

## Ein Extremalproblem aus der Statistik: Lineare Regression

Messungen haben die  $n$  Wertepaare  $(x_i, y_i)$   $i = 1 \dots n$  ergeben. Liegen die Punkte angenähert auf einer Geraden, so kann man rechnerisch eine Ausgleichsgerade (to fit) bestimmen, die von den  $n$  Punkten "möglichst wenig" abweicht.

einführendes Beispiel: (2,3), (6,4), (10,8)

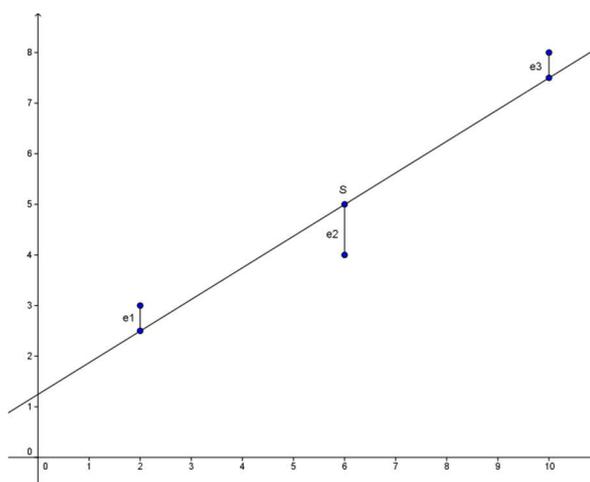
$i$	$x_i$	$y_i$	$x_i y_i$	$x_i^2$
1	2	3	6	4
2	6	4	24	36
3	10	8	80	100
$\Sigma$	18	15	110	140

Mittelwert der x-Koordinaten:  $\bar{x} = 6$

Mittelwert der y-Koordinaten:  $\bar{y} = 5$

Die Skizze liefert als vermutete Ausgleichsgerade

$$g: y = 0.6x + 1.2$$



Die vertikal gemessenen Abweichungen der Punkte von der Ausgleichsgeraden heissen Residuen  $e_i$

Wir machen die plausible Annahme, dass die Ausgleichsgerade durch den Schwerpunkt  $S(\bar{x}, \bar{y})$  der Punkte geht und setzen die Gleichung der Ausgleichsgeraden in der folgenden Form an:

$$y = mx + q$$

S liegt auf der Ausgleichsgeraden  $q = \bar{y} - m\bar{x}$   $y = m \cdot (x - \bar{x}) + \bar{y}$

Als Mass für die Abweichungen der Punkte von der Ausgleichsgeraden wird auf Vorschlag von C.F. Gauss die Summe  $D$  der Residuenquadrate  $e_i^2$  verwendet. Diese Summe ist von der Steigung  $m$  der Geraden abhängig, d.h.  $D$  ist eine Funktion von  $m$ .

$$D(m) = \sum e_i^2 = (m \cdot (x_i - \bar{x}) + \bar{y} - y_i)^2$$

$i$	$x_i$	$y_i$	$x_i - 6$	$m \cdot (x_i - 6)$	$5 - y_i$	$e_i = m \cdot (x_i - 6) + 5 - y_i$	$e_i^2$
1	2	3	-4	-4m	2	-4m + 2	$(-4m + 2)^2$
2	6	4	0	0	1	1	0
3	10	8	4	4m	-3	4m - 3	$(4m - 3)^2$

$$D(m) = (-4m + 2)^2 + 1^2 + (4m - 3)^2 = 32m^2 - 40m + 14$$

Der Graph dieser quadratischen Funktion ist eine nach oben geöffnete Parabel. Die Bedingung  $D'(m) = 0$  liefert  $64m - 40 = 0$  und daraus  $m = \frac{5}{8}$ .

Die Ausgleichsgerade hat damit die Gleichung

$$y = \frac{5}{8}x + \frac{5}{4}$$

allg. Fall

Mit dem folgenden Ansatz für die Gleichung der Ausgleichsgeraden

$$y = m \cdot (x - \bar{x}) + \bar{y} \text{ erhalten wir}$$

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

Der Graph der Funktion D ist eine nach oben geöffnete Parabel. D ist genau dann minimal, wenn  $D'(m) = 0$  ist.

$$D'(m) = 2 \sum_{i=1}^n (m(x_i - \bar{x}) - (y_i - \bar{y})) \cdot (x_i - \bar{x}) = 0$$

Die Gleichung wird durch 2 dividiert und die Summe unterteilt.

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

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

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

denn

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

$$\sum_{i=1}^n \bar{x} \bar{y} = n \bar{x} \bar{y} \quad \text{bzw.} \quad \sum_{i=1}^n \bar{x}^2 = n \bar{x}^2$$

Damit gilt:

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

Die Formel kann auch folgendermassen geschrieben werden:

$$m = \frac{s_{xy}}{s_x^2}$$

Der Term im Zähler heisst Kovarianz der Stichprobe, der im Nenner Varianz der x-Werte.

Für das einführende Beispiel ergibt sich erneut mit  $\bar{x} = 6$   $\bar{y} = 5$

$$m = \frac{110 - 3 \cdot 6 \cdot 5}{140 - 3 \cdot 36} = \frac{5}{8}$$

Übungsaufgaben:

a)

$x_i$	$y_i$	$x_i y_i$	$x_i^2$
-1	1	-1	1
2	2	4	4
3	3	9	9
6	4	24	36
8	5	40	64

18    15    76    114

Als Gleichung für die Ausgleichsgerade ergibt sich grafisch  $y = 0.45x + 1.4$

$$m = \frac{76 - 54}{114 - 64.8} = 0.447 \quad \text{und in die Geradengleichung eingesetzt } y = 0.447x + 1.39.$$

b)

$x_i$	$y_i$	$x_i y_i$	$x_i^2$
1	3	3	1
2	6	12	4
3	7	21	9
4	11	44	16

10    27    80    30

Als Gleichung für die Ausgleichsgerade ergibt sich  $y = 2.5x + 0.5$