Lineare Regression

Theorie & Aufgaben

1 Beispiel

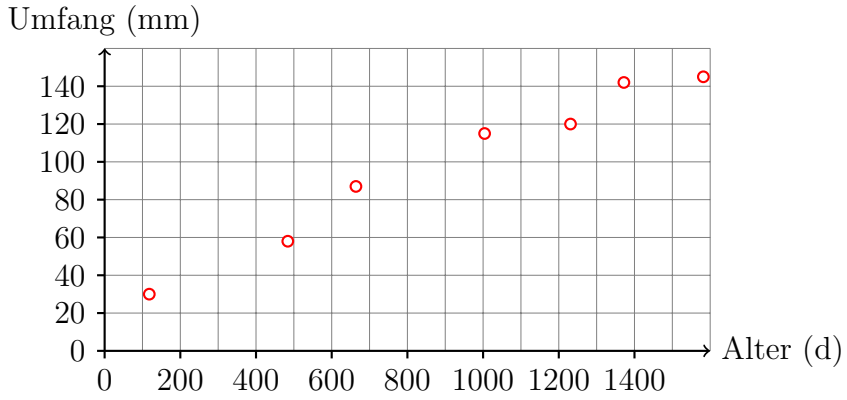
In einer Orangenplantage wurden Alter (Anzahl Tage seit dem 31.12.1968) und Umfang (in mm auf Brusthöhe) gemessen.

Alter (d)	Umfang (mm)
118	30
484	58
664	87
1004	115
1231	120
1372	142
1582	145

Beachte, dass Alter und Umfang jeweils *paarweise* auftreten.

2 Streudiagramm

Stelle die Daten als *Streudiagramm* im vorbereiteten Koordinatensystem dar.



Aufgrund der Grafik vermuten wir einen linearen Zusammenhang zwischen Alter und Umfang. Diese Beziehung können wir durch eine Gleichung der Form

$$g\colon y = a \cdot x + b$$

beschreiben, wobei a die Steigung und b der y-Achsenabschnitt der zur Gleichung gehörenden *Ausgleichsgeraden* sind.

3 Die Bestimmung der Ausgleichsgeraden

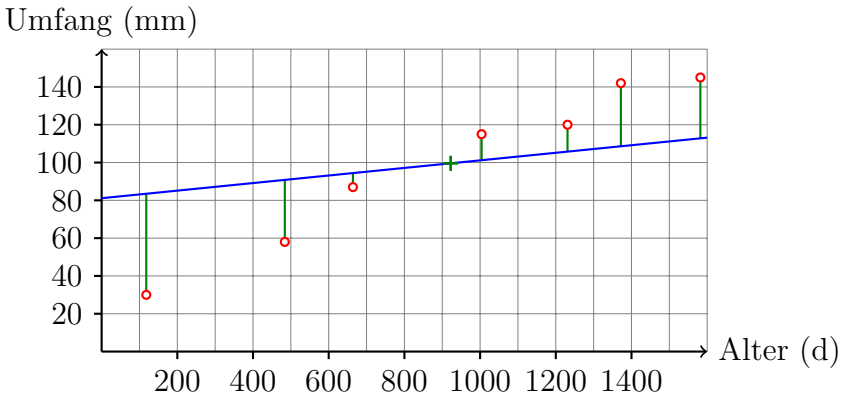
3.1 Erstes Kriterium

Die Abweichungen nach oben und nach unten sollen sich gegenseitig aufheben:

$$\sum_{i=1}^{n} \left[y_i - \underbrace{(a \cdot x_i + b)}_{g(x_i)} \right] = 0$$

Man kann zeigen, dass dies gleichbedeutend damit ist, dass die Ausgleichsgerade durch den *Datenschwerpunkt* $S(\overline{x}, \overline{y})$ geht.

Etwa so:



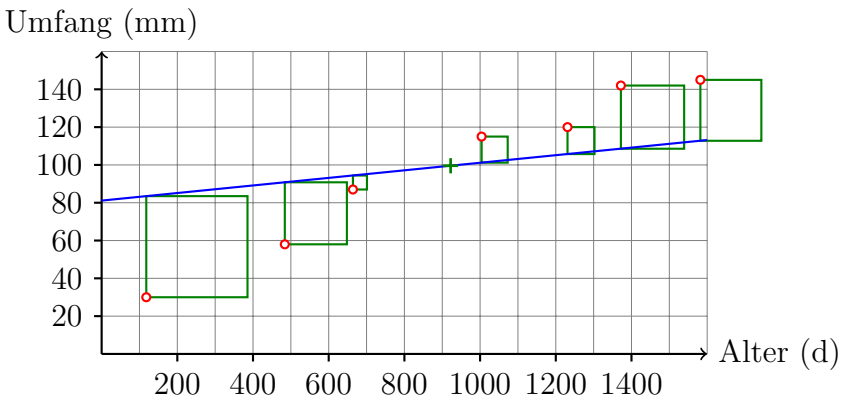
3.2 Zweites Kriterium

Das zweite Kriterium liegt nicht auf der Hand aber es ist vernünftig. Es verlangt, dass für die gesuchte Gerade die Summe der Abstandsquadrate möglichst klein werden soll. Das bedeutet: Wähle a und b so, dass die Summe

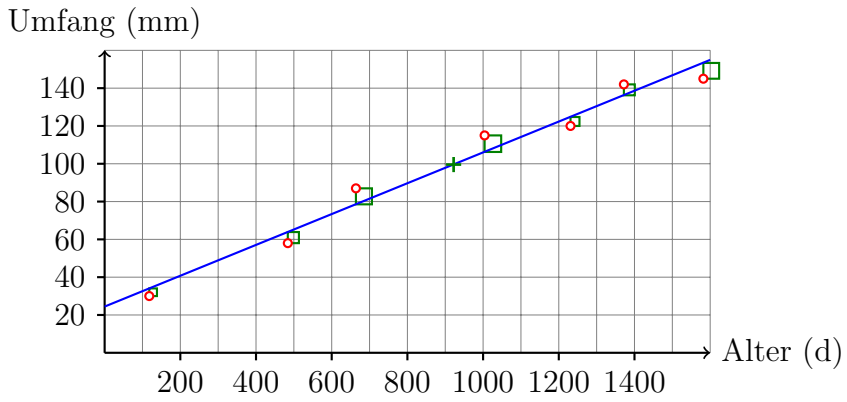
$$\sum_{i=1}^{n} \left[y_i - \underbrace{(ax_i + b)}_{g(x_i)} \right]^2$$

minimal wird.

Schlechte Anpassung



Optimale Anpassung



Mit dem Wissen aus aus der 5. Klasse kann man zeigen, dass folgendes a die gewünschte Eigenschaft hat:

$$a = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2}$$

s_{xy}: empirische Kovarianz

s_x^2: empirische Varianz (sozusagen s_{xx})

$S(\overline{x}, \overline{y})$ muss auf der Ausgleichsgeraden liegen:

$\overline{y} = a \cdot \overline{x} + b \quad \Rightarrow \quad \boxed{b = \overline{y} - a \cdot \overline{x}}$

3.3 Zahlenbeispiel

$\overline{x} = 922$, $\overline{y} = 100$

x	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	y	$y_i - \overline{y}$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
118	−804	646416	30	−70	4900	56280
484	−438	191844	58	−42	1764	18396
664	−258	66564	87	−13	169	3354
1004	82	6724	115	15	225	1230
1231	309	95481	120	20	400	6180
1372	450	202500	142	42	1764	18900
1582	660	435600	145	45	2025	29700
922	0	1645129	100	0	11247	134040

$$a = \frac{s_{xy}}{s_{xx}} = \frac{134\,140}{1\,645\,129} = 0.0815$$

$b = \overline{y} - a \cdot \overline{x} = 100 - 0.0815 \cdot 922 = 24.8$

Ausgleichsgerade: $y = 0.0815 \cdot x + 24.8$

4 Der Korrelationskoeffizient

Wenn man die empirische Kovarianz durch die empirischen Standardabweichungen der x- und y-Werte dividiert, so erhält man den *Korrelationskoeffizienten*

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

Man kann beweisen, dass r_{xy} die Ungleichungen

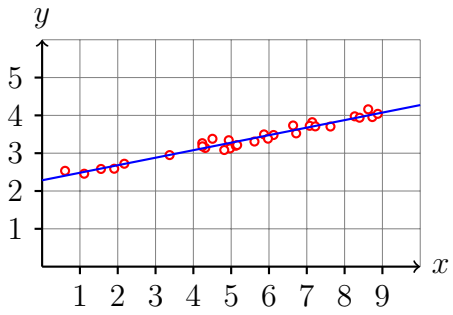
$$-1 \leq r_{xy} \leq 1$$

erfüllt.

Der Korrelationskoeffizient ist ein Mass für die Güte des linearen Zusammenhangs.

- Je näher r_{xy} bei $+1$ liegt, desto besser ist der lineare Zusammenhang (Korrelation). $r_{xy} = 1$ bedeutet, dass alle (x_i, y_i) auf der (steigenden) Regressionsgeraden liegen.
- Je näher r_{xy} bei 0 liegt, desto schlechter ist der lineare Zusammenhang.
- Je näher r_{xy} bei -1 liegt, desto besser ist der lineare Zusammenhang. $r_{xy} = -1$ bedeutet, dass alle (x_i, y_i) auf der (fallenden) Regressionsgeraden liegen.

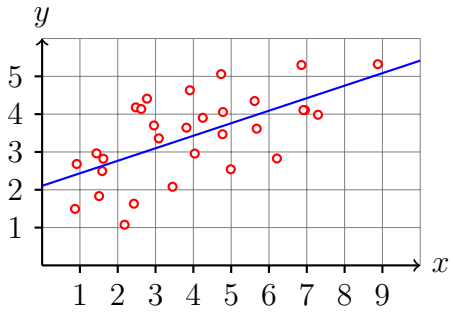
Beispiel 1



$y = 0.2x + 2.28$

$r_{xy} = 0.98$

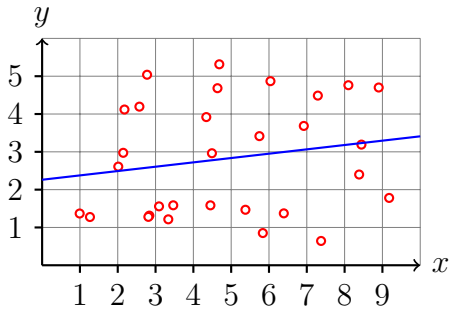
Beispiel 2



$y = 0.33x + 2.1$

$r_{xy} = 0.63$

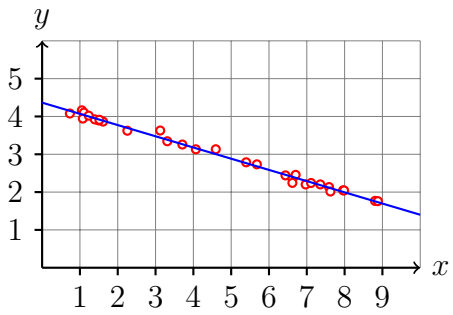
Beispiel 3



$y = 0.11x + 2.26$

$r_{xy} = 0.19$

Beispiel 4



$y = -0.29x + 4.37$

$r_{xy} = -0.99$

5 Korrelation und Kausalität

Im Buch von Bortz und Schuster findet man dazu (S. 159):

„Hat man zwischen zwei Variablen x und y eine Korrelation gefunden, kann diese im kausalen Sinne folgendermassen interpretiert werden:

- x beeinflusst y kausal,
- y beeinflusst x kausal,
- x und y werden von einer dritten oder weiteren Variablen kausal beeinflusst,
- x und y beeinflussen sich wechselseitig kausal.

Der Korrelationskoeffizient liefert keine Informationen darüber, welche dieser Interpretationen richtig ist. (. . .)

Merke: Korrelationen dürfen ohne zusätzliche Informationen nicht kausal interpretiert werden.“

6 Übungen

Aufgabe 1

Gegeben ist folgende gepaarte Stichprobe:

x	3	9	18
y	12	5	4

- Bestimme die Gleichung der Regressionsgeraden.
- Berechne den Korrelationskoeffizienten.
- Skizziere das Streudiagramm und die Ausgleichsgerade.

Aufgabe 2

Gegeben ist folgende gepaarte Stichprobe:

x	9	13	1	5
y	14	15	4	7

- Bestimme die Gleichung der Regressionsgeraden.
- Berechne den Korrelationskoeffizienten.
- Skizziere das Streudiagramm und die Ausgleichsgerade.

Aufgabe 3

Gegeben ist folgende gepaarte Stichprobe:

x	16	13	9	10
y	9	12	14	13

- Bestimme die Gleichung der Regressionsgeraden.
- Berechne den Korrelationskoeffizienten.
- Skizziere das Streudiagramm und die Ausgleichsgerade.

Aufgabe 4

Gegeben ist folgende gepaarte Stichprobe:

x	7	3	6	4
y	17	13	15	11

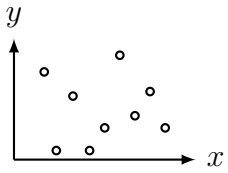
- Bestimme die Gleichung der Regressionsgeraden.
- Berechne den Korrelationskoeffizienten.
- Skizziere das Streudiagramm und die Ausgleichsgerade.

Aufgabe 5

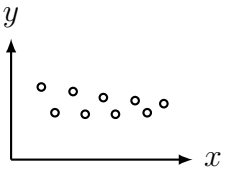
Für vier Messreihen wurden die empirischen Korrelationen ermittelt. Ordne diese Werte den folgenden Grafiken zu.

$r_{xy} = 0.946$ $r_{xy} = -0.996$ $r_{xy} = 0.021$ $r_{xy} = -0.344$

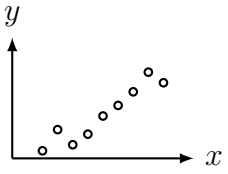
(a)



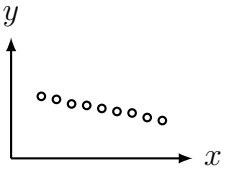
(b)



(c)



(d)



Aufgabe 6

Rex Boggs, Glenmore State High School, Rockhampton, Queensland, Australia hat untersucht, wie sich das Gewicht eines Stücks Seife im Laufe der Zeit verändert.

Date	Day	Weight	Date	Day	Weight
30.8.1999	0	124	10.9.1999	11	58
31.8.1999	1	121	11.9.1999	12	50
3.9.1999	4	103	16.9.1999	17	27
4.9.1999	5	96	18.9.1999	19	16
5.9.1999	6	90	19.9.1999	20	12
6.9.1999	7	84	20.9.1999	21	8
7.9.1999	8	78	21.9.1999	22	6
8.9.1999	9	71			

Am 22.9.1999 zerbrach das Seifenstück in zwei Teile und eines davon verschwand im Abfluss.

Erstelle ein lineares Modell für die Abhängigkeit des Seifengewichts von ihrer Lebensdauer und berechne das Bestimmheitsmass.

7 Lösungen

Aufgabe 1

Regressionsgerade: $y = -\frac{1}{2} \cdot x + 12$; Korrelationskoeffzient: $r_{xy} = -0.865$

Aufgabe 2

Regressionsgerade: $y = 1 \cdot x + 3$; Korrelationskoeffzient: $r_{xy} = 0.964$

Aufgabe 3

Regressionsgerade: $y = -\frac{2}{3} \cdot x + 20$; Korrelationskoeffzient: $r_{xy} = -0.975$

Aufgabe 4

Regressionsgerade: $y = \frac{6}{5} \cdot x + 8$; Korrelationskoeffzient: $r_{xy} = 0.849$

Aufgabe 5

(a) $r_{xy} = 0.021$ (b) $r_{xy} = -0.344$ (c) $r_{xy} = 0.946$ (d) $r_{xy} = -0.996$

Aufgabe 6

$\hat{y} = 123.1 - 5.575 \cdot t$; $r = 0.9953$