Residuen

- Jedem Datenpaar (y_i, x_i), $i = 1, \ldots, n$ der Stichprobe wird der Wert auf der Regressionsgeraden

$$\hat{y}_i = b_0 + b_1 x_i$$

zugeordnet. Die **Residuen**

$$\hat{e}_i = y_i - \hat{y}_i$$

sind die vorzeichenbehafteten vertikalen Abstände der Punkte (y_i, x_i) zur Regressionsgeraden.

- Mit Hilfe der Residuen kann man für jeden einzelnen Datenpunkt überprüfen, wie gut er aufgrund des Modells vorhergesagt wurde.

Bsp. Residuen Schlafverhalten (1)

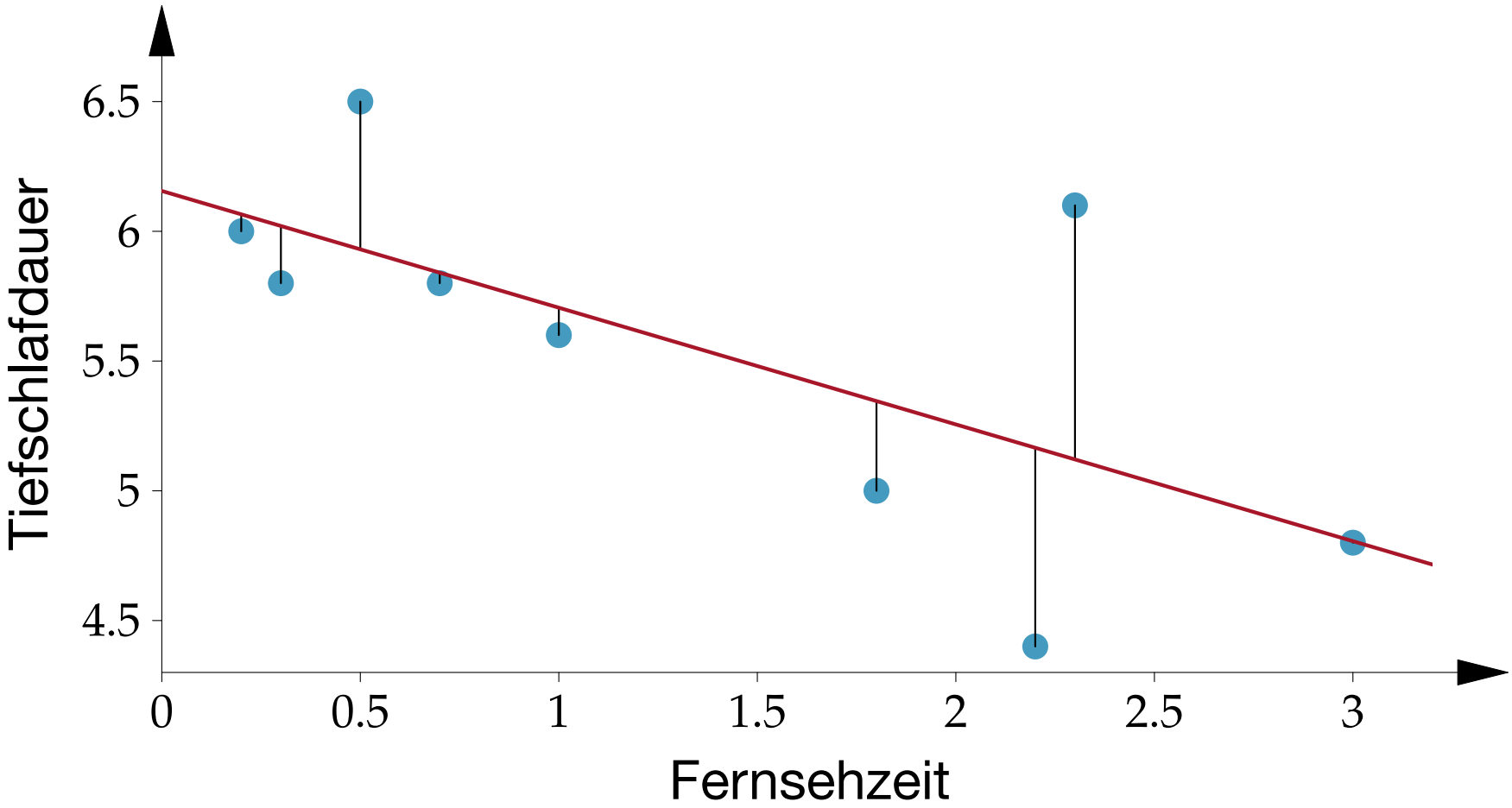

Abb.: Streudiagramm mit Regressionsgerade und Residuen

$$\hat{e}_i = y_i - \hat{y}_i$$

Bsp. Residuen Schlafverhalten (2)

- Hilfstabelle:

i	x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
1	0.30	5.80	6.02	-0.22
2	2.20	4.40	5.17	-0.77
3	0.50	6.50	5.93	0.57
4	0.70	5.80	5.84	-0.04
5	1.00	5.60	5.71	-0.11
6	1.80	5.00	5.35	-0.35
7	3.00	4.80	4.81	-0.01
8	0.20	6.00	6.07	-0.07
9	2.30	6.10	5.12	0.98

Wie gut ist das Modell?

- Frage: Welcher Anteil der Streuung der y-Werte lässt sich durch die Regression von Y auf X erklären?

Gesamtstreuung SQT

- Die Summe der Quadrate der Abweichungen der y_i von ihrem Mittelwert wird als **Gesamtstreuung** bezeichnet:

$$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2 \,,$$

wobei SQT die Abkürzung für ***S**um of S**q**uares **T**otal* ist.

Bsp. Gesamtstreuung Schlafverh. (1)

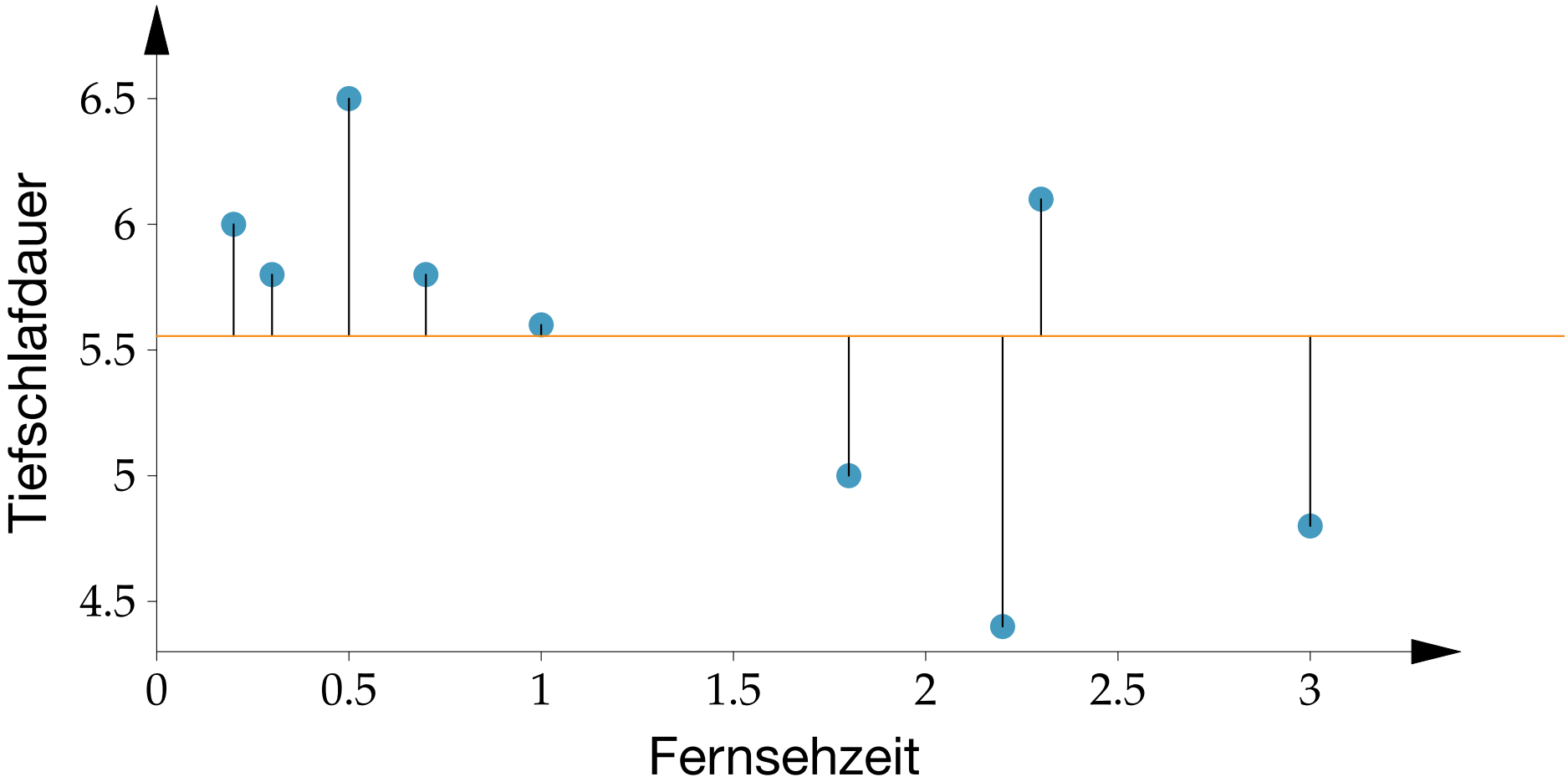

Abb.: Streudiagramm mit Gerade $y = \bar{y}$ und $y_i - \bar{y}$

Bsp. Gesamtstreuung Schlafverh. (2)

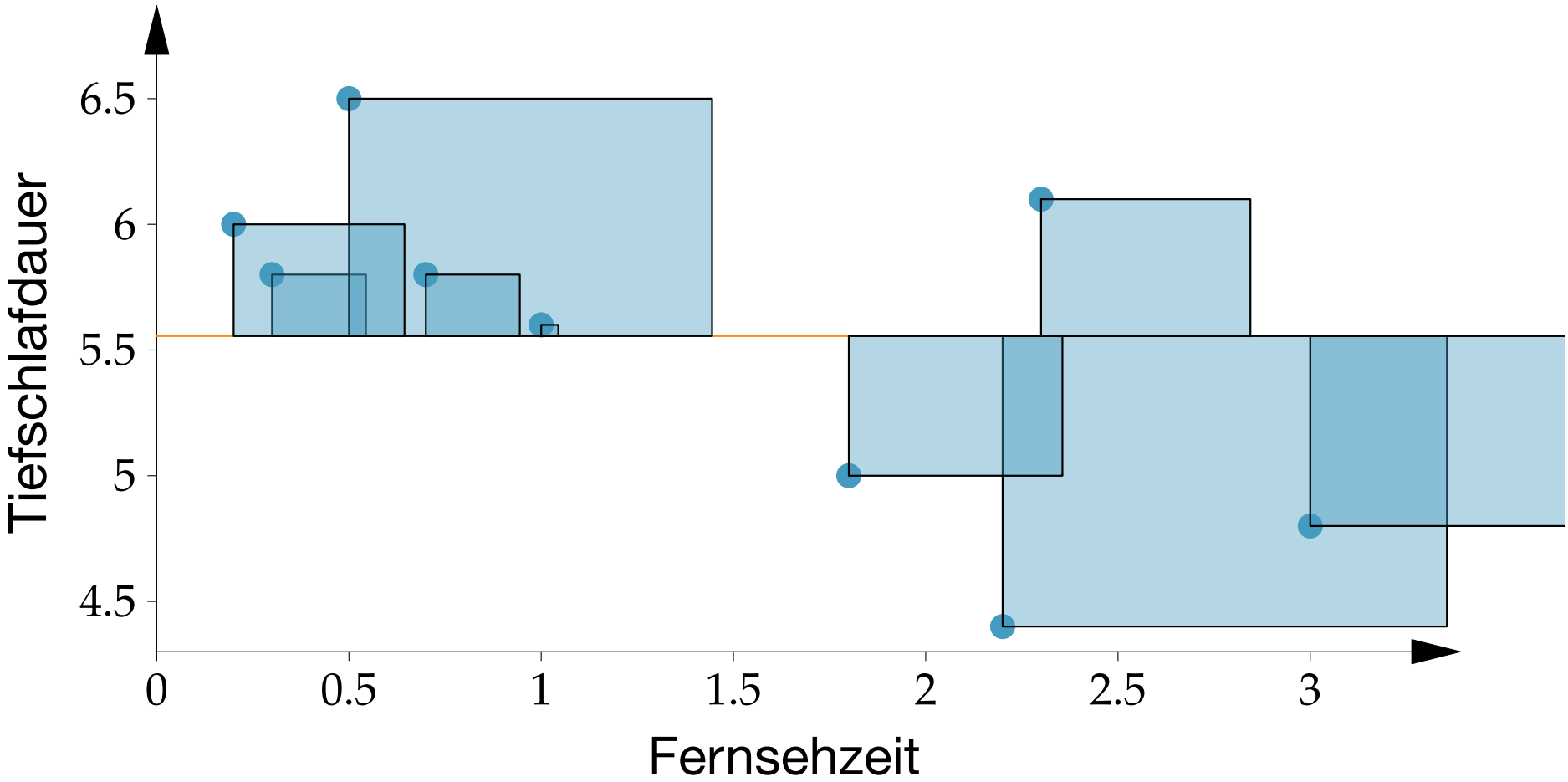

Abb.: Streudiagramm mit Gerade $y = \bar{y}$ und $(y_i - \bar{y})^2$

Bsp. Gesamtstreuung Schlafverh. (3)

- Hilfstabelle:

i	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	5.8	0.24	0.06
2	4.4	-1.16	1.34
3	6.5	0.94	0.89
4	5.8	0.24	0.06
5	5.6	0.04	0
6	5	-0.56	0.31
7	4.8	-0.76	0.57
8	6	0.44	0.2
9	6.1	0.54	0.3
Σ			3.72
$\bar{y}$		5.56	

Erklärte Streuung SQE

- Die Summe der Quadrate der Abweichungen der $\hat{y}_i$ auf der Regressionsgeraden von ihrem Mittelwert wird als **erklärte Streuung** bezeichnet:

$$SQE = \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \,,$$

wobei SQE die Abkürzung für ***S**um of S**q**uares **E**xplained* ist. Wegen $\bar{\hat{y}} = \bar{y}$ besitzen die y-Werte auf der Regressionsgeraden den gleichen Mittelwert wie y-Werte der Stichprobe.

Bsp. Erklärte Streuung Schlafverh. (1)

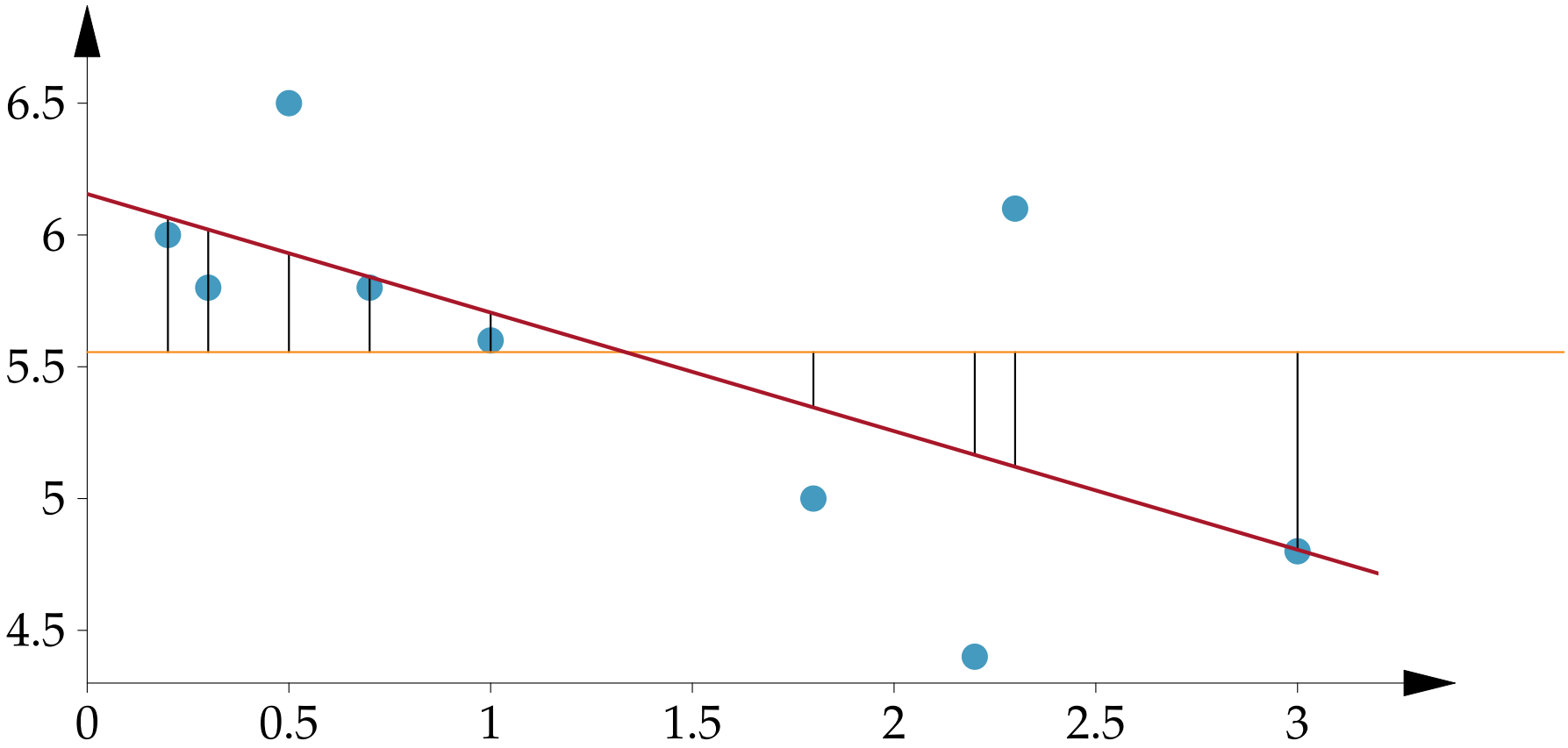

Abb.: Streudiagramm mit Gerade $y = \bar{y}$, Regressionsgerade und $\hat{y}_i - \bar{y}$

Bsp. Erklärte Streuung Schlafverh. (2)

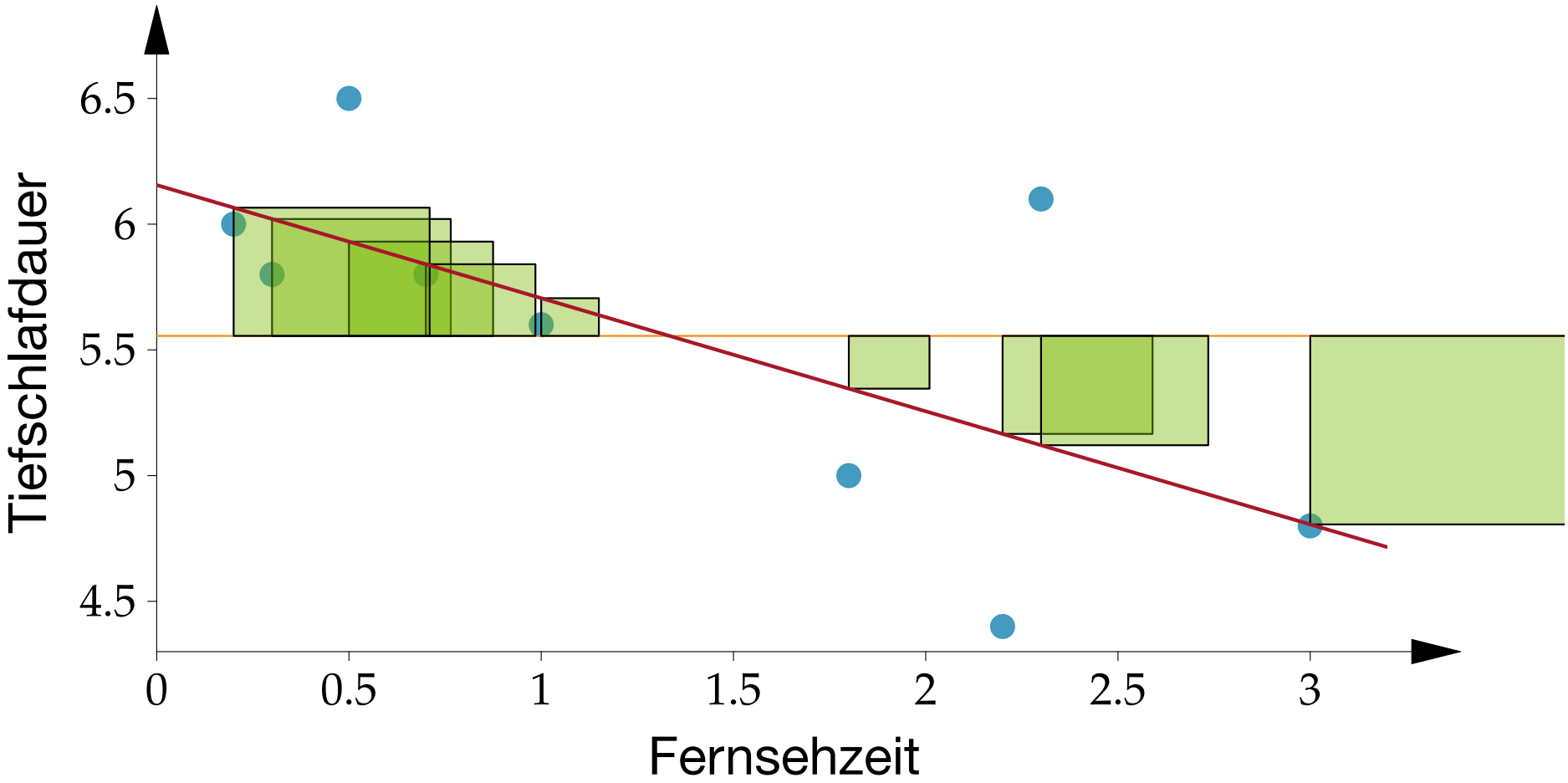

Abb.: Streudiagramm mit Gerade $y = \bar{y}$, Regressionsgerade und $(\hat{y}_i - \bar{y})^2$

Bsp. Erklärte Streuung Schlafverh. (3)

- Hilfstabelle:

i	$\hat{y}_i$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$
1	6.02	0.46	0.22
2	5.17	-0.39	0.15
3	5.93	0.37	0.14
4	5.84	0.28	0.08
5	5.71	0.15	0.02
6	5.35	-0.21	0.04
7	4.81	-0.75	0.56
8	6.07	0.51	0.26
9	5.12	-0.43	0.19
Σ			1.67

Reststreuung SQR

- Die Summe der Quadrate der Abweichungen der y_i zur Regressionsgeraden wird als **Reststreuung** bezeichnet:

$$SQR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{e}_i^2 \,,$$

wobei SQR die Abkürzung für ***S**um of S**q**uares **R**esiduals* ist.

Bsp. Reststreuung Schlafverhalten (1)

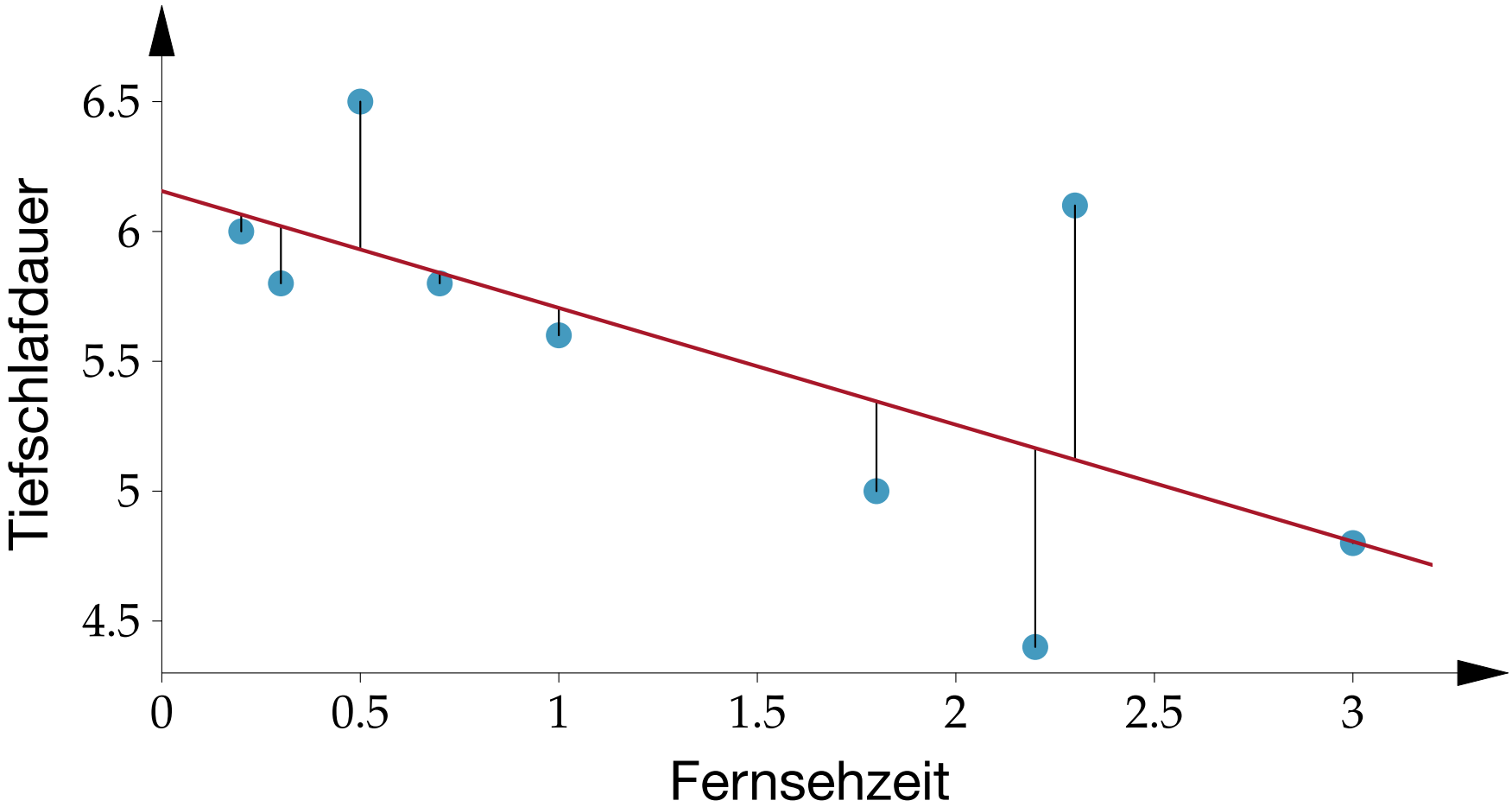

Abb.: Streudiagramm mit Regressionsgerade und Residuen

Bsp. Reststreuung Schlafverhalten (2)

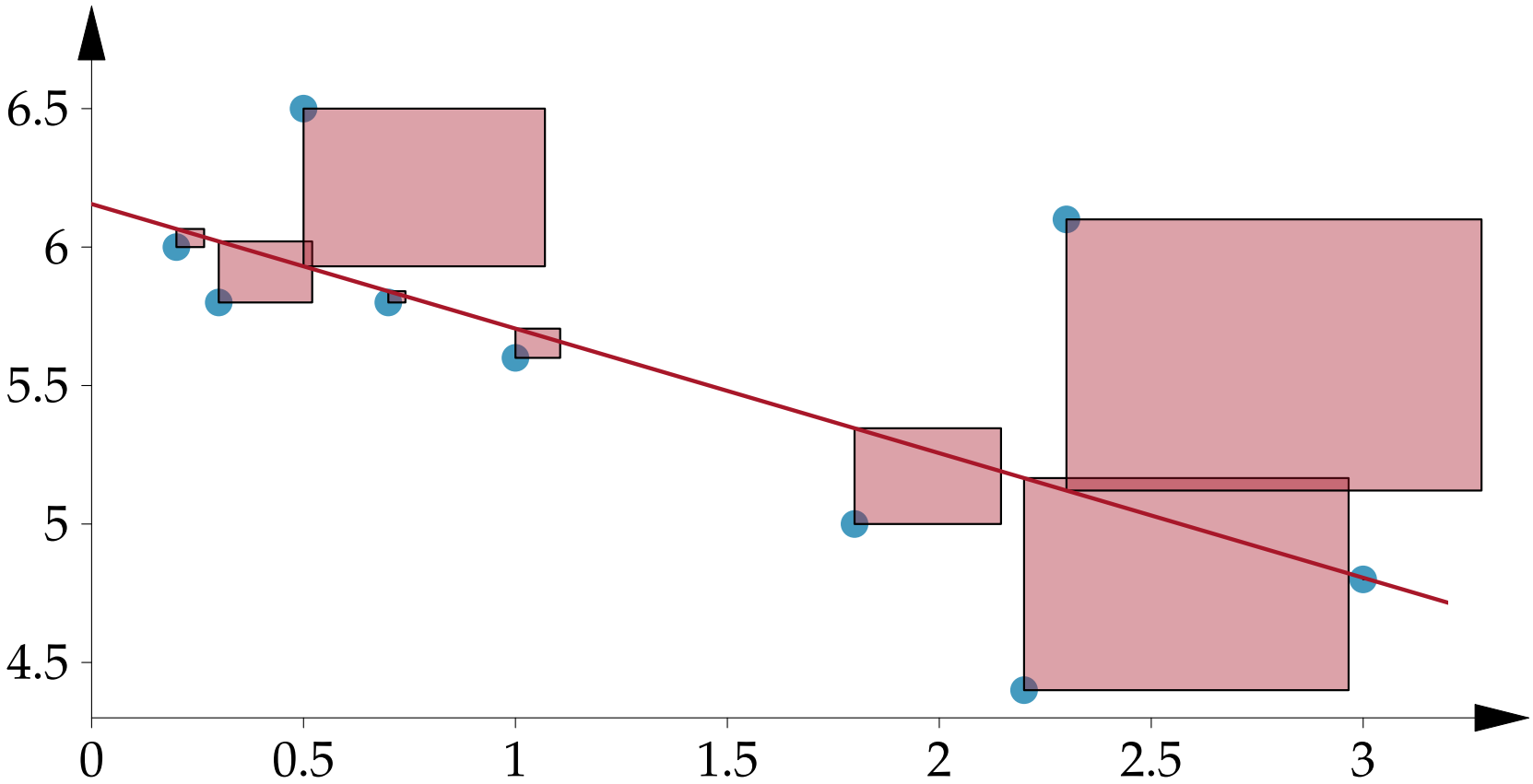

Abb.: Streudiagramm mit Regressionsgerade und $(y_i - \hat{y}_i)^2$

Bsp. Reststreuung Schlafverhalten (3)

- Hilfstabelle:

i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	5.8	6.02	-0.22	0.05
2	4.4	5.17	-0.77	0.59
3	6.5	5.93	0.57	0.32
4	5.8	5.84	-0.04	0
5	5.6	5.71	-0.11	0.01
6	5	5.35	-0.35	0.12
7	4.8	4.81	-0.01	0
8	6	6.07	-0.07	0
9	6.1	5.12	0.98	0.96
Σ				2.05

Streuungszerlegung (1)

- Für die Gesamtstreuung gilt also die folgende Zerlegung:

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

$$SQT = SQE + SQR\,.$$

- Die erklärte Streuung SQE enthält die Variation der Datenpunkte auf der Geraden um $\bar{y}$.
- Die Reststreuung SQR entspricht dem verbleibenden Rest an Variation der y-Werte.

Streuungszerlegung (2)

- Liegen alle beobachteten Punkte exakt auf einer Geraden, so sind die Residuen alle null und ebenso die Reststreuung.
- Je größer nun die Reststreuung ist, desto schlechter beschreibt das Modell die Daten, d. h. desto weniger wird die in den Daten vorhandene Streuung durch das Modell erklärt.
- Als Maßzahl für die Güte der Modellanpassung verwendet man eine Größe, die auf der Streuungszerlegung aufbaut, und zwar das sogenannte *Bestimmtheitsmaß*.

Bestimmtheitsmaß R^2

- Das Bestimmtheitsmaß R^2 ist der Anteil der Gesamtstreuung der y-Werte, der durch die Regression von Y auf X erklärt wird, und ist somit der Quotient aus erklärter und Gesamtstreuung, d. h.

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

- Das Bestimmtheitsmaß nimmt Werte zwischen null und eins an, also $0 \leq R^2 \leq 1$.
- Außerdem gilt $R^2 = r^2$, d. h. der quadrierte Korrelationskoeffizient entspricht dem Anteil der erklärten Streuung an der Gesamtstreuung.

Bsp. Erklärte und Gesamtstreuung

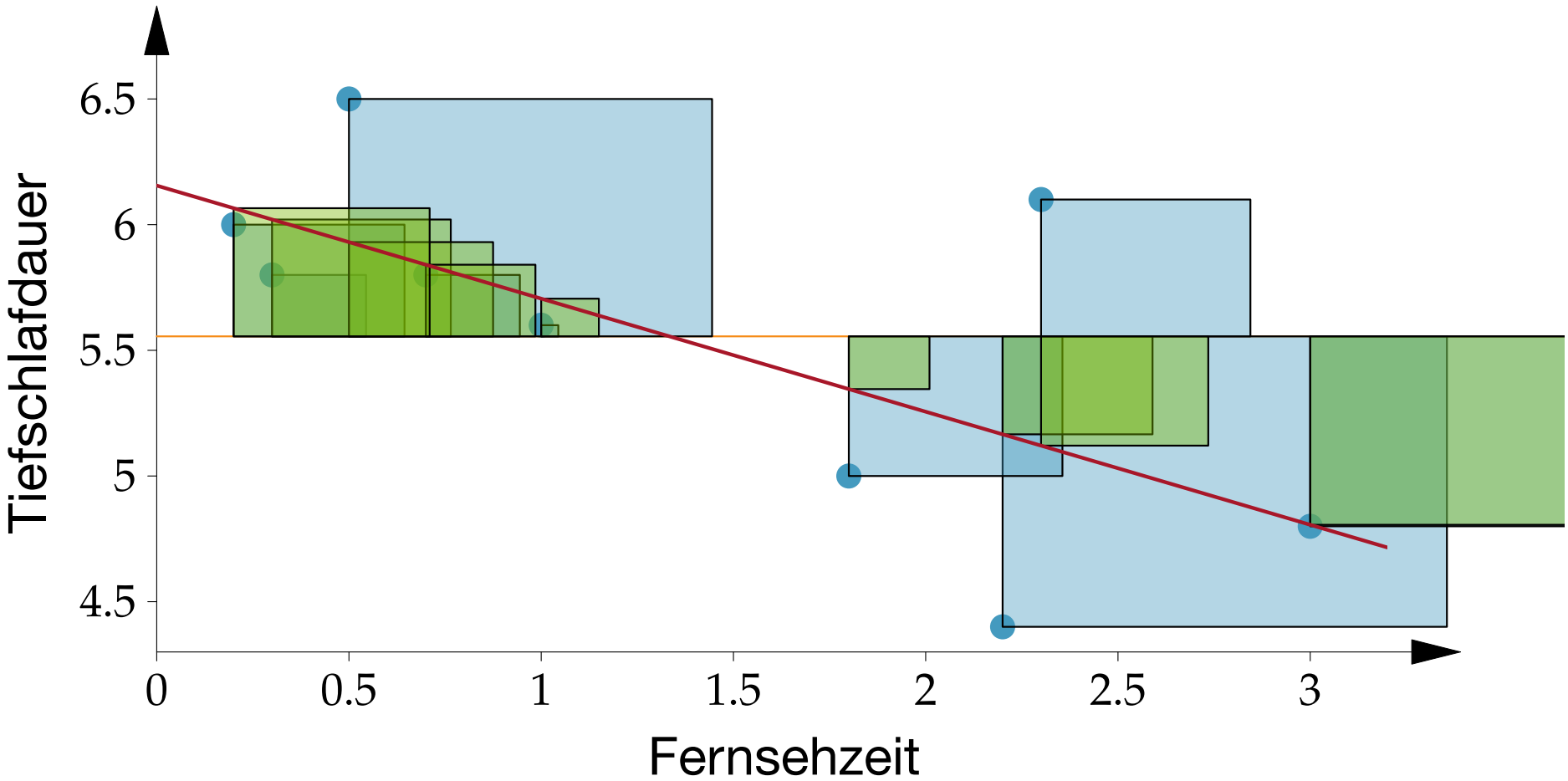

Abb.: Streudiagramm mit Gerade $y = \bar{y}$, Regressionsgerade, Gesamtstreuung (blau) und erklärter Streuung (grün)

Residuenplots

- Verletzungen der Modellannahmen lassen sich durch sogenannte *Residuenplots* aufdecken.
- Das Streudiagramm der $(\hat{y}_i, \hat{e}_i)$-Werte sollte kein systematisches Muster aufweisen.

Bsp. Residuenplots (1)

- Schwanken die Residuen unsystematisch um die horizontale Achse und sind nahe bei null, deutet dies auf eine gute Modellanpassung hin.

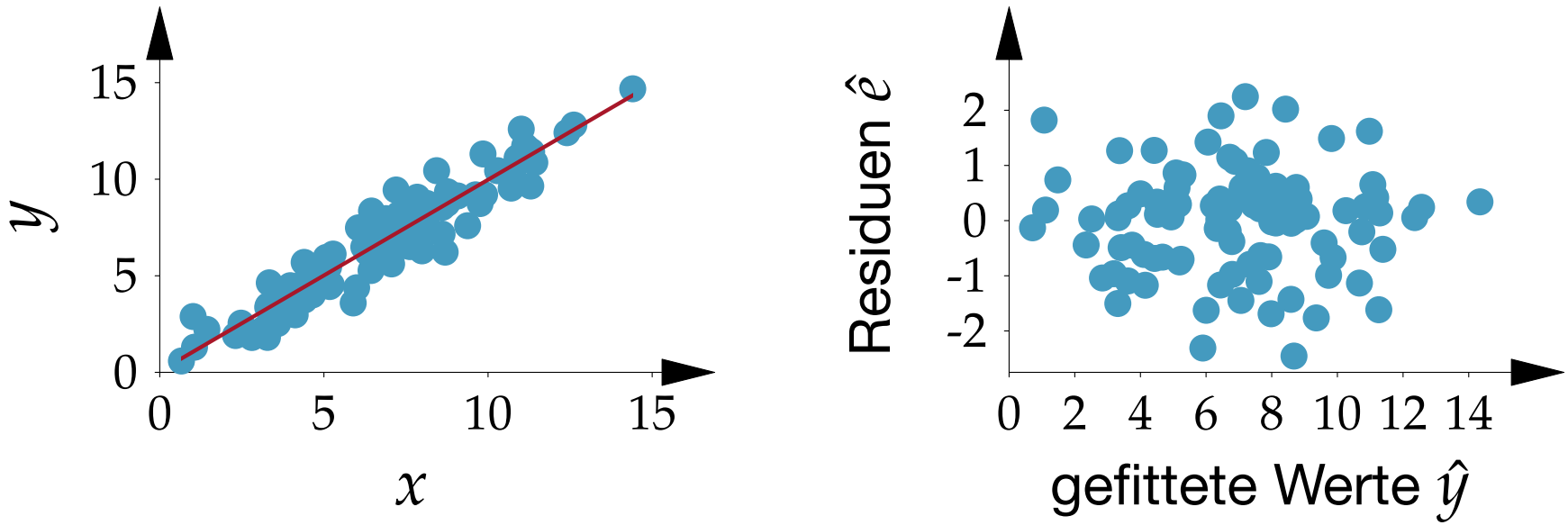

Abb.: Streudiagramm mit Regressionsgerade (links) und zugehörigem Residuenplot (rechts)

Bsp. Residuenplots (2)

- Zeigen die Residuen ein Muster, liegt die Vermutung nahe, dass eine nicht lineare Abhängigkeit zwischen den Merkmalen besteht, die nicht durch das Modell erfasst wird.

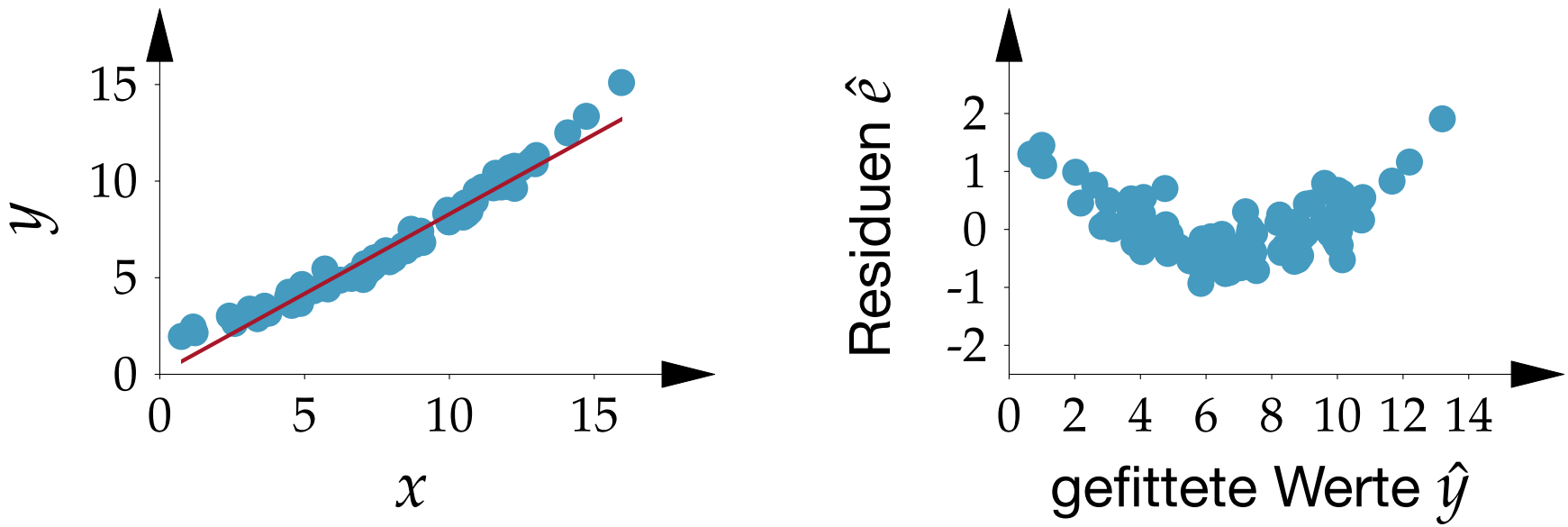

Abb.: Streudiagramm mit Regressionsgerade (links) und zugehörigem Residuenplot (rechts)

Bsp. Residuenplots (3)

- Verändert sich die Streuung der Residuen, besteht keine Varianzhomogenität (Voraussetzung).

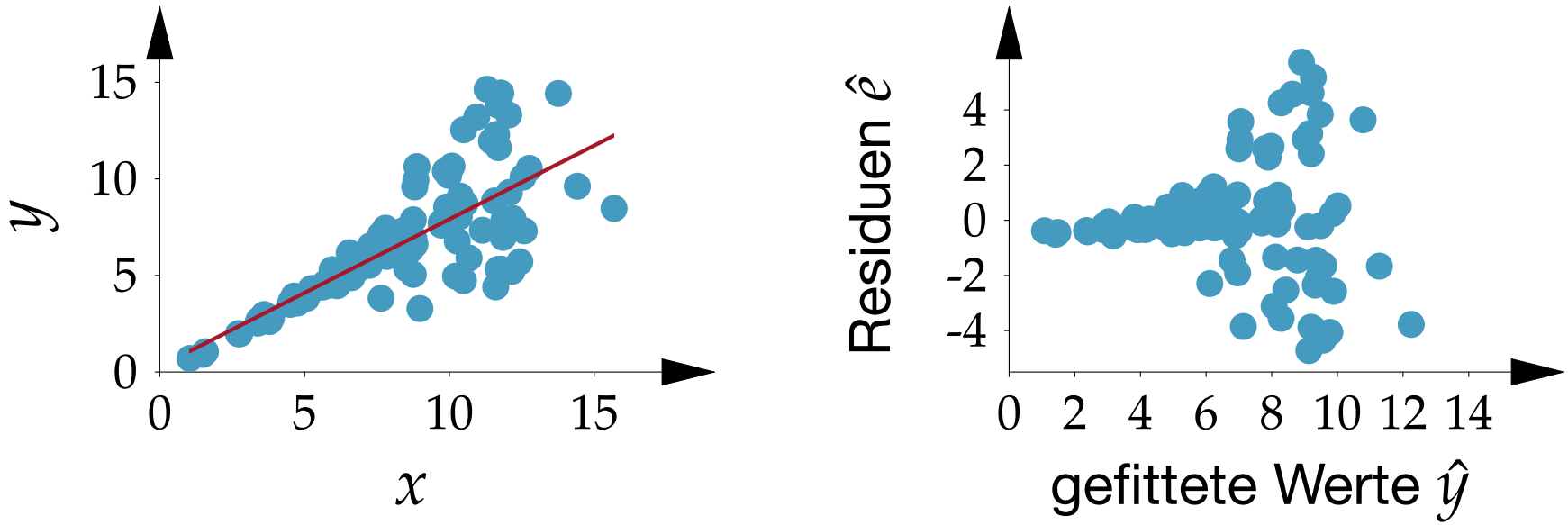

Abb.: Streudiagramm mit Regressionsgerade (links) und zugehörigem Residuenplot (rechts)

Bsp. Residuenplot Schlafverhalten

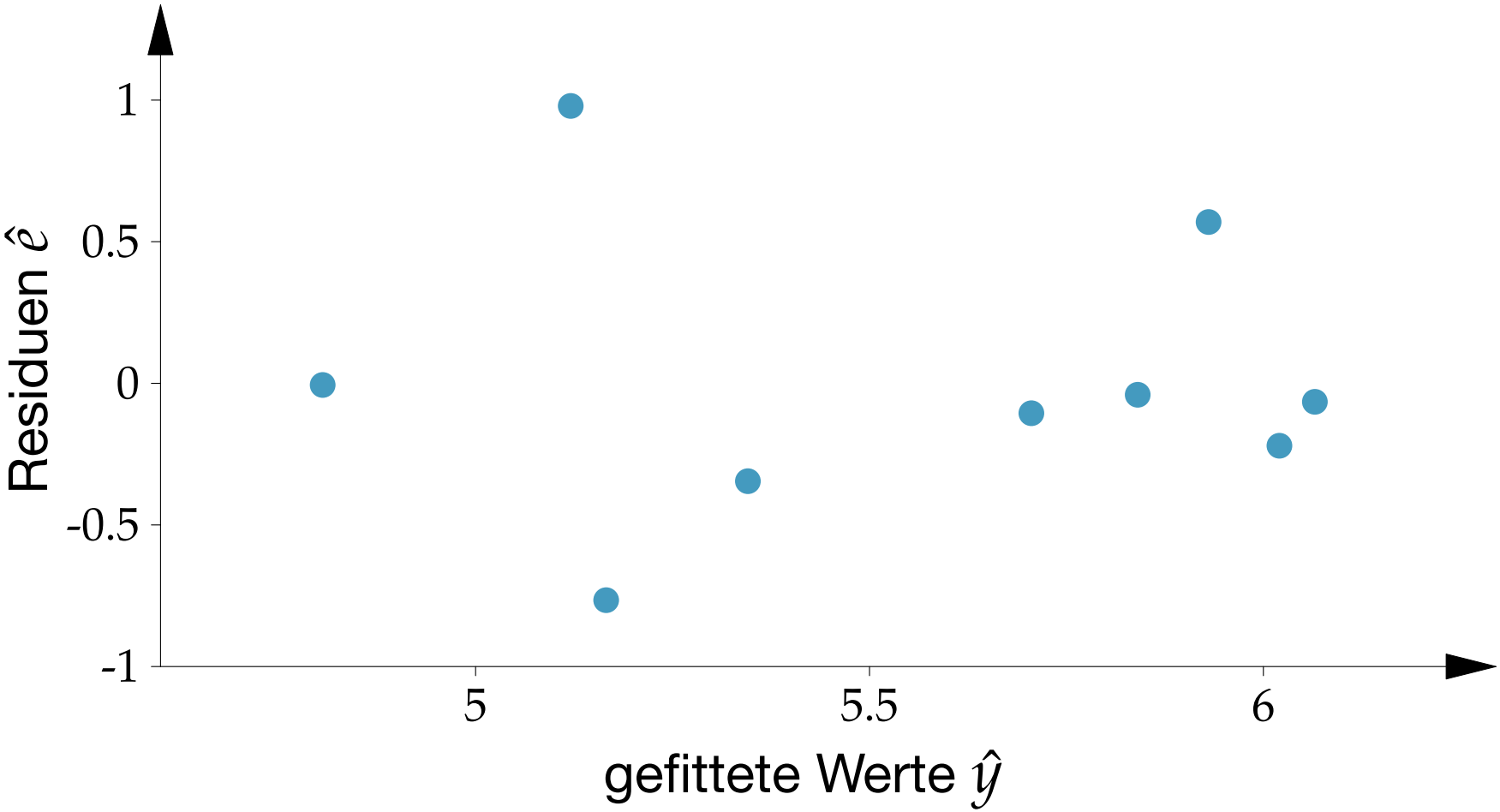

Abb.: Residuenplot

4. Wahrscheinlich-keitsrechnung

Wahrscheinlichkeitsrechnung

- deskriptive Statistik: basiert auf Daten
- Wahrscheinlichkeitsrechnung: Wahrheiten aus Gedankenexperimenten, ohne Daten!
  - Beschäftigt sich mit der mathematischen Beschreibung von Zufall.
  - Gegenstand sind Zufallsexperimente, Zufallsereignisse und deren Wahrscheinlichkeit.

Bsp. Wahrscheinlichkeitsrechnung

- Ein Studierender, der keine Zeit hat, sich auf einen 20-Fragen-Single-Choice-Test vorzubereiten, beschließt, bei jeder Frage aufs Geratewohl zu raten. Dabei besitzt jede Frage fünf Antwortmöglichkeiten, wobei nur eine Antwort richtig ist. Der Test gilt als bestanden, wenn zehn Fragen richtig beantwortet sind.

  Wie groß ist die Wahrscheinlichkeit des Studierenden, den Test zu bestehen?

Grundbegriffe (1)

- **Zufallsexperiment:** Experiment mit nicht vorhersagbarem Ergebnis
- Beispiele:
  - Werfen einer Münze, Werfen eines Würfels
  - Messen der Körpergröße, des Blutdrucks und des Gewichts einer zufällig ausgewählten Person
  - zufällige Auswahl einer Glühbirne und Bestimmung ihrer Lebensdauer

Grundbegriffe (2)

- **Ergebnismenge Ω:** Menge $\Omega = \{\omega_1, \ldots, \omega_n\}$ der möglichen Ergebnisse eines Zufallsexperiments
- Beispiele:
  - Werfen einer Münze: $\Omega = \{\text{Kopf, Zahl}\}$
  - Werfen eines Würfels: $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - zweimaliges Werfen eines Würfels: $\Omega = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), \ldots, (2,6), \ldots, (6,6)\}$
  - Bestimmung der Lebensdauer einer Glühbirne: $\Omega = \{x \in \mathbb{R} \mid x \geq 0\} = [0, \infty)$

Grundbegriffe (3)

- **Ergebnisse:** Elemente von Ω
- **Ereignisse:** Teilmengen von Ω, also Zusammenfassungen von Ergebnissen eines Zufallsexperiments
- Beispiele:
  - Werfen eines Würfels:
    Das Ereignis „Eine gerade Zahl wird geworfen“ entspricht der Menge $\{2, 4, 6\}$.
  - Bestimmung der Lebensdauer einer Glühbirne:
    Das Ereignis „Die Glühbirne brennt höchstens 200 Stunden“ ist das Intervall $[0, 200]$.

Grundbegriffe (4)

- **Elementarereignisse** sind einelementige Teilmengen von Ω, d. h. $\{\omega_1\}, \ldots, \{\omega_n\}$.
- Ein **unmögliches Ereignis** wird mit $\{\ \}$ bezeichnet, ein **sicheres Ereignis** mit Ω.
- Man sagt „Das Ereignis A tritt ein“, wenn das Ergebnis ω des Zufallsvorgangs in A liegt, also $\omega \in A$ gilt.

Grundlagen Mengen (1)

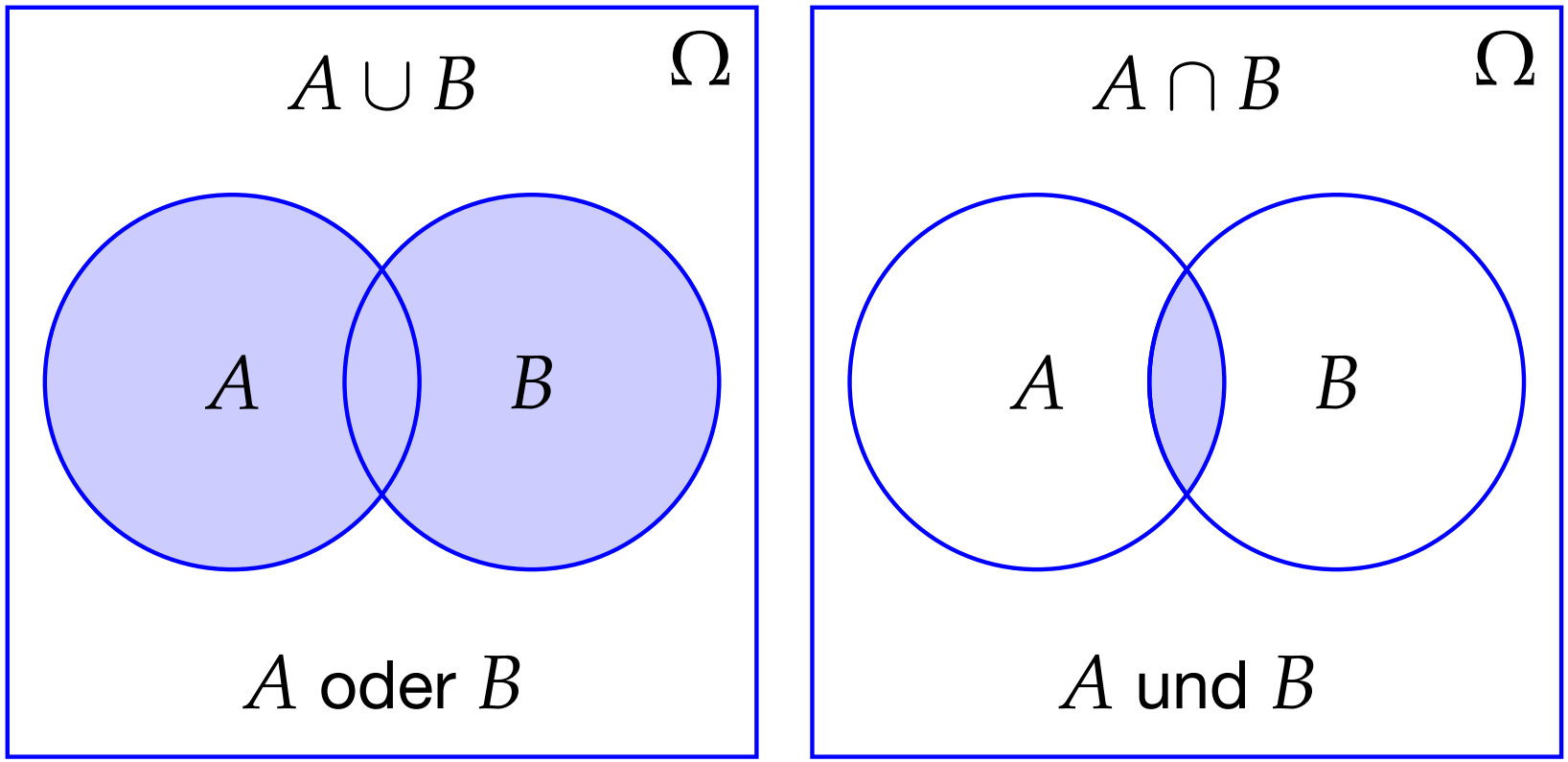

Abb.: Venn-Diagramme zur Vereinigungsmenge (links) und Schnittmenge (rechts) zweier Mengen A und B

Grundlagen Mengen (2)

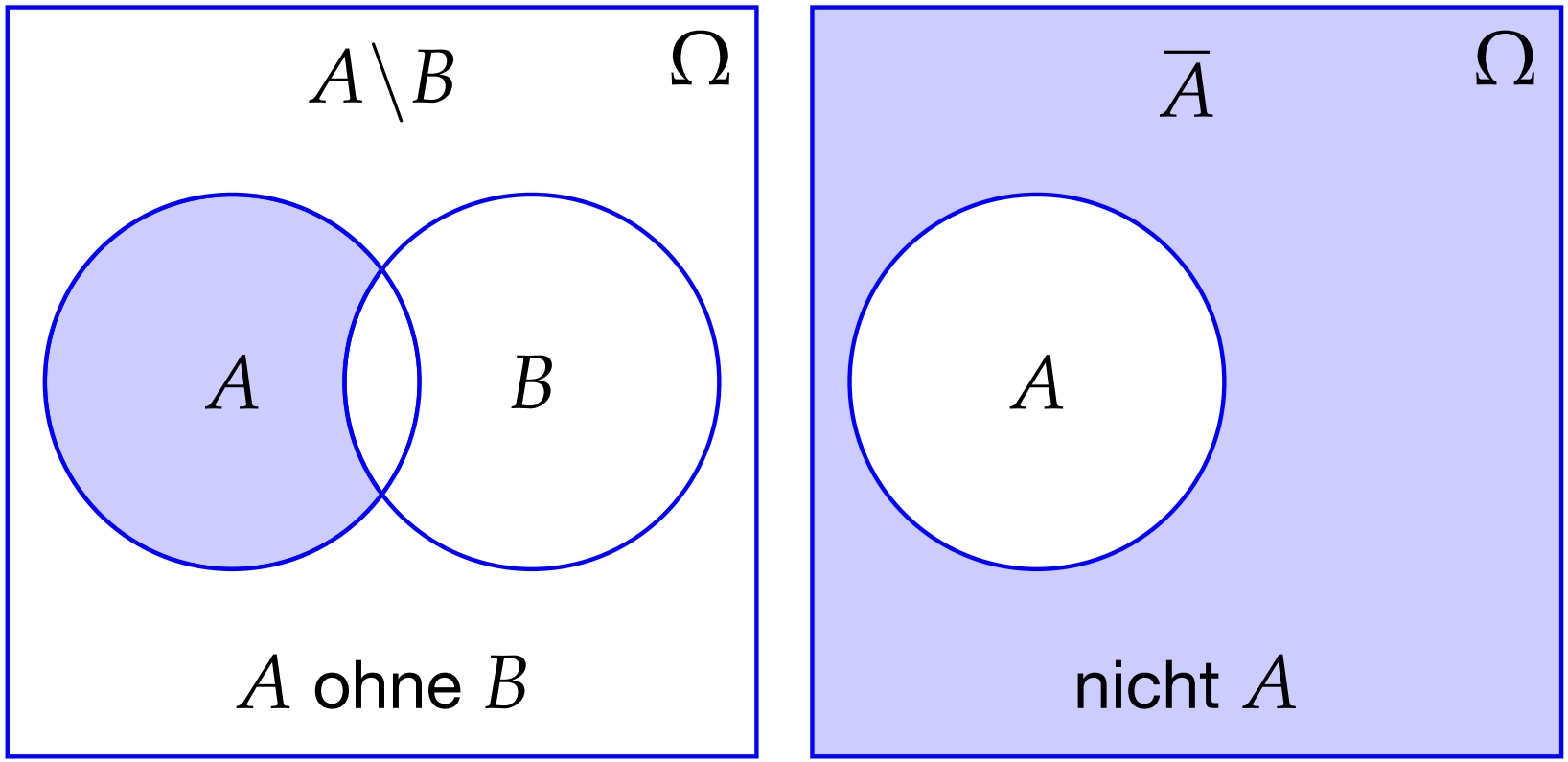

Abb.: Venn-Diagramme zur Differenzmenge (links) zweier Mengen A, B und zur Komplementärmenge (rechts) einer Menge A

Grundlagen Mengen (3)

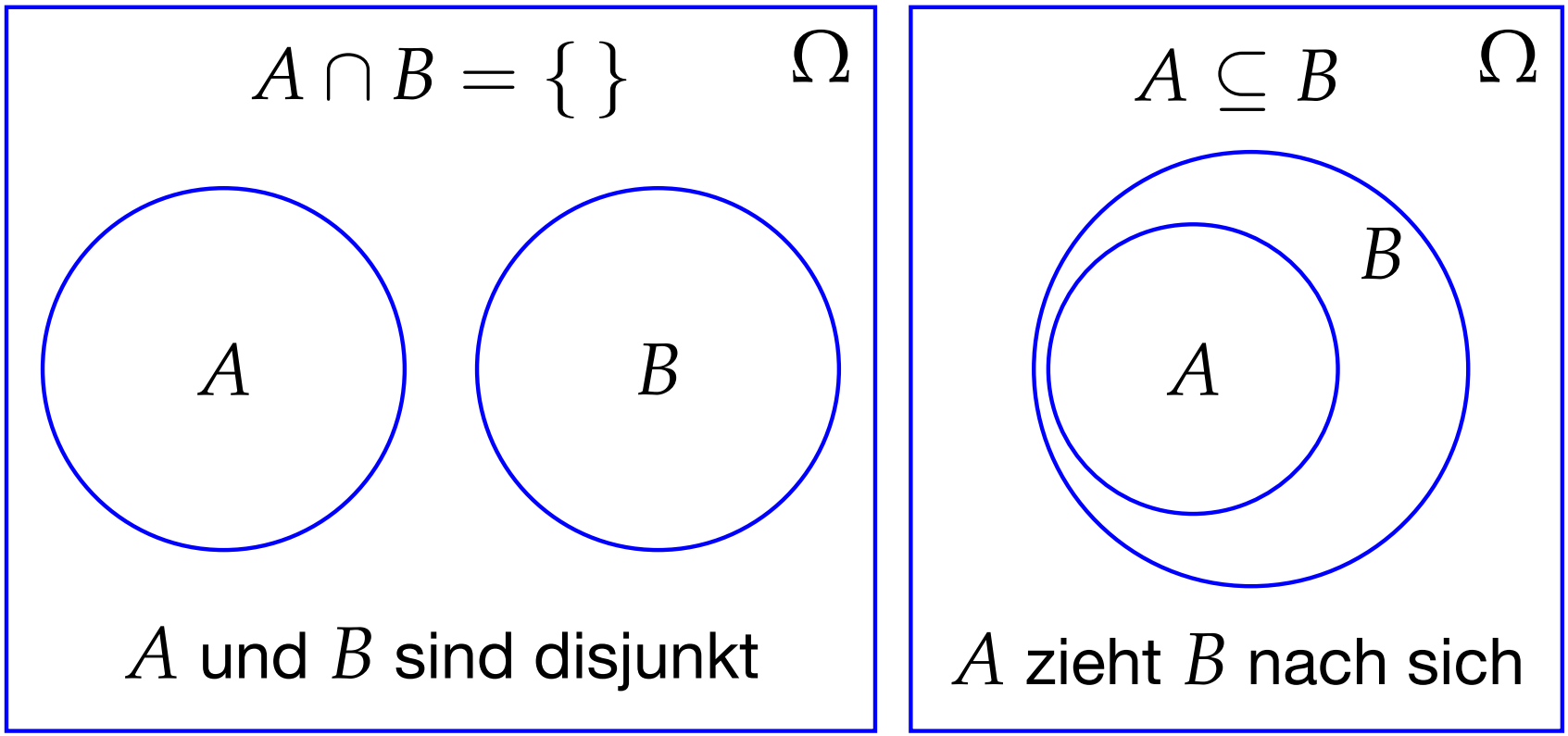

Abb.: Venn-Diagramme zur Disjunktheit (links) und Teilmenge (rechts) zweier Mengen A und B

Wahrscheinlichkeit (1)

- Den Ereignissen werden Wahrscheinlichkeiten zugeordnet, die den Axiomen von Kolmogoroff (1903–1987) zu genügen haben:

(K1) Jedem Ereignis ist eine Wahrscheinlichkeit – eine Zahl von Null bis Eins – zugeordnet (Nichtnegativitätsaxiom):

$$0 \leq P(A) \leq 1$$

(K2) Das sichere Ereignis hat die Wahrscheinlichkeit Eins (Normierungsaxiom):

$$P(\Omega) = 1$$

Wahrscheinlichkeit (2)

(K3) Die Wahrscheinlichkeit dafür, dass von mehreren, paarweise einander sich ausschließenden Ereignissen eines eintritt, ist gleich der Summe der Wahrscheinlichkeiten der Ereignisse (Additivitätsaxiom):

Falls $A \cap B = \{\}$, so ist

$$P(A \cup B) = P(A) + P(B)\,.$$

Rechenregeln Wahrscheinlichkeiten

- Sei Ω ein Ergebnisraum und seien A, B beliebige Ereignisse. Dann gilt:

$$P(\{\}) = 0$$

$$P(\overline{A}) = 1 - P(A)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(B \backslash A) = P(B) - P(A \cap B)$$

Bsp. Rechenregeln Wahrscheinlichkeiten

- A und B seien Ereignisse, für die gilt: $P(A) = 0.4$, $P(B) = 0.7$, $P(A \cap B) = 0.2$

  Wie groß sind die Wahrscheinlichkeiten der folgenden Ereignisse:

  - Von beiden Ereignissen A und B
    a) treten beide ein.
    b) tritt mindestens eines ein.
    c) tritt keines ein.
    d) tritt höchstens eines ein.
  - Nur das Ereignis B tritt ein.

Berechnung der Wahrscheinlichkeit

- Sei $A \subseteq \Omega$, Ω endlich, und seien die Elementarereignisse bezeichnet mit $\{\omega_1\}, \ldots, \{\omega_n\}$, dann gilt:

$$P(\{\omega_i\}) \geq 0, \quad i = 1, \ldots, n$$

$$P(\Omega) = P(\{\omega_1\}) + \ldots + P(\{\omega_n\}) = 1$$

$$P(A) = \sum_{\omega \in A} P(\{\omega\})$$

Laplace-Wahrscheinlichkeit (1)

- Ein Laplace-Experiment ist ein Zufallsexperiment mit gleich wahrscheinlichen Elementarereignissen.
- Für ein Ereignis A ergibt sich die Laplace-Wahrscheinlichkeit $P(A)$ durch

$$P(A) = \frac{\text{Anzahl der für } A \text{ günstigen Ergebnisse}}{\text{Anzahl aller möglichen Ergebnisse}}.$$

- Bedeutung: Wahrscheinlichkeiten für komplexere Experimente werden durch Abzählen ermittelt. (Kunst des Abzählens: Gebiet der Kombinatorik)

Laplace-Wahrscheinlichkeit (2)

- In einem Laplace-Experiment gilt für ein beliebiges Ereignis A

$$P(A) = \frac{|A|}{|\Omega|}$$

mit:

$|A|$: Anzahl der Elemente der Menge A

$|\Omega|$: Anzahl der Elemente der Menge Ω

- Ein unmögliches Ereignis hat die Wahrscheinlichkeit Null, ein sicheres Ereignis die Wahrscheinlichkeit Eins.

Bedingte Wahrscheinlichkeit

- Seien $A, B \subseteq \Omega$ zwei Ereignisse und es gelte $P(B) > 0$. Dann heißt

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

die **bedingte Wahrscheinlichkeit von A unter der Bedingung B**.

Produktsatz

- Seien $A, B \subseteq \Omega$ zwei Ereignisse und es gelte $P(B) > 0$. Durch Umformen der Definition der bedingten Wahrscheinlichkeit erhält man

$$P(A \cap B) = P(A|B) \cdot P(B)\,.$$

2×2-Kontingenztafel zu zwei Ereignissen

- Zwei Ereignisse $A, B \subseteq \Omega$ erzeugen vier disjunkte Fälle:

	B	$\overline{B}$	
A	$P(A \cap B)$	$P(A \cap \overline{B})$	$P(A)$
$\overline{A}$	$P(\overline{A} \cap B)$	$P(\overline{A} \cap \overline{B})$	$P(\overline{A})$
	$P(B)$	$P(\overline{B})$	1

Zellen sind gemeinsame Wahrscheinlichkeiten,
Randwerte sind Randwahrscheinlichkeiten.

Stochastische Unabhängigkeit

- Seien $A, B \subseteq \Omega$ zwei Ereignisse. A und B heißen **stochastisch unabhängig**, wenn

$$P(A \cap B) = P(A) \cdot P(B)\,,$$

$$P(A \mid B) = P(A) \quad \text{mit} \quad P(B) > 0 \quad \text{bzw.}$$

$$P(B \mid A) = P(B) \quad \text{mit} \quad P(A) > 0 \quad \text{gilt.}$$

Andernfalls heißen die Ereignisse **stochastisch abhängig**.