Statistik

Themengebiet: Deskriptive Statistik

VE 04: Korrelation und Regressionsanalyse

Wintersemester 2024/2025

Dozent: Prof. Dr. Sören Gröttrup

Vorlage von Prof. Dr. Max Krüger

Fakultät Informatik, Technische Hochschule Ingolstadt (THI)

Der vorliegende Foliensatz ist ausschließlich für den persönlichen, vorlesungsinternen Gebrauch im Rahmen der Vorlesungen zur Mathematik und Statistik für anwendungsorientierte Informatik an der Fakultät Informatik der Technischen Hochschule Ingolstadt (THI) bestimmt.

Der Foliensatz wird kontinuierlich korrigiert, aktualisiert und erweitert.

– Urheberrechtlich geschütztes Material –

Lerninhalte VE 04

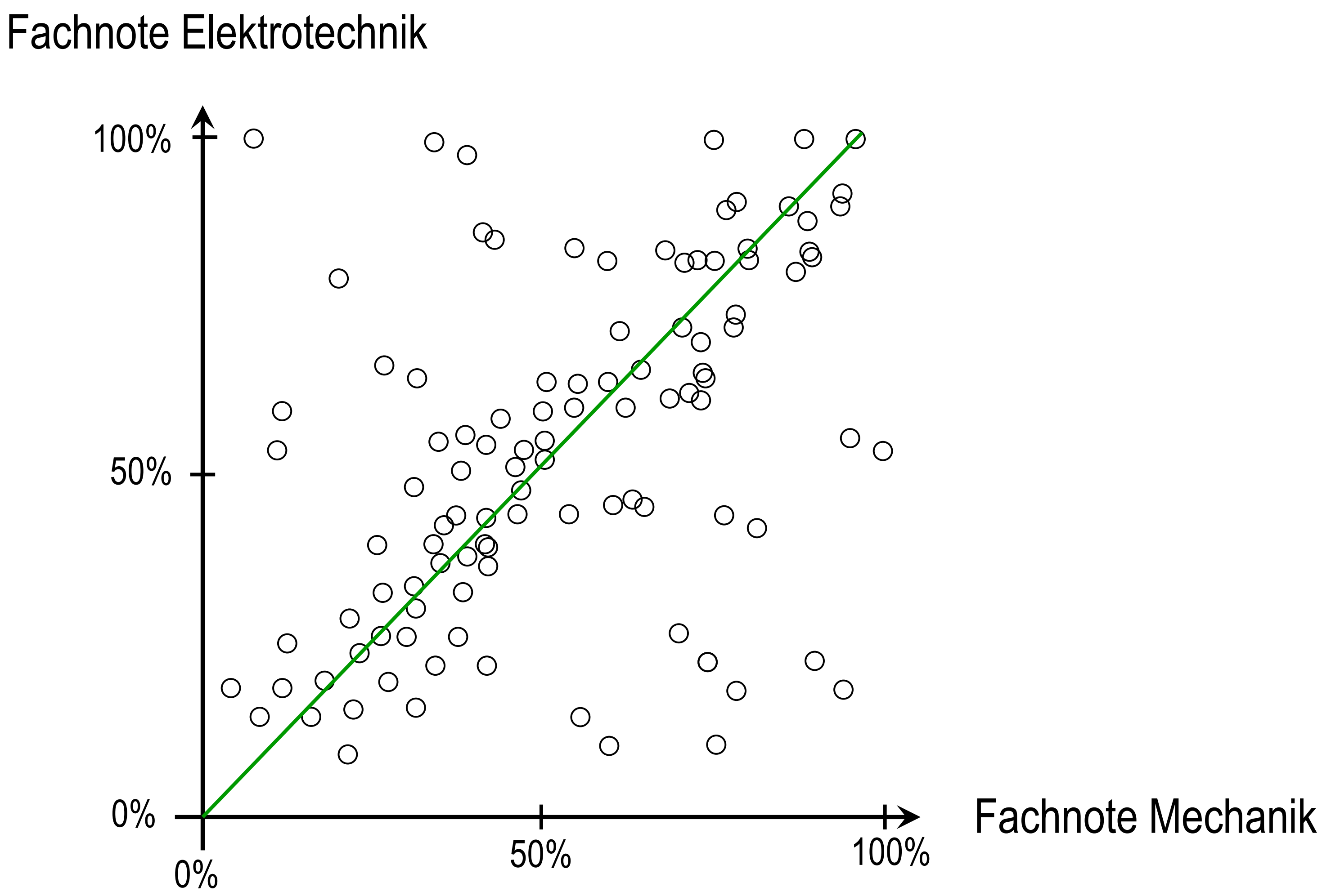
Thema: Korrelation und Regressionsanalyse

- Arten kausaler Zusammenhänge
- Empirische Kovarianzen
- Zielsetzung der Korrelationsbetrachtung
- Korrelationskoeffizienten
- Zielsetzung der Regressionsanalyse
- Regression, Regressor und Regressand
- Lineare Regression und Regressionsgrade
- Methode der kleinsten Quadrate zur Regressionsgeradenberechnung
- Bestimmtheitsmaß für die lineare Regression
- Quadratische und mehrfache Regression

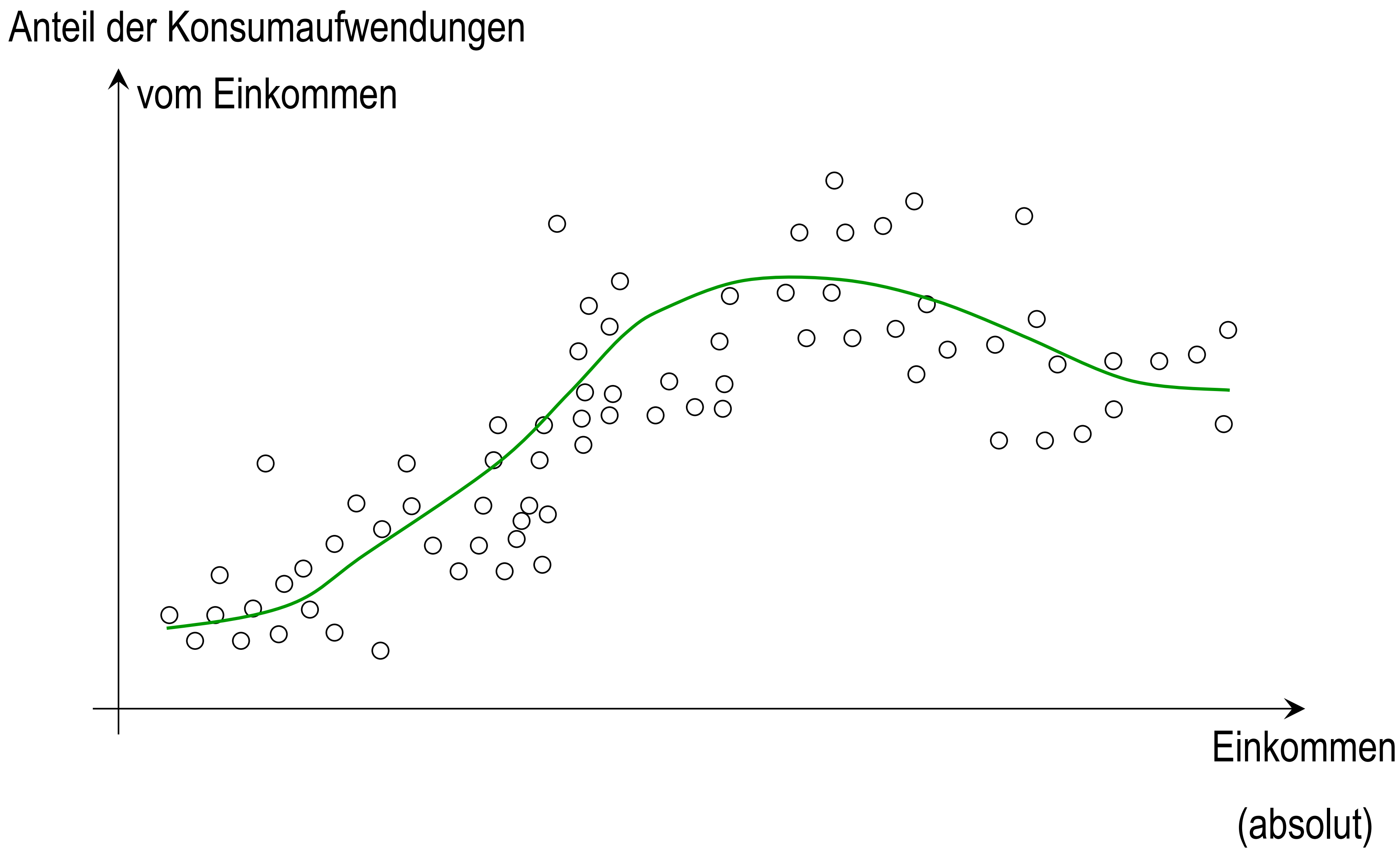
Inhaltsverzeichnis VE 04

VE 04: Korrelation und Regressionsanalyse

Einstieg (1): Abhängigkeit von Fachnoten



Einstieg (2): Abhängigkeit Einkommen und Konsum



Inhaltsverzeichnis VE 04

VE 04: Korrelation und Regressionsanalyse

Zielsetzung von Korrelationsbetrachtungen

Bei der Korrelationsbetrachtung werden mögliche Zusammenhänge zwischen gemeinsam auftretenden Merkmalen erfasst:

- Qualitative Beschreibung des (tendenziellen) Zusammenhangs des Auftretens der Merkmalsausprägungen
- Quantifizierung der gegenseitigen Abhängigkeit/Unabhängigkeit des Auftretens der Merkmalsausprägungen
- Keine Betrachtung von Kausalitätsrichtungen, die Merkmale werden gleichberechtigt erfasst.

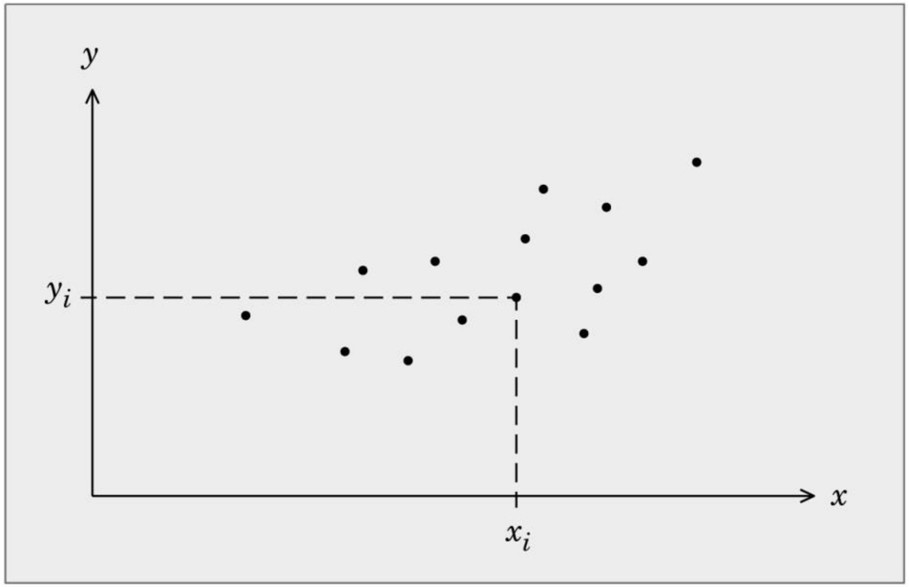


BILD 3.1 Punkte im Streudiagramm

Bildquelle: [7,p.84 Bild 3.1]

Empirische Kovarianz

(Empirische) Kovarianz

Sei $(x, y) = ((x_1, y_1), \ldots, (x_n, y_n))$ eine Stichprobe vom Umfang n mit $n \geq 2$, bei der für jedes Stichprobenelement zwei Merkmale x_i und y_i erfasst werden, dann ist die **empirische Kovarianz** $\operatorname{cov}(x, y)$ der Stichprobe (x, y) definiert durch

$$\operatorname{cov}(x, y) := \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y}),$$

wobei $\overline{x}$ und $\overline{y}$ die empirischen arithmetischen Mittel der Einzelmerkmale $x = (x_1, \ldots, x_n)$ und $y = (y_1, \ldots, y_n)$ darstellen.

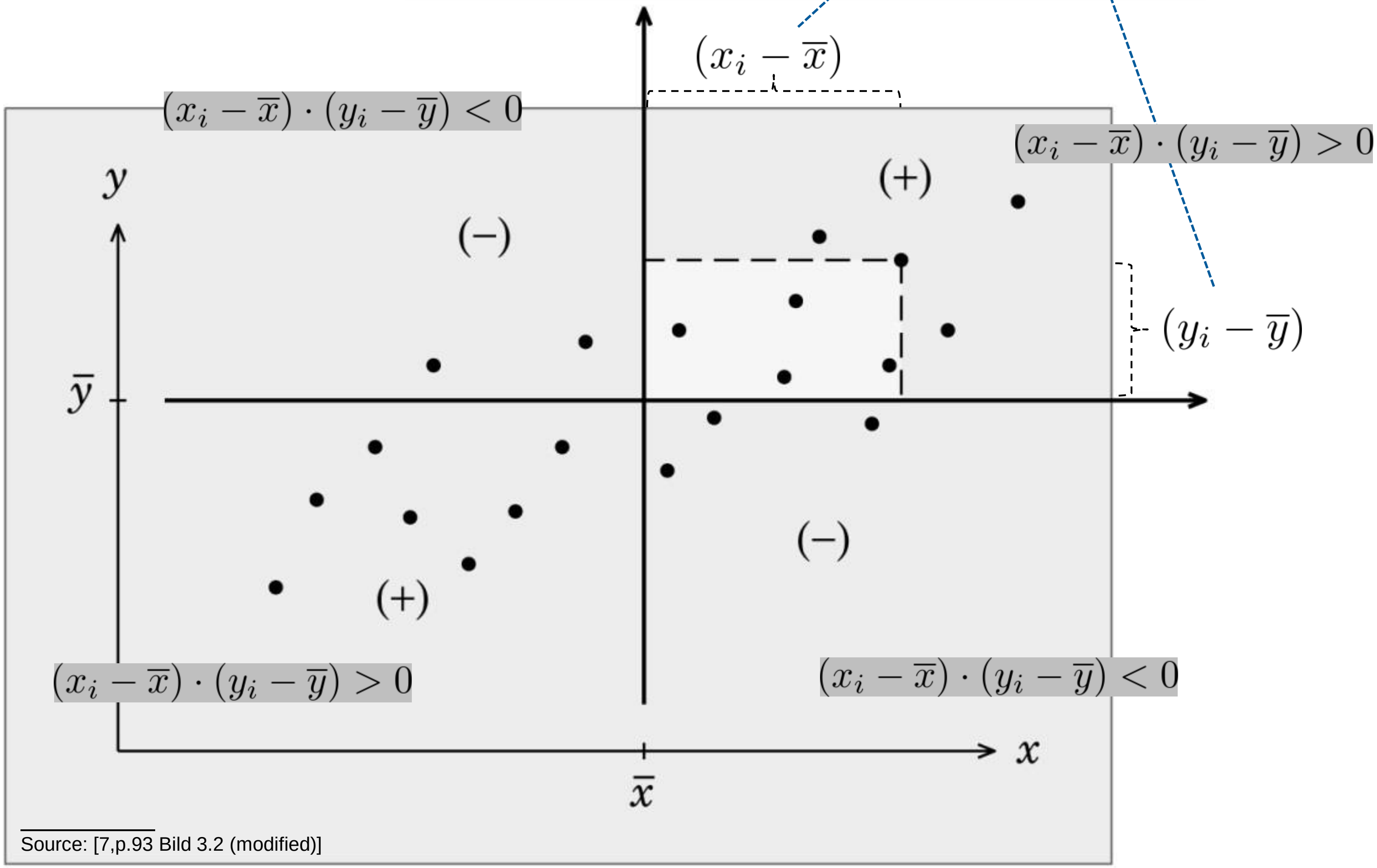
Anmerkung:

Eine vereinfachte Berechnung wird durch folgenden Zusammenhang ermöglicht:

$$\operatorname{cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i \cdot y_i) - \overline{x} \cdot \overline{y}$$

Erläuterung der empirischen Kovarianz

$$\mathrm{cov}(\mathbf{x},\mathbf{y}) := \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\cdot(y_i-\overline{y})$$



Source: [7,p.93 Bild 3.2 (modified)]

Interpretation der empirischen Kovarianz

Interpretation:

- Ein positiver Wert der Kovarianz $\operatorname{cov}(x, y) > 0$ beschreibt eine gemeinsame Tendenz der Merkmalsausprägungen (x_i, y_i) .
- Ein negativer Wert der Kovarianz $\operatorname{cov}(x, y) < 0$ beschreibt eine entgegengesetzte Tendenz der Merkmalsausprägungen (x_i, y_i) .
- Ein Wert der Kovarianz $\operatorname{cov}(x, y) \approx 0$ nahe bei Null lässt keine Tendenz von gemeinsamen Merkmalsausprägungen erschließen.

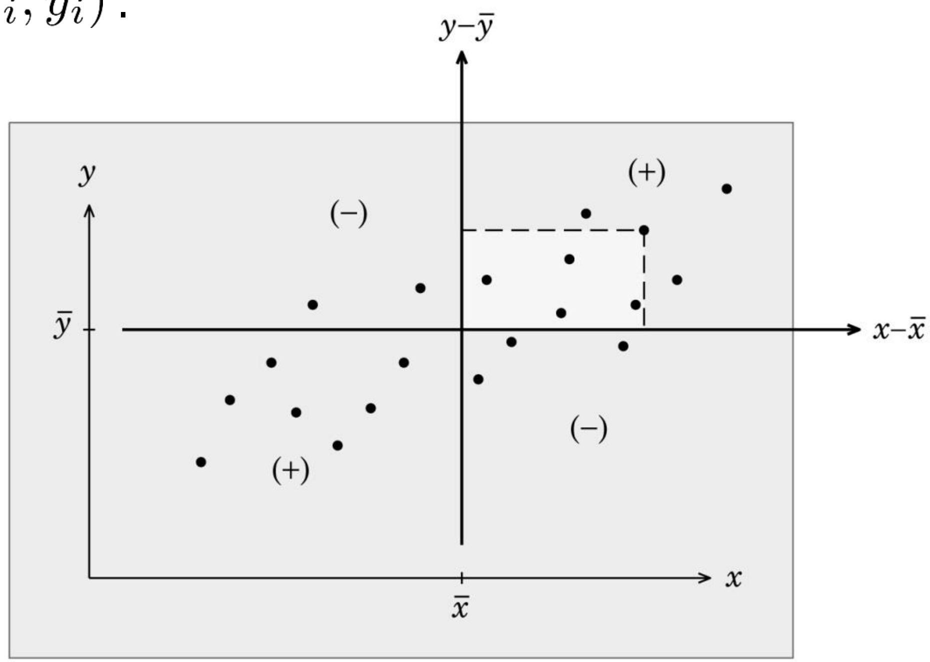


BILD 3.2 Illustration der Kovarianz

Bildquelle: [7,p.93 Bild 3.2]

Empirische Korrelationskoeffizient

(Empirischer) Korrelationskoeffizient

Sei $(x, y) = ((x_1, y_1), \ldots, (x_n, y_n))$ eine Stichprobe vom Umfang n mit $n \geq 2$, bei der für jedes Stichprobenelement zwei Merkmale x_i und y_i erfasst werden, dann ist der **empirische Korrelationskoeffizient** $\mathrm{r}_{x,y}$ der Stichprobe (x, y) definiert durch

$$\mathrm{r}_{x,y} := \frac{\mathrm{cov}(x,y)}{\sqrt{s_x^2 \cdot s_y^2}},$$

wobei s_x^2 und s_y^2 die empirischen Varianzen der Einzelmerkmale $x = (x_1, \ldots, x_n)$ und $y = (y_1, \ldots, y_n)$ sind.

Anmerkung:

Der empirische Korrelationskoeffizient ist ein normiertes Maß für den linearen Zusammenhang:

Es gilt $-1 \leq r_{x,y} \leq +1$ und $r_{x,y} = r_{y,x}$.

Interpretation des empirischen Korrelationskoeffizienten (1)

<u>Qualitative Interpretation</u>:

- Ein positiver Wert $r_{x,y} >> 0$ des empirischen Korrelationskoeffizienten beschreibt eine gemeinsame lineareTendenz der Merkmalsausprägungen (x_i, y_i).
- Ein negativer Wert $r_{x,y} << 0$ des empirischen Korrelationskoeffizienten beschreibt eine entgegengesetzte lineare Tendenz der Merkmalsausprägungen (x_i, y_i).
- Ein Wert $r_{x,y} \approx 0$ des empirischen Korrelationskoeffizienten nahe bei Null lässt keine Tendenz von gemeinsamen Merkmalsausprägungen (x_i, y_i) erschließen.

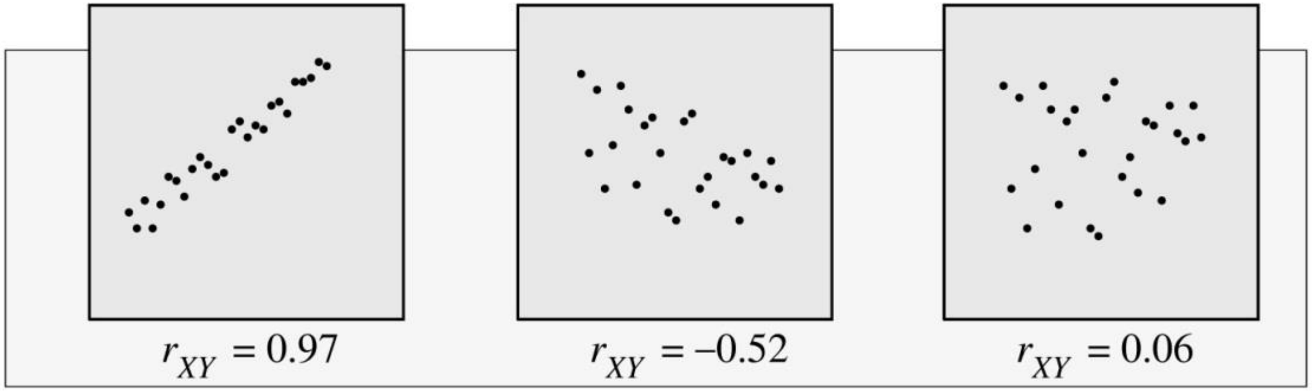


BILD 3.3 Punktewolken und Korrelationskoeffizienten

Bildquelle: [7,p.95 Bild 3.3]

Interpretation des empirischen Korrelationskoeffizienten (2)

Anmerkung:

Je ausgeprägter die gemeinsame bzw. entgegengesetzte lineare Tendenz der Merkmalsausprägungen (x_i, y_i) , um so dichter liegt der Wert des empirischen Korrelationskoeffizienten $r_{x,y}$ bei +1 bzw. -1.

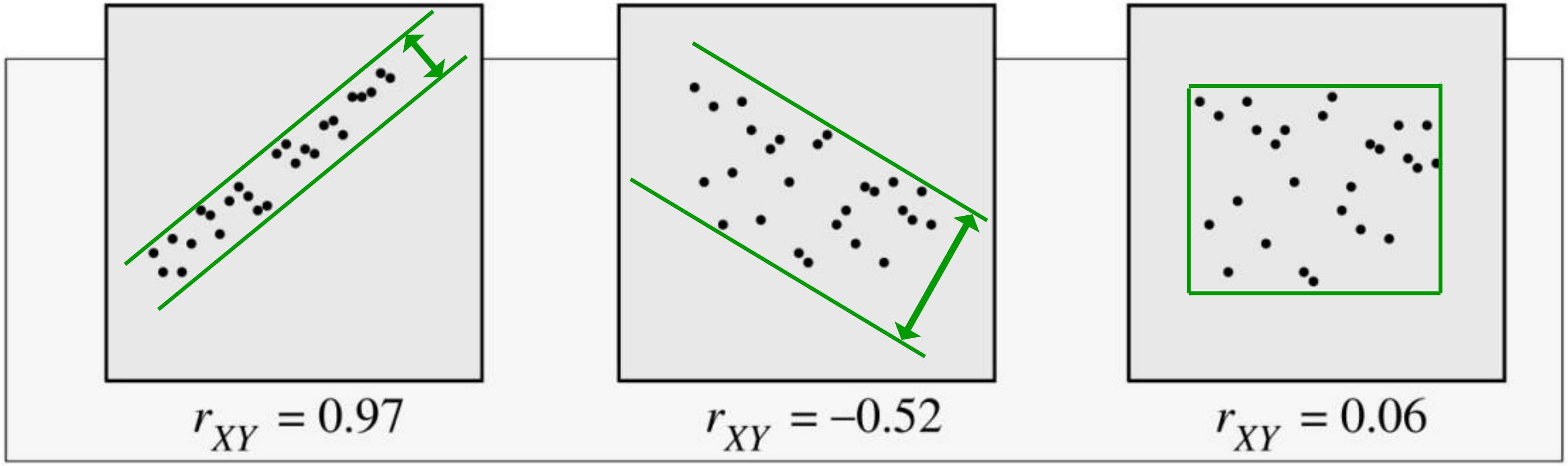
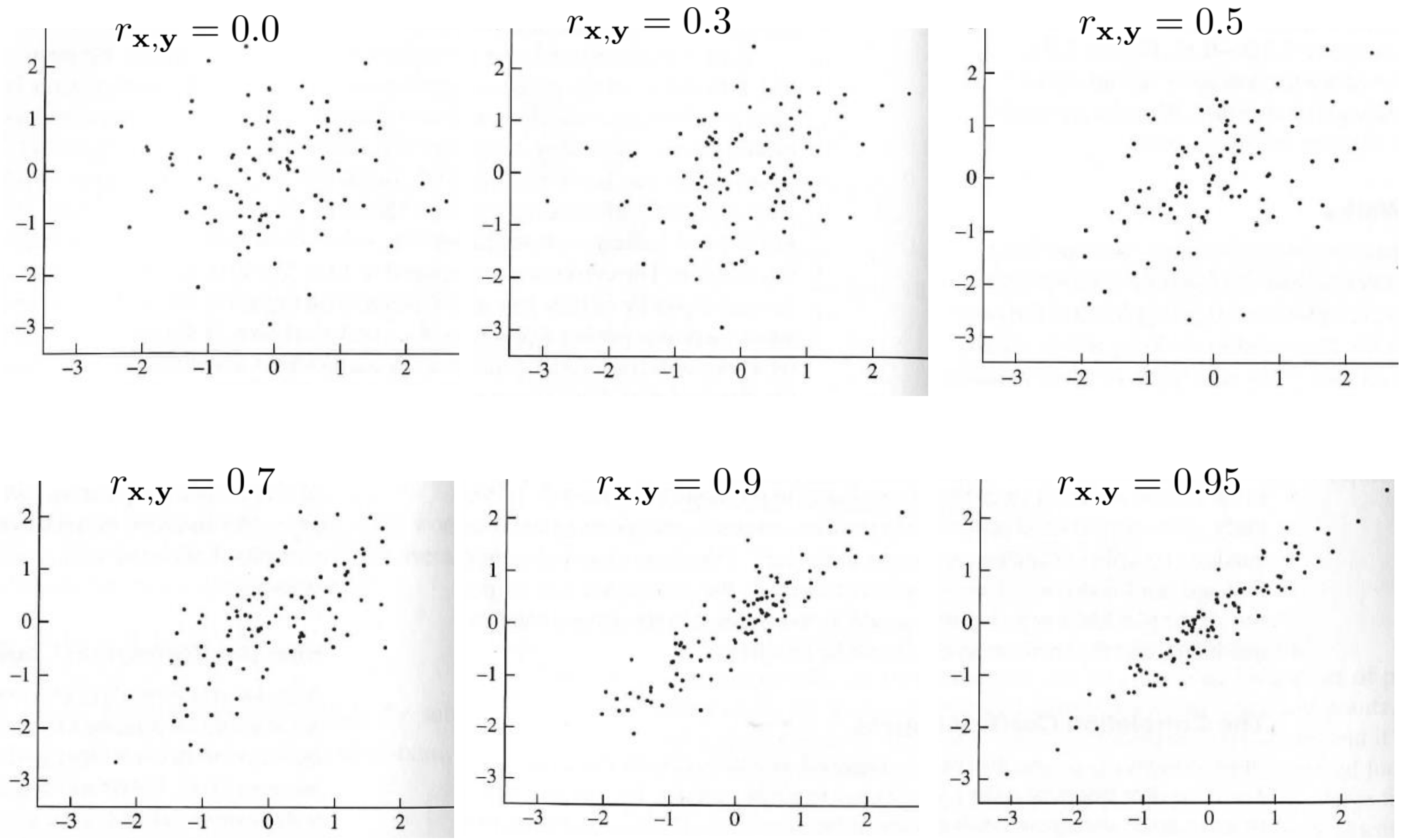


BILD 3.3 Punktewolken und Korrelationskoeffizienten

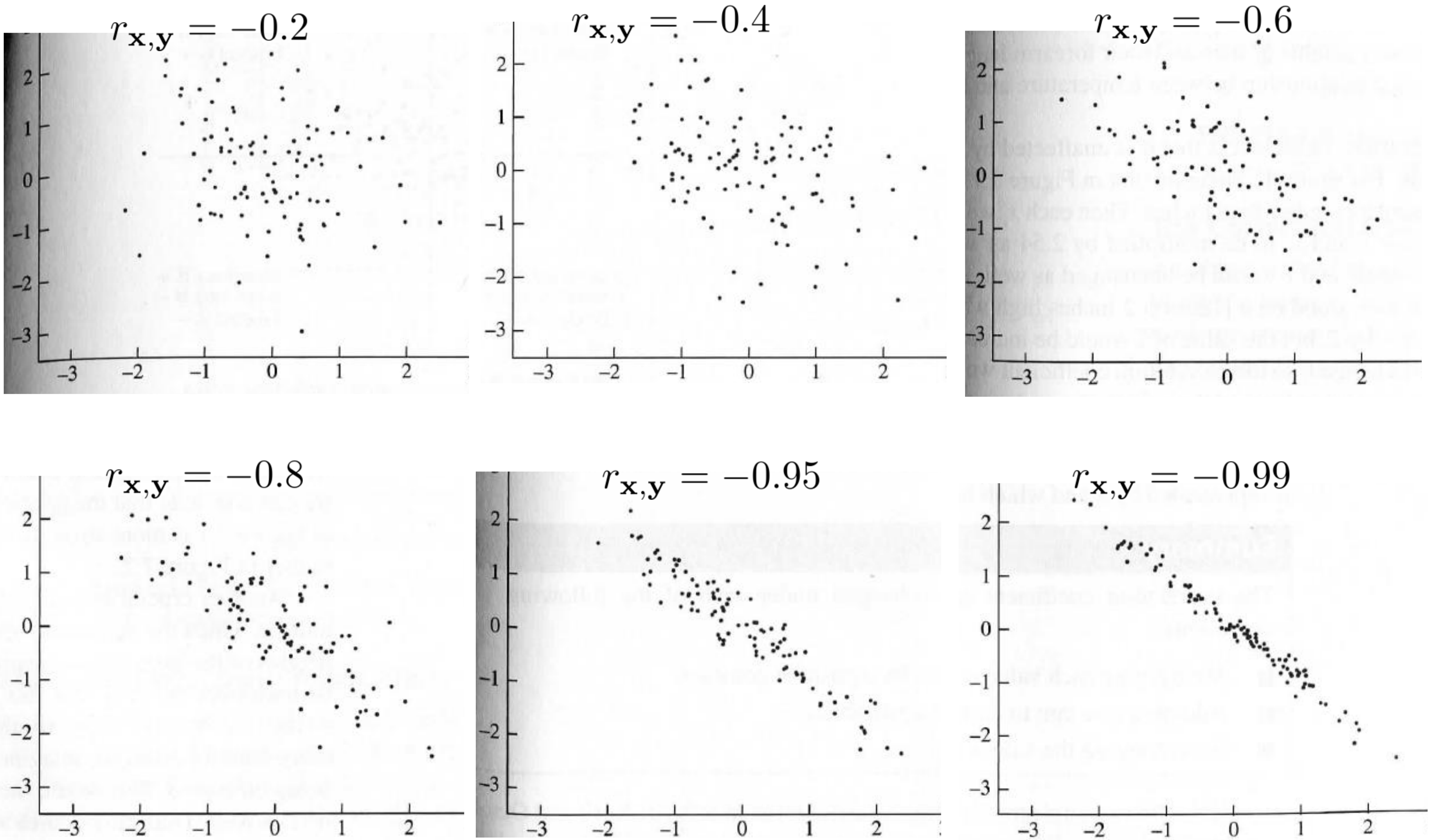
Bildquelle: [7,p.95 Bild 3.3, modifiziert]

Beispiele für den emprischen Korrelationskoeffizienten (1)



Source: [15,p.518 Figure 7.3, modified]

Beispiele für den emprischen Korrelationskoeffizienten (2)

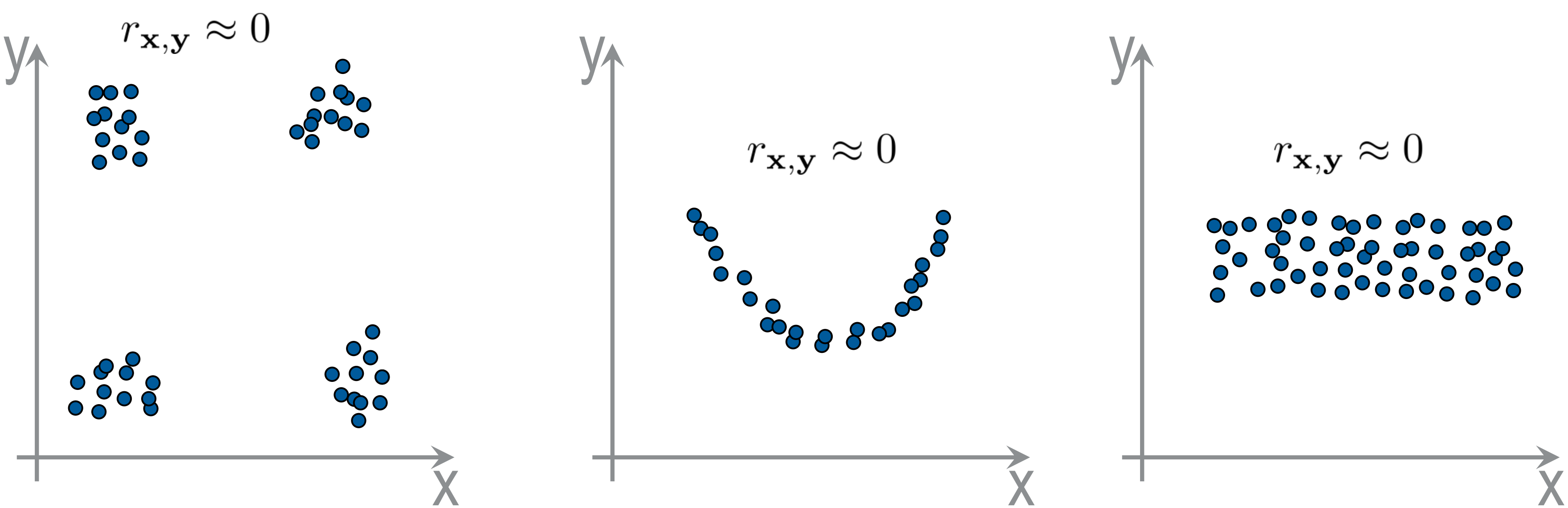


Source: [15,p.519 Figure 7.4, modified]

Versagen des empirischen Korrelationskoeffizienten

<u>Bemerkung</u>: ([2])

Der empirische Korrelationskoeffizient (und analog die empirische Kovarianz) misst nur die **linearen Anteile** der Abhängigkeit und ignoriert andere Arten von Abhängigkeiten. Beispielsweise gilt in den folgenden Streudiagrammen mit offentsichtlichen Abhängigkeiten überall $r_{\mathbf{x},\mathbf{y}} \approx 0$:



Inhaltsverzeichnis VE 04

VE 04: Korrelation und Regressionsanalyse

Zielsetzung der Regressionsanalyse

In der Regressionsanalyse wird ein funktionaler Zusammenhang behandelt. Ziele sind dabei gem. [8] u.a.:

- Erkennen einer Ursache-Wirkung-Beziehung
- Schätzen des Parameters einer bekannten funktionalen Beziehung
- Empirische Repräsentation großer Datenmengen (deskriptiv!)
- Interpolation fehlender bzw. Prognose zukünftiger Werte.

Regressionsarten

Einen ersten Eindruck bzgl. der Art des Regressesionszusammenhangs kann man sich anhand der grafischen Darstellung der Ausprägungskombinationen der Werte $((x_1, y_1), \ldots, (x_n, y_n))$ einer Stichprobe verschaffen.

Die Ausprägungen $(x_1, \ldots, x_n)$ des Merkmals X können dabei vorgegeben worden sein und damit die Auswahl bedingt haben oder alternativ miterhoben worden sein.

Beispiele:

- lineare Regression,
- logarithmisch lineare und halblogarithmische Regression,
- quadratische Regression und
- mehrfache (auch: multiple) Regression.

Grundproblemstellung der Regression

Es werden n Paare (x_i, y_i) von Messungen durchgeführt und als Punkte in ein Koordinatensystem eingetragen.

Beispiel: $x_i =$ Temperatur eines Stahlstabs, $y_i =$ Länge des Stahlstabs.

Fragestellung:
Wie kann man diesen n Punkten eine möglichst einfache Kurve möglichst gut anpassen, d.h. die Parameter passend wählen?

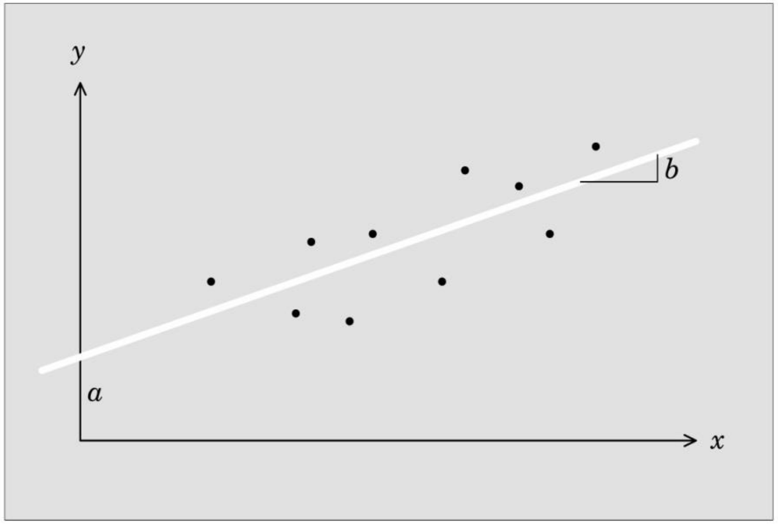


BILD 4.1 Punktewolke und Gerade im Streudiagramm

Bildquelle: [7,p.107 Bild 4.1]

Regression von Y auf X

Definition:

Kann der Zusammenhang zweier Merkmale X und Y mit den Ausprägungen x und y durch die funktionale Beziehung $y = f(x)$ beschrieben werden, so spricht man von einer Regression von Y auf X.

Bezeichnungen und Namensvielfalt in der Regression:

Man bezeichnet X als **Regressor** (auch: unabhängige Variable, exogene Variable, erklärende Variable, Einflussfaktor) und Y als **Regressand** (auch: endogene Variable, Zielvariable, erklärte Variable).

Lineare Regression

Bei der **linearen Regression**, wird der wesentliche Regressionszusammenhang $y = f(x)$ durch eine lineare Gleichung beschrieben: $y = f(x) = a + b \cdot x$.

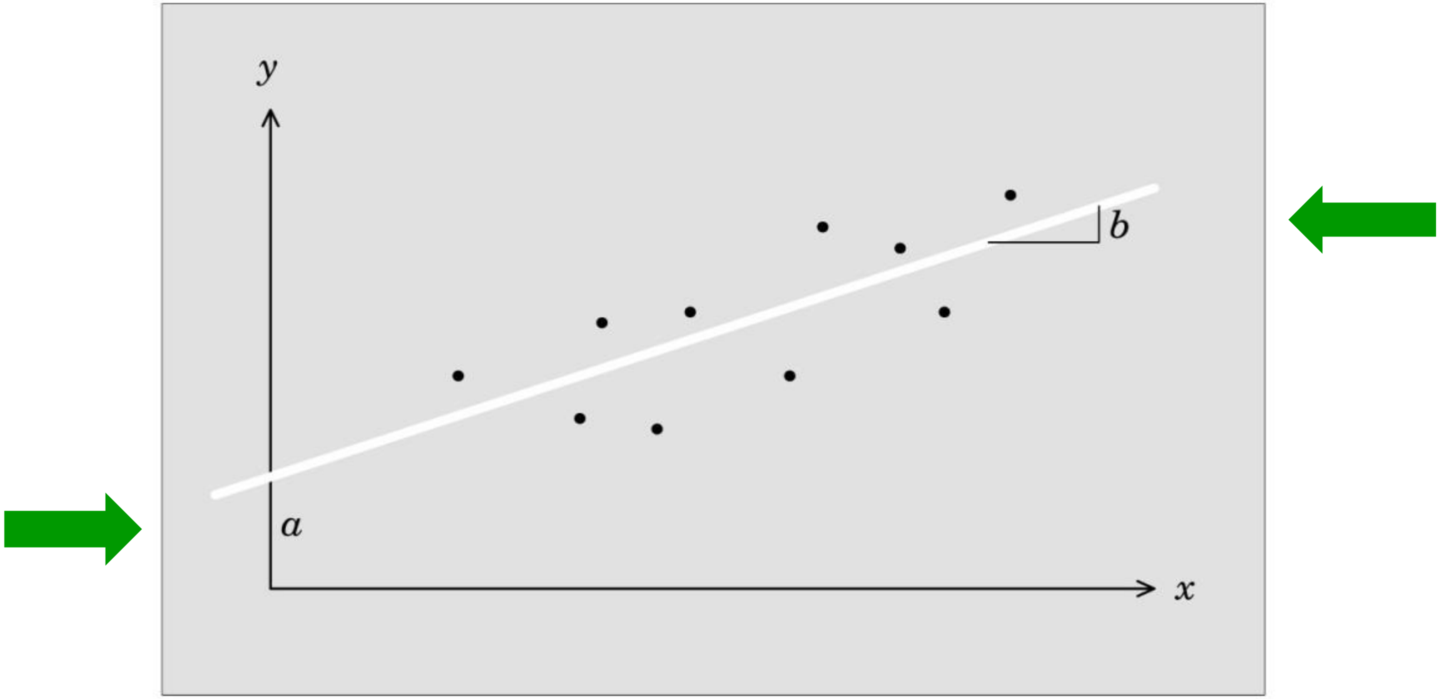


BILD 4.1 Punktewolke und Gerade im Streudiagramm

Bildquelle: [7,p.107 Bild 4.1]

Grundmodell der linearen Regressionsanalyse

Vermutet man einen (zumindest näherungsweise) linearen Zusammenhang zwischen den Merkmalen X und Y, so kann dieser Zusammenhang mittels des folgenden linearen Modells näher spezifiziert und untersucht werden:

$$y = a + b \cdot x + e \qquad \text{„}\textit{Regressionsgerade}\text{“}$$

mit den Bezeichnungen:

- x bzw. y Ausprägungen der Merkmale X bzw. Y,
- $a, b \in \mathbb{R}$ (wahre Werte der) Parameter der linearen Beziehung und
- e Ausprägung eines zufälligen Fehlers E (Störvariable, latente Variable).

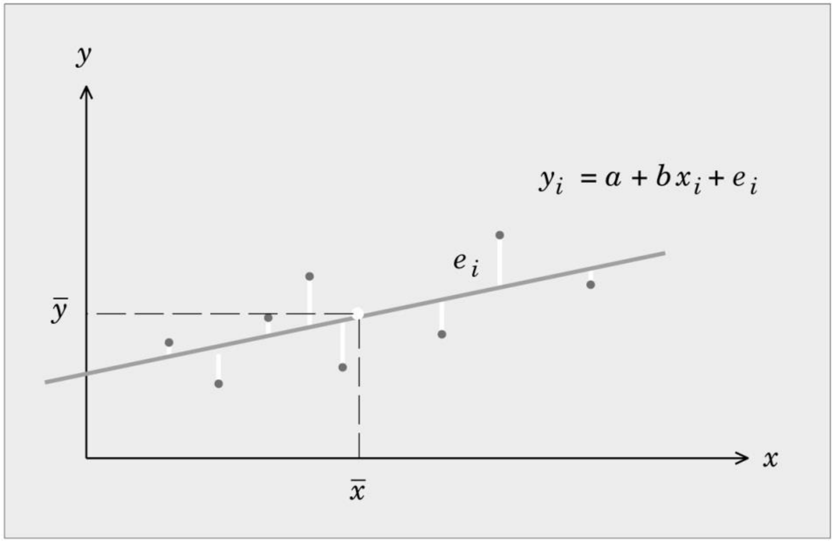


BILD 4.5 Regression von Y auf X

Bildquelle: [7,p.119 Bild 4.5]

Interpretation des Grundmodell der linearen Regressionsanalyse

Interpretation:

Die latente Variable e_i bündelt in der Modellvorstellung weitere, in der exakte Spezifikation fehlende, exogene Variablen, Messfehler der Y-Ausprägungen und andere unvorhersagbare Zufälligkeiten.

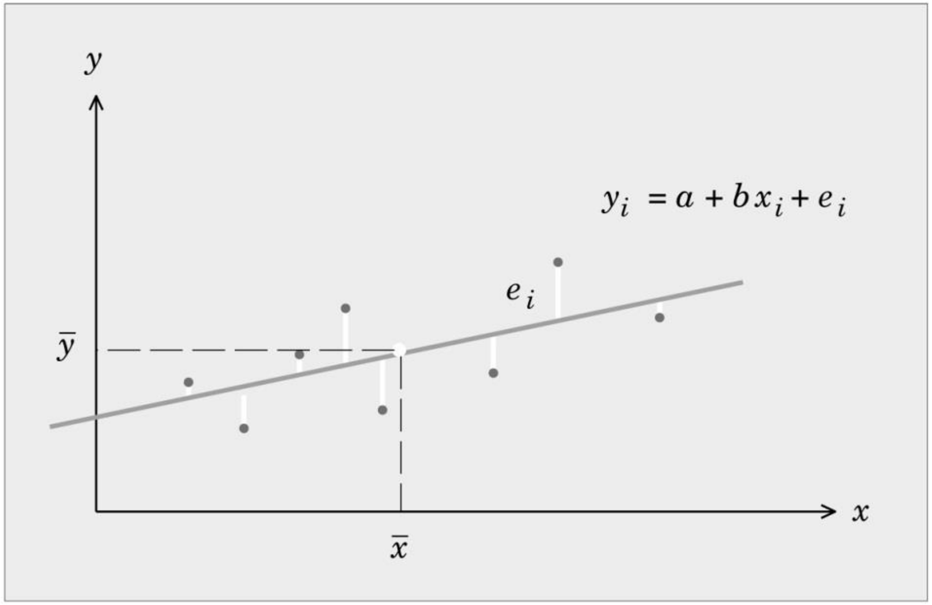


BILD 4.5 Regression von Y auf X

Bildquelle: [7,p.119 Bild 4.5]

Methode der kleinsten Quadrate zur Regressionsgeradenberechnung

Für die Ausprägungen einer Stichprobe $((x_1, y_1), \ldots, (x_n, y_n))$ der kombinierten Merkmale (X, Y) können die Parameter der Regressionsgerade mit Hilfe der Methode der kleinsten Quadrate geschätzt werden :

Vorgehen:

Die Parameter a, b werden so gewählt, dass sie die Summe der Quadrate der Abweichungen $e_i = y_i - a - b \cdot x_i$ für $i = 1, \ldots, n$ minimieren:

$$S^2(a,b) = \sum_{i=1}^{n} \left(y_i - a - b \cdot x_i\right)^2 \longrightarrow min!$$

Ergebnis:

Es ergeben sich folgende Schätzwerte:

$$b = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \quad \text{und} \quad a = \overline{y} - b\overline{x}\,.$$

„Methode der kleinsten Quadrate“
bzw.
„Kleinste-Quadrate-Schätzer
des linearen Modells“

Beispiel: Berechnung der Regressiongerade

Bestimmen Sie die Regressionsgerade für die Messwertpaare (x_i, y_i) mit

(2; 1630), (2.5; 1644), (3; 1661), (4; 1681), (5; 1710), (6; 1737), (6.5; 1738) und (8; 1786).

Bestimmtheitsmaß für die lineare Regression

Bestimmtheitsmaß

Als Maß für die Güte der Anpassung, die eine Regression für eine Stichprobe $((x_1, y_1), \ldots, (x_n, y_n))$ erzielt, wird das **Bestimmtheitsmaß der linearen Regression** definiert:

$$B_{X,Y} = \frac{\sum_{i=1}^{n} (a + b \cdot x_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \quad .$$

<u>**Eigenschaften**</u>**:**

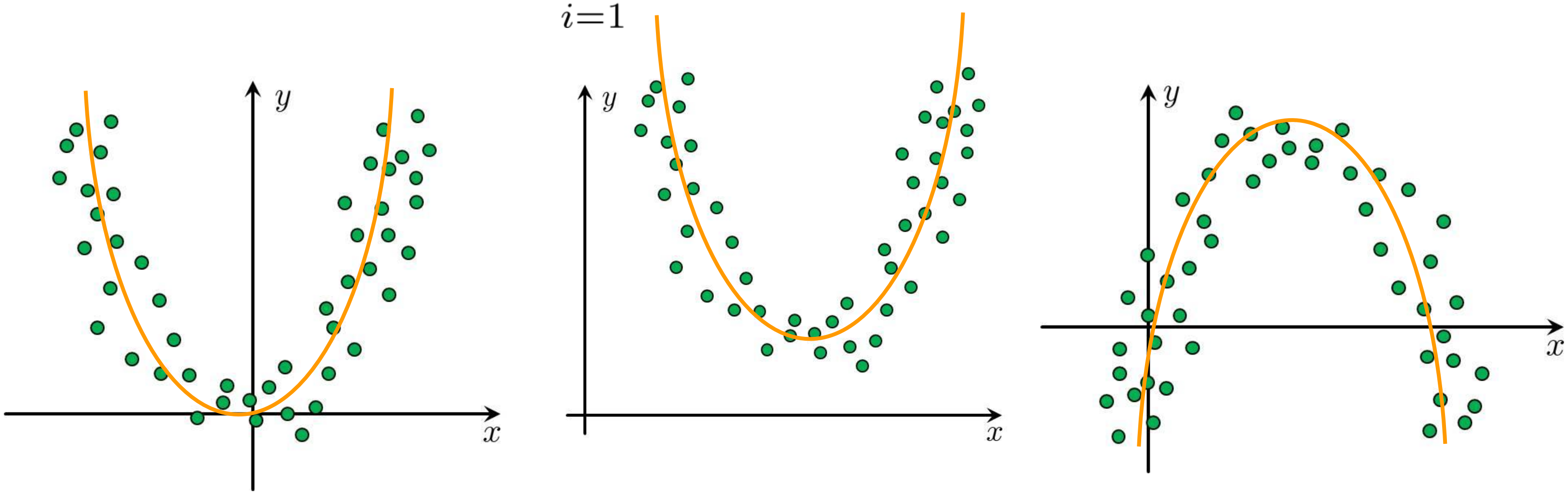
- Es gilt stets $0 \leq B_{X,Y} \leq 1$.
- Gilt $B_{X,Y} = 1$, so wird (für diese Stichprobe) der Zusammenhang zwischen den Merkmalen vollständig erklärt.

Ausblick: Quadratische Regression

Bei der **quadratischen Regression**, wird der wesentliche Regressionszusammenhang durch eine quadratische Gleichung beschrieben: $y = f(x) = a \cdot (x-b)^2 + c$.

Die Parameter $a, b, c \in \mathbb{R}$ werden wiederum mit der *Methode der kleinsten Quadrate* bestimmt:

$$S^2(a,b,c) = \sum_{i=1}^{n} \left(y_i - a \cdot (x_i - b)^2 - c\right)^2 \longrightarrow min!$$



Bildquelle: eigene Darstellungen

Ausblick: Mehrfache Regression

Bei der **mehrfachen Regression** hängt die Zielvariable Y von mehreren erklärenden Variablen $X_1, \ldots, X_M$ ab. Im Fall der 2-dimensionalen linearen Regression gilt beispielsweise der Zusammenhang $y = f(x_1, x_2) = a + b_1 \cdot x_1 + b_2 \cdot x_2$.

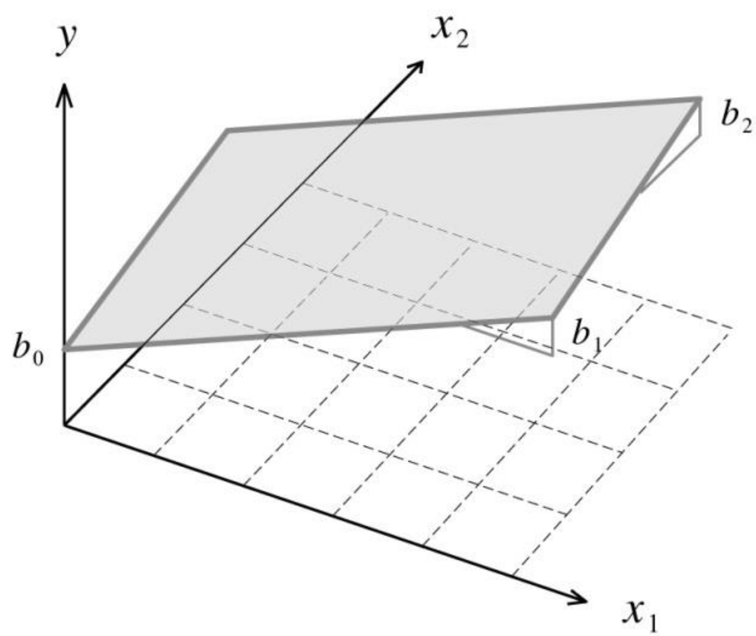


BILD 4.9 Regressionsebene

Bildquelle: [7,p.124 Bild 4.9]