Regression und Korrelation

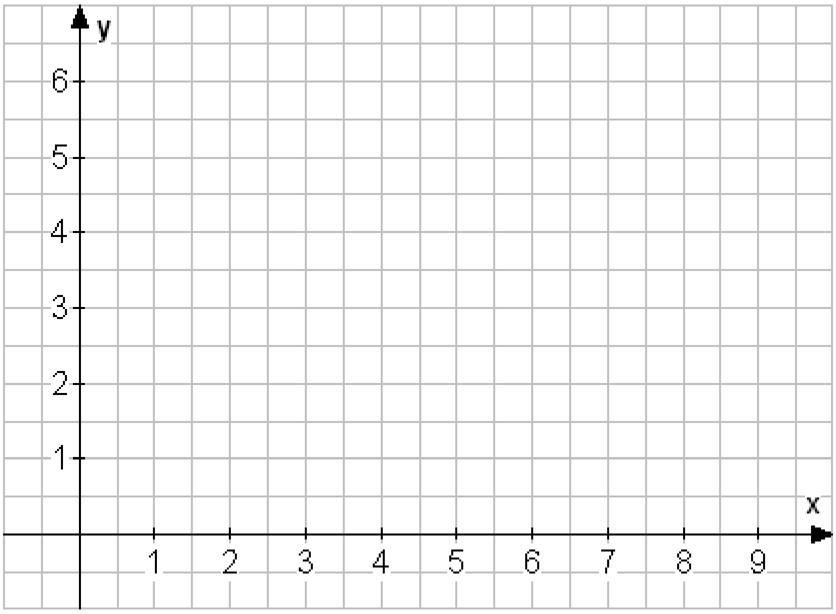
von Ac

<u>Einstiegsaufgabe (lineare Regression): (*)</u>

Durch die 3 Punkte P1(2/1), P2(4/5), P3(9/6) ist eine „Mini-Punktwolke" gegeben .
Gesucht ist diejenige Gerade g, welche in der „Nähe" der Punkte verläuft und die durch die Punkte vorgegebene Tendenz möglichst gut wiedergibt .

<u>Erster Schritt:</u>
Der „Schwerpunkt" der Punkte ist S(5/4) . (Begründung ?)

Es macht Sinn, die gesuchte optimale Gerade g durch diesen Schwerpunkt laufen zu lassen. Somit ist nur noch die Steigung m von g unbekannt .



<u>Aufgabe:</u>
Zeichne die 3 Punkte ein und zusätzlich eine Gerade durch den Punkt S(5/4) mit m=0,5 . Wie weit liegen die Daten von der Gerade g entfernt ??
Überlege dabei: Wie könnte man eine Summe der Entfernung der Punkte von g berechnen ?

<u>Erweiterungen:</u>

1) Löse die obige Aufgabe auch mit den Geradensteigungen m = 1 und m = 0,3 .
Welches Modell ist bisher am besten ?

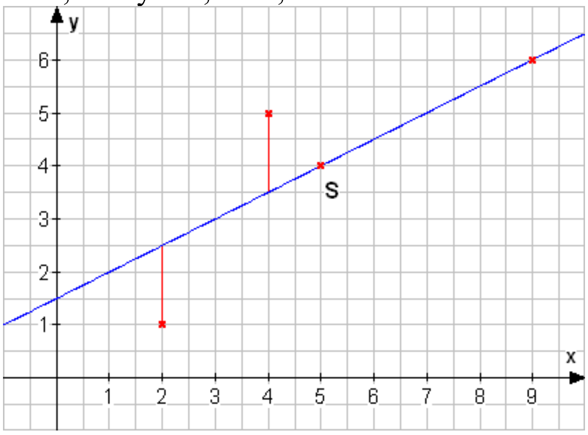
2) Begründe, dass für jede Gerade g durch den Punkt S gilt: $y = m\cdot(x-5) + 4$

<u>Ausblick:</u>
Wie könnte man zu einer „optimalen" Geraden kommen ? Welche Steigung könnte diese haben ?
Wie löst der Grafikrechner TI83 diese Aufgabe ?

(*) Das <u>Beispiel</u> (Punktwolke aus 3 Punkten) stammt von http://www.helmholtz-bi.de/... , wurde jedoch von mir komplett neu bearbeitet .

Lösungshinweise:

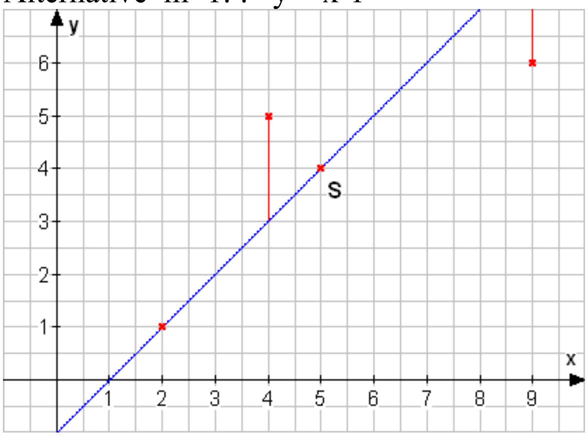
m=0,5: :   y = 0,5x+1,5



Die vertikalen Entfernungen der Punkte von g eignen sich u.a. als Grundlage für die gesuchte Abstandssumme :

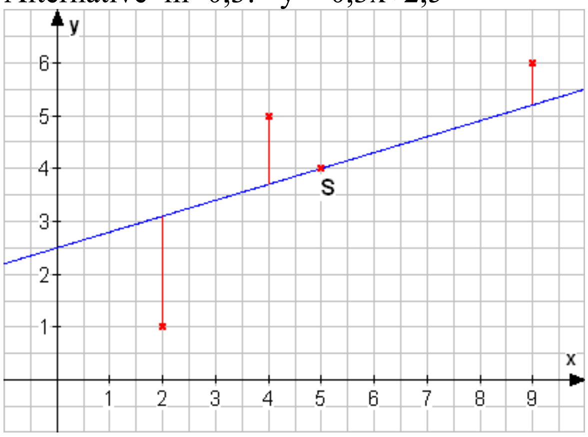
Summe = (2,5-1) + (5-3,5) + 0 = **3**

Zusatzaufgabe 1):
Alternative m=1: :   y = x-1



Summe = 0 + (5-3) + (8-6) = **4**

Alternative m=0,3:   y = 0,3x+2,5



Summe = (3,1-1) + (5-3,7) + (6-5,2) = **4,2**

Das Modell mit m=0,5 ist **bisher** am besten !

Zur Zusatzaufgabe 2):
y = m·(x-5) + 4 ergibt sich aus der Punktsteigungsform bei gegebenem Punkt S(5/4) . Diese Form lautet bekanntlich $m = \frac{y - y1}{x - x1}$, Setzt man nun x1=5 und y1=4, so ergibt sich $m = \frac{y - 4}{x - 5}$, also m(x-5) = y-4 und somit die obige Geradengleichung !

Zu „Ausblick“:
Eine optimale Gerade ergibt sich daraus, dass man das Minimum der Summe der Abstände sucht, also ein Problem der Analysis. Ob der oben berechnete Wert von 3 wirklich das Minimum ist lässt sich nicht voraussagen. Man müsste erst alle Geraden mit y = m·(x-5) + 4   „ausprobieren“ oder (wesentlich besser) ein relatives Minimum der Summenfunktion der Abstände bestimmen (siehe unten) !

Der TI83 besitzt bekanntlich ein Regressionsmenü (über STAT CALC zu erreichen):
Verwendet man LinReg (ax+b), so erhält man   y = 0,61538…x + 0,923…
Offensichtlich geht diese Gerade auch durch S(5/4), jedoch ist die Steigung m eine ganz andere als bei den obigen 3 Beispielen .

Zusatzinformation:

Warum verwendet man nicht die Abstände (Lote) der Punkte von der Geraden zur Bestimmung einer Regressionsgeraden ? Man müsste dann das Minimum der Summe dieser Abstände für alle möglichen Kombinationen von m und b suchen.
Dies ist rechnerisch ziemlich aufwändig, weil man den jeweiligen Lotfußpunkt benötigt !

Durchführung der Berechnung der Lote:
Wenn die ermittelte Gerade die Form $y = mx+b$ besitzt, so hat die Lotgerade durch $P_i(x_i ; y_i)$ die Gleichung $y = -\frac{1}{m}(x - x_i)+y_i$. Schneidet man beide Geraden, so erhält man $mx+b = -\frac{1}{m}(x - x_i) + y_i$ und somit

$$x(m+\frac{1}{m}) = \frac{1}{m} x_i + y_i - b \Rightarrow x(\frac{m^2+1}{m}) = \frac{1}{m} x_i + y_i - b \Rightarrow x_s = \frac{1}{m^2+1} x_i + \frac{m}{m^2+1} (y_i-b)$$

Mit $y_s = mx_s+b$ erhält man den jeweiligen Abstand nach Pythagoras: $d_i = \sqrt{(x_i - x_s)^2 + (y_i - y_s)^2}$

Für obiges Beispiel mit $y = 0{,}5x+1{,}5$; also $m = 0{,}5$ und $b =1{,}5$: $\frac{1}{m^2+1} = 0{,}8$ $\quad \frac{m}{m^2+1} = 0{,}4$

P1(2;1): xs = 1,6-0,2 = 1,4 ys = 0,7+1,5 = 2,2 $d_1 = \sqrt{(2 - 1{,}4)^2 + (1 - 2{,}2)^2} = \sqrt{1{,}8} \approx 1{,}34$

P2(4;5): xs = 3,2+1,4 = 4,6 ys = 2,3+1,5 = 3,8 $d_1 = \sqrt{(4 - 4{,}6)^2 + (5 - 3{,}8)^2} = \sqrt{1{,}8} \approx 1{,}34$

P3(9;6): xs = 7,2+1,8 = 9,0 ys = 4,5+1,5 = 6,0 $d_1 = \sqrt{(9 - 9)^2 + (6 - 6)^2} = 0$

Als Summe der Abstände erhalten wir : $\Sigma \approx 2{,}68$
Dies ist deutlich weniger als bei der Summe der vertikalen Entfernungen (=3) !

Info: Eine Mimimierungsrechnung ist illusorisch, weil der Analysis-Aufwand unvertretbar wäre !

Erläuterung der TI83-Regressionsmethode am Beispiel der Daten P1(2/1), P2(4/5), P3(9/6) :

Der TI83 berechnet nicht das Minimum der Summe der Entfernungen der Datenpunkte von der Geraden, sondern das Minimum der **Summe der Quadrate** der Entfernungen der Datenpunkte von der Geraden.
Der Grund liegt darin, dass man Quadrate besser ableiten kann als Beträge (solche würden entstehen bei der Bildung der bloßen Entfernung, weil die Entfernung stets positiv sein muss !).

Die Vorgehensweise ist die folgende:

1) Berechnen der Mittelwerte $\bar{x}$ und $\bar{y}$ (Koordinaten des Schwerpunktes S) der Daten x_i , y_i :

Für obige Daten gilt $\bar{x} = \frac{2+4+9}{3} = 5$ sowie $\bar{y} = \frac{1+5+6}{3} = 4$

Die Geradengleichung ist dann $y = m(x-5) + 4$.

2) Quadrate der vertikalen Entfernungen der Datenpunkte bezüglich g berechnen:
Für x=2 besitzt g den y-Wert $m(2-5) + 4$, also $-3m+4$. P1 hat den y-Wert 1 . Differenz = $-3m+4-1 = -3m+3$.
Quadriert man diesen Wert, so ergibt sich $\underline{9m^2-18m+9}$.
Ebenso erhält man für P2 die quadrierte Entfernung $(-m+4-5)^2 = \underline{m^2+2m+1}$
und für P3: $(4m+4-6)^2 = \underline{16m^2-16m+4}$.

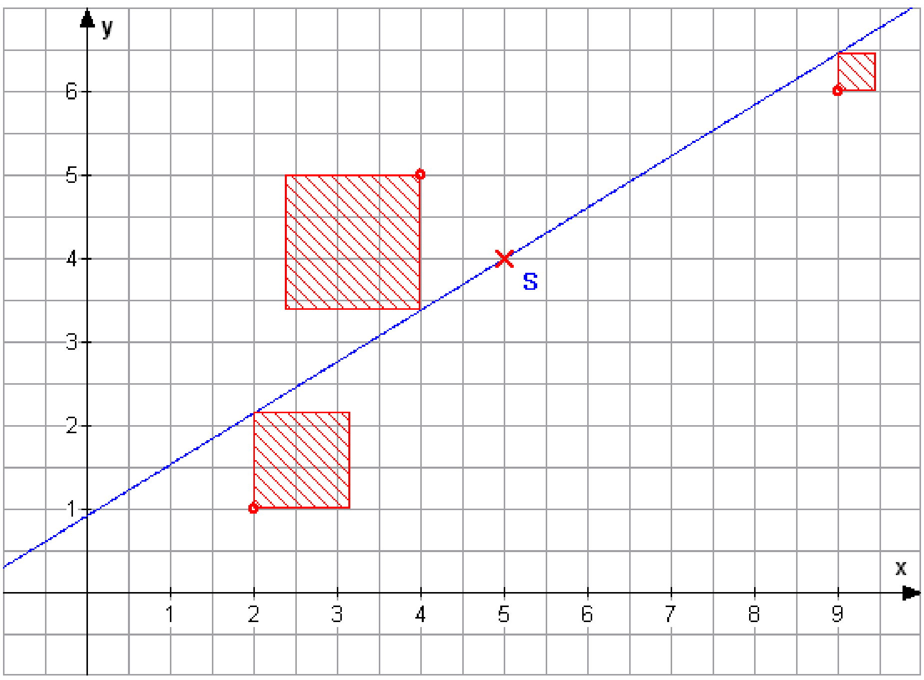
3) Summiert man alle Entfernungsquadrate, so ergibt sich die Summe SEQ = $26m^2-32m+14$.
(SEQ bedeutet: Summe der Entfernungsquadrate)
Man erkennt eine quadratische Funktion (Parabel) für die Variable m .

Das Minimum dieser quadratischen Funktion liegt bei $m = \frac{32}{52} = \frac{8}{13} \approx 0{,}61538$

Mit diesem optimalen m lautet dann die Geradengleichung $y = \frac{8}{13} \cdot (x-5) + 4$ bzw. umgeformt

$y = \frac{8}{13} x + \frac{12}{13}$. Dies entspricht dem Ergebnis des TI(83) !

Das vertikale Entfernungsminimum (SEQ) ist übrigens dann SEQ = $26m^2-32m+14 = \frac{54}{13} \approx \underline{\underline{4{,}15}}$



Betrachtung des **horizontalen** Entfernungsminimums:
(Das Einführungsbeispiel sei wieder zu Grunde gelegt)

Die Geradengleichung (mit Schwerpunkt S) ist dann wieder **y = m(x-5) + 4** . Gesucht ist m .

Benötigt wird die Umkehrfunktion $g^I(y) = x = \frac{y-4}{m} + 5$, weil y gegeben und x (von g) gesucht ist.

Nun werden die Quadrate der **horizontalen** Entfernungen der Datenpunkte bezüglich g berechnet :

P1(2/1): $g = x = \frac{1-4}{m} + 5$, also $-\frac{3}{m} + 5$. Dies ist der x-Wert auf der Geraden bzgl. P1.

Da P1 den x-Wert 2 besitzt, berechnet sich die Differenz der x-Werte zu $-\frac{3}{m} + 5 - 2 = -\frac{3}{m} + 3$.

Quadriert man diesen Wert, so ergibt sich $\boxed{\frac{9}{m^2} - \frac{18}{m} + 9}$ (horiz. Abweichungsquadrat für P1)

P2(4/5): $g = x = \frac{5-4}{m} + 5$, also $\frac{1}{m} + 5$.

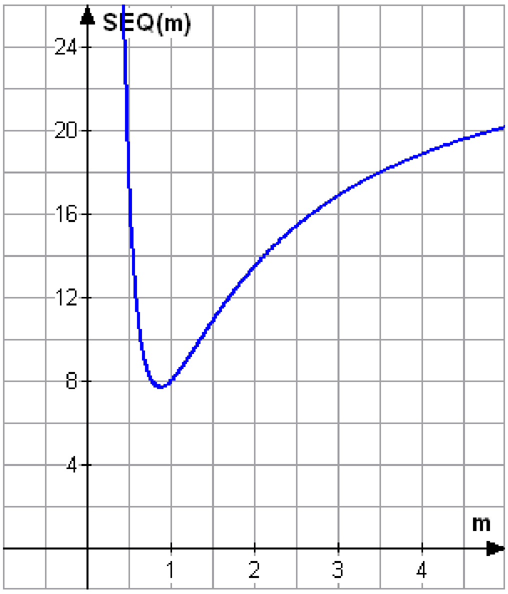
Das horiz. Abweichungsquadrat für P2 berechnet sich zu $(\frac{1}{m} + 5 - 4)^2 = (\frac{1}{m} + 1)^2 = \boxed{\frac{1}{m^2} + \frac{2}{m} + 1}$

P3(9/6): Horiz. Abweichungsquadrat $= (\frac{2}{m} + 5 - 9)^2 = (\frac{2}{m} - 4)^2 = \boxed{\frac{4}{m^2} - \frac{16}{m} + 16}$

Summiert man alle Entfernungsquadrate, so ergibt sich die Summe $SEQ(m) = \boxed{\frac{14}{m^2} - \frac{32}{m} + 26}$.

(SEQ bedeutet: Summe der Entfernungsquadrate)

Das Minimum dieser gebr. rat. Funktion liegt bei
$m = \frac{28}{32} = \frac{7}{8} = 0{,}875$ (siehe Grafik rechts)



Mit diesem optimalen m (horizontal) lautet dann die

Geradengleichung $y = \frac{7}{8} \cdot (x-5) + 4$ bzw. umgeformt

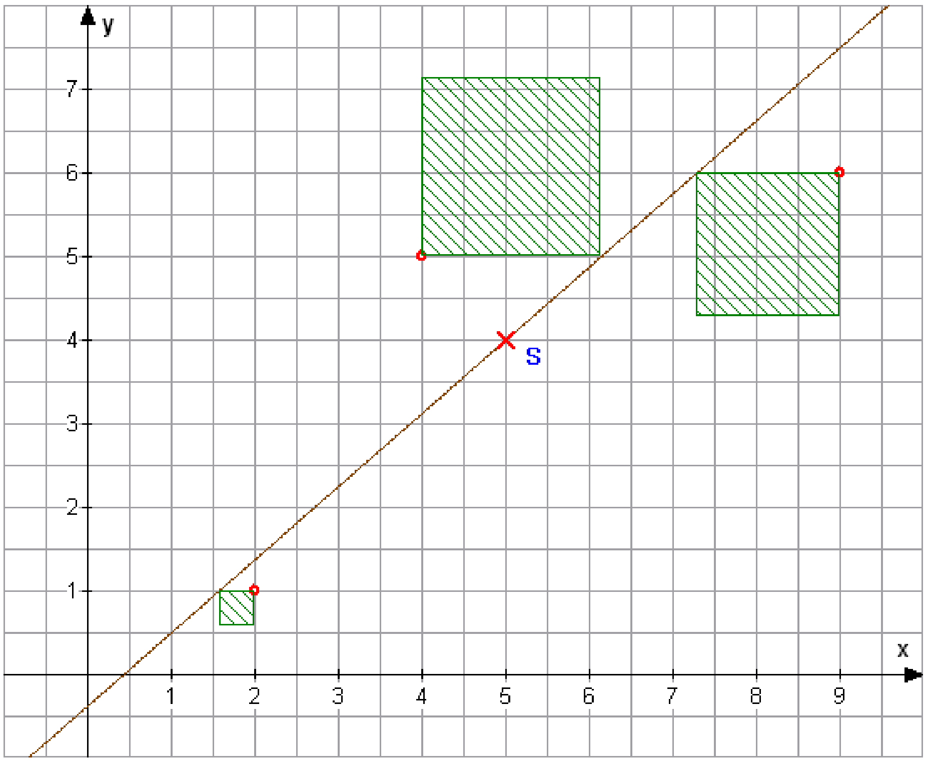
$\boxed{y = \frac{7}{8}x - \frac{3}{8}}$ („horiz. Regr.gerade")

(nach x aufgelöst: $x = \frac{8}{7}y + \frac{3}{7}$)

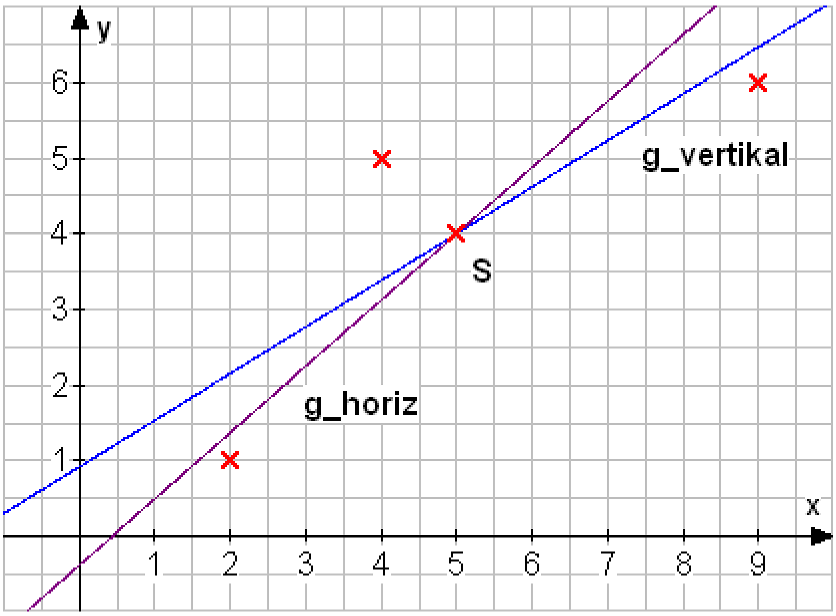
Das **horizontale** Entfernungsminimum ist dann

$$SEQ(\frac{7}{8}) = \frac{14}{(\frac{7}{8})^2} - \frac{32}{\frac{7}{8}} + 26 = \frac{54}{7} \approx 7{,}71 \quad \text{(horiz.)}$$

Grafik für die „horizontale Regressionsgerade" :



Die beiden „optimalen“ Geraden im Vergleich :



Allgemeine Formeln zur Berechnung der beiden Regressionsgeraden

Gegeben seien n Datenpaare x_i und y_i :

1.Fall: **Gesucht sind 2 Regressionsgeraden der Form y = mx+b**

Dann gelten: $\boxed{\bar{x} = \frac{\sum x_i}{n}}$ und $\boxed{\bar{y} = \frac{\sum y_i}{n}}$. Schwerpunktes S der „Punktwolke“ : $S(\bar{x} / \bar{y})$

Regressionsgerade „bezüglich x“: (Betrachtung **vertikaler** Entfernungen – der „Normalfall“)

$m = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ b = $\bar{y} - m \cdot \bar{x}$ Geradengleichung: $y = m \cdot x + b$

$\Rightarrow y = m \cdot x + \bar{y} - m \cdot \bar{x}$ „vertikale“ Geradengleichung

Regressionsgerade „bezüglich y“: (Betrachtung horizontaler Entfernungen)

$m^* = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2}$ $b^* = \bar{x} - m^* \cdot \bar{y}$ Geradengleichung: $x = m^* \cdot y + b^*$ $\Rightarrow y = \frac{1}{m^*} x - \frac{b^*}{m^*}$

$\Rightarrow y = \frac{1}{m^*} x - \frac{\bar{x} - m^* \cdot \bar{y}}{m^*}$ Vereinfachung: $\Rightarrow y = \frac{1}{m^*} x + \bar{y} - \frac{1}{m^*} \bar{x}$ „horizontale“ Geradengleichung

Alternativ könnte man auch die x_i und y_i vertauschen und dann nach der Methode „bezüglich x“ die Geradengleichung der Umkehrfunktion $x = m^* \cdot y + b^*$ bestimmen und diese nach y auflösen.

Korrelationskoeffizient r : $\boxed{r = \sqrt{m \cdot m^*}}$ bzw. $\boxed{r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}}$

Anmerkung: Der Term $\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$ heißt **Kovarianz** !

Die Kovarianz ist ein Maß für den Zusammenhang zwischen den Merkmalen x und y .

2.Fall: **Der Spezialfall b = 0 (y = mx) :**

Der Schwerpunkt ist dann der **Ursprung (0/0)**, weil sich dort die beiden Geraden treffen !

Mit $\bar{x} = 0$ und $\bar{y} = 0$ folgt dann aus den obigen Formeln:

Vertikal: $m = \frac{\sum x_i \cdot y_i}{\sum x_i^2}$ $\Rightarrow y = \frac{\sum x_i \cdot y_i}{\sum x_i^2} \cdot x$ Horizontal: $m^* = \frac{\sum x_i \cdot y_i}{\sum y_i^2}$ $\Rightarrow y = \frac{\sum y_i^2}{\sum x_i \cdot y_i} \cdot x$

Korrelationskoeffizient: $r = \frac{\sum x_i \cdot y_i}{\sqrt{\sum x_i^2 \cdot \sum y_i^2}}$

Tabellarische Rechnung (manuell) für das Einführungsbeispiel:

Erst „vertikal“

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	1	-3	-3	9	9
2	4	5	-1	1	-1	1
3	9	6	4	2	8	16
Summe	15	12			16	26

$\bar{x} = \frac{15}{3} = 5$ $\bar{y} = \frac{12}{3} = 4$ $\mathbf{m} = \frac{16}{26} = \frac{8}{13}$ $b = 4 - \frac{8}{13} \cdot 5 = \frac{12}{13}$ Ergebnis: $y = \frac{8}{13} x + \frac{12}{13}$

Dann „horizontal“ :

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(y_i - \bar{y})^2$
1	2	1	-3	-3	9	9
2	4	5	-1	1	-1	1
3	9	6	4	2	8	4
Summe	15	12			16	14

$\bar{x} = \frac{15}{3} = 5$ $\bar{y} = \frac{12}{3} = 4$ $\mathbf{m^*} = \frac{16}{14} = \frac{8}{7}$ $b^* = 5 - \frac{8}{7} \cdot 4 = \frac{3}{7}$ Zwischenergebnis: $x = \frac{8}{7} y + \frac{3}{7}$

Umgeformt nach y: Endergebnis: $y = \frac{7}{8} x - \frac{3}{8}$

Aufgabe zum Thema Regressionsgerade (Arbeitsblatt):

Die Körpergröße x und das Körpergewicht y von 10 Schülern wurde bestimmt :

Größe x (in cm)	Gewicht y (in kg)
155	47
157	47
159	50
163	55
164	52
167	54
168	58
170	53
172	61
176	65

Wir nehmen einen linearen Zusammenhang zwischen Größe und Körpergewicht an:
Das Koordinatensystem muss (im 1.Quadranten) zunächst skaliert werden:
x-Achse: Achsenkreuz = 140cm und dann 5cm je Skalenzentimeter . (also 140 bis 190)
y-Achse: Achsenkreuz = 40kg und dann 5kg je Skalenzentimeter . (also 40 bis 70)

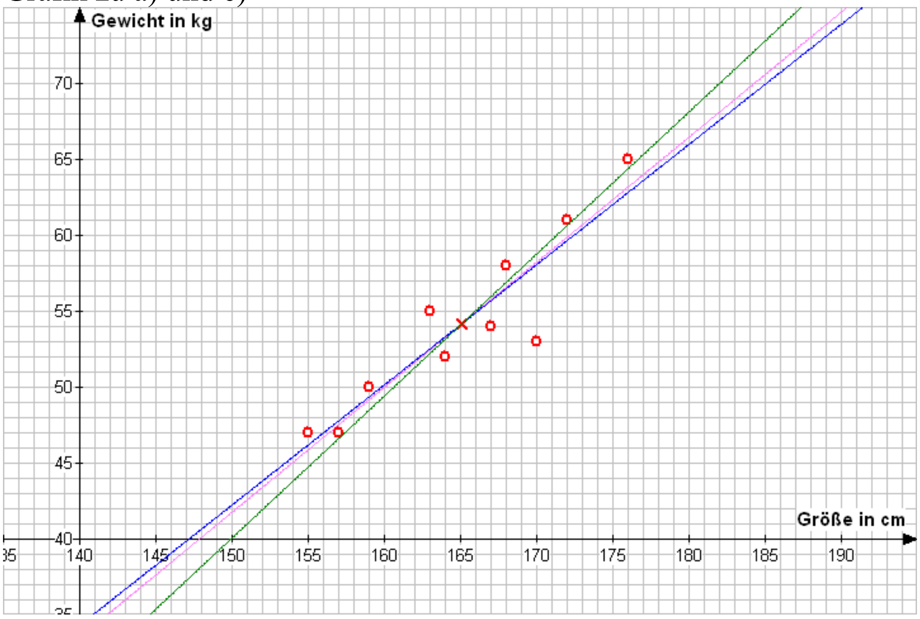
a) Trage die entsprechenden Punkte in das Koordinatensystem ein und bestimme den Schwerpunkt S der „Punktwolke“.
b) Zeichne diejenige Gerade ein, welche durch S und durch P(160/50) verläuft und bestimme ihre Gleichung.
c) Bestimme
- das zur Körpergröße 155cm gehörende Körpergewicht,
- die zum Körpergewicht 60kg gehörende Körpergröße.

d) Bestimme mit dem TI83 eine Regressionsgerade für obige Daten und beantworte nun nochmals c).
e) Bestimme eine Regressionsgerade bzgl. der „horizontalen Minima“ (TI83). Zeichne diese und die in d) bestimmte Regressionsgerade in das Koordinatensystem ein. Welche Interpretation lassen diese beiden Geraden zu ?

Lösungen:

Lösungen:

Grafik zu a) und e)



a) Schwerpunkt S(165,1 / 54,2); siehe $\bar{x}, \bar{y}$ rechts .

b) Gerade durch S und P(160/50):

$$y-50 = \frac{54{,}2 - 50}{165{,}1 - 160}(x-160) \quad \Rightarrow \quad y \approx 0{,}823529x-81{,}76$$

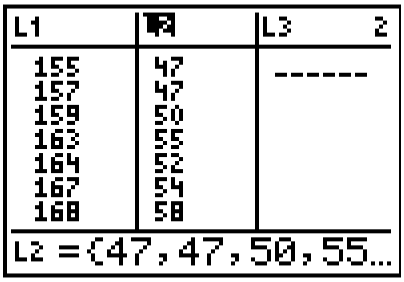
c) Zu x=155(cm) gehört y = 45,9 (kg) und
zu y=60(kg) gehört x = 172,1 (cm) [über Tabelle !]

d) **__vertikal:__** **y = 0,7915x – 76,4725** (siehe LinReg)
Zu x=155(cm) gehört y = 46,2 (kg) und zu
y=60(kg) gehört x = 172,4 (cm) [über Tabelle / Solver !]

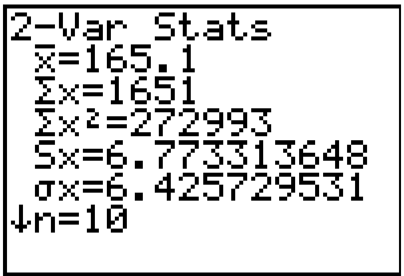
e) **__horizontal:__** zunächst x = 1,06937x+107,14 (siehe LinReg)
Dann auflösen: **y = 0,9351x – 100,189**

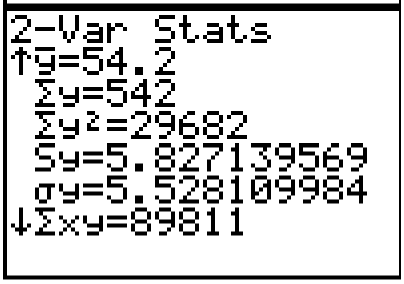
Grafik siehe oben

TI83-Eintragungen:

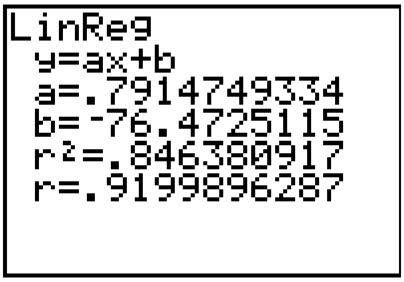


STAT CALC 2-Var Stats ENTER liefert

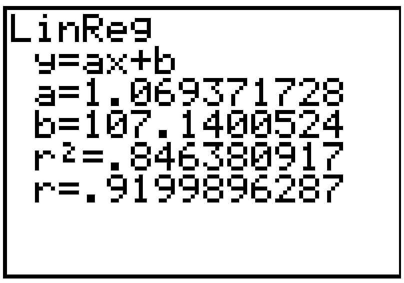




STAT CALC 4 ENTER liefert



STAT CALC 4 L2,L1 ENTER liefert



Interpretation: Sowohl die eine wie auch die andere Gerade könnte „optimal“ sein .

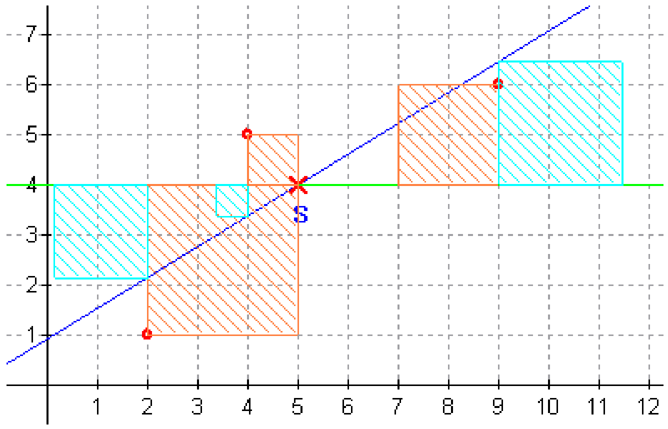
Der Korrelationskoeffizient r

Die Minimierung der Abweichungsquadrate in vertikaler bzw. horizontaler Richtung führt im Allgemeinen zu 2 verschiedenen Ausgleichsgeraden. Diese können nur dann übereinstimmen, wenn alle Punkte der Punktwolke bereits auf einer der Geraden liegen. In diesem Fall besteht zwischen den Merkmalen x und y ein vollständiger linearer Zusammenhang. Je stärker die Punkte einer Punktwolke „auseinanderdriften", desto größer wird der Winkel zwischen den beiden Regressionsgeraden sein; im Extremfall 90° !

Korrelation ist ein Maß dafür, wie stark die Merkmalswerte (Daten) von dem gewählten Regressionsmodell (z.B. Regressionsgerade) abweichen.

Man definiert den Betrag der Korrelation r als Verhältnis (Anteil) der Standardabweichung, die durch das Regressionsmodell festgelegt ist zur Standardabweichung, die durch die Daten gegeben ist,

kurz $\boxed{|r| = \frac{s_{Modell}}{s_{Daten}}}$



Die Grafik erläutert beispielhaft für 3 Punkte P1(2/1) , P2(4/5) , P3(9/6) , was unter s_{Modell} und s_{Daten} zu verstehen ist .

Dabei werden zunächst die Quadrate der vertikalen Abweichungen der Daten vom Schwerpunkt S betrachtet (rot) und ihre Flächeninhalte aufsummiert. Teilt man diese Summe durch sie Anzahl der Daten (hier = 3) und zieht dann die Wurzel, so ergibt sich die eine der beiden Standardabweichungen s_{Daten} !

Entsprechend geht man vor bei der Bestimmung von s_{Modell} .
Dort werden die Quadrate der vertikalen Abweichungen des Modells „Gerade" vom Schwerpunkt S betrachtet (blau) .

Hier gelten: $y = \frac{8}{13}x + \frac{12}{13}$ (Regressionsgerade)

$s_{Daten} = \sqrt{\frac{14}{3}}$ $\qquad s_{Modell} = \sqrt{\frac{128}{13 \cdot 3}}$

Begründung der beiden Ergebnisse:

Für die roten Quadrate kann man die Summe der Flächeninhalte am Grafikraster ablesen (14 Einheiten).

Für die blauen Quadrate rechnet man so:

Summe der Flächeninhalte = $(\frac{8}{13} \cdot 2 + \frac{12}{13} - 4)^2 + (\frac{8}{13} \cdot 4 + \frac{12}{13} - 4)^2 + (\frac{8}{13} \cdot 9 + \frac{12}{13} - 4)^2 = \frac{128}{13} \approx 9{,}85$

Teilt man jetzt s_{Modell} durch s_{Daten} , so erhält man $r = \sqrt{\frac{64}{91}} \approx \underline{\underline{0{,}8386}}$

Allgemeine Formulierungen zur Berechnung des Korrelationskoeffizienten:

Die Abweichung des Modells (hier: Gerade) vom Schwerpunkt S ist : $a \cdot x_i + b - \bar{y}$

Die Abweichung der Daten P_i vom Schwerpunkt S ist : $y_i - \bar{y}$

Berechnet man die Summe der Quadrate dieser Abweichungen für alle Daten, so erhält man:

$\text{Quadratsumme}_{\text{Modellabweichugen}} = \sum (ax_i + b - \bar{y})^2$ $\text{Quadratsumme}_{\text{Datenabweichugen}} = \sum (y_i - \bar{y})^2$

Für die Standardabweichungen s erhält man dann:

$$s_{\text{Modellabweichugen}} = \sqrt{\frac{\sum (ax_i + b - \bar{y})^2}{n}} \qquad s_{\text{Datenabweichugen}} = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}}$$

Da |r| als Quotient der beiden Größen definiert ist ergibt sich vereinfacht: $|r| = \sqrt{\frac{\sum (ax_i + b - \bar{y})^2}{\sum (y_i - \bar{y})^2}}$

Anmerkung: **Für fallende Geraden als Modell wird r als negativ definiert, sonst positiv !!**

> Bei linearer Regression liegt r im Bereich [-1 ; 1] (s. oben: r kann auch negativ sein).
>
> Je näher der Betrag von r dem Wert 1 kommt, um so besser ist das Modell .

Bezug zu den beiden möglichen Regressionsgeraden:

Hat man die beiden Steigungen m_v (=m) und m_h (= m*) bzgl. der vertikalen und der horizontalen Entfernungsmaxima bestimmt, so lässt sich r auch auf andere Weise ganz einfach ermitteln.
r ist dann gleich dem geometrischen Mittel der beiden Zahlen m_v und m_h , also $r = \sqrt{m \cdot m^*}$.
(Achtung: m_h ist der Steigungswert m* der **Umkehrfunktion** mit $x = m^* \cdot y + b^*$!!)

Für das Einführungsbeispiel galt: $m_v = \frac{8}{13}$ $m^* = m_h = \frac{8}{7}$.

Dann ist $r = \sqrt{\frac{8}{13} \cdot \frac{8}{7}} = \sqrt{\frac{64}{91}} \approx 0{,}8386$

Bei **linearer** Regression gilt dann: $|r| = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$

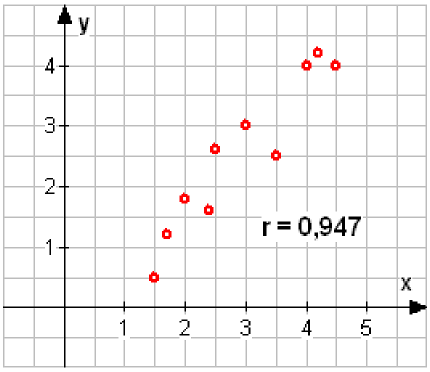
Für das Einführungsbeispiel mit $x_1=2; x_2=4; x_3=9$ $y_1=1; y_2=5; y_3=6$

ist $$r = \frac{(2-5)(1-4) + (4-5)(5-4) + (9-5)(6-4)}{\sqrt{[(2-5)^2 + (4-5)^2 + (9-5)^2] \cdot [(1-4)^2 + (5-4)^2 + (6-4)^2]}}$$

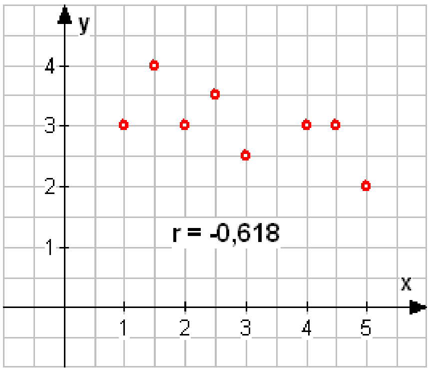
$$= \frac{9 - 1 + 8}{\sqrt{[9+1+16] \cdot [9+1+4]}} = \frac{16}{\sqrt{26 \cdot 14}} = \frac{16}{\sqrt{364}} \approx 0{,}8386$$ (vgl. Ergebnis oben)

Bewertung des Zusammenhangs (der Korrelation) zwischen 2 Größen:

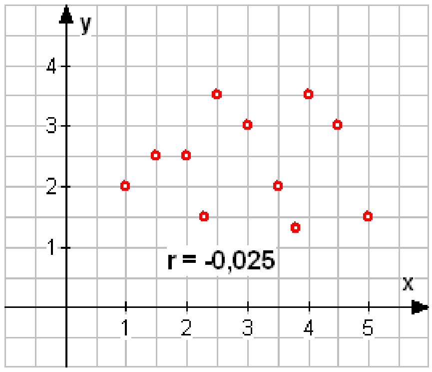
$-1 \le r < -0{,}8$ oder $0{,}8 < r \le 1$
starke Korrelation



$-0{,}8 \le r < -0{,}3$ oder $0{,}3 < r \le 0{,}8$
schwache Korrelation



$-0{,}3 \le r \le 0$ oder $0 \le r \le 0{,}3$
keine Korrelation



Die Grafiken sind lediglich willkürlich gewählte Beispiele !
Anmerkung: In der Literatur findet man auch noch andere Klassifizierungen der Korrelation !

Der Korrelationskoeffizient beim TI83:

r ist beim TI83 nicht ohne weiteres sichtbar, kann aber „nachgerüstet“ werden.
Erst **CATALOG** (2nd 0) wählen, dann dort zu **DiagnosticOn** und **ENTER ENTER** drücken .
Von jetzt ab wird r immer zusammen mit der Regressionsfunktion angezeigt.