Kurvenanpassung durch Regression (1)

Ac 2017-2020

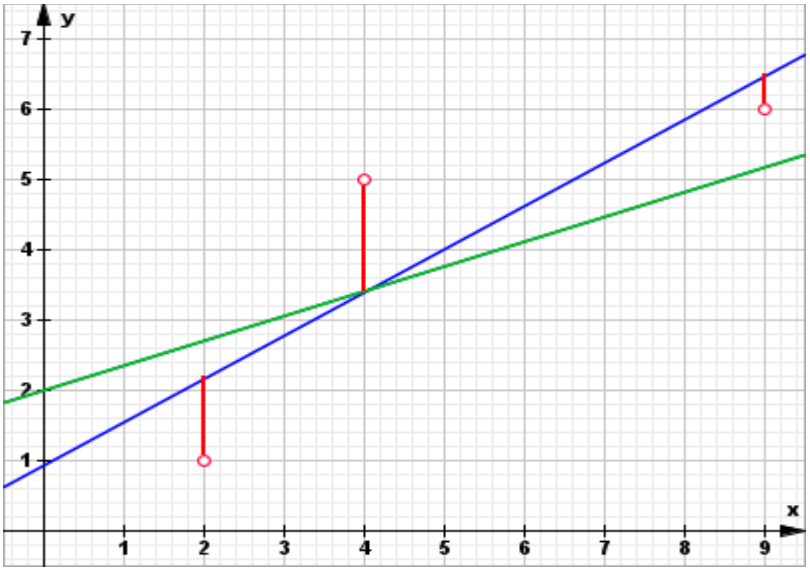
- Einführung / Lineare Regression -

Problemstellung:

Es sind n Wertepaare (Messwerte bzw. Punkte) $\mathbf{P_i\ (x_i\ ;\ y_i)}$ gegeben.
Gesucht ist eine Funktion f, die sich den Punkten P_i möglichst gut anpasst (Regression).

Lösungsidee („Methode der kleinsten Quadrate“):

Man wählt bzw. sucht eine Funktion f, bei der die Summe der Quadrate aller senkrechten Entfernungen der P_i von den Funktionwerten $f(x_i)$ minimal ist.



Beispiel: $P_1(2;1)$ $P_2(4;5)$ $P_3(9;6)$

Hier wurden 2 verschiedene lineare Funktionstypen zur Approximation gewählt.

Bei der blauen Geraden beträgt die Summe der senkrechten Entfernungsquadrate ca. 3,23.

Bei der grünen (Quadrate nicht eingezeichnet) sind es ca. 4,15.

Daher ist die blaue Gerade besser zur Approximation geeignet.

Ob die blaue Gerade aber die beste Approximation für die Daten ist, das muss erst rechnerisch geprüft werden.

Herleitung der Formeln (zunächst die allgemeine Formel für beliebige Funktionen f):

Bezeichnet man die o.g. Summe mit SQ, so gilt:

$$\mathrm{SQ} = \sum_{i=1}^{n} (y_i - f(x_i))^2 \quad \text{soll minimal werden !}$$

Weil SQ in der Regel von Parametern a, b, c, ... abhängt, müssen die partiellen Ableitungen bezüglich der Parameter gebildet und Null gesetzt werden (notw. Kriterium für Extrema).

$$\frac{d\mathrm{SQ}}{d\mathrm{a}} = 0 \ \wedge\ \frac{d\mathrm{SQ}}{d\mathrm{b}} = 0 \ \wedge\ \frac{d\mathrm{SQ}}{d\mathrm{c}} = 0 \ \wedge\ \ldots$$

Dies führt auf ein Gleichungssystem (GS) zur Lösung der Parameter a, b, c, ... , welches nicht notwendig linear sein muss !

Im folgenden werden die Regressionsformeln für verschiedene Funktionsklassen hergeleitet:

Lineare Regression: f(x) = m·x + b

Die oben definierte Summe der Fehler-Quadrate ist : $\boxed{SQ = \sum_{i=1}^{n} (y_i - (mx_i + b))^2}$

Für die partiellen Ableitungen ergibt sich:

$$\frac{dSQ}{dm} = \sum_{i=1}^{n} 2(y_i - (mx_i + b)) \cdot (-x_i) \quad \wedge \quad \frac{dSQ}{db} = \sum_{i=1}^{n} 2(y_i - (mx_i + b)) \cdot (-1)$$

Setzt man diese gleich 0, so erhält man ein lineares Gleichungssystem (LGS):

$$\left| \begin{array}{c} \sum_{i=1}^{n} (y_i - (mx_i + b)) \cdot (-x_i) = 0 \\ \sum_{i=1}^{n} (y_i - (mx_i + b)) = 0 \end{array} \right| \Leftrightarrow \left| \begin{array}{c} \sum_{i=1}^{n} (-x_i y_i + mx_i^2 + bx_i) = 0 \\ \sum_{i=1}^{n} (-y_i + mx_i + b) = 0 \end{array} \right| \Leftrightarrow \left| \begin{array}{c} m \cdot \sum_{i=1}^{n} x_i^2 + b \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ m \cdot \sum_{i=1}^{n} x_i + b \cdot \sum_{i=1}^{n} 1 = \sum_{i=1}^{n} y_i \end{array} \right|$$

Unter Berücksichtigung von $\sum_{i=1}^{n} 1 = n$ ergibt sich folgende Lösung

$$m = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \quad \wedge \quad b = \frac{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}, \quad falls\ der\ Nenner \neq 0$$

Dies lässt sich noch vereinfachen, wenn man die arithmetischen Mittelwerte der x-Daten ($\bar{x}$) sowie der y-Daten ($\bar{y}$) verwendet:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Lösungsformel (Lineare Regression):

$$\boxed{y = m \cdot x + b \quad mit \quad m = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \wedge \quad b = \bar{y} - m \cdot \bar{x}}$$

Anmerkungen:

- Der Nenner von m ist das n-fache der sog. „Varianz“ der x-Daten .
- Der Zähler von m ist das n-fache der sog. „Kovarianz“ der Daten .

Diese Formel lässt sich noch so umformen, dass die Rechenschritte vereinfacht werden:

$$m = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i\cdot y_i-\bar{x}\cdot y_i-x_i\cdot\bar{y}+\bar{x}\cdot\bar{y})}{\sum_{i=1}^{n}(x_i^{\,2}+\bar{x}^2-2\cdot x_i\cdot\bar{x})} =$$

$$\frac{\sum_{i=1}^{n}x_i\cdot y_i-\bar{x}\cdot\sum_{i=1}^{n}y_i-\bar{y}\cdot\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}\cdot\bar{y}}{\sum_{i=1}^{n}x_i^{\,2}+\sum_{i=1}^{n}\bar{x}^2-2\cdot\sum_{i=1}^{n}x_i\cdot\bar{x}} \quad \begin{matrix}\sum_{i=1}^{n}y_i=n\cdot\bar{y}\\ =\\ \sum_{i=1}^{n}\bar{x}\cdot\bar{y}=n\cdot\bar{x}\cdot\bar{y}\end{matrix} \quad \frac{\sum_{i=1}^{n}x_i\cdot y_i-n\cdot\bar{x}\cdot\bar{y}-n\cdot\bar{x}\cdot\bar{y}+n\cdot\bar{x}\cdot\bar{y}}{\sum_{i=1}^{n}x_i^{\,2}+n\cdot\bar{x}^2-2\cdot\bar{x}\cdot\sum_{i=1}^{n}x_i} = \quad \sum_{i=1}^{n}x_i=n\cdot\bar{x}$$

$$\frac{\sum_{i=1}^{n}x_i\cdot y_i-n\cdot\bar{x}\cdot\bar{y}}{\sum_{i=1}^{n}x_i^{\,2}+n\cdot\bar{x}^2-2\cdot n\cdot\bar{x}\cdot\bar{x}} = \frac{\sum_{i=1}^{n}x_i\cdot y_i-n\cdot\bar{x}\cdot\bar{y}}{\sum_{i=1}^{n}x_i^{\,2}+n\cdot\bar{x}^2-2\cdot n\cdot\bar{x}^2}$$

Ergebnisformel (Lineare Regression) :

$$y = m\cdot x + b \quad \text{mit} \quad m = \frac{(\sum_{i=1}^{n}x_i\cdot y_i)-n\cdot\bar{x}\cdot\bar{y}}{(\sum_{i=1}^{n}x_i^{\,2})-n\cdot\bar{x}^2} \quad \text{und} \quad b=\bar{y}-m\cdot\bar{x}$$

Beispielrechnung für P1(2;1) P2(4;5) P3(9;6) :

$\bar{x}$ = (2+4+9) / 3 = 5

$\bar{y}$ = (1+5+6) / 3 = 4

Also Schwerpunkt **S(5 ; 4)**

∑ xi·yi = 2 + 20 + 54 = 76

∑ xi² = 4 + 16 + 81 = 101

m = (76 – 3·5·4) / (101 - 3·5²) = 16 / 26 = **8 / 13**

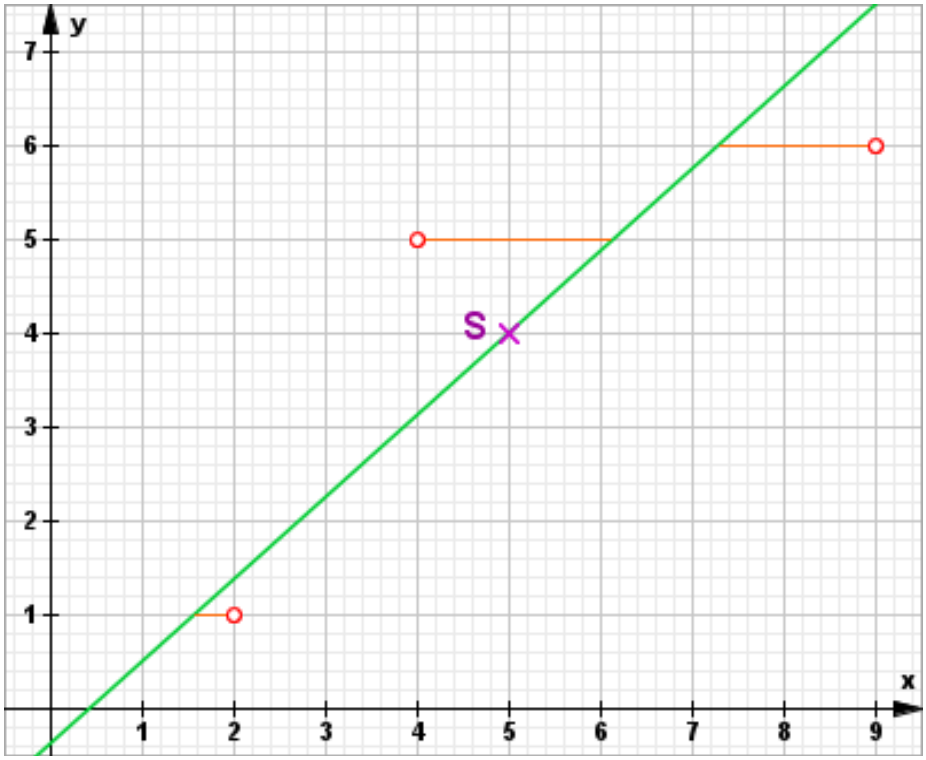
b = 4 – 8 / 13 · 5 = 52 / 13 – 40 / 13 = **12 / 13**

Die gesuchte lineare Funktion ist: **y = 8 / 13 · x + 12 / 13**

bzw. **y ≈ 0,615384615 ·x + 0,923076923**

Eine zweite Gerade für die lineare Regression:

Das Problem der Minimumsuche kann auch auf die Summe der Quadrate aller **waagerechten Entfernungen** der Punkte P_i von den Funktionwerten $f(x_i)$ bezogen werden (siehe Grafik) .



Hierzu muss die horizontale Gerade $y = y_i$ mit $y = \mathbf{f(x) = m_2 \cdot x + b_2}$ geschnitten werden, also $y_i = m_2 \cdot x + b_2$. Die Schnittstelle ist dann $x = (y_i - b_2) / m_2$.
Die horizontale Entfernung von P_i zur Regressionsgerade ist daher $x_i - x = x_i - (y_i - b_2) / m_2$.
Die Summe der Quadrate aller dieser Differenzen muss minimal sein, d.h.

$$SQ_h = \sum_{i=1}^{n} \left(x_i - \frac{1}{m_2} \cdot (y_i - b_2)\right)^2 \quad \text{muss minimal sein !}$$

Diese Formel zu SQ_h ist analog zur Formel für SQ bei den „vertikalen“ Geraden .
Hier ist lediglich x_i mit y_i vertauscht und die Steigung ist $1/m_2$.

Analog erhält man: $\frac{1}{m_2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2}$ und somit als Ergebnis:

$$y = m_2 \cdot x + b_2 \text{ mit } m_2 = \frac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})(y_i - \bar{y})} ; \; b_2 = \bar{y} - m_2 \cdot \bar{x}$$ Gerade für waagerechte Abstände

Vereinfachung (analog zur „vertikalen“ Gerade) :

Ergebnis ("horizontale" Gerade) :

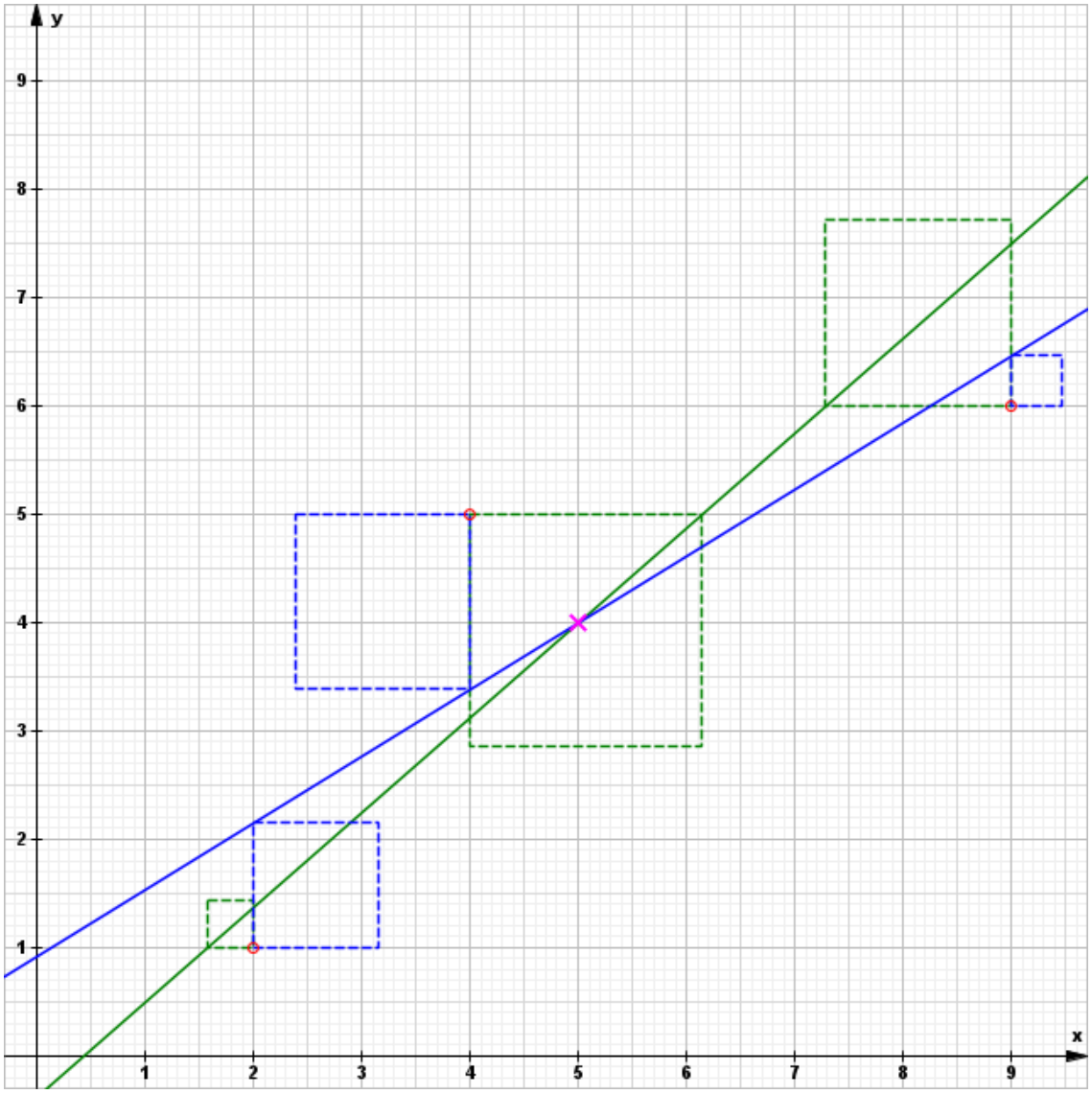
$$y = m_2 \cdot x + b_2 \quad \text{mit} \quad m_2 = \frac{(\sum_{i=1}^{n} y_i^2) - n \cdot \bar{y}^2}{(\sum_{i=1}^{n} x_i \cdot y_i) - n \cdot \bar{x} \cdot \bar{y}} \quad \text{und} \quad b_2 = \bar{y} - m_2 \cdot \bar{x}$$

Für das Beispiel von oben erhält man:

$m_2 = (62 - 3 \cdot 4^2) / (76 - 3 \cdot 5 \cdot 4) = 14 / 16 = 7/8$ $\quad\quad$ $b_2 = 4 - 7/8 \cdot 5 = -3/8$

Dann ist g_horizontal: **$y = 7/8 \cdot x - 3/8 = 0{,}875 \cdot x - 0{,}375$**

In der folgenden Grafik ist der Unterschied in der Fehlerquadratsumme gut zu erkennen:



blau: „vertikale“ Gerade

grün: „horizontale“ Gerade

Wir berechnen noch die Summe der Fehlerquadrate, um den Unterschied auch zahlenmäßig herausstellen zu können:

g_vertikal: y = 8 / 13 · x + 12 / 13

$$SQ=\sum_{i=1}^{n}(y_i-m(x_i+b))^2=\sum_{i=1}^{n}(y_i-\frac{8}{13}x_i-\frac{12}{13})^2=(1-\frac{8}{13}\cdot 2-\frac{12}{13})^2+(5-\frac{8}{13}\cdot 4-\frac{12}{13})^2+(6-\frac{8}{13}\cdot 9-\frac{12}{13})^2=$$

$$\frac{225}{169}+\frac{441}{169}+\frac{36}{169}=\frac{702}{169}=\frac{54}{13}\approx 4{,}154$$

g_horizontal: y = 7 / 8 · x – 3 / 8

$$SQ=\sum_{i=1}^{n}(x_i-\frac{1}{m}(y_i-b))^2=\sum_{i=1}^{n}(x_i-\frac{8}{7}(y_i+\frac{3}{8}))^2=(2-\frac{8}{7}\cdot(1+\frac{3}{8}))^2+(4-\frac{8}{7}\cdot(5+\frac{3}{8}))^2+(9-\frac{8}{7}\cdot(6+\frac{3}{8}))^2=$$

$$\frac{9}{49}+\frac{225}{49}+\frac{144}{49}=\frac{378}{49}=\frac{54}{7}\approx 7{,}714$$

Da die Summe der Fehlerquadrate sich bei der „vertikalen“ Gerade als kleiner herausstellt, ist diese als Regressionsgerade besser geeignet !

Spezialfall der linearen Regression: f(x) = m·x (**Ursprungsgerade**)

Die Summe der Quadrate ist : $$SQ=\sum_{i=1}^{n}(y_i-mx_i)^2$$

Für die einzige Ableitung nach m ergibt sich:

$$\frac{dSQ}{dm}=\sum_{i=1}^{n}2(y_i-mx_i)\cdot(-x_i)$$

Nullsetzen dieser Ableitung liefert bereits die Lösung:

$$\sum_{i=1}^{n}(y_i-mx_i)\cdot(-x_i)=0 \Leftrightarrow \sum_{i=1}^{n}(-x_iy_i+mx_i^2)=0 \Leftrightarrow m\cdot\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}x_iy_i \Leftrightarrow m=\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2}\ ,$$

Ergebnis also: $$y=\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2}\cdot x$$ "vertikale" Ursprungsgerade

Wie man sieht, spielt hier der Schwerpunkt S($\bar{x}$; $\bar{y}$) keine Rolle mehr.

Beispiel von oben: P1(2;1) P2(4;5) P3(9;6) :

m = (2+20+54) / (4+16+81) = 76 / 101

Die gesuchte lineare Funktion ist: $y=\frac{76}{101}\cdot x \approx \underline{\underline{0,7524752475\cdot x}}$

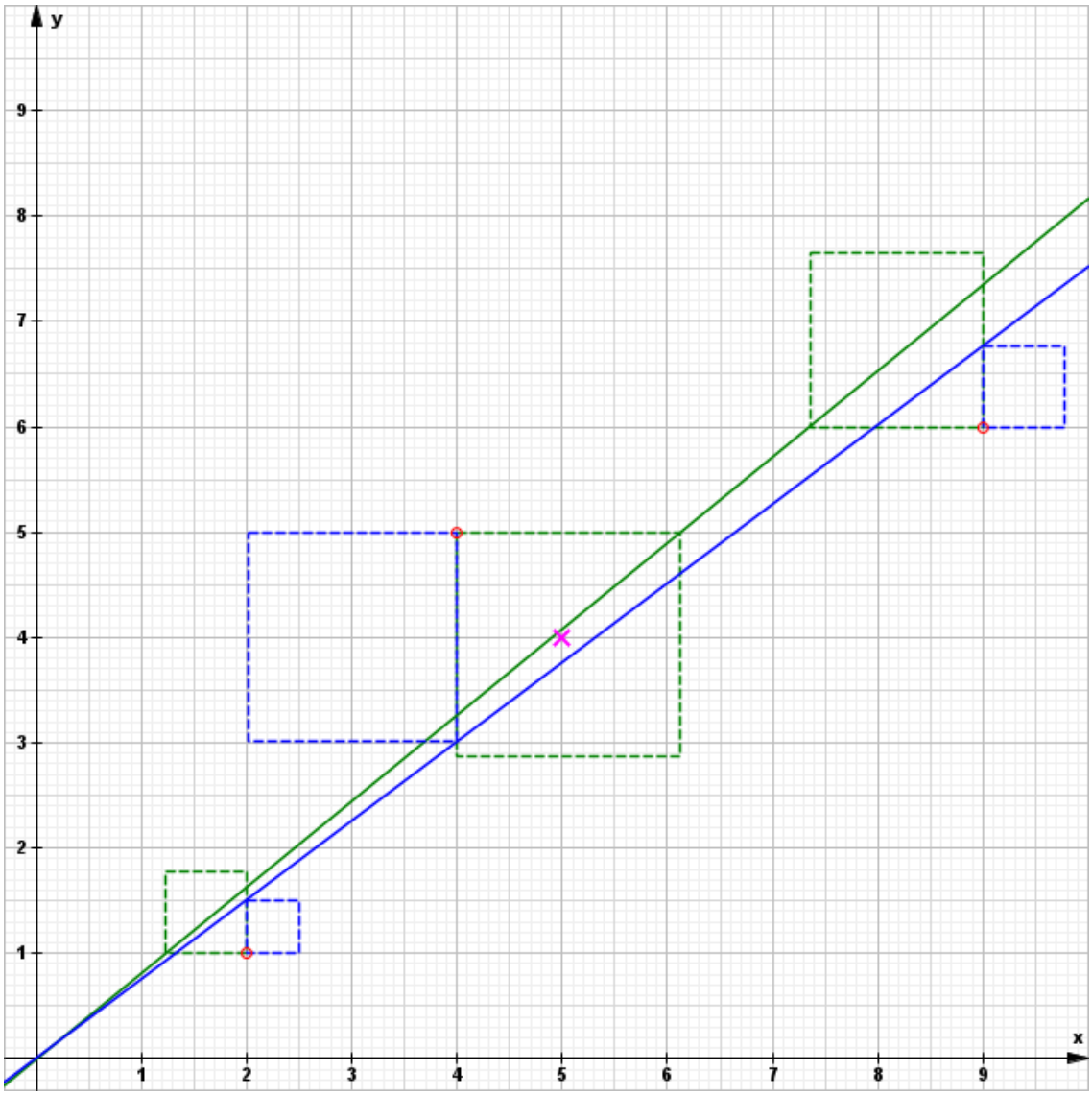
Anm.: Für die „horizontale“ Gerade erhält man analog folgende Formel:

$$y=\frac{\sum_{i=1}^{n}y_i^2}{\sum_{i=1}^{n}x_iy_i}\cdot x$$ "horizontale" Ursprungsgerade

Beispiel von oben: P1(2;1) P2(4;5) P3(9;6) :

Die gesuchte Funktionsgleichung der „horizontalen“ Ursprungsgerade ist dann :

$$y=\frac{62}{76}\cdot x=\frac{31}{38}\cdot x \approx 0,8157894737\cdot x$$



Man kann erkennen, dass der Schwerpunkt S der Daten (als Kreuz eingezeichnet) nichts mehr mit den Geradengleichungen zu tun hat !

Für die Summe der Fehlerquadrate erhält man hier:

g_vertikal (blau):

$$SQ=\sum_{i=1}^{n}(y_i-mx_i))^2=\sum_{i=1}^{n}(y_i-\frac{76}{101}x_i)^2=(1-\frac{76}{101}\cdot 2)^2+(5-\frac{76}{101}\cdot 4)^2+(6-\frac{76}{101}\cdot 9)^2=$$

$$\frac{2601}{10201}+\frac{40401}{10201}+\frac{6084}{10201}=\frac{49086}{10201}=\frac{486}{101}\approx 4,812$$

g_horizontal (grün):

$$SQ=\sum_{i=1}^{n}(x_i-\frac{1}{m}y_i))^2=\sum_{i=1}^{n}(x_i-\frac{38}{31}y_i)^2=(2-\frac{38}{31}\cdot 1)^2+(4-\frac{38}{31}\cdot 5)^2+(9-\frac{38}{31}\cdot 6)^2=$$

$$\frac{576}{961}+\frac{4356}{961}+\frac{2601}{961}=\frac{7533}{961}=\frac{243}{31}\approx 7,839$$

Auch hier ist die „vertikale“ Gerade besser geeignet !