REGRESSIONE LINEARE E POLINOMIALE

Nota una tabella di dati relativi alle osservazioni di due grandezze X e Y, è naturale formulare ipotesi su quale possa essere una ragionevole funzione che rappresenti o che approssimi la relazione tra X e Y. Il metodo dei minimi quadrati è una risposta largamente condivisa a tale problema. Di seguito si presenta inizialmente il modello più semplice di regressione, quello lineare, per trattare poi casi più complessi in cui il modello di regressione è di tipo polionomiale.

Il modello classico di regressione lineare semplice

Il modello di *regressione lineare semplice* suppone una relazione lineare tra *x* e *y*, ovvero

$$y = \beta_1 + \beta_2 x + e \qquad (1)$$

dove $ß_1$ e $ß_2$ sono i parametri della cosiddetta *retta di regressione*, i quali devono essere opportunamente valutati sulla base delle osservazioni ed *e* rappresenta un termine d'errore.

Le ipotesi del modello classico di regressione lineare semplice implicano che la y_i sia costituita dalla somma di una componente deterministica $ß_1 + ß_2 x_i$ e una termine di scarto e_i

$$y_i = \beta_1 + \beta_2 x_i + e_i \qquad (2)$$

infatti i valori x_i della variabile esplicativa sono fissati e $ß_1$e $ß_2$ sono parametri e quindi costanti.

Per stimare i parametri $ß_1$ e $ß_2$ del modello di regressione si considera un campione costituito da n coppie di valori (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), dove $x_1, x_2, ..., x_n$ sono i valori degli ingressi e $y_1, y_2, ..., y_n$ sono i valori delle osservazioni. Le osservazioni possono essere rappresentate in un grafico a dispersione come esemplificato nella Figura 1.

Figura 1. Osservazioni sul modello di regressione

Le stime sono costituite dai valori dei parametri cui corrisponde la retta che approssima al meglio i dati. A tal fine si considerano le distanze dei punti (x_i, y_i) dalla retta di regressione, ovvero gli scarti

$$e_i = y_i - (\beta_1 + \beta_2 x_i) \qquad \text{per } i = 1,2,\dots,n \qquad (3)$$

e i valori di ß$_1$ e ß$_2$ sono scelti in modo tale da minimizzare le distanze dei punti (x_i, y_i) dalla retta di regressione stimata. Poiché alcune distanze sono positive e altre negative, si considera la somma delle distanze al quadrato

$$Q(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2 \qquad (4)$$

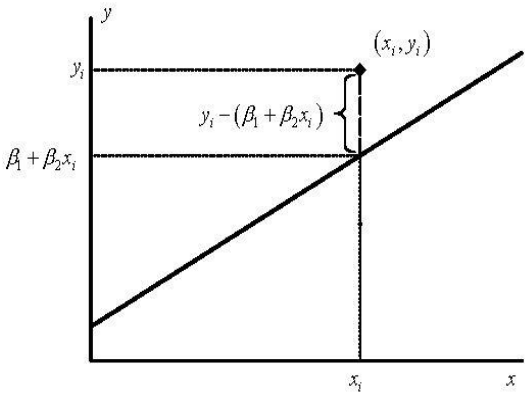

Figura 2. Distanza del punto (x_i, y_i) dalla retta di regressione

Le stime del metodo dei minimi quadrati sono quei valori ß$_i$ che minimizzano la somma dei quadrati delle distanze, per le quali cioè si ha

$$Q(\widehat{\beta_1}, \widehat{\beta_2}) = \min_{\beta_1, \beta_2} Q(\beta_1, \beta_2) \qquad (5)$$

Derivando si ottiene

$$\frac{\partial Q(\beta_1, \beta_2)}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)$$

$$\frac{\partial Q(\beta_1, \beta_2)}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i) x_i$$

(6)

Ponendo le derivate uguali a zero e dividendo entrambi i membri per -2 si ottengono le equazioni:

$$\sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i) = 0$$

$$\sum_{i=1}^{n}(y_i-\beta_1-\beta_2 x_i)x_i = 0 \qquad (7)$$

Distribuendo la sommatoria nella prima equazione si ha

$$\sum_{i=1}^{n} y_i - n\widehat{\beta_1} - \widehat{\beta_2}\sum_{i=1}^{n} x_i = n\bar{y} - n\hat{\beta}_1 - n\hat{\beta}_2\bar{x} = 0 \qquad (8)$$

dove $\bar{x}$ e $\bar{y}$ sono le medie aritmetiche dei valori degli ingressi e delle osservazioni. Dividendo per n e risolvendo rispetto a $\hat{\beta}_1$ si ottiene la stima dell'intercetta:

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x} \qquad (9)$$

e quindi

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})x_i} \qquad (10)$$

La retta di regressione stimata ha la seguente equazione

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i \qquad (11)$$

Essa passa per il punto di coordinate $(\bar{x},\bar{y})$; infatti dalla formula della stima dell'intercetta si ha

$\bar{y} = ß_1 + ß_2\bar{x}$

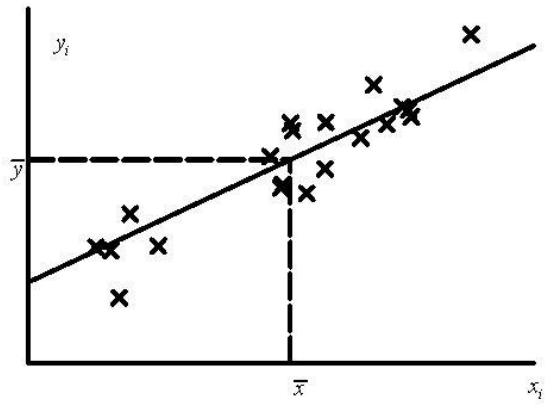

Figura 3. La retta stimata passa per il punto di coordinate (x, y)

Notazione vettoriale della soluzione dei minimi quadrati

Si consideri la somma delle distanze al quadrato:

$$\varepsilon^2 = [\, y - (\beta_1 1 + \beta_2 x)] \qquad (12)$$

dove: y corrisponde al vettore dei dati misurati;

$\beta_1 1 + \beta_2 x$ rappresenta la retta di regressione;

β_1 e β_2 rappresentano i parametri di regressione lineare (vettori colonna);

N è la lunghezza del vettore;

La (12) diventa:

$$\varepsilon^2 = [\, y - (\beta_1 1 + \beta_2 x)\,]^T [y - (\beta_1 1 + \beta_2 x)] =$$

$$y^T y - \beta_2 y^T x - \beta_1 y^T 1 + \beta_2{}^2 x^T x - \beta_2 x^T y + \beta_2\beta_1 x^T 1 - \beta_1 1^T y + \beta_2\beta_1 1^T x + \beta_1{}^2 1^T 1 =$$

$$y^T y - \beta_2 y^T x - \beta_1 y^T 1 + \beta_2{}^2 x^T x - \beta_2 (x^T y)^T + \beta_2\beta_1 x^T 1 - \beta_1 (1^T y)^T + +\beta_2\beta_1 (1^T x)^T + \beta_1{}^2 1^T 1 =$$

$$y^T y - 2\beta_1 y^T 1 - 2\beta_2 y^T x + \beta_2{}^2 x^T x + \beta_1{}^2 N + 2\beta_2\beta_1 x^T 1$$

I coefficienti β_2 e β_1 si determinano in modo da minimizzare ε^2, ponendo a zero le derivate parziali di ε^2 rispetto ad β_2 e rispetto a β_1 :

$$\begin{cases} \frac{\partial\, \varepsilon^2}{\partial \beta_2} = -2\, y^T x + 2\beta_2\, x^T x + 2\beta_1\, x^T 1 = 0 \\ \frac{\partial\, \varepsilon^2}{\partial \beta_1} = -2\, y^T 1 + 2\beta_2 x^T 1 + 2\beta_1 N = 0 \end{cases}$$

$$\begin{cases} \beta_2\, x^T x + \beta_1\, x^T 1 = y^T x \\ \beta_2 x^T 1 + \beta_1 N = y^T 1 \end{cases}$$

$$\beta_2 = \frac{\begin{vmatrix} y^T x & x^T 1 \\ y^T 1 & N \end{vmatrix}}{\begin{vmatrix} x^T x & x^T 1 \\ x^T 1 & N \end{vmatrix}} = \frac{N\, y^T x - (x^T 1)(y^T 1)}{N x^T x - x^T 1} = \frac{N \sum_k x_k y_k - \sum_k x_k * \sum_k y_k}{N \sum_{k=1}^{N} x_k^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\beta_1 = \frac{\begin{vmatrix} x^T x & y^T x \\ x^T 1 & y^T 1 \end{vmatrix}}{\begin{vmatrix} x^T x & x^T 1 \\ x^T 1 & N \end{vmatrix}} = \frac{(x^T x)(y^T 1) - (y^T x)(x^T 1)}{N x^T x - x^T 1} = \frac{\sum_k x_k^2 * \sum_k y_k - \sum_k x_k y_k \sum_k x_k}{N \sum_{k=1}^N x_k^2 - (\sum_{i=1}^n x_i)^2}$$

oppure si ricava β_2 in funzione di β_1 :

$$\beta_2 = \frac{y^T 1 - \beta_1 x^T 1}{N}$$

Le espressioni simboliche sembrano inguardabili, ma si possono riscrivere in una forma più leggibile. Se si indica con

- $\overline{X}$ la media aritmetica di $x = \{x_i, \dots, x_n\}$ cioè $\frac{1}{n}\sum_{i=1}^n x_i$
- $\overline{Y}$ la media aritmetica di $y = \{y_i, \dots, y_n\}$ cioè $\frac{1}{n}\sum_{i=1}^n y_i$
- $\overline{X^2}$ la media aritmetica di $x^2 = \{x_i^2, \dots, x_n^2\}$ cioè $\frac{1}{n}\sum_{i=1}^n x_i^2$
- $\overline{XY}$ la media aritmetica di $xy = \{x_i y_i, \dots, x_n y_n\}$ cioè $\frac{1}{n}\sum_{i=1}^n x_i y_i$

allora dividendo per n il sistema diventa

$$\begin{cases} \beta_2 \overline{X^2} + \beta_1 \overline{X} = \overline{XY} \\ \beta_2 \overline{X} + \beta_1 = \overline{Y} \end{cases} \qquad (13)$$

La seconda equazione mette in luce che il baricentro $(\overline{X}, \overline{Y})$ cioè il punto le cui coordinate sono la media delle ascisse e la media delle ordinate, appartiene alla retta di regressione, perché soddisfa l'equazione $y = \beta_1 + \beta_2 x$.

Si può ora scrivere la soluzione mediante un'espressione simbolica più semplice:

$$\beta_2 = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2} \qquad (14)$$

$$\beta_1 = \overline{Y} - \beta_2 \overline{X} \qquad (15)$$

Regressione lineare multipla

Nel modello di regressione semplice le variazioni delle osservazioni sono spiegate mediante una sola variabile d'ingresso. Si ottiene così un modello molto semplice che tuttavia non è sempre in grado di spiegare i fenomeni di interesse in maniera adeguata.

Un modello di regressione multipla spiega la variabile dipendente *y* in funzione di k variabili esplicative o *regressori*, con k > 2,

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon \qquad (16)$$

Per convenzione la prima variabile esplicativa e costante $x_1 = 1$. Il primo coefficiente di regressione β_1 rappresenta quindi l'intercetta del modello.

Il modello di regressione multipla può essere rappresentato in termini matriciali:

Si consideri un campione di numerosità n sul modello di regressione lineare multipla

$$y = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + \varepsilon_i \qquad \text{per } i = 1,2,\ldots n. \qquad (17)$$

Sia Y un vettore le cui componenti sono costituite dalle n variabili y_1, y_2, ..., y_n e X una matrice di dimensioni n × k con i valori delle variabili esplicative,

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \; X = \begin{bmatrix} 1 & x_{21} & \cdots & x_{k1} \\ 1 & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2n} & \cdots & x_{kn} \end{bmatrix} \qquad (18)$$

Nella matrice X ogni colonna corrisponde ad un regressore: la prima colonna ha tutti elementi unitari, la seconda contiene i valori osservati di x_2 e così via fino all'ultima colonna che contiene i valori di x_k. Si definisce quindi il vettore β, di dimensione n, con i parametri del modello di regressione e il vettore ε, di dimensioni n, con gli scarti

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad (19)$$

In forma matriciale il problema è rappresentato da

$$Y = X\beta + \varepsilon \qquad (20)$$

che corrisponde a

$$y_1 = \beta_1 + \beta_2 x_{21} + \beta_3 x_{31} + \cdots \beta_k x_{k1} + \varepsilon_1$$

$$y_2 = \beta_1 + \beta_2 x_{22} + \beta_3 x_{32} + \cdots \beta_k x_{k2} + \varepsilon_2$$

$$\ldots \qquad (21)$$

$$y_n = \beta_1 + \beta_2 x_{2n} + \beta_3 x_{3n} + \cdots \beta_k x_{kn} + \varepsilon_n$$

Regressione polinomiale

Si propone di descrivere alcune curve di adattamento con il metodo dei minimi quadrati e di fornire un metodo iterativo per generalizzare tali funzioni a polinomi di grado M.

Spesso si può esprimere una variabile, y, come polinomiale di una seconda variabile x:

$$y = A + Bx + C x^2 + \ldots + Z x^M + \varepsilon \qquad (22)$$

Si supponga per esempio di avere una polinomiale di forma quadratica,

$$y = A + Bx + C x^2 + \varepsilon \qquad (23)$$

nota una serie di valori (x_i, y_i), i = 1,..,N per ogni x_i il valore y_i si ottiene dalla (23) dove A,B,C sono ancora incognite.

La miglior stima per *A,B,C* è data da quei valori per cui la sommatoria degli scarti quadratici ε^2 (in notazione vettoriale) è minima. Si differenzi quindi ε^2 rispetto a *A,B,C* :

$$\frac{\partial \varepsilon^2}{\partial A} = \sum_{i=1}^{N} (2A - 2y_i + 2Bx_i + Cx_i^2)$$

$$\frac{\partial \varepsilon^2}{\partial B} = \frac{1}{\sigma_y^2} \sum_{i=1}^{N} (2Bx_i^2 - 2x_i y_i + 2Ax_i + Cx_i^3)$$

$$\frac{\partial \varepsilon^2}{\partial C} = \frac{1}{\sigma_y^2} \sum_{i=1}^{N} (2C x_i^4 - 2y_i x_i^2 + 2A x_i^2 + 2Bx_i^3) \qquad (24)$$

ponendo uguale a zero, omettendo gli estremi di sommatoria:

$$\sum y_i = AN + B \sum x_i + C \sum x_i^2$$

$$\sum y_i x_i = A \sum x_i + B \sum x_i^2 + C \sum x_i^3$$

$$\sum y_i x_i^2 = A \sum x_i^2 + B \sum x_i^3 + C \sum x_i^4 \qquad (25)$$

si tratta si un sistema 3x3 che in forma matriciale diventa

$$\begin{bmatrix} N & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i x_i \\ \sum y_i x_i^2 \end{bmatrix} \qquad (26)$$

che è del tipo

$$A\,X = B \qquad (27)$$

e si può risolvere in diversi modi tra cui:

$$X = A^{-1} B \qquad (28)$$

Trovati i valori di *A,B,C* si sostituiscono nella (23). In generale volendo ricavare una polinomiale di grado *M* si avranno *M+1* equazioni in *M+1* incognite, la matrice *A* sarà di dimensiioni *(M+1)x(M+1),* i vettori *B* e *X* avranno lunghezza *(M+1).* I coefficienti della matrice da invertire e i termini noti del problema risultano quelli del sistema:

$$\begin{bmatrix} N & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^n \\ \sum x_i & \sum x_i^2 & \cdots & \cdots & \sum x_i^{n+1} \\ \sum x_i^2 & \sum x_i^3 & \cdots & \cdots & \sum x_i^{n+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_i^n & \sum x_i^{n+1} & \cdots & \cdots & \sum x_i^{2n} \end{bmatrix} \begin{bmatrix} A \\ B \\ C \\ \vdots \\ Z \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i x_i \\ \sum y_i x_i^2 \\ \vdots \\ \sum y_i x_i^n \end{bmatrix} \qquad (29)$$

Risolvendo tale sistema si ottengono i valori *A,B,C,...,Z* da sostituire nella (22), questa equazione rappresenta la regressione polinomiale di adattamento ai dati (x_i,y_i).