Kurvenanpassung durch Regression (1)

Ac 2017

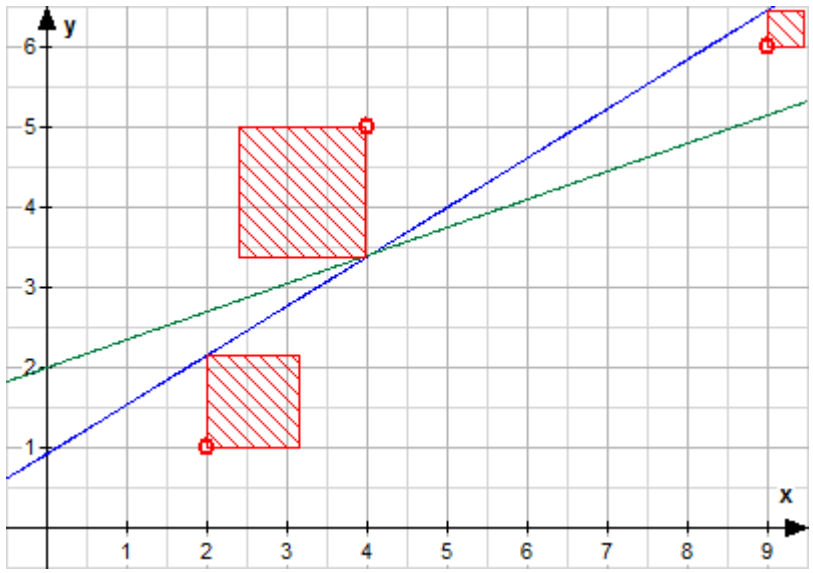
- Einführung / Lineare Regression -

Problemstellung:

Es sind n Wertepaare (Messwerte bzw. Punkte) **Pi (xi ; yi)** gegeben.
Gesucht ist eine Funktion f, die sich den Punkten Pi möglichst gut anpasst.

Lösungsidee („Methode der kleinsten Quadrate“):

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



Beispiel: P1(2;1) P2(4;5) P3(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:

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

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

SQ 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).

$$\frac{dSQ}{da} = 0 \quad \wedge \quad \frac{dSQ}{db} = 0 \quad \wedge \quad \frac{dSQ}{dc} = 0 \quad \wedge \quad \ldots$$

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

Im folgenden werden die Regressionsformeln für verschiedene Funktionsklassen hergeleitet:

Lineare Regression: f(x) = mx + b

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

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):

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

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 (xquer) sowie der y-Daten (yquer) 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):

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

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 .

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

xquer = (2+4+9) / 3 = 5 yquer = (1+5+6) / 3 = 4 ==> Schwerpunkt **S(5 ; 4)**
m = [(2-5)(1-4)+(4-5)(5-4)+(9-5)(6-4)] / [(2-5)²+(4-5)²+(9-5)²] = (9-1+8) / (9+1+16) = **8 / 13 = m**
b = 4 – 8 / 13 * 5 = 52 / 13 – 40 / 13 = **12 / 13 = b**
Die gesuchte lineare Funktion ist: **y = 8 / 13 * x + 12 / 13**

Spezialfall der linearen Regression: f(x) = mx (**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:

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

Wie man sieht, spielt hier der Schwerpunkt S(xquer ; yquer) keine Rolle mehr.

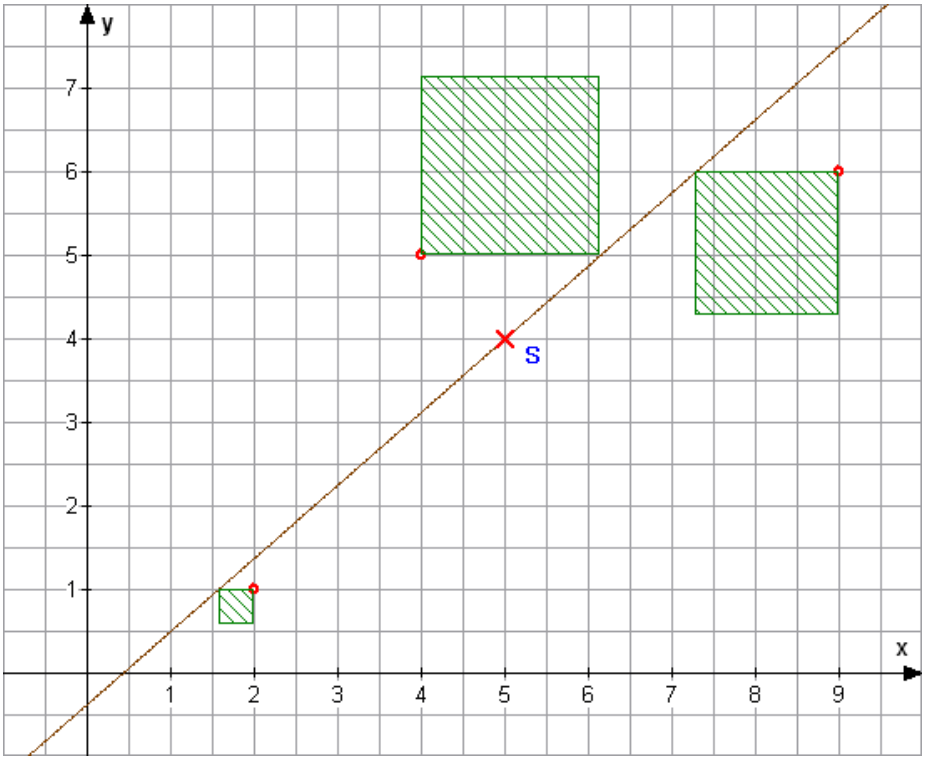
Beispiel von oben: P1(2;1) P2(4;5) P3(9;6) :

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

Die gesuchte lineare Funktion ist: **y = 76 / 101 * x**

Eine zweite Gerade für die lineare Regression:

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



Dadurch ergeben sich andere Formeln und in der Regel auch eine andere Gerade.

Für **y = m2 · x + b2** erhält man :

$$m2 = \frac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})(y_i - \bar{y})} \qquad b = \bar{y} - m2 \cdot \bar{x}$$ Gerade für waagerechte Abstände

Im Gegensatz zur „Hauptlösung“ mit den senkrechte Entfernungen, bei der die Kovarianz durch die x-Varianz dividiert wird, ist hier die y-Varianz durch die Kovarianz zu dividieren !

Vergleich der beiden Geraden: