Lineare Regression

Im Prinzip geht es darum, zwischen zwei Größen X und Y einen bestmöglichen Zusammenhang herzustellen. Meist wird die Größe Y in Abhängigkeit der Größe X gemessen (beobachtet). Man nennt Y die von der unabhängigen Größe X abhängige Größe. Dann liegen mehrere Messwertpaare (xi,yi) der jeweiligen Größen vor. Man kann deshalb die Messwerte der Größen X und Y auch durch Vektoren beschreiben: $\vec{x} = \begin{pmatrix} x_1 \\ ... \\ x_n \end{pmatrix}$ und $\vec{y} = \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix}$.

Häufig besteht - wenn überhaupt - eine lineare, quadratische oder exponentielle Abhängigkeit. Im Folgenden soll es um einen linearen Zusammenhang gehen.

Als Beispiel diene der $C0_2$-Gehalt y der Atmosphäre in Abhängigkeit der Zeit x. Der CO_2-Gehalt y ist in diesem Beispiel die von der Zeit x abhängige Größe.

Jahr	**Zeit x in Jahren ab 2015**	**CO_2Gehalt y in ppm (parts per million)**
2015	0	400 = 400 + 0
2016	1	403 = 400 + 3
2017	2	406 = 400 + 6
2018	3	408 = 400 + 8
2019	4	411 = 400 + 11
2020	5	413 = 400 + 13
2021	6	416 = 400 +16

Es ist praktisch für eine mathematische Betrachtung, die Zeit x=0 auf den Beginn der Zeitreihe (hier das Jahr 2015) zu setzen. Zum Zwecke kleinerer Zahlen wird für den CO_2-Gehalt ein Offset von 400 verwendet. Vermutet man einen linearen Zusammenhang für die Zuordnung f zwischen der Zeit x und dem CO_2-Gehalt

f : Zeit x $\rightarrow$ CO_2-Gehalt (in ppm) y, sind also die beiden Parameter s und t gesucht mit

$f(x) = y = s \cdot x + t$.

Die (Mess-)Punkte liegen in der Regel nicht auf einer Geraden. Tun sie es, ist die Bestimmung der Parameter s und t einfach. Aber aufgrund von Messfehlern (in der Physik) oder zufälligen Effekten (in der Statistik) sind die Punkte häufig nicht perfekt linear angeordnet.

Natürlich könnte man wie in unserem Beispiel eine Funktion 6. Grades $f(x) = a_6 x^6 + a_5 x^5 + \ldots + a_1 x + a_0$ (also mit 7 Freiheitsgraden) finden, auf denen alle Messpunkte liegen. Dann wird aber wahrscheinlich ein 8. Messpunkt nicht auf dieser Kurve liegen, und man müsste eine neue Funktion 7. Grades bestimmen. Dies wird der vermuteten Abhängigkeit zwischen den Größen aber nicht gerecht.

In der Graphik sind (fiktive) Messwerte als schwarze Punkte dargestellt, die blaue Gerade soll die ‚optimale' Ausgleichsgerade darstellen.

Wie findet man nun diese Gerade?

Dazu gibt es folgende Ideen:

0. Optisch – also nach Augenmaß – wird eine Ausgleichsgerade eingezeichnet

1. lineare Abweichung α in y-Richtung minimieren (Abb. 1)

2. quadratische Abweichung β in y- Richtung minimieren (Abb. 1)

3. orthogonale Abweichung γ minimieren (Abb. 2)

4. Verwendung der Vektorrechnung

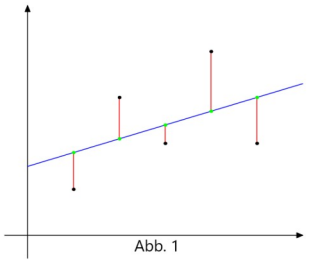


Abb. 1

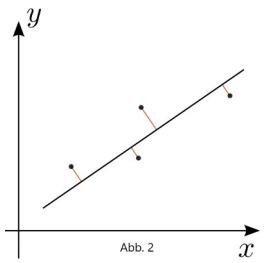


Abb. 2

Fall 0: Diese Methode ist einfach zu realisieren und bietet oftmals eine brauchbare Näherung. Die Parameter werden mit Hilfe eines Steigungsdreiecks (Steigung s) und als y-Achsenabschnitt t bestimmt. Dieses Verfahren wird gern im Schulbereich angewandt.

Fall 1: Die Summe der rot eingezeichneten Abstände (Abb. 1) in y-Richtung der Werte der abhängigen Variablen von der ‚optimalen' Geraden müssen minimiert werden:

$$\alpha = \sum_{i=1}^{n} \left| \hat{y}_i - y_i \right| \rightarrow \min$$

y_i sind Messwerte, $\hat{y}_i$ sind die zu den jeweiligen grünen Punkten gehörenden y-Werte auf der Ausgleichsgeraden.
Das Rechnen mit Beträgen ist im Allgemeinen wegen auftretender Fallunterscheidungen (Vorzeichen beachten) unerfreulich.
Diese Idee wird nicht weiter verfolgt.

Fall 2: Die Summe der Quadrate der rot eingezeichneten Abstände (Abb. 1) in y-Richtung der Werte der abhängigen Variablen von der ‚optimalen' Geraden müssen minimiert werden:

$$\beta = \sum_{i=1}^{n} \left(\hat{y}_i - y_i \right)^2 \rightarrow \min$$

Durch die Bildung der Abstandsquadrate erreicht man eine höhere Gewichtung der größeren Abweichungen von der ‚optimalen' Geraden. Fallunterscheidungen werden vermieden.

$\hat{y}_i$ sind die zu den y_i – Werten gehörenden y-Werte der grünen Punkte auf der Ausgleichsgeraden, demzufolge gilt auch hier: $\hat{y}_i = s \cdot x_i + t$
Nun kann β minimiert werden. Dazu bildet man die partiellen Ableitungen nach s und t und bestimmt die Nullstellen dieser Ableitungen.

Man erhält mit den arithmetischen Mittelwerten $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ und $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$:

$$s = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i} \qquad \text{und} \qquad t = \bar{y} - s \cdot \bar{x}$$

Angewendet auf unser Beispiel lautet die Ausgleichsgerade (gerundet) : $f(x) = 2{,}61 \cdot x + 0{,}32$

Fall 3: Die Summe der Quadrate der rot eingezeichneten Abstände (Abb. 2) der Messpunkte von der ‚optimalen' Geraden müssen minimiert werden: $\gamma = \sum_{i=1}^{n} \left(\hat{x}_i - x_i\right)^2 + \left(\hat{y}_i - y_i\right)^2$

$(\hat{x}_i\ ,\ \hat{y}_i)$ sind die zu den Messpunkten $(x_i\ ,\ y_i)$ gehörenden Fußpunkte auf der Ausgleichsgeraden. Dieses Verfahren kann verwendet werden, wenn beide Größen fehlerbehaftet sind. Das ist beispielsweise bei physikalischen Messungen häufig der Fall. Es wird nicht mehr zwischen einer abhängigen und einer unabhängigen Größe unterschieden.

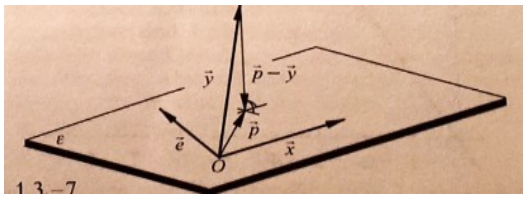
Mit den Mittelwerten $\bar{x}$ und $\bar{y}$ sowie

$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$, $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$, $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$ erhält man

$$s = \frac{s_y^2 - s_x^2 + \sqrt{\left(s_y^2 - s_x^2\right)^2 + 4\,s_{xy}^2}}{2\,s_{xy}} \quad \text{und} \qquad t = \bar{y} - s \cdot \bar{x}$$

Angewendet auf unser Beispiel lautet die Ausgleichsgerade (gerundet) : $f(x) = 2{,}27 \cdot x + 1{,}34$

Fall 4: Die Erstellung einer Regressionsgeraden mit Hilfe der Vektorrechnung bietet eine interessante Möglichkeit, den dreidimensionalen Anschauungsraum auf höhere Dimensionen zu erweitern. Bei n Wertepaaren befindet man sich im n-dimensionalen Raum: in dem oben angeführten Beispiel demzufolge im 7-dimensionalen Raum.



Das n-dimensionale Gleichungssystem wird als Vektorgleichung geschrieben:

$$\begin{aligned} y_1 &= s \cdot x_1 + t \\ &\ldots \qquad \rightarrow \qquad \vec{y} = s \cdot \vec{x} + t \begin{pmatrix} 1 \\ \ldots \\ 1 \end{pmatrix} = s \cdot \vec{x} + t \cdot \vec{e} \\ y_n &= s \cdot x_n + t \end{aligned}$$

Durch $\vec{x}$ und $\vec{e}$ wird eine 2-dimensionale Ursprungsebene im n-dimensionalen Raum aufgespannt, in der idealerweise der Punkt $Y(y_1, \ldots, y_n)$ liegt. Liegt Y nicht in der Ebene, sucht man als bestmögliche Näherung den (Lotfuß)Punkt P in der Ebene, der dem Punkt Y am nächsten liegt:

$\vec{p} = s \cdot \vec{x} + t \cdot \vec{e}$

Es gilt $\vec{p} - \vec{y}$ ist sowohl orthogonal zu $\vec{x}$ als auch zu $\vec{e}$: $\vec{p} - \vec{y} \perp \vec{x}, \vec{e}$

Diese Orthogonalitätsbedingungen führen zu den Skalarproduktgleichungen

$(\vec{p} - \vec{y}) \cdot \vec{x} = 0$ und $(\vec{p} - \vec{y}) \cdot \vec{e} = 0$

Setzt man für $\vec{p}$ ein, erhält man $(s \cdot \vec{x} + t \cdot \vec{e} - \vec{y}) \cdot \vec{x} = 0$

sowie $(s \cdot \vec{x} + t \cdot \vec{e} - \vec{y}) \cdot \vec{e} = 0$

Beachtet man, dass $\vec{x} \cdot \vec{x} = \sum_{i=1}^{n} x_i^2$, $\vec{x} \cdot \vec{y} = \sum_{i=1}^{n} x_i y_i$, $\vec{x} \cdot \vec{e} = n \cdot \bar{x}$, $\vec{y} \cdot \vec{e} = n \cdot \bar{y}$
lassen sich s und t leicht errechnen. Das Ergebnis ist identisch zu Fall 2.
Durch die Minimierung des Abstands von P zur Ebene wird also auch hier die Summe der Abstandsquadrate in y-Richtung minimiert.
Die Ausgleichsgerade lautet wie in Fall 2: $f(x) = 2{,}61 \cdot x + 0{,}32$

Weiteres Beispiel:

X in Monaten, Y in Jahren gemessen

	X = Dauer der Schwangerschaft	Y = Lebens-erwartung
Lemur	18	18
Makak	24	26
Gibbon	30	30
Schimpanse	34	40
Mensch	38	70