Regressionsgerade

Die Regressionsgerade (Ausgleichsgerade) verwendet man, um in möglichst guter Näherung einen linearen Zusammenhang zwischen zwei verschiedenen Größen (z. B. Gewicht und Körpergröße) darzustellen.

Gegeben sind n Wertepaare (x_1/y_1), (x_2/y_2),.....(x_n/y_n) zweier Größen x und y, die voneinander abhängen. Die Darstellung dieser Wertepaare im Koordinatensystem ergibt eine sog. Punktwolke.
Durch diese Punktwolke soll eine Gerade $y = m\,x + b$ so gelegt werden, dass ein linearer Zusammenhang zwischen x und y so genau wie möglich beschrieben wird. Das ist nur möglich, wenn die Regressionsgerade bestimmte, sinnvolle Bedingungen erfüllt.

1) Die Summe der quadratischen Abstände SAQ in y-Richtung aller Punkte $P_i = (x_i/y_i)$ von der Regressionsgeraden soll minimal sein.

$$SAQ = \sum_{i=1}^{n} [y_i - (m\,x_i - b)]^2 \qquad (*)$$

In dieser Gleichung ist y_i die y-Koordinate eines Punktes $P_i = (x_i/y_i)$ der Punktwolke; $m\,x_i - b$ ist die y-Koordinate eines Punktes mit der x-Koordinate x_i, der auf der Regressionsgeraden liegt.
SAQ hängt von den beiden Variablen m und b ab.

2) Die Regressionsgerade soll durch den Schwerpunkt S der Sunktwolke verlaufen. Der Schwerpunkt S hat die Koordinaten

$$S = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i \;\middle|\; \frac{1}{n} \cdot \sum_{i\,1}^{n} y_i \right) = \left(\bar{x} \;\middle|\; \bar{y} \right)$$

Die Schwerpunktkoordinaten müssen also die Gleichung der Regressionsgeraden $y = m\,x + b$ erfüllen. Folglich gilt:
$\bar{y} = m\,\bar{x} + b \quad \Leftrightarrow \quad b = \bar{y} - m\,\bar{x} \;\; (\alpha)$ Durch Einsetzen in (*) erhält man:

$$SAQ = f(m) = \sum_{i=1}^{n} [(y_i - (m\,x_i + \bar{y} - m\,\bar{x})]^2$$

Die Summe der Abweichungsquadrate ist eine Funktio, die nun nur noch von der einzigen Variablen m abhängt.

$$f(m) = \sum_{i=1}^{n} [y_i - a\,m\,x_i - \bar{y} + m\,\bar{x}]^2 = \sum_{i=1}^{n} [(y_i - \bar{y}) - m\,(x_i - \bar{x})]^2$$

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

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

Die $\sum_{i=1}^{n} (y_i - \bar{y})^2$ ist das n-fache der mittleren quadratischen Abweichung $\overline{S_y^2}$ vom Mittelwert $\bar{y}$.

Die $\sum_{i=1}^{n} (x_i - \bar{x})^2$ ist das n-fache der mittleren quadratischen Abweichung $\overline{S_x^2}$ vom Mittelwert $\bar{x}$.

Damit ergibt sich:

$$f(m) = n\,\overline{S_y^2} - 2m \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + m^2 n\,\overline{S_x^2}$$

$$= n\left[\overline{S_y^2} - 2m \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + m^2\,\overline{S_x^2}\right]$$

Für den Term $\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ schreiben wir die Abkürzung $\overline{S_{xy}}$ und erhalten damit

$$f(m) = n\,(\overline{S_y^2} - 2m\,\overline{S_{xy}} + m^2\,\overline{S_x^2})$$

Durch diese Gleichung ist eine quadratische Funktion mit der Variablen m festgelegt. Um das Minimum dieser quadratischen Funktion zu bestimmen, berechnet man die m-Koordinate $P_{S,m}$ ihres Scheitelpunktes P_S.

Diese Koordinate hängt nicht von der Zahl n ab und ist folglich identisch mit der m-Koordinate des Scheitelpunktes der Funktion

$f^*(m) = \overline{S_y^2} - 2m\,\overline{S_{xy}} + m^2\,\overline{S_x^2}$ Dievidiert man diese Funktionsgleichung durch $\overline{S_x^2}$, so ändert das ebenfalls nichts an der m-Koordinate des Scheitelpunktes.

Bestimmung der m-Koordinate des Scheitelpunkts P_S

$$m^2 - 2 \cdot \frac{\overline{S_{xy}}}{\overline{S_x^2}} \cdot m + \frac{\overline{S_y^2}}{\overline{S_x^2}} = m^2 - 2 \cdot \frac{\overline{S_{xy}}}{\overline{S_x^2}} + \left(\frac{\overline{S_{xy}}}{\overline{S_x^2}}\right)^2 + \frac{\overline{S_y^2}}{\overline{S_x^2}} - \left(\frac{\overline{S_{xy}}}{\overline{S_x^2}}\right)^2 =$$

$$\left(m - \frac{\overline{S_{xy}}}{\overline{S_x^2}}\right)^2 + \frac{\overline{S_y^2}}{\overline{S_x^2}} - \frac{\overline{S_{xy}}}{\overline{S_x^2}}$$

$m = \frac{\overline{S_{xy}}}{\overline{S_x^2}}$ ist die m-Koordinate des Scheitelpunktes und folglich die Steigung der Regressionsgeraden.

Den y-Achsenabschnitt b der Regressionsgeraden erhält man, indem man den Term für die Steigung m in Gleichung (α) einsetzt.

$$b = \bar{y} - m\bar{x} = \bar{y} - \frac{\overline{S_{xy}}}{\overline{S_x^2}} \cdot \bar{x}$$

Damit erhält man nun insgesamt für die Regressionsgerade die Gleichung

$$\underline{\underline{y = \frac{\overline{S_{xy}}}{\overline{S_x^2}} \cdot x + \left(\bar{y} - \frac{\overline{S_{xy}}}{\overline{S_x^2}}\right)}}$$