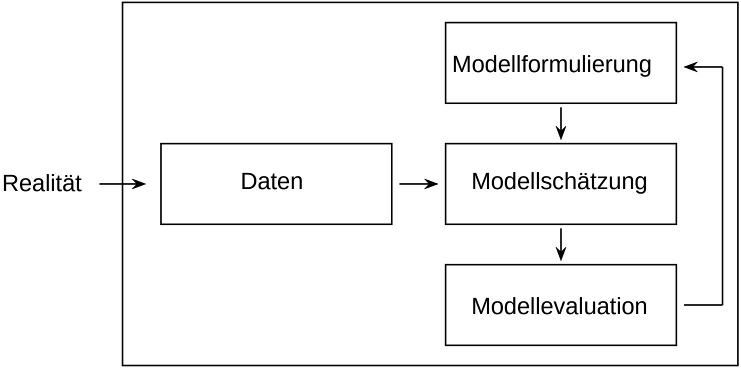
Allgemeines Lineares Modell

BSc Psychologie SoSe 2023

Prof. Dr. Dirk Ostwald

(7) T-Statistiken

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

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

Standardprobleme Frequentistischer Inferenz

(1) Parameterschätzung

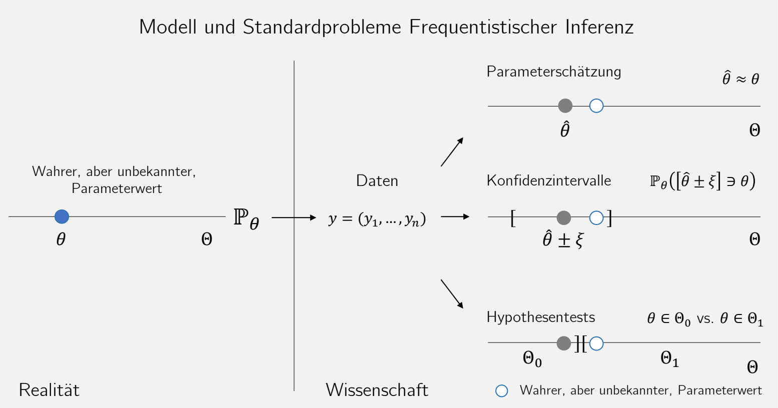
Ziel der Parameterschätzung ist es, einen möglichst guten Tipp für wahre, aber unbekannte, Parameterwerte oder Funktionen dieser abzugeben, typischerweise mithilfe von Daten.

(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.



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

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

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

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

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

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

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

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

- In diesem Abschnitt 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 im Rahmen von Einstich- und Zweistichproben T-Tests ist das Thema von Einheit (9) T-Tests.

T-Zufallsvariablen

T-Statistiken

Konfidenzintervalle

Selbstkontrollfragen

T-Zufallsvariablen

T-Statistiken

Konfidenzintervalle

Selbstkontrollfragen

T-Zufallsvariablen

Definition (t-Zufallsvariable)

ξ sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}$ und WDF

$$p_\xi : \mathbb{R} \to \mathbb{R}_{>0}, x \mapsto p_\xi(x) := \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \tag{4}$$

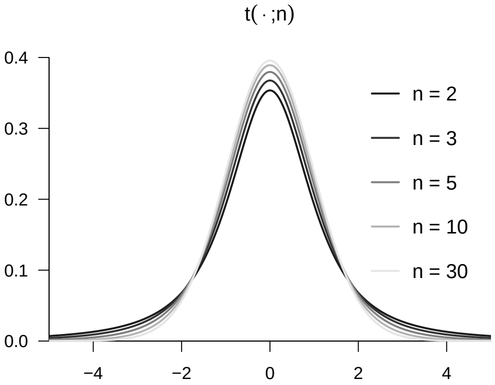
wobei Γ die Gammafunktion bezeichne. Dann sagen wir, dass ξ einer t-Verteilung mit Freiheitsgradparameter n unterliegt und nennen ξ eine t-Zufallsvariable mit Freiheitsgradparameter n. Wir kürzen dies mit $\xi \sim t(n)$ ab. Die WDF einer t-Zufallsvariable bezeichnen wir mit $t(\cdot; n)$. Die KVF und inverse KVF einer nichtzentralen t-Zufallsvariable bezeichnen wir mit $\Psi(\cdot; n)$ und $\Psi^{-1}(\cdot; n)$, respektive.

Bemerkungen

- Die Verteilung ist um 0 symmetrisch
- Steigendes n verschiebt Wahrscheinlichkeitsmasse aus den Ausläufen zum Zentrum.
- Ab $n = 30$ gilt $t(n) \approx N(0, 1)$.

T-Zufallsvariablen

Wahrscheinlichkeitsdichtefunktionen von t-Zufallsvariablen



Theorem (T-Transformation)

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

$$T := \frac{Z}{\sqrt{U/n}} \tag{5}$$

eine t-verteilte Zufallsvariable mit Freiheitsgradparameter n, es gilt also $T \sim t(n)$.

Bemerkungen

- Das Theorem geht auf Student (1908) zurück.
- Das Theorem ist eines der zentralen Resultate der Frequentistischen Statistik.
- Zabell (2008) gibt hierzu einen historischen Überblick.

T-Zufallsvariablen

Definition (Nichtzentrale t-Zufallsvariable)

ξ sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}$ und WDF

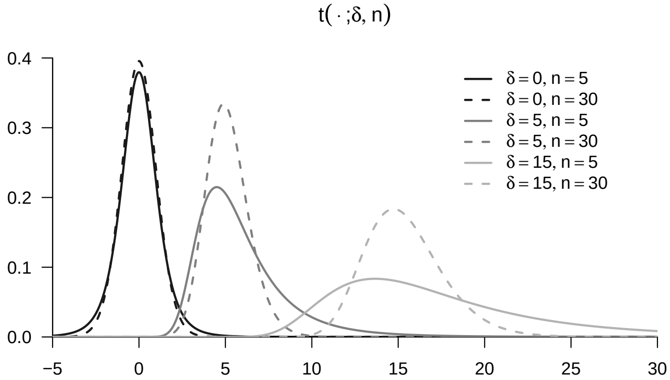
$$p_\xi : \mathbb{R} \to \mathbb{R}_{>0}, x \mapsto p_\xi(x) := \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n}{2}\right)(n\pi)^{\frac{1}{2}}} \times \int_0^\infty \tau^{\frac{n-1}{2}} \exp\left(-\frac{\tau}{2}\right) \exp\left(-\frac{1}{2}\left(x\left(\frac{\tau}{n}\right)^{\frac{1}{2}} - \delta\right)^2\right) d\tau. \quad (6)$$

Dann sagen wir, dass ξ einer nichtzentralen t-Verteilung mit Nichtzentralitätsparameter δ und Freiheitsgradparameter n unterliegt und nennen ξ eine *nichtzentrale t-Zufallsvariable mit Nichtzentralitätsparameter δ und Freiheitsgradparameter n*. Wir kürzen dies mit $\xi \sim t(\delta, n)$ ab. Die WDF einer nichtzentralen t-Zufallsvariable bezeichnen wir mit $t(\cdot; \delta, n)$. Die KVF und inverse KVF einer nichtzentralen t-Zufallsvariable bezeichnen wir mit $\Psi(\cdot; \delta, n)$ und $\Psi^{-1}(\cdot; \delta, n)$, respektive.

Bemerkungen

- Eine nichtzentrale t-Zufallsvariable mit $\delta = 0$ ist eine t-Zufallsvariable.
- Es gilt also $t(\cdot; 0, n) = t(\cdot; n)$.
- Die funktionale Form der WDF findet sich zum Beispiel in Lehmann (1986), Seite 254, Gl. (80).

Wahrscheinlichkeitsdichtefunktionen nichtzentraler t-Zufallsvariablen



T-Zufallsvariablen

Theorem (Nichtzentrale T-Transformation)

$\xi \sim N(\mu, 1)$ sei eine normalverteilte Zufallsvariable, $U \sim \chi^2(n)$ sei eine χ^2 Zufallsvariable mit Freiheitsgradparameter n, und ξ und U seien unabhängige Zufallsvariablen. Dann ist die Zufallsvariable

$$T := \frac{\xi}{\sqrt{U/n}} \tag{7}$$

eine nichtzentrale t-Zufallsvariable mit Nichtzentralitätsparameter μ und Freiheitsgradparameter n, es gilt also $T \sim t(\mu, n)$

Bemerkung

- Wir verzichten auf einen Beweis.

T-Zufallsvariablen

T-Statistiken

Konfidenzintervalle

Selbstkontrollfragen

Definition (T-Statistik)

Es sei

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

das ALM. Weiterhin seien

$$\hat{\beta} := (X^T X)^{-1} X^T \upsilon \text{ und } \hat{\sigma}^2 := \frac{(\upsilon - X\hat{\beta})^T(\upsilon - X\hat{\beta})}{n-p} \tag{9}$$

die Betaparameter- und Varianzparameterschätzer, respektive. Dann ist für einen *Kontrastgewichtsvektor* $c \in \mathbb{R}^p$ und einen Parameter $\beta_0 \in \mathbb{R}^p$ die *T-Statistik* definiert als

$$T := \frac{c^T\hat{\beta} - c^T\beta_0}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}}. \tag{10}$$

Bemerkungen

- Die T-Statistik hängt via $\hat{\beta}$ und $\hat{\sigma}^2$ von den Daten υ ab.
- Der Kontrastgewichtsvektor projiziert $\hat{\beta}$ auf einen Skalar $c^T\hat{\beta} \in \mathbb{R}$.
- Die Wahl p-dimensionaler Einheitsvektoren für c erlaubt die Auswahl einzelner Komponenten von $\hat{\beta}$ bzw. β_0.
- Eine generelle Wahl von c erlaubt die Evaluation beliebiger Linearkombinationen von $\hat{\beta}$ bzw. β_0.

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 adequate (Ko)Standardabweichung der entsprechenden Betaparameterkomponentenkombination als Bezugsgröße dient, da es sich bei $\hat{\sigma}^2 \left(X^T X\right)^{-1}$ bekanntlich um die Kovarianz des Betaparameterschätzers handelt. Folgende erste Intuition ist in diesem Kontext hilfreich:
$$T = \frac{\text{Geschätzte Effektstärke}}{\text{Geschätzte stichprobenumfangskalierte Datenvariabilität}} \tag{11}$$
- Wählt man für $\beta_0 = \beta$, also den wahren, aber unbekannten, Betaparameterwert, so eröffnet die T-Statistik die Möglichkeit, für einzelen 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 Hypothesentest-basierte Inferenz über Betaparameterkomponenten und ihrer Linearkombinationen. des ALMs.

Theorem (T-Statistik)

Es sei

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

das ALM. Weiterhin seien

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

die Betaparameter- und Varianzparameterschätzer, respektive. Schließlich sei für einen Kontrastgewichtsvektor $c \in \mathbb{R}^p$ und einen Parameter $\beta_0 \in \mathbb{R}^p$

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

die T-Statistik. Dann gilt

$$T \sim t(\delta, n-p) \text{ mit } \delta := \frac{c^T\beta - c^T\beta_0}{\sqrt{\sigma^2 c^T (X^T X)^{-1} c}}. \tag{15}$$

Bemerkungen

- Wir verzichten auf einen Beweis.
- T ist eine Funktion der Parameterschätzer, δ ist eine Funktion der wahren, aber unbekannten, Parameter
- Für $c^T\beta = c^T\beta_0$, also bei Zutreffen der Nullhypothese, gilt $\delta = 0$ und damit $T \sim t(n-p)$.
- Für $c^T\beta \neq c^T\beta_0$ kann die Verteilung von T zur Herleitung von Powerfunktionen benutzt werden.

T-Statistiken

Beispiel (1) Unabhängige und identisch normalverteilte Zufallsvariablen

Es sei

$$\upsilon \sim N(X\beta, \sigma^2 I_n) \text{ mit } X := 1_n \in \mathbb{R}^{n \times 1}, \beta := \mu \in \mathbb{R} \text{ und } \sigma^2 > 0. \tag{16}$$

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

$$T = \frac{c^T \hat{\beta} - c^T \mu_0}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}} = \frac{1^T \bar{\upsilon} - 1^T \mu_0}{\sqrt{s_y^2 1^T (1_n^T 1_n)^{-1} 1}} = \sqrt{n} \frac{\bar{\upsilon} - \mu_0}{s_\upsilon} \tag{17}$$

was der Einstichproben-T-Teststatistik für den Fall $\mu_0 = 0$ entspricht (vgl. Einheit (12) Hypothesentest in Wahrscheinlichkeitstheorie und Frequentistische Inferenz und Einheit (9) T-Tests in Allgemeines Lineares Modell). Die hier betrachtete T-Statistik nimmt hohe Werte für hohe Werte von $\bar{\upsilon}$ (Effekt), kleine Werte von s_υ^2 (Datenvariabilität) und hohe Werte von n (Stichprobenumfang) an.

Eine beliebte Definition in diesem Zusammenhang ist *Cohen's* d als *Effektstärkenmaß*. Es gilt

$$d := \frac{\bar{\upsilon}}{s_\upsilon}, \tag{18}$$

so dass für $\mu_0 := 0$ gilt, dass

$$T = \sqrt{n} d \text{ bzw. } d = \frac{1}{\sqrt{n}} T. \tag{19}$$

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

Simulation (1) Unabhängig und identische normalverteilte Zufallsvariablen

Wahre, aber unbekannte, Hypothesenszenarien $c^T\beta = c^T\beta_0$ und $c^T\beta \neq c^T\beta_0$

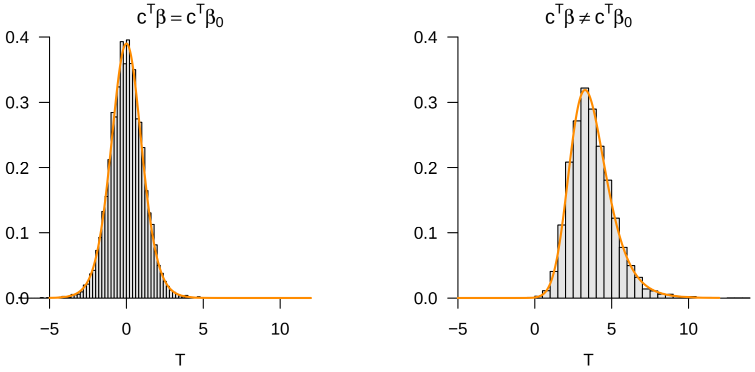
```
# Libraries
library(MASS)                                                  # multivariate Normalverteilung

# Modellformulierung
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
nscn        = length(beta)                                     # Anzahl wahrer, aber unbekannter, Hypothesenszenarien
sigsqr      = 1                                                # wahrer, aber unbekannter, Varianzparameter
c           = 1                                                # Kontrastvektor von Interessse
beta_0      = 0                                                # Nullhypothesenbetaparameter

# Frequentistische Simulation
nsim        = 1e4                                              # Anzahl Simulationen
delta       = rep(NaN, nscn)                                   # Anzahl Nichtzentralitätsparameter
Tee         = matrix(rep(NaN, nscn*nsim), ncol = nscn)         # T-Teststatistik Realisierungsarray
for(s in 1:nscn){                                              # Hypothesenszenarien
  delta[s]     = ((t(c) %*% beta[s] - t(c) %*% beta_0)/        # Nichtzentralitätsparameter
                 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 $c^T\beta = c^T\beta_0$ und $c^T\beta \neq c^T\beta_0$



Simulation (1) Einfache lineare Regression

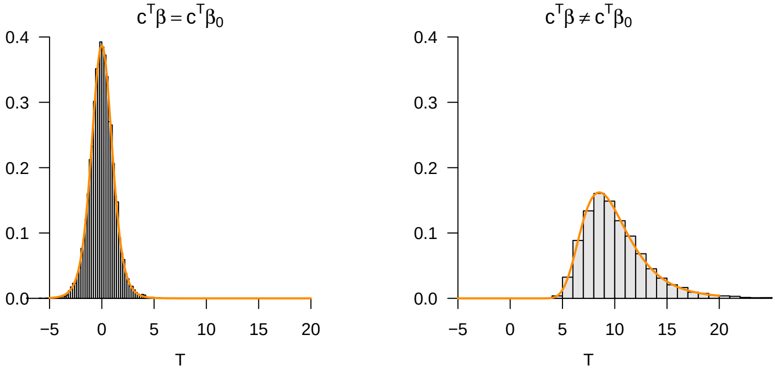
Wahre, aber unbekannte, Hypothesenszenarien $c^T\beta = c^T\beta_0$ und $c^T\beta \neq c^T\beta_0$

```
# 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
nscn        = ncol(beta)                               # Anzahl wahrer, aber unbekannter, Hypothesenszenarien
sigsqr      = 1                                        # wahrer, aber unbekannter, Varianzparameter
c           = matrix(c(0,1), nrow = 2)                 # Kontrastvektor von Interessse
beta_0      = matrix(c(0,0), nrow = 2)                 # Nullhypothesenbetaparameter

# Frequentistische Simulation
nsim        = 1e4                                      # Anzahl Simulationen
delta       = rep(NaN, nscn)                           # Anzahl Nichtzentralitätsparameter
Tee         = matrix(rep(NaN, nscn*nsim), ncol = nscn) # T-Teststatistik Realisierungsarray
for(s in 1:nscn){                                      # Hypothesenszenarien
  delta[s]    = ((t(c) %*% beta[,s] - t(c) %*% beta_0)/ # Nichtzentralitätsparameter
                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) Einfache lineare Regression

Wahre, aber unbekannte, Hypothesenszenarien $c^T\beta = c^T\beta_0$ und $c^T\beta \neq c^T\beta_0$



T-Zufallsvariablen

T-Statistiken

Konfidenzintervalle

Selbstkontrollfragen

Theorem (Konfidenzintervalle für Betaparameterkomponenten)

Es sei

$$\upsilon = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \tag{20}$$

das ALM, $\hat{\beta}$ und $\hat{\sigma}^2$ seien die Betaparameter- und Varianzparameterschätzer, respektive und und für ein $\delta \in]0,1[$ sei

$$t_\delta := \Psi^{-1}\left(\frac{1+\delta}{2}; n-p\right). \tag{21}$$

Schließlich sei für $j = 1, ...p$

$$\lambda_j := \left(\left(X^T X\right)^{-1}\right)_{jj} \text{ das } j\text{te Diagonalelement von } \left(X^T X\right)^{-1}. \tag{22}$$

Dann ist für $j = 1, ...p$

$$\kappa_j := \left[\hat{\beta}_j - \hat{\sigma}\sqrt{\lambda_j} t_\delta, \hat{\beta}_j + \hat{\sigma}\sqrt{\lambda_j} t_\delta\right] \tag{23}$$

ein δ-Konfidenzintervall für die $j$te Komponente β_j des Betaparameters $\beta = (\beta_1, ..., \beta_p)^T$.

Bemerkungen

- Intuitiv gilt im Vergleich zum Konfidenzintervall für den Erwartungswertparameter bei Normalverteilung

$$\hat{\beta}_j \approx \bar{\upsilon}, \hat{\sigma} \approx S, \sqrt{\lambda_j} \approx \sqrt{n^{-1}} \text{ und } t_\delta = t_\delta, \tag{24}$$

vgl. (11) Konfidenzintervalle in Wahrscheinlichkeitstheorie und Frequentistische Inferenz.

Beweis

Wir müssen zeigen, dass

$$\mathbb{P}(\kappa_j \ni \beta_j) = \delta. \tag{25}$$

Dazu halten wir zunächst fest, dass für alle $j = 1, ..., p$ 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

$$T = \frac{e_j^T \hat{\beta} - e_j^T \beta}{\sqrt{\hat{\sigma}^2 e_j^T \left(X^T X\right)^{-1} e_j}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 \left(\left(X^T X\right)^{-1}\right)_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{\lambda_j}} =: T_j. \tag{26}$$

und

$$\delta = \frac{e_j \beta - e_j^T \beta}{\sqrt{\hat{\sigma}^2 e_j^T \left(X^T X\right)^{-1} e_j}} = 0 \tag{27}$$

Damit gilt dann auch sofort, dass $T_j \sim t(n-p)$. Weiterhin erinnern erinnern wir daran (vgl. (11) Konfidenzintervalle in Wahrscheinlichkeitstheorie und Frequentistischer Inferenz), dass per Definition von t_δ gilt, dass

$$\mathbb{P}\left(-t_\delta \le T_j \le t_\delta\right) \tag{28}$$

Beweis (fortgeführt)

Aus der Definition eines δ-Konfidenzintervalls folgt dann

$$
\begin{aligned}
\delta & = \mathbb{P}\left(-t_\delta \le T_j \le t_\delta\right) \\
& = \mathbb{P}\left(-t_\delta \le \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{\lambda_j}} \le t_\delta\right) \\
& = \mathbb{P}\left(-t_\delta\hat{\sigma}\sqrt{\lambda_j} \le \hat{\beta}_j - \beta_j \le t_\delta\hat{\sigma}\sqrt{\lambda_j}\right) \\
& = \mathbb{P}\left(-\hat{\beta}_j - t_\delta\hat{\sigma}\sqrt{\lambda_j} \le -\beta_j \le -\hat{\beta}_j + t_\delta\hat{\sigma}\sqrt{\lambda_j}\right) \\
& = \mathbb{P}\left(\hat{\beta}_j + t_\delta\hat{\sigma}\sqrt{\lambda_j} \ge \beta_j \ge \hat{\beta}_j - t_\delta\hat{\sigma}\sqrt{\lambda_j}\right) \\
& = \mathbb{P}\left(\hat{\beta}_j - t_\delta\hat{\sigma}\sqrt{\lambda_j} \le \beta_j \le \hat{\beta}_j + t_\delta\hat{\sigma}\sqrt{\lambda_j}\right) \\
& = \mathbb{P}\left(\left[\hat{\beta}_j - \hat{\sigma}\sqrt{\lambda_j}t_\delta, \hat{\beta}_j + \hat{\sigma}\sqrt{\lambda_j}t_\delta\right] \ni \beta_j\right) \\
& = \mathbb{P}(\kappa_j \ni \beta_j)
\end{aligned}
\tag{29}
$$

und damit ist alles gezeigt.

Konfidenzintervall bei unabhängigen und identische normalverteilten Zufallsvariablen

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

$$\upsilon \sim N(X\beta, \sigma^2 I_n) \text{ mit } X := 1_n \in \mathbb{R}^n, \beta := \mu \in \mathbb{R}, \sigma^2 > 0 \tag{30}$$

Dann gelten, wie bereits gesehen

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} \upsilon_i =: \bar{\upsilon}, \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\upsilon_i - \bar{\upsilon})^2 =: s^2 \text{ und } \lambda_1 = \left(1_n^T 1_n\right)^{-1} = \frac{1}{n} \tag{31}$$

Nach dem Theorem zu Konfidenzintervallen für Betaparameterkomponenten gilt dann, dass

$$\kappa := \left[\bar{\upsilon} - \frac{s}{\sqrt{n}} t_\delta, \bar{\upsilon} + \frac{s}{\sqrt{n}} t_\delta\right] \tag{32}$$

ein δ-Konfidenzintervall für β ist und dieses ist offenbar identisch mit dem δ-Konfidenzintervall für den Erwartungsparameter der Normalverteilung, welches wir in (9) Konfidenzintervalle in Wahrscheinlichkeitstheorie und Frequentistische Inferenz eingeführt haben.

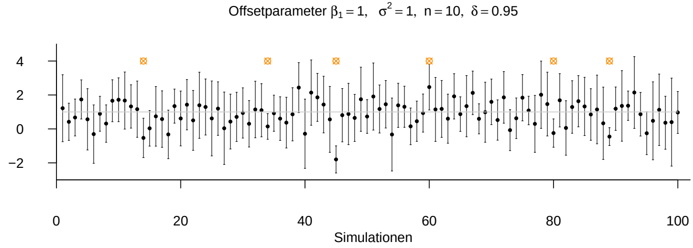
Konfidenzintervalle

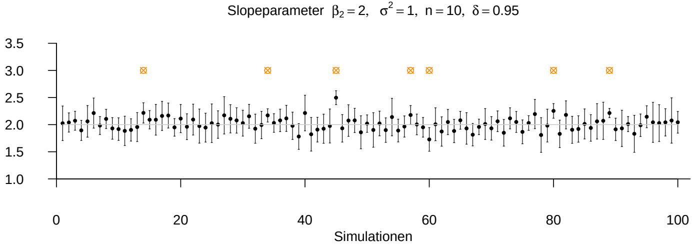
Simulation von Konfidenzintervallen bei einfacher linearer Regression

```r
# Modellformulierung
library(MASS)                                                              # multivariate Normalverteilung
set.seed(0)                                                                # random number generator seed
ns          = 1e2                                                          # Anzahl Simulationen
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
delta       = 0.95                                                         # Konfidenzbedingung
t_delta     = qt((1+delta)/2,n-1)                                          # \Psi^{-1}((1+\delta)/2,n-1)
lambda      = diag(solve(t(X) %*% X))                                      # \lambda_j Werte

# Simulation
kappa       = array(rep(NaN, ns*p*p), dim=c(ns,2,2))                       # Konfidenzintervallarray
beta_hat    = matrix(rep(NaN,p*ns), nrow = p)                              # Betaparameterschätzer
for(i in 1:ns){                                                            # Iteration über Realisierungen
  y             = mvrnorm(1, X %*% beta, sigsqr*I_n)                       # Datenrealisierung
  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 Betaarraykomponenten
    kappa[i,1,j] = beta_hat[j,i]-sqrt(sigsqr_hat*lambda[j])*t_delta # untere KI Grenze
    kappa[i,2,j] = beta_hat[j,i]+sqrt(sigsqr_hat*lambda[j])*t_delta # obere  KI Grenze
  }
}
```

Simulation von Konfidenzintervallen bei einfacher linearer Regression





T-Verteilungen

T-Statistiken

Konfidenzintervalle

Selbstkontrollfragen

Selbstkontrollfragen

1. Skizzieren Sie die WDFen von t-Zufallsvariablen mit 2, 10 und 30 Freiheitsgraden.
2. Skizzieren Sie die WDFen von nichtzentralen t-Zufallsvariablen mit Nichtzentralitätsparametern 0,5 und 15.
3. Geben Sie die Definition der T-Statistik wieder.
4. Erläutern Sie für die T-Statistik die Bedeutung der Wahl von $c \in \mathbb{R}^p$.
5. Erläutern Sie für die T-Statistik die Bedeutung der Wahl von $\beta_0 \in \mathbb{R}^p$.
6. Wann und warum kann die T-Statistik als Signal-zu-Rauschen Verhältnis interpretiert werden?
7. Geben Sie das Theorem zur T-Statistik wieder.
8. Geben Sie die Form der T-Statistik im Szenario von n u.i.n.v. Zufallsvariablen wieder.
9. Erläutern Sie den Zusammenhang der T-Statistik und Cohen's d.
10. Geben Sie das Theorem zu Konfidenintervallen für Betaparameterkomponenten 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.