G.C. Barozzi: Primo Corso di Analisi Matematica, Zanichelli (Bologna), 1998
ISBN 88-08-01169-0

Complemento all'esempio 4.4-8: la retta dei minimi quadrati

Riprendiamo in considerazione gli *errori* (o *scarti*, oppure ancora *residui*)

$$e_i = y_i - \hat{y}_i = y_i - mx_i - q,$$

dove s'intende che m e q sono i valori calcolati in base alle (9) e (10). Si ha

$$\sum_{i=1}^{n} e_i = 0, \tag{$*$}$$

$$\sum_{i=1}^{n} e_i\,\hat{y}_i = 0. \tag{$**$}$$

La (*) è conseguenza immediata della formula

$$\frac{d}{dq}\sum_{i=1}^{n}(y_i - mx_i - q)^2 = -2\sum_{i=1}^{n}(y_i - mx_i - q) = -2\sum_{i=1}^{n} e_i = 0,$$

equivalente alla (9). Ne segue che $\overline{y}$ non è soltanto la media aritmetica delle "ordinate sperimentali" y_i ma anche la media delle "ordinate stimate" dal modello, cioè le $\hat{y}_i$:

$$\sum_{i=1}^{n} e_i = 0 \quad \Longleftrightarrow \quad \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i \quad \Longleftrightarrow \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i.$$

Quanto alla (**), riprendiamo il calcolo che ci ha condotto alla (10). Da $s = \sum_i e_i^2 = \sum_i (y_i - mx_i - q)^2$, derivando rispetto ad m e uguagliando a 0 si trova

$$-2m\sum_{i=1}^{n}(y_i - mx_i - q)\,x_i = 0 \quad \Longleftrightarrow \quad \sum_{i=1}^{n} e_i\,x_i = 0.$$

Ne segue, combinando l'ultima auguaglianza con la (*),

$$\sum_{i=1}^{n} e_i\,\hat{y}_i = \sum_{i=1}^{n} e_i\,(mx_i + q) = m\sum_{i=1}^{n} e_i\,x_i + q\sum_{i=1}^{n} e_i = 0.$$

Un modo equivalente per formulare l'uguaglianza (**) si ottiene considerando il vettore n-dimensionale degli errori e_i, $i = 1, 2, \ldots, n$, cioè quello avente come componenti le differenze $y_i - \hat{y}_i$, ed il vettore degli scarti delle ordinate stimate dalla loro media, dunque il vettore di componenti $\hat{y}_i - \overline{y}$; questi vettori sono *ortogonali* tra loro, nel senso che la somma dei prodotti delle componenti di uguale indice (il loro *prodotto scalare*) è nullo. Infatti

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \sum_{i=1}^{n} e_i\,(\hat{y}_i - \overline{y}) = \sum_{i=1}^{n} e_i\,\hat{y}_i - \overline{y}\sum_{i=1}^{n} e_i = 0. \tag{$***$}$$

Un misura della variabilità delle ordinate y_i è data dalla loro *devianza*, cioè dalla somma dei quadrati degli scarti dalla media:

$$\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

Se scriviamo ciascun scarto facendo intervenire l'ordinata stimata $\hat{y}_i$, abbiamo la seguente espressione per la devianza:

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}[(y_i-\hat{y}_i)+(\hat{y}_i-\overline{y})]^2=$$
$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\overline{y})=$$
$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2,$$

in quanto la terza somma nella penultima riga è nulla in forza della (***).

Dunque la devianza delle y_i si spezza in due somme: la devianza delle $\hat{y}_i$, cioè la devianza "spiegata dal modello", ed una somma residua, $\sum_i(y_i-\hat{y}_i)^2$, che non è "spiegata" dal modello della retta dei minimi quadrati. Il modello è tanto migliore quanto più questa seconda parte è una piccola frazione della devianza totale.

Ciò induce a scegliere come indice della bontà del modello il rapporto tra la devianza spiegata e la devianza totale, dunque il rapporto

$$R^2:=\frac{\sum_i(\hat{y}_i-\overline{y})^2}{\sum_i(y_i-\overline{y})^2}.$$

Evidentemente R^2 è compreso tra 0 e 1: il modello è tanto migliore quanto più R è prossimo a 1.

Possiamo dare un'espressione diversa ad R^2 se ricordiamo che le differenze $y_i-\overline{y}$ si scrivono $m(x_i-\overline{x})$ e successivamente utilizziamo il valore di m fornito dalla formula (10) del testo. Otteniamo

$$R^2=\frac{\sum_i(\hat{y}_i-\overline{y})^2}{\sum_i(y_i-\overline{y})^2}=m^2\,\frac{\sum_i(x_i-\overline{x})^2}{\sum_i(y_i-\overline{y})^2}=$$
$$=\left[\frac{\sum_i(x_i-\overline{x})(y_i-\overline{y})}{\sum_i(x_i-\overline{x})^2}\right]^2\,\frac{\sum_i(x_i-\overline{x})^2}{\sum_i(y_i-\overline{y})^2}=$$
$$=\frac{\left[\sum_i(x_i-\overline{x})(y_i-\overline{y})\right]^2}{\sum_i(x_i-\overline{x})^2\ \sum_i(y_i-\overline{y})^2}.$$

Dunque R^2 può essere considerato come il quadrato del rapporto (compreso tra -1 e 1)

$$r:=\frac{\sum_i(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_i(x_i-\overline{x})^2}\,\sqrt{\sum_i(y_i-\overline{y})^2}};$$

esso viene chiamato *coefficiente di correlazione* tra le x_i e le y_i. Nello spazio n-dimensionale esso può essere interpretato come il coseno dell'angolo formato tra il vettore di componenti $x_i-\overline{x}$ e quello di componenti $y_i-\overline{y}$. Evidentemente si ha $|r|=R$.

Se $r>0$ si ha una *correlazione positiva* tra le x_i e le y_i (intuitivamente: le ordinate crescono al crescere delle ascisse); il contrario accade se il coefficiente di correlazione è negativo.

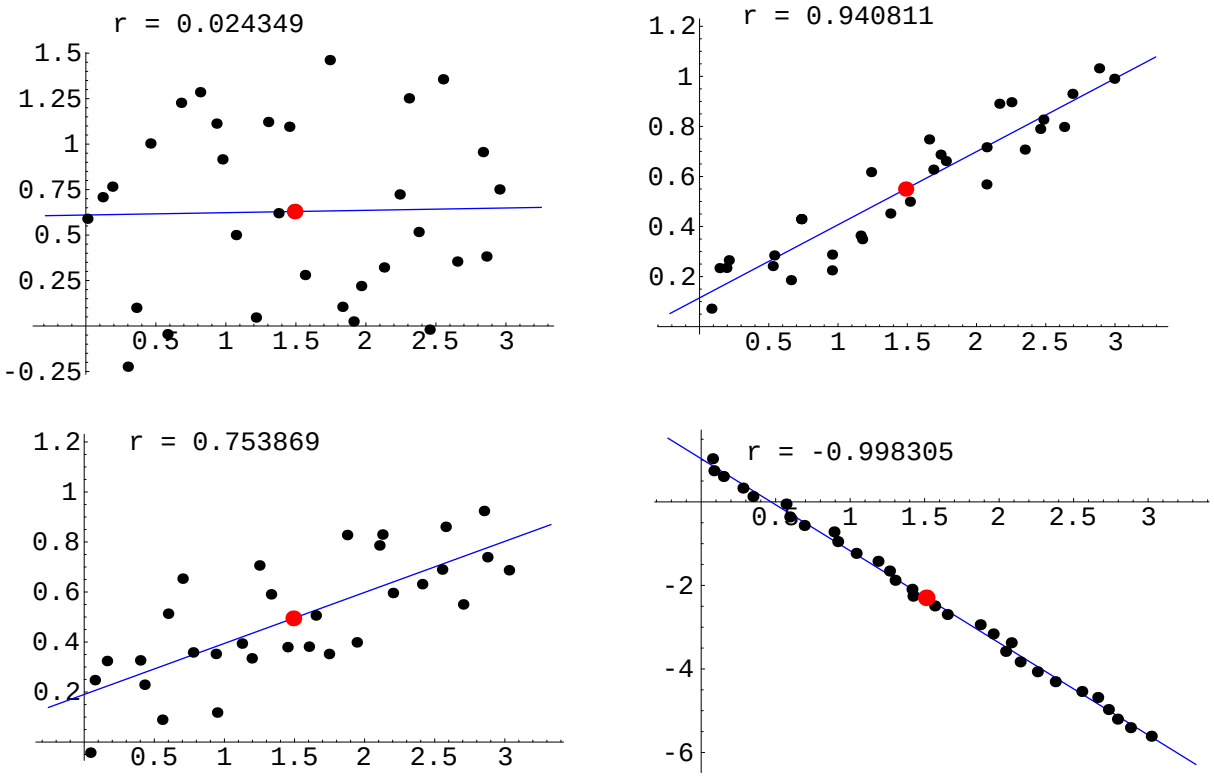
Ricordiamo al lettore (v. Laboratorio 4.4-1 a pag. 532 del testo) che le quantità a numeratore e denominatore del coefficiente di correlazione si possono più agevolmente calcolare mediante le identità

$$\sum_i(x_i-\overline{x})^2=\sum_i x_i^2-\frac{1}{n}\left(\sum_i x_i\right)^2,\quad \sum_i(y_i-\overline{y})^2=\sum_i y_i^2-\frac{1}{n}\left(\sum_i y_i\right)^2,$$

$$\sum_i(x_i-\overline{x})(y_i-\overline{y})=\sum_i x_iy_i-\frac{1}{n}\sum_i y_i\sum_i y_i.$$

Per concludere, osserviamo che le due variabili x e y giocano due ruoli completamente distinti. La retta che abbiamo calcolato va sotto il nome di *retta di regressione* di y rispetto a x, secondo una terminologia che risale al biologo inglese Francis Galton (1822-1911).

Essa viene utilizzata per stimare i valori delle ordinate y a partire da valori misurati della x. Si pensi ad una categoria di pazienti su cui è necessario rilevare un dato clinico y, di difficile misurazione: se esso è "fortemente correlato" con un dato x di facile misurazione (nel senso che la retta di regressione di y rispetto



ad x, sulla base di dati sperimentali ottenuti suun campione di pazienti, presenta un valore di R prossimo a 1), allora è conveniente una misura indiretta di y a partire da una misura diretta di x.

Si dice anche che la variabile x gioca il ruolo di "predittore".