Vorlesung 2

Strukturen-prüfende Verfahren: Regressionsanalyse

SoSe2007

Dr.Silika Prohl
Lehrstuhl für Statistik, Ökonometrie und Empirische Wirtschaftsforschung,
Universität Tübingen

Literatur:

✓Backhaus, Erichson, Plinke, und Weiber (2006): Multivariate Analysemethoden: Eine anwendungsorientierte Einfuehrung. 11. Auflage, Springer. Kapitel 1, S. 45-94.

- Homepage: http://www.innovation.uni-trier.de/multivariate

✓ Hair, Black, Babin, Anderson, and Tatham (2006): Multivariate Data Analysis. Chapter 4.

- Homepage: http://www.mvstats.com/

Regressionsanalyse

- Problemstellung und Vorgehensweise bei der Regressionsanalyse
- Einfache Lineare Regression
- Multiple Lineare Regression
- Prüfung der Regressionsfunktion
- Prüfung der Regressionskoeffizienten
- Prüfung der Modellprämissen
- Anwendungsbeispiel

Einfache Regressionsanalyse: Schätzung der Regressionsfunktion

Das einfache Regressionsmodell:

$$y_n = \beta_0 + \beta_1 x_n + e_n .$$

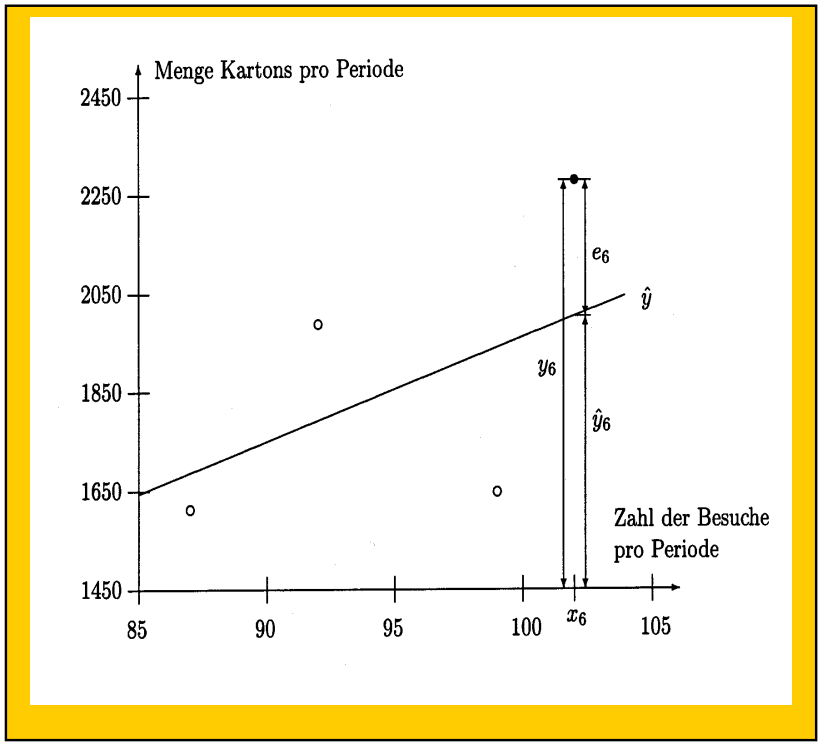
Der Teil $\beta_0 + \beta_1 x_n$ beschreibt die Gerade,

Der Fehler e_n beschreibt die Abweichung von der Gerade

Notation für den Parameter:

β_0 ist Achsenabschnitt, β_1 ist Steigung.

Was ist e_n ?



Einfache Regression

Annahmen des linearen Regressionsmodells:

A1. $y_n = \beta_0 + \beta_1 x_n + e_n$ **mit** $n = 1,2,...,N$ **und** $N > J + 1$

Das Modell ist *richtig* spezifiziert. Linearitätsannahme.

A2. $E(e_n|x) = 0$ **für** $n = 1,2,...,N$ **. Störgroessen haben den Erwartungswert Null.**

A3. $\mathrm{Cov}(e_n, x_n) = 0$ **Erklärende Variable ist nicht mit der Störgrösse korreliert**

A4. $\mathrm{Var}(e_n) = \sigma^2$ **Störgrössen haben eine konstante Varianz *(Homoskedastizität)***

A5. Die Störgrössen e_n **sind *normalverteilt.***

Einfache Regressionsanalyse: Zielfunktion

Kleinste-Quadrat-Summen-Schätzung: Zielfunktion:

$$\sum_{n=1}^{N} \hat{e}_n^2 = \sum_{n=1}^{N} [y_n - (\hat{\beta}_0 + \hat{\beta}_1 x_n)]^2 \rightarrow \min!$$

Kleinste-Quadrat-Summen-Schätzung (Beweis: Backhaus et al. (2006), S.114):

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \frac{\sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})}{\sum_{n=1}^{N} (x_n - \bar{x})^2} = \frac{\sum_{n=1}^{N} (x_n y_n - N\,\bar{x} \cdot \bar{y})}{\sum_{n=1}^{N} x_n^2 - N\bar{x}^2},$$

und

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Einfache Regression: Anwendungsbeispiel

Beobachtungen N	MENGE y_n	BESUCHE x_n	xy	x^2
1	2585	109	281765	11881
.....	.....	.....	.....	.....
10	1913	79	151127	6241
$\sum$	18068	936	1724403	89370
	$\overline{y}$ =1806,8	$\overline{x}$ =93,6		

Regressionsmodell:

$\hat{Y} = \beta_0 + \beta_1 \text{BESUCHE} + e_n$.

KQ-Schaetzer:

$$\hat{\beta}_1 = \frac{10 \cdot 1724403 - 936 \cdot 18068}{10 \cdot 89370 - (936)^2} = 18,881 \quad , \quad \hat{\beta}_0 = 1806,8 - (18,881 \cdot 93,6) = 39,5$$

Anwendungsbeispiel für geschätzte einfache Regressionsfunktion:

$\hat{y}_n = 39,5 + 18,881\, x_n$, Beispiel: $\hat{y}_6 = 39,5 + 18,881\, x_6 = 39,5 + 18,881 \cdot 102 = 1965$

Residuum: $y_6 - \hat{y}_6 = 2278\text{-}1965 = 313$

Einfache Regression: Anwendungsbeispiel

Software Output (n=37)

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,507[a]	,258	,236	331,84495

a. Einflußvariablen : (Konstante), besuche

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	1336925	1	1336925,4	12,141	,001[a]
	Residuen	3854238	35	110121,07		
	Gesamt	5191163	36			

a. Einflußvariablen : (Konstante), besuche

b. Abhängige Variable: menge

Koeffizienten [a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	751,158	320,624		2,343	,025
	besuche	12,195	3,500	,507	3,484	,001

a. Abhängige Variable: menge

Multiple Regression: Zielfunktion

Motivation fuer eine Multiple Regression

Ziel: Bestimme den *ceteris paribus* Effekt des Regressors (hier $x_1, x_2 \ldots x_n$) auf die abhängige Variable y.

Modell 1: $y_n = \beta_0 + \beta_1 x_1 + e_n, \; E(e_n | x_1) = 0$

$$\partial E(y | x_1) / \partial x_1 = \beta_1$$

Modell 2: $y_n = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e_n, \; E(e_n | x_1, x_2) = 0$

$$\partial E(y | x_1, x_2) / \partial x_1 = \beta_1$$

x_2 wird explizit konstant gehalten.

Anwendungsbeispiel:

(1) $\text{MENGE} = \beta_0 + \beta_1 \text{BESUCHE} + e_n$

(2) $\text{MENGE} = \beta_0 + \beta_1 \text{BESUCHE} + \beta_2 \text{PREIS} + e_n$

Regressionsanalyse: Prüfung der Modellprämissen

Annahmen des linearen Regressionsmodells:

A1. $y_n = \beta_0 + \sum_{j=1}^{J} \beta_j x_{jn} + e_n$ **mit** $n = 1,2,...,N$ **und** $N > J + 1$

Das Modell ist *richtig* spezifiziert. Linearitätsannahme.

A2. $E(e_n | x_1,...,x_1) = 0$ **für** $n = 1,2,...,N$ **. Störgroessen haben den Erwartungswert Null.**

A3. $Cov(e_n, x_{jn}) = 0$ **Erklärende Variable ist nicht mit der Störgrösse korreliert**

A4. $Var(e_n) = \sigma^2$ **Störgrössen haben eine konstante Varianz *(Homoskedastizität)***

A5. $Cov(e_n, e_{n+i}) = 0$ **mit** $i \neq 0$ **Störgrösse sind unkorreliert *(keine Autokorrelation)***

A6. Zwischen den erklaerenden Variablen X_j **besteht keine lineare Abhaengigkeit**

(keine perfekte Multikollinearitaet)

A7. Die Störgrössen e_n **sind *normalverteilt.***

Multiple Regressionsanalyse: Zielfunktion

Multiples Regressionmodell:

$$y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + ... + \beta_j x_{jn} + e_n$$

Zielfunktion der KQ-Schätzung

$$\sum_{n=1}^{N} \hat{e}_n^2 = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 = \sum_{n=1}^{N} [y_n - (\hat{\beta}_0 + \hat{\beta}_1 x_{1n} + \hat{\beta}_2 x_{2n} + ... + \hat{\beta}_j x_{jn})]^2 \rightarrow \min!$$

Die FOC besteht aus j+1 linearen Gleichungen in j+1 Unbekannten.

KQ-Schätzung (Beweis: Backhaus et al. (2006), S.114):

Z.B.: Im Fall j=2 kann man zeigen, dass

$$\hat{\beta}_1 = \frac{S_{y1} S_{22} - S_{y2} S_{12}}{S_{11} S_{22} - S_{12}^2} \text{ und } \hat{\beta}_2 = \frac{S_{y2} S_{11} - S_{y1} S_{12}}{S_{11} S_{22} - S_{12}^2}$$

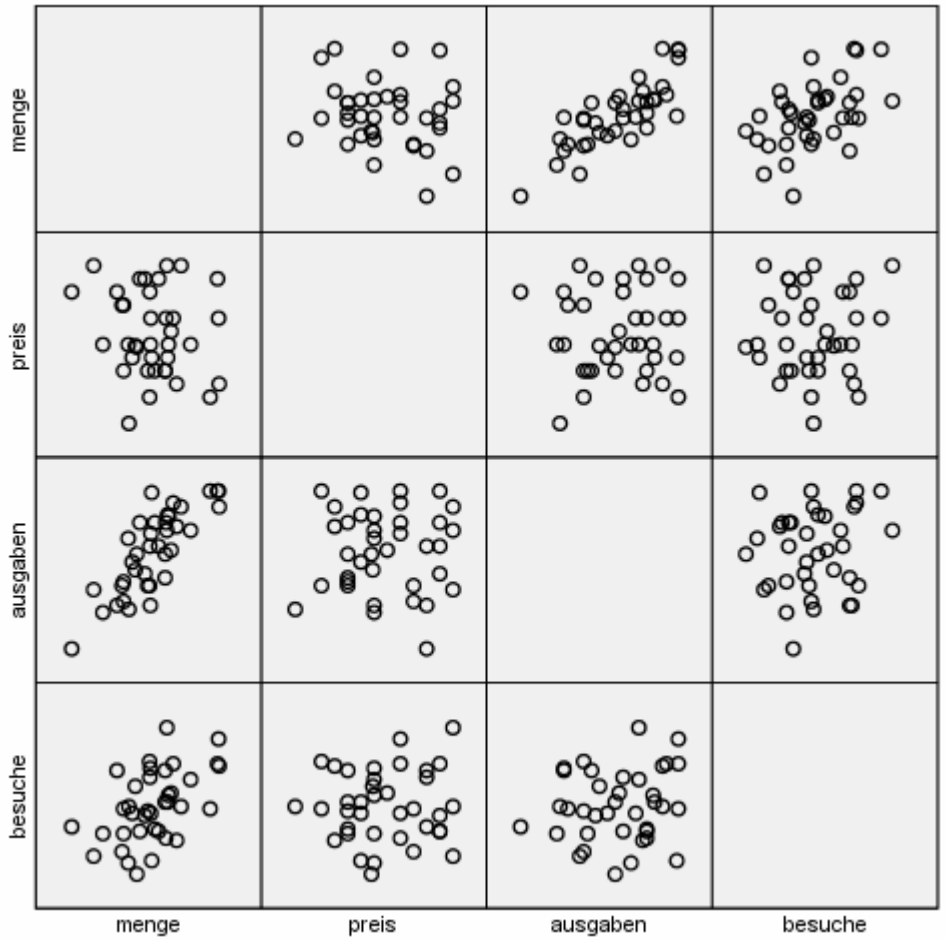
wobei

$$S_{y1} = \sum_{n=1}^{N} (y_n - \bar{y})(x_{1n} - \bar{x}_1), \quad S_{y2} = \sum_{n=1}^{N} (y_n - \bar{y})(x_{1n} - \bar{x}_2)$$

$$S_{11} = \sum_{n=1}^{N} (x_{1n} - \bar{x}_1)^2, \quad S_{12} = \sum_{n=1}^{N} (x_{1n} - \bar{x}_1)(x_{2n} - \bar{x}_2), \quad S_{22} = \sum_{n=1}^{N} (x_{2n} - \bar{x}_2)^2$$

Multiple Regressionsanalyse

Software Output (n=37)



Multiple Regressionsanalyse: Anwendungsbeispiele

Multiple Regression: Anwendungsbeispiel (n=10)

$$\hat{Y} = -6.9 + 11.085 \cdot \text{BESUCHE} + 9.927 \cdot \text{PREIS} + 0.655 \cdot \text{AUSGABEN} + e_n$$

Angenommen, wir interessieren uns für 6. Beobachtung:

$$\hat{Y}_6 = -6.9 + 11.085 \cdot 102 + 9.927 \cdot 10 + 0.655 \cdot 1500 = 2206$$

Residuum: 2278-2206=72

Multiple Regression: Anwendungsbeispiel (n=37)

Software Output (n=37)

Koeffizienten [a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		B	Standardfehler	Beta	T	Signifikanz
1	(Konstante)	763,650	223,946		3,410	,002
	besuche	9,705	1,658	,404	5,854	,000
	preis	-45,177	16,102	-,191	-2,806	,008
	ausgaben	,551	,050	,753	10,925	,000

a. Abhängige Variable: menge

Prüfung der Regressionsfunktion

Globale Guetermasse zur Pruefung der Regressionsfunktion:

- ***Das Bestimmtheitsmass***
- ***Die F - Statistik***
- ***Der Standardfehler***

Masse zur Pruefung der Regressionskoeffizienten:

- ***Der t-Wert***
- ***Der Beta-Wert***

Regressionsanalyse: Bestimmtheitsmass

Bestimmtheitsmass: misst die Güte der Anpassung der Regressionsfunktion an die empirischen Daten („goodness of fit“). Ein Basis dafuer sind Residualgroessen, d.h., die Abweichungen zwischen den Beobachtungswerten und den geschätzten Werten von Y. Diese lassen sich wie folgt zerlegen:

Gesamtabweichung = Erklärte Abweichung + Residuum

TSS = ESS + RSS

$(y_n - \bar{y}) = (\hat{y}_n - \bar{y}) + (y_n - \hat{y}_n)$

Beispiel für einfache Regression:

$(y_6 - \bar{y}) = (\hat{y}_6 - \bar{y}) + (y_6 - \hat{y}_6)$

471,2 = 158,6 + 312,6

Als **Gesamtstreuung** wird die Summe der quadrierten Gesamtabweichungen aller Beobachtungen bezeichnet.

Gesamtstreuung = Erklärte Streuung + nicht erklärte Streuung

$$\sum_{n=1}^{N}(y_n - \bar{y})^2 = \sum_{n=1}^{N}(\hat{y}_n - \bar{y})^2 + \sum_{n=1}^{N}(y_n - \hat{y}_n)^2$$

Multiple Regressionsfunktion: Anwendungsbeispiel

Software Output (n=37)

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	4395066	3	1465022,0	60,728	,000[a]
	Residuen	796097,0	33	24124,152		
	Gesamt	5191163	36			

a. Einflußvariablen : (Konstante), ausgaben, preis, besuche

b. Abhängige Variable: menge

Regressionsanalyse: Standardfehler der Schätzung

Standardfehler der Schaetzung:

$$s = \sqrt{\frac{\sum_{n=1}^{N} \hat{e}^2_n}{N - J - 1}}$$

Anwendungsbeispiel für einfache Regression (n=10):

$$s = \sqrt{\frac{1188685}{10 - 1 - 1}} \qquad = 385$$

Regressionsanalyse: Bestimmtheitsmass

Ein Modell ist umso besser, je grösser der Anteil der Gesamtvariation ist, der durch das Modell erklärt wird

Bestimmtheitsmass: $$R^2 = \frac{\sum_{n=1}^{N}(\hat{y}_n - \bar{y})^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} = \frac{ESS}{TSS} = \frac{\text{erklaerte Streuung}}{\text{Gesamtstreuung}}, \quad 0 < R^2 \leq 1$$

äquivalent

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} = 1 - \frac{RSS}{TSS} = 1 - \frac{\text{nicht erklaerte Streuung}}{\text{Gesamtstreuung}}$$

Aber: R^2 **sagt wenig über die Güte der Schätzung eines partiellen Effektes aus**

R^2 **ignoriert die (wichtige?) Frage, ob ein Effekt kausal ist**

Im Falle einfacher Regression: $R^2 = r^2$**.**

.

Regressionsanalyse: Bestimmtheitsmass

Software Output (n=37)

Einfache Regression

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,507[a]	,258	,236	331,84495

a. Einflußvariablen : (Konstante), besuche

Multiple Regression

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,920[a]	,847	,833	155,31952

a. Einflußvariablen : (Konstante), ausgaben, preis, besuche

Regressionsanalyse: F-Statistik

F-Statistik:

$$F_{emp} = \frac{\sum_{n=1}^{N} (\hat{y}_n - \bar{y})^2 / J}{\sum_{n=1}^{N} (y_n - \hat{y}_n)^2 / (N - J - 1)} = \frac{\text{erklaerte Streuung / J}}{\text{nicht erklaerte Streuung / (N - J - 1)}} = \frac{R^2 / J}{(1 - R^2)/(N - J - 1)}$$

$H_0 : \beta_1 = \ldots = \beta_j = 0$ kein Zusammenhang zwischen Y und X_j

$H_1 :$ nicht alle β_j sind Null

Anwendungsbeispiel für einfache Regression (n=10):

$$F_{emp} = \frac{0{,}3455/1}{(1-0{,}3455)/(10-1-1)} = 4{,}223, \quad F_{theor.} = \mathbf{5{,}32}$$ (siehe Abbildung 1.17, S. 71)

Wichtig: $F_{emp} > F_{theor.}$ H_0 wird verworfen *Zusammenhang ist signifikant*

$F_{emp} \leq F_{theor.}$ H_0 wird nicht verworfen

Regressionsanalyse: Prüfung der Regressionskoeffizienten

Hypothesentest fuer β_0

Hypothese: $H_0 : \beta_0 = b_0$,

$$H_1: \begin{cases} (a) & \beta_0 > b_0 & \text{(einseitig)} \\ (b) & \beta_0 < b_0 & \text{(einseitig)} \\ (c) & \beta_0 \neq b_0 & \text{(zweisetig)} \end{cases}$$

t-Statistik: $$z = t(\hat{\beta}_0)_{emp(N-J-1)} = \frac{\hat{\beta}_0 - b_0}{se(\hat{\beta}_0)}$$, wobei $$SE(\hat{\beta}_0) = \sqrt{s^2 \left(\frac{1}{N} + \frac{\sum_{n=1}^{N} \overline{x}_n^{\,2}}{\sum_{n=1}^{N} (x_n - \overline{x})^2} \right)}$$

p-Wert: $$PW = \begin{cases} (a) & P(Z \geq z) & \text{(einseitig)} \\ (b) & P(Z \leq z) & \text{(einseitig)} \\ (c) & 2P(Z \geq |z|) & \text{(zweisetig)} \end{cases}$$

wobei: $\hat{\beta}_j$ ist Regressionskoeffizient, und $se(\hat{\beta}_j)$ ist Standardfehler von $\hat{\beta}_j$ im Falle einfacher Regression (siehe Anhang des Kapitels 1), b_j ist wahrer Regressionskoeffizient (unbekannt).

Regressionsanalyse: Prüfung der Regressionskoeffizienten

Hypothesentest fuer β_1

Hypothese: $H_0 : \beta_1 = b_1$,

$$H_1: \begin{cases} (a) & \beta_1 > b_1 \quad (\text{einseitig}) \\ (b) & \beta_1 < b_1 \quad (\text{einseitig}) \\ (c) & \beta_1 \neq b_1 \quad (\text{zweisetig}) \end{cases}$$

t-Statistik: $$z = t(\hat{\beta}_1)_{emp(N-J-1)} = \frac{\hat{\beta}_1 - b_1}{se(\hat{\beta}_1)}, \text{ wobei } SE(\hat{\beta}_1) = \sqrt{\frac{s^2}{\sum_{n=1}^{N}(x_n - \bar{x})^2}}$$

p-Wert : $$PW = \begin{cases} (a) & P(Z \geq z) \quad (\text{einseitig}) \\ (b) & P(Z \leq z) \quad (\text{einseitig}) \\ (c) & 2P(Z \geq |z|) \quad (\text{zweisetig}) \end{cases}$$

wobei: $\hat{\beta}_j$ ist Regressionskoeffizient, und $se(\hat{\beta}_j)$ ist Standardfehler von $\hat{\beta}_j$ im Falle einfacher Regression (siehe Anhang des Kapitels 1), b_j ist wahrer Regressionskoeffizient (unbekannt).

Regressionsanalyse: Prüfung der Regressionskoeffizienten

Wichtig: $|t_{emp}| > t_{theor.}$ H_0 wird verworfen *Einfluss ist signifikant*

$|t_{emp}| \leq t_{theor.}$ H_0 wird nicht verworfen

Ablehnungsregionen:

Hypothese	Kritischer Wert	Ablehnungsregion
$H_0 : \beta_j = 0$, $H_1 : \beta_j > 0$	$c = z_{1-\alpha}$	(c, ∞)
$H_0 : \beta_j = 0$, $H_1 : \beta_j < 0$	$c = z_{\alpha}$	$(-\infty, c)$
$H_0 : \beta_j = 0$, $H_1 : \beta_j \neq 0$	$c = z_{1-\alpha/2}$	$(-\infty, -c), (c, \infty)$

Fuer Inferenz wichtig:

- Grosser Stichprobenumfang: $n \geq 50$ (fuer kleine n geht die Inferenz schief)
- Zufalls-Stichprobe

Anwendungsbeispiel für einfache Regression:

$$t(\hat{\beta}_1)_{emp} = \frac{18{,}881}{9{,}187} = \mathbf{2{,}055}\ ; \quad t_{theor.} = \mathbf{2{,}306}$$ (siehe Abbildung 1.19, S. 75), $\alpha = 0{,}05$, $df = (N - J - 1) = 8$

Regressionsanalyse: Konfidenzintervall des Regressionskoeffizienten

Man benuetzt die t_{n-2}-Verteilung statt der $N(0,1)$, um Konfidenz-Intervalle und p-Werte zu berechnen. Seien $t_{\alpha/2}$ und $t_{1-\alpha/2}$, die $\alpha/2$- und $1-\alpha/2$-Quantile der t-Verteilung. Dann gilt:

$$P(t_{\alpha/2} \le \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \le t_{1-\alpha/2}) = 1-\alpha$$

Ein (standard) Konfidenzintervall fuer β_0 hat die allgemeine Form:

$$P(\hat{\beta}_0 - t_{1-\alpha/2} \cdot SE(\hat{\beta}_0) \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2} \cdot SE(\hat{\beta}_0)) = 1-\alpha$$

wobei t der t-Wert aus der Student-Verteilung ist:

Typischerweise: $\alpha = 0.05$, $t_{0.975} \approx z_{0.975} = 1.96$

Regressionsanalyse: Konfidenzintervall des Regressionskoeffizienten

Ein (standard) Konfidenzintervall fuer β_1 hat die allgemeine Form:

$$P(\hat{\beta}_1 - t_{1-\alpha/2} \cdot SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2} \cdot SE(\hat{\beta}_1)) = 1 - \alpha$$

wobei t der t-Wert aus der Student-Verteilung ist:

Typischerweise: $\alpha = 0.05$, $t_{0.975} \approx z_{0.975} = 1.96$

Anwendungsbeispiel für einfache Regression:

$$18{,}881 - 2{,}306 \cdot 9{,}187 \leq \beta_1 \leq 18{,}881 + 2{,}306 \cdot 9{,}187$$

$$-2{,}304 \leq \beta_1 \leq 40{,}066$$