Regression

Leonardo da Vinci (1452-1519) formulierte folgende Zusammenhänge:

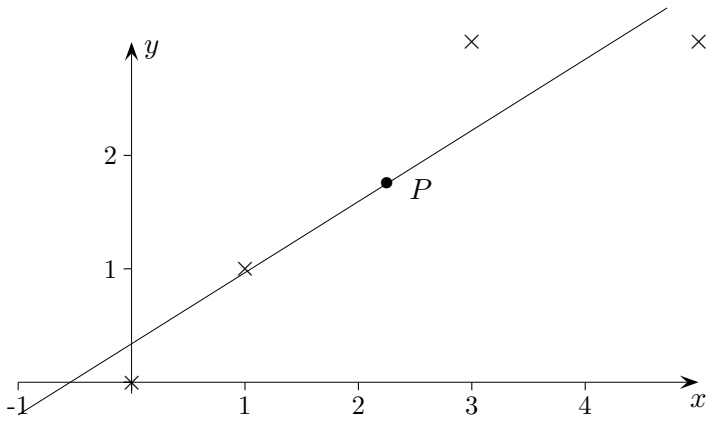
1. Die Körpergröße ist gleich der Spannweite der Arme.
2. Die Höhe einer knienden Person ist $\frac{3}{4}$ der Körpergröße.
3. Die Handlänge ist $\frac{1}{9}$ der Körpergröße.

Dies kann durch Messungen überprüft werden.
Allgemein werden wir Datenpaare $(x_i \mid y_i)$ auf eine möglicherweise vorliegende lineare Beziehung hin untersuchen. In der Regressionsanalyse ist die bestmögliche approximierende Gerade gesucht.

x	x_1	x_2	x_3	$\dots$	x_n
y	y_1	y_2	y_3	$\dots$	y_n

Wir gehen zunächst von folgenden einfachen Daten aus:

x	y
0	0
1	1
3	3
5	3



Es erscheint plausibel, dass der Schwerpunkt $P(\overline{x} \mid \overline{y})$ auf der Ausgleichsgeraden $y = mx + b$ liegt,
$\overline{x} = \frac{1}{n}\sum x_i$, $\overline{y} = \frac{1}{n}\sum y_i$. Dies soll an dieser Stelle nicht bewiesen werden (siehe Verschiedenes, Mittelstufe).

Mit $\overline{y} = m\overline{x} + b$ erhalten wir $b = \overline{y} - m\overline{x}$. Dies ergibt den Ansatz $y = mx + \overline{y} - m\overline{x}$.
Um die Steigung m der optimalen Ausgleichsgeraden zu ermitteln, muss präzisiert werden, was unter optimal zu verstehen ist. Wir erinnern uns an die Definition der Standardabweichung. Optimal wäre also z. B., wenn die Summe Q der Abweichungsquadrate minimal wäre. Mit Quadraten lässt sich leichter rechnen, als mit Wurzeln oder Beträgen.

$Q = (mx_1 + \overline{y} - m\overline{x} - y_1)^2 + (mx_2 + \overline{y} - m\overline{x} - y_2)^2 + \ldots + (mx_n + \overline{y} - m\overline{x} - y_n)^2$ oder übersichtlicher

$Q = \big(m(x_1 - \overline{x}) - (y_1 - \overline{y})\big)^2 + \big(m(x_2 - \overline{x}) - (y_2 - \overline{y})\big)^2 + \ldots + \big(m(x_n - \overline{x}) - (y_n - \overline{y})\big)^2$

Bestimme für unser Beispiel $m_{\min}$.

Regression

Etwas Übersicht ist - wie meistens im Leben - hilfreich:

$\overline{x} = 2{,}25$
$\overline{y} = 1{,}75$

x	y	$x-\overline{x}$	$y-\overline{y}$
0	0	−2,25	−1,75
1	1	−1,25	−0,75
3	3	0,75	1,25
5	3	2,75	1,25

$Q(m) = (-2{,}25m + 1{,}75)^2 + (-1{,}25m + 0{,}75)^2 + (0{,}75m - 1{,}25)^2 + (2{,}75m - 1{,}25)^2$

Für die Berechnung des Minimums $m_{\min} = 0{,}627$ gibt es mehrere Möglichkeiten (GTR, Scheitel einer Parabel, Differenzialrechnung). Wir erhalten $y = 0{,}627x + 0{,}339$.

Eine allgemeine Rechnung führt zu

$$m = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} \qquad \left(= \frac{\frac{1}{n}\sum (x_i - \overline{x})(y_i - \overline{y})}{\sigma_x^2} = \frac{Cov_{xy}}{\sigma_x^2} \right)$$

Die Gleichung der Regressionsgeraden lautet: $y = m(x - \overline{x}) + \overline{y}$

Die Kovarianz $Cov_{xy} = \frac{1}{n}\sum (x_i - \overline{x})(y_i - \overline{y})$
ist umso größer, je mehr positive Summanden vorhanden sind.
Ein Summand ist positiv, falls beide Werte x_i und y_i kleiner oder beide größer als der entsprechende Mittelwert sind. Es liegt dann eine steigende Tendenz vor.
In welchem Fall ist diese fallend?

GTR
Mit STAT | EDIT x-Werte in L1 und y-Werte in L2 eingeben.
Das Streudiagramm wird mit STAT PLOT gezeichnet.
STAT | CALC 4: LinReg(ax+b) aufrufen, vorher DiagnosticOn.
Mit LinReg(ax+b) Y1 (z. B.) wird das Ergebnis in Y1 gespeichert.
Falls die Werte nicht in L1 und L2 stehen, sind die Listen anzugeben, z. B. LinReg(ax+b) L2, L3, Y1
Nützlich ist auch STAT | CALC 2: 2−Var Stats.

Aufg.
An einer Tankstelle werden für die letzten vier Monate folgende Benzinpreise und Absatzmengen notiert:

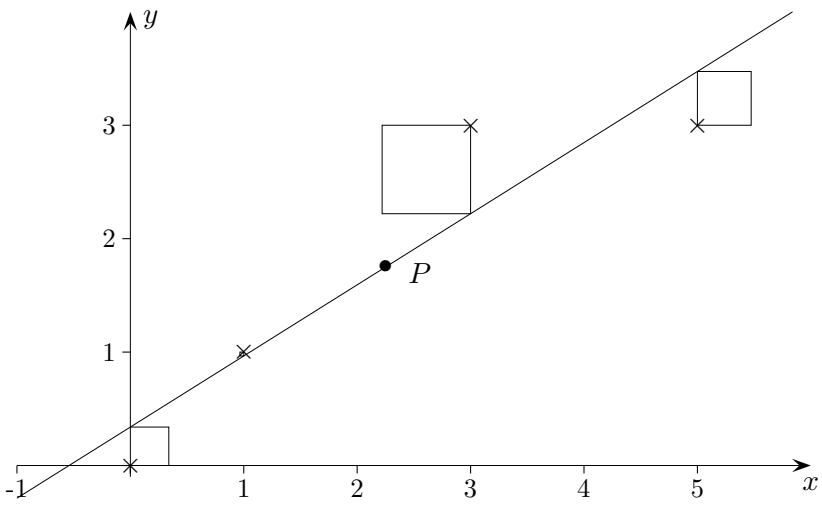
Preis (in €)	1,45	1,41	1,55	1,50
Menge (in 1000 l)	150	165	140	144

Bei welchem Preis ergäbe sich ein Absatz von 100 000 l? Lösung: 1,77 €

Regression alternativer Einstieg

Wir gehen zunächst von folgenden einfachen Daten aus:

x	y
0	0
1	1
3	3
5	3



Um von der optimalen Ausgleichsgeraden $y = mx + b$ die Steigung m und den y-Achsenabschnitt b zu ermitteln, muss präzisiert werden, was unter optimal zu verstehen ist. Die Summe der senkrechten Abstände zur Geraden zu minimieren erscheint aufgrund der Komplexität wenig aussichtsreich. Stattdessen könnten die Abstände in y-Richtung betrachtet werden. Die Beträge sind hinderlich. Wir erinnern uns an die Definition der Standardabweichung. Optimal wäre, wenn die Summe Q der Abweichungsquadrate minimal wäre. Mit Quadraten lässt sich leichter rechnen, als mit Wurzeln oder Beträgen.

$Q(m, b) = b^2 + (m + b - 1)^2 + (3m + b - 3)^2 + (5m + b - 3)^2$

Stellen wir uns vor, wir hätten schon b. Dann müsste für m gelten: $Q'(m) = 0$. Dies führt zu:
$35m + 9b - 25 = 0$
Wäre umgekehrt m schon gegeben, so müsste gelten: $Q'(b) = 0$. Dies führt zu:
$9m + 4b - 7 = 0$

Die Lösung dieses Gleichungssystems führt zur gesuchten Ausgleichsgeraden $y = 0{,}627x + 0{,}339$.

Und nun allgemein:

x	x_1	x_2	x_3	$\dots$	x_n
y	y_1	y_2	y_3	$\dots$	y_n

$Q(m, b) = \sum (mx_i + b - y_i)^2$ Die Summe erstreckt sich stets von 1 bis n.

$Q'(b) = 0$

$\Longrightarrow \quad b_{\min} = \frac{1}{n} \sum_{i=1}^{n} (y_i - mx_i)$

$\Longrightarrow \quad \overline{y} = m\overline{x} + b$ Mittelwerte $\overline{x} = \frac{1}{n} \sum x_i$, $\overline{y} = \frac{1}{n} \sum y_i$

$P(\overline{x} \mid \overline{y})$ liegt auf der Ausgleichsgeraden.

Regression alternativer Einstieg

m ist noch zu bestimmen.

Wir können in $Q(m, b)$ das b durch $\overline{y} - m\overline{x}$ ersetzen.

$$Q(m) = \sum (mx_i + b - y_i)^2 = \sum (m(x_i - \overline{x}) - (y_i - \overline{y}))^2$$

$$\Longrightarrow \quad m = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} \qquad \left(= \frac{\frac{1}{n}\sum (x_i - \overline{x})(y_i - \overline{y})}{\sigma_x^2} = \frac{Cov_{xy}}{\sigma_x^2} \right)$$

Die Gleichung der Regressionsgeraden lautet: $y = m(x - \overline{x}) + \overline{y}$

Der Übergang von x_i zu $x_i - \overline{x}$ bzw. von y_i zu $y_i - \overline{y}$ bedeutet eine Verschiebung der Punktwolke, so dass der Schwerpunkt $P(\overline{x} \,|\, \overline{y})$ in den Ursprung fällt.

Die Kovarianz $Cov_{xy} = \frac{1}{n}\sum (x_i - \overline{x})(y_i - \overline{y})$
ist umso größer, je mehr positive Summanden vorhanden sind.
Ein Summand ist positiv, falls beide Werte x_i und y_i kleiner oder beide größer als der entsprechende Mittelwert sind. Es liegt dann eine steigende Tendenz vor.
In welchem Fall ist diese fallend?

GTR

Mit STAT | EDIT x-Werte in L1 und y-Werte in L2 eingeben.
Das Streudiagramm wird mit STAT PLOT gezeichnet.
STAT | CALC 4: LinReg(ax+b) aufrufen, vorher DiagnosticOn.
Mit LinReg(ax+b) Y1 (z. B.) wird das Ergebnis in Y1 gespeichert.
Falls die Werte nicht in L1 und L2 stehen, sind die Listen anzugeben, z. B. LinReg(ax+b) L2, L3, Y1 .
Nützlich ist auch STAT | CALC 2: 2−Var Stats.

Aufg.

An einer Tankstelle werden für die letzten vier Monate folgende Benzinpreise und Absatzmengen notiert:

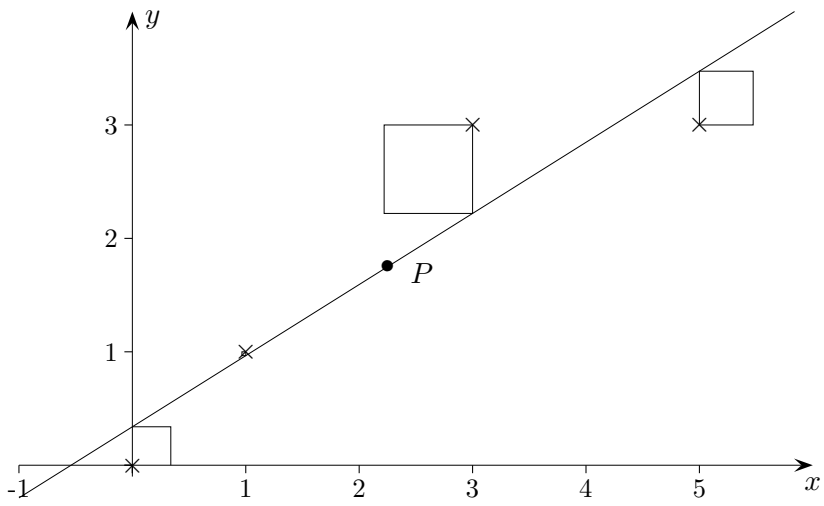
Preis (in €)	1,45	1,41	1,55	1,50
Menge (in 1000 l)	150	165	140	144

Bei welchem Preis ergäbe sich ein Absatz von 100 000 l? Lösung: 1,77 €

Regression weitere Alternative

Wir gehen zunächst von folgenden einfachen Daten aus:

x	y
0	0
1	1
3	3
5	3



Um von der optimalen Ausgleichsgeraden durch den Ursprung $y = mx$ die Steigung m zu ermitteln, muss präzisiert werden, was unter optimal zu verstehen ist. Die Summe der senkrechten Abstände zur Geraden zu minimieren erscheint aufgrund der Komplexität wenig aussichtsreich. Stattdessen könnten die Abstände in y-Richtung betrachtet werden. Die Beträge sind hinderlich. Wir erinnern uns an die Definition der Standardabweichung. Optimal wäre, wenn die Summe Q der Abweichungsquadrate minimal wäre. Mit Quadraten lässt sich leichter rechnen, als mit Wurzeln oder Beträgen.

$Q(m) = (m-1)^2 + (3m-3)^2 + (5m-3)^2$

Für das Minimum müsste gelten: $Q'(m) = 0$. Dies führt zu $m = \frac{5}{7}$ und der Ausgleichsgeraden $y = \frac{5}{7}x$.

Und nun allgemein:

x	x_1	x_2	x_3	$\dots$	x_n
y	y_1	y_2	y_3	$\dots$	y_n

$Q(m) = \sum (mx_i - y_i)^2$ Die Summe erstreckt sich stets von 1 bis n.

$Q'(m) = 0 \quad \Longrightarrow \quad m = \dfrac{\sum x_i y_i}{\sum x_i^2}$

Die Punktwolke hat den Schwerpunkt $P(\overline{x} \mid \overline{y})$ mit den Mittelwerten $\overline{x} = \frac{1}{n}\sum x_i$, $\overline{y} = \frac{1}{n}\sum y_i$.

Um die Ausgleichsgerade $y = mx + b$ (muss nicht durch den Ursprung verlaufen) zu ermitteln, kann die Punktwolke mit $x_i - \overline{x}$ und $y_i - \overline{y}$ so verschoben werden, dass der Schwerpunkt P in den Ursprung fällt.

Für diese Werte gilt, wie wir gerade gesehen haben: $m = \dfrac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$

Die Gleichung der Regressionsgeraden lautet dann: $y = m(x - \overline{x}) + \overline{y}$

Korrelationskoeffizient

Wir benötigen ein Maß dafür, wie stark die Datenpunkte um die Regressionsgerade streuen.
Dazu rechnen wir die quadratische Abweichung aus.

$$Q = \sum \big(m(x_1 - \overline{x}) - (y_1 - \overline{y}) \big)^2 \qquad m = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

... Klammern auflösen, m einsetzen, kürzen, wir halten uns damit nicht auf

$$= \sum (y_i - \overline{y})^2 - \frac{(\sum (x_i - \overline{x})(y_i - \overline{y}))^2}{\sum (x_i - \overline{x})^2}$$

Je größer der 2. Term $\frac{(\sum (x_i - \overline{x})(y_i - \overline{y}))^2}{\sum (x_i - \overline{x})^2}$ ist, desto kleiner ist die Quadratsumme.

Da $Q \geq 0$ ist, folgt $\quad 0 \leq \frac{(\sum (x_i - \overline{x})(y_i - \overline{y}))^2}{\sum (x_i - \overline{x})^2} \leq \sum (y_i - \overline{y})^2$

$$0 \leq \frac{(\sum (x_i - \overline{x})(y_i - \overline{y}))^2}{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2} \leq 1 \quad \Big| \text{ mit } \tfrac{1}{n^2} \text{ erweitern, } \sqrt{\ }$$

$$\iff \frac{|\, Cov_{xy} \,|}{\sigma_x \sigma_y} \leq 1$$

$$\iff -1 \leq \underbrace{\frac{Cov_{xy}}{\sigma_x \sigma_y}}_{= r} \leq 1$$

r heißt *Korrelationskoeffizient*.

Für $r = 0$ ist die Steigung m auch Null. Es liegt kein linearer Zusammenhang vor.
Für $r = 1$ oder $r = -1$ ist $Q = 0$, die Datenpunkte liegen auf der Regressionsgeraden.
Eine starke Korrelation (linearer Zusammenhang) der Daten besteht für $r \geq 0{,}8$ bzw. für $r \leq -0{,}8$.
Zu beachten ist, dass ein hoher Korrelationskoeffizient allein noch keine *kausale* Abhängigkeit bedeuten muss.
Die lineare Abhängigkeit von Daten x_i, y_i wird nicht verändert,
wenn sie verschoben werden: $x_i - \overline{x}$, $y_i - \overline{y}$. Die neuen Mittelwerte sind dann null.
Die lineare Abhängigkeit bleibt auch dann weiter erhalten, wenn diese Daten gestreckt oder gestaucht werden: $\frac{x_i - \overline{x}}{\sigma_x}$, $\frac{y_i - \overline{y}}{\sigma_y}$.
Die Standardabweichung dieser Daten ist 1.
Der Korrelationskoeffizient der Daten x_i, y_i kann nun veranschaulicht werden.
Er ist die Steigung der Regressionsgeraden der standardisierten Daten (siehe Excel-Blatt Korrelationskoeffizient).

Standardisierung

Um die Standardisierung noch etwas zu beleuchten, gehen wir von folgenden Daten aus:

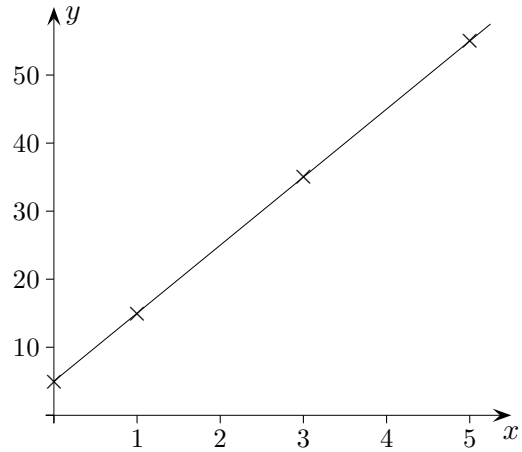
x	y	$\frac{x-\overline{x}}{\sigma_x}$	$\frac{y-\overline{y}}{\sigma_y}$
0	5	−1,01	−1,01
1	15	−0,56	−0,56
3	35	0,34	0,34
5	55	1,24	1,24

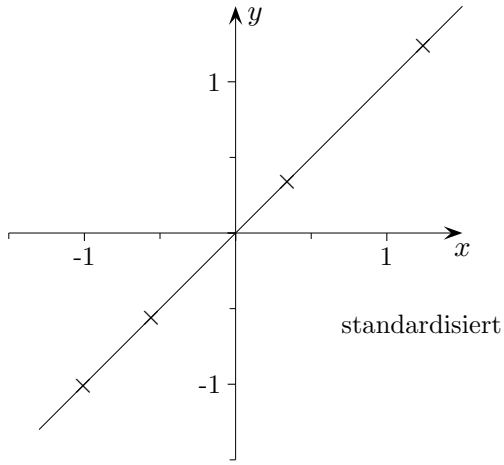
$\overline{x} = 2{,}25$
$\overline{y} = 27{,}50$
$\sigma_x = 2{,}22$
$\sigma_y = 22{,}17$

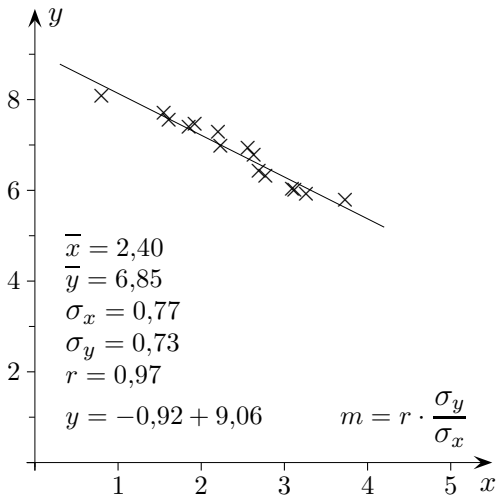
$$k = \frac{x-\overline{x}}{\sigma_x} \iff x - \overline{x} = k \cdot \sigma_x \iff x = \overline{x} + k \cdot \sigma_x$$

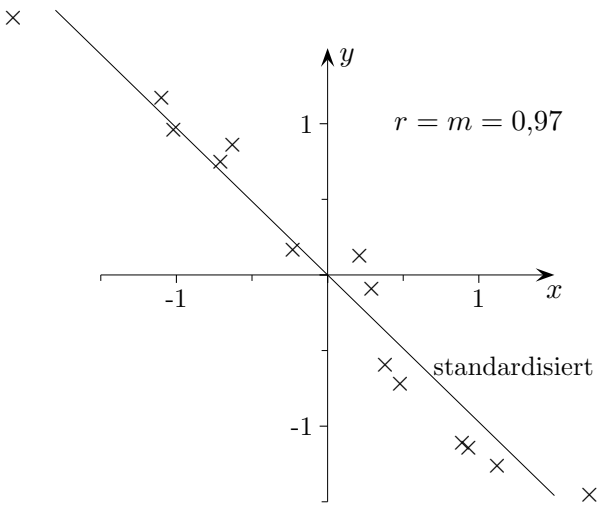
$$1{,}24 = \frac{5-\overline{x}}{\sigma_x} \iff 5 - \overline{x} = 1{,}24 \cdot \sigma_x \iff 5 = \overline{x} + 1{,}24 \cdot \sigma_x$$

Der k-Wert gibt an, um welches Vielfache von σ_x x vom Erwartungswert $\overline{x}$ abweicht.
An den standardisierten Werten ist die lineare Abhängigkeit unmittelbar zu erkennen: $r = m = 1$.
Obwohl die Daten $(x \mid y)$ völlig verschieden sind, sind die Vielfachen k identisch: $(\overline{x} + k \cdot \sigma_x \mid \overline{y} + k \cdot \sigma_y)$.





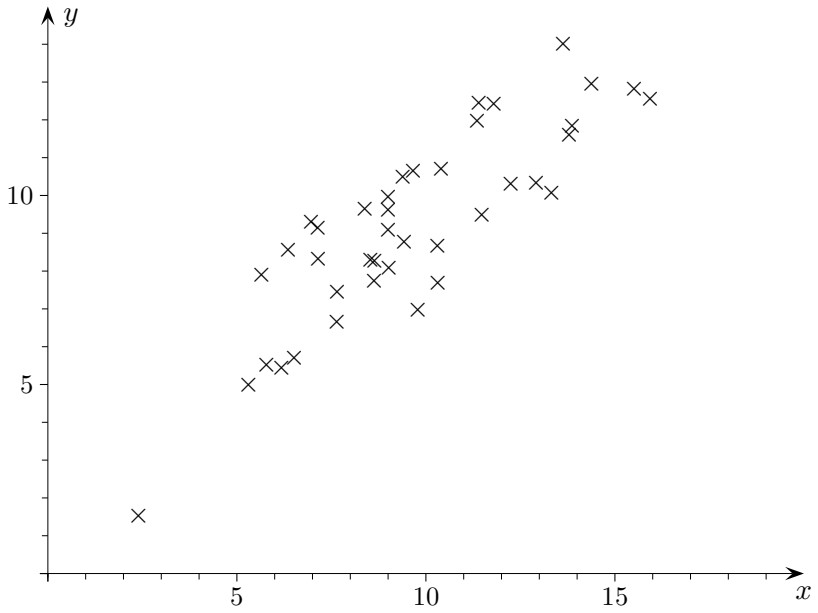




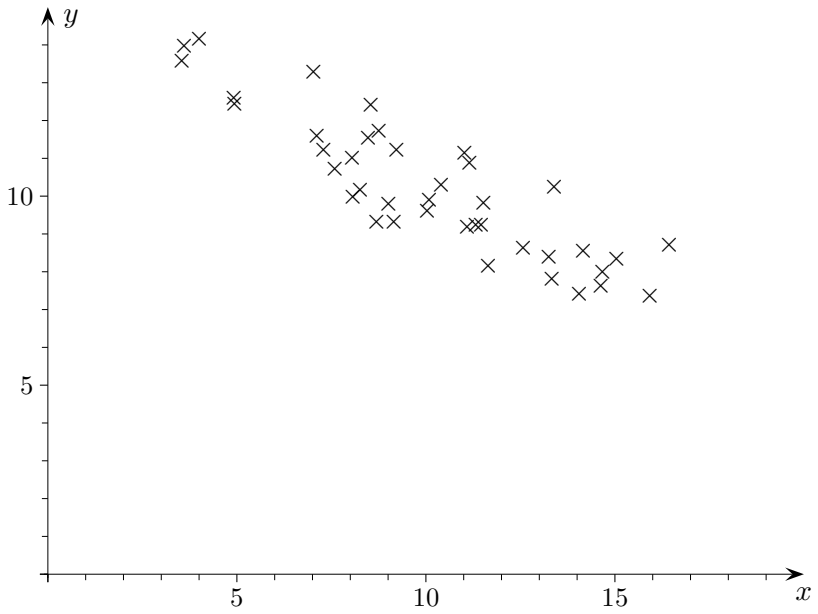
Streudiagramm

Ermittle die Gleichung einer Geraden, die möglichst gut mit der Richtung der Punktwolke übereinstimmt.

a)



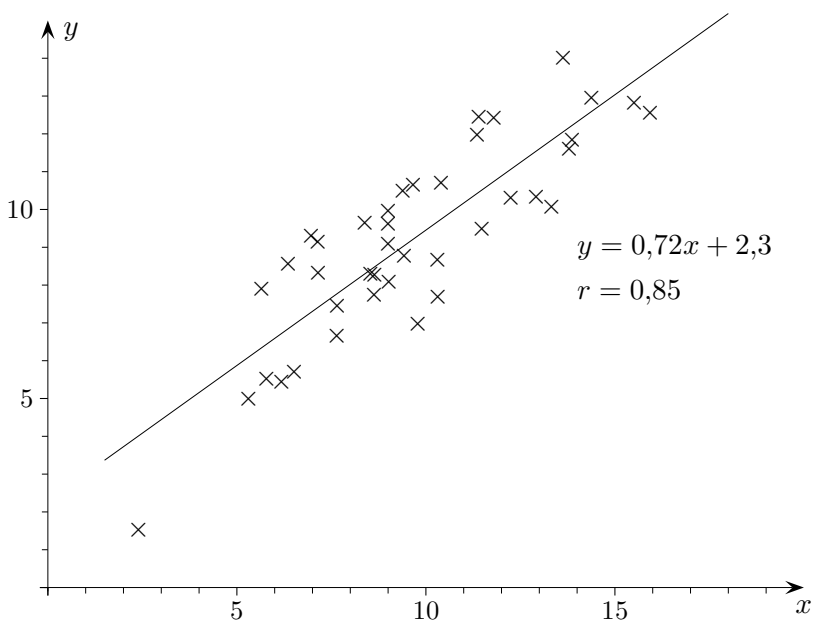
b)



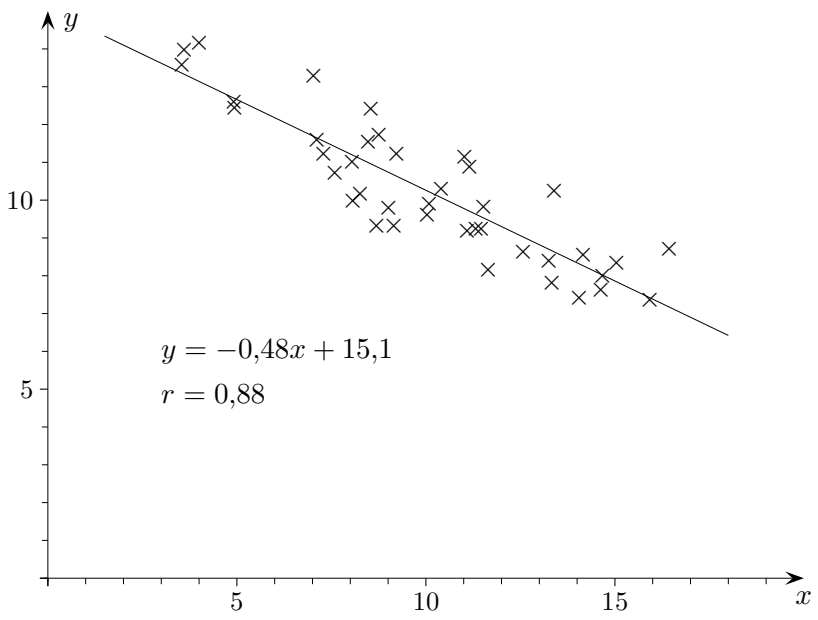
Streudiagramm

Ermittle die Gleichung einer Geraden, die möglichst gut mit der Richtung der Punktwolke übereinstimmt.

a)



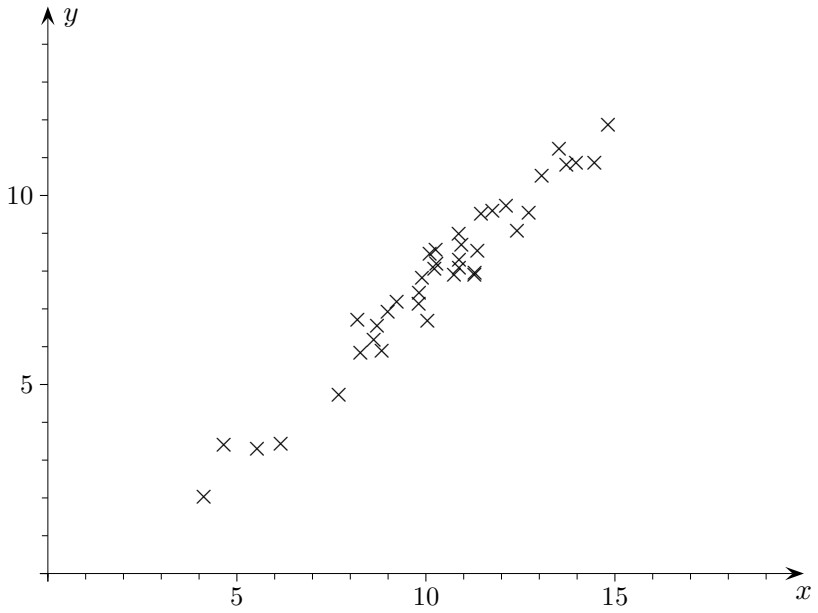
b)



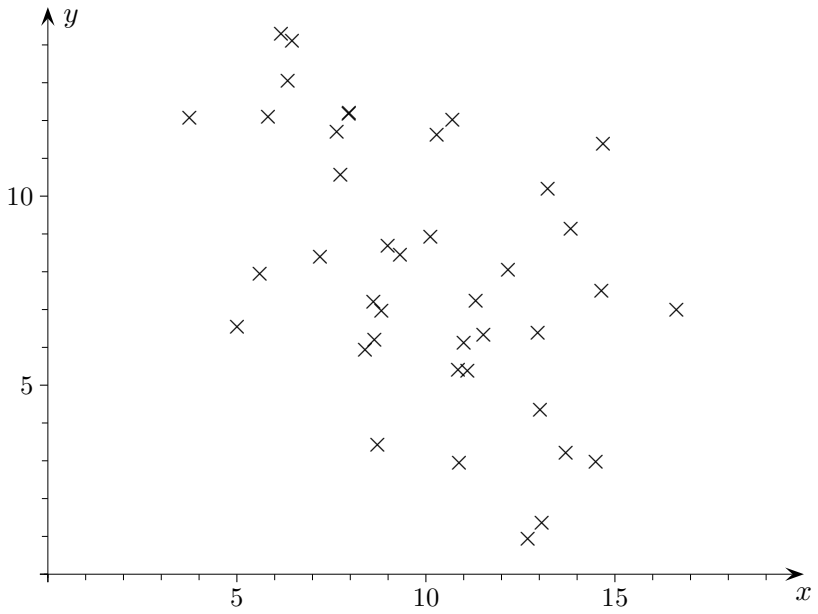
Streudiagramm

Ermittle die Gleichung einer Geraden, die möglichst gut mit der Richtung der Punktwolke übereinstimmt.

c)



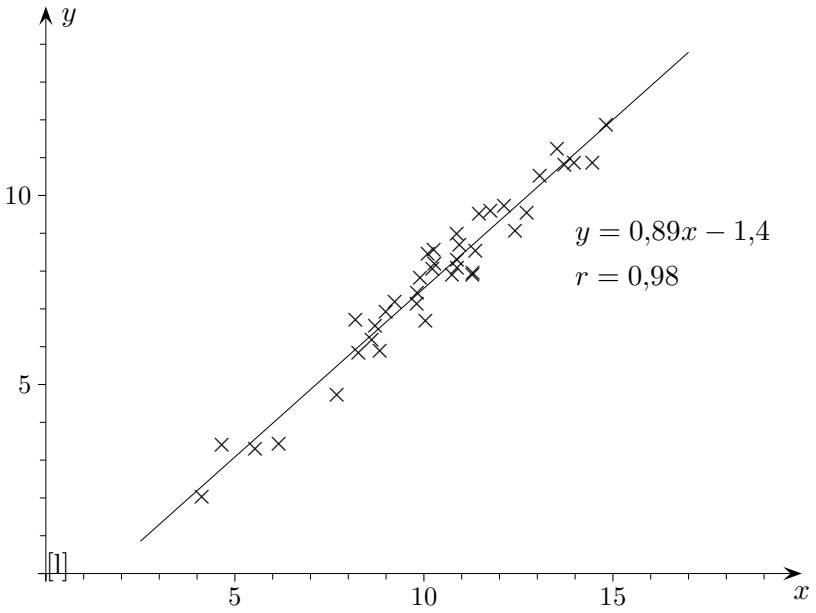
d)



Streudiagramm

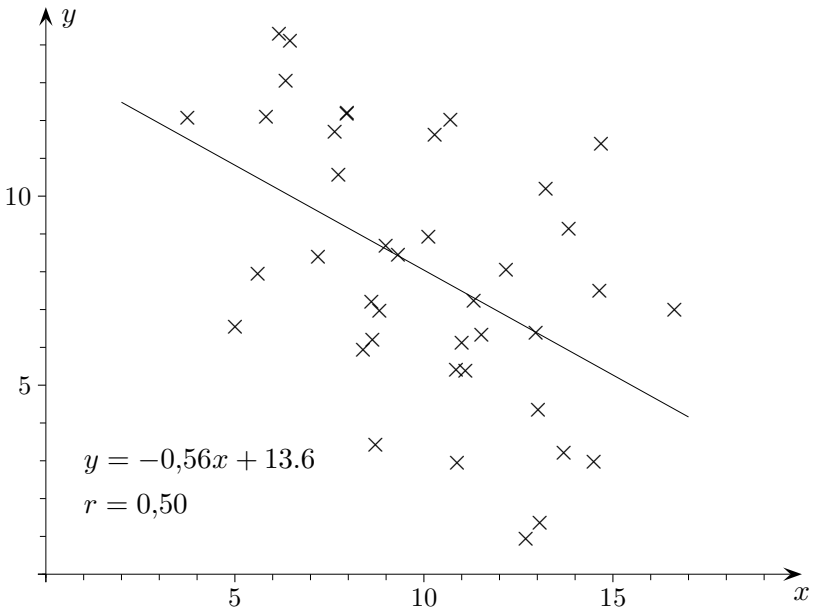
Ermittle die Gleichung einer Geraden, die möglichst gut mit der Richtung der Punktwolke übereinstimmt.

c)



[1]

d)



Eigenschaften des Mittelwerts

Unter dem arithmetischen Mittel (Mittelwert) $\overline{x}$ von n Zahlen verstehen wir:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} (x_1 + x_2 + \ldots + x_n)$$

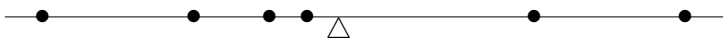
Diesen Mittelwert untersuchen wir etwas genauer.

1. Zeige für $n = 3$:

   $$\sum_{i=1}^{n} (\overline{x} - x_i) = 0$$

   d. h. die Summe der Abweichungen vom Mittelwert ist Null.

2. $\overline{x}$ kann als Schwerpunkt interpretiert werden, erläutere dies.



$$\sum_{x_i < x_s} (x_s - x_i) \; = \sum_{x_s < x_i} (x_i - x_s)$$

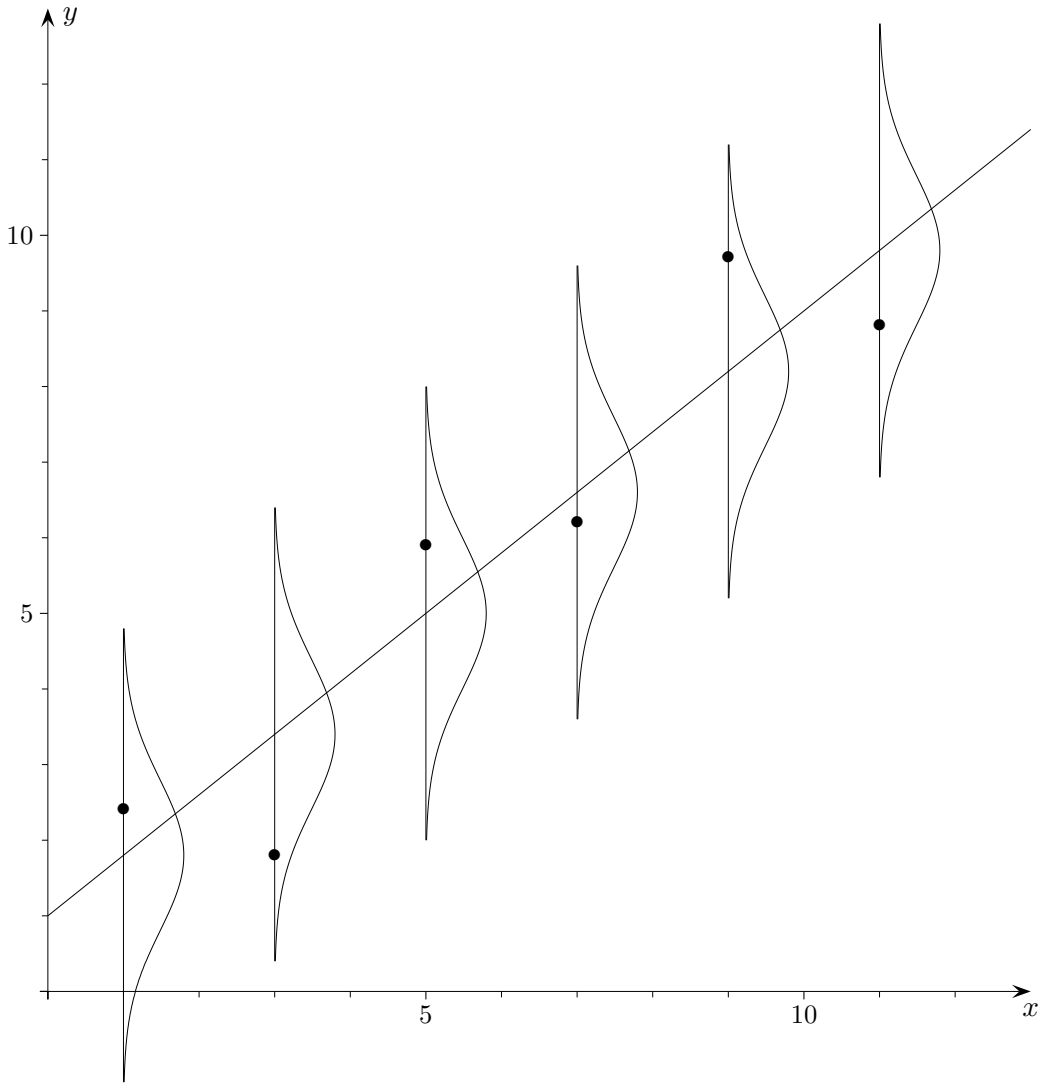
$$\Longrightarrow \quad x_s = \frac{1}{n} \sum_{i=1}^{n} x_i$$

3. Zeige: $x = \overline{x}$ minimiert die quadratische Funktion:

   $$f(x) = \sum_{i=1}^{n} (x - x_i)^2$$

   d. h. die Summe der quadratischen Abweichungen wird für $x = \overline{x}$ minimal.

Regression und Normalverteilung



Erläutere die Grafik.

Regression, Korrelation Aufgaben

1. In einer Versuchsreihe wurden sieben Schüler bezüglich ihres Stressverhaltens vor einer Mathematikarbeit untersucht. Die Parameter des Stressverhaltens werden in relativer Angabe auf einer Skala von 1 bis 10 bezogen, die Mathematikleistung in Prozent der erreichbaren Punktzahl.

Stressverhalten	6,5	4,0	2,5	7,2	8,1	3,4	5,5
Ergebnis der Arbeit	81	96	93	68	63	84	71

   a) Zeichnen Sie die Daten in ein geeignetes Koordinatensystem und beurteilen Sie, ob ein linearer Zusammenhang vorliegt.

   b) Bestimmen Sie die Regressionsgerade und zeichnen Sie diese in das Koordinatensystem ein. Erläutern Sie an diesem Beispiel die Methode der kleinsten Fehlerquadrate.

   c) Erläutern Sie, inwiefern sich die Regressionsgerade für Vorhersagen eignet.

2. Bestimmen Sie für folgende Daten den Korrelationskoeffizienten und die Gleichung der Regressionsgeraden und beurteilen Sie den Zusammenhang.

Alkoholgehalt [‰]	0,2	0,3	0,4	0,6	0,8	1,0
Reaktionszeit [s]	0,13	0,158	0,18	0,23	0,27	0,33

3. In alten Aufzeichnungen finden sich verlässliche Daten über die Anzahl der Storchenpaare und der Einwohnerzahl einer Region.

Jahr	1930	1931	1932	1933	1934	1935	1936
Anzahl der Storchenpaare	132	142	166	188	240	250	252
Anzahl der Einwohner	55400	55400	65000	67700	69800	73300	76000

   Bestimmen Sie den Korrelationskoeffizienten und interpretieren Sie das Ergebnis. Erläutern Sie in diesem Zusammenhang, was der Korrelationskoeffizient aussagt.

4. Gehen Sie anhand der Daten (Bayerischer Wald) der Frage nach, ob die dümmsten Bauern die dicksten Kartoffeln haben. Der Umfang ist der mittlere Kartoffelumfang in cm.

Bauer	1	2	3	4	5	6	7	8	9	10
IQ	90	103	98	88	123	108	89	95	116	100
Umfang	14	18	19	21	16	13	12	17	15	15

korrelierte Zufallsvariablen

Seien X und Y Zufallsvariablen mit den Erwartungswerten $E(X) = \mu_X$ und $E(Y) = \mu_Y$.

Für die Varianzen gilt dann:

$$\begin{aligned} V(X) &= E(X-\mu_X)^2 \\ V(Y) &= E(Y-\mu_Y)^2 \end{aligned}$$

$$\begin{aligned} V(X+Y) &= E\big((X+Y)-(\mu_X+\mu_Y)\big)^2 \\ &= E\big((X-\mu_X)+(Y-\mu_Y)\big)^2 \\ &= E\big((X-\mu_X)^2+2(X-\mu_X)(Y-\mu_Y)+(Y-\mu_Y)^2\big) \\ &= V(X)+2\underbrace{E\big((X-\mu_X)(Y-\mu_Y)\big)}_{Cov(X,Y)}+V(Y) \end{aligned}$$

$$V(X+Y) = V(X)+V(Y)+2Cov(X,Y)$$

Die Bezeichnung Kovarianz sollte nun verständlich sein.
Für unabhängige Zufallsvariablen gilt:

$$V(X+Y) = V(X)+V(Y)$$

Für $Cov(X,Y) > 0$ überwiegen gleichsinnige Abweichungen der Zufallsvariablen von ihren Erwartungswerten, d. h. ihre Werte liegen gemeinsam drüber bzw. drunter, X und Y heißen dann positiv korreliert, betrachte $(X-\mu_X)(Y-\mu_Y)$,

$Cov(X,Y) < 0$ überwiegen gegensinnige Abweichungen der Zufallsvariablen von ihren Erwartungswerten, X und Y heißen dann negativ korreliert,

$Cov(X,Y) = 0$ heißen die Zufallsvariablen unkorreliert.

Das Gleiche gilt für den Korrelationskoeffizienten $r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$.

Für ein Gut liegen 15 Beobachtungspaare aus 3 aufeinanderfolgenden Jahren
für den Stückpreis p und die zugehörige nachgefragte Menge y vor:

	1. Jahr					2. Jahr					3. Jahr				
p	2	3	4	5	7	10	12	14	15	16	19	21	23	25	26
y	7	5	6	3	2	9	7	6	4	3	13	10	9	7	5

a) Ermitteln Sie die Regressionsgerade zu den 15 Wertepaaren.

b) Spricht das Ergebnis für einen positiven Zusammenhang zwischen Preis und Nachfrage?

c) Sehen Sie bei diesem Regressionsansatz ein Problem?