La retta di regressione + Esercizio Tipo 2

Analisi matematica e Calcolo delle probabilità
corso di laurea SIA, a.a. 2014/2015
G.Parmeggiani

October 24, 2014

1 La retta di regressione

Supponiamo che nello studio dell'eventuale dipendenza tra due variabili x ed y, i punti corrispondenti alle singole osservazioni si distribuiscano "quasi" su di una retta nel piano Oxy. Si cerca una retta che "più si avvicina" a tutti i punti.

Fissato nel piano un sistema di riferimento ortogonale e monometrico Oxy, siano

$$(x_1, y_1), \quad (x_2, y_2), \quad \dots \quad , (x_n, y_n)$$

le coordinate dei punti

$$P_1, \quad P_2, \quad \dots \quad , P_n.$$

Siano r una retta nel piano e $Q_1, Q_2, \dots, Q_n$ i punti di r di ascisse $x_1, x_2, \dots, x_n$ rispettivamente (ossia di ascisse uguali a quelle dei punti $P_1, P_2, \dots, P_n$). Dunque se

$$y = mx + q$$

è l'equazione di r, le coordinate dei punti $Q_1, Q_2, \dots, Q_n$ sono:

$$(x_1, mx_1 + q), \quad (x_2, mx_2 + q), \quad \dots \quad , (x_n, mx_n + q).$$

Per ogni $i = 1, \dots, n$, indichiamo con $|P_iQ_i|$ la distanza tra il punto P_i ed il punto Q_i:

$$\begin{aligned} |P_iQ_i| &= \text{distanza tra il punto } P_i \text{ ed il punto } Q_i \\ &= \text{lunghezza del segmento (verticale) che congiunge i punti } P_i \text{ e } Q_i \\ &= |y_i - (mx_i + q)| = |y_i - mx_i - q|. \end{aligned}$$

Si prova che **se i punti $P_1, P_2, \ldots, P_n$ non sono tutti allineati verticalmente** (ossia si prova che se esistono $i, j \in \{1, \ldots, n\}$ con $i \neq j$ tali che $x_i \neq x_j$) allora

1. esiste un'unica retta r, detta **retta di regressione lineare** o **retta dei minimi quadrati**, tale che sia minima la somma dei quadrati delle lunghezze dei segmenti $|P_iQ_i|$, ossia tale che sia minima

$$\sum_{i=1}^{n} |P_iQ_i|^2$$

2. tale retta si ottiene prendendo m e q soluzioni del seguente sistema (detto **sistema delle equazioni normali**):

$$\begin{cases} m \cdot (\sum_{i=1}^n x_i) + \quad n \cdot q \quad = \sum_{i=1}^n y_i \\ m \cdot (\sum_{i=1}^n x_i^2) + q \cdot (\sum_{i=1}^n x_i) = \sum_{i=1}^n x_i y_i \end{cases}$$

N.B. Mentre esiste un'unica retta $y = mx + q$ che rende minima $\sum_{i=1}^n |y_i - mx_i - q|^2$, può esistere più di una retta $y = mx + q$ che rende minima $\sum_{i=1}^n |y_i - mx_i - q|$, e può esistere più di una retta $y = mx + q$ che rende minima $\sum_{i=1}^n (y_i - mx_i - q)$.

2 ESERCIZIO TIPO 2

(a) Determinare la retta di regressione per i dati riportati nella seguente tabella:

x_i	y_i
-2	8
0	2
2	0
3	$-\frac{8}{3}$

(b) Pesando un gattino ogni tre settimane si ottengono i suguenti valori:

età	**peso**
3 settimane	$0,5$ Kg
6 settimane	$0,7$ Kg
9 settimane	$1,2$ Kg
12 settimane	$1,8$ Kg
15 settimane	$2,2$ Kg

Prendendo come variabile x l'età del gattino, con unità di misura uguale a 3 settimane e come variabile y il peso del gattino, con unità di misura uguale ad 1 etto, si scriva l'equazione della retta di regressione dei dati.

(a) La retta di regressione ha equazione del tipo $y = mx + q$ dove m e q sono le soluzioni del sistema delle equazioni normali

$$\begin{cases} m \cdot \left(\sum_{i=1}^{n} x_i\right) + \quad n \cdot q \quad = \sum_{i=1}^{n} y_i \\ m \cdot \left(\sum_{i=1}^{n} x_i^2\right) + q \cdot \left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Per nostra comodità aggiungiamo alla tabella che ha come colonne i valori delle x_i e delle y_i, altre due colonne (che calcoliamo): la colonna dei valori $x_i y_i$ e la colonna dei valori x_i^2. Aggiungiamo inoltre un'ultima riga alla tabella riportando le somme delle righe precedenti: quindi la riga (in questo caso $n = 4$)

$$\sum_{i=1}^{4} x_i \qquad \sum_{i=1}^{4} y_i \qquad \sum_{i=1}^{4} x_i y_i \qquad \sum_{i=1}^{4} x_i^2.$$

Otteniamo:

x_i	y_i	$x_i y_i$	x_i^2
-2	8	-16	4
0	2	0	0
2	0	0	4
3	$-\frac{8}{3}$	-8	9
3	$\frac{22}{3}$	-24	17

Il sistema delle equazioni normali è quindi:

$$\begin{cases} m \cdot 3 \; + 4 \cdot q = \frac{22}{3} \\ m \cdot 17 + q \cdot 3 = -24 \end{cases}$$

da cui $m = -2$ e $q = \frac{10}{3}$.

Infatti: dalla 1^a equazione otteniamo

$$m = \frac{22}{9} - \frac{4}{3} \cdot q$$

che sostituito nella 2^a equazione dà

$$17 \cdot \left(\frac{22}{9} - \frac{4}{3} \cdot q \right) + 3 \cdot q = -24.$$

Svolgendo i conti:

$$\begin{aligned} & 17 \cdot (\tfrac{22}{9} - \tfrac{4}{3} \cdot q) + 3 \cdot q = -24 \\ \iff \quad & 17 \cdot \tfrac{22}{9} - 17 \cdot \tfrac{4}{3} \cdot q + 3 \cdot q = -24 \\ \iff \quad & q \cdot (-17 \cdot \tfrac{4}{3} + 3) = -24 - 17 \cdot \tfrac{22}{9} \\ \iff \quad & q \cdot \tfrac{-17 \cdot 4 + 3 \cdot 3}{3} = \tfrac{-24 \cdot 9 - 17 \cdot 22}{9} \\ \iff \quad & q \cdot \tfrac{-68+9}{3} = \tfrac{-216-374}{9} \\ \iff \quad & q \cdot (-\tfrac{59}{3}) = -\tfrac{590}{9} \\ \iff \quad & q = (-\tfrac{590}{9}) \cdot (-\tfrac{3}{59}) = \tfrac{590}{9} \cdot \tfrac{3}{59} = \tfrac{10}{9} \cdot 3 = \tfrac{10}{3}. \end{aligned}$$

Sostituendo $q = \frac{10}{3}$ nella 1^a equazione otteniamo infine

$$m = \frac{22}{9} - \frac{4}{3} \cdot \frac{10}{3} = \frac{22-40}{9} = -\frac{18}{9} = -2.$$

Quindi l'equazione della retta di regressione è:

$$y = -2x + \frac{10}{3}.$$

(b) Cominciamo con lo scrivere la tabella che ha come 1^a colonna le età del gattino misurate rispetto all'unità di misura considerata

$$x_1 = \frac{3 \text{ sett.}}{3 \text{ sett.}} = 1, \quad x_2 = \frac{6 \text{ sett.}}{3 \text{ sett.}} = 2, \quad x_3 = \frac{9 \text{ sett.}}{3 \text{ sett.}} = 3,$$

$$x_4 = \frac{12 \text{ sett.}}{3 \text{ sett.}} = 4, \quad x_5 = \frac{15 \text{ sett.}}{3 \text{ sett.}} = 5$$

e come 2^a colonna i pesi del gattino misurati rispetto all'unità di misura considerata

$$y_1 = 5, \quad y_2 = 7, \quad y_3 = 12, \quad y_4 = 18, \quad y_5 = 22.$$

La retta di regressione ha equazione del tipo $y = mx + q$ dove m e q sono le soluzioni del sistema delle equazioni normali

$$\begin{cases} m \cdot (\sum_{i=1}^n x_i) + \quad n \cdot q = \sum_{i=1}^n y_i \\ m \cdot (\sum_{i=1}^n x_i^2) + q \cdot (\sum_{i=1}^n x_i) = \sum_{i=1}^n x_i y_i \end{cases}$$

Per nostra comodità aggiungiamo alla tabella che ha come colonne i valori delle x_i e delle y_i, altre due colonne (che calcoliamo): la colonna dei valori $x_i y_i$ e la colonna dei valori x_i^2. Aggiungiamo inoltre un'ultima riga alla tabella riportando le somme delle righe precedenti: quindi la riga (in questo caso $n = 5$)

$$\sum_{i=1}^{5} x_i \qquad \sum_{i=1}^{5} y_i \qquad \sum_{i=1}^{5} x_i y_i \qquad \sum_{i=1}^{5} x_i^2.$$

Otteniamo:

x_i	y_i	$x_i y_i$	x_i^2
1	5	5	1
2	7	14	4
3	12	36	9
4	18	72	16
5	22	110	25
15	64	237	55

Il sistema delle equazioni normali è quindi:

$$\begin{cases} m \cdot 15 + 5 \cdot q = 64 \\ m \cdot 55 + q \cdot 15 = 237 \end{cases}$$

da cui $m = \frac{9}{2}$ e $q = -\frac{7}{10}$.

Infatti: dalla 1^a equazione otteniamo

$$q = \frac{1}{5}(64 - 15 \cdot m) = \frac{64}{5} - 3 \cdot m$$

che sostituito nella 2^a equazione dà

$$55 \cdot m + 15 \cdot \left(\frac{64}{5} - 3 \cdot m\right) = 237,$$

da cui

$$55 \cdot m - 45 \cdot m = 237 - 192$$

e quindi

$$m = \frac{45}{10} = \frac{9}{2}.$$

Sostituendo $m = \frac{9}{2}$ nella 1^a equazione otteniamo infine

$$q = \frac{64}{5} - 3 \cdot \frac{9}{2} = \frac{128 - 135}{10} = -\frac{7}{10}.$$

Dunque l'equazione della retta di regressione è:

$$y = \frac{9}{2}x - \frac{7}{10}.$$