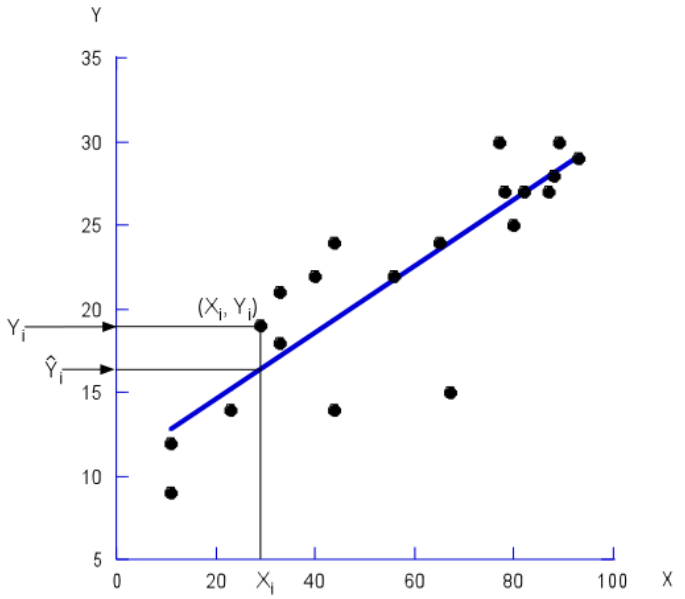
Psicometria 2
1. Regressione lineare bivariata

Michele Grassi

Dipartimento di Psicologia, Università di Trieste

Modulo 2, 4 C.F.U.

Marzo – Aprile 2012



La regressione

Il problema della regressione si pone quando il valore di una variabile aleatoria *y*, chiamata *variabile dipendente*, è funzione di una variabile non aleatoria *x*, chiamata *variabile indipendente*.

Qui ci soffermeremo su un'unica classe di modelli, detti *modelli statistici lineari*.

Si veda la seguente figura:

La regressione

Per i dati della figura, ad esempio, il modello della regressione lineare che mette in relazione il valore della variabile dipendente con quello della variabile indipendente è

$$y = \alpha + \beta x + \varepsilon$$

Qui α è l'*intercetta*, β la *pendenza* della retta, e ε una variabile aleatoria con una specifica distribuzione di probabilità con media uguale a zero.

Abbiamo così la somma di una componente deterministica, completamente predicibile dalla variabile indipendente, e di una aleatoria, ε.

Regressione e principio dei minimi quadrati

Qui, per ogni *i*-esimo soggetto, i valori predetti della retta di regressione sono

$$\hat{y}_i = \alpha + \beta x_i$$

e i valori osservati in funzione dei coefficienti α e β sono

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

Assumendo una distribuzione normale (univariata) per le osservazioni y_i, la stima di *massima verosimiglianza* dei coefficienti si ottiene massimizzando la seguente funzione:

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2}\left(y_i - (\alpha + \beta x_i)\right)^2 / \sigma^2\right\};$$

ossia minimizzando le differenze y_i - $\hat{y}_i$.

Regressione e principio dei minimi quadrati

La funzione di cui occorre trovare il minimo è

$$SQ_{err} = \sum \left(y_i - a - bx_i\right)^2,$$

che sviluppata dà

$$SQ_{err} = \sum_{i=1}^{n} \left(y_i^2 + a^2 + b^2 x_i^2 - 2ay_i - 2by_i x_i + 2abx_i\right),$$

e le cui derivate parziali rispetto ad *a* e *b* sono:

$$\frac{\delta SQ_{err}}{\delta a} = \sum_{i=1}^{n} \left(2a - 2y_i + 2bx_i\right) = 2\left(-\sum_{1}^{n} y_i + na + b\sum_{1}^{n} x_i\right);$$

$$\frac{\delta SQ_{err}}{\delta b} = \sum_{i=1}^{n} \left(2bx_i^2 - 2y_i x_i + 2ax_i\right) = 2\left(-\sum_{1}^{n} y_i x_i + b\sum_{1}^{n} x_i^2 + a\sum_{1}^{n} x_i\right).$$

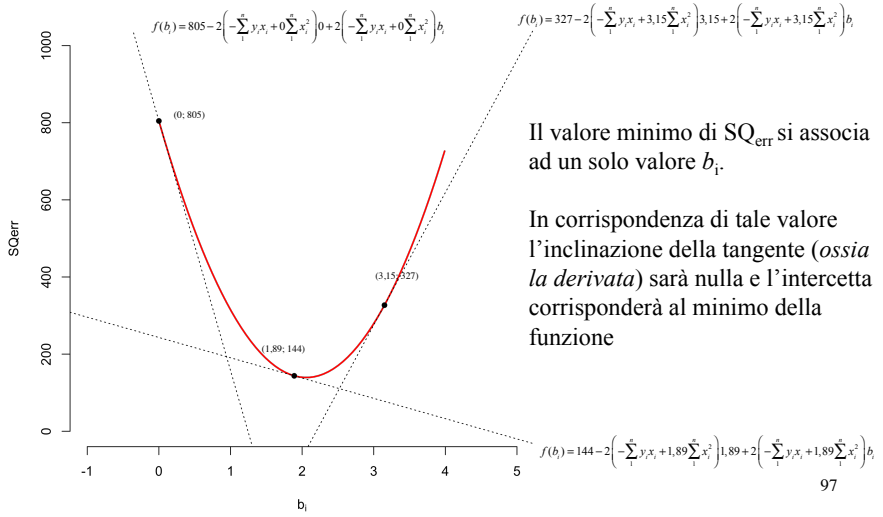
Regressione e principio dei minimi quadrati

Anticipando che il coefficiente *a* (*o intercetta)* assumerà valore zero se si lavora con gli scarti dalla media, possiamo analizzare l'andamento di SQ_{err} in funzione dei possibili valori del coefficiente *b*.
Il valore b_i in corrispondenza del quale troveremo il minimo $SQerr_i$, sarà la nostra stima di massima verosimiglianza *(confronta figura successiva).*

- La derivata parziale in *b* rappresenta l'*inclinazione* della (retta) tangente la funzione per un determinato punto ($SQerr_i$, b_i);
- l'*intercetta* della tangente si troverà per differenza:

$$SQ_{err_i} - \frac{\delta SQ_{err}}{\delta b} b_i.$$

Regressione e principio dei minimi quadrati



Il valore minimo di SQ_{err} si associa ad un solo valore b_i.

In corrispondenza di tale valore l'inclinazione della tangente (*ossia la derivata*) sarà nulla e l'intercetta corrisponderà al minimo della funzione

Regressione e principio dei minimi quadrati

Quindi, ponendo entrambe le derivate per a e b uguali a 0, otteniamo:

$$\left(\frac{\delta SQ_{err}}{\delta a} = 0\right)^* \quad \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i.$$

$$\left(\frac{\delta SQ_{err}}{\delta b} = 0\right) \quad \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2.$$

* da cui si ricava facilmente la soluzione $\bar{y} - b\bar{x} = a$.

Regressione e principio dei minimi quadrati

Scrivendo in forma di matrice queste equazioni,

$$\begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix};$$

i coefficienti a e b possono essere trovati come:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i \\ \sum y_i x_i \end{bmatrix}$$

$$= \frac{1}{n\sum x_i^2 - \left(\sum x_i\right)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum y_i x_i \end{bmatrix}$$

Regressione e principio dei minimi quadrati

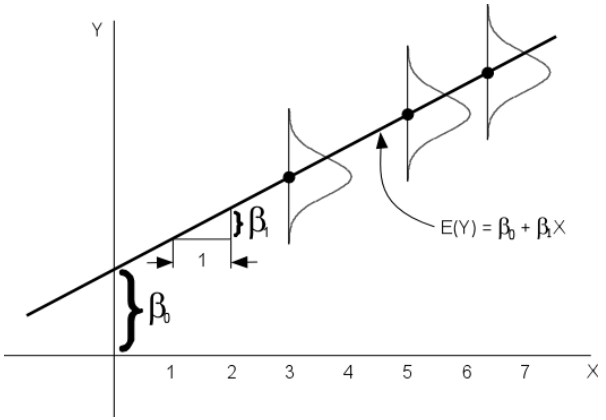
Di qui:

$$a = \frac{\sum y_i}{n} - b\frac{\sum x_i}{n};$$

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}.$$

La regressione

Si osservi ora questa figura



Modello probabilistico della regressione

La figura fornisce una rappresentazione grafica del modello probabilistico di $y = a + bx + e$.

In corrispondenza di ogni valore di x c'è una popolazione di possibili valori y.

Questa popolazione ha media $E(y)$ e varianza σ_y^2.

Per ciascuno dei possibili valori che X può assumere varierà solo $E(y)$, ma non la varianza σ_y^2.

Il modello probabilistico della regressione assume dunque che vi sia una diversa popolazione di risposte (y) per ciascuno dei valori della variabile indipendente (x), la cui media dipende dal valore della variabile indipendente ed è predetta dal modello della regressione.

I coefficienti di regressione

I coefficienti di regressione possono essere considerati delle *variabili aleatorie* in quanto assumono valori diversi in campione diversi.

Poniamoci quindi il problema di determinare le proprietà della *distribuzione campionaria* di a e b calcolati con il metodo dei minimi quadrati.

Iniziamo riarrangiando la formula di b:

$$b = \frac{S_{xy}}{S_x^2} = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{\sum(x_i - \bar{X})^2} =$$

$$= \frac{\sum(x_i - \bar{X})y_i - \bar{Y}\cancel{\sum(x_i - \bar{X})}}{\sum(x_i - \bar{X})^2} = \frac{\sum(x_i - \bar{X})y_i}{\sum(x_i - \bar{X})^2}.$$

I coefficienti di regressione

In questo modo, il valore atteso di b diventa:

$$E(b) = E\left(\frac{\sum(x_i - \bar{X})y_i}{\sum(x_i - \bar{X})^2}\right) = E\left(\frac{\sum(x_i - \bar{X})(\alpha + \beta x_i + e_i)}{\sum(x_i - \bar{X})^2}\right) =$$

$$= \alpha\frac{\sum(x_i - \bar{X})}{\sum(x_i - \bar{X})^2} + \beta\frac{\sum(x_i - \bar{X})x_i}{\sum(x_i - \bar{X})^2} + \left(\frac{E(x_i e_i)}{\sum(x_i - \bar{X})^2} - \frac{\bar{X}E(e_i)}{\sum(x_i - \bar{X})^2}\right) = \beta.$$

dato che:

$$E(x_i e_i) = \frac{1}{n}\sum_i\left(x\left(y - \alpha - \beta x_i\right)\right) = \frac{1}{n}\left(\sum_i(xy) - \alpha\sum_i x_i - \beta\sum_i x_i^2\right) = 0$$

$$E(e_i) = \frac{1}{n}\sum_i\left(y - \alpha - \beta x_i\right) = \frac{1}{n}\sum_i\left(y - \bar{y} + \beta\bar{x}_i - \beta x_i\right) = \frac{1}{n}\left(\sum_i(y - \bar{y}) - \beta\sum_i(x_i - \bar{x})\right) = 0$$

Si dimostra quindi che b è uno stimatore privo di errore sistematico di β: $E(b) = \beta$.

In modo analogo si può mostrare che $E(a) = \alpha$.

I coefficienti di regressione

Si dimostra inoltre che

$$V(b) = \frac{\sigma_\varepsilon^2}{\sum(x_i - \bar{X})^2}$$

$$V(a) = \frac{\sigma_\varepsilon^2\sum x_i^2}{n\sum(x_i - \bar{X})^2}$$

dove

$$\sigma_\varepsilon^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}.$$

I coefficienti di regressione

Si dimostra inoltre che

$$V(b) = \frac{\sigma_\varepsilon^2}{\sum (x_i - \bar{X})^2}$$

$$V(a) = \frac{\sigma_\varepsilon^2 \sum x_i^2}{n \sum (x_i - \bar{X})^2}$$

Se le osservazioni Y_i sono distribuite normalmente, allora anche a e b lo saranno.

In conclusione,

$$b \sim N\left(\beta, \frac{\sigma_\varepsilon^2}{\sum (x_i - \bar{X})^2}\right)$$

$$a \sim N\left(\alpha, \frac{\sigma_\varepsilon^2 \sum x_i^2}{n \sum (x_i - \bar{X})^2}\right)$$