- Im einfachen linearen Regressionsmodell sind also (neben σ^2) insbesondere β_1 und β_2 Parameter, deren Schätzung für die Quantifizierung des linearen Zusammenhangs zwischen x_i und y_i nötig ist.
- Die Schätzung dieser beiden Parameter führt wieder zum Problem der Suche nach Absolutglied und Steigung einer geeigneten Geradengleichung

$$y = f_{\beta_1,\beta_2}(x) = \beta_1 + \beta_2 \cdot x \ .$$

Satz 10.1 (Satz von Gauß-Markov)

Unter den getroffenen Annahmen liefert die aus dem deskriptiven Ansatz bekannte Verwendung der **KQ-Methode**, *also die Minimierung der Summe der quadrierten vertikalen Abstände zur durch β_1 und β_2 bestimmten Geraden, in Zeichen*

$$\sum_{i=1}^{n}(y_i - (\widehat{\beta}_1 + \widehat{\beta}_2 \cdot x_i))^2 \stackrel{!}{=} \min_{\beta_1,\beta_2\in\mathbb{R}} \sum_{i=1}^{n}(y_i - (\beta_1 + \beta_2 \cdot x_i))^2 \ ,$$

die **beste** *(varianzminimale)* **lineare** *(in y_i)* **erwartungstreue** *Schätzfunktion $\widehat{\beta}_1$ für β_1 bzw. $\widehat{\beta}_2$ für β_2.*

- Dies rechtfertigt letztendlich die Verwendung des Optimalitätskriteriums „Minimierung der quadrierten vertikalen Abstände“.

- Man erhält also — ganz analog zum deskriptiven Ansatz — die folgenden Parameterschätzer:

Parameterschätzer im einfachen linearen Regressionsmodell

$$\widehat{\beta}_2 = \frac{n\left(\sum_{i=1}^n x_i y_i\right) - \left(\sum_{i=1}^n x_i\right)\cdot\left(\sum_{i=1}^n y_i\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{s_{X,Y}}{s_X^2} = r_{X,Y}\cdot\frac{s_Y}{s_X}\ ,$$

$$\widehat{\beta}_1 = \frac{1}{n}\left(\sum_{i=1}^n y_i\right) - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)\cdot\widehat{\beta}_2 = \overline{y} - \overline{x}\widehat{\beta}_2\ .$$

- Wegen der Abhängigkeit von y_i handelt es sich bei $\widehat{\beta}_1$ und $\widehat{\beta}_2$ (wie in der schließenden Statistik gewohnt) um (Realisationen von) *Zufallsvariablen*.
- Die resultierenden vertikalen Abweichungen $\widehat{u}_i := y_i - (\widehat{\beta}_1 + \widehat{\beta}_2\cdot x_i) = y_i - \widehat{y}_i$ der y_i von den auf der Regressionsgeraden liegenden Werten $\widehat{y}_i := \widehat{\beta}_1 + \widehat{\beta}_2\cdot x_i$ nennt man **Residuen**.
- Wie im deskriptiven Ansatz gelten die Beziehungen

$$\sum_{i=1}^n \widehat{u}_i = 0,\quad \sum_{i=1}^n y_i = \sum_{i=1}^n \widehat{y}_i,\quad \sum_{i=1}^n x_i\widehat{u}_i = 0,\quad \sum_{i=1}^n \widehat{y}_i\widehat{u}_i = 0$$

sowie die Varianzzerlegung

$$\frac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^n (\widehat{y}_i - \overline{\widehat{y}})^2 + \frac{1}{n}\sum_{i=1}^n \widehat{u}_i^2\ .$$

Das (multiple) Bestimmtheitsmaß R^2

- Auch im linearen Regressionsmodell wird die Stärke des linearen Zusammenhangs mit dem Anteil der erklärten Varianz an der Gesamtvarianz gemessen und mit

$$R^2 = \frac{\sum_{i=1}^n (\widehat{y}_i - \overline{\widehat{y}})^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^n \widehat{u}_i^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

  bezeichnet. R^2 wird auch **(multiples) Bestimmtheitsmaß** genannt.
- Es gilt $0 \leq R^2 \leq 1$ sowie der (bekannte) Zusammenhang $R^2 = r_{X,Y}^2 = \frac{s_{X,Y}^2}{s_X^2 \cdot s_Y^2}$.
- Größere Werte von R^2 (in der Nähe von 1) sprechen für eine hohe Modellgüte, niedrige Werte (in der Nähe von 0) für eine geringe Modellgüte.

Vorsicht!

s_X^2, s_Y^2 sowie $s_{X,Y}$ bezeichnen in diesem Kapitel die **empirischen** Größen

$$s_X^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2 = \overline{x^2} - \overline{x}^2, \qquad s_Y^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2 = \overline{y^2} - \overline{y}^2$$

$$\text{und } s_{X,Y} = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}) \cdot (y_i - \overline{y}) = \overline{xy} - \overline{x} \cdot \overline{y} \ .$$

Beispiel: Ausgaben in Abhängigkeit vom Einkommen (I)

- Es wird angenommen, dass die Ausgaben eines Haushalts für Nahrungs- und Genussmittel y_i linear vom jeweiligen Haushaltseinkommen x_i (jeweils in 100 €) in der Form

$$y_i = \beta_1 + \beta_2 \cdot x_i + u_i, \qquad u_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \qquad i \in \{1, \ldots, n\}$$

  abhängen. Für $n = 7$ Haushalte beobachte man nun neben dem Einkommen x_i auch die (Realisation der) Ausgaben für Nahrungs- und Genussmittel y_i wie folgt:

Haushalt i	1	2	3	4	5	6	7
Einkommen x_i	35	49	21	39	15	28	25
NuG-Ausgaben y_i	9	15	7	11	5	8	9

- Mit Hilfe dieser Stichprobeninformation sollen nun die Parameter β_1 und β_2 der linearen Modellbeziehung geschätzt sowie die Werte $\widehat{y}_i$, die Residuen $\widehat{u}_i$ und das Bestimmtheitsmaß R^2 bestimmt werden.

- Berechnete (deskriptive/empirische) Größen:

$$\overline{x} = 30.28571 \quad \overline{y} = 9.14286 \quad \overline{x^2} = 1031.71429 \quad \overline{y^2} = 92.28571$$
$$s_X^2 = 114.4901 \quad s_Y^2 = 8.6938 \quad s_{X,Y} = 30.2449 \quad r_{X,Y} = 0.9587$$

- Damit erhält man die Parameterschätzer $\widehat{\beta}_1$ und $\widehat{\beta}_2$ als

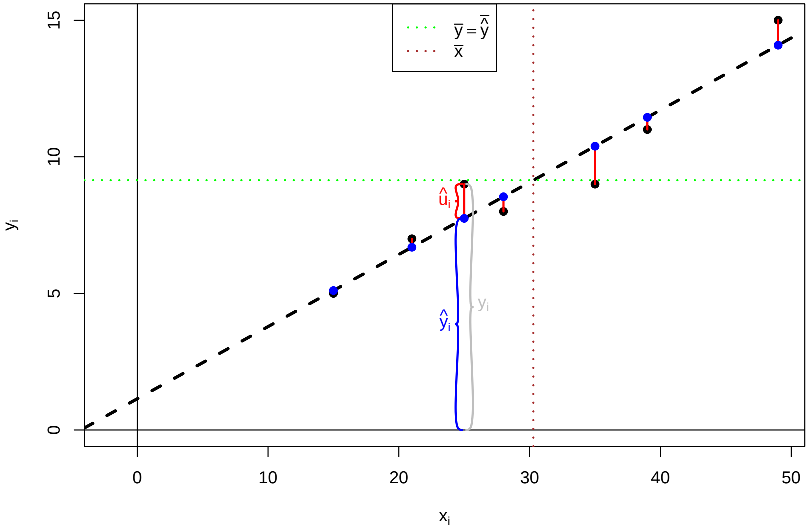
$$\widehat{\beta}_2 = \frac{s_{X,Y}}{s_X^2} = \frac{30.2449}{114.4901} = 0.26417$$
$$\widehat{\beta}_1 = \overline{y} - \widehat{\beta}_2 \cdot \overline{x} = 9.14286 - 0.26417 \cdot 30.28571 = 1.14228 \ .$$

- Als Bestimmtheitsmaß erhält man $R^2 = r_{X,Y}^2 = 0.9587^2 = 0.9191$.
- Für $\widehat{y}_i$ und $\widehat{u}_i$ erhält man durch Einsetzen ($\widehat{y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 \cdot x_i$, $\widehat{u}_i = y_i - \widehat{y}_i$):

i	1	2	3	4	5	6	7
x_i	35	49	21	39	15	28	25
y_i	9	15	7	11	5	8	9
$\widehat{y}_i$	10.39	14.09	6.69	11.44	5.1	8.54	7.75
$\widehat{u}_i$	−1.39	0.91	0.31	−0.44	−0.1	−0.54	1.25

Grafik: Ausgaben in Abhängigkeit vom Einkommen

$\widehat{\beta}_1 = 1.14228$, $\widehat{\beta}_2 = 0.26417$, $R^2 = 0.9191$



Eigenschaften der Schätzfunktionen $\widehat{\beta}_1$ und $\widehat{\beta}_2$

- $\widehat{\beta}_1$ und $\widehat{\beta}_2$ sind **linear in** y_i, man kann genauer zeigen:

$$\widehat{\beta}_1 = \sum_{i=1}^{n} \frac{\overline{x^2} - \overline{x} \cdot x_i}{n s_X^2} \cdot y_i \qquad \text{und} \qquad \widehat{\beta}_2 = \sum_{i=1}^{n} \frac{x_i - \overline{x}}{n s_X^2} \cdot y_i$$

- $\widehat{\beta}_1$ und $\widehat{\beta}_2$ sind **erwartungstreu für** β_1 **und** β_2, denn wegen $\mathrm{E}(u_i) = 0$ gilt
  - $\mathrm{E}(y_i) = \beta_1 + \beta_2 \cdot x_i + \mathrm{E}(u_i) = \beta_1 + \beta_2 \cdot x_i$,
  - $\mathrm{E}(\overline{y}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(y_i) = \frac{1}{n}\sum_{i=1}^{n}(\beta_1 + \beta_2 \cdot x_i) = \beta_1 + \beta_2 \cdot \overline{x}$,
  - $\mathrm{E}(\overline{xy}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) = \frac{1}{n}\sum_{i=1}^{n} x_i(\beta_1 + \beta_2 \cdot x_i) = \beta_1 \cdot \overline{x} + \beta_2 \cdot \overline{x^2}$

  und damit

$$\mathrm{E}(\widehat{\beta}_2) = \mathrm{E}\left(\frac{s_{X,Y}}{s_X^2}\right) = \frac{\mathrm{E}\left(\overline{xy} - \overline{x} \cdot \overline{y}\right)}{s_X^2} = \frac{\mathrm{E}(\overline{xy}) - \overline{x} \cdot \mathrm{E}(\overline{y})}{s_X^2}$$

$$= \frac{\beta_1 \cdot \overline{x} + \beta_2 \cdot \overline{x^2} - \overline{x} \cdot (\beta_1 + \beta_2 \cdot \overline{x})}{s_X^2} = \frac{\beta_2 \cdot (\overline{x^2} - \overline{x}^2)}{s_X^2} = \beta_2$$

  sowie

$$\mathrm{E}(\widehat{\beta}_1) = \mathrm{E}(\overline{y} - \overline{x}\widehat{\beta}_2) = \mathrm{E}(\overline{y}) - \overline{x}\,\mathrm{E}(\widehat{\beta}_2) = \beta_1 + \beta_2 \cdot \overline{x} - \overline{x} \cdot \beta_2 = \beta_1 \; .$$

(Diese Eigenschaften folgen bereits mit dem Satz von Gauß-Markov.)

- Für die Varianzen der Schätzfunktionen erhält man:

$$\mathrm{Var}(\widehat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sigma^2}{n \cdot (\overline{x^2} - \overline{x}^2)} = \frac{\sigma^2}{n \cdot s_X^2}$$

$$\mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{n} \cdot \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sigma^2 \cdot \overline{x^2}}{n \cdot (\overline{x^2} - \overline{x}^2)} = \frac{\sigma^2 \cdot \overline{x^2}}{n \cdot s_X^2}$$

  Diese hängen von der unbekannten Varianz σ^2 der u_i ab.
- Eine erwartungstreue Schätzfunktion für σ^2 ist gegeben durch

$$\begin{aligned}\widehat{\sigma^2} := \widehat{\mathrm{Var}(u_i)} &= \frac{1}{n-2} \sum_{i=1}^n \widehat{u}_i^2 \\ &= \frac{n}{n-2} \cdot s_Y^2 \cdot (1 - R^2) = \frac{n}{n-2} \cdot (s_Y^2 - \widehat{\beta}_2 \cdot s_{X,Y})\end{aligned}$$

- Die positive Wurzel $\widehat{\sigma} = +\sqrt{\widehat{\sigma^2}}$ dieser Schätzfunktion heißt auch **S**tandard **E**rror of the **R**egression (SER) oder **residual standard error**.

- Einsetzen des Schätzers $\widehat{\sigma^2}$ für σ^2 liefert die geschätzten Varianzen der Parameterschätzer

$$\widehat{\sigma^2}_{\widehat{\beta}_2} := \widehat{\mathrm{Var}(\widehat{\beta}_2)} = \frac{\widehat{\sigma^2}}{n \cdot (\overline{x^2} - \overline{x}^2)} = \frac{\widehat{\sigma^2}}{n \cdot s_X^2} = \frac{s_Y^2 - \widehat{\beta}_2 \cdot s_{X,Y}}{(n-2) \cdot s_X^2}$$

und

$$\widehat{\sigma^2}_{\widehat{\beta}_1} := \widehat{\mathrm{Var}(\widehat{\beta}_1)} = \frac{\widehat{\sigma^2} \cdot \overline{x^2}}{n \cdot (\overline{x^2} - \overline{x}^2)} = \frac{\widehat{\sigma^2} \cdot \overline{x^2}}{n \cdot s_X^2} = \frac{(s_Y^2 - \widehat{\beta}_2 \cdot s_{X,Y}) \cdot \overline{x^2}}{(n-2) \cdot s_X^2} .$$

- Die positiven Wurzeln $\widehat{\sigma}_{\widehat{\beta}_1} = \sqrt{\widehat{\sigma^2}_{\widehat{\beta}_1}}$ und $\widehat{\sigma}_{\widehat{\beta}_2} = \sqrt{\widehat{\sigma^2}_{\widehat{\beta}_2}}$ dieser geschätzten Varianzen werden wie üblich als (geschätzte) **Standardfehler** von $\widehat{\beta}_1$ und $\widehat{\beta}_2$ bezeichnet.
- Trifft man eine weitergehende Verteilungannahme für u_i und damit für y_i, so lassen sich auch die Verteilungen von $\widehat{\beta}_1$ und $\widehat{\beta}_2$ weiter untersuchen und zur Konstruktion von Tests, Konfidenzintervallen und *Prognoseintervallen* verwenden.

Konfidenzintervalle und Tests

unter Normalverteilungsannahme für u_i

- Häufig nimmt man für die Störgrößen an, dass speziell

$$u_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

  gilt, d.h. dass alle u_i (für $i \in \{1, \dots, n\}$) unabhängig identisch normalverteilt sind mit Erwartungswert 0 und (unbekannter) Varianz σ^2.
- In diesem Fall sind offensichtlich auch $y_1, \dots, y_n$ stochastisch unabhängig und jeweils normalverteilt mit Erwartungswert $\mathrm{E}(y_i) = \beta_1 + \beta_2 \cdot x_i$ und Varianz $\mathrm{Var}(y_i) = \sigma^2$.
- Da $\widehat{\beta}_1$ und $\widehat{\beta}_2$ linear in y_i sind, folgt insgesamt mit den bereits berechneten Momenten von $\widehat{\beta}_1$ und $\widehat{\beta}_2$:

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2 \cdot \overline{x^2}}{n \cdot s_X^2}\right) \qquad \text{und} \qquad \widehat{\beta}_2 \sim N\left(\beta_2, \frac{\sigma^2}{n \cdot s_X^2}\right)$$

Konfidenzintervalle

unter Normalverteilungsannahme für u_i

- Da σ^2 unbekannt ist, ist für Anwendungen wesentlich relevanter, dass im Falle unabhängig identisch normalverteilter Störgrößen u_i mit den Schätzfunktionen $\widehat{\sigma^2}_{\widehat{\beta}_1}$ für $\mathrm{Var}(\widehat{\beta}_1)$ und $\widehat{\sigma^2}_{\widehat{\beta}_2}$ für $\mathrm{Var}(\widehat{\beta}_2)$ gilt:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\widehat{\beta}_1}} \sim t(n-2) \qquad \text{und} \qquad \frac{\widehat{\beta}_2 - \beta_2}{\widehat{\sigma}_{\widehat{\beta}_2}} \sim t(n-2)$$

- Hieraus erhält man unmittelbar die „Formeln"

$$\left[\widehat{\beta}_1 - t_{n-2;1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{\widehat{\beta}_1}, \widehat{\beta}_1 + t_{n-2;1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{\widehat{\beta}_1}\right]$$

für (symmetrische) Konfidenzintervalle zur Vertrauenswahrscheinlichkeit $1-\alpha$ für β_1 bzw.

$$\left[\widehat{\beta}_2 - t_{n-2;1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{\widehat{\beta}_2}, \widehat{\beta}_2 + t_{n-2;1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{\widehat{\beta}_2}\right]$$

für (symmetrische) Konfidenzintervalle zur Vertrauenswahrscheinlichkeit $1-\alpha$ für β_2.

Beispiel: Ausgaben in Abhängigkeit vom Einkommen (II)

- Im bereits erläuterten Beispiel erhält man als Schätzwert für σ^2:

$$\widehat{\sigma^2} = \frac{n \cdot (s_Y^2 - \widehat{\beta}_2 \cdot s_{X,Y})}{n-2} = \frac{7 \cdot (8.6938 - 0.26417 \cdot 30.2449)}{7-2} = 0.9856$$

- Die (geschätzten) Standardfehler für $\widehat{\beta}_1$ und $\widehat{\beta}_2$ sind damit

$$\widehat{\sigma}_{\widehat{\beta}_1} = \sqrt{\frac{\widehat{\sigma^2} \cdot \overline{x^2}}{n \cdot s_X^2}} = \sqrt{\frac{0.9856 \cdot 1031.71429}{7 \cdot 114.4901}} = 1.1264 \,,$$

$$\widehat{\sigma}_{\widehat{\beta}_2} = \sqrt{\frac{\widehat{\sigma^2}}{n \cdot s_X^2}} = \sqrt{\frac{0.9856}{7 \cdot 114.4901}} = 0.0351 \,.$$

- Für $\alpha = 0.05$ erhält man mit $t_{n-2;1-\frac{\alpha}{2}} = t_{5;0.975} = 2.571$ für β_1 also

$$[1.14228 - 2.571 \cdot 1.1264, 1.14228 + 2.571 \cdot 1.1264] = [-1.7537, 4.0383]$$

als Konfidenzintervall zur Vertrauenswahrscheinlichkeit $1 - \alpha = 0.95$ bzw.

$$[0.26417 - 2.571 \cdot 0.0351, 0.26417 + 2.571 \cdot 0.0351] = [0.1739, 0.3544]$$

als Konfidenzintervall zur Vertrauenswahrscheinlichkeit $1 - \alpha = 0.95$ für β_2.

Hypothesentests

unter Normalverteilungsannahme für u_i

- Genauso lassen sich unter der Normalverteilungsannahme (exakte) t-Tests für die Parameter β_1 und β_2 konstruieren.
- Trotz unterschiedlicher Problemstellung weisen die Tests Ähnlichkeiten zum t-Test für den Mittelwert einer normalverteilten Zufallsvariablen bei unbekannter Varianz auf.
- Untersucht werden können die Hypothesenpaare

$$\begin{array}{ccc} H_0 : \beta_1 = \beta_1^0 & H_0 : \beta_1 \leq \beta_1^0 & H_0 : \beta_1 \geq \beta_1^0 \\ \text{gegen} & \text{gegen} & \text{gegen} \\ H_1 : \beta_1 \neq \beta_1^0 & H_1 : \beta_1 > \beta_1^0 & H_1 : \beta_1 < \beta_1^0 \end{array}$$

bzw.

$$\begin{array}{ccc} H_0 : \beta_2 = \beta_2^0 & H_0 : \beta_2 \leq \beta_2^0 & H_0 : \beta_2 \geq \beta_2^0 \\ \text{gegen} & \text{gegen} & \text{gegen} \\ H_1 : \beta_2 \neq \beta_2^0 & H_1 : \beta_2 > \beta_2^0 & H_1 : \beta_2 < \beta_2^0 \end{array}$$

- Besonders anwendungsrelevant sind Tests auf die „Signifikanz“ der Parameter (insbesondere β_2), die den zweiseitigen Tests mit $\beta_1^0 = 0$ bzw. $\beta_2^0 = 0$ entsprechen.

Zusammenfassung: t-Test für den Parameter β_1

im einfachen linearen Regressionsmodell mit Normalverteilungsannahme

Anwendungs-voraussetzungen	exakt: $y_i = \beta_1 + \beta_2 \cdot x_i + u_i$ mit $u_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ für $i \in \{1, \dots, n\}$, σ^2 unbekannt, $x_1, \dots, x_n$ deterministisch und bekannt, Realisation $y_1, \dots, y_n$ beobachtet		
Nullhypothese Gegenhypothese	$H_0 : \beta_1 = \beta_1^0$ $H_1 : \beta_1 \neq \beta_1^0$	$H_0 : \beta_1 \leq \beta_1^0$ $H_1 : \beta_1 > \beta_1^0$	$H_0 : \beta_1 \geq \beta_1^0$ $H_1 : \beta_1 < \beta_1^0$
Teststatistik Verteilung (H_0)	$t = \frac{\widehat{\beta}_1 - \beta_1^0}{\widehat{\sigma}_{\widehat{\beta}_1}}$ t für $\beta_1 = \beta_1^0$ $t(n-2)$-verteilt		
Benötigte Größen	$\widehat{\beta}_2 = \frac{s_{X,Y}}{s_X^2}, \widehat{\beta}_1 = \overline{y} - \widehat{\beta}_2 \cdot \overline{x}, \widehat{\sigma}_{\widehat{\beta}_1} = \sqrt{\frac{(s_Y^2 - \widehat{\beta}_2 \cdot s_{X,Y}) \cdot \overline{x^2}}{(n-2) \cdot s_X^2}}$		
Kritischer Bereich zum Niveau α	$(-\infty, -t_{n-2;1-\frac{\alpha}{2}}) \cup (t_{n-2;1-\frac{\alpha}{2}}, \infty)$	$(t_{n-2;1-\alpha}, \infty)$	$(-\infty, -t_{n-2;1-\alpha})$
p-Wert	$2 \cdot (1 - F_{t(n-2)}(\lvert t \rvert))$	$1 - F_{t(n-2)}(t)$	$F_{t(n-2)}(t)$

Zusammenfassung: t-Test für den Parameter β_2

im einfachen linearen Regressionsmodell mit Normalverteilungsannahme

Anwendungsvoraussetzungen	exakt: $y_i = \beta_1 + \beta_2 \cdot x_i + u_i$ mit $u_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ für $i \in \{1, \ldots, n\}$, σ^2 unbekannt, $x_1, \ldots, x_n$ deterministisch und bekannt, Realisation $y_1, \ldots, y_n$ beobachtet		
Nullhypothese Gegenhypothese	$H_0 : \beta_2 = \beta_2^0$ $H_1 : \beta_2 \neq \beta_2^0$	$H_0 : \beta_2 \leq \beta_2^0$ $H_1 : \beta_2 > \beta_2^0$	$H_0 : \beta_2 \geq \beta_2^0$ $H_1 : \beta_2 < \beta_2^0$
Teststatistik Verteilung (H_0)	$t = \frac{\widehat{\beta}_2 - \beta_2^0}{\widehat{\sigma}_{\widehat{\beta}_2}}$ t für $\beta_2 = \beta_2^0$ $t(n-2)$-verteilt		
Benötigte Größen	$\widehat{\beta}_2 = \frac{s_{X,Y}}{s_X^2}, \widehat{\sigma}_{\widehat{\beta}_2} = \sqrt{\frac{s_Y^2 - \widehat{\beta}_2 \cdot s_{X,Y}}{(n-2) \cdot s_X^2}}$		
Kritischer Bereich zum Niveau α	$(-\infty, -t_{n-2;1-\frac{\alpha}{2}}) \cup (t_{n-2;1-\frac{\alpha}{2}}, \infty)$	$(t_{n-2;1-\alpha}, \infty)$	$(-\infty, -t_{n-2;1-\alpha})$
p-Wert	$2 \cdot (1 - F_{t(n-2)}(\lvert t \rvert))$	$1 - F_{t(n-2)}(t)$	$F_{t(n-2)}(t)$

Beispiel: Ausgaben in Abhängigkeit vom Einkommen (III)

- Im bereits erläuterten Beispiel soll zum Signifikanzniveau $\alpha = 0.05$ getestet werden, ob β_1 signifikant von Null verschieden ist. Geeigneter Test: t-**Test für den Regressionsparameter** β_1
  1. **Hypothesen:**
     $$H_0 : \beta_1 = 0 \qquad \text{gegen} \qquad H_1 : \beta_1 \neq 0$$
  2. **Teststatistik:**
     $$t = \frac{\widehat{\beta}_1 - 0}{\widehat{\sigma}_{\widehat{\beta}_1}} \text{ ist unter } H_0 \text{ (für } \beta_1 = 0\text{) } t(n-2)\text{-verteilt.}$$
  3. **Kritischer Bereich zum Niveau $\alpha = 0.05$:**
     $$K = (-\infty, -t_{n-2;1-\frac{\alpha}{2}}) \cup (t_{n-2;1-\frac{\alpha}{2}}, +\infty) = (-\infty, -t_{5;0.975}) \cup (t_{5;0.975}, +\infty)$$
     $$= (-\infty, -2.571) \cup (2.571, +\infty)$$
  4. **Berechnung der realisierten Teststatistik:**
     $$t = \frac{\widehat{\beta}_1 - 0}{\widehat{\sigma}_{\widehat{\beta}_1}} = \frac{1.14228 - 0}{1.1264} = 1.014$$
  5. **Entscheidung:**
     $$t = 1.014 \notin (-\infty, -2.571) \cup (2.571, +\infty) = K \ \Rightarrow \ H_0 \text{ wird nicht abgelehnt!}$$
     (p-Wert: $2 - 2 \cdot F_{t(5)}(|t|) = 2 - 2 \cdot F_{t(5)}(|1.014|) = 2 - 2 \cdot 0.8215 = 0.357$)

  Der Test kann für β_1 keine signifikante Abweichung von Null feststellen.

Beispiel: Ausgaben in Abhängigkeit vom Einkommen (IV)

- Nun soll zum Signifikanzniveau $\alpha = 0.01$ getestet werden, ob β_2 **positiv** ist.
  Geeigneter Test:
  ***t*-Test für den Regressionsparameter β_2**

  1. **Hypothesen:**
     $$H_0 : \beta_2 \leq 0 \qquad \text{gegen} \qquad H_1 : \beta_2 > 0$$
  2. **Teststatistik:**
     $$t = \frac{\widehat{\beta}_2 - 0}{\widehat{\sigma}_{\widehat{\beta}_2}} \text{ ist unter } H_0 \text{ (für } \beta_2 = 0) \ t(n-2)\text{-verteilt.}$$
  3. **Kritischer Bereich zum Niveau $\alpha = 0.01$:**
     $$K = (t_{n-2;1-\alpha}, +\infty) = (t_{5;0.99}, +\infty) = (3.365, +\infty)$$
  4. **Berechnung der realisierten Teststatistik:**
     $$t = \frac{\widehat{\beta}_2 - 0}{\widehat{\sigma}_{\widehat{\beta}_2}} = \frac{0.26417 - 0}{0.0351} = 7.5262$$
  5. **Entscheidung:**
     $$t = 7.5262 \in (3.365, +\infty) = K \quad \Rightarrow \quad H_0 \text{ wird abgelehnt!}$$
     (p-Wert: $1 - F_{t(5)}(t) = 1 - F_{t(5)}(7.5262) = 1 - 0.9997 = 0.0003$)

  Der Test stellt fest, dass β_2 signifikant positiv ist.

Punkt- und Intervallprognosen

im einfachen linearen Regressionsmodell mit Normalverteilungsannahme

- Neben Konfidenzintervallen und Tests für die Parameter β_1 und β_2 in linearen Regressionsmodellen vor allem **Prognosen** wichtige Anwendung.
- Zur Erstellung von Prognosen: Erweiterung der Modellannahme

$$y_i = \beta_1 + \beta_2 \cdot x_i + u_i, \quad u_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \quad i \in \{1, \ldots, n\}$$

  auf (zumindest) einen weiteren, hier mit (x_0, y_0) bezeichneten Datenpunkt, bei dem jedoch y_0 **nicht** beobachtet wird, sondern lediglich der Wert des Regressors x_0 bekannt ist.
- *Ziel:* „Schätzung" (Prognose) von $y_0 = \beta_1 + \beta_2 \cdot x_0 + u_0$ bzw. $\mathrm{E}(y_0) = \beta_1 + \beta_2 \cdot x_0$ auf Grundlage von x_0.
- Wegen $\mathrm{E}(u_0) = 0$ und der Erwartungstreue von $\widehat{\beta}_1$ für β_1 bzw. $\widehat{\beta}_2$ für β_2 ist

$$\widehat{y}_0 := \widehat{\beta}_1 + \widehat{\beta}_2 \cdot x_0 =: \widehat{\mathrm{E}(y_0)}$$

  offensichtlich erwartungstreu für y_0 bzw. $\mathrm{E}(y_0)$ gegeben x_0.
- $\widehat{y}_0$ bzw. $\widehat{\mathrm{E}(y_0)}$ wird auch **(bedingte) Punktprognose für** y_0 **bzw.** $\mathrm{E}(y_0)$ **gegeben** x_0 genannt.

Prognosefehler

- Zur Beurteilung der Genauigkeit der Prognosen:
  Untersuchung der sogenannten Prognosefehler

$$\widehat{y}_0 - y_0 \qquad \text{bzw.} \qquad \widehat{\mathrm{E}(y_0)} - \mathrm{E}(y_0)\ .$$

- Qualitativer Unterschied:
  - Prognosefehler

    $$\widehat{\mathrm{E}(y_0)} - \mathrm{E}(y_0) = \widehat{\beta}_1 + \widehat{\beta}_2 \cdot x_0 - (\beta_1 + \beta_2 \cdot x_0) = (\widehat{\beta}_1 - \beta_1) + (\widehat{\beta}_2 - \beta_2) \cdot x_0$$

    resultiert **nur** aus Fehler bei der Schätzung von β_1 bzw. β_2 durch $\widehat{\beta}_1$ bzw. $\widehat{\beta}_2$.
  - Prognosefehler

    $$\widehat{y}_0 - y_0 = \widehat{\beta}_1 + \widehat{\beta}_2 \cdot x_0 - (\beta_1 + \beta_2 \cdot x_0 + u_0) = (\widehat{\beta}_1 - \beta_1) + (\widehat{\beta}_2 - \beta_2) \cdot x_0 - u_0$$

    ist Kombination von Schätzfehlern (für β_1 und β_2) sowie zufälliger Schwankung von $u_0 \sim N(0, \sigma^2)$.
- Zunächst: Untersuchung von $e_E := \widehat{\mathrm{E}(y_0)} - \mathrm{E}(y_0)$

- Wegen der Erwartungstreue stimmen mittlerer quadratischer (Prognose-) Fehler und Varianz von $e_E = \widehat{\mathrm{E}(y_0)} - \mathrm{E}(y_0)$ überein und man erhält

$$\begin{aligned}\mathrm{Var}(\widehat{\mathrm{E}(y_0)} - \mathrm{E}(y_0)) = \mathrm{Var}(\widehat{\mathrm{E}(y_0)}) &= \mathrm{Var}(\widehat{\beta}_1 + \widehat{\beta}_2 \cdot x_0) \\ &= \mathrm{Var}(\widehat{\beta}_1) + x_0^2 \,\mathrm{Var}(\widehat{\beta}_2) + 2 \cdot x_0 \cdot \mathrm{Cov}(\widehat{\beta}_1, \widehat{\beta}_2).\end{aligned}$$

- Es kann gezeigt werden, dass für die Kovarianz von $\widehat{\beta}_1$ und $\widehat{\beta}_2$ gilt:

$$\mathrm{Cov}(\widehat{\beta}_1, \widehat{\beta}_2) = -\sigma^2 \cdot \frac{\overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = -\sigma^2 \cdot \frac{\overline{x}}{n \cdot s_X^2}$$

- Insgesamt berechnet man so die Varianz des Prognosefehlers

$$\begin{aligned}\sigma_{e_E}^2 := \mathrm{Var}(e_E) &= \frac{\sigma^2 \cdot \overline{x^2}}{n \cdot s_X^2} + x_0^2 \cdot \frac{\sigma^2}{n \cdot s_X^2} - 2 \cdot x_0 \cdot \frac{\sigma^2 \cdot \overline{x}}{n \cdot s_X^2} \\ &= \sigma^2 \cdot \frac{\overline{x^2} + x_0^2 - 2 \cdot x_0 \cdot \overline{x}}{n \cdot s_X^2} \\ &= \sigma^2 \cdot \frac{(\overline{x^2} - \overline{x}^2) + (\overline{x}^2 + x_0^2 - 2 \cdot x_0 \cdot \overline{x})}{n \cdot s_X^2} \\ &= \sigma^2 \cdot \frac{s_X^2 + (x_0 - \overline{x})^2}{n \cdot s_X^2} = \sigma^2 \cdot \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n \cdot s_X^2}\right) .\end{aligned}$$

- Die Linearität von $\widehat{\beta}_1$ und $\widehat{\beta}_2$ (in y_i) überträgt sich (natürlich) auch auf $\widehat{\mathrm{E}(y_0)}$, damit gilt offensichtlich

$$e_E = \widehat{\mathrm{E}(y_0)} - \mathrm{E}(y_0) \sim N\left(0, \sigma^2_{e_E}\right) \qquad \text{bzw.} \qquad \frac{\widehat{\mathrm{E}(y_0)} - \mathrm{E}(y_0)}{\sigma_{e_E}} \sim N(0,1)\ .$$

- Da σ^2 unbekannt ist, erhält man durch Ersetzen von σ^2 durch die erwartungstreue Schätzfunktion $\widehat{\sigma^2}$ die geschätzte Varianz

$$\widehat{\sigma^2}_{e_E} := \widehat{\mathrm{Var}}(e_E) = \widehat{\sigma^2} \cdot \frac{s_X^2 + (x_0 - \overline{x})^2}{n \cdot s_X^2} = \widehat{\sigma^2} \cdot \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n \cdot s_X^2}\right)$$

von $\widehat{\mathrm{E}(y_0)}$ und damit die praktisch wesentlich relevantere Verteilungsaussage

$$\frac{e_E}{\widehat{\sigma}_{e_E}} = \frac{\widehat{\mathrm{E}(y_0)} - \mathrm{E}(y_0)}{\widehat{\sigma}_{e_E}} \sim t(n-2)\ ,$$

aus der sich in bekannter Weise (symmetrische) Konfidenzintervalle (und Tests) konstruieren lassen.

Prognoseintervalle für $\mathrm{E}(y_0)$ gegeben x_0

- Intervallprognosen zur Vertrauenswahrscheinlichkeit $1-\alpha$ erhält man also als Konfidenzintervalle zum Konfidenzniveau $1-\alpha$ für $\mathrm{E}(y_0)$ in der Form

$$\left[\widehat{\mathrm{E}(y_0)} - t_{n-2;1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{e_E}\, , \widehat{\mathrm{E}(y_0)} + t_{n-2;1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{e_E}\right]$$
$$= \left[(\widehat{\beta}_1 + \widehat{\beta}_2 \cdot x_0) - t_{n-2;1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{e_E}\, , (\widehat{\beta}_1 + \widehat{\beta}_2 \cdot x_0) + t_{n-2;1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{e_E}\right] .$$

- Im Beispiel (Ausgaben in Abhängigkeit vom Einkommen) erhält man zu gegebenem $x_0 = 38$ (in 100 €)

$$\widehat{\sigma^2}_{e_E} = \widehat{\sigma^2} \cdot \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n \cdot s_X^2}\right) = 0.9856 \cdot \left(\frac{1}{7} + \frac{(38 - 30.28571)^2}{7 \cdot 114.4901}\right) = 0.214$$

die Punktprognose $\widehat{\mathrm{E}(y_0)} = \widehat{\beta}_1 + \widehat{\beta}_2 \cdot x_0 = 1.14228 + 0.26417 \cdot 38 = 11.1807$ (in 100 €) sowie die Intervallprognose zur Vertrauenswahrscheinlichkeit 0.95

$$\left[11.1807 - 2.571 \cdot \sqrt{0.214}\, , 11.1807 + 2.571 \cdot \sqrt{0.214}\right]$$
$$= [9.9914\, , 12.37] \text{ (in 100 €)} .$$

Prognosefehler $e_0 := \widehat{y_0} - y_0$

- *Nun:* Untersuchung des Prognosefehlers $e_0 := \widehat{y_0} - y_0$
- Offensichtlich gilt für $e_0 = \widehat{y_0} - y_0$ die Zerlegung

$$\begin{aligned}\widehat{y_0} - y_0 &= \underbrace{(\widehat{\beta}_1 + \widehat{\beta}_2 \cdot x_0)}_{=\widehat{\mathrm{E}(y_0)}} - (\underbrace{\beta_1 + \beta_2 \cdot x_0}_{=\mathrm{E}(y_0)} + u_0) \\ &= \underbrace{\widehat{\mathrm{E}(y_0)} - \mathrm{E}(y_0)}_{\substack{\text{Fehler aus Schätzung von} \\ \beta_1 \text{ und } \beta_2}} - \underbrace{u_0}_{\substack{\text{zufällige Schwankung} \\ \text{der Störgröße}}} \quad .\end{aligned}$$

- $\widehat{\mathrm{E}(y_0)}$ hängt nur von $u_1, \ldots, u_n$ ab (über $y_1, \ldots, y_n$ bzw. $\widehat{\beta}_1$ und $\widehat{\beta}_2$) und ist wegen der Annahme $u_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ **unabhängig** von u_0.
- Damit sind die beiden Bestandteile des Prognosefehlers insbesondere auch unkorreliert und man erhält:

$$\begin{aligned}\sigma_{e_0}^2 := \mathrm{Var}(\widehat{y_0} - y_0) &= \mathrm{Var}(\widehat{\mathrm{E}(y_0)} - \mathrm{E}(y_0)) + \mathrm{Var}(u_0) \\ &= \sigma^2 \cdot \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n \cdot s_X^2}\right) + \sigma^2 = \sigma^2 \cdot \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n \cdot s_X^2}\right)\end{aligned}$$

- Aus der Unkorreliertheit der beiden Komponenten des Prognosefehlers folgt auch sofort die Normalverteilungseigenschaft des Prognosefehlers $e_0 = y_0 - \widehat{y_0}$, genauer gilt:

$$e_0 = \widehat{y_0} - y_0 \sim N\left(0, \sigma_{e_0}^2\right) \qquad \text{bzw.} \qquad \frac{\widehat{y_0} - y_0}{\sigma_{e_0}} \sim N(0,1)\ .$$

- Wieder muss σ^2 durch $\widehat{\sigma^2}$ ersetzt werden, um mit Hilfe der geschätzen Varianz

$$\widehat{\sigma^2}_{e_0} := \widehat{\text{Var}}(\widehat{y_0} - y_0) = \widehat{\sigma^2} \cdot \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n \cdot s_X^2}\right)$$

des Prognosefehlers die für die Praxis relevante Verteilungsaussage

$$\frac{e_0}{\widehat{\sigma}_{e_0}} = \frac{\widehat{y_0} - y_0}{\widehat{\sigma}_{e_0}} \sim t(n-2)\ ,$$

zu erhalten, aus der sich dann wieder Prognoseintervalle konstruieren lassen.