9. Regressionsanalyse

9.1 Regressionsfunktion

- Zielsetzung der Regressionsanalyse: Quantifizierung des Einflusses einer Variablen auf eine andere Variable
- Unterstellte Beziehung in der Regression:

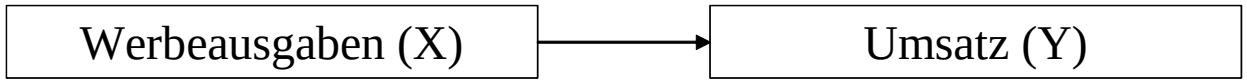
$$\boxed{X} \longrightarrow \boxed{Y}$$

- Korrelationsanalyse: Hier ist es unerheblich, welches Merkmal mit X oder Y bezeichnet wird
- Regressionsanalyse: Betrachtung des Einflusses von X auf Y, d.h. X ist die unabhängige Variable und Y ist die abhängige Variable
- Man spricht auch von einer Regression von Y auf X
- Verschiedene synonyme Begriffe für die Variablen:

Y	X
Regressand	Regressor
abhängige Variable	unabhängige Variable
endogene Variable	exogene Variable
zu erklärende Variable	erklärende Variable

Beispiel (Fortsetzung): Werbeausgaben und Umsatz

Soll eine Regression für den Zusammenhang zwischen Werbeausgaben und Umsatz durchgeführt werden, dann ist im ersten Schritt festzulegen, welche Variable die abhängige Variable darstellt. Da die Werbeausgaben zur Steigerung des Umsatzes getätigt werden, verwenden wir den Umsatz (Y) als abhängige Variable und die Werbeausgaben (X) als unabhängige Variable.



- Die Form der Beziehung zwischen X und Y ist durch eine ökonomische Theorie bzw. Hypothese vorgegeben oder aus dem Streuungsdiagramm ersichtlich.
- Hier: Beschränkung auf lineare Beziehungen zwischen X und Y, d.h. lineare Regression:
    - Lineare Regressionsmodelle sind einfach zu schätzen.
    - Sehr viele Beziehungen in der Realität sind tatsächlich linear.
    - Nichtlineare Zusammenhänge lassen sich durch eine geeignete Transformation in lineare Regressionsmodelle überführen.

Lineare Regressionsfunktion

- Die abhängige Variable Y soll durch eine lineare Funktion mit der unabhängigen Variablen X erklärt werden ($i = 1, 2, \ldots, n$):

$$\hat{y}_i = a + b \cdot x_i$$

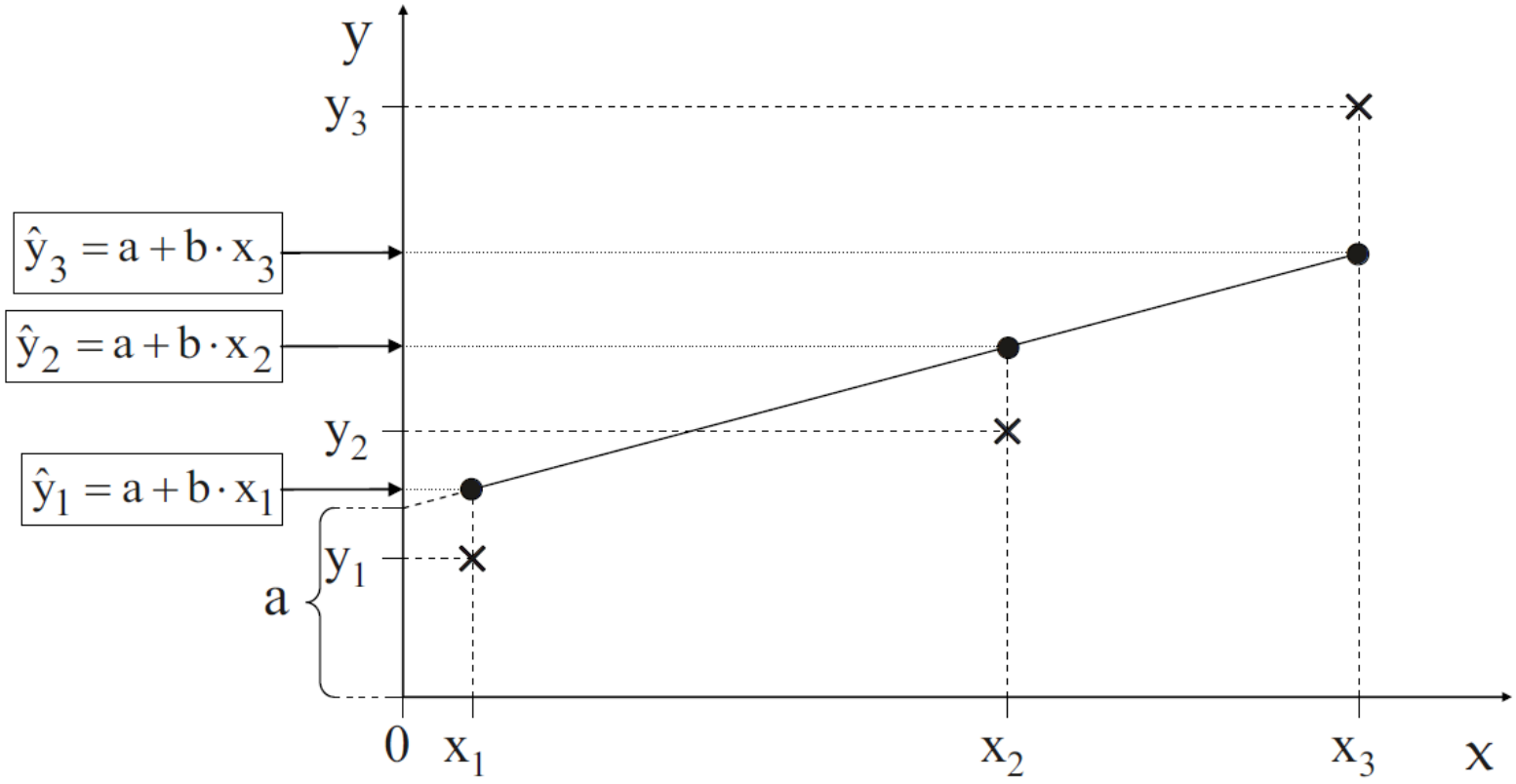
- x_i: beobachteter Wert der unabhängigen Variablen bei der i-ten statistischen Einheit
- y_i: beobachteter Wert der abhängigen Variablen bei der i-ten statistischen Einheit
- $\hat{y}_i$: Regressionswert, d.h. geschätzter Wert, für die i-ten statistische Einheit
- Die Größen a (= absolutes Glied) und b (= Steigungsparameter) heißen **Regressionskoeffizienten**. Beide sind zunächst unbekannt und müssen geschätzt werden.
- Die **Regressionsgerade** ordnet allen x-Werten eines Bereichs Funktionswerte $\hat{y}$ zu:

$$\hat{y} = a + b \cdot x$$

- Da diese Gerade nicht nur für beobachtete Werte definiert ist, entfällt der Index.

Funktionswerte

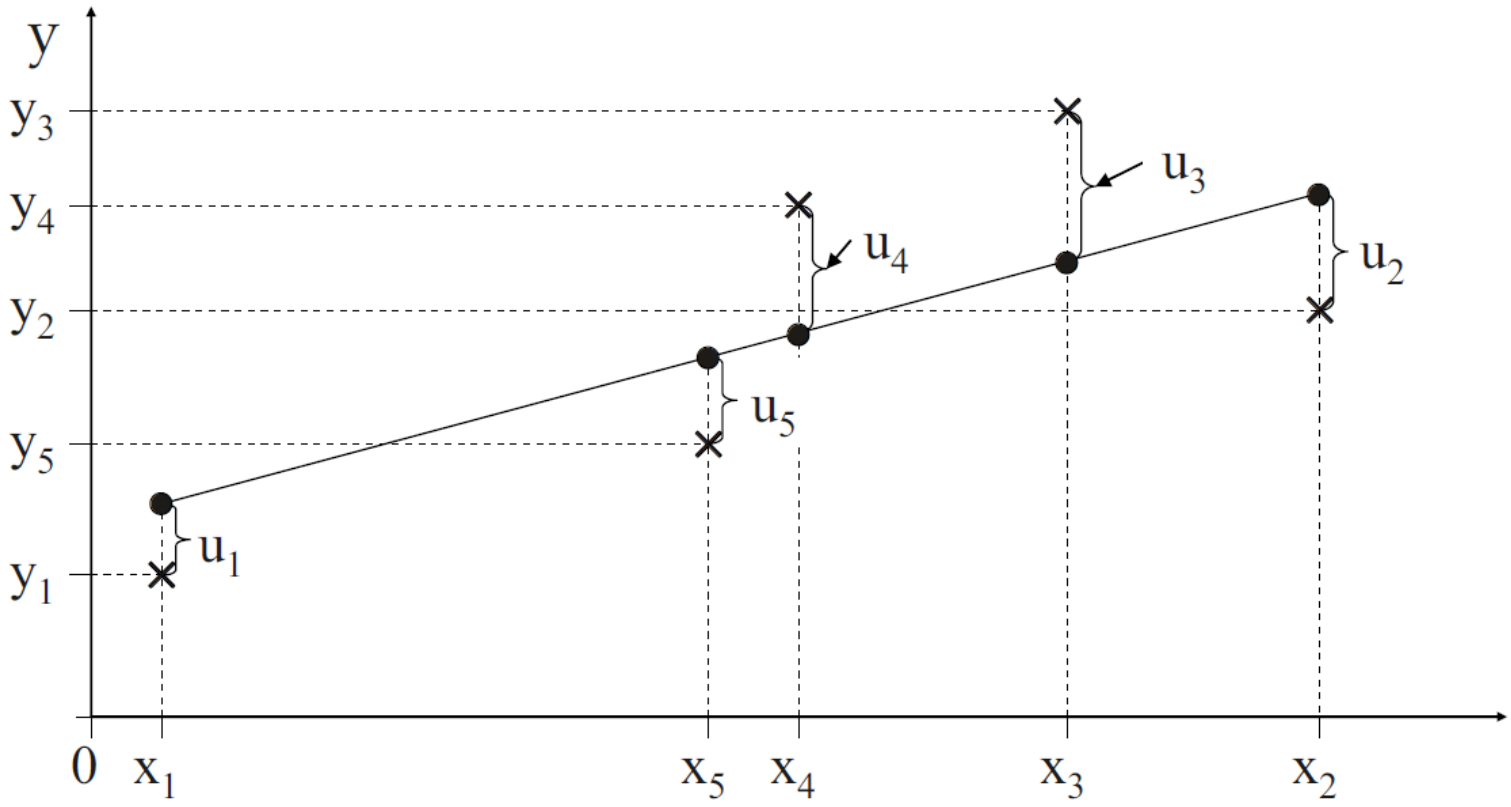
Die Funktionswerte auf der Regressionsgerade $\hat{y}_i$ werden als Regressionswerte oder geschätzte y-Werte bezeichnet.



9.2 Methode der kleinsten Quadrate und Regressionsgerade

- Zur eindeutigen Bestimmung der Regressionsgeraden benötigt man ein objektives Kriterium.
- **Fehlergrößen** (= Differenz zwischen den beobachteten und den geschätzten Werten der abhängigen Variablen):

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



Objektives Kriterium zur Bestimmung der Regressionsgeraden (I)

- $u_i > 0$: beobachteter y-Wert liegt oberhalb der Regressionsgeraden
- $u_i < 0$: beobachteter y-Wert liegt unterhalb der Regressionsgeraden
- $u_i = 0$: beobachteter y-Wert liegt auf der Regressionsgeraden
- Potenzielles Kriterium: Ausgleich der positiven und negativen Residuen

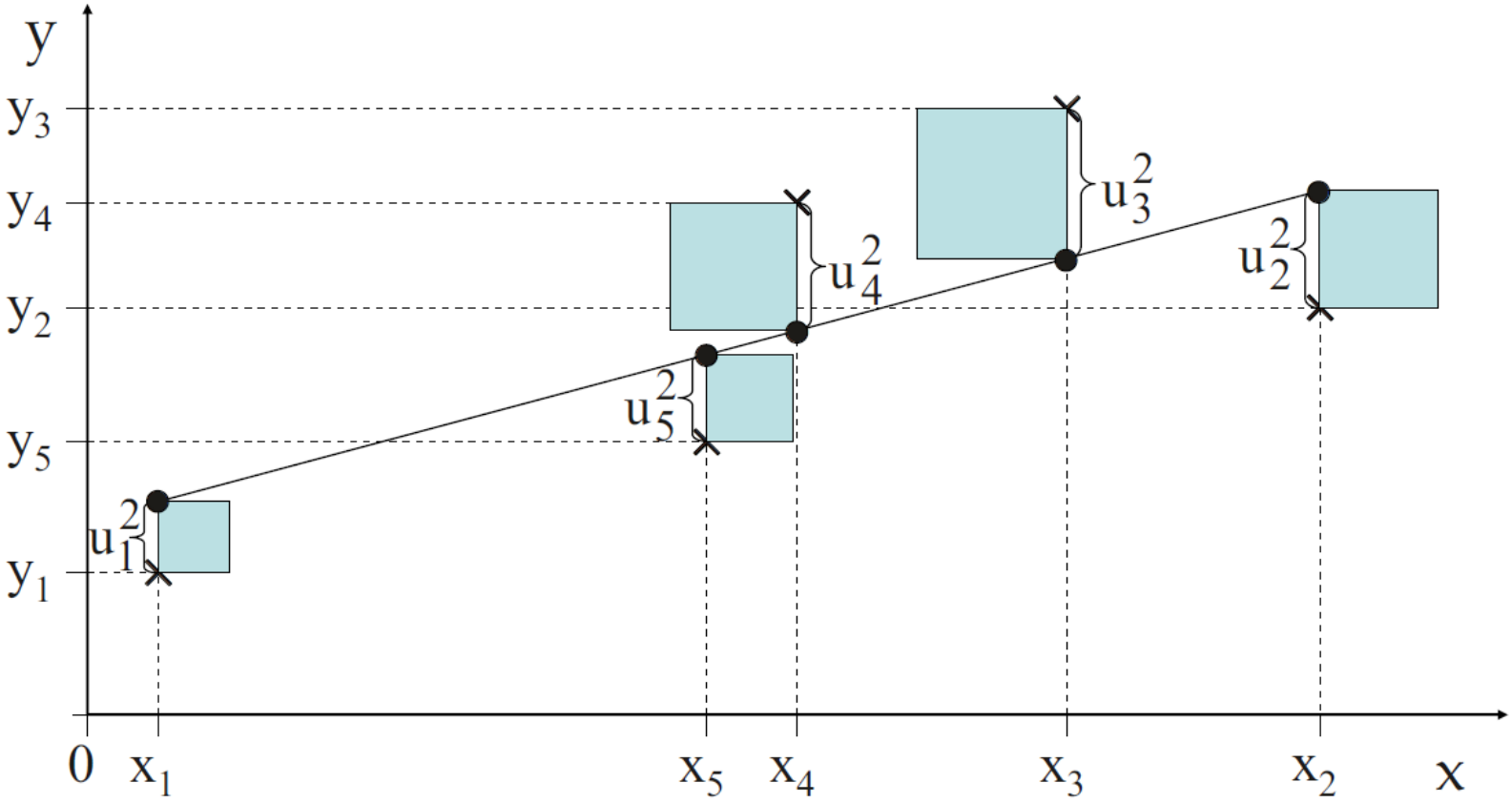
$$\sum_{i=1}^{n} u_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

- Problem: Es gibt unendlich viele Geraden, die diese Bedingung erfüllen. Sie verlaufen alle durch den Schwerpunkt.
- Kleinst-Quadrate-Kriterium: Minimierung der Summe der quadrierten Fehlergrößen (= Abweichungsquadrate)

$$Q(a,b) = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \longrightarrow \quad \min_{a,b}$$

Objektives Kriterium zur Bestimmung der Regressionsgeraden (II)

- Grafisch können die quadrierten Fehlergrößen als Quadrate dargestellt werden.
- Die Minimierung der Summe der Quadrate führt aufgrund der Quadrierung dazu, dass größere Abweichungen überproportional berücksichtigt werden.



Methode der kleinsten Quadrate

- Die Anwendung der Methode der kleinsten Quadrate führt zu einer eindeutigen Schätzung der Regressionsgeraden.
- Außerdem besitzen die Kleinst-Quadrate-Schätzer $\hat{a}$ und $\hat{b}$ bestimmte Optimalitätseigenschaften.
- Die Bestimmung der Kleinst-Quadrate-Schätzer $\hat{a}$ und $\hat{b}$ für die (unbekannten) Regressionskoeffizienten a und b erfolgt mit Hilfe der Differenzialrechnung:

  1. **Schritt:** Partielles Ableiten der Funktion Q nach a und b
  2. **Schritt:** Nullsetzen der partiellen Ableitungen (= notwendige Bedingungen für Extremwerte)
  3. **Schritt:** Auflösen der beiden sich ergebenden Gleichungen (= Normal-gleichungen) nach $\hat{a}$ und $\hat{b}$:

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

(Detaillierte Herleitung im Lehrbuch)

Interpretation der Kleinst-Quadrate-Schätzer

- Regressionswerte für die Beobachtungswerte x_i der unabhängigen Variablen X:

$$\hat{y}_i = \hat{a} + \hat{b} \cdot x_i$$

- Regressionswerte für beliebige x-Werte innerhalb des Stützbereichs (d.h. im Bereich, für den beobachtete x-Werte vorliegen):

$$\hat{y} = \hat{a} + \hat{b} \cdot x$$

- Residuen (= geschätzte Fehlergrößen):

$$\hat{u}_i = y_i - \hat{y} = y_i - \hat{a} - \hat{b} \cdot x_i$$

- Das **absolute Glied** gibt an, welchen Wert Y annimmt, wenn X null ist. Das absolute Glied lässt sich häufig nicht sinnvoll interpretieren, insbesondere dann nicht, wenn x = 0 außerhalb des Stützbereichs liegt.

- **Steigungsmaß**: Das Steigungsmaß weist aus, um wie viele Einheiten Y durchschnittlich ansteigt, wenn sich X um eine Einheit erhöht.

- **Regressionsgerade**: Ohne zusätzliche Annahmen ist die Regressionsgerade nur innerhalb des Stützbereichs interpretierbar. Sie sollte daher nur innerhalb dieses Bereichs in das Streuungsdiagramm eingezeichnet werden.

Beispiel (Fortsetzung): Werbeausgaben und Umsatz (I)

Der Umsatz ist als Zielgröße die abhängige Variable, während die Werbeausgaben als absatzpolitisches Instrument die unabhängige Variable ist. Zur Berechnung der Regressionskoeffizienten der Regression vom Umsatz (Y) auf die Werbeausgaben (X) stellen wir eine Arbeitstabelle auf:

i	x_i	y_i	$x_i \cdot y_i$	x_i^2
1	3	1	3·1 = 3	$3^2 = 9$
2	5	4	5·4 = 20	$5^2 = 25$
3	4	2	4·2 = 8	$4^2 = 16$
4	7	4	7·4 = 28	$7^2 = 49$
5	8	5	8·5 = 40	$8^2 = 64$
6	10	6	10·6 = 60	$10^2 = 100$
∑	37	22	159	263

Mit den Summen lassen sich die Regressionskoeffizienten ermitteln:

$$\hat{b} = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{6 \cdot 159 - 37 \cdot 22}{6 \cdot 263 - 37^2} = \frac{140}{209} = 0{,}670$$

Beispiel (Fortsetzung): Werbeausgaben und Umsatz (II)

$$\hat{a} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i - \hat{b} \cdot \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{6} \cdot 22 - 0{,}670 \cdot \frac{1}{6} \cdot 37 = 3{,}667 - 4{,}132 = -0{,}465$$

Bei der Interpretation ist auf die Maßeinheiten der Variablen zu achten:

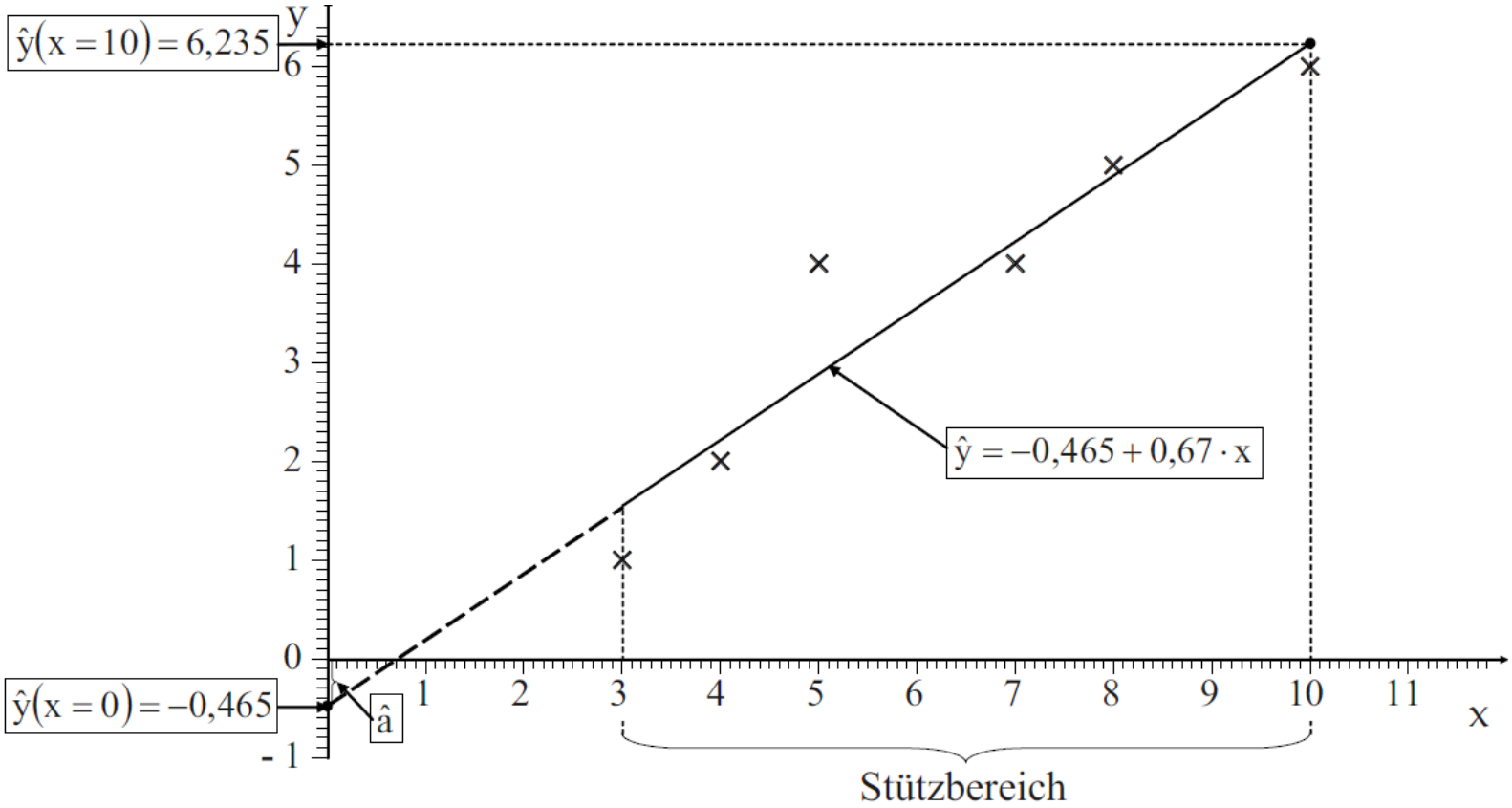
- Steigungsmaß: Erhöhen sich die Werbeausgaben um 1000€, dann steigt der Umsatz durchschnittlich um 0,67 Mio. €.
- Absolutes Glied: Das absolute Glied ist hier allein eine technische Größe, die die Lage der Regressionsgeraden festlegt. Es ist nicht weiter zu interpretieren.
- Regressionsgerade:

$$\hat{y} = \hat{a} + \hat{b} \cdot x = -0{,}465 + 0{,}670 \cdot x$$

Beispiel (Fortsetzung): Werbeausgaben und Umsatz (III)

Koordinaten zur Zeichnung der Regressionsgeraden:

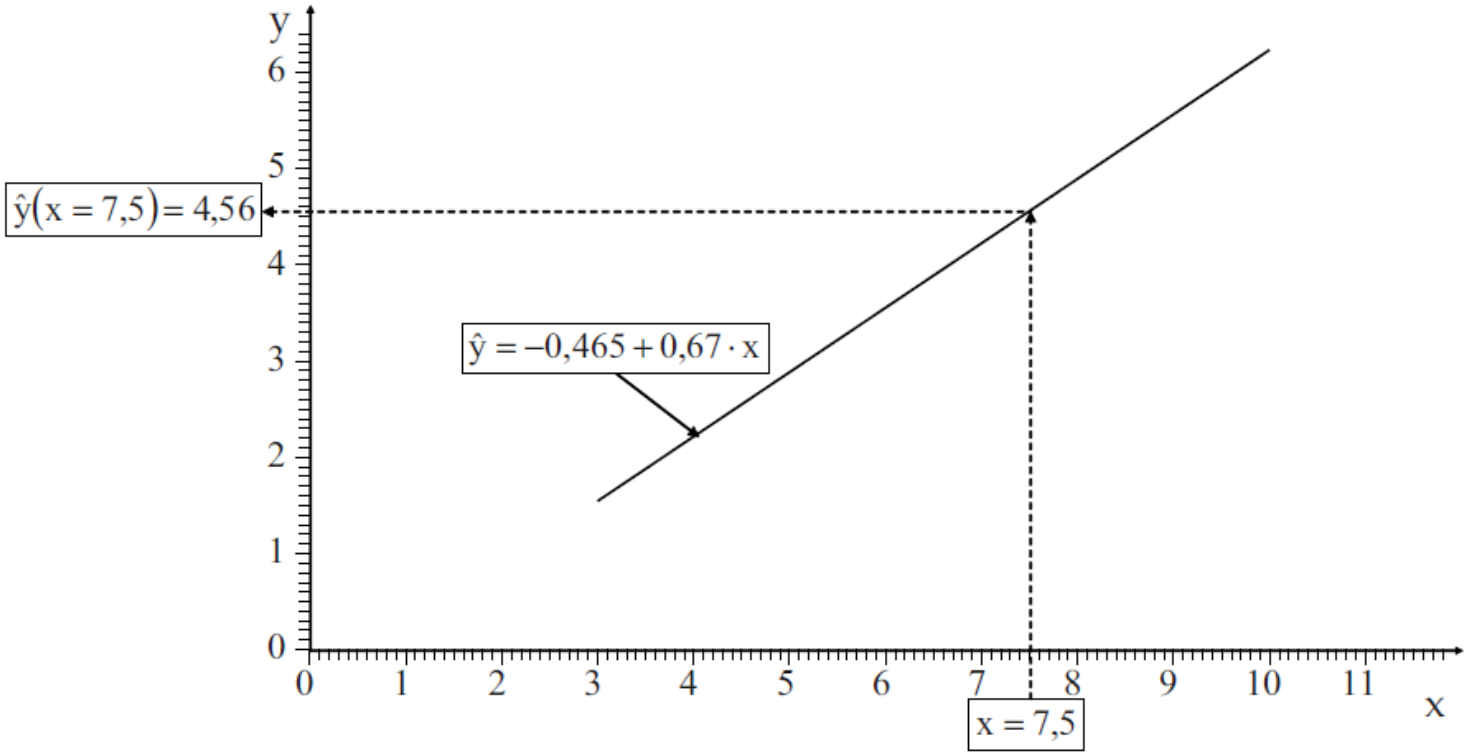
- $x = 0$: $\hat{y}(x = 0) = -0{,}465 + 0{,}670 \cdot 0 = -0{,}465$
- $x = 10$: $\hat{y}(x = 10) = -0{,}465 + 0{,}670 \cdot 10 = 6{,}235$



Beispiel (Fortsetzung): Werbeausgaben und Umsatz (IV)

Welcher Umsatz ist zu erwarten, wenn die Werbeausgaben 7,5 Tsd. € betragen?

- Der Wert x = 7,5 wird in die Regressionsfunktion eingesetzt:
- $x = 7{,}5$: $\hat{y}(x = 7{,}5) = -0{,}465 + 0{,}670 \cdot 7{,}5 = 4{,}56$
- Bei Werbeausgaben von 7,5 Tsd. € ist also mit einem Umsatz von 4,56 Mio. € zu rechnen.



Beispiel (Fortsetzung): Heimwerkerartikel (I)

Zur Ermittlung der Regressionsbeziehung zwischen dem Absatz (Y) und dem Preis (X) erstellen wir eine Arbeitstabelle:

i	x_i	y_i	$x_i \cdot y_i$	x_i^2
1	44	4	44·4 = 176	44^2 = 1.936
2	40	8	40·8 = 320	40^2 = 1.600
3	42	6	42·6 = 252	42^2 = 1.764
4	46	3	46·3 = 138	46^2 = 2.116
5	36	12	36·12 = 432	36^2 = 1.296
6	37	11	37·11 = 407	37^2 = 1.369
7	39	7	39·7 = 273	39^2 = 1.521
∑	284	51	1998	11602

Regressionskoeffizienten:

$$\hat{b} = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{7 \cdot 1998 - 284 \cdot 51}{7 \cdot 11602 - 284^2} = \frac{-498}{558} = -0{,}892$$

Beispiel (Fortsetzung): Heimwerkerartikel (II)

$$\hat{a} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i - \hat{b} \cdot \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{7} \cdot 51 - (-0{,}892) \cdot \frac{1}{7} \cdot 284 = 7{,}286 + 36{,}190 = 43{,}476$$

Interpretation:

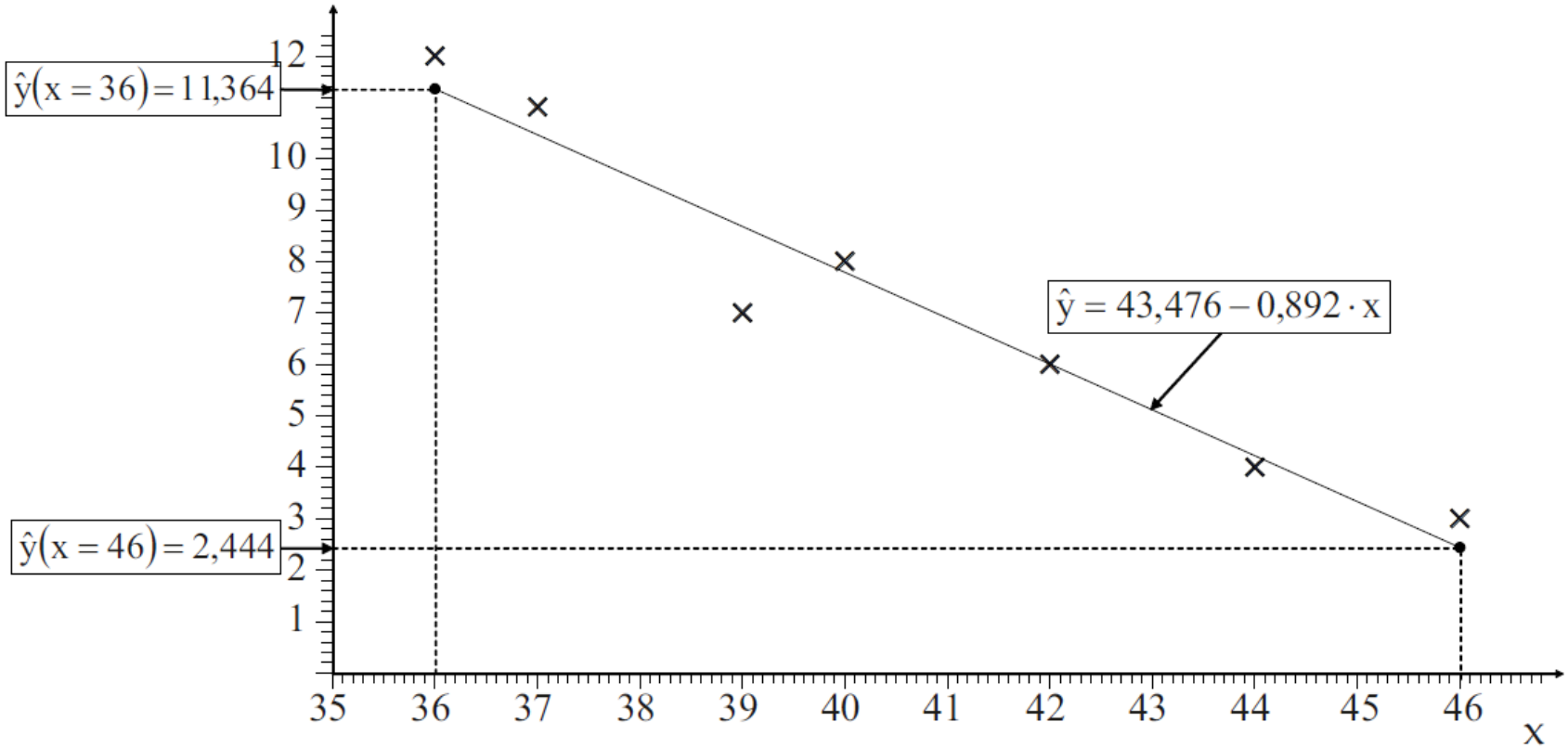
- Steigungsmaß: Erhöht sich der Preis um einen Euro, dann sinkt der Absatz durchschnittlich um 0,892 Tsd. Stück.
- Absolutes Glied: Das absolute Glied ist hier allein eine technische Größe, die die Lage der Regressionsgeraden festlegt. Es ist nicht weiter zu interpretieren.
- Regressionsgerade:

$$\hat{y} = \hat{a} + \hat{b} \cdot x = 43{,}476 - 0{,}892 \cdot x$$

Beispiel (Fortsetzung): Heimwerkerartikel (III)

Koordinaten zur Zeichnung der Regressionsgeraden:

- $x = 36$: $\hat{y}(x = 36) = 43{,}476 - 0{,}892 \cdot 36 = 11{,}364$
- $x = 46$: $\hat{y}(x = 46) = 43{,}476 - 0{,}892 \cdot 46 = 2{,}444$



Steigungsmaß und Korrelationskoeffizient (I)

- Durch einfache Umformung lässt sich der Regressionskoeffizient $\hat{b}$ auf die Kovarianz s_{xy} und die Varianz s_x^2 zurückführen.

- Steigungsmaß:

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

- Division des Zählers und Nenners durch n^2:

$$\hat{b} = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - \bar{x} \cdot \bar{y}}{\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \bar{x}^2}$$

→ Alternativ kann das Steigungsmaß mithilfe der Kovarianz und der Varianz der unabhängigen Variablen (X) berechnet werden.

Steigungsmaß und Korrelationskoeffizient (II)

- Darstellung des Steigungsmaßes:

$$\hat{b} = \frac{s_{xy}}{s_x^2}$$

- Das Vorzeichen des Steigungsmaßes $\hat{b}$ wird durch die Kovarianz bestimmt. Immer dann, wenn die Kovarianz negativ ist, nimmt auch das Steigungsmaß einen negativen Wert an und umgekehrt.

- Darstellung des Korrelationskoeffizienten:

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

- Das Vorzeichen des Korrelationskoeffizienten wird ebenfalls durch die Kovarianz bestimmt. Der Korrelationskoeffizient r und das Steigungsmaß $\hat{b}$ haben also das gleiche Vorzeichen.

- Beziehung zwischen Steigungsmaß $\hat{b}$ und dem Korrelationskoeffizienten r:

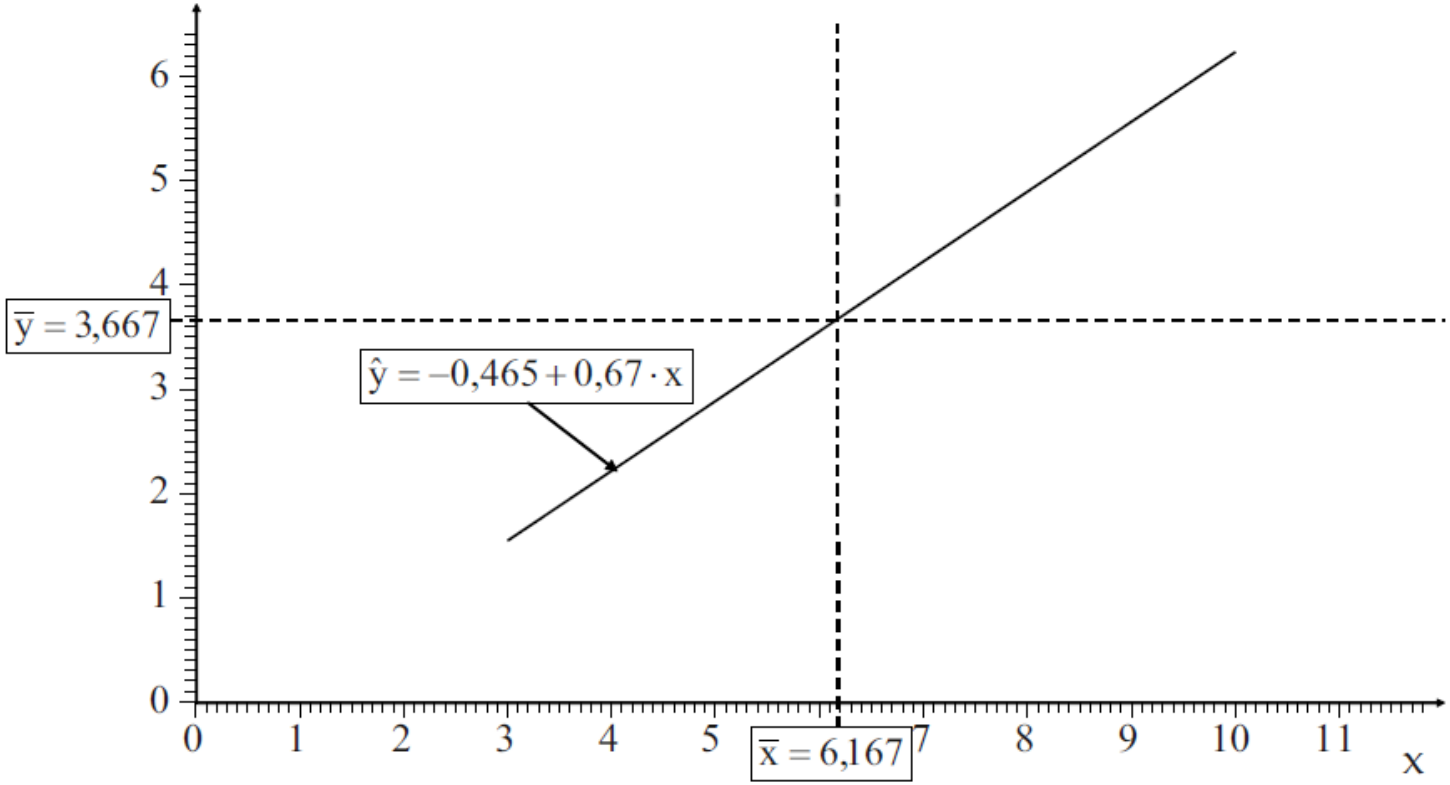
$$\hat{b} = \frac{s_{xy}}{s_x^2} = \frac{s_{xy}}{s_x \cdot s_x} = \frac{s_{xy}}{s_x \cdot s_y} \cdot \frac{s_y}{s_x} = r \cdot \frac{s_y}{s_x}$$

9.3 Eigenschaften der Kleinst-Quadrate-Regression

(Herleitungen: s. Lehrbuch)

1. Die Regressionsgerade verläuft durch den Schnittpunkt der beiden Variablenmittelwerte, dem sogenannten Datenschwerpunkt $(\bar{x}, \bar{y})$. Für $x = \bar{x}$ nimmt die Regressionsfunktion $\hat{y}$ genau den Wert $\bar{y}$ an

Beispiel (Fortsetzung): Werbeausgaben und Umsatz



2. Die **Ausgleichseigenschaft** besagt, dass die Summe der positiven und negativen Residuen ($\hat{u}_i = y_i - \hat{y}_i$) gleich groß ist. Die Summe aller Residuen ist deshalb null:

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

Damit ist auch das arithmetische Mittel der Residuen gleich null:

$$\bar{\hat{u}}_i = \frac{1}{n} \cdot \sum_{i=1}^{n} \hat{u}_i = 0$$

3. **Gleiche Mittelwerte**: Die geschätzten Werte der abhängigen Variablen haben das gleiche arithmetische Mittel wie die beobachteten y-Werte:

$$\bar{\hat{y}} = \bar{y}$$

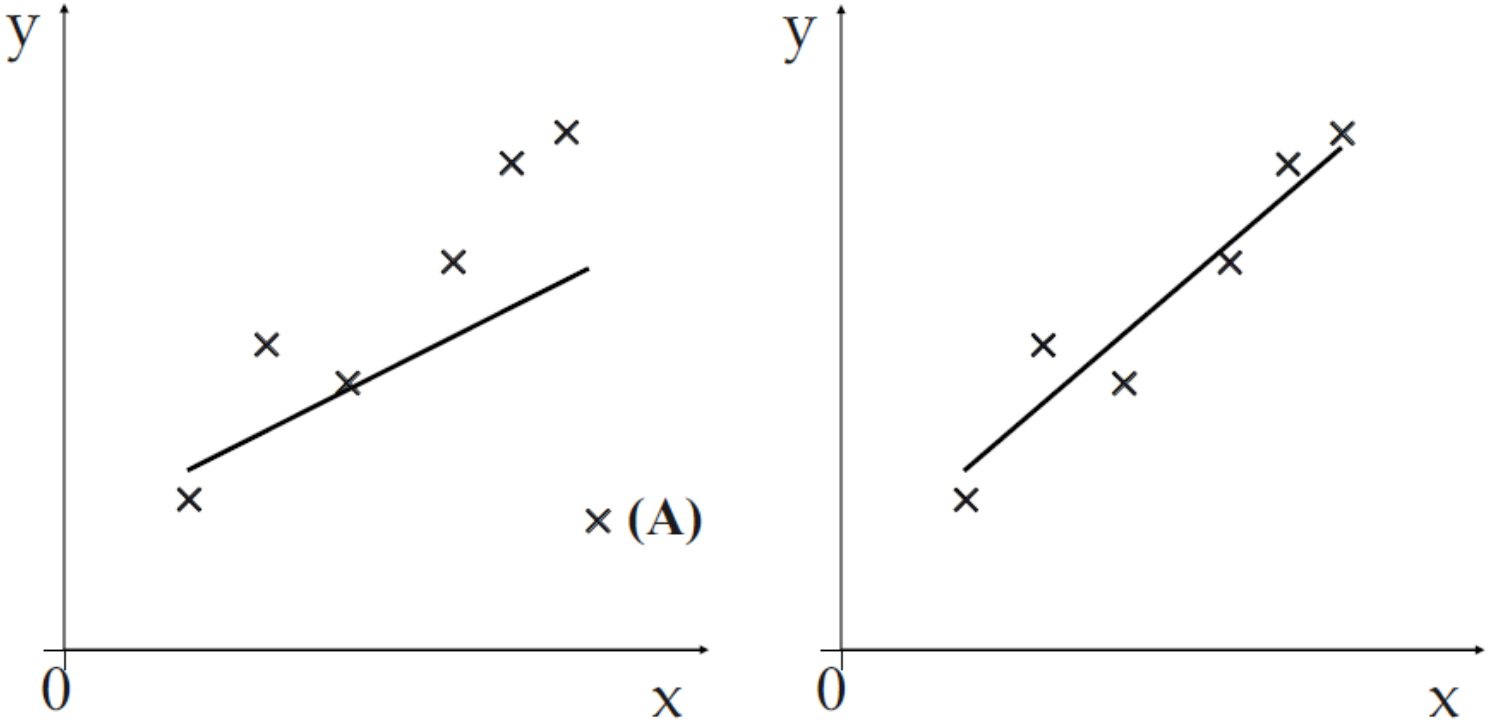
4. Die Residuen sind mit den Werten der unabhängigen Variablen nicht korreliert. Die Kovarianz zwischen beiden Größen ist deshalb null:

$$s_{\hat{u}x} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\hat{u}_i - \bar{\hat{u}}\right) \cdot (x_i - \bar{x}) = 0$$

Sensitivität gegenüber Ausreißern

Ausreißerwerte mit sehr großen absoluten Residuen beeinflussen den Verlauf der Regressionsgeraden (Sensitivität gegenüber Ausreißern) deutlich.

Regressionsgerade mit und ohne Ausreißer



a) mit Ausreißer (A)

b) ohne Ausreißer

Beispiel (Fortsetzung): Werbeausgaben und Umsatz (I)

Gehen wir davon aus, dass in einer siebten Periode trotz hoher Werbeausgaben von 11 Tsd. € nur ein relativ geringer Umsatz von 1 Mio. € erzielt wurde.

i	x_i	y_i	$x_i \cdot yi$	x_i^2
1	3	1	3·1 = 3	$3^2 = 9$
2	5	4	5·4 = 20	$5^2 = 25$
3	4	2	4·2 = 8	$4^2 = 16$
4	7	4	7·4 = 28	$7^2 = 49$
5	8	5	8·5 = 40	$8^2 = 64$
6	10	6	10·6 = 60	$10^2 = 100$
7	11	1	11·1 = 11	$11^2 = 121$
∑	48	23	170	384

Beispiel (Fortsetzung): Werbeausgaben und Umsatz (II)

- Neue Regressionskoeffizienten:

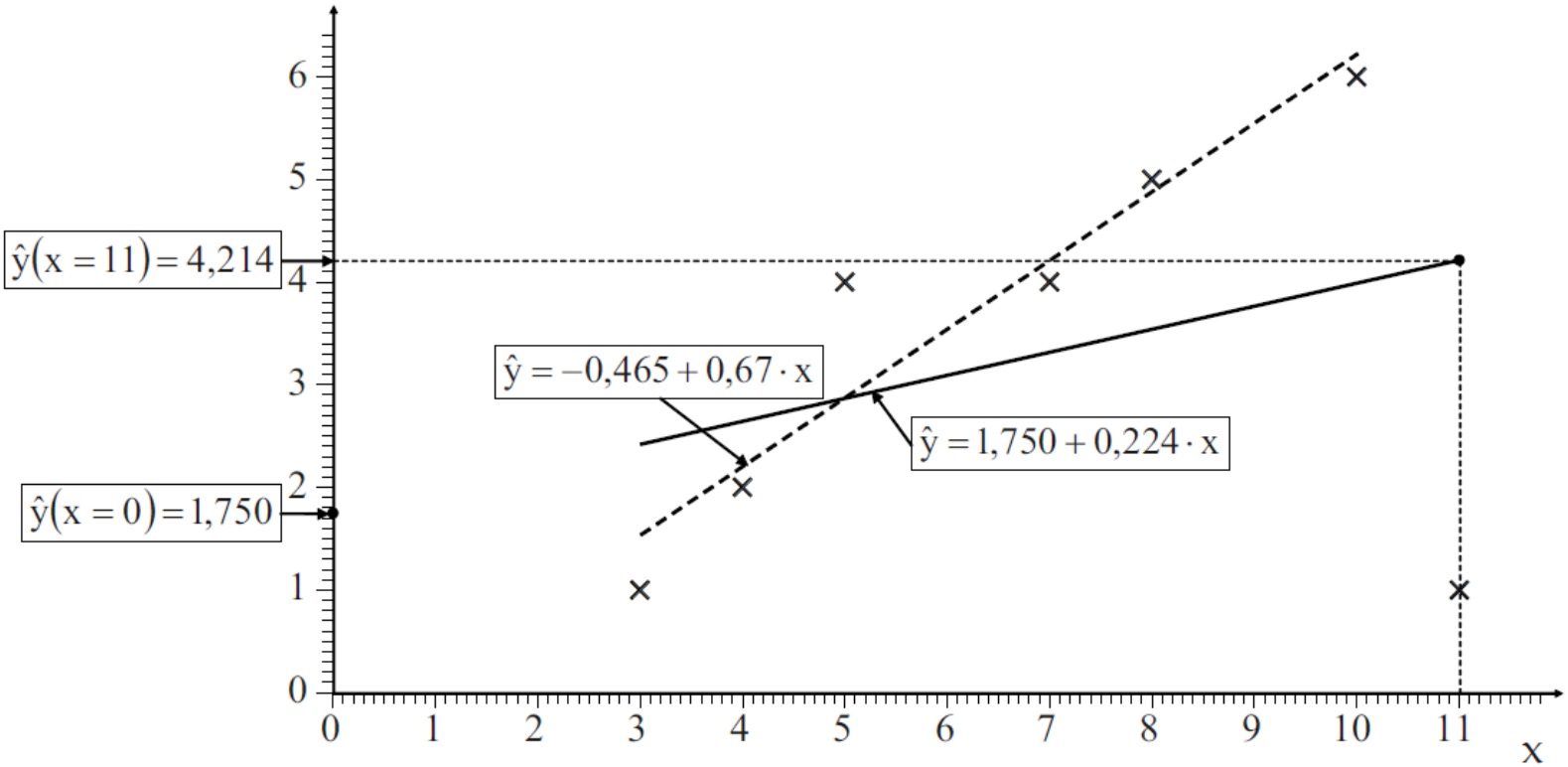
$$\hat{b} = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{7 \cdot 170 - 48 \cdot 23}{7 \cdot 384 - 48^2} = \frac{86}{384} = 0{,}224$$

$$\hat{a} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i - \hat{b} \cdot \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{7} \cdot 23 - 0{,}224 \cdot \frac{1}{7} \cdot 48 = 3{,}286 - 1{,}536 = 1{,}750$$

- Neue Regressionsgerade:

$$\hat{y} = \hat{a} + \hat{b} \cdot x = 1{,}750 + 0{,}224 \cdot x$$

Beispiel (Fortsetzung): Werbeausgaben und Umsatz (III)



Es fällt auf, dass die Regressionsgerade ohne Berücksichtigung des Ausreißerwertes (gestrichelte Linie) deutlich steiler verläuft als die neue Regressionsgerade (durchgehende Linie).

9.4 Güte der Anpassung

- Probleme der Anpassung einer Regression werden visuell bereits aus einem Streuungsdiagramm mit der Regressionsgeraden offenbar.
- Bei Vorliegen von Ausreißern spiegelt die Regressionsgerade nicht mehr unbedingt die Beziehung zwischen der Masse der Beobachtungen wider.
- Globales Maß für die Güte der Anpassung der Regression wünschenswert.
- Ein solches Maß ist der **Determinationskoeffizient** (das **Bestimmtheitsmaß**).
- Zur Konstruktion wird zunächst die Varianz der abhängigen Variablen zerlegt:

$$s_y^2 = s_{\hat{y}}^2 + s_{\hat{u}}^2$$

- 1. Komponente: Varianz der Regressionswerte

$$s_{\hat{y}}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\hat{y}_i - \bar{\hat{y}}\right)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad \text{da} \quad \bar{\hat{y}} = \hat{y}$$

- 2. Komponente: Varianz der Residuen

$$s_{\hat{u}}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\hat{u}_i - \bar{\hat{u}}\right)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \hat{u}_i^2 \qquad \text{da} \quad \bar{\hat{u}} = 0$$

Interpretation der Varianzzerlegung (I)

- Die zerlegte Varianz lässt sich wie folgt schreiben:

$$s_y^2 = s_{\hat{y}}^2 + s_{\hat{u}}^2$$

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

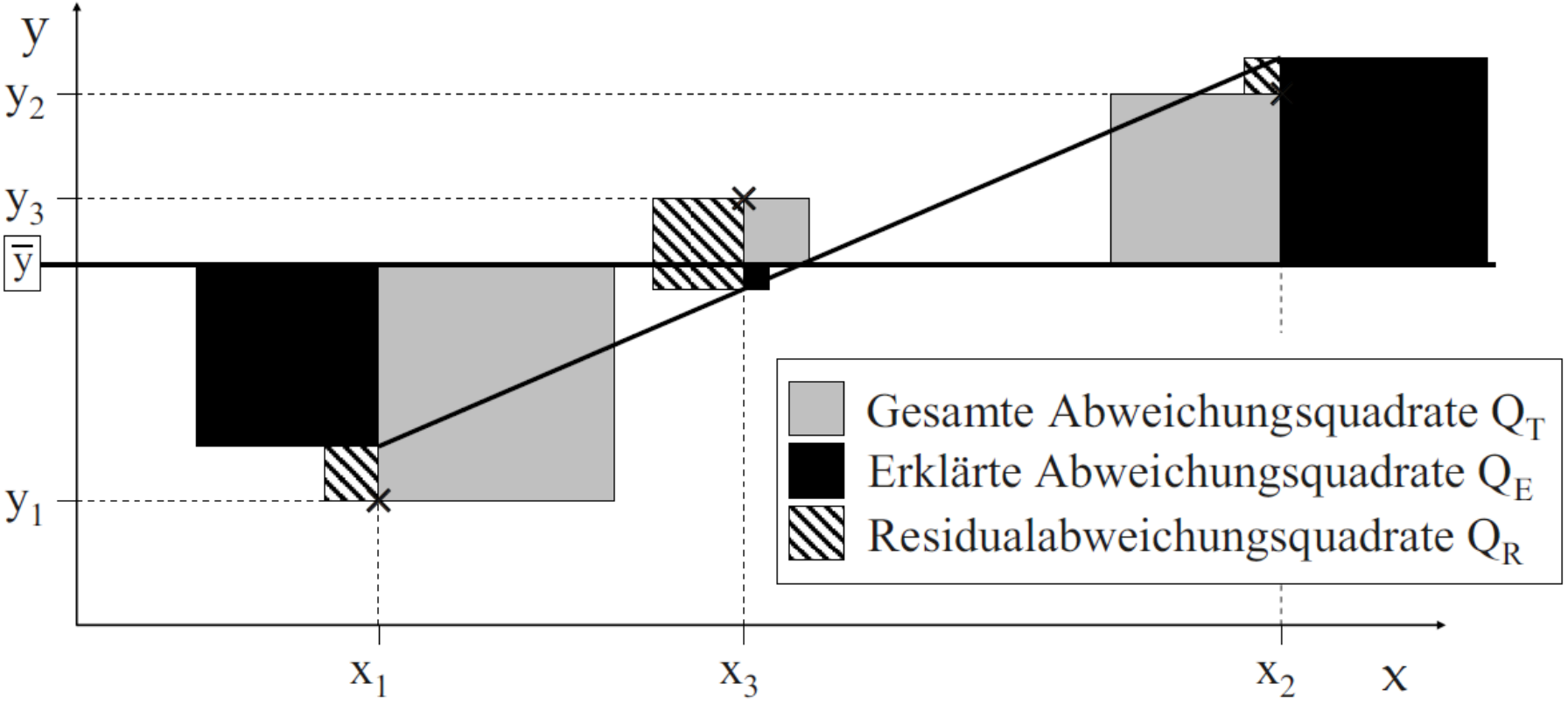
- Multipliziert mit n:

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{Q_T} = \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{Q_E} + \underbrace{\sum_{i=1}^{n} \hat{u}_i^2}_{Q_R}$$

- Q_T: Totale Abweichungsquadratsumme
- Q_E: Durch die Regression erklärte Abweichungsquadratsumme
- Q_R: Nicht erklärte Abweichungsquadratsumme (Abweichungsquadratsumme der Residuale)

Interpretation der Varianzzerlegung (II)

Abweichungsquadrate



Interpretation der Varianzzerlegung (III)

- Ohne Kenntnis des jeweiligen x-Wertes wäre der beste Vorhersagewert (Prädiktor) für alle y-Werte das arithmetische Mittel $\bar{y}$.
- Wenn X keinen Einfluss auf Y hat, würde ein y-Wert auch bei Kenntnis der x-Werte stets durch $\bar{y}$ geschätzt werden, d.h. es würde $\hat{y}_i = \bar{y}$ gelten.
- Die Regressionsgerade wäre dann eine Parallele zur x-Achse, so dass das Steigungsmaß den Wert null annehmen würde:

$$\hat{y}_i = \hat{a} + \left(\hat{b} = 0\right) \cdot x_i = \hat{a} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i - \hat{b} \cdot \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i = \bar{y}$$

- In diesem Fall ist die durch die Regression erklärte Abweichungsquadratsumme Q_E gleich null und Q_R stimmt mit der totalen Abweichungsquadratsumme Q_T überein.
- Die Regressionsanalyse liefert in dieser Situation keinen zusätzlichen Erklärungsbeitrag.

Definition des Determinationskoeffizienten

- Die Anpassung der Regression ist dann als besonders gut zu beurteilen, wenn die quadrierten Residuen möglichst niedrig ausfallen.
- Dann nimmt Q_R einen geringen und Q_E einen hohen Wert an.
- **Determinationskoeffizient:**

$$R^2 = \frac{Q_E}{Q_T} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n} \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{s_{\hat{y}}^2}{s_y^2}$$

- Der Determinationskoeffizient (das Bestimmtheitsmaß) gibt den Anteil der Varianz der abhängigen Variablen an, der durch die Regression von Y auf X erklärt wird. Der Determinationskoeffizient lässt sich prozentual interpretieren.
- Wertebereich: $0 \leq R^2 \leq 1$
- Rechenformel:

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

Beispiel (Fortsetzung): Werbeausgaben und Umsatz

Für die Regression des Umsatzes (Y) auf die Werbeausgaben (X) wird der Determinationskoeffizient bestimmt.

i	x_i	y_i	$\hat{y}_i = -0{,}465 + 0{,}670 \cdot x_i$	y_i^2	$\hat{y}_i^2$
1	3	1	-0,465 + 0,670·3 = 1,545	1	2,387
2	5	4	-0,465 + 0,670·5 = 2,885	16	8,323
3	4	2	-0,465 + 0,670·4 = 2,215	4	4,906
4	7	4	-0,465 + 0,670·7 = 4,225	16	17,851
5	8	5	-0,465 + 0,670·8 = 4,895	25	23,961
6	10	6	-0,465 + 0,670·10 = 6,235	36	38,875
∑	37	22	22,000	98	96,303

Rechenformel:

$$R^2 = \frac{n \cdot \sum_{i=1}^{n} \hat{y}_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}{n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2} = \frac{6 \cdot 96{,}303 - 22^2}{6 \cdot 98 - 22^2} = \frac{93{,}818}{104} = 0{,}902$$

Interpretation: Durch den Regressionsansatz werden 90,2% der Varianz des Umsatzes erklärt.

Berechnung von R^2 mit dem Steigungsmaß $\hat{b}$

- Ausgangspunkt:

$$s_{\hat{y}}^2 = \hat{b}^2 \cdot s_x^2$$

- R^2 kann dann ohne Berechnung der Regressionswerte bestimmt werden:

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{\hat{b}^2 \cdot s_x^2}{s_y^2} = \hat{b}^2 \cdot \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2} = \hat{b}^2 \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Berechnung von R^2 mit dem Korrelationskoeffizienten r

- Setzt man in

$$R^2 = \frac{\hat{b}^2 \cdot s_x^2}{s_y^2}$$

- die Beziehung

$$r \cdot \frac{s_y}{s_x}$$

- ein, erhält man

$$R^2 = \frac{\left(r \cdot \frac{s_y}{s_x}\right)^2 \cdot s_x^2}{s_y^2} = r^2 \cdot \frac{s_y^2}{s_x^2} \cdot \frac{s_x^2}{s_y^2} = r^2$$

Beispiel (Fortsetzung): Werbeausgaben und Umsatz

- Arithmetisches Mittel und Abweichungsquadrate von X:

$$\bar{x} = \frac{1}{6} \cdot \sum_{i=1}^{6} x_i = \frac{1}{6} \cdot (3 + 5 + 4 + 7 + 8 + 10) = \frac{1}{6} \cdot 37 = 6{,}167$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = (3 - 6{,}167)^2 + (5 - 6{,}167)^2 + (4 - 6{,}167)^2 + (7 - 6{,}167)^2 + (8 - 6{,}167)^2 + (10 - 6{,}167)^2 = 34{,}833$$

- Analoge Berechnung des arithmetischen Mittels und der Abweichungsquadrate von Y

- Determinationskoeffizient:

$$R^2 = \hat{b}^2 \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 0{,}670^2 \cdot \frac{34{,}833}{17{,}333} = 0{,}902$$

- Zudem lässt sich der Determinationskoeffizient auch mit Hilfe des Korrelationskoeffizienten berechnen:

$$R^2 = r^2 = 0{,}95^2 \approx 0{,}902$$