, \sigma^2 c^{\mathrm{T}} (X^{\mathrm{T}} X)^{-1} c\right) \tag{21}$$

sowie

$$X = \frac{1}{\sqrt{\sigma^2 c^{\mathrm{T}} (X^{\mathrm{T}} X)^{-1} c}} \left(c^{\mathrm{T}} \hat{\beta} - c^{\mathrm{T}} \beta_0\right) \sim N\left(\frac{c^{\mathrm{T}} \beta - c^{\mathrm{T}} \beta_0}{\sqrt{\sigma^2 c^{\mathrm{T}} (X^{\mathrm{T}} X)^{-1} c}}, 1\right) \ . \tag{22}$$

Wir wissen darüber hinaus, dass folgende Funktion des Varianzparameterschätzers einer Chi-Quadrat-Verteilung folgt (siehe Einheit (6) in *Allgemeines Lineares Modell*):

$$U = \frac{n-p}{\sigma^2} \hat{\sigma}^2 \sim \chi^2(n-p) \ . \tag{23}$$

Beweis (fortgeführt)

Definitionsgemäß ist der Quotient aus der Zufallsvariable X und der Wurzel der Zufallsvariable U, geteilt durch den Freiheitsgradparameter, eine nichtzentrale t-Zufallsvariable

$$T = \frac{X}{\sqrt{U/(n-p)}} \sim t(\delta, n-p) \ , \tag{24}$$

wobei der Nichtzentralitätsparameter der nichtzentralen t-Verteilung von T durch den Erwartungswertparameter der Normalverteilung von X gegeben ist

$$\delta = \frac{c^{\mathrm{T}}\beta - c^{\mathrm{T}}\beta_0}{\sqrt{\sigma^2 c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}} \tag{25}$$

und sich die T-Statistik wie folgt ergibt:

$$\begin{aligned} T &= \frac{\frac{1}{\sqrt{\sigma^2 c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}}\left(c^{\mathrm{T}}\hat{\beta} - c^{\mathrm{T}}\beta_0\right)}{\sqrt{\frac{n-p}{\sigma^2}\hat{\sigma}^2/(n-p)}} \\ &= \frac{c^{\mathrm{T}}\hat{\beta} - c^{\mathrm{T}}\beta_0}{\sqrt{\sigma^2 c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}\,\sqrt{\hat{\sigma}^2/\sigma^2}} \\ &= \frac{c^{\mathrm{T}}\hat{\beta} - c^{\mathrm{T}}\beta_0}{\sqrt{\hat{\sigma}^2 c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}} \ . \end{aligned} \tag{26}$$

□

Beispiel (1) Unabhängige und identisch normalverteilte Zufallsvariablen

Es sei

$$y \sim N(X\beta, \sigma^2 I_n) \text{ mit } X := 1_n \in \mathbb{R}^n, \beta := \mu \in \mathbb{R} \text{ und } \sigma^2 > 0. \tag{27}$$

das ALM-Szenario unabhängiger und identisch normalverteilter Zufallsvariablen. Weiterhin seien $c := 1$ und $\beta_0 := \mu_0$. Dann gilt für die T-Statistik

$$T = \frac{c^\mathrm{T}\hat{\beta} - c^\mathrm{T}\mu_0}{\sqrt{\hat{\sigma}^2 c^\mathrm{T}(X^\mathrm{T}X)^{-1}c}} = \frac{1^\mathrm{T}\bar{y} - 1^\mathrm{T}\mu_0}{\sqrt{s_y^2 1^\mathrm{T}(1_n^\mathrm{T}1_n)^{-1}1}} = \sqrt{n}\frac{\bar{y} - \mu_0}{s_y} \ , \tag{28}$$

was der Teststatistik für den Einstichproben-T-Test entspricht (vgl. Einheit (11) in *Wahrscheinlichkeitstheorie und Frequentistische Inferenz* und Einheit (9) in *Allgemeines Lineares Modell*). Die hier betrachtete T-Statistik nimmt hohe Werte für hohe Werte von $\bar{y}$ (Effekt), kleine Werte von s_y^2 (Datenvariabilität) und hohe Werte von n (Stichprobenumfang) an.

In diesem Zusammenhang ist *Cohen's* d ein beliebtes *Effektstärkenmaß*. Es ist definiert als

$$d := \frac{\bar{y}}{s_y} \ , \tag{29}$$

sodass für $\mu_0 := 0$ gilt, dass

$$T = \sqrt{n}d \quad \text{bzw.} \quad d = \frac{1}{\sqrt{n}}T \ . \tag{30}$$

Cohen's d ist also ein Stichprobenumfang-unabhängiges Signal-zu-Rauschen-Verhältnis.

Simulation (1) Unabhängig und identische normalverteilte Zufallsvariablen

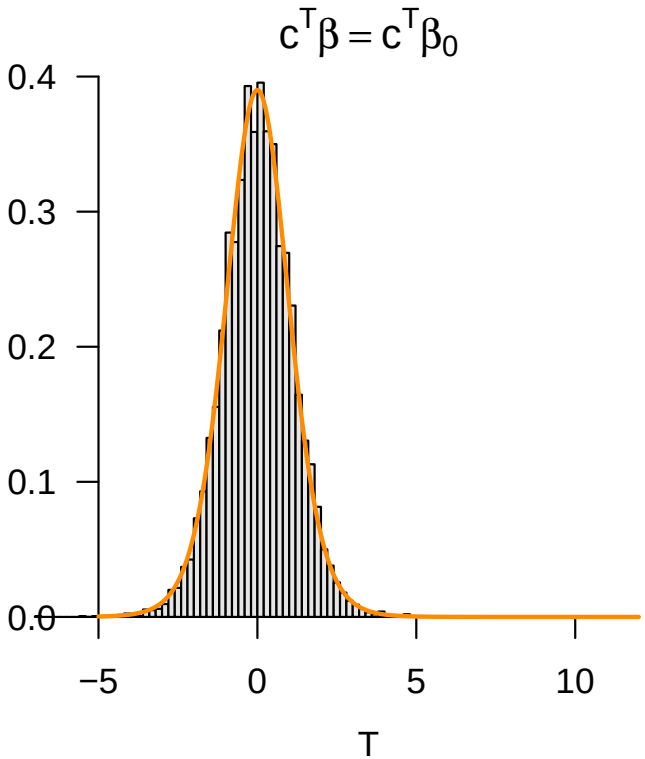
wahre, aber unbekannte Hypothesenszenarien $c^{\mathrm{T}}\beta = c^{\mathrm{T}}\beta_0$ und $c^{\mathrm{T}}\beta \neq c^{\mathrm{T}}\beta_0$

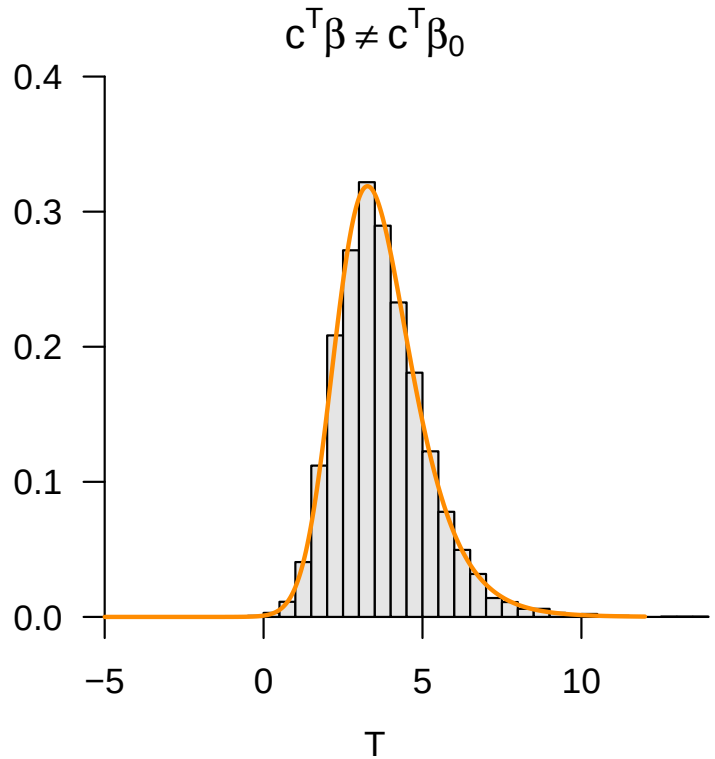
```
# Modellformulierung
library(MASS)                                                  # multivariate Normalverteilung
n           = 12                                               # Anzahl von Datenpunkten
p           = 1                                                # Anzahl von Betaparametern
X           = matrix(c(rep(1,n)), nrow = n)                    # Designmatrix
I_n         = diag(n)                                          # Einheitsmatrix
beta        = c(0,1)                                           # wahre,  aber unbekannte  Betaparameter
sigsqr      = 1                                                # wahrer, aber unbekannter Varianzparameter
nscn        = length(beta)                                     # Anzahl der Hypothesenszenarien
c           = 1                                                # Kontrastvektor von Interessse
beta_0      = 0                                                # Betaparameter gemäß Nullhypothese

# Frequentistische Simulation
nsim        = 1e4                                              # Anzahl der Simulationen
delta       = rep(NaN, nscn)                                   # Array der Nichtzentralitätsparameter
Tee         = matrix(rep(NaN, nscn*nsim), ncol = nscn)         # Array der T-Teststatistik-Realisierungen
for(s in 1:nscn){                                              # Hypothesenszenarien
    delta[s]       = ((t(c) %*% beta[s] - t(c) %*% beta_0)/    # \delta
                      sqrt(sigsqr*t(c)%*%solve(t(X)%*%X)%*%c))
    for(i in 1:nsim){                                          # Simulationsiterationen
        y          = mvrnorm(1, X %*% beta[s], sigsqr*I_n)     # y
        beta_hat   = solve(t(X) %*% X) %*% t(X) %*% y          # \hat{\beta}
        eps_hat    = y - X %*% beta_hat                        # \hat{\eps}
        sigsqr_hat = (t(eps_hat) %*% eps_hat)/(n-p)            # \hat{\sigma}^2
        Tee[i,s]   = ((t(c) %*% beta_hat - t(c) %*% beta_0)/   # T
                      sqrt(sigsqr_hat*t(c)%*%solve(t(X)%*%X)%*%c))
    }
}
```

Simulation (1) Unabhängig und identische normalverteilte Zufallsvariablen

wahre, aber unbekannte, Hypothesenszenarien $c^{\mathrm{T}}\beta = c^{\mathrm{T}}\beta_0$ und $c^{\mathrm{T}}\beta \neq c^{\mathrm{T}}\beta_0$





Simulation (2) Einfache lineare Regression

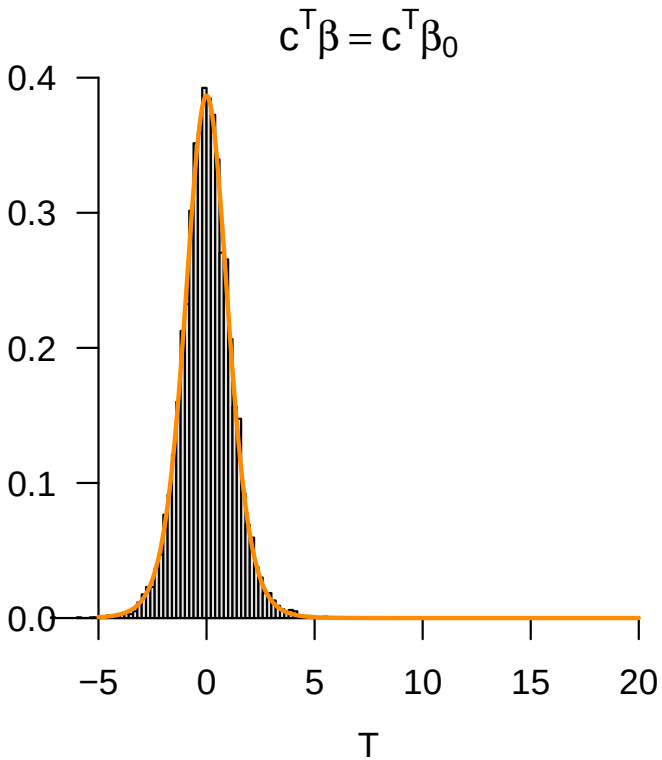
wahre, aber unbekannte, Hypothesenszenarien $c^{\mathrm{T}}\beta = c^{\mathrm{T}}\beta_0$ und $c^{\mathrm{T}}\beta \neq c^{\mathrm{T}}\beta_0$

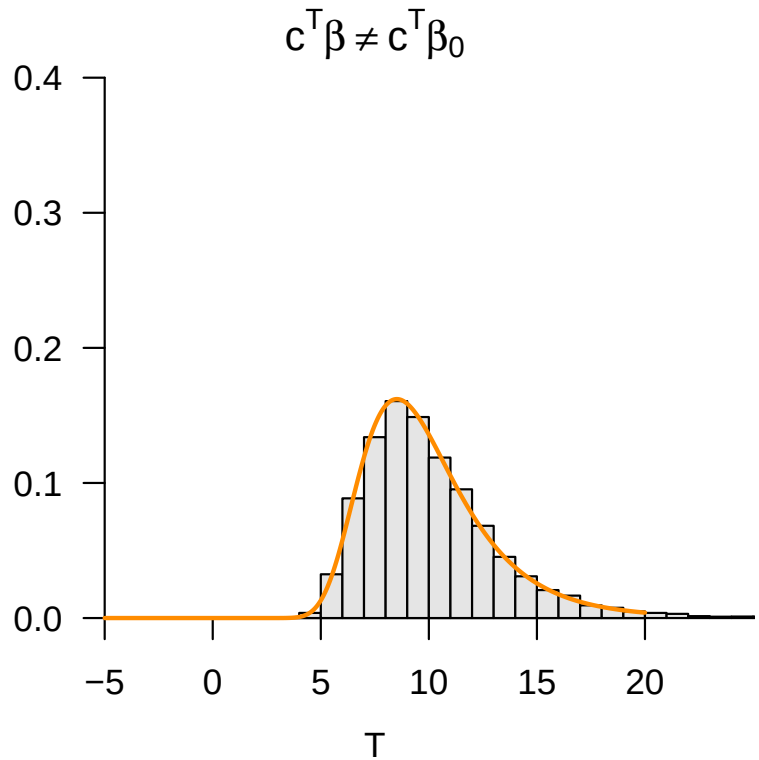
```
# Modellformulierung
library(MASS)                                                  # multivariate Normalverteilung
n          = 10                                                # Anzahl von Datenpunkten
p          = 2                                                 # Anzahl von Betaparametern
x          = 1:n                                               # Prädiktorwerte
X          = matrix(c(rep(1,n),x), ncol = p)                   # Designmatrix
I_n        = diag(n)                                           # Einheitsmatrix
beta       = matrix(c(1,0,
                      1,1), nrow = 2)                          # wahre,  aber unbekannte  Betaparameter
sigsqr     = 1                                                 # wahrer, aber unbekannter Varianzparameter
nscn       = ncol(beta)                                        # Anzahl der Hypothesenszenarien
c          = matrix(c(0,1), nrow = 2)                          # Kontrastvektor von Interessse
beta_0     = matrix(c(0,0), nrow = 2)                          # Betaparameter gemäß Nullhypothese

# Frequentistische Simulation
nsim       = 1e4                                               # Anzahl der Simulationen
delta      = rep(NaN, nscn)                                    # Array der Nichtzentralitätsparameter
Tee        = matrix(rep(NaN, nscn*nsim), ncol = nscn)          # Array der T-Teststatistik-Realisierungen
for(s in 1:nscn){                                              # Hypothesenszenarien
    delta[s]       = ((t(c) %*% beta[,s] - t(c) %*% beta_0)/   # \delta
                      sqrt(sigsqr*t(c)%*%solve(t(X)%*%X)%*%c))
    for(i in 1:nsim){                                          # Simulationsiterationen
        y          = mvrnorm(1, X %*% beta[,s], sigsqr*I_n)    # y
        beta_hat   = solve(t(X) %*% X) %*% t(X) %*% y          # \hat{\beta}
        eps_hat    = y - X %*% beta_hat                        # \hat{\eps}
        sigsqr_hat = (t(eps_hat) %*% eps_hat)/(n-p)            # \hat{\sigma}^2
        Tee[i,s]   = ((t(c) %*% beta_hat - t(c) %*% beta_0)/   # T
                      sqrt(sigsqr_hat*t(c)%*%solve(t(X)%*%X)%*%c))
    }
}
```

Simulation (2) Einfache lineare Regression

wahre, aber unbekannte, Hypothesenszenarien $c^{\mathrm{T}}\beta = c^{\mathrm{T}}\beta_0$ und $c^{\mathrm{T}}\beta \neq c^{\mathrm{T}}\beta_0$





Anekdote: Herkunft der "Student'schen t-Verteilung"

William Sealy Gosset (1876 – 1937)

(Quelle: *Wikimedia Commons*: "William_Sealy_Gosset.jpg"; Lizenz: gemeinfrei.)

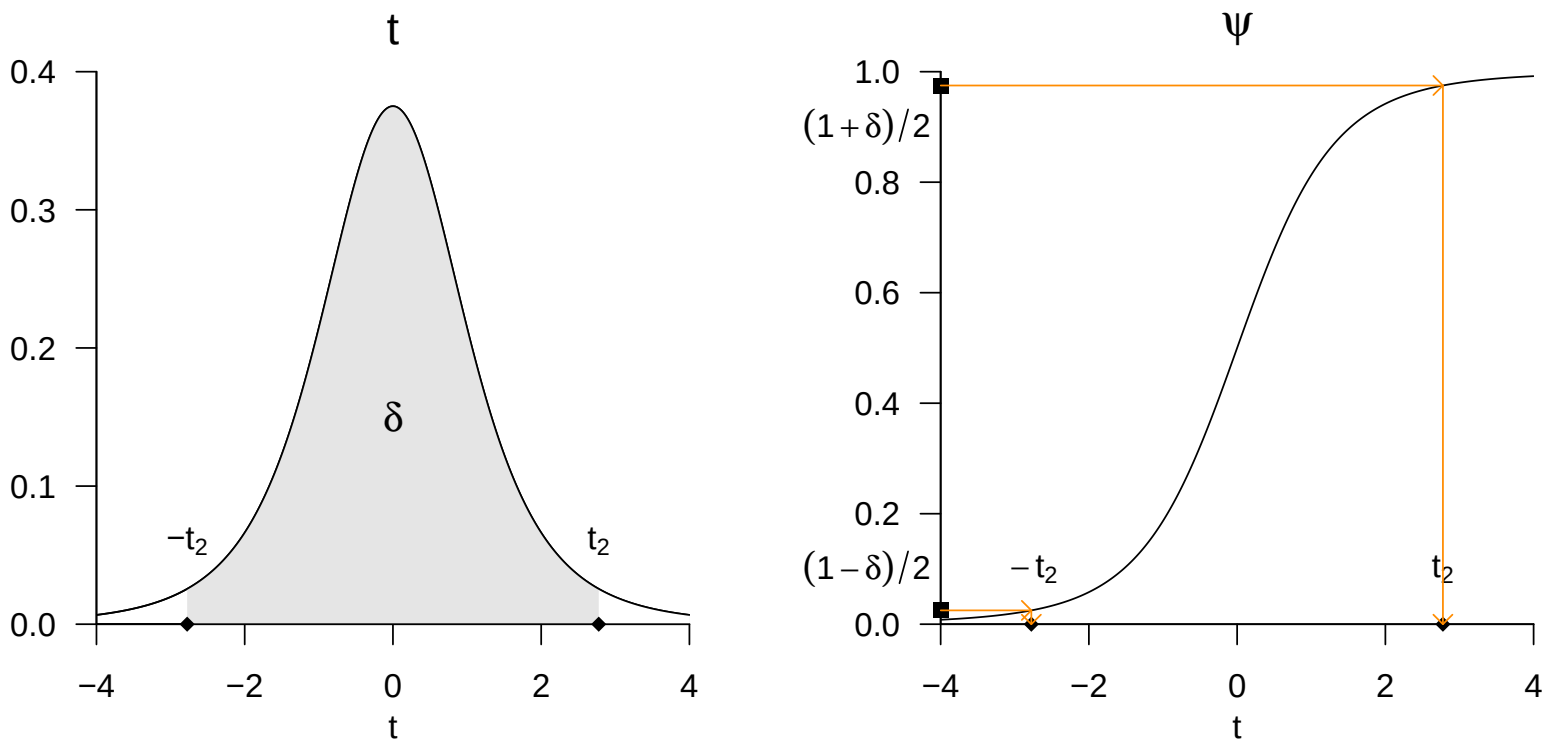
T-Zufallsvariablen

T-Statistiken

Konfidenzintervalle

Selbstkontrollfragen

Wiederholung: Konfidenzbedingung für die T-Konfidenzintervallstatistik



(siehe Einheit (10) in *Wahrscheinlichkeitstheorie und Frequentistische Inferenz*)

Theorem (Konfidenzintervalle für Betaparameterkomponenten)

Gegeben seien das ALM

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \tag{31}$$

sowie der Betaparameterschätzer $\hat{\beta}$ und der Varianzparameterschätzer $\hat{\sigma}^2$. Für ein $\gamma \in]0, 1[$ sei

$$t_\gamma := \Psi^{-1}\left(\frac{1+\gamma}{2}; n-p\right) . \tag{32}$$

Schließlich sei λ_j das das $j$te Diagonalelement von $\left(X^{\mathrm{T}}X\right)^{-1}$ für $j = 1, \ldots, p$:

$$\lambda_j := \left(\left(X^{\mathrm{T}}X\right)^{-1}\right)_{jj} . \tag{33}$$

Dann ist

$$\kappa_j := \left[\hat{\beta}_j - \hat{\sigma}\sqrt{\lambda_j}t_\gamma, \hat{\beta}_j + \hat{\sigma}\sqrt{\lambda_j}t_\gamma\right] \tag{34}$$

ein γ-Konfidenzintervall für die $j$te Komponente β_j des Betaparameters $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, für $j = 1, \ldots, p$.

Bemerkungen

- Intuitiv gilt im Vergleich mit dem Konfidenzintervall für den Erwartungswertparameter bei der Normalverteilung

$$\hat{\beta}_j \approx \bar{y}, \quad \hat{\sigma} \approx S, \quad \sqrt{\lambda_j} \approx \sqrt{n^{-1}} \quad \text{und} \quad t_\gamma = t_\delta \tag{35}$$

  (vgl. Einheit (10) in *Wahrscheinlichkeitstheorie und Frequentistische Inferenz*).

Beweis

Wir müssen zeigen, dass

$$\mathbb{P}(\kappa_j \ni \beta_j) = \gamma \, . \tag{36}$$

Dazu halten wir zunächst fest, dass für alle $j = 1, \dots, p$ bei Wahl von $\beta_0 = \beta$ und $c := e_j$ nach dem Theorem zur T-Statistik für $T \sim t(\delta, n-p)$ gilt, dass

$$T = \frac{e_j^{\mathrm{T}} \hat{\beta} - e_j^{\mathrm{T}} \beta}{\sqrt{\hat{\sigma}^2 e_j^{\mathrm{T}} \left(X^{\mathrm{T}} X\right)^{-1} e_j}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 \left(\left(X^{\mathrm{T}} X\right)^{-1}\right)_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{\lambda_j}} =: T_j \, . \tag{37}$$

und

$$\delta = \frac{e_j^{\mathrm{T}} \beta - e_j^{\mathrm{T}} \beta}{\sqrt{\hat{\sigma}^2 e_j^{\mathrm{T}} \left(X^{\mathrm{T}} X\right)^{-1} e_j}} = 0 \tag{38}$$

Damit gilt dann auch sofort, dass $T_j \sim t(n-p)$. Weiterhin erinnern wir daran (vgl. Einheit (10) in *Wahrscheinlichkeitstheorie und Frequentistischer Inferenz*), dass per Definition von t_γ gilt, dass

$$\mathbb{P}\left(-t_\gamma \leq T_j \leq t_\gamma\right) = \gamma \, . \tag{39}$$

Beweis (fortgeführt)

Aus der Definition eines γ-Konfidenzintervalls folgt dann

$$\begin{aligned}
\gamma &= \mathbb{P}\left(-t_\gamma \leq T_j \leq t_\gamma\right) \\
&= \mathbb{P}\left(-t_\gamma \leq \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{\lambda_j}} \leq t_\gamma\right) \\
&= \mathbb{P}\left(-t_\gamma \hat{\sigma}\sqrt{\lambda_j} \leq \hat{\beta}_j - \beta_j \leq t_\gamma \hat{\sigma}\sqrt{\lambda_j}\right) \\
&= \mathbb{P}\left(-\hat{\beta}_j - t_\gamma \hat{\sigma}\sqrt{\lambda_j} \leq -\beta_j \leq -\hat{\beta}_j + t_\gamma \hat{\sigma}\sqrt{\lambda_j}\right) \\
&= \mathbb{P}\left(\hat{\beta}_j + t_\gamma \hat{\sigma}\sqrt{\lambda_j} \geq \beta_j \geq \hat{\beta}_j - t_\gamma \hat{\sigma}\sqrt{\lambda_j}\right) \\
&= \mathbb{P}\left(\hat{\beta}_j - t_\gamma \hat{\sigma}\sqrt{\lambda_j} \leq \beta_j \leq \hat{\beta}_j + t_\gamma \hat{\sigma}\sqrt{\lambda_j}\right) \\
&= \mathbb{P}\left(\left[\hat{\beta}_j - \hat{\sigma}\sqrt{\lambda_j} t_\gamma, \hat{\beta}_j + \hat{\sigma}\sqrt{\lambda_j} t_\gamma\right] \ni \beta_j\right) \\
&= \mathbb{P}(\kappa_j \ni \beta_j)
\end{aligned} \tag{40}$$

und damit ist alles gezeigt. □

Beispiel (1) Unabhängig und identische normalverteilte Zufallsvariablen

Wir betrachten die ALM-Form des Szenarios unabhängig und identisch normalverteilter Zufallsvariablen:

$$y \sim N(X\beta, \sigma^2 I_n) \text{ mit } X := 1_n \in \mathbb{R}^n, \beta := \mu \in \mathbb{R}, \sigma^2 > 0 \; . \tag{41}$$

Wie bereits gesehen erhalten wir dann:

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} y_i =: \bar{y}, \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 =: s^2 \text{ und } \lambda_1 = \left(1_n^{\mathrm{T}} 1_n\right)^{-1} = \frac{1}{n} \; . \tag{42}$$

Nach dem Theorem zu Konfidenzintervallen für Betaparameterkomponenten gilt dann, dass

$$\kappa := \left[\bar{y} - \frac{s}{\sqrt{n}} t_\gamma, \bar{y} + \frac{s}{\sqrt{n}} t_\gamma\right] \tag{43}$$

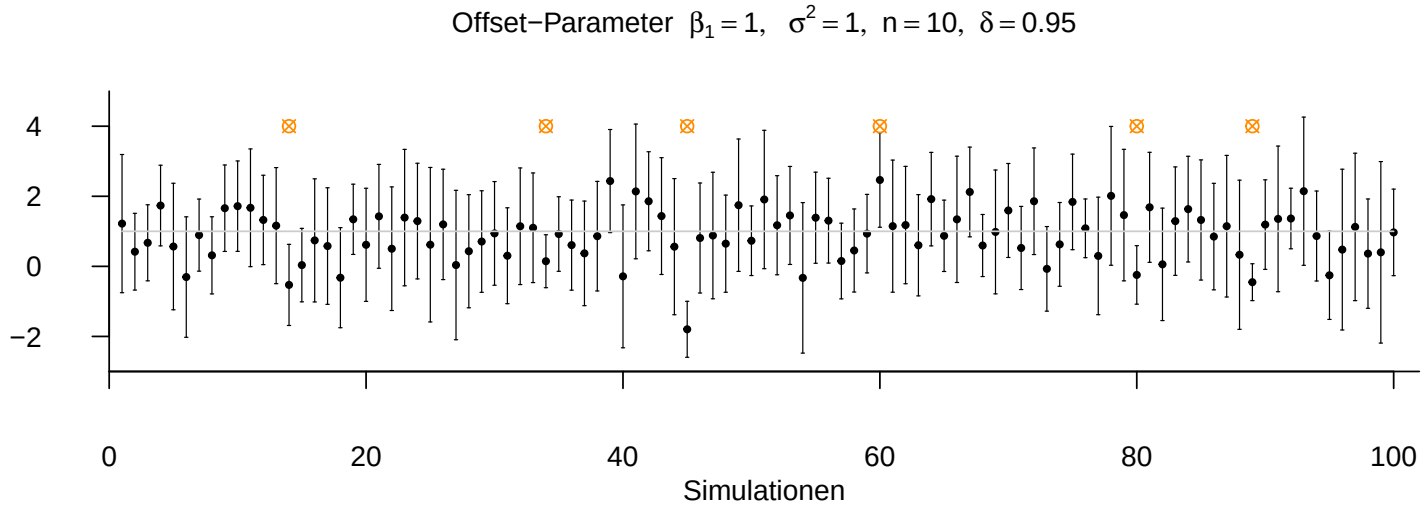
ein γ-Konfidenzintervall für β ist und dieses ist offenbar identisch mit dem Konfidenzintervall für den Erwartungsparameter der Normalverteilung, welches wir bereits eingeführt haben (siehe Einheit (10) in *Wahrscheinlichkeitstheorie und Frequentistische Inferenz*).

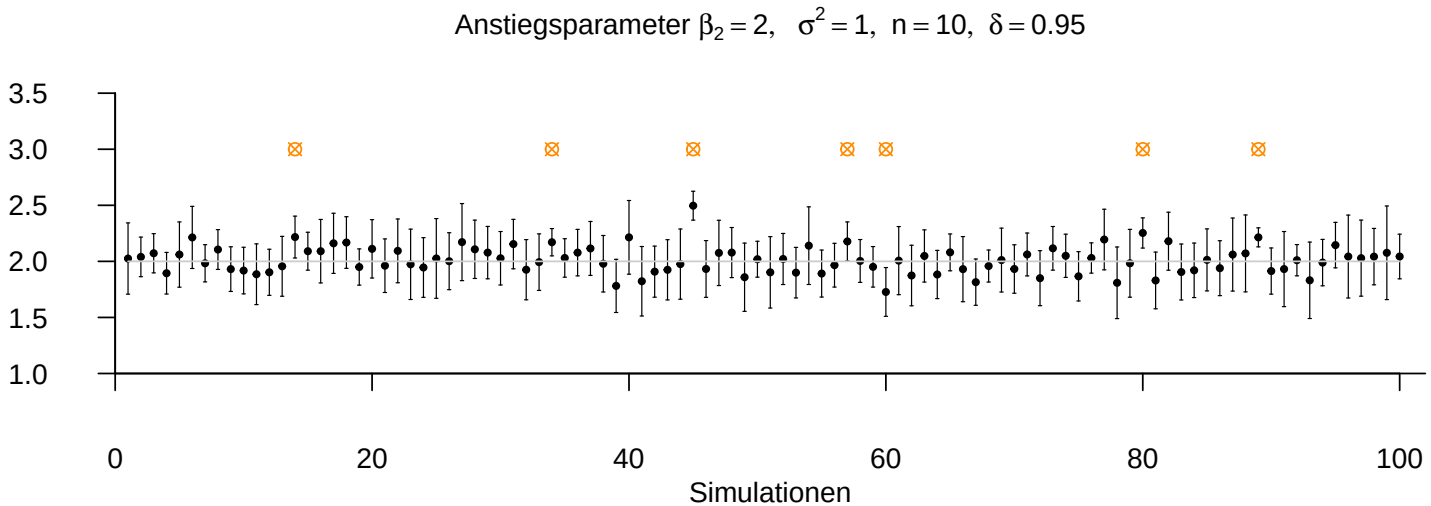
Beispiel (2) Einfache lineare Regression

```
# Modellformulierung
library(MASS)                                                    # multivariate Normalverteilung
set.seed(0)                                                      # Zufallszahlengenerator initialisieren
n          = 10                                                  # Anzahl von Datenpunkten
p          = 2                                                   # Anzahl von Betaparametern
x          = 1:n                                                 # Prädiktorwerte
X          = matrix(c(rep(1,n),x), ncol = p)                     # Designmatrix
I_n        = diag(n)                                             # Einheitsmatrix
beta       = matrix(c(1,2), nrow = 2)                            # wahre,  aber unbekannte  Betaparameter
sigsqr     = 1                                                   # wahrer, aber unbekannter Varianzparameter
gamma      = 0.95                                                # Konfidenzbedingung
t_gamma    = qt((1+gamma)/2,n-1)                                 # \Psi^{-1}((1+\gamma)/2,n-1)
lambda     = diag(solve(t(X) %*% X))                             # \lambda_j, j = 1,...,p

# Simulation
nsim       = 1e2                                                 # Anzahl der Simulationen
kappa      = array(rep(NaN, nsim*p*p), dim=c(nsim,2,2))          # Konfidenzintervallarray
beta_hat   = matrix(rep(NaN,p*nsim), nrow = p)                   # Betaparameterschätzerarray
for(i in 1:nsim){                                                # Iteration über Realisierungen
    y              = mvrnorm(1, X %*% beta, sigsqr*I_n)          # y
    beta_hat[,i]   = solve(t(X) %*% X) %*% t(X) %*% y            # \hat{\beta}
    eps_hat        = y - X %*% beta_hat[,i]                      # \hat{\varepsilon}
    sigsqr_hat     = (t(eps_hat) %*% eps_hat)/(n-p)              # \hat{\sigma}^2
    for(j in 1:p){                                               # Iteration über Betaparameterkomponenten
        kappa[i,1,j] = beta_hat[j,i]-sqrt(sigsqr_hat*lambda[j])*t_gamma # untere KI-Grenze
        kappa[i,2,j] = beta_hat[j,i]+sqrt(sigsqr_hat*lambda[j])*t_gamma # obere  KI-Grenze
    }
}
```

Simulation von Konfidenzintervallen bei einfacher linearer Regression





T-Zufallsvariablen

T-Statistiken

Konfidenzintervalle

Selbstkontrollfragen

Selbstkontrollfragen

1. Geben Sie die Definition einer t-Zufallsvariable wieder.
2. Geben Sie die Definition einer nichtzentralen t-Zufallsvariable wieder.
3. Skizzieren Sie die WDFen von t-Zufallsvariablen mit Freiheitsgradparametern 2, 10 und 30.
4. Skizzieren Sie die WDFen von nichtzentralen t-Zufallsvariablen mit Nichtzentralitätsparametern 0, 5 und 15.
5. Erläutern Sie den Begriff des Kontrastgewichtsvektors.
6. Geben Sie die Definition der T-Statistik wieder.
7. Erläutern Sie für die T-Statistik die Bedeutung der Wahl von $c \in \mathbb{R}^p$.
8. Erläutern Sie für die T-Statistik die Bedeutung der Wahl von $\beta_0 \in \mathbb{R}^p$.
9. Wann und warum kann die T-Statistik als Signal-zu-Rauschen-Verhältnis interpretiert werden?
10. Geben Sie das Theorem zur Verteilung der T-Statistik wieder.
11. Geben Sie die Formel für die T-Statistik im ALM-Szenario von unabhängig und identisch normalverteilten Zufallsvariablen wieder.
12. Erläutern Sie den Zusammenhang zwischen der T-Statistik und Cohen's d.
13. Geben Sie das Theorem zu Konfidenintervallen für Betaparameterkomponenten wieder.
14. Geben Sie die Formel für das Konfidenzintervall des Erwartungswertparameters im ALM-Szenario von unabhängig und identisch normalverteilten Zufallsvariablen wieder.

Referenzen

Lehmann, E. L. 1986. *Testing Statistical Hypotheses*. Wiley Series in Probability and Statistics.

Student. 1908. "The Probable Error of a Mean." *Biometrika* 6 (1): 1–25.

Zabell, S. L. 2008. "On Student's 1908 Article 'The Probable Error of a Mean'." *Journal of the American Statistical Association* 103 (481): 1–7. https://doi.org/10.1198/016214508000000030.