

Lineare Regression Teil 1 – Die Regressionsgerade

In Erinnerung an den 170. Todestag von Carl-Friedrich Gauß

Der TI-30X Prio MathPrint™ verfügt auf Veranlassung der KMK und des IQB¹ nicht über eine Anwendung, die zu zwei Datenmengen X und Y die Gleichung einer Regressionsgeraden automatisch zurückgibt.

Dennoch ist es möglich, mit den wenigen Kenngrößen, die der TI-30X Prio MathPrint™ zu zwei solchen Datenmengen erzeugt, die Gleichung der Regressionsgeraden zu berechnen. Im Folgenden werden die benötigten Formeln bereitgestellt und auf ein einfaches Beispiel angewendet sowie einige Vorschläge zur Behandlung im Unterricht unterbreitet. Die Herleitung der benötigten Gleichungen ist im Anhang zu finden.

1. Beispiel für die lineare Regression mit dem TI-30X Prio MathPrint™

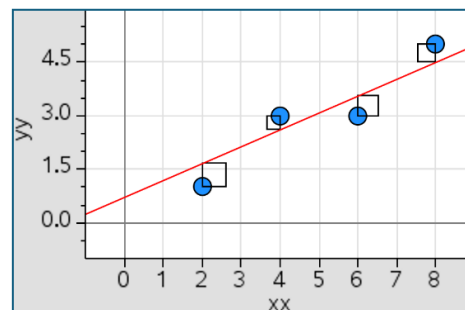
Grundlegende Idee

Gegeben sind n Paare $(x;y)$ reeller Zahlen, die z. B. durch eine statistische Erhebung oder Messung gewonnen wurden. Lässt eine grafische Darstellung einen annähernd linearen Zusammenhang zwischen den unabhängigen x-Werten und den abhängigen y-Werten vermuten, so ist die Gleichung der Regressionsgeraden $\hat{y} = a \cdot x + b$, für die die Summe der Quadrate der Abweichungen der gegebenen y-Werte von den Funktionswerten \hat{y} minimal ist, von Interesse:

$$\sum(\hat{y} - y)^2 \rightarrow \min \quad \Rightarrow \quad \sum(a \cdot x + b - y)^2 \rightarrow \min$$

Geometrisch veranschaulicht soll die Summe der Flächeninhalte der Quadrate für die Regressionsgerade möglichst klein sein (siehe Abbildung²).

Die Seitenlänge jedes der Quadrate entspricht dem Betrag der Differenz $\hat{y} - y$, sein Flächeninhalt ist also $(\hat{y} - y)^2$. Durch das Quadrieren spielen unterschiedliche Vorzeichen der Differenzen keine Rolle mehr und größere Abweichungen fallen stärker, kleinere weniger stark ins Gewicht.



Die Idee der Minimierung der Summe der Fehlerquadrate geht auf Carl Friedrich Gauß zurück. Die Grundlagen der Methode der kleinsten Quadrate hatte Gauß schon 1795, also vor 230 Jahren, im Alter von 18 Jahren entwickelt. Sein Todestag jährt sich am 23.02.2025 zum 170. Male.

¹ <https://www.iqb.hu-berlin.de/abitur/dokumente/mathematik/>

² Die grafischen Darstellungen wurden mit der Software von TI-Nspire erstellt.

Die Formeln für die Regressionskoeffizienten a und b

Für die Regressionskoeffizienten a und b der Regressionsgeraden $\hat{y} = a \cdot x + b$ gelten folgende Formeln:

$$(1) a = \frac{\sum x \cdot y - \frac{\sum x \cdot \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \quad (2) b = \frac{\sum y}{n} - \frac{\sum x}{n} \cdot a$$

Beispiel:

Hier werden bewusst nur vier Zahlenpaare aus ganzen Zahlen verwendet, um notwendige Rechnungen überschaubar und nachvollziehbar zu demonstrieren.

Gegeben sind die Zahlenpaare:

x	2	4	6	8
y	1	3	3	5

2	1	
4	3	
6	3	
8	5	
L3(1)=		

Die gegebenen Daten werden in den Listeneditor des TI-30X Prio MathPrint™ unter `[data]` eingegeben.

Mit `[2nd][data][3]` wird die 2-Variablenstatistik geöffnet:

2-VAR STATS				
xDATA:	L1	L2	L3	
yDATA:	L1	L2	L3	
FREQ:	ONE	L1	L2	L3

2-Var:L1,L2,1	
1:	n=4
2:	Σx=20
3:	Σx²=120

2-Var:L1,L2,1	
4:	Σy=12
5:	Σy²=44
6:	Σxy=72

Die in die Formeln (1) und (2) einzusetzenden Termwerte lassen sich der obigen Taschenrechneranzeige entnehmen und die Rechnung kann in diesem einfachen Fall sogar durch Kopfrechnen erfolgen:

$$n = 4; \sum x = 20; \sum x^2 = 120; \sum y = 12; \sum xy = 72$$

$$\text{Damit wird } a = \frac{\sum x \cdot y - \frac{\sum x \cdot \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{72 - \frac{20 \cdot 12}{4}}{120 - \frac{(20)^2}{4}} = \frac{72 - 60}{20} = \frac{12}{20} = \frac{3}{5} = 0,6 \text{ und}$$

$$b = \frac{\sum y}{n} - \frac{\sum x}{n} \cdot a = \frac{12}{4} - \frac{20}{4} \cdot \frac{3}{5} = 3 - 3 = 0.$$

Die Regressionsgerade hat die Gleichung $\hat{y} = 0,6x$.

Natürlich lassen sich die Werte für a und b auch mit dem Taschenrechner berechnen. Im einfachsten Fall werden die vom WTR angezeigten Termwerte auf Papier notiert und dann in die Formeln (1) und (2) eingesetzt. Die für a und b berechneten Werte können unter den Variablen a bzw. b gespeichert werden. Die Gleichung der Regressionsgeraden wird unter a gespeichert, um ggf. weiter mit ihr arbeiten zu können.

$\frac{72 - \frac{20 \cdot 12}{4}}{120 - \frac{20^2}{4}} \rightarrow a$

$\frac{12}{4} - \frac{20}{4} \cdot a \rightarrow b$

$f(x) = a \cdot x + b$

2. Einige Vorschläge zur Behandlung im Unterricht

a) Summen von Fehlerquadraten verschiedener Ausgleichsgeraden vergleichen

Nach Vorgabe eines konkreten Beispiels, das auch durch eigene Datenerhebungen (Messungen) gewonnen werden kann und deren grafische Darstellung auf einen möglichen linearen Zusammenhang hindeutet, werden die Schülerinnen und Schüler gebeten, Ausgleichsgeraden nach Augenmaß in die „Punktwolke“ einzulegen, deren Gleichung zu ermitteln und die Summe der Fehlerquadrate mit dem TI-30X Prio MathPrint™ zu berechnen. Das kann auch als Wettbewerb gestaltet werden: Wer findet die Ausgleichsgerade mit der kleinsten Summe der Fehlerquadrate?

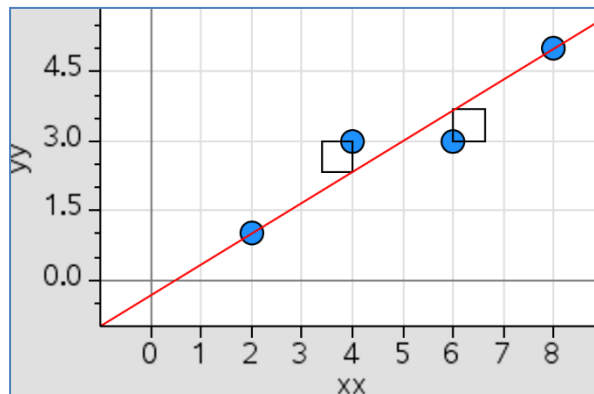
Im obigen Beispiel kann u.a. eine Ausgleichsgerade beispielsweise durch A(2|1) und D(8|5) gelegt werden. Die Berechnung anhand dieses Beispiels könnte auch im Unterrichtsgespräch als Vorlage für die Berechnungen zu den Messwerten durch die Schülerinnen und Schüler dienen.

Zwei-Punkte-Form der Geradengleichung:

$$\frac{\hat{y}-5}{x-8} = \frac{1-5}{2-8} \Rightarrow \hat{y} - 5 = \frac{4}{6} \cdot (x - 8)$$

Diese spezielle Ausgleichsgerade hat die

$$\text{Gleichung } \hat{y} = \frac{2}{3}x - \frac{1}{3}.$$



Berechnung der Summe der Fehlerquadrate durch $\left(\frac{2}{3} \cdot L1 - \frac{1}{3} - L2\right)^2$ gelingt mit dem TI-30X Prio MathPrint™ in der Spalte L3:

LI	LE	DEG	LI
2	1		
4	3		
6	3		
8	5		

L3=(2/3*L1-1/3-L2)²

LI	LE	DEG	LI
2	1		0
4	3		4/9
6	3		4/9
8	5		0

L3(1)=0

Berechnung der Summe dieser Quadrate in der Spalte L3 mit $\boxed{\text{data}}$ OPS $\boxed{4}$ [Sum List]:

DEG
CLR FORMULA OPS
2↑Sort Lg-Sm...
3:Sequence...
4:Sum List...

DEG
SUM LIST
SUM LIST: L1 L2 $\boxed{L3}$
CALC

DEG
SUM LIST
SUM OF LIST=8/9
STORE: \boxed{NO} x y z t a b c d
DONE

Die Summe der Fehlerquadrate ist für diese spezielle Ausgleichsgerade gleich $\frac{8}{9} \approx 0,89$.

Die Ausgleichsgerade muss aber nicht durch die beiden hier gewählten Punkte gehen. Es können auch andere oder gar keine der gegebenen Punkte verwendet werden.

Je nach verwendeter Ausgleichsgerade gibt es unterschiedliche Summen von Fehlerquadraten. Hier kann die Ausgleichsgerade mit der kleinsten Summe ermittelt werden. Die Lehrkraft kann nun „noch eines draufsetzen“ und behaupten, dass es möglich ist, genau eine Ausgleichsgerade zu finden, deren Summe der Fehlerquadrate nicht mehr unterboten werden kann. Daran könnte sich die Bekanntgabe der Formeln und ggf. deren Herleitung anschließen. Die Formeln könnten auch in der Gestalt der Merkhilfen mitgeteilt werden, die im nächsten Abschnitt thematisiert werden.

b) Merkhilfen für die Formeln (1) und (2) erarbeiten und einprägen

$$(1) \quad a = \frac{E(X \cdot Y) - E(X) \cdot E(Y)}{E(X \cdot X) - E(X) \cdot E(X)} = \frac{E(X \cdot Y) - E(X) \cdot E(Y)}{E(X^2) - (E(X))^2}$$

Man beachte die Ähnlichkeit der Strukturen im Zähler und Nenner.

Nachweis:

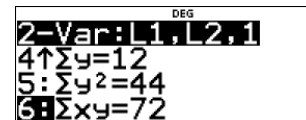
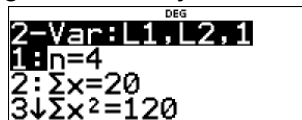
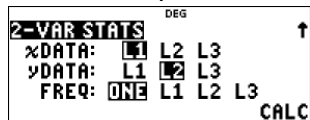
$$a = \frac{E(X \cdot Y) - E(X) \cdot E(Y)}{E(X^2) - (E(X))^2} = \frac{\frac{\sum x \cdot y}{n} - \frac{\sum x}{n} \cdot \frac{\sum y}{n}}{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} \Rightarrow a = \frac{\sum x \cdot y - \frac{\sum x \cdot \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$(2) \quad b = E(Y) - a \cdot E(X) \Rightarrow b = \frac{\sum y}{n} - \frac{\sum x}{n} \cdot a$$

Die Formeln für die Regressionskoeffizienten lassen sich also im Wesentlichen durch verschiedene Erwartungswerte (Mittelwerte) ausdrücken. Mittelwerte sind hier durch die Quotienten der Summen der Zufallsgrößen durch die Anzahl der zu den Zufallsgrößen gehörenden Werte gekennzeichnet.

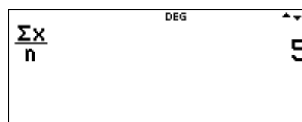
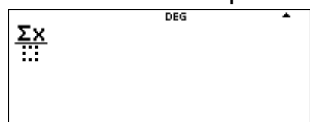
c) Kopieren der Werte aus der 2-Variablenstatistik in den Hauptbildschirm

Mit $\boxed{2\text{nd}} \boxed{\text{data}} \boxed{3}$ wird die 2-Variablenstatistik geöffnet. Die Größen stehen in den Listen L1 und L2. Als Frequenz wird „One“ gewählt, weil jedes Zahlenpaar einzeln erscheint.



Es wird nun gezeigt, wie man $E(X)$ auf diese Weise berechnet:

Um die angezeigten Werte für die benötigten Mittelwerte in den Hauptbildschirm zu bringen, wird die Taste mit der Zahl vor dem Term gedrückt, z. B. für $\sum x$ die Taste $\boxed{2}$. Das Symbol erscheint im Hauptbildschirm.



Dann wird die Taste $\boxed{\frac{\square}{\square}}$ betätigt. Der Term $\sum x$ rückt in den Zähler der Bruchvorlage.

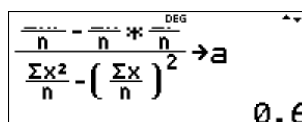
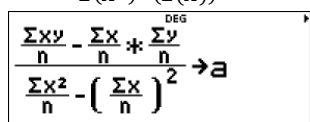
Mit $\boxed{2\text{nd}} \boxed{\text{data}} \boxed{1}$ geht es zurück in die 2-Variablenstatistik.

(Hinweis: Sollte man mit $\boxed{2\text{nd}} \boxed{\text{data}} \boxed{1}$ nicht in den Bildschirm mit den Kenngrößen der 2-Variablenstatistik gelangen, kann dieser mit $\boxed{2\text{nd}} \boxed{\text{data}} \boxed{3}$ wieder aufgerufen werden.)

Nun wird mit $\boxed{1}$ der Term „n“ und damit auch sein Wert in den Nenner des Bruches im Hauptbildschirm exportiert. Mit $\boxed{\text{enter}}$ wird das Ergebnis 5 für diesen Mittelwert angezeigt.

Der folgende Bildschirm zeigt, wie man für das Eingangsbeispiel den Wert für den Regressionskoeffizienten a auf diese Weise berechnet. Wir orientieren uns an der Formel

$$a = \frac{E(X \cdot Y) - E(X) \cdot E(Y)}{E(X^2) - (E(X))^2}$$



d) Regressionsgerade für eine Wertetabelle, in der Zahlenpaare mehr als einmal mit den gleichen Werten auftauchen

Als eine mögliche Vertiefung bietet sich an, mit dem TI-30X Prio MathPrint™ zu untersuchen, wie die Regressionsgeraden für Datenmengen ermittelt werden können, in denen gleiche Zahlenpaare mehrfach vorkommen.

Beispiel für gegebene Zahlenpaare: Wir verwenden die Tabelle von Seite 1, allerdings soll das Zahlenpaar (4; 3) nun zweimal, alle anderen Zahlenpaare genau einmal auftauchen.

x	2	4	6	8
y	1	3	3	5
Anzahl	1	2	1	1

Die Anzahlen der Zahlenpaare werden in der Liste L3 vermerkt.

Nach Aufrufen der 2-Variablenstatistik wird als Frequenz nicht „One“, sondern „L3“ gewählt. Man sieht, dass n nun den Wert 5 hat, alle anderen Kenngrößen haben sich ebenfalls verändert.

2	1	DEG	1
4	3		2
6	3		1
8	5		1
L3(1)=1			

2-VAR STATS			
%DATA:	L1	L2	L3
yDATA:	L1	L2	L3
FREQ:	ONE	L1	L2
CALC			

2-Var:L1,L2,L3	
1:	n=5
2:	Σx=24
3↓:	Σx ² =136

$$\frac{\frac{\sum xy}{n} - \frac{\sum x}{n} \cdot \frac{\sum y}{n}}{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} \rightarrow a$$

$$\frac{\frac{\sum xy}{n} - \frac{\sum x}{n} \cdot \frac{\sum y}{n}}{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} \rightarrow a$$

0.576923077

$$\frac{\sum y}{n} - \frac{\sum x}{n} \cdot a \rightarrow b$$

0.230769231

Für die Berechnung der Regressionskoeffizienten wie oben beschrieben, erhält man nun $a \approx 0,577$ und $b \approx 0,231$.

e) Beispiel: Zusammenhang von Außentemperatur und der Anzahl verkaufter Eiskugeln

Die Besitzerin einer kleinen Eisdiele hat über einige Tage notiert, welche Außentemperatur um 13 Uhr herrschte und wie viele Kugeln Eis sie an diesen Tagen verkauft hatte. Die Ergebnisse sind in der Tabelle zusammengestellt:



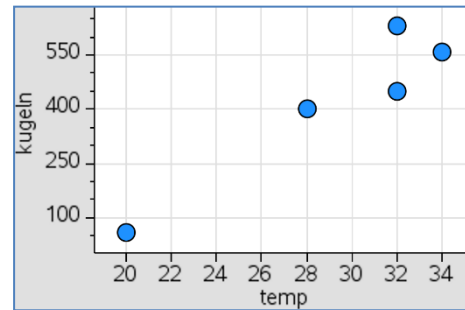
(Quelle: erzeugt mit KI)

Temperatur in °C	28	20	32	34	32
Anzahl verkaufter Kugeln Eis	400	60	630	560	450

Kann man anhand dieser Daten abschätzen, wie viele Kugeln Eis bei einer Temperatur von 30°C oder 40°C verkauft werden könnten?

- Grafische Darstellung der Daten erstellen und interpretieren

Die grafische Darstellung der Messergebnisse lässt einen linearen Zusammenhang zwischen der Temperatur und der Anzahl verkaufter Eiskugeln vermuten. Wenn man diesen Zusammenhang durch eine Gleichung beschreiben kann, dann lassen sich Abschätzungen auch für Temperaturen treffen, die nicht in der Tabelle enthalten sind. Dabei ist zu bedenken, welcher Definitionsbereich für diesen Sachverhalt sinnvoll sein könnte.



- Übertragen der Messergebnisse in den Listeneditor des TI-30X Prio MathPrint™:

28	400
20	60
32	630
34	560

L3(1)=

- Erstellen der Kenngrößen der 2-Variablenstatistik mit 2nd data :

STAT DISTR

1: StatVars

2: 1-VAR STATS

3: 2-VAR STATS

2-VAR STATS

%DATA: L1 L2 L3

yDATA: L1 L2 L3

FREQ: ONE L1 L2 L3

CALC

2-Var: L1, L2, 1

1: n=5

2: $\Sigma x=146$

3: $\Sigma x^2=4388$

2-Var: L1, L2, 1

4: $\Sigma y=2100$

5: $\Sigma y^2=1076600$

6: $\Sigma xy=66000$

- Berechnen der Regressionskoeffizienten a und b:

$$a = \frac{E(X \cdot Y) - E(X) \cdot E(Y)}{E(X^2) - (E(X))^2} \text{ und } b = E(Y) - a \cdot E(X):$$

$$\frac{\frac{\Sigma xy}{5} - \frac{\Sigma x}{5} \cdot \frac{\Sigma y}{5}}{\frac{\Sigma x^2}{5} - \left(\frac{\Sigma x}{5}\right)^2} \rightarrow a$$

$$\frac{\frac{66000}{5} - \frac{146}{5} \cdot \frac{2100}{5}}{\frac{4388}{5} - \left(\frac{146}{5}\right)^2} \rightarrow a$$

37.5

$$\frac{\Sigma y}{5} - a \cdot \frac{\Sigma x}{5} \rightarrow b$$

37.5

-675

- Zeichnen der Regressionsgeraden

$$\hat{y} \approx 37,5x - 675:$$

$$f(x) = a \cdot x + b$$

TABLE SETUP

Start=0

Step=1

Auto %=?

CALC

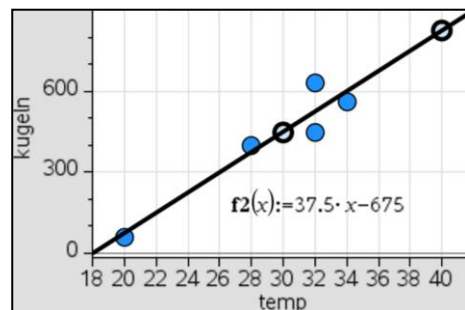
x	f(x)
30	450
40	825

x=

- Interpretation der Lösung

Bei 30°C könnte man mit etwa 450, bei 40°C mit etwa 825 verkauften Kugeln Speiseeis rechnen, wenn man die lineare Regression als Grundlage nimmt. Ob das aber wirklich so eintritt, lässt sich mit dieser Methode nicht eindeutig feststellen, weil die Verkaufszahlen wohl nicht allein von der Außentemperatur abhängen.

Insgesamt macht die Interpretation dieser linearen Regression nur einen Sinn, wenn sommerlich hohe Temperaturen zugrunde gelegt werden, also vielleicht etwa im Intervall von 15°C bis 42°C.



Anhang: Herleitung der Gleichungen (1) und (2)

Gesucht ist die Gleichung der Geraden $\hat{y} = a \cdot x + b$, für die die Summe der Quadrate der Abweichungen der gegebenen y -Werte von den Funktionswerten \hat{y} minimal ist.

Diese Summe lässt sich schreiben als $f(a, b) = \sum_{k=1}^n [(a \cdot x_k + b) - y_k]^2$.

Die x_k sind die gegebenen (oft gemessenen) unabhängigen Werte der Daten.

Die y_k sind die gegebenen (oft gemessenen) abhängigen Werte der Daten.

Im Folgenden schreibe ich statt $\sum_{k=1}^n \dots$ der Einfachheit halber $\sum \dots$.

Die Funktion $f(a, b)$ ist eine Funktion mit zwei unabhängigen Variablen, deren Mathematik im Schulunterricht nicht behandelt wird. Aber es ist vermutlich nicht unmöglich, diese Herleitung nachzuvollziehen, wenn man grundlegende Kenntnisse zur Differentialrechnung besitzt.

Für die Bestimmung der Werte von a und b für ein Minimum der Funktionswerte $f(a, b)$ wird $f(a, b)$ einmal nach a und zum anderen nach b nach den bekannten Ableitungsregeln differenziert. Dabei wird die jeweils andere Variable von a oder b als Konstante betrachtet.

Die Ableitungen werden gleich null gesetzt und das entstehende Gleichungssystem nach a und b gelöst.

So kann zumindest die notwendige Bedingung für die Existenz lokaler Extrema nachvollzogen werden.

Ableitung von $f(a, b) = \sum [(a \cdot x_k + b) - y_k]^2$ nach a (b wird als konstant angesehen):

$$f'_a(a, b) = \sum 2 \cdot [(a \cdot x_k + b) - y_k] \cdot x_k \quad (\text{Kettenregel})$$

$$f'_a(a, b) = 2 \cdot \sum [(a \cdot x_k + b) - y_k] \cdot x_k = 2 \cdot \sum [a \cdot x_k^2 + b \cdot x_k - x_k \cdot y_k]$$

Ableitung von $f(a, b) = \sum [(a \cdot x_k + b) - y_k]^2$ nach b (a wird als konstant angesehen):

$$f'_b(a, b) = \sum 2 \cdot [(a \cdot x_k + b) - y_k] \cdot 1 \quad (\text{Kettenregel})$$

$$f'_b(a, b) = 2 \cdot \sum [(a \cdot x_k + b) - y_k] = 2 \cdot \sum [a \cdot x_k + b - y_k]$$

Beide Ableitungen gleich null setzen:

$$2 \cdot \sum [a \cdot x_k^2 + b \cdot x_k - x_k \cdot y_k] = 0 \quad \text{und} \quad 2 \cdot \sum [a \cdot x_k + b - y_k] = 0$$

Jede Gleichung durch 2 dividieren und Summanden weise schreiben, konstante Faktoren ausklammern:

$$2 \cdot \sum [a \cdot x_k^2 + b \cdot x_k - x_k \cdot y_k] = 0 \quad \text{und} \quad 2 \cdot \sum [a \cdot x_k + b - y_k] = 0$$

$$a \cdot \sum x_k^2 + b \cdot \sum x_k - \sum x_k \cdot y_k = 0 \quad \text{und} \quad a \cdot \sum x_k + \sum b - \sum y_k = 0$$

Die zweite Gleichung wegen $\sum b = n \cdot b$ nach b umstellen:

$$n \cdot b = \sum y_k - a \cdot \sum x_k \implies b = \frac{\sum y_k - a \cdot \sum x_k}{n} \quad (*)$$

Diese Gleichung in die andere einsetzen:

$$a \cdot \sum x_k^2 + \frac{\sum y_k - a \cdot \sum x_k}{n} \cdot \sum x_k - \sum x_k \cdot y_k = 0$$

$$a \cdot \sum x_k^2 + \frac{\sum x_k \cdot \sum y_k - a \cdot (\sum x_k)^2}{n} - \sum x_k \cdot y_k = 0$$

Nach a umstellen:

$$a \cdot \sum x_k^2 + \frac{\sum x_k \cdot \sum y_k}{n} - \frac{a \cdot (\sum x_k)^2}{n} - \sum x_k \cdot y_k = 0$$

$$a \cdot \left(\sum x_k^2 - \frac{(\sum x_k)^2}{n} \right) = \sum x_k \cdot y_k - \frac{\sum x_k \cdot \sum y_k}{n}$$

$$a = \frac{\sum x_k \cdot y_k - \frac{\sum x_k \cdot \sum y_k}{n}}{\sum x_k^2 - \frac{(\sum x_k)^2}{n}}$$

Das ist die auf Seite 2 angegebene Formel (1) für die Berechnung von a.

Setzen wir nun dieses Ergebnis in obige Gleichung (*) ein:

$$b = \frac{\sum y_k - a \cdot \sum x_k}{n} = \frac{\sum y_k}{n} - a \cdot \frac{\sum x_k}{n}$$

Das ist die auf Seite 2 angegebene Formel (2) für die Berechnung von b.

Autor:

Dr. Wilfried Zappe