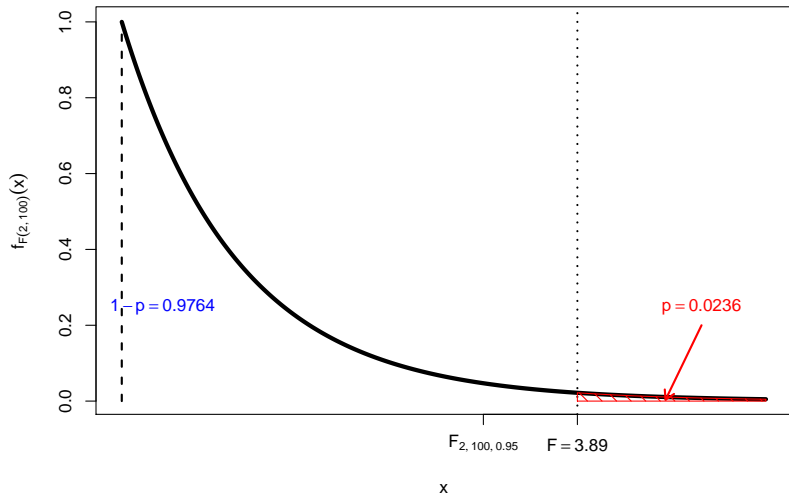


# Beispiel: $p$ -Wert bei Varianzanalyse (Grafik)

Bedienungszeiten-Beispiel, realisierte Teststatistik  $F = 3.89$ ,  $p$ -Wert: 0.0236



# Varianzanalyse und 2-Stichproben- $t$ -Test

- Varianzanalyse zwar für  $k > 2$  unabhängige Stichproben eingeführt, Anwendung aber auch für  $k = 2$  möglich.
- Nach Zuordnung der beteiligten Größen in den unterschiedlichen Notationen ( $\mu_A \equiv \mu_1, \mu_B \equiv \mu_2, X_i^A \equiv X_{1,i}, X_i^B \equiv X_{2,i}, n_A \equiv n_1, n_B \equiv n_2, n = n_A + n_B$ ) enger Zusammenhang zum 2-Stichproben- $t$ -Test erkennbar:
  - ▶ Fragestellungen (Hypothesenpaare) und Anwendungsvoraussetzungen identisch mit denen des zweiseitigen 2-Stichproben- $t$ -Tests für den Mittelwertvergleich bei unbekanntem, aber übereinstimmenden Varianzen.
  - ▶ Man kann zeigen: Für Teststatistik  $F$  der Varianzanalyse im Fall  $k = 2$  und Teststatistik  $t$  des 2-Stichproben- $t$ -Tests gilt  $F = t^2$ .
  - ▶ Es gilt außerdem zwischen Quantilen der  $F(1, n)$  und der  $t(n)$ -Verteilung der Zusammenhang  $F_{1, n; 1-\alpha} = t_{n, 1-\frac{\alpha}{2}}^2$ . Damit:

$$x \in (-\infty, -t_{n, 1-\frac{\alpha}{2}}) \cup (t_{n, 1-\frac{\alpha}{2}}, \infty) \iff x^2 \in (F_{1, n; 1-\alpha}, \infty)$$

- Insgesamt sind damit die Varianzanalyse mit  $k = 2$  Faktorstufen und der zweiseitige 2-Stichproben- $t$ -Test für den Mittelwertvergleich bei unbekanntem, aber übereinstimmenden Varianzen also äquivalent in dem Sinn, dass Sie stets übereinstimmende Testentscheidungen liefern!

# Deskriptive Beschreibung linearer Zusammenhänge

- Aus deskriptiver Statistik bekannt: Pearsonscher Korrelationskoeffizient als Maß der Stärke des *linearen* Zusammenhangs zwischen zwei (kardinalskalierten) Merkmalen  $X$  und  $Y$ .
- *Nun*: Ausführlichere Betrachtung linearer Zusammenhänge zwischen Merkmalen (zunächst rein deskriptiv!):  
Liegt ein linearer Zusammenhang zwischen zwei Merkmalen  $X$  und  $Y$  nahe, ist nicht nur die Stärke dieses Zusammenhangs interessant, sondern auch die genauere „Form“ des Zusammenhangs.
- „Form“ linearer Zusammenhänge kann durch Geraden(gleichungen) spezifiziert werden.
- *Problemstellung*: Wie kann zu einer Urliste  $(x_1, y_1), \dots, (x_n, y_n)$  der Länge  $n$  zu  $(X, Y)$  eine sog. **Regressionsgerade** (auch: Ausgleichsgerade) gefunden werden, die den linearen Zusammenhang zwischen  $X$  und  $Y$  „möglichst gut“ widerspiegelt?
- *Wichtig*: Was soll „möglichst gut“ überhaupt bedeuten?  
*Hier*: Summe der quadrierten Abstände von der Geraden zu den Datenpunkten  $(x_i, y_i)$  in **vertikaler** Richtung soll möglichst gering sein.  
(*Begründung für Verwendung dieses „Qualitätskriteriums“ wird nachgeliefert!*)

- Geraden (eindeutig) bestimmt (zum Beispiel) durch Absolutglied  $a$  und Steigung  $b$  in der bekannten Darstellung

$$y = f_{a,b}(x) := a + b \cdot x .$$

- Für den  $i$ -ten Datenpunkt  $(x_i, y_i)$  erhält man damit den vertikalen Abstand

$$u_i(a, b) := y_i - f_{a,b}(x_i) = y_i - (a + b \cdot x_i)$$

von der Geraden mit Absolutglied  $a$  und Steigung  $b$ .

- Gesucht werden  $a$  und  $b$  so, dass die Summe der quadrierten vertikalen Abstände der „Punktwolke“  $(x_i, y_i)$  von der durch  $a$  und  $b$  festgelegten Geraden,

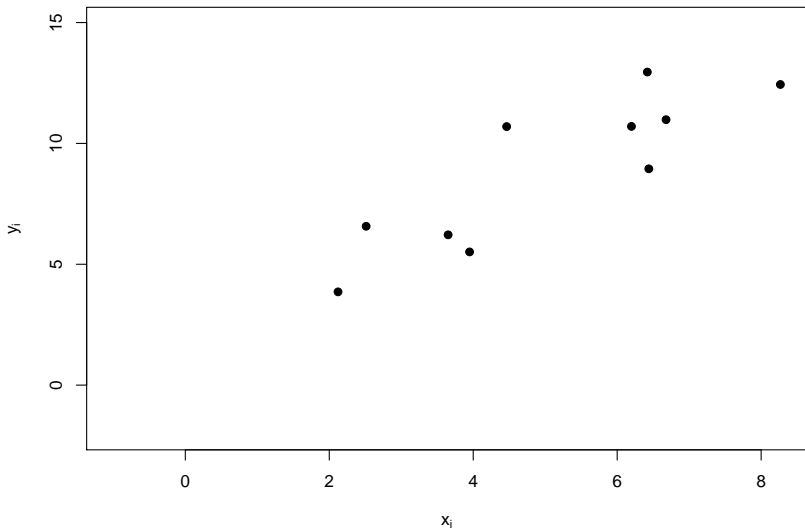
$$\sum_{i=1}^n (u_i(a, b))^2 = \sum_{i=1}^n (y_i - f_{a,b}(x_i))^2 = \sum_{i=1}^n (y_i - (a + b \cdot x_i))^2 ,$$

möglichst klein wird.

- Verwendung dieses Kriteriums heißt auch **Methode der kleinsten Quadrate (KQ-Methode)** oder **Least-Squares-Methode (LS-Methode)**.

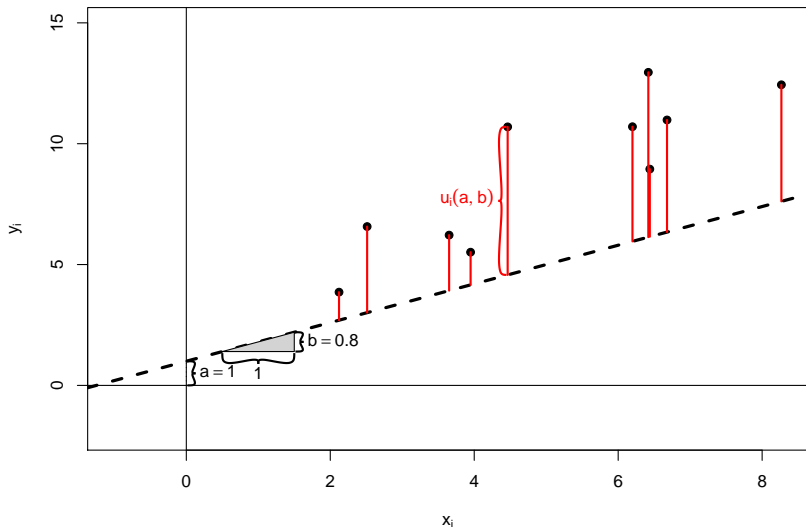
# Beispiel: „Punktwolke“

aus  $n = 10$  Paaren  $(x_i, y_i)$



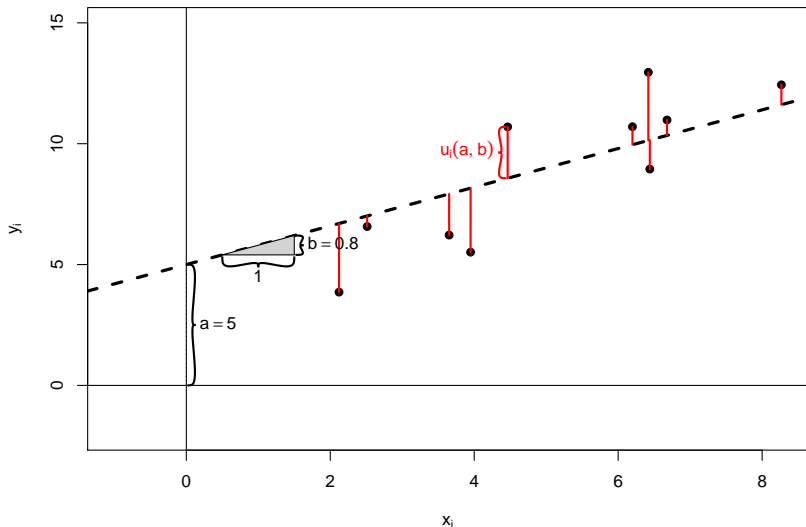
# Beispiel: „Punktwolke“ und verschiedene Geraden (I)

$$a = 1, b = 0.8, \sum_{i=1}^n (u_i(a, b))^2 = 180.32$$



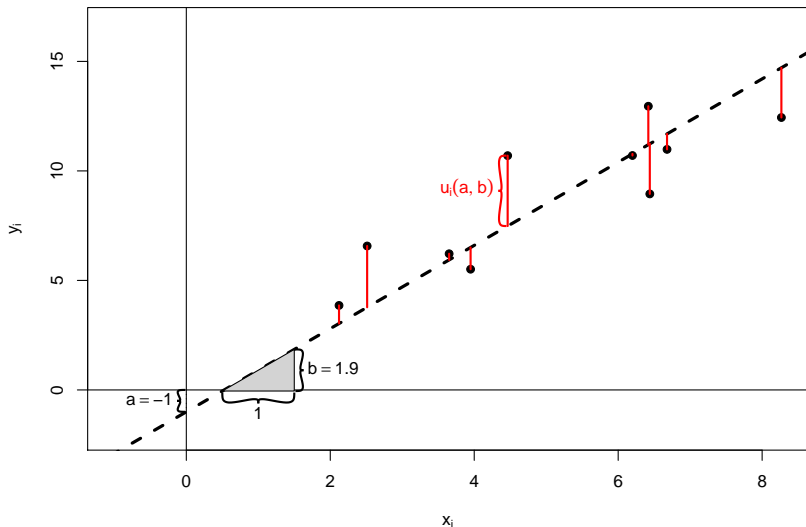
# Beispiel: „Punktvolke“ und verschiedene Geraden (II)

$$a = 5, b = 0.8, \sum_{i=1}^n (u_i(a, b))^2 = 33.71$$



# Beispiel: „Punktwolke“ und verschiedene Geraden (III)

$$a = -1, b = 1.9, \sum_{i=1}^n (u_i(a, b))^2 = 33.89$$





# Rechnerische Bestimmung der Regressionsgeraden (I)

- Gesucht sind also  $\hat{a}, \hat{b} \in \mathbb{R}$  mit

$$\sum_{i=1}^n (y_i - (\hat{a} + \hat{b}x_i))^2 = \min_{a, b \in \mathbb{R}} \sum_{i=1}^n (y_i - (a + bx_i))^2$$

- Lösung dieses Optimierungsproblems durch Nullsetzen des Gradienten, also

$$\frac{\partial \sum_{i=1}^n (y_i - (a + bx_i))^2}{\partial a} = -2 \sum_{i=1}^n (y_i - a - bx_i) \stackrel{!}{=} 0$$

$$\frac{\partial \sum_{i=1}^n (y_i - (a + bx_i))^2}{\partial b} = -2 \sum_{i=1}^n (y_i - a - bx_i)x_i \stackrel{!}{=} 0 ,$$

führt zu sogenannten **Normalgleichungen**:

$$na + \left( \sum_{i=1}^n x_i \right) b \stackrel{!}{=} \sum_{i=1}^n y_i$$

$$\left( \sum_{i=1}^n x_i \right) a + \left( \sum_{i=1}^n x_i^2 \right) b \stackrel{!}{=} \sum_{i=1}^n x_i y_i$$

## Rechnerische Bestimmung der Regressionsgeraden (II)

- Aufgelöst nach  $a$  und  $b$  erhält man die Lösungen

$$\hat{b} = \frac{n \left( \sum_{i=1}^n x_i y_i \right) - \left( \sum_{i=1}^n x_i \right) \cdot \left( \sum_{i=1}^n y_i \right)}{n \left( \sum_{i=1}^n x_i^2 \right) - \left( \sum_{i=1}^n x_i \right)^2}$$

$$\hat{a} = \frac{1}{n} \left( \sum_{i=1}^n y_i \right) - \frac{1}{n} \left( \sum_{i=1}^n x_i \right) \cdot \hat{b}$$

oder kürzer mit den aus der deskr. Statistik bekannten Bezeichnungen

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \overline{x^2} = \frac{1}{n} \sum_{i=1}^n x_i^2, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad \text{und} \quad \overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i$$

bzw. den empirischen Momenten  $s_{X,Y} = \overline{xy} - \bar{x} \cdot \bar{y}$  und  $s_X^2 = \overline{x^2} - \bar{x}^2$ :

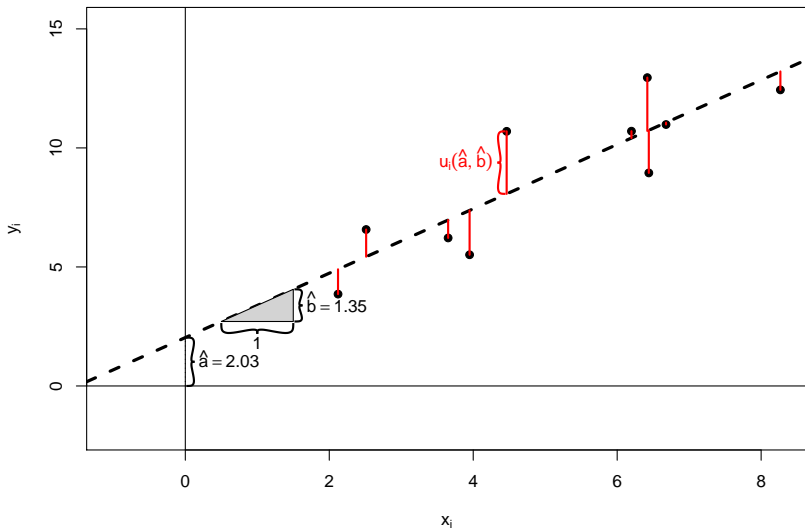
$$\hat{b} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{s_{X,Y}}{s_X^2}$$

$$\hat{a} = \bar{y} - \bar{x} \hat{b}$$

- Die erhaltenen Werte  $\hat{a}$  und  $\hat{b}$  minimieren tatsächlich die Summe der quadrierten vertikalen Abstände, da die Hesse-Matrix positiv definit ist.

# Beispiel: „Punktwolke“ und Regressionsgerade

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



- Zu  $\hat{a}$  und  $\hat{b}$  kann man offensichtlich die folgende, durch die Regressionsgerade erzeugte Zerlegung der Merkmalswerte  $y_i$  betrachten:

$$y_i = \underbrace{\hat{a} + \hat{b} \cdot x_i}_{=: \hat{y}_i} + \underbrace{y_i - (\hat{a} + \hat{b} \cdot x_i)}_{=: u_i(\hat{a}, \hat{b}) =: \hat{u}_i}$$

- Aus den Normalgleichungen lassen sich leicht einige wichtige Eigenschaften für die so definierten  $\hat{u}_i$  und  $\hat{y}_i$  herleiten, insbesondere:
  - ▶  $\sum_{i=1}^n \hat{u}_i = 0$  und damit  $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$  bzw.  $\bar{y} = \bar{\hat{y}} := \frac{1}{n} \sum_{i=1}^n \hat{y}_i$ .
  - ▶  $\sum_{i=1}^n x_i \hat{u}_i = 0$ .
  - ▶ Mit  $\sum_{i=1}^n \hat{u}_i = 0$  und  $\sum_{i=1}^n x_i \hat{u}_i = 0$  folgt auch  $\sum_{i=1}^n \hat{y}_i \hat{u}_i = 0$ .

Mit diesen Eigenschaften erhält man die folgende Varianzzerlegung:

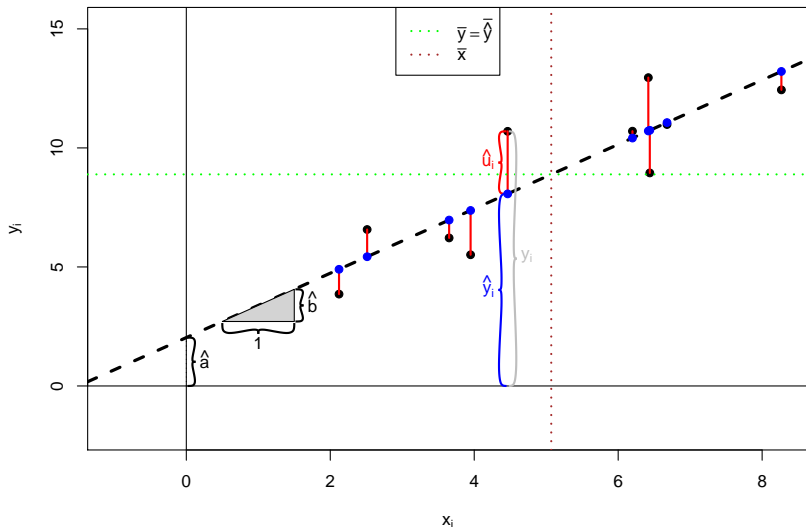
$$\underbrace{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}_{\text{Gesamtvarianz der } y_i} = \underbrace{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}_{\text{erklärte Varianz}} + \underbrace{\frac{1}{n} \sum_{i=1}^n \hat{u}_i^2}_{\text{unerklärte Varianz}}$$

- Die als Anteil der erklärten Varianz an der Gesamtvarianz gemessene Stärke des linearen Zusammenhangs steht in engem Zusammenhang mit  $r_{X,Y}$ ; es gilt:

$$r_{X,Y}^2 = \frac{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$$

# 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**.