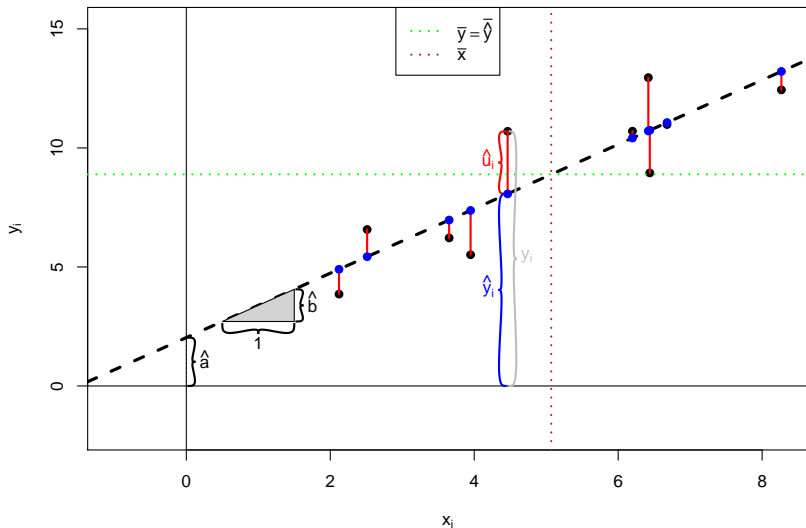


# Beispiel: Regressionsgerade mit Zerlegung $y_i = \hat{y}_i + \hat{u}_i$

$$\hat{a} = 2.03, \hat{b} = 1.35, \sum_{i=1}^n (\hat{u}_i)^2 = 22.25$$



## Beispiel: Berechnung von $\hat{a}$ und $\hat{b}$

- Daten im Beispiel:

$i$	1	2	3	4	5	6	7	8	9	10
$x_i$	2.51	8.27	4.46	3.95	6.42	6.44	2.12	3.65	6.2	6.68
$y_i$	6.57	12.44	10.7	5.51	12.95	8.95	3.86	6.22	10.7	10.98

- Berechnete (deskriptive/empirische) Größen:

$$\begin{aligned} \bar{x} &= 5.0703 & \bar{y} &= 8.8889 & \overline{x^2} &= 29.3729 & \overline{y^2} &= 87.9398 \\ s_X^2 &= 3.665 & s_Y^2 &= 8.927 & s_{X,Y} &= 4.956 & r_{X,Y} &= 0.866 \end{aligned}$$

- Damit erhält man Absolutglied  $\hat{a}$  und Steigung  $\hat{b}$  als

$$\hat{b} = \frac{s_{X,Y}}{s_X^2} = \frac{4.956}{3.665} = 1.352$$

$$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x} = 8.8889 - 1.352 \cdot 5.0703 = 2.03$$

und damit die Regressionsgerade

$$y = f(x) = 2.03 + 1.352 \cdot x .$$

- *Bisher: rein deskriptive Betrachtung linearer Zusammenhänge*
- Bereits erläutert/bekannt: Korrelation  $\neq$  Kausalität:  
Aus einem beobachteten (linearen) Zusammenhang zwischen zwei Merkmalen lässt sich **nicht** schließen, dass der Wert eines Merkmals den des anderen beeinflusst.
- Bereits durch die Symmetrieeigenschaft  $r_{X,Y} = r_{Y,X}$  bei der Berechnung von Pearsonschen Korrelationskoeffizienten wird klar, dass diese Kennzahl alleine auch keine Wirkungsrichtung erkennen lassen **kann**.
- *Nun: statistische Modelle für lineare Zusammenhänge*
- **Keine** symmetrische Behandlung von  $X$  und  $Y$  mehr, sondern:
  - ▶ Interpretation von  $X$  („Regressor“) als **erklärende deterministische** Variable.
  - ▶ Interpretation von  $Y$  („Regressand“) als **abhängige, zu erklärende** (Zufalls-)Variable.
- Es wird angenommen, dass  $Y$  in linearer Form von  $X$  abhängt, diese Abhängigkeit jedoch nicht „perfekt“ ist, sondern durch zufällige Einflüsse „gestört“ wird.
- Anwendung in Experimenten: Festlegung von  $X$  durch Versuchsplaner, Untersuchung des Effekts auf  $Y$
- Damit auch Kausalitätsanalysen möglich!

# Das einfache lineare Regressionsmodell

- Es wird genauer angenommen, dass für  $i \in \{1, \dots, n\}$  die Beziehung

$$y_i = \beta_1 + \beta_2 \cdot x_i + u_i$$

gilt, wobei

- ▶  $u_1, \dots, u_n$  (Realisationen von) Zufallsvariablen mit  $E(u_i) = 0$ ,  $\text{Var}(u_i) = \sigma^2$  (unbekannt) und  $\text{Cov}(u_i, u_j) = 0$  für  $i \neq j$  sind, die zufällige Störungen der linearen Beziehung („**Störgrößen**“) beschreiben,
- ▶  $x_1, \dots, x_n$  deterministisch sind mit  $s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 > 0$  (d.h. nicht alle  $x_i$  sind gleich),
- ▶  $\beta_1, \beta_2$  feste, **unbekannte** reelle Parameter sind.
- Man nimmt an, dass man neben  $x_1, \dots, x_n$  auch  $y_1, \dots, y_n$  beobachtet, die wegen der Abhängigkeit von den Zufallsvariablen  $u_1, \dots, u_n$  ebenfalls (Realisationen von) Zufallsvariablen sind. Dies bedeutet **nicht**, dass man auch (Realisationen von)  $u_1, \dots, u_n$  beobachten kann ( $\beta_1$  und  $\beta_2$  unbekannt!).
- Für die Erwartungswerte von  $y_i$  gilt

$$E(y_i) = \beta_1 + \beta_2 \cdot x_i \text{ für } i \in \{1, \dots, n\} .$$

- Das durch obige Annahmen beschriebene Modell heißt auch **einfaches lineares Regressionsmodell**.

- Im einfachen linearen Regressionsmodell sind also (neben  $\sigma^2$ ) insbesondere  $\beta_1$  und  $\beta_2$  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  $\beta_1$  und  $\beta_2$  bestimmten Geraden, in Zeichen*

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

die **beste** (varianzminimale) **lineare** (in  $y_i$ ) **erwartungstreue** Schätzfunktion  $\hat{\beta}_1$  für  $\beta_1$  bzw.  $\hat{\beta}_2$  für  $\beta_2$ .

- 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

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

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

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

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

sowie die Varianzzerlegung

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

## 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

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

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

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

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

# 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

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

abhängen. Für  $n = 7$  Haushalte beobachtet 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  $\beta_1$  und  $\beta_2$  der linearen Modellbeziehung geschätzt sowie die Werte  $\hat{y}_i$ , die Residuen  $\hat{u}_i$  und das Bestimmtheitsmaß  $R^2$  bestimmt werden.



- Berechnete (deskriptive/empirische) Größen:

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

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

$$\hat{\beta}_2 = \frac{s_{X,Y}}{s_X^2} = \frac{30.2449}{114.4901} = 0.26417$$

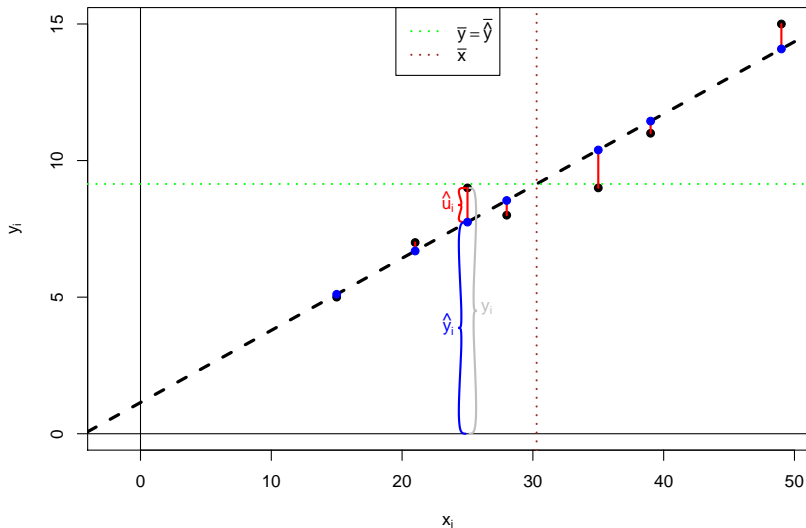
$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \cdot \bar{x} = 9.14286 - 0.26417 \cdot 30.28571 = 1.14228 .$$

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

$i$	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
$\hat{y}_i$	10.39	14.09	6.69	11.44	5.1	8.54	7.75
$\hat{u}_i$	-1.39	0.91	0.31	-0.44	-0.1	-0.54	1.25

# Grafik: Ausgaben in Abhängigkeit vom Einkommen

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



# Eigenschaften der Schätzfunktionen $\hat{\beta}_1$ und $\hat{\beta}_2$

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

$$\hat{\beta}_1 = \sum_{i=1}^n \frac{\overline{x^2} - \bar{x} \cdot x_i}{ns_X^2} \cdot y_i \quad \text{und} \quad \hat{\beta}_2 = \sum_{i=1}^n \frac{x_i - \bar{x}}{ns_X^2} \cdot y_i$$

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

und damit

$$\begin{aligned} E(\hat{\beta}_2) &= E\left(\frac{s_{X,Y}}{s_X^2}\right) = \frac{E(\overline{xy} - \bar{x} \cdot \bar{y})}{s_X^2} = \frac{E(\overline{xy}) - \bar{x} \cdot E(\bar{y})}{s_X^2} \\ &= \frac{\beta_1 \cdot \bar{x} + \beta_2 \cdot \overline{x^2} - \bar{x} \cdot (\beta_1 + \beta_2 \cdot \bar{x})}{s_X^2} = \frac{\beta_2 \cdot (\overline{x^2} - \bar{x}^2)}{s_X^2} = \beta_2 \end{aligned}$$

sowie

$$E(\hat{\beta}_1) = E(\bar{y} - \bar{x}\hat{\beta}_2) = E(\bar{y}) - \bar{x}E(\hat{\beta}_2) = \beta_1 + \beta_2 \cdot \bar{x} - \bar{x} \cdot \beta_2 = \beta_1 .$$

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

- Für die Varianzen der Schätzfunktionen erhält man:

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

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

Diese hängen von der unbekanntem Varianz  $\sigma^2$  der  $u_i$  ab.

- Eine erwartungstreue Schätzfunktion für  $\sigma^2$  ist gegeben durch

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

- Die positive Wurzel  $\hat{\sigma} = +\sqrt{\hat{\sigma}^2}$  dieser Schätzfunktion heißt auch **Standard Error of the Regression (SER)** oder **residual standard error**.

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

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

und

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

- Die positiven Wurzeln  $\widehat{\sigma}_{\widehat{\beta}_1} = \sqrt{\widehat{\sigma}_{\widehat{\beta}_1}^2}$  und  $\widehat{\sigma}_{\widehat{\beta}_2} = \sqrt{\widehat{\sigma}_{\widehat{\beta}_2}^2}$  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

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

gilt, d.h. dass alle  $u_i$  (für  $i \in \{1, \dots, n\}$ ) unabhängig identisch normalverteilt sind mit Erwartungswert 0 und (unbekannter) Varianz  $\sigma^2$ .

- In diesem Fall sind offensichtlich auch  $y_1, \dots, y_n$  stochastisch unabhängig und jeweils normalverteilt mit Erwartungswert  $E(y_i) = \beta_1 + \beta_2 \cdot x_i$  und Varianz  $\text{Var}(y_i) = \sigma^2$ .
- Da  $\hat{\beta}_1$  und  $\hat{\beta}_2$  linear in  $y_i$  sind, folgt insgesamt mit den bereits berechneten Momenten von  $\hat{\beta}_1$  und  $\hat{\beta}_2$ :

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

# Konfidenzintervalle

unter Normalverteilungsannahme für  $u_i$

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

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

- 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  $\beta_1$  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  $\beta_2$ .

## Beispiel: Ausgaben in Abhängigkeit vom Einkommen (II)

- Im bereits erläuterten Beispiel erhält man als Schätzwert für  $\sigma^2$ :

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

- 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  $\beta_1$  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  $\beta_2$ .



# Hypothesentests

unter Normalverteilungsannahme für  $u_i$

- Genauso lassen sich unter der Normalverteilungsannahme (exakte)  $t$ -Tests für die Parameter  $\beta_1$  und  $\beta_2$  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

$$H_0 : \beta_1 = \beta_1^0$$

gegen

$$H_1 : \beta_1 \neq \beta_1^0$$

$$H_0 : \beta_1 \leq \beta_1^0$$

gegen

$$H_1 : \beta_1 > \beta_1^0$$

$$H_0 : \beta_1 \geq \beta_1^0$$

gegen

$$H_1 : \beta_1 < \beta_1^0$$

bzw.

$$H_0 : \beta_2 = \beta_2^0$$

gegen

$$H_1 : \beta_2 \neq \beta_2^0$$

$$H_0 : \beta_2 \leq \beta_2^0$$

gegen

$$H_1 : \beta_2 > \beta_2^0$$

$$H_0 : \beta_2 \geq \beta_2^0$$

gegen

$$H_1 : \beta_2 < \beta_2^0$$

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

# Zusammenfassung: $t$ -Test für den Parameter $\beta_1$

im einfachen linearen Regressionsmodell mit Normalverteilungsannahme

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

# Zusammenfassung: $t$ -Test für den Parameter $\beta_2$

im einfachen linearen Regressionsmodell mit Normalverteilungsannahme

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

## Beispiel: Ausgaben in Abhängigkeit vom Einkommen (III)

- Im bereits erläuterten Beispiel soll zum Signifikanzniveau  $\alpha = 0.05$  getestet werden, ob  $\beta_1$  signifikant von Null verschieden ist. Geeigneter Test:

**t-Test für den Regressionsparameter  $\beta_1$**

1 **Hypothesen:**

$$H_0 : \beta_1 = 0 \quad \text{gegen} \quad H_1 : \beta_1 \neq 0$$

2 **Teststatistik:**

$$t = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}} \text{ ist unter } H_0 \text{ (für } \beta_1 = 0) \text{ } t(n-2)\text{-verteilt.}$$

3 **Kritischer Bereich zum Niveau  $\alpha = 0.05$ :**

$$\begin{aligned} K &= (-\infty, -t_{n-2; 1-\frac{\alpha}{2}}) \cup (t_{n-2; 1-\frac{\alpha}{2}}, +\infty) = (-\infty, -t_{5; 0.975}) \cup (t_{5; 0.975}, +\infty) \\ &= (-\infty, -2.571) \cup (2.571, +\infty) \end{aligned}$$

4 **Berechnung der realisierten Teststatistik:**

$$t = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{1.14228 - 0}{1.1264} = 1.014$$

5 **Entscheidung:**

$$t = 1.014 \notin (-\infty, -2.571) \cup (2.571, +\infty) = K \Rightarrow H_0 \text{ wird nicht abgelehnt!}$$

$$(\text{p-Wert: } 2 - 2 \cdot F_{t(5)}(|t|) = 2 - 2 \cdot F_{t(5)}(|1.014|) = 2 - 2 \cdot 0.8215 = 0.357)$$

Der Test kann für  $\beta_1$  keine signifikante Abweichung von Null feststellen.

# Beispiel: Ausgaben in Abhängigkeit vom Einkommen (IV)

- Nun soll zum Signifikanzniveau  $\alpha = 0.01$  getestet werden, ob  $\beta_2$  **positiv** ist.  
Geeigneter Test:

## **t-Test für den Regressionsparameter $\beta_2$**

### 1 Hypothesen:

$$H_0 : \beta_2 \leq 0 \quad \text{gegen} \quad H_1 : \beta_2 > 0$$

### 2 Teststatistik:

$$t = \frac{\hat{\beta}_2 - 0}{\hat{\sigma}_{\hat{\beta}_2}} \text{ ist unter } H_0 \text{ (für } \beta_2 = 0) \text{ } t(n-2)\text{-verteilt.}$$

### 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:

$$t = \frac{\hat{\beta}_2 - 0}{\hat{\sigma}_{\hat{\beta}_2}} = \frac{0.26417 - 0}{0.0351} = 7.5262$$

### 5 Entscheidung:

$$t = 7.5262 \in (3.365, +\infty) = K \quad \Rightarrow \quad H_0 \text{ wird abgelehnt!}$$

$$(\text{p-Wert: } 1 - F_{t(5)}(t) = 1 - F_{t(5)}(7.5262) = 1 - 0.9997 = 0.0003)$$

Der Test stellt fest, dass  $\beta_2$  signifikant positiv ist.

# Punkt- und Intervallprognosen

im einfachen linearen Regressionsmodell mit Normalverteilungsannahme

- Neben Konfidenzintervallen und Tests für die Parameter  $\beta_1$  und  $\beta_2$  in linearen Regressionsmodellen vor allem **Prognosen** wichtige Anwendung.
- Zur Erstellung von Prognosen: Erweiterung der Modellannahme

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

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.  $E(y_0) = \beta_1 + \beta_2 \cdot x_0$  auf Grundlage von  $x_0$ .
- Wegen  $E(u_0) = 0$  und der Erwartungstreue von  $\hat{\beta}_1$  für  $\beta_1$  bzw.  $\hat{\beta}_2$  für  $\beta_2$  ist

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

offensichtlich erwartungstreu für  $y_0$  bzw.  $E(y_0)$  gegeben  $x_0$ .

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

# Prognosefehler

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

$$\hat{y}_0 - y_0 \quad \text{bzw.} \quad \widehat{E(y_0)} - E(y_0) .$$

- Qualitativer Unterschied:
  - Prognosefehler

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

resultiert **nur** aus Fehler bei der Schätzung von  $\beta_1$  bzw.  $\beta_2$  durch  $\hat{\beta}_1$  bzw.  $\hat{\beta}_2$ .

- Prognosefehler

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

ist Kombination von Schätzfehlern (für  $\beta_1$  und  $\beta_2$ ) sowie zufälliger Schwankung von  $u_0 \sim N(0, \sigma^2)$ .

- Zunächst: Untersuchung von  $e_E := \widehat{E(y_0)} - E(y_0)$

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

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

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

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

- Insgesamt berechnet man so die Varianz des Prognosefehlers

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



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

$$e_E = \widehat{E}(y_0) - E(y_0) \sim N(0, \sigma_{e_E}^2) \quad \text{bzw.} \quad \frac{\widehat{E}(y_0) - E(y_0)}{\sigma_{e_E}} \sim N(0, 1).$$

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

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

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

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

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

## Prognoseintervalle für $E(y_0)$ gegeben $x_0$

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

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

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

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

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

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

# Prognosefehler $e_0 := \hat{y}_0 - y_0$

- *Nun:* Untersuchung des Prognosefehlers  $e_0 := \hat{y}_0 - y_0$
- Offensichtlich gilt für  $e_0 = \hat{y}_0 - y_0$  die Zerlegung

$$\begin{aligned} \hat{y}_0 - y_0 &= \underbrace{(\hat{\beta}_1 + \hat{\beta}_2 \cdot x_0)}_{=E(\hat{y}_0)} - \underbrace{(\beta_1 + \beta_2 \cdot x_0 + u_0)}_{=E(y_0)} \\ &= \underbrace{E(\hat{y}_0) - E(y_0)}_{\text{Fehler aus Schätzung von } \beta_1 \text{ und } \beta_2} - \underbrace{u_0}_{\text{zufällige Schwankung der Störgröße}}. \end{aligned}$$

- $E(\hat{y}_0)$  hängt nur von  $u_1, \dots, u_n$  ab (über  $y_1, \dots, y_n$  bzw.  $\hat{\beta}_1$  und  $\hat{\beta}_2$ ) und ist wegen der Annahme  $u_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$  **unabhängig** von  $u_0$ .
- Damit sind die beiden Bestandteile des Prognosefehlers insbesondere auch unkorreliert und man erhält:

$$\begin{aligned} \sigma_{e_0}^2 &:= \text{Var}(\hat{y}_0 - y_0) = \text{Var}(E(\hat{y}_0) - E(y_0)) + \text{Var}(u_0) \\ &= \sigma^2 \cdot \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n \cdot s_X^2} \right) + \sigma^2 = \sigma^2 \cdot \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n \cdot s_X^2} \right) \end{aligned}$$

- Aus der Unkorreliertheit der beiden Komponenten des Prognosefehlers folgt auch sofort die Normalverteilungseigenschaft des Prognosefehlers  $e_0 = y_0 - \hat{y}_0$ , genauer gilt:

$$e_0 = \hat{y}_0 - y_0 \sim N(0, \sigma_{e_0}^2) \quad \text{bzw.} \quad \frac{\hat{y}_0 - y_0}{\sigma_{e_0}} \sim N(0, 1) .$$

- Wieder muss  $\sigma^2$  durch  $\hat{\sigma}^2$  ersetzt werden, um mit Hilfe der geschätzten Varianz

$$\hat{\sigma}_{e_0}^2 := \widehat{\text{Var}}(\hat{y}_0 - y_0) = \hat{\sigma}^2 \cdot \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n \cdot s_X^2} \right)$$

des Prognosefehlers die für die Praxis relevante Verteilungsaussage

$$\frac{e_0}{\hat{\sigma}_{e_0}} = \frac{\hat{y}_0 - y_0}{\hat{\sigma}_{e_0}} \sim t(n - 2) ,$$

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