

Coppie o vettori di dati

Spesso i dati osservati sono di tipo vettoriale. Ad esempio studiamo 222 osservazioni relative alle eruzioni del geyser Old Faithful.



Old Faithful, Yellowstone Park (Wyoming, USA).

Old Faithful

Qui una sequenza di immagini dell'inizio di una eruzione.



La tabella dei dati

Una parte è elencata in questa tabella (che comprende altre pagine che non riportiamo):

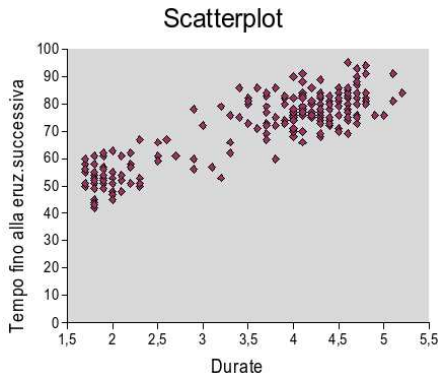
Giorno n.	Durata eruz.	Tempo prima della succ.
1	4,4	78
1	3,9	74
1	4	68
1	4	76
1	3,5	80
1	4,1	84
1	2,3	50
1	4,7	93
1	1,7	55
1	4,9	76
1	1,7	58
1	4,6	74
1	3,4	75
2	4,3	80
2	1,7	56
2	3,9	80
2	3,7	69
2	3,1	57
2	4	90
2	1,8	42
2	4,1	91
2	1,8	51
2	3,2	79
2	1,9	53
2	4,6	82
2	2	51
3	4,5	76
3	3,9	82
3	4,3	84
3	2,3	53
3	3,8	86

Prima componente = giorno di osservazione,
 seconda componente = minuti di durata dell'eruzione,
 terza componente = minuti intercorsi fra quell'eruzione e la successiva.

Scatterplot

Il tipico grafico usato per rappresentare i dati accoppiati è lo **scatterplot** (o diagramma di dispersione).

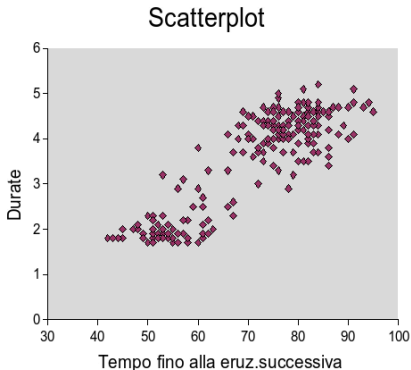
I punti hanno coordinate $(x, y) = (\text{prima var.}, \text{seconda var.})$. Qui $x = \text{durate}$ (seconda colonna), $y = \text{tempi inattività successiva}$ (terza colonna).



A brevi durate sembrano associati brevi periodi di inattività successiva e a lunghe durate lunghe inattività.

Scambiando x e y

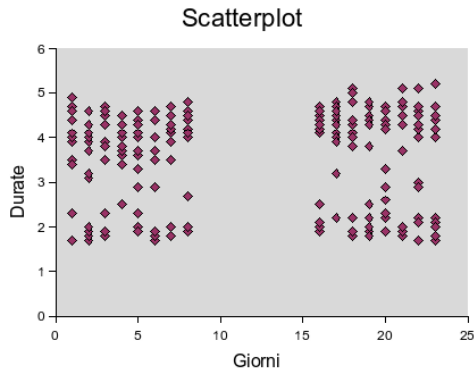
Qui x = tempi inattività successiva (terza colonna), y = durate (seconda colonna).



A brevi periodi di inattività sembrano associati brevi durate dell'eruzione precedente e a lunghe inattività lunghe durate dell'eruzione precedente.

Giorni e durate

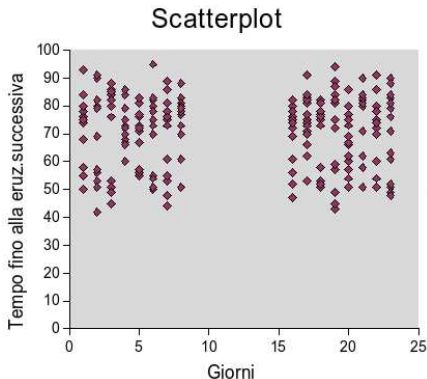
Qui x = numero progressivo del giorno di osservazione (prima colonna), y = durate (seconda colonna).



Non pare esserci alcun legame fra le due quantità.

Giorni e tempi inattività

Qui x = numero progressivo del giorno di osservazione (prima colonna), y = tempi inattività successiva (terza colonna).



Anche qui non pare esserci alcun legame fra le due quantità.

Covarianza

Per quantificare il legame fra due variabili x e y , si possono calcolare la **covarianza** e il **coefficiente di correlazione**.

DEFINIZIONE DI COVARIANZA (di un insieme di dati accoppiati)

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \cdot \bar{y}.$$

Covarianza di un insieme di dati = caso particolare di covarianza di v.a.

La covarianza di un insieme di dati accoppiati coincide con la covarianza delle variabili X e Y dove (X, Y) = coordinate di un valore scelto a caso (cioè con uguale probabilità) fra i dati.

Segno della covarianza

Se $\sigma_{xy} > 0$ si hanno variabili **positivamente correlate** (a valori piccoli di x corrispondono valori piccoli di y e idem per i valori grandi).

Se $\sigma_{xy} < 0$ si hanno variabili **negativamente correlate** (a valori piccoli di x corrispondono valori grandi di y e viceversa).

Se $\sigma_{xy} = 0$ si hanno variabili **scorrelate**.

Correlazione

La covarianza può risultare grande (in valore assoluto) semplicemente perché le quantità x e y assumono valori grandi e/o con varianza grande.

Per ovviare a questo inconveniente si considera il coefficiente di correlazione.

DEFINIZIONE DI COEFFICIENTE DI CORRELAZIONE (di un insieme dati accoppiati)

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

dove σ_{xy} è la covarianza mentre σ_x e σ_y sono rispettivamente la radice della varianza della componente x e la radice della varianza della componente y .

Il coefficiente di correlazione è un numero puro, compreso fra -1 e 1.

Se il coefficiente di correlazione

- è vicino a zero, indica che non c'è legame fra le due variabili;
- è vicino a +1 oppure a -1, indica che fra le due variabili c'è un legame *lineare*, cioè sono legate da una relazione del tipo

$$Y = aX + b, \quad \text{con } a \text{ e } b \text{ numeri reali.}$$

Esempio

La covarianza tra le durate e i tempi di inattività vale 12.16 (minuti al quadrato), mentre la correlazione vale 0.88.

La covarianza tra i giorni e le durate vale 0.71 (minuti al quadrato), mentre la correlazione vale 0.08.

La covarianza tra i giorni e i tempi di inattività vale -0.3 (minuti al quadrato), mentre la correlazione vale -0.003.

Dunque è ragionevole pensare che non ci sia legame fra giorni e durate e fra giorni e tempi di inattività, mentre fra durate e tempi di inattività pare esserci un legame lineare.

Modello lineare

Modello

Si pensa ogni Y_i come una **variabile dipendente**, funzione lineare affine di X_i (detto **predittore**) più una **perturbazione casuale** W_i :

$$Y_i = \beta_0 + \beta_1 X_i + W_i.$$

Si hanno n osservazioni accoppiate

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

che si pensano provenire da

$$y_i = \beta_0 + \beta_1 x_i + w_i.$$

Ipotesi necessarie

Modello

Si suppone che:

- β_0 e β_1 sono numeri reali (incogniti);
- le W_i sono v.a. indipendenti e tutte $\mathcal{N}(0, \sigma^2)$.

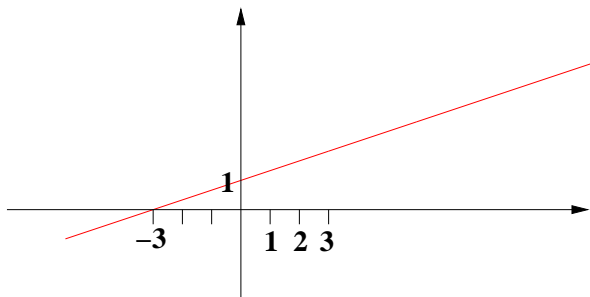
Il modello potrebbe comunque non valere perché:

- il legame fra le x e le y potrebbe essere più complesso;
- le W_i potrebbero non essere normali;
- la varianza delle W_i potrebbe dipendere da i (non essere dunque identicamente distribuite).

Regressione lineare

Se il modello lineare è valido, significa che, se non ci fossero perturbazioni (cioè se fosse $w_i = 0$ per ogni i), allora tutti i punti si troverebbero sulla retta

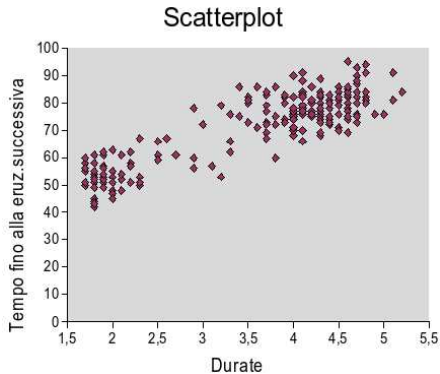
$$y = \beta_0 + \beta_1 x.$$



Esempio: la retta $y = 1 + \frac{1}{3}x$

Con le perturbazioni

In realtà le perturbazioni non sono nulle, perciò i grafici che troviamo sono di questo tipo:



La retta dei minimi quadrati

Siccome nella situazione ideale i punti giacciono tutti su una stessa retta,

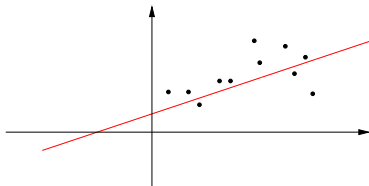
l'idea è che la retta $y = \beta_0 + \beta_1 x$ sarà quella **più vicina** ai punti dello *scatterplot*.

Dobbiamo quindi definire un concetto di **distanza fra una retta e un insieme di punti**

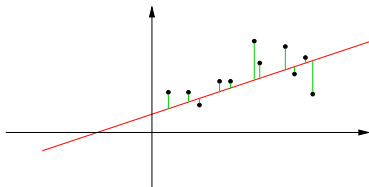
⇒ la *retta dei minimi quadrati* sarà la retta che minimizza tale distanza.

Distanza retta - punti

Supponiamo di avere la retta rossa e i punti:



La distanza è la somma dei quadrati delle lunghezze dei segmenti verdi:



Definizione di distanza

DEFINIZIONE DI DISTANZA RETTA - PUNTI

Dati i punti $\{(x_1, y_1), \dots, (x_n, y_n)\}$ e la retta $y = ax + b$, la loro **distanza** è

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

Si tratta della somma dei quadrati della distanza verticale fra y_i e il corrispondente punto sulla retta, $ax_i + b$.

La retta dei minimi quadrati

DEFINIZIONE DI RETTA DEI MINIMI QUADRATI

Date le osservazioni accoppiate

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

la **retta dei minimi quadrati** (o **retta di regressione**) è la retta che ha la minima distanza da questi punti.

Tale retta dipende dalle osservazioni, e precisamente dipende da

la media delle x : \bar{x}_n ;

la media delle y : \bar{y}_n ;

la varianza dei dati x : $\sigma_x^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x}_n)^2$;

la covarianza di x e y : $\sigma_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \cdot \bar{y}$.

Equazione

EQUAZIONE

La retta dei minimi quadrati ha equazione

$$y = \left(\bar{y}_n - \frac{\sigma_{xy}}{\sigma_x^2} \bar{x}_n \right) + \frac{\sigma_{xy}}{\sigma_x^2} x,$$

e quindi gli stimatori per β_0 e per β_1 sono rispettivamente

$$b_0 = \left(\bar{y}_n - \frac{\sigma_{xy}}{\sigma_x^2} \bar{x}_n \right) \quad \text{e} \quad b_1 = \frac{\sigma_{xy}}{\sigma_x^2}.$$

Sono entrambi non distorti e consistenti.

Esempio: Old Faithful

Consideriamo x = durate eruzione; y = tempo inattività successiva. Il calcolo fornisce:

\bar{x}_n	\bar{y}_n	σ_x^2	σ_{xy}
3.58	70.99	1.17	12.16

Da questo si ricavano le stime:

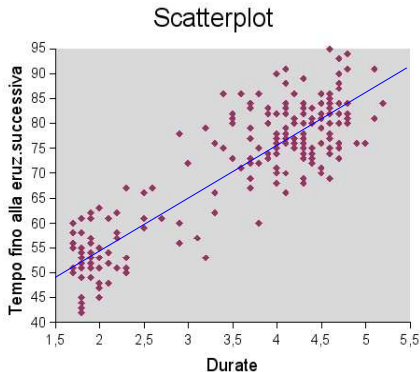
$$b_1 = 10.4, \quad b_0 = 33.8$$

e la retta di regressione:

$$y = 33.8 + 10.4x.$$

Scatterplot e retta

Vediamo la retta disegnata in blu sullo scatterplot:



Valori previsti

Una volta calcolata la retta di regressione la si può utilizzare per stimare i valori di y corrispondenti a x date.

DEFINIZIONE DI VALORE PREVISTO

Data la retta dei minimi quadrati $y=b_0+b_1x$, il **valore previsto** di y per $x = x_i$ è

$$\hat{y}_i = b_0 + b_1 x_i.$$

Quindi y_i è il valore osservato in corrispondenza di x_i , mentre \hat{y}_i è il valore previsto dal modello (quello che sta sulla retta). Se il modello è buono questi valori dovranno essere vicini.

Altro uso

La formula può essere usata per prevedere valori y in corrispondenza di x non osservate. Ad esempio: calcolo una regressione fra anni e temperature medie. Posso usare la formula per ottenere il valore previsto fra 10 anni.

Previsione eruzione Old Faithful

Ricordiamo la retta dei minimi quadrati

$$y = 33.8 + 10.4x.$$

Se abbiamo appena osservato un'eruzione della durata di 4'45" allora la stima del successivo intervallo di inattività sarà

$$y = 33.8 + 10.4 \cdot 4.75 = 83.2$$

cioè 83'12".

DEFINIZIONE DI RESIDUO

Se y_i sono i valori osservati e \hat{y}_i quelli previsti, le quantità seguenti sono dette **residui**

$$r_i = y_i - \hat{y}_i.$$

Stima della varianza delle W_i

Uno stimatore per la varianza σ^2 delle perturbazioni W_i è

$$T = \frac{1}{n-2} \sum_{i=1}^n r_i^2.$$

R quadro

Un numero che ci aiuta a valutare la bontà del modello lineare è il cosiddetto R quadro.

DEFINIZIONE DI R^2

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

Si potrebbe vedere che R^2 è il coefficiente di correlazione dei campioni y e \hat{y} .

Interpretazione

Si interpreta come proporzione di varianza di y spiegata dal modello lineare.

Tanto più è vicino a 1, tanto migliore è l'adattamento del modello ai dati osservati.

Test per la regressione

Esistono test sui coefficienti β_0 e β_1 .

In particolare è interessante il test con

$$H_0 : \beta_1 = 0; \quad H_1 : \beta_1 \neq 0.$$

Infatti tale ipotesi significa che non c'è dipendenza di y da x : rifiutarla vuol dire che c'è evidenza statistica di una dipendenza di y da x .

Un altro test è quello con

$$H_0 : \beta_0 = 0; \quad H_1 : \beta_0 \neq 0.$$

Non vediamo come si fanno questi test (entrambi con quantili di Student), ma i pacchetti statistici in genere per la regressione danno un output molto ricco, compreso il risultato di questi test.

Regressione multipla

Con le stesse idee esposte fin qui, si affronta anche la regressione multipla, dove ci sono più predittori.

Modello

Si pensa ogni Y_i come una **variabile dipendente**, funzione lineare affine dei **predittori** Z_1, \dots, Z_k più **perturbazioni casuali** W_i . Si hanno n osservazioni Y_1, \dots, Y_n :

$$Y_1 = \beta_0 + \beta_1 Z_{11} + \beta_2 Z_{21} + \dots + \beta_k Z_{k1} + W_1$$

$$Y_2 = \beta_0 + \beta_1 Z_{12} + \beta_2 Z_{22} + \dots + \beta_k Z_{k2} + W_2$$

$$\dots = \dots$$

$$Y_n = \beta_0 + \beta_1 Z_{1n} + \beta_2 Z_{2n} + \dots + \beta_k Z_{kn} + W_n$$

dove Z_{ij} è la osservazione j del predittore Z_i , β_0, \dots, β_k sono k parametri e le W_1, \dots, W_n sono v.a. i.i.d. con legge $\mathcal{N}(0, \sigma^2)$.

Tipici output

La regressione multipla è ormai affidata ai calcolatori (i conti a mano sono possibili ma lunghi e tediosi). Mostriamo dunque a titolo di esempio **uno studio di regressione per la concentrazione di ozono in funzione della temperatura**, fornito dal software S-Plus.

Dunque y = concentrazione di ozono e x = temperatura.

*** Linear Model ***

Coefficients:

	Value	Std. Error	t value	Pr(> t)
(Intercept)	-147.6461	18.7553	-7.8723	0.0000
Temperature	2.4391	0.2393	10.1919	0.0000

Residual standard error: 23.92 on 109 degrees of freedom

Multiple R-Squared: 0.488

Leggiamo l'output

*** Linear Model ***				
Coefficients:				
	Value	Std. Error	t value	Pr(> t)
(Intercept)	-147.6461	18.7553	-7.8723	0.0000
Temperature	2.4391	0.2393	10.1919	0.0000
Residual standard error: 23.92 on 109 degrees of freedom				
Multiple R-Squared: 0.488				

$b_0 = -147.6461$ è la stima di β_0 ; $b_1 = 2.4391$ è la stima di β_1 ;
23.92 è la stima di σ^2 ; R^2 vale 0.488.

La terza colonna riporta la deviazione standard (stimata) dei due stimatori di β_0 e di β_1 : 18.7553 per b_0 e 0.2393 per b_1 .

La quarta e quinta colonna riportano il calcolo della statistica e il p -value del test con $H_0 : \beta_0 = 0$ (prima riga) e del test con $H_0 : \beta_1 = 0$ (seconda riga).

Poiché il p -value è pressoché nullo (inferiore a 10^{-4}), c'è evidenza che $\beta_0 \neq 0$ e $\beta_1 \neq 0$.

Aggiungiamo il vento

Ecco l'output con **due predittori: la temperatura e il vento**.

*** Linear Model ***				
Coefficients:				
	Value	Std. Error	t value	Pr(> t)
(Intercept)	-67.2008	23.6083	-2.8465	0.0053
Temperature	1.8265	0.2504	7.2931	0.0000
Wind	-3.2993	0.6706	-4.9201	0.0000
Residual standard error: 21.72 on 108 degrees of freedom				
Multiple R-Squared: 0.5817				

R^2 aumenta: dunque il modello con anche il vento spiega meglio la concentrazione di ozono del modello con la sola temperatura.

Aggiungiamo la radiazione

Ecco l'output con **tre predittori: la temperatura, il vento e la radiazione**.

*** Linear Model ***				
Coefficients:				
	Value	Std. Error	t value	Pr(> t)
(Intercept)	-64.2321	23.0420	-2.7876	0.0063
Radiation	0.0598	0.0232	2.5800	0.0112
Temperature	1.6512	0.2534	6.5159	0.0000
Wind	-3.3376	0.6538	-5.1046	0.0000
Residual standard error: 21.17 on 107 degrees of freedom				
Multiple R-Squared: 0.6062				

Anche se R^2 è aumentato, risulta un poco più dubbia la dipendenza dalla radiazione
(il p -value del test con $H_0 : \beta_2 = 0$ è 0.0112).

Come superare l'esame?

Rubo qualche consiglio a Francesco Rovetto (psicologo, autore de *Il piacere di apprendere*).

- 1 **INTERESSE.** Ogni materia ha qualcosa di curioso, di utile. Le informazioni che rispondono a una nostra curiosità si fissano più rapidamente e più facilmente.
- 2 **VISIONE D'INSIEME.** Un grosso errore è studiare i dettagli senza aver prima capito il contesto generale.
- 3 **COMPRENSIONE.** Non imparate a memoria! Noi non ricordiamo le cose come le abbiamo lette ma ricordiamo i significati e le relazioni tra di esse. La comprensione richiede uno sforzo iniziale, ma poi ripaga.
- 4 **MOTIVAZIONE.** Uno studente disinteressato non ricorda ciò che studia.
- 5 **FIDUCIA.** Bisogna aver fiducia nelle proprie possibilità.