LORE LORENTZ SCHULE

Glossar: Lineare Regression die Regressionsgerade

Regressionsgerade [Statistik, beschreibende]

Gegeben sind zwei quantitative Merkmale (Größen) x und y, zwischen denen ein statistischer Zusammenhang vermutet wird.
Dazu liegen n aneinander gekoppelte Merkmalsausprägungen vor: $\{(x_i|y_i) \mid 1 \leq i \leq n\}$
Anders ausgedrückt: Man verfügt z.B. aus einer Messreihe über n Wertepaare $(x_1|y_1)$, ..., $(x_n|y_n)$.
Diejenige Gerade, für die dabei die mittlere quadratische Abweichung möglichst klein wird, ist die *Regressionsgerade bezüglich* x.

Die Regressionsgerade entspricht derjenigen lineare Funktion, die zu einem gegebenen Wert von x die optimale Schätzung des Wertes von y angibt – dabei heißt optimal eben, dass die mittlere quadratische Abweichung minimal ist.

Eigenschaften: Die Regressionsgerade geht durch den Schwerpunkt der Punktwolke – also durch $(\bar{x} \mid \bar{y})$
($\bar{x}$ ist das arithmetische Mittel der Messwerte x_i,
$\bar{y}$ ist das arithmetische Mittel der Messwerte y_i).

Bem.: Die Regressionsgerade bzgl. x entspricht nicht der Regressionsgerade bzgl. y (außer der Zusammenhang ist exakt linear, also alle Punkte liegen exakt auf einer Geraden).

Beispielrechnung:

Gegeben sind drei Wertepaare
x_1=2 und y_1=7,
x_2=4 und y_2=8 und
x_3=9 und y_3=3
Zuerst berechnet man das jeweilige arithmetische Mittel:
$\bar{x}$ = 5, $\bar{y}$ =6.
Dann betrachtet man diejenigen Geraden, die durch den Schwerpunkt $(5 \mid 6)$ gehen.

$r(x) = m\,x + b$
$r(5) = 5\,m + b = 6$ (weil r durch den Schwerpunkt gehen soll)
$\Leftrightarrow\ b = -5\,m + 6$
$r(x) = m\,x - 5\,m + 6 = (x-5)\cdot m + 6$

Die Berechnung von m geht am schnellsten auf Basis der Varianz und Covarianz:

i	x_i	y_i	x_i-$\bar{x}$	y_i-$\bar{y}$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	2	7	-3	1	9	-3
2	4	8	-1	2	1	-2
3	9	3	4	-3	16	-12
∑	15	18			26	-17
Mittel	$\bar{x}$=5	$\bar{y}$=6			**Varianz bzgl. x:** $\frac{26}{3}$	**Covarianz:** $-\frac{17}{3}$

$$m = \frac{Covarianz}{Varianz} = \frac{17}{26} \approx -0{,}6538$$

Alternativ kann man sich verdeutlichen, dass quadratische Abweichungen minimiert werden, indem man diese als Funktion mit der Variablen m betrachtet:

i (Nr.)	x_i	y_i	quadratische Abweichung	
1	2	7	$\big((2-5)\cdot m + 6 - 7\big)^2 = (-3m-1)^2$	$= 9\,m^2 + 6\,m + 1$
2	4	8	$\big((4-5)\cdot m + 6 - 8\big)^2 = (-m-2)^2$	$= m^2 + 4\,m + 4$
3	9	3	$\big((9-5)\cdot m + 6 - 3\big)^2 = (4m+3)^2$	$= 16\,m^2 + 24\,m + 9$
∑	15	18		$= 26\,m^2 + 34\,m + 14$
Mittel	**5**	**6**		

Die Summe der quadratischen Abweichungen
– also $s(m) = 26\,m^2 + 34\,m + 14$ – muss nun minimiert werden.
(das geht mit Hilfe der Differentialrechnung
oder indem man die Scheitelpunktform bestimmt
oder indem man zwei Stellen berechnet, die denselben Funktionswert haben und die Stelle in der Mitte von beiden bestimmt.)

Hier wird nun die Minimierung mit Hilfe der

Differentialrechnung vorgeführt:
Ableitung:
$s'(m) = 52m + 34$
notwendige Bedingung: $q'(x) = 0$
$52m + 34 = 0$
$\Leftrightarrow 52m = -34$
$\Leftrightarrow m \approx -0{,}6538$
(Dass es sich wirklich um eine Minimierung handelt und nicht aus Versehen maximiert wurde, ist sofort klar, wenn man sich vergegenwärtigt, dass zur Funktion s eine nach oben geöffnete Parabel gehört.)

Nun erhält man die Gleichung der Regressionsgeraden durch Einsetzen:
$r(x) = (x - 5) \cdot (-0{,}6538) + 6$
$= -0{,}6538\,x + 5 \cdot 0{,}6538 + 6$
$= -0{,}6538\,x + 9{,}269.$

Beispiel für eine Berechnung unter Benutzung der Scheitelpunktform: Helmholtz-BI

Spätestens, wenn man mehr als fünf Wertepaare hat, hat man in der Regel keine Lust mehr, die Regression selbst durchzurechnen. Da hilft z.B. der Nspire oder der TI30X-II oder Geogebra (Trendlinie())