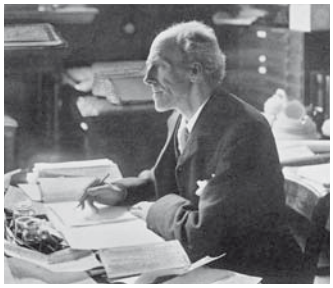
2 Regression und Korrelation

Regressionsanalysen sind statistische Methoden zur mathematischen Beschreibung von Beziehungen zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen. Dabei greift man auf die Theorie der kleinsten Fehlerquadrate von **Carl Friedrich Gauß** (1777 – 1855) zurück, die erlaubt, Abweichungen von einer „Modellkurve" mithilfe der Differenzialrechnung zu minimieren.



Carl Pearson

Korrelation bewertet den Zusammenhang zwischen zwei variablen Größen. Hier bewährt sich unter anderem der nach **Carl Pearson** (1857 – 1936) benannte Korrelationskoeffizient. Pearson war ein britischer Gelehrter, der nicht nur in Naturwissenschaft, sondern in den unterschiedlichsten Wissenschaftsbereichen sehr erfolgreich tätig war.

2.1 Regression (Ausgleichsrechnung)

Bei Verwendung von Formeln aus der Geometrie oder der Natur- und Wirtschaftswissenschaft hast du es häufig mit dem Zusammenhang zwischen zwei Größen an demselben Untersuchungsobjekt zu tun gehabt. Beispielsweise gibt die Formel $A = r^2\pi$ einen exakten Zusammenhang zwischen dem Radius und der Fläche eines Kreises wieder.

Wenn du aber zB eine Versuchsreihe über den von dir als möglich angesehenen Zusammenhang zwischen den Messgrößen Körpermasse und Körpergröße der Kolleginnen und Kollegen in deiner Klasse machst, so kann mit den Mitteln der Statistik dieser Zusammenhang in Form einer Gleichung bzw. einer Funktion beschrieben werden. Dadurch können auch Prognosen über zukünftige bzw. noch nicht gemessene Entwicklungen gemacht werden.

2.1.1 Der lineare Zusammenhang zweier Größen

ABD **2.1** Die Messung der Körpergröße in Zentimeter (cm) und der Körpermasse in Kilogramm (kg) in einer ausgewählten Stichprobe aus deiner Klasse ergab folgende Wertetabelle:

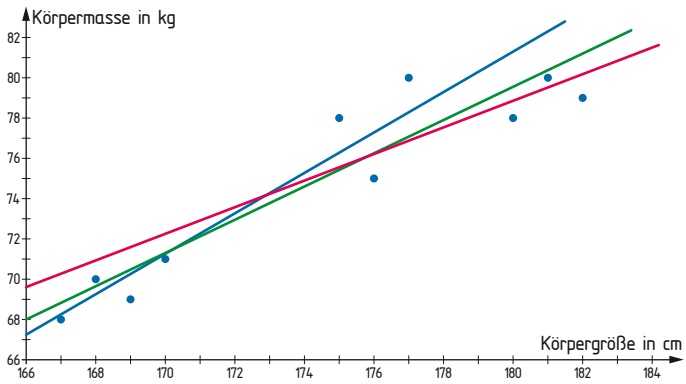
Größe (cm)	167	168	169	170	175	176	177	180	181	182
Masse (kg)	68	70	69	71	78	75	80	78	80	79

Beurteile mit Hilfe einer Grafik, ob ein Zusammenhang zwischen den Messgrößen besteht.
Drücke diesen Zusammenhang näherungsweise mithilfe einer linearen Gleichung aus.
Erkläre, wie man zu dieser Gleichung mithilfe der Differenzialrechnung kommt.

Zeichne zunächst die Messpunkte in ein Koordinatensystem ein. Du erhältst ein so genanntes **Streudiagramm** (eine **Punktewolke**). Das Diagramm zeigt den Zusammenhang in losen Punkten an und bestätigt die Tendenz, dass mit einer größeren Körpergröße auch eine größere Körpermasse verbunden ist.

Die Punkte liegen aber nicht auf einem erkennbaren Funktionsgraphen, weder auf einer Geraden noch auf einer Parabel noch auf einer exponentiellen Wachstumskurve.

Mit Augenmaß könntest du eine Gerade in die Punktewolke hineinlegen, die den „Trend“ des Zusammenhangs grob beschreibt.



Das Problem: Alle 3 Geraden in der Grafik würden in Frage kommen. Wir brauchen jene Modellgerade, von der die Punkte so wenig wie möglich abweichen sollen.

Es stellt sich zunächst die Frage, was unter **„Abweichung“** zu verstehen ist. Da du nicht weißt, wo genau die beste Gerade liegt, können ihre Funktionswerte nur allgemein mit $\boldsymbol{y = k \cdot x + d}$ bezeichnet werden.

Als Abweichung der einzelnen Punkte $(x_i|y_i)$ von der von uns gesuchten Modellgeraden definieren wir den Abstand, den der einzelne y_i-Wert eines Punkts vom $\boldsymbol{y}$-Wert der Modellgeraden hat, also $\boldsymbol{(y_i - y)}$. Diesen Abstand nennt man **Residuum**.

Die Residuen der einzelnen Punkte können positiv oder negativ sein, je nachdem, ob der einzelne Punkt oberhalb oder unterhalb der Geraden liegt. Beim Aufsummieren der Abweichungen heben sie sich gegenseitig auf. Durch Quadrieren kann diese Vorzeichen-Problematik beseitigt werden. Außerdem werden größere Abweichungen stärker als kleinere berücksichtigt. Also quadrieren wir die einzelnen Residuen und bilden ihre Summe s_n.

Auf der Bedingung, dass die Summe der quadratischen Abweichungen ein Minimum ist, beruht die **Gauß’sche „Methode der kleinsten Fehlerquadrate“**.

Erinnere dich, dass ein Funktionsgraph nur dann ein lokales Minimum aufweist, wenn die 1. Ableitung null ist. Das ist eine notwendige Bedingung für ein Minimum. Daher muss die Summe der quadratischen Abweichungen differenziert und null gesetzt werden.

Wir suchen das $\boldsymbol{k}$ und das $\boldsymbol{d}$ der Modellgeraden. Daher musst du die Fehlerquadratsumme einmal nach der Variablen k und einmal nach der Variablen d differenzieren. Beide Ableitungen werden gleich null gesetzt. Das dadurch entstehende Gleichungssystem kann anschließend gelöst werden.

Es wird der Rechengang nun im Einzelnen vorgeführt. Dabei verwenden wir die Kurzform des Summenzeichens ohne Indizes:

$\sum x_i$ statt $\sum_{i=1}^{n} x_i$, n ist die Anzahl der Datenpunkte.

$s_n = \sum (y_i - (k \cdot x_i + d))^2$ ... Fehlerquadratsumme $\rightarrow$ Minimum

Es handelt sich hier um eine verkettete Funktion, die zum Ableiten die Kettenregel benötigt. Die innere Funktion ist $y_i - (k \cdot x_i + d)$, die äußere Funktion ist das Quadrat.

Ableitung von s_n nach k:

$$2 \cdot \sum (y_i - (k \cdot x_i + d)) \cdot (-x_i) = 2 \cdot \sum (-x_i \cdot y_i + k \cdot x_i^2 + d \cdot x_i) =$$
$$= 2 \left(-\sum x_i \cdot y_i + k \cdot \sum x_i^2 + d \cdot \sum x_i\right)$$

Ableitung von s_n nach d:

$$2 \cdot \sum (y_i - (k \cdot x_i + d)) \cdot (-1) = 2 \left(-\sum y_i + k \cdot \sum x_i + d \cdot n\right)$$

Setze nun beide Ableitungen gleich null, dividiere durch 2 und ordne nach k und nach d.

Das Gleichungssystem zur Berechnung der Parameter k und d der so genannten **linearen Regressionslinie oder Trendlinie**, die sich am besten den Punkten des Streudiagramms anpasst:

$$\text{I: } k \cdot \sum x_i^2 + d \cdot \sum x_i = \sum x_i \cdot y_i$$
$$\text{II: } k \cdot \sum x_i + d \cdot n = \sum y_i$$

Mithilfe dieses Gleichungssystems kannst du k und d bestimmen.
Wir erhalten für die angegebene Tabelle die folgenden Summen:

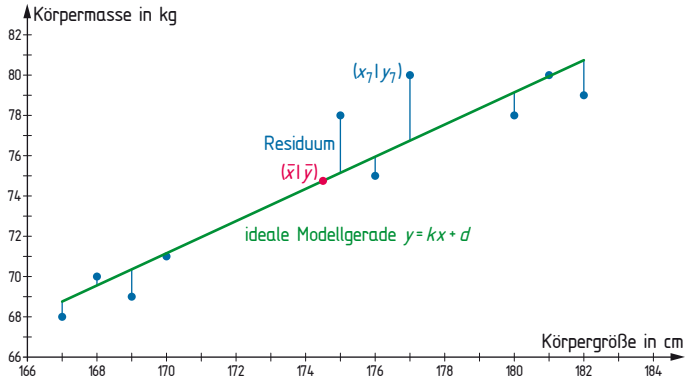
$$\sum x_i = 1\,745; \quad \sum x_i^2 = 304\,789; \quad \sum y_i = 748; \quad \sum x_i \cdot y_i = 130\,755$$

Eingesetzt in das Gleichungssystem ergibt dies:

$$\left.\begin{array}{l} \text{I: } 304\,789k + 1\,745d = 130\,755 \\ \text{II: } 1\,745k + 10d = 748 \end{array}\right\} \Rightarrow k = 0{,}799 \text{ und } d = -64{,}678$$

Die optimale Gerade – **die Regressionsgerade (= Trendlinie)** – hat daher die Gleichung $y = 0{,}799x - 64{,}678$.

In der nachstehenden Grafik ist die Regressionsgerade (= Trendlinie) grün eingezeichnet. Diese Linie ist die ideale Modellgerade, sie ist mithilfe der Gauß'schen Methode den Punkten optimal angenähert. Die Residuen, das sind die Abstände $\mathbf{(y_i - y)}$ der Punkte von der idealen Geraden, sind eingezeichnet.



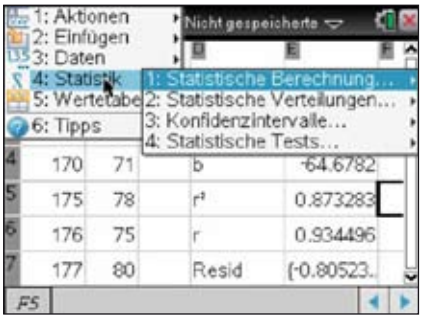
Die Trendlinie hat noch eine zusätzliche Besonderheit:
Der Punkt $(\bar{x}|\bar{y})$ liegt auf der Regressionsgeraden. Dabei ist $\bar{x}$ der arithmetische Mittelwert aller x_i-Werte und $\bar{y}$ der arithmetische Mittelwert aller y_i-Werte.

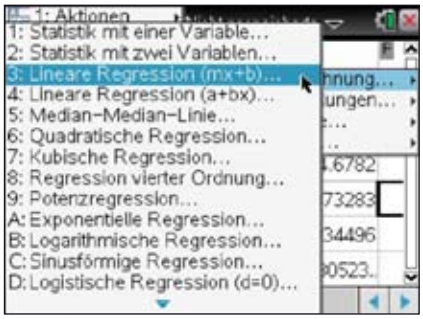
Rechengeräte bzw. Computer haben diesen Rechenvorgang im Statistik-Anwendungsprogramm gespeichert.

Wir setzen daher für die Berechnung der Regression ab nun stets Technologie ein.

TI-Nspire:

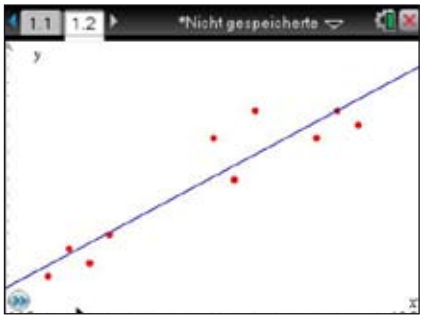
Wir geben die Werte beider Größen in zwei Listen l1 und l2 einer Tabellenkalkulation ein und verwenden **Menu/4: Statistik/1: Statistische Berechnungen/3: Lineare Regression (mx+b)**





Es öffnet sich ein Fenster, in das die Listen eingegeben werden können, in denen die *x*- bzw. die *y*-Werte der Punkte zu finden sind. In unserem Falle geben wir l1 und l2 ein, die als Variable zu verstehen sind (Variablenverweis!). In f1 wird die berechnete **Regressionslinie (= Trendlinie)** als Grafik gespeichert.

m steht hier für den Anstieg *k* und *b* steht für den Ordinatenabschnitt *d*. Wir erhalten die „beste Gerade" durch die Punktewolke als **Regressionslinie (= Trendlinie)** mit ***y* = 0,799*x* – 64,678**.



Streudiagramm mit TI-Nspire:
2: Graphs/Menu/3: Grafiktyp/
4: Streudiagramm/x...l1;y...l2

Zeichnen der Regressionslinie:
3: Grafiktyp/1: Funktion/f2(x)=f1(x)

Technologieeinsatz zu diesem Thema für TI 82-84, EXCEL und Geogebra-CAS siehe www.hpt.at (Schulbuch Plus für Schüler/innen)

Mithilfe der Trendlinie können verschiedenste **Prognosen** erstellt werden:
Bezogen auf das Eingangsbeispiel zB: Wie viel Kilogramm wiegt ein Jugendlicher in deiner Klasse mit einer Körpergröße von 172 cm? Setze $x = 172$ in die Gleichung der Trendlinie ein.
Man erwartet nach dem berechneten Modell eine Körpermasse von 72,75 kg $\approx$ 73 kg.
Wie groß wäre ein Jugendlicher in deiner Klasse, wenn sein Gewicht 79 kg beträgt?
Setze 79 für y ein und berechne $x = 179{,}82$ cm $\approx$ 180 cm.
Selbstverständlich handelt es sich hier nur um **Schätzwerte**, weil die tatsächlichen Messwerte im Allgemeinen nie exakt im Trend liegen.

B **2.2** Der Zusammenhang von 2 Merkmalen A und B wurde mit der folgenden Tabelle erfasst:

A	2	3	4	6	7	8	10	12
B	4	3	5	6	8	7	9	13

Berechne die Parameter der Regressionsgeraden.
Berechne den fehlenden Wert für $A = 9$ und für $B = 14$.

B **2.3** Der Zusammenhang von 2 Merkmalen C und D wurde mit der folgenden Tabelle erfasst:

C	1	3	5	6	9	11	14	15
D	10	12	15	14	18	23	27	30

Berechne die Parameter der Regressionsgeraden.
Berechne den fehlenden Wert für $C = 8$ und für $D = 32$.

AB **2.4** Aus Angeboten verschiedener Hersteller wurden zufällig 10 Autos ausgewählt und ihre Leistung in Kilowatt (kW) mit der Spitzengeschwindigkeit in Kilometer pro Stunde (km/h) verglichen.

Leistung	44	74	55	136	128	74	59	110	85	115
Spitzen-geschwindigkeit	157	188	160	240	215	195	163	225	200	200

Erstelle den Zusammenhang der beiden Größen mithilfe einer linearen Regression.
Ermittle eine Prognose, welche Spitzengeschwindigkeit nach diesem Modell bei einer Leistung von 139 kW zu erwarten ist.
Schätze mithilfe des Modells ab, welche Leistung bei einer Spitzengeschwindigkeit von 220 km/h benötigt wird.

ABD **2.5** Für die Schwefeldioxid-Emission (SO_2) in Österreich liegen für eine Periode von 8 Jahren folgende Angaben (in 1 000 t) vor:

Jahr	1	2	3	4	5	6	7	8
SO_2	60,4	56,3	53,8	52,8	50,7	45,8	41,4	40,8

a) Zeichne die Punktewolke und ein Liniendiagramm.
b) Ermittle die Trendlinie.
Berechne die Vorhersagewerte für das 10. Jahr.
Erkläre, wie man aus der Trendlinie ermitteln kann, in welchem Jahr ein Wert von 35 000 t unterschritten wird.

2.6 ABD

Das Bruttoinlandsprodukt (kurz BIP) wird als Indikator für die wirtschaftliche Leistung einer Volkswirtschaft verwendet. Es setzt sich aus dem ökonomischen Wert aller im Inland innerhalb eines bestimmten Zeitraums produzierten Güter und Dienstleistungen zusammen.
Für eine bestimmte bereits vergangene Periode lagen in Österreich für 6 Jahre die folgenden Daten für das BIP in Milliarden Euro vor:

Jahr	1	2	3	4	5	6
BIP	123,48	133,60	143,23	151,83	156,94	165,41

a) Zeichne ein Liniendiagramm.
b) Bestimme die Trendlinie.
Berechne die Vorhersagewerte für das Jahr 10.
c) Suche auf der Website der STATISTIK AUSTRIA die aktuellen Werte der letzten 6 Jahre und berechne die Vorhersage für das 10. Jahr neu.

2.7 AB

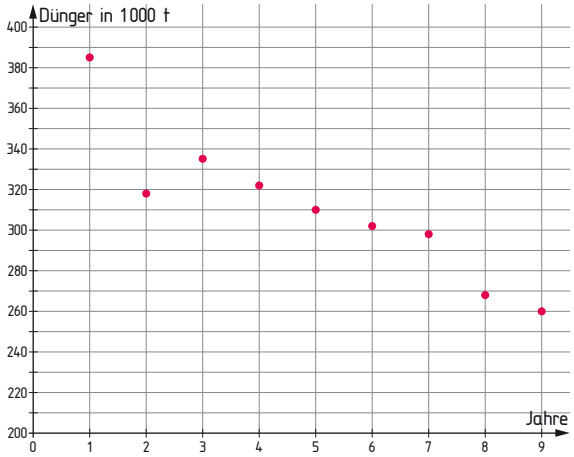
Ein Lichtstrahl wird in eine Serie von verdünnten Lösungen geschickt. Man misst den Schwächungsgrad der Lichtintensität anhand des (dimensionslosen) Extinktionskoeffizienten E.
Die Konzentration c der Lösungen ist in Mol pro Liter (mol/ℓ) angegeben.
Die Messung ergab die folgenden Werte:

Konzentration	0,01	0,05	0,1	0,5	1,0	1,5	2,0
Extinktion	0,003	0,016	0,03	0,15	0,3	0,45	0,6

a) Erstelle die Gleichung für den linearen Zusammenhang der beiden Größen.
b) Schätze, welche Konzentration für eine Lösung mit der Extinktion $E = 0{,}22$ zu erwarten ist.
c) Schätze, welche Extinktion für eine Lösung mit der Konzentration $c = 0{,}75$ mol/ℓ zu erwarten ist.

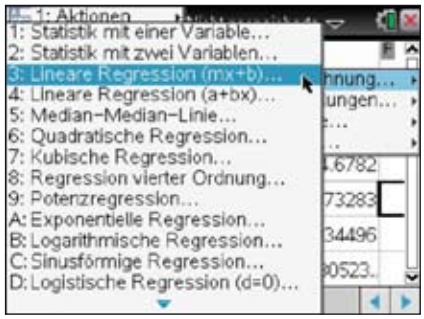
2.8 ABD

In einer bestimmten Region liegen für 10 Jahre die Daten zum Verbrauch von Dünger in 1 000 Tonnen (t) in Form eines Streudiagramms vor:
a) Erstelle die Gleichung der Trendlinie.
Zeichne sie in das Diagramm.
b) Berechne die Vorhersagewerte für das 12. Jahr.
Erkläre, wie man aus der Trendlinie ermitteln kann, in welchem Jahr ein Wert von 250 000 t unterschritten wird.



2.1.2 Nichtlineare Zusammenhänge zweier Größen

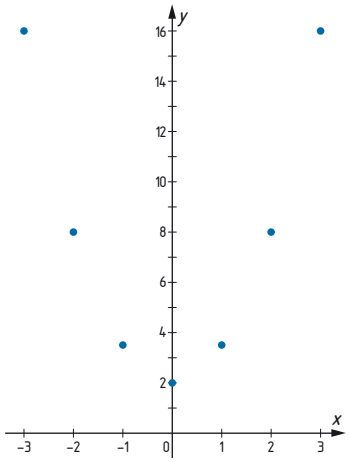
Die Programme der Rechner liefern auch Regressionskurven für nichtlineare Zusammenhänge. Sie können daher mithilfe von Technologie im Prinzip genauso ermittelt werden wie lineare Zusammenhänge. Das einzige, was du an Wissen beisteuern musst, damit du die richtige Auswahl treffen kannst, ist die Grundkenntnis über den Verlauf einiger häufig vorkommender Kurven. Einige Beispiele, die du möglicherweise mithilfe deiner verfügbaren Technologie einer Punktewolke zuordnen kannst:



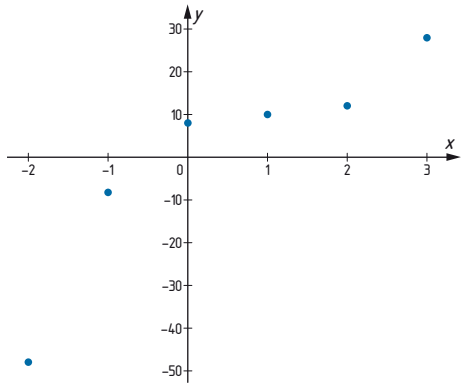
Lists/4: Statistik/1: Statistische Berechnungen

Polynomfunktion 2. Grads: $y = ax^2 + bx + c$ ... parabelförmig
Polynomfunktion 3. Grads: $y = ax^3 + bx^2 + cx + d$ ... s-förmig
Exponentialfunktion: $y = a \cdot e^{bx}$ ... monoton steigend oder fallend mit einer waagrechten Asymptote
Logistisches Wachstum: $y = \frac{c}{1 + a \cdot e^{-bx}}$ ... zusammengesetzt aus einer progressiv steigenden, durch einen Wendepunkt gehenden und schließlich degressiv verlaufenden Kurve mit waagrechten Asymptoten.

Polynomfunktionen

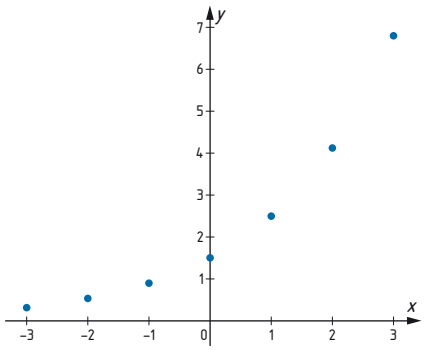


quadratisch: $y = ax^2 + bx + c$



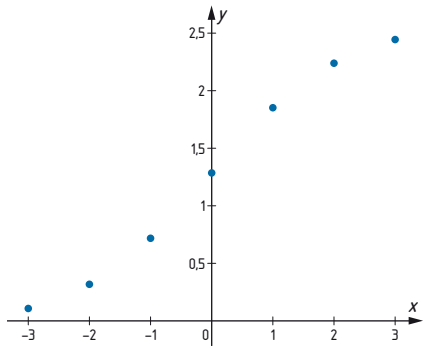
kubisch: $y = ax^3 + bx^2 + cx + d$

exponentielle Funktion



exponentiell: $y = a \cdot e^{bx}$

logistische Kurve



logistisch: $y = \frac{c}{1 + a \cdot e^{-bx}}$

2.9 Die Auswertung der Untersuchung des Merkmals Y in Abhängigkeit vom Merkmal X ergab die folgenden Messwerte:

ABD

X	1	2	3	4	5	6	7
Y	0,38	1,15	2,71	3,92	5,93	8,56	11,24

a) Zeichne die Punktewolke.
b) Ermittle eine lineare Regressionslinie dieser Punktewolke.
c) Ermittle eine exponentielle Regressionslinie.
d) Argumentiere, welche der beiden Linien ***b)*** oder ***c)*** optisch besser an die Punkte angenähert ist.

2.10 Ein Betrieb hat die folgenden Daten zu den abgesetzten Mengen eines Produkts in Mengeneinheiten (ME) und dem Erlös in Geldeinheiten (GE):

BD

Menge x	0	1	2	3	4	5	6
Erlöse E	0	127	199	227	196	124	10

a) Zeichne ein Streudiagramm.
b) Ermittle die Gleichung der Erlösfunktion E mithilfe einer quadratischen Regression.
Zeichne die Trendlinie in das Streudiagramm.
Beurteile anhand der Grafik die Güte der Anpassung.
c) Berechne die obere Erlösgrenze.

2.11 Die Gesamtkosten in Geldeinheiten (GE) bei der Erzeugung eines bestimmten Produkts wurden in Abhängigkeit von der Absatzmenge in Mengeneinheiten (ME) untersucht und ergaben die folgenden Daten:

BD

Menge x	0	1	2	3	4	5	6
Kosten K	100	117	134	147	163	200	244

a) Zeichne ein Streudiagramm.
b) Ermittle die Gleichung der Kostenfunktion K mithilfe einer kubischen Regression.
Zeichne die Trendlinie in das Streudiagramm.
Beurteile anhand der Grafik die Güte der Anpassung.
c) Berechne die Ableitung von K, dh die Grenzkostenfunktion K'.

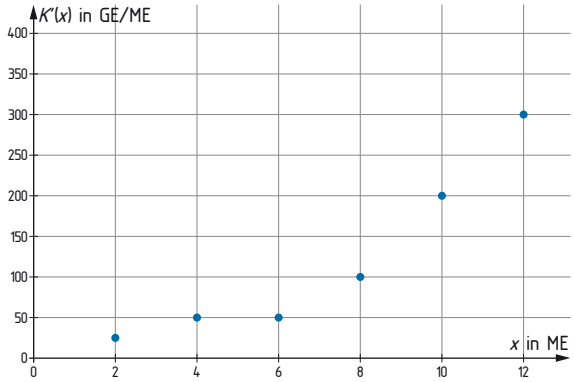
2.12 Die Preisgestaltung in Euro pro Mengeneinheit (€/ME) eines Produkts hängt mit der auf dem Markt nachgefragten Menge in Mengeneinheiten (ME) zusammen.
Man stellt für dieses Produkt eine diesbezügliche Untersuchung an und findet die folgenden Werte:

ABD

Menge	0	10	20	30	40	50	60	70	80
Preis	330	324	310	288	258	220	174	120	58

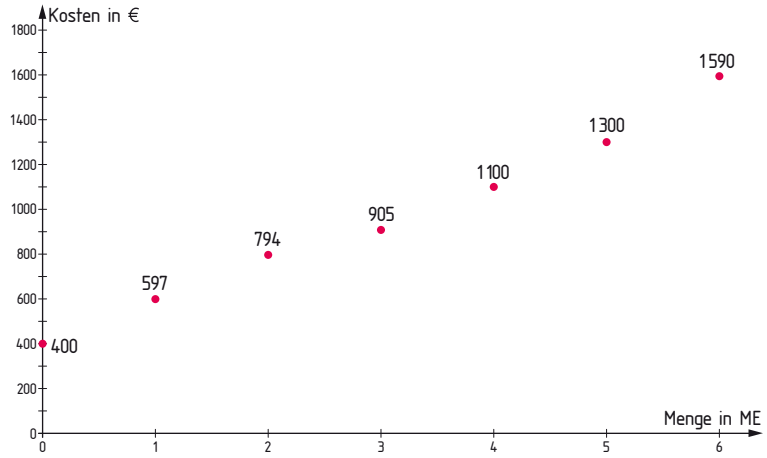
a) Zeichne die Punktewolke.
b) Erstelle die Gleichung der quadratischen Trendlinie.
Berechne die Nullstelle dieser Funktion und interpretiere das Ergebnis.
Argumentiere, weshalb die von dir berechnete quadratische Regression im Sachzusammenhang angemessen ist.

ABCD **2.13** Die Gesamtkostenfunktion K beschreibt alle anfallenden Herstellkosten in Abhängigkeit von der Produktionsmenge x. Unter der Grenzkostenfunktion K' versteht man die Ableitungsfunktion von K. K' gibt näherungsweise Auskunft über die Kostenänderungen, wenn man die Produktionsmengen um eine Einheit steigert oder vermindert.
Die Erhebung in einer Firma hat das folgende Streudiagramm für K' ergeben:



Erstelle anhand des Diagramms die quadratische Gleichung der Grenzkostenfunktion.
Zeichne die Regressionslinie in das Diagramm.
Beurteile anhand der Grafik die Güte der Anpassung an die Datenpunkte.
Schätze ab, wo die Regressionslinie ein Minimum hat.
Lies die Koordinaten des Minimums der Grenzkostenfunktion ab.
Wie kann man diese Stelle interpretieren?

ABCD **2.14** Die Gesamtkosten K, die bei der Erzeugung eines Produkts anfallen, sind in Abhängigkeit von der erzeugten Menge untersucht worden.
x ... erzeugte Menge in Mengeneinheiten (ME)
$K(x)$ ... Gesamtkosten in Euro (€) bei einer Produktion von x ME
Die Untersuchung ergab das folgende Streudiagramm:



a) Beschreibe, welche Funktionen in diesem Fall eventuell passend wären.
Wähle die beste Kurve. Begründe deine Auswahl.
b) Erstelle die Gleichung einer Polynomfunktion 3. Grads.
Berechne die mittlere Änderungsrate der Funktion im Intervall [2; 3].
Was fällt dir bei dieser mittleren Änderungsrate auf?

2.2 Korrelation

Die **Korrelation** (Latein: relatio = Beziehung) beschreibt, wie stark die Beziehung zwischen zwei oder mehreren Größen ist.
Allerdings lassen sich aus der Korrelation keine Schlüsse ziehen, ob eine der Größen die andere kausal beeinflusst, das heißt, ob sie diese Größe bzw. ihre Ausprägung verursacht, oder ob eine so genannte Schein-Korrelation vorliegt. So lässt sich der absurde Zusammenhang für das gemeinsame Auftreten von Störchen und Geburten rechnerisch zeigen, wenn zB in einer ländlichen Region viele Störche und viele Geburten vorkommen. Einen kausalen Zusammenhang gibt es hier aber natürlich nicht.
Aber auch das Gegenteil von Schein-Korrelationen kann auftreten: Tatsächlich vorhandene Korrelationen werden oft nicht erkannt.

2.15 Zu den beiden Merkmalen A und B liegen folgende Daten einer Stichprobe vor:

BC

A: x	19	20	33	44	45
B: y	3,38	3,18	2,79	2,18	1,94

a) Zeichne das Streudiagramm.
b) Berechne den Korrelationskoeffizienten nach Bravais und Pearson zwischen den beiden Merkmalen (auf 3 Dezimalstellen gerundet).
Interpretiere die Aussage des Korrelationskoeffizienten.

Carl Pearson (britischer Mathematiker, 1857 – 1936) und Auguste Bravais (französischer Physiker, 1811 – 1863) entwickelten eine dimensionslose Maßzahl r, deren Wert einen Schätzwert für die Richtung und Ausprägung eines **linearen Zusammenhangs** zwischen zwei quantitativen Merkmalen darstellt. Man nennt den Wert r **Pearson-Korrelationskoeffizient** bzw. **Korrelationskoeffizient nach Bravais und Pearson**.
Man berechnet zunächst die arithmetischen Mittelwerte der x- und y-Daten.

$\overline{x} = \frac{1}{5} \cdot (19 + 20 + 33 + 44 + 45) = 32{,}2$

$\overline{y} = \frac{1}{5} \cdot (3{,}38 + 3{,}18 + 2{,}79 + 2{,}18 + 1{,}94) = 2{,}694$

Die Varianzen für die Stichprobe aller x- und y-Daten werden mit $n - 1$ gemittelt, daher gilt:

$s_x^2 = \frac{1}{4} \cdot ((32{,}2 - 19)^2 + (32{,}2 - 20)^2 + \ldots + (32{,}2 - 45)^2) = 156{,}7$

$s_y^2 = \frac{1}{4} \cdot ((2{,}694 - 3{,}38)^2 + (2{,}694 - 3{,}18)^2 + \ldots + (2{,}694 - 1{,}94)^2) = 0{,}38718$

Aus den Varianzen zieht man die Wurzel und erhält die Standardabweichungen s_x und s_y.
Anschließend berechnet man die sogenannte **Kovarianz** s_{xy} der Stichprobe. Sie ist ähnlich definiert wie die Varianz bei einer Variablen, nur berücksichtigt sie nun beide Variablen, also die Abweichungen $(x_i - \overline{x})$ und $(y_i - \overline{y})$.

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

$s_{xy} = \frac{1}{4} \cdot ((32{,}2 - 19) \cdot (2{,}694 - 3{,}38) + (32{,}2 - 20) \cdot (2{,}694 - 3{,}18) + \ldots = -7{,}656$

Damit erhalten wir nun mithilfe der Kovarianz und der Standardabweichungen:

Korrelationskoeffizient nach Bravais und Pearson:

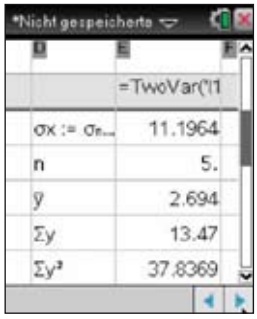
$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

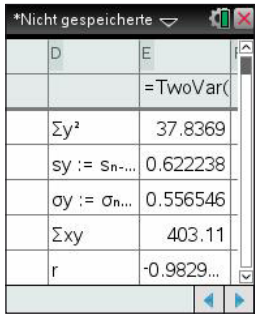
Eingesetzt in die Formel ergibt dies den Wert $r = -0{,}9829\ldots \approx -0{,}983$.

Dein Rechengerät bietet den Korrelationskoeffizienten *r* sowohl unter „Statistik mit 2 Variablen“ als auch bei der Bestimmung der Regressionslinie.

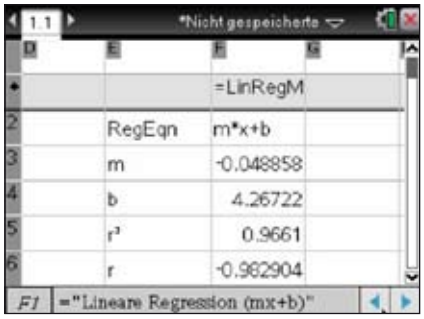
TI-Nspire:

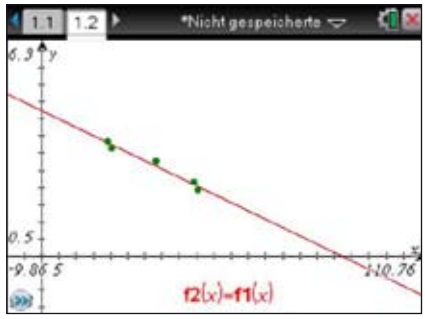
1. Variante der Berechnung: Statistik mit 2 Variablen:
Die Werte in die Listen mit Namen l1 und l2 eingeben,
Menu/4: Statistik/1: Statistische Berechnung/2: Statistik mit 2 Variablen/l1,l2
jeweils „Variablenverweis“





2. Variante: **Menu/4: Statistik/1: Statistische Berechnungen/3: Lineare Regression/l1,l2**
jeweils „Variablenverweis“
Grafik: In f1 wird die Regressionsgerade gespeichert, daher **f2(x) = f1(x)** eingeben, die Gerade wird gezeichnet. Um die Punktewolke zu ergänzen:
Graphs/3: Grafiktyp/4: Streudiagramm/x...l1;y...l2

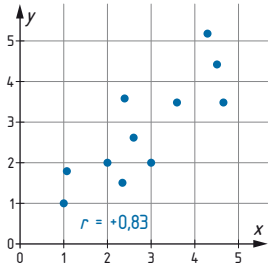


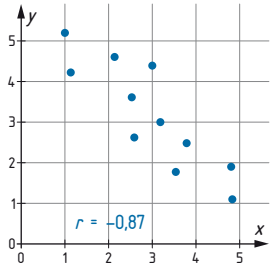


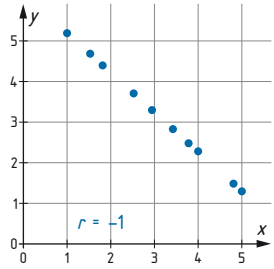
Technologieeinsatz zu diesem Thema für TI 82-84, EXCEL und Geogebra-CAS siehe www.hpt.at (Schulbuch Plus für Schüler/innen)

Wie interpretiert man den Korrelationskoeffizienten?

Der Betrag der Kovarianz kann höchstens gleich groß wie das Produkt der beiden Standardabweichungen sein. Für *r* gilt daher $-1 \leq r \leq 1$. Das **Vorzeichen** wird durch die Kovarianz bestimmt und sagt aus, ob der Trend im Streudiagramm **steigend oder fallend** ist.







Die Bedeutung, die wir aus dem Betrag von r ablesen können:
$|r| = 1$ ... In diesem Fall ist der Betrag der Kovarianz gleich groß wie das Produkt der beiden Standardabweichungen s_x und s_y. Das bedeutet, dass die Punkte des Streudiagramms alle genau auf der Regressionsgeraden liegen.
$|r| < 1$ ... Der Betrag der Kovarianz ist kleiner als das Produkt der Standardabweichungen s_x und s_y, dh die Punkte streuen um die Modellgerade. Je näher der $|r|$ bei 1 liegt, desto stärker ist der lineare Zusammenhang zwischen den x- und den y-Werten ausgebildet. Im Beispiel bedeutet $r \approx -0{,}983$ daher einen starken Zusammenhang der beiden Merkmale A und B.
$|r| = 0$ ... Die Kovarianz ist null, dh, dass die beiden Größen voneinander unabhängig sind. Es besteht kein linearer Zusammenhang.

Der Korrelationskoeffizient nach Bravais und Pearson gilt nur für den linearen Zusammenhang von metrisch skalierten Merkmalen.

Liegen **nichtmetrische Merkmale** vor, so kann ein Korrelationskoeffizient dann berechnet werden, wenn beide Merkmale in eine Rangordnung gebracht werden können: zB Zufriedenheit am Arbeitsplatz (mit der Bewertung 1 ... sehr gut; 2 ... gut; 3 ... befriedigend; 4 ... mäßig; 5 ... schlecht) könnte mit der Höhe der Einkommen, die ebenfalls nach der Größe geordnet wird, korrelieren. Bei gleichen Einkommen werden die Ränge geteilt. Befinden sich beispielsweise auf Platz 2, 3, 4, 5 gleich hohe Einkommen, so erhalten alle 4 die „Platznummer" = Rang 3,5. Der Korrelationskoeffizient wird mit der Pearson-Formel berechnet, man rechnet aber mit den Rangstufen und erhält den **Rangkorrelationskoeffizienten nach Spearman r_{SP}** (Charles Edward Spearman, britischer Psychologe, 1863 – 1945).

Für **nichtlineare Regressionskurven** wird als Maß für die Güte der Korrelation das **Bestimmtheitsmaß R^2** angegeben. Da dieses aber nur angibt, wie viel Prozent der Varianz der y-Werte durch die Regression erfasst werden, ist es kein verlässliches Maß und lässt keine eindeutigen Schlüsse zu. Die Güte muss daher zusätzlich grafisch überprüft werden.
(Beim linearen Zusammenhang ist das Bestimmtheitsmaß gleich $R^2 = r^2$.)

ABC

2.16 X und Y sind Punktebewertungen von 2 Punktrichtern bei einem Sportwettbewerb:
Datenmenge $M = \{(2; 3), (3; 3), (4; 6), (6; 6)\}$.

a) Berechne den Korrelationskoeffizienten nach Bravais und Pearson mit der Formel.

b) Berechne den Korrelationskoeffizienten nach Spearman.
Interpretiere das Ergebnis bezüglich der Stärke des Zusammenhangs der beiden Punktebewertungen X und Y.

Lösung:

a) $\overline{x} = 3{,}75$; $\overline{y} = 4{,}5$; $s_x = 1{,}7078...$; $s_y = 1{,}732...$

$s_{xy} = \frac{1}{3} \cdot (1{,}75 \cdot 1{,}5 + 0{,}75 \cdot 1{,}5 - 0{,}25 \cdot (-1{,}5) - 2{,}25 \cdot (-1{,}5)) = 2{,}5$ ... $\Rightarrow$ $r \approx 0{,}845$

Da Punkte zwischen 1 und 2 nicht gleich einzustufen sind wie die Punkte zwischen 3 und 4 usw. ist hier die Untersuchung der Rangkorrelation eher angemessen!

b) Punkte nach Größe ordnen $\rightarrow$ Platznummer geben, sind 2 Werte gleich, dann werden aufeinanderfolgende Plätze „geteilt".
Punktrichter 2 gibt 2-mal 3 Punkte und 2-mal 6 Punkte, daher gibt es für beide 3 den Platz 1,5 (Rang 1 und 2) und für 6 gibt es 3,5 (Rang 3 und 4 aufgeteilt).

Richter 1	1	2	3	4
Richter 2	1,5	1,5	3,5	3,5

Mit Technologie (Regression): $r_{SP} = 0{,}894...$
Die beiden Punktebewertungen X und Y weisen eine gute Korrelation auf.

ABD **2.17** Stelle die Datenmenge M der beiden Merkmale X und Y grafisch dar.
Schätze anhand der Grafik ungefähr die Größe des Pearson-Korrelationskoeffizienten.
Berechne den Korrelationskoeffizienten mithilfe der Formel.

a) $M = \{(2; 4), (3; 3), (4; 5), (6; 6), (7; 8), (8; 7), (10; 9), (12; 13)\}$

b) $M = \{(1; 10), (2; 9), (3; 12), (5; 15), (6; 14), (7; 15), (9; 18), (11; 23), (14; 27), (15; 30)\}$

ABC **2.18** 10 internationale Konzerne einer Branche hatten in einem bestimmten Jahr die in der Tabelle ausgewiesenen Werbeausgaben und Jahresumsätze.

Werbeausgaben (in Millionen €)	3,15	3,05	1,75	0,78	1,52	1,60	2,12	0,81	0,91	2,12
Jahresumsatz (in Milliarden €)	12,04	11,05	6,45	1,25	5,25	4,65	8,90	1,62	2,24	7,32

Zeichne die Punktewolke.
Ermittle die Regressionsgerade.
Berechne den Korrelationskoeffizienten und interpretiere das Ergebnis.

ABC **2.19** Es wird untersucht, ob die Anzahl der Beschäftigten einer Firma mit dem Umsatz der Firma korrelieren.
Dazu gibt es die folgenden ermittelten Werte.

Anzahl der Beschäftigten	7 232	15 604	23 055	46 840	17 050	18 536	2 600	3 355
Umsatz in Mio. €	10 880	7 420	6 879	4 509	4 373	4 094	3 578	2 927

Zeichne das Streudiagramm zu diesen Daten.
Ermittle die Regressionsgerade.
Berechne den Korrelationskoeffizienten nach Bravais und Pearson.
Interpretiere die Aussage des Koeffizienten.

ABC **2.20** Der Treibstoffverbrauch in Liter/100 km hängt mit der Motorleistung in Kilowatt (kW) zusammen.
Es wurden die folgenden Werte gemessen:

Leistung	55	74	77	85	110	150
Verbrauch	6,4	7,6	6,8	7,9	9,3	10,8

Untersuche mithilfe des Korrelationskoeffizienten nach Bravais und Pearson sowie nach Spearman den Zusammenhang der beiden Größen.

ABCD **2.21** Der Bruttogehalt eines Angestellten soll angeblich mit den Studienjahren (Schulen und tertiäre Bildungseinrichtungen) korrelieren.
Eine entsprechende Untersuchung in einem Betrieb lieferte die folgenden Ergebnisse:

Gehalt in €	1 500	2 000	2 500	4 000	5 000
Anzahl der Studienjahre	9	8	10	16	16

a) Berechne die Kovarianz der beiden Größen.
Interpretiere, was aus dem Wert der Kovarianz geschlossen werden kann.

b) Berechne den Korrelationskoeffizienten nach Bravais und Pearson sowie nach Spearman.
Interpretiere die Aussage der Korrelationskoeffizienten.

c) Argumentiere, welche nichtlineare Regression durch diese Punkte denkbar wäre.
Wie groß wäre das Bestimmtheitsmaß? Was sagt es aus?

Zusammenfassung

Die **Regressionsrechnung** (= Ausgleichsrechnung) erfasst mathematisch den Zusammenhang zwischen zwei messbaren Merkmalen X und Y.

Hat man die Daten in einer **Punktwolke** oder in einem **Streudiagramm** dargestellt, erkennt man in vielen Fällen bereits die Funktion, die den Zusammenhang zwischen den Variablen am besten wiedergibt.
Falls sich die Daten gut linear annähern lassen, so nennen wir die Gerade **Regressionsgerade** oder **Trendlinie** von Y bezüglich X. Die Abweichung $(y_i - y)$ des Funktionswerts y der Trendlinie von einem einzelnen y_i-Wert eines Datenpunkts nennt man **Residuum**.

Mathematisch gesehen muss man sich für die Berechnung der **Parameter *k* und *d*** der Geraden eines objektiven Verfahrens bedienen, zB der **Gauß'schen „Methode der kleinsten Fehlerquadrate"**:
Es ist die Gerade so zu legen, dass die Summe der Quadrate der Abstände aller Punkte von der Geraden möglichst klein ist.

Mit Methoden der Differenzialrechnung wird ein Gleichungssystem gefunden, mit dessen Hilfe man den Anstieg k und den Ordinatenabschnitt d der Geraden berechnen kann.
Die **Parameter** der idealen Modellgeraden, der **Regressionsgeraden (= Trendlinie)** erhält man aus:

$$\text{I: } k \cdot \sum x_i^2 + d \cdot \sum x_i = \sum x_i \cdot y_i$$
$$\text{II: } k \cdot \sum x_i + d \cdot n = \sum y_i$$

Der Punkt $(\overline{x}|\overline{y})$ liegt auf der Regressionsgeraden.

Die **Korrelation** (Latein: relatio = Beziehung) beschreibt, wie stark die Beziehung zwischen zwei oder mehreren Größen, die sich zB durch eine **lineare** Regressionsfunktion darstellen lässt, ist.

Korrelationskoeffizient nach Bravais und Pearson:

$$r = \frac{s_{xy}}{s_x \cdot s_y} \text{ mit } -1 \leq r \leq 1$$

s_{xy} ... **Kovarianz**. Sie wird ähnlich gebildet wie die Varianz in der 1-Variablen-Statistik, nur sind jetzt beide Variablen berücksichtigt. Man berechnet die Produkte der Abweichungen der x_i vom arithmetischen Mittel $\overline{x}$ und der y_i vom arithmetischen Mittel $\overline{y}$, bildet die Summe und dividiert durch $(n - 1)$.

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

***r* gilt nur für den linearen Zusammenhang von metrisch skalierten Merkmalen.**
***r* = −1 und 1 ... alle Punkte liegen auf der Geraden, *r* = 0 ... kein linearer Zusammenhang**

Für nichtlineare Regressionskurven wird als Maß für die Güte der Anpassung die Grafik und das **Bestimmtheitsmaß R^2** angesehen. $R^2 \leq 1$

Liegen **nichtmetrische Merkmale** vor, die mit einem bestimmten anderen monotonen Merkmal korrelieren, dann kann ein Korrelationskoeffizient dann berechnet werden, wenn beide Merkmale in eine Rangordnung gebracht werden können. Der Korrelationskoeffizient wird mit der Pearson'schen Formel berechnet, man verwendet aber die Rangplätze zur Berechnung und bezeichnet ihn als **Rangkorrelationskoeffizient nach Spearman r_{SP}**.

Vermischte Aufgaben zur Vertiefung

ABC **2.22** Für ein bestimmtes Gut wurden die anfallenden Gesamtkosten K in Euro (€) in Abhängigkeit von der erzeugten Menge x in Mengeneinheiten (ME) festgestellt:

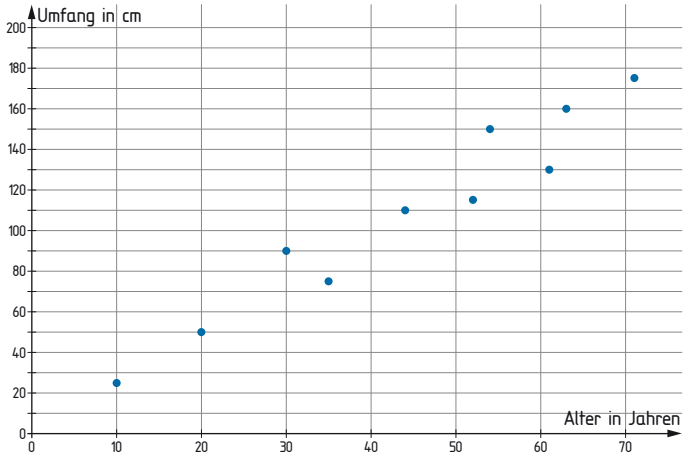
Menge x	100	200	300	400	500
Kosten K	3.450	5.210	7.400	9.180	10.940

a) Zeichne die Punktewolke.
Ermittle eine lineare Trendlinie und zeichne sie in die Punktwolke ein.

b) Schätze mithilfe der Trendlinie, wie hoch die Gesamtkosten bei einer Produktion von 600 ME des Guts wären.

B **2.23** In einer Datenmenge aus zusammengehörenden Paaren ist die Standardabweichung der x-Werte 2,1, die Standardabweichung der y-Werte beträgt 1,9 und der Korrelationskoeffizient −0,98.
Berechne die Kovarianz der Datenmenge.

ABD **2.24** Es wird der Umfang von verschieden alten Bäumen der gleichen Art gemessen.
Die Messpunkte sind in das Streudiagramm eingezeichnet.
Lies die Punkte aus dem Diagramm ab und fertige für die weiteren Aufgaben eine Tabelle an.



a) Skizziere die Regressionsgerade mit Augenmaß in das Streudiagramm.
Lies die Gleichung der Regressionsgeraden ab.

b) Berechne die Regressionsgerade mithilfe der abgelesenen Punktkoordinaten.

c) Berechne den Korrelationskoeffizienten nach Bravais und Pearson.
Interpretiere die Information, die du dem Wert des Korrelationskoeffizienten entnehmen kannst.

d) Berechne die Kovarianz der Punktewolke.

e) Argumentiere, ob und welche nichtlineare Regressionslinie noch in Frage kommen könnte.

2.25 Die Kaufpreise für Gebrauchtwagen desselben Typs wurden in Abhängigkeit vom Alter des Autos erhoben:

ABCD

Alter in Jahren	7	4	10	8	1	5	9	7	3	1
Preis in 100 €	98	125	36	72	213	110	43	92	163	220

a) Erstelle ein Streudiagramm.
Ermittle die Gleichung der Regressionsgeraden.
Erstelle eine Prognose für den Preis eines 6 Jahre alten Autos.

b) Berechne den Korrelationskoeffizienten nach Bravais und Pearson.
Interpretiere die Aussage des Korrelationskoeffizienten.
Erkläre, wie man die Angabe verändern müsste, um den Spearman-Korrelationskoeffizienten berechnen zu können.

2.26 In einer medizinischen Studie wurde der Fettgehalt im Gewebe des menschlichen Körpers in Abhängigkeit vom Alter in einer Stichprobe von männlichen Testpersonen ermittelt.
In der Tabelle wird das Alter in Jahren angegeben, der Fettgehalt in Prozent.

ABCD

Alter	23	39	46	50	54	58	60	62
Fett	9,8	31	27,6	31,5	29,1	33,8	41,1	35

a) Zeichne ein Streudiagramm.
Beurteile, ob das Diagramm auf einen Zusammenhang der beiden Größen hinweist.
Ermittle die Gleichung der Regressionsgeraden.
Berechne den Korrelationskoeffizienten nach Bravais und Pearson.
Interpretiere die Aussage des Korrelationskoeffizienten.

b) Erstelle eine passende kubische Regressionslinie durch die Messpunkte.
Argumentiere, welche der beiden Regressionslinien eine bessere Beschreibung liefert.

2.27 Arbeiter in einer Firma werden nach der Höhe ihres monatlichen Lohns und nach der Zufriedenheit an ihrem Arbeitsplatz befragt.
1 ... sehr zufrieden;
2 ... eher zufrieden;
3 ... eher unzufrieden;
4 ... unzufrieden
Das Ergebnis:

ABCD

Lohn in 100 €	15	17	19	19,5	20	20	21	21,5	21,7	22
Zufriedenheit	3	3	4	3	1	3	2	3	1	2

Ermittle die Rangordnung der monatlichen Löhne.
Berechne den Rangkorrelationskoeffizienten nach Spearman.
Interpretiere die Aussage des Spearman-Koeffizienten.
Beurteile, ob sich die Vermutung bestätigt, dass ein starker Zusammenhang zwischen der Höhe des Lohns und der Zufriedenheit am Arbeitsplatz besteht.

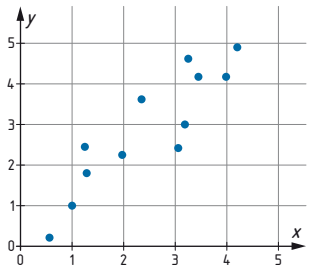
Wissens-Check

Bearbeite die Aufgaben. **Begründe** jeweils deine Auswahl.

		gelöst

1

Ergänze die Textlücken durch Ankreuzen der jeweils richtigen Satzteile so, dass eine korrekte Aussage entsteht.



Das dargestellte Streudiagramm lässt sich durch die Regressionsgerade f mit $f(x) = k \cdot x + d$ annähernd beschreiben. Für diese Gerade gilt _____①_____, weil dann _____②_____ ist.

①		②	
$k = -1$ und $d = 5$	A ☐	die Summe der quadratischen Abweichungen der Punkte von der Geraden f minimal	A ☐
$k = 1$ und $d = 0{,}27$	B ☐	die Summe der Abweichungen der Punkte von der Geraden f minimal	B ☐
$k = 2$ und $d = -0{,}5$	C ☐	die Summe der Abweichungen der Punkte von der Geraden f null	C ☐

2

Die Körpergröße L und die Körpermasse M von Studierenden werden gemessen und die Korrelationskoeffizienten der Regressionsgeraden bestimmt.
r_1 ... Korrelationskoeffizient der Funktion L in Abhängigkeit von M
r_2 ... Korrelationskoeffizient der Funktion M in Abhängigkeit von L
Kreuze an, welche Aussage zutrifft:

A ☐	r_1 ist größer als r_2.
B ☐	r_1 ist kleiner als r_2.
C ☐	r_1 ist genau gleich groß wie r_2.
D ☐	r_1 und r_2 sind gleich groß, haben aber unterschiedliches Vorzeichen.
E ☐	r_1 ist der Kehrwert von r_2.

3

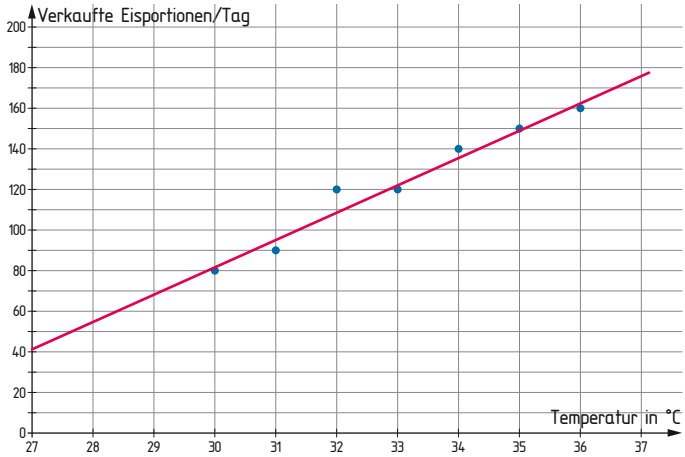
Durch Einnahme einer Droge wurde die Körpertemperatur erhöht.
Es wurde untersucht, welchen Einfluss dies auf den oberen Blutdruckwert des Patienten hatte. Das Ergebnis:

Temperatur in °C	36	37	40
Blutdruck in mm HG	120	122,5	130

Kreuze an, welchem Wert der Korrelationskoeffizient zwischen den beiden Größen am nächsten kommt.

A ☐	B ☐	C ☐	D ☐	E ☐
1	0,8	0,5	0	−0,5

		gelöst
4	Eine Datenmenge aus den Merkmalen X und Y lässt sich mit einer Regressionsgeraden darstellen, deren Steigung 1,5 beträgt. Die arithmetischen Mittelwerte betragen $\bar{x} = 10$ und $\bar{y} = 9$. Kreuze an, wie groß der Achsenabschnitt d der Regressionsgeraden ist. **A** ☐ $d = -6$ — **D** ☐ $d = 8$ **B** ☐ $d = -8$ — **E** ☐ $d = 6$ **C** ☐ $d = 0$	
5	Die Gleichung der quadratischen Regressionslinie lautet $y = 1{,}5x^2 + 3$. Kreuze den richtigen Wert des Residuums für den Messpunkt $(4\vert26)$ an. **A** ☐ −3 — **B** ☐ −2 — **C** ☐ −1 — **D** ☐ 0 — **E** ☐ 1	
6	Der Zusammenhang zwischen der Außentemperatur an bestimmten Sommertagen und der Menge an verkauften Eisportionen pro Tag in einem Eissalon ist in der folgenden Grafik mithilfe von Punkten und der Regressionsgeraden dargestellt. Markiere die Temperatur von 28 °C in der Grafik. Kreuze an, welche gültige Aussage über die Anzahl von Eisportionen bei einer Temperatur von 28 °C gemacht werden kann.	



A ☐	Bei 28° ist keine Aussage über den Verkauf an Eisportionen möglich.
B ☐	Bei 28° sind keine Eisportionen verkauft worden.
C ☐	Bei 28° verläuft die Regressionsgerade nicht durch die Messpunkte. Daher gilt der Wert der Trendlinie hier nicht.
D ☐	Bei 28° liefert die Regressionsgerade keinen gültigen Wert, weil eine Scheinkorrelation vorliegt.
E ☐	Bei 28° kann man ungefähr 55 Eisportionen pro Tag verkaufen.

Lösung:
1) 1 → B; 2 → A 2) C 3) A 4) A 5) C 6) E

CLIL-Review: Regression and correlation

Get in pairs and work on the following tasks (2.E1, 2.E2). You may use your formula collection and an online dictionary.

Please find important vocabulary and other supporting material to these tasks as well as solutions and example answers in the solutions book. For further information have a look at page 2 of this book.

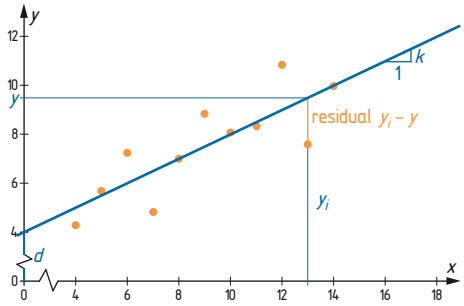
BCD **2.E1**

First read the text about linear regression. Translate all new mathematical terms. Get familiar with the content.

If the n pairs (x_i, y_i) for an experiment are plotted on a graph, then the points are scattered about a straight line. The method of least squares fits the best line to the points by making the sum of the squared deviations $(y_i - y)$ of each point from this line a minimum. This line is called the **line of regression** and is used for predicting y-values by given x-values.

The equation of the line of regression is $y = k \cdot x + d$. It can be calculated by the statistic programme of your calculator or computer.

The differences $(y_i - y)$ between the observed values and the fitted values of the regression line are called **residuals**.



Plotting the residuals to the regression line is a check for the adequacy of the model used. If the plot of residuals shows a small random scatter, this is a good indication that the regression line is an adequate model.

Solve the following task on your own.
Then compare your solution with your partner.
Prepare this topic for a presentation in class.

Carbon dioxide is an important greenhouse gas. The burning of carbon-based fuels has rapidly increased the concentration of CO_2 in the atmosphere.
The maximum value of atmospheric carbon dioxide in parts per million (ppm) for each given year is shown in the table below.

Year x	1995	1998	2001	2004	2007	2010	2013
Carbon dioxide y	318	320	322	325	329	333	335

Plot these points on a scatter diagram.
Calculate the equation of the line of regression.
Draw the line of regression into the scatter diagram.
Estimate the amount of carbon dioxide, when $x = 2\,020$.

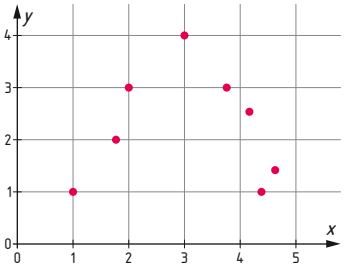
2.E2 First solve the following task about correlation on your own.
Then compare your solution with your partner.
Finally show your findings to your classmates.

BCD

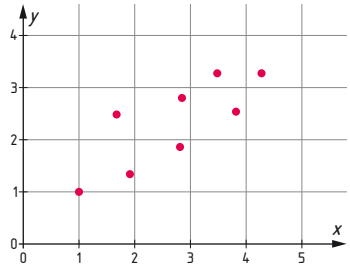
The word "correlation" comes from "co" with the meaning of together and "relation". Correlation shows the dependence between two random variables or two sets of data. If two sets of data are strongly linked together, they have a high correlation. Correlation is positive, when the values increase together, they are negative, when one value decreases as the other increases.

Look at the following seven scatter diagrams.
Match the appropriate special terms and the values of the so called Pearson correlation coefficient (red numbers) in the green box to the diagrams.

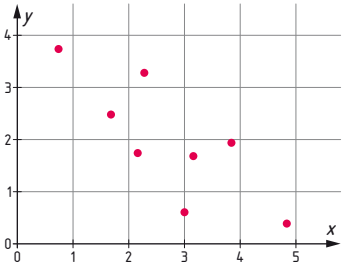
no correlation −0.3
0 high positive correlation high negative correlation
low positive correlation −0.8
0.8 low negative correlation 0.3



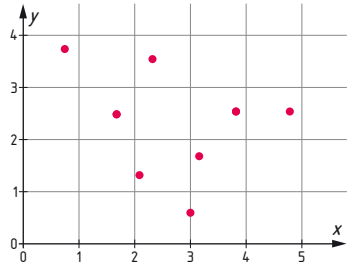
→



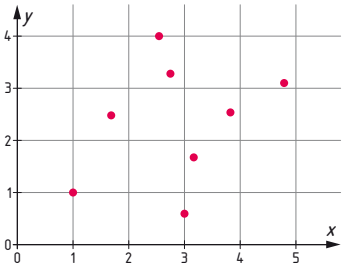
→



→



→



→

Describe how to calculate the Pearson correlation coefficient for the following data:

{(10|1), (7|7), (4|4), (2|6), (1|4), (0|2)}

Interpret the result.