

Berechnung der bivariaten linearen Regression

Eine Gerade hat die folgende allgemeine Form: $y = ax + b$

Wir möchten die Gerade möglichst gut an die Punkte anpassen. Als Mass der Anpassung wählen wir die Summe der quadrierten Abstände, wobei wir die Abstände parallel zur y-Achse messen. Diese Summe sollte minimal sein. Dazu müssen wir a und b so wählen, dass diese Summe minimal wird.

Die Summe der quadrierten Abstände ist: $S(a, b) = \sum_{i=1}^n (y_i - (ax_i + b))^2$

Wir müssen nach a und nach b ableiten, um das Minimum zu berechnen:

$$(a) \frac{\partial}{\partial a} \sum_{i=1}^n (y_i - (ax_i + b))^2 = -2 \sum_{i=1}^n (y_i - (ax_i + b))x_i$$

$$(b) \frac{\partial}{\partial b} \sum_{i=1}^n (y_i - (ax_i + b))^2 = -2 \sum_{i=1}^n (y_i - (ax_i + b))$$

Wenn $S(a, b)$ ein Minimum hat, dann liegt dieses an den Nullstellen der 1. Ableitung von (a) und (b) (Wir verzichten auf den Beweis, dass $S(a, b)$ an den Nullstellen tatsächlich ein Minimum und nicht etwa ein Maximum oder einen Sattelpunkt annimmt):

$$(a) -2 \sum_{i=1}^n (y_i - (ax_i + b))x_i = 0 \text{ genau dann, wenn } \sum_{i=1}^n (y_i - (ax_i + b))x_i = 0$$

$$\sum_{i=1}^n (y_i - (ax_i + b))x_i = \sum_{i=1}^n (x_i y_i - ax_i^2 - bx_i) = \sum_{i=1}^n x_i y_i - \sum_{i=1}^n ax_i^2 - \sum_{i=1}^n bx_i = 0$$

$$\text{Damit gilt: } \sum_{i=1}^n bx_i + \sum_{i=1}^n ax_i^2 = \sum_{i=1}^n x_i y_i$$

$$(b) -2 \sum_{i=1}^n (y_i - (ax_i + b)) = 0 \text{ genau dann, wenn } \sum_{i=1}^n (y_i - (ax_i + b)) = 0$$

$$\sum_{i=1}^n (y_i - (ax_i + b)) = \sum_{i=1}^n y_i - \sum_{i=1}^n ax_i - \sum_{i=1}^n b = 0$$

$$\text{Damit gilt (Da } \sum_{i=1}^n b = nb): \quad nb + \sum_{i=1}^n ax_i = \sum_{i=1}^n y_i$$

Diese beiden Gleichungen können wir in Matrixschreibweise darstellen:

$$\begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

Beispiel aus der Pisastudie

Die Daten der Pisa-Studie erfüllen die Anforderungen, die zu Beginn gestellt wurden nicht (Symmetrie der Punktwolke, ungefähr auf einer Geraden, die Leistungspunkte sind wohl ordinal, nicht jedoch metrisch skaliert). Trotzdem wollen wir die Berechnung an diesem Beispiel vorführen:

	Land	x_i Ausgaben pro Kopf	y_i Mittlere Fähigkeit	$x_i \cdot y_i$	$(x_i^2) \cdot y_i$	x_i^2
1	Australia	44'623	528	23573604.47	1'051'932'361'049	1991240178
2	Austria	71'387	507	36202303.74	2'584'360'609'650	5096051525
3	Belgium	46'338	507	23499389.94	1'088'914'731'040	2147210244
4	Czech Republic	21'384	492	10512172.71	224'797'133'757	457295117.3
5	Denmark	65'794	497	32691173.41	2'150'890'211'030	4328879206
6	Finland	45'363	546	24789473.39	1'124'522'829'903	2057794261
7	France	50'481	505	25479639.83	1'286'230'625'681	2548303336
8	Germany	41'978	484	20316977.18	852'867'955'342	1762160283
9	Greece	27'356	474	12961410.66	354'576'121'248	748366655.2
10	Hungary	20'277	480	9732218.92	197'337'510'992	411145511.2
11	Ireland	31'015	527	16334747.43	506'624'591'442	961939338.6
12	Italy	60'824	487	29649881.45	1'803'424'764'954	3699560516
13	Japan	53'255	522	27811183.21	1'481'075'218'896	2836059244
14	Korea	30'844	525	16185493.78	499'228'601'996	951364653.6
15	Mexico	11'239	422	4742208.62	53'295'437'095	126304477
16	Norway	61'677	505	31164186.60	1'922'115'512'852	3804060151
17	Poland	16'154	479	7739669.39	125'026'052'551	260949350
18	Portugal	36'521	470	17170362.21	627'079'311'802	1333785625
19	Spain	36'699	493	18075881.32	663'359'020'208	1346785139
20	Sweden	53'386	516	27564977.74	1'471'593'743'624	2850103119
21	Switzerland	64'266	494	31771151.90	2'041'802'886'964	4130110822
22	United Kingdom	42'793	523	22399616.65	958'548'880'284	1831248816
23	United States	67'313	504	33953874.52	2'285'527'129'993	4531000218
	Summen	1000967	11489	504321599	25355131242351	50211717785

Wir erhalten somit das Gleichungssystem in Matrizenform:

$$\begin{pmatrix} 23 & 1000967 \\ 1000967 & 50211717785 \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} 11489 \\ 504321599.07 \end{pmatrix}$$

(oder in Gleichungsform:

$$23b + 1000967a = 11489$$

$$1000967b + 50211717785a = 504321599.07$$

Lösung: $a = 6.4921 \times 10^{-4}$, $b = 471.27$

Wir erhalten somit die Regressionsgleichung:

$$y = 0.00064921x + 471.27$$

Bivariate, quadratische Regression

Wir können auch ein quadratisches Polynom durch Daten legen.

Die Funktion hat die allgemeine Form: $y = ax^2 + bx + c$

Die Summe der quadrierten Abweichungen ist: $S(a, b, c) = \sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c))^2$

Wiederum leiten wir nach a , b und c ab, um das Minimum von $S(a, b, c)$ zu bestimmen.

$$(a) \frac{\partial}{\partial a} \sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c))^2 = -2 \sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c))x_i^2$$

$$(b) \frac{\partial}{\partial b} \sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c))^2 = -2 \sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c))x_i$$

$$(c) \frac{\partial}{\partial c} \sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c))^2 = -2 \sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c))$$

Im Minimum sind die Ableitungen mit 0 identisch:

$$(a) -2 \sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c))x_i^2 = 0 \text{ genau dann, wenn } \sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c))x_i^2 = 0$$

$$\sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c))x_i^2 = \sum_{i=1}^n (x_i^2 y_i - ax_i^4 - bx_i^3 - cx_i^2) = \sum_{i=1}^n x_i^2 y_i - \sum_{i=1}^n ax_i^4 - \sum_{i=1}^n bx_i^3 - \sum_{i=1}^n cx_i^2$$

$$\text{Somit gilt: } \sum_{i=1}^n cx_i^2 + \sum_{i=1}^n bx_i^3 + \sum_{i=1}^n ax_i^4 = \sum_{i=1}^n x_i^2 y_i$$

$$(b) -2 \sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c))x_i = 0 \text{ genau dann, wenn } \sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c))x_i = 0$$

$$\sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c))x_i = \sum_{i=1}^n (x_i y_i - ax_i^3 - bx_i^2 - cx_i) = 0$$

$$\sum_{i=1}^n ax_i^3 + \sum_{i=1}^n bx_i^2 + \sum_{i=1}^n cx_i = \sum_{i=1}^n x_i y_i$$

$$(c) -2 \sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c)) = 0 \text{ genau dann, wenn } \sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c)) = 0$$

$$\sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c)) = \sum_{i=1}^n y_i - \sum_{i=1}^n ax_i^2 - \sum_{i=1}^n bx_i - \sum_{i=1}^n c = 0$$

$$\sum_{i=1}^n c + \sum_{i=1}^n bx_i + \sum_{i=1}^n ax_i^2 = \sum_{i=1}^n y_i$$

In Matrizenform ergibt dies:

$$\begin{pmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^4 \end{pmatrix} \begin{pmatrix} c \\ b \\ a \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i^2 y_i \end{pmatrix}$$

Man sieht an der Matrix sofort, wie die Berechnungsmethode vermutlich auf Polynome höheren Grades zu verallgemeinern ist.

Am Beispiel der Pisa-Daten:

$$\begin{pmatrix} 23 & 1000967 & 50211717785 \\ 1000967 & 50211717785 & 2746075883034200 \\ 50211717785 & 2746075883034200 & 158828749434821000000 \end{pmatrix} \begin{pmatrix} c \\ b \\ a \end{pmatrix} = \begin{pmatrix} 11489 \\ 504321599.07 \\ 25355131242351 \end{pmatrix},$$

Lösung: $a = -4.7669 * 10^{-8}$; $b = 4.6699 * 10^{-3}$ $c = 400.35$

$y = -4.7669 * 10^{-8}x^2 + 4.6699 * 10^{-3}x + 400.35$