LINEARE REGRESSION UND KORRELATION - EIN ELEMENTARER ZUGANG

von S.M. Goode und E.J. Gold, Bristol

Originaltitel in "Teaching Statistics" Vol. 9 (1987) Nr. 2:
Linear Regression and Correlation - an enlightening approach

Übersetzung: Hans-Joachim Bentz, Osnabrück

Zusammenfassung: In diesem Artikel werden die Parameter der Regressionsgleichung sowie einige wesentliche mathematische Eigenschaften des Korrelationskoeffizienten auf *elementarem* Wege hergeleitet.

Gegeben seien die beiden Variablen x und y:

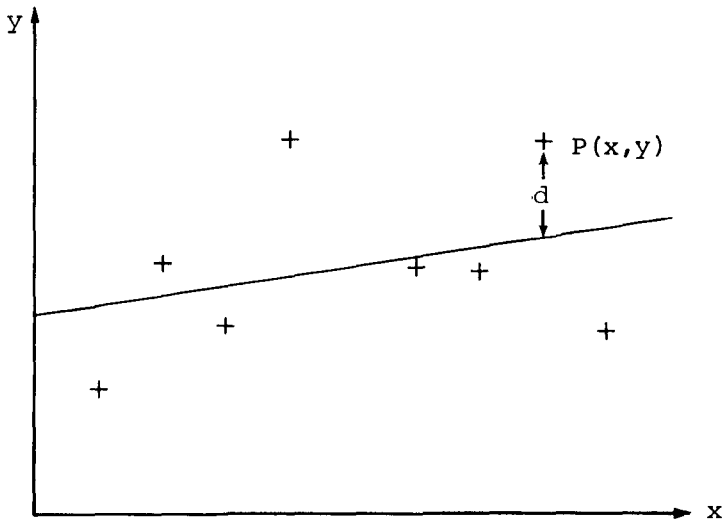

Fig.1

Die Regressionsgerade von y auf x ist so definiert, daß Σd_i^2 minimiert wird. Nimmt man folgende Variablentransformation vor:

$$X = x-\bar{x}, \quad Y = y-\bar{y}, \quad \text{wobei} \quad \bar{x} = \frac{\Sigma x_i}{n}, \quad \bar{y} = \frac{\Sigma y_i}{n},$$

so erhält man folgendes Bild:

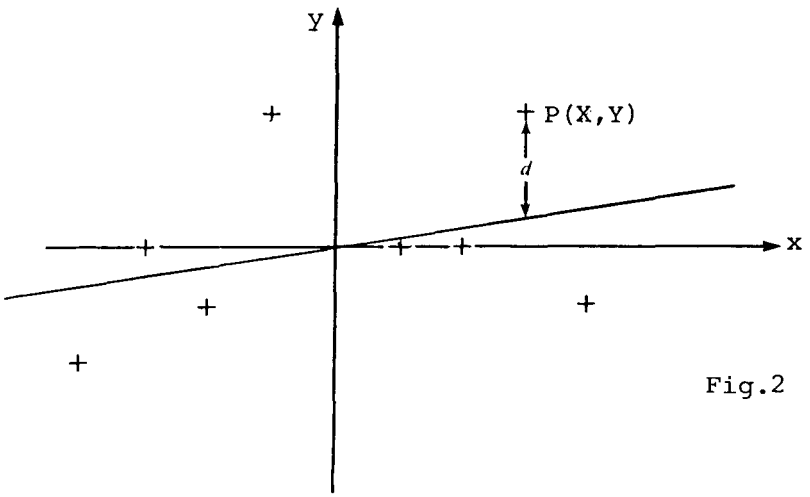

Fig.2

Bemerkung: In der deutschen Literatur ist manches Mal die Unterscheidung zwischen X und x üblich: X bezeichnet die Zufallsvariable, x einen konkreten Wert von X, etwa x = 14. Die Bezeichnungen oben haben damit nichts gemeinsam.

In einigen Lehrbüchern werden die Themengebiete Regression und Korrelation sehr zu unserem Bedauern vollständig getrennt behandelt. In anderen wiederum wird vermieden, auf die Theorie einzugehen, wie man zu den Gleichungen für die Regressionsgerade bzw. zu einer Rechtfertigung für die Formel für den Produktmoment-Korrelationskoeffizienten kommt. Manchmal wird in Lehrbüchern auch auf partielle Differentiation zurückgegriffen, um die ersteren Gleichungen zu erhalten - ein Stoffgebiet, das in den meisten Lehrplänen für Reine Mathematik auf dem G/E Advanced Level nicht vorkommt. Dabei gibt es einen ganz elementaren und trotzdem vollständig verständlichen Zugang, dieser soll im folgenden dargestellt werden:

Die Regressionsgerade von Y auf X sei festgelegt durch:

$$Y = mX + c$$

Zu minimieren ist Σd_i^2, d.h.

$$\Sigma d_i^2 = \Sigma (Y_i - mX_i - c)^2.$$

Der Übersicht wegen werden im folgenden die Indices i weggelassen. Ausquadrieren liefert:

$$\Sigma d^2 = \Sigma Y^2 + m^2 \Sigma X^2 + \Sigma c^2 - 2m \Sigma XY + 2mc \Sigma X - 2c \Sigma Y =$$
$$= \Sigma Y^2 + m^2 \Sigma X^2 - 2m \Sigma XY + nc^2$$

Dies deswegen, weil ΣX und ΣY gleich Null sind. Vervollständigen des Quadrats in m ergibt ferner:

$$\Sigma d^2 = (\Sigma X^2)[m - \frac{\Sigma XY}{\Sigma X^2}]^2 + \Sigma Y^2 - \frac{(\Sigma XY)^2}{\Sigma X^2} + nc^2$$

Nun sieht man unmittelbar, daß dieser Ausdruck minimal wird, wenn

$c = 0$ (d.h. die Regressionsgerade geht durch $(\bar{x},\bar{y})$ und

$$m = \frac{\Sigma XY}{\Sigma X^2} = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}$$

ist.

Die Regressionsgerade von y auf x ist daher

$$Y = \frac{\Sigma XY}{\Sigma X^2} \cdot X \quad \text{bzw.} \quad y-\bar{y} = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\Sigma (x-\bar{x})^2}(x-\bar{x}).$$

Die Regressionsgerade von y auf x kann man analog bestimmen.

Der minimale Wert der Abweichungsquadrate der Daten von der Regressionsgeraden Σd^2 ist nun:

$$\Sigma d^2 = \Sigma Y^2 - \frac{(\Sigma XY)^2}{\Sigma X^2}$$

Dies ist eine Summe von Quadraten, daher folgt:

$$\Sigma Y^2 - \frac{(\Sigma XY)^2}{\Sigma X^2} \geq 0 \qquad \text{d.h.} \qquad \Sigma Y^2 \geq \frac{(\Sigma XY)^2}{\Sigma X^2} \quad \text{bzw.}$$

$$\frac{(\Sigma XY)^2}{\Sigma X^2 \Sigma Y^2} \leq 1, \qquad \text{also:} \qquad -1 \leq \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}} \leq 1$$

Die Definition des Produkt-Moment-Korrelationskoeffizienten mit

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}} = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}}$$

folgt daraus in natürlicher Weise, wobei $r = \pm 1$ gilt, falls $\Sigma d^2 = 0$ (was wiederum bedeutet, daß die Punkte (x,y) alle auf einer Geraden liegen), und $r = 0$ gilt, falls die Kovarianz, cov (x,y) gleich Null ist (d.h. die Punkte sind "gleichmäßig" verstreut in den vier Quadranten zum Punkt $(\bar{x},\bar{y})$.