Le relazioni statistiche

Pasquale Sarnacchiaro

Ricercatore di Statistica presso l'Università degli studi di Roma Unitelma Sapienza

Struttura Generale

Distribuzioni doppie

c modalità del carattere Y

X / Y	y_1	y_2	....	y_j	....	y_c	Totale
x_1	n_{11}	n_{12}	....	n_{1j}	....	n_{1c}	$n_{1\circ}$
x_2	n_{21}	n_{22}	....	n_{2j}	....	n_{2c}	$n_{2\circ}$
⋮	....	....	....	....	....	....	⋮
x_i	n_{i1}	n_{i2}	....	n_{ij}	....	n_{ic}	$n_{i\circ}$
⋮	....	....	....	....	....	....	⋮
x_r	n_{r1}	n_{r2}	....	n_{rj}	....	n_{rc}	$n_{r\circ}$
Totale	$n_{\circ 1}$	$n_{\circ 2}$	....	$n_{\circ j}$	....	$n_{\circ c}$	N

Numero totale di unità statistiche

r modalità del carattere X

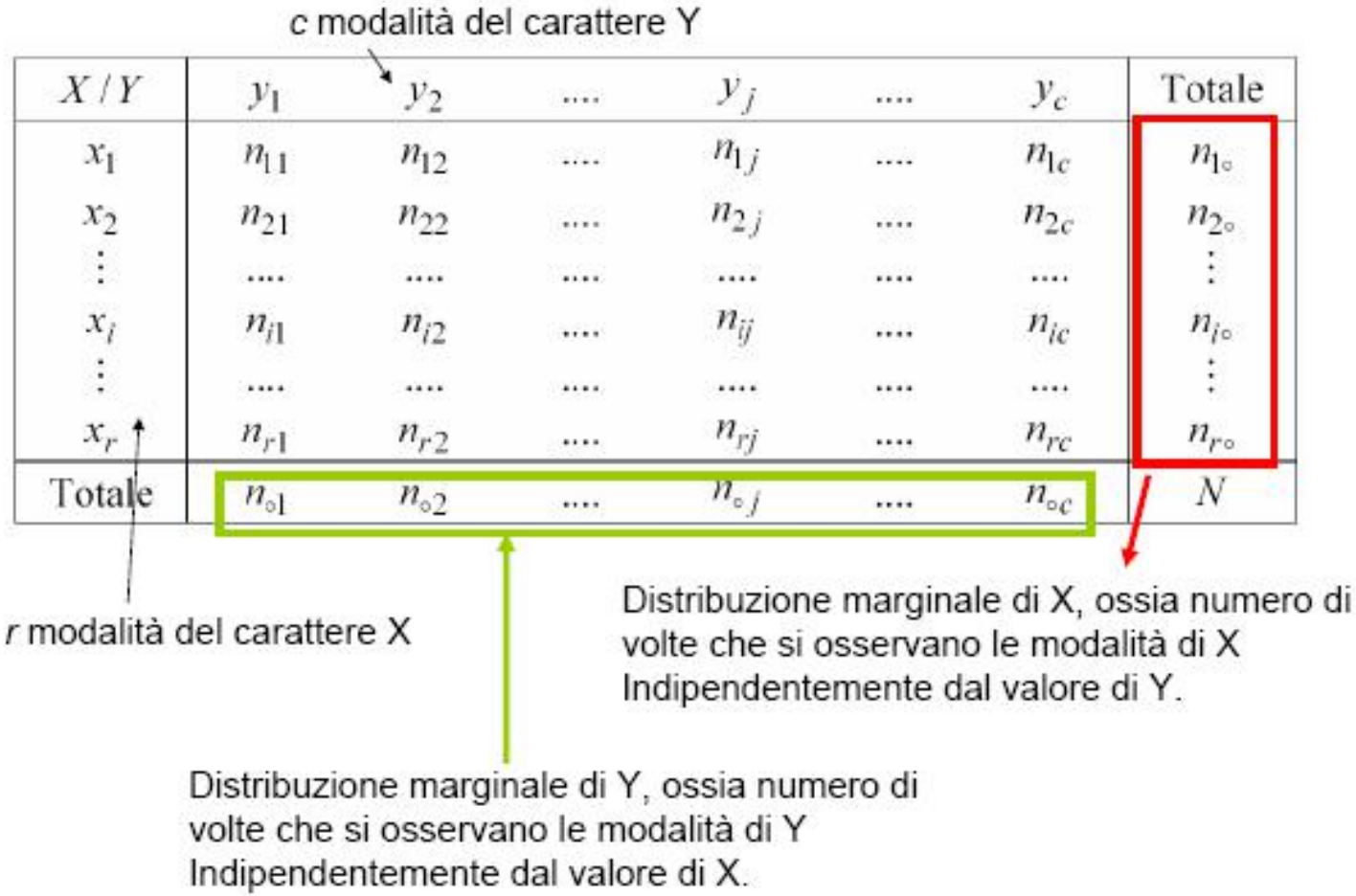
Frequenza congiunta (assoluta) con cui si osserva la modalità i-ima di X congiuntamente alla modalità j-ima di Y, ossia la coppia (*xi*,*yj*) è osservata *nij* volte!!

Tabelle a doppia entrata

Dopo il disastro, una commissione d'inchiesta del *British Board of Trade* ha compilato una lista di tutti i 1316 passeggeri del Titanic con alcune informazioni aggiuntive riguardanti: l'esito (salvato, non salvato), la classe (I, II, III) in cui viaggiavano, il sesso, l'età…

Passeggero	Classe	Esito
nome 1	II	salvato
nome 2	III	non salvato
nome 3	I	non salvato
⋮	⋮	⋮
nome 1316	III	salvato

Distribuzioni marginali



X / Y	y_1	y_2	....	y_j	....	y_c	Totale
x_1	n_{11}	n_{12}	....	n_{1j}	....	n_{1c}	$n_{1\circ}$
x_2	n_{21}	n_{22}	....	n_{2j}	....	n_{2c}	$n_{2\circ}$
⋮	....	....	....	....	....	....	⋮
x_i	n_{i1}	n_{i2}	....	n_{ij}	....	n_{ic}	$n_{i\circ}$
⋮	....	....	....	....	....	....	⋮
x_r	n_{r1}	n_{r2}	....	n_{rj}	....	n_{rc}	$n_{r\circ}$
Totale	$n_{\circ 1}$	$n_{\circ 2}$	....	$n_{\circ j}$	....	$n_{\circ c}$	N

- Quindi avremo:

$$n_{i.} = \sum_{j=1}^{c} n_{ij}$$ Frequenze marginali di X

$$n_{.j} = \sum_{i=1}^{r} n_{ij}$$ Frequenze marginali di Y

$$N = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{c} n_{.j}$$

La numerosità del colletivo è data dalla somma di tutte le frequenze congiunte, o di tutte le frequenze marginali.

Indici di Connessione

- Un operatore statistico bivariato è una procedura di calcolo che considera due variabili e sintetizza l'informazione sulla loro distribuzione congiunta in uno scalare.
- Gli operatori di connessione producono uno scalare sempre positivo; assumono valore zero in assenza di connessione e maggiore di zero in presenza di connessione tra le due variabili.

La natura della relazione deve essere ipotizzata dal ricercatore

Quello che dobbiamo notare è che la natura della relazione causale tra due o più variabili discende da un ragionamento che si pone a livello teorico e non a livello empirico: un operatore statistico di per sé non informa sulla natura logica di una relazione causale tra le variabili empiriche.

L’indice Chi-quadrato

- Il più importante operatore di connessione che rappresenta l’intensità della relazione tra due variabili categoriali è chiamato “chi quadrato” (χ^2).
- Per costruire un operatore di connessione tra le due variabili prendiamo come modello di riferimento *l’assenza di relazione* e calcoliamo quanto le frequenze osservate si discostano dalle frequenze teoriche calcolate sulla base dell’ipotesi di completa indipendenza.

L’indice Chi-quadrato

- Più le frequenze empiriche si allontanano dalle frequenze teoriche più è elevato il grado di connessione tra le variabili.
- L’operatore chi quadrato si basa proprio sulla differenza tra frequenze empiriche e teoriche.
- Più le precisamente, l’operatore chi quadrato si calcola come:

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}}$$

Frequenze Teoriche

Le frequenze teoriche ($\hat{n}_{ij}$) sono riportate tra parentesi e sono calcolate come:

$$\hat{n}_{ij} = \frac{n_{i.} \times n_{.j}}{N}$$

Campo di variazione

- L'operatore chi quadrato assume come valore minimo zero e come valore massimo il minore dei seguenti due valori: $N(I-1)$ e $N(J-1)$.

- Il valore massimo dipende dunque dall'ampiezza del collettivo e dal numero di righe e colonne della tabella.

Misure basate sul chi quadrato

Una misura di connessione basata sul chi quadrato e indipendente dal numero dei casi è stata proposta da Pearson e si calcola come:

$$\Phi^2 = \frac{X^2}{N}$$

Questa misura di connessione viene chiamata phi quadrato, assume coma valore minimo 0. Il suo valore massimo è funzione del numero di modalità delle variabili:

$$\min (I - 1; J - 1)$$

Misure basate sul chi quadrato

Un'altra misura di connessione è data dal *T* di Tschuprov:

$$T = \frac{\Phi^2}{\sqrt{(J-1)(I-1)}}$$

Il *T* di Tschuprov assume il valore di 1 (con tabelle quadrate) nel caso di dipendenza reciproca perfetta.

Misure basate sul chi quadrato

Un ultimo operatore basato sulla relativizzazione del chi quadrato è il *V* di Cramer:

$$V = \frac{\chi^2}{N \times \min\left[(J-1);(I-1)\right]}$$

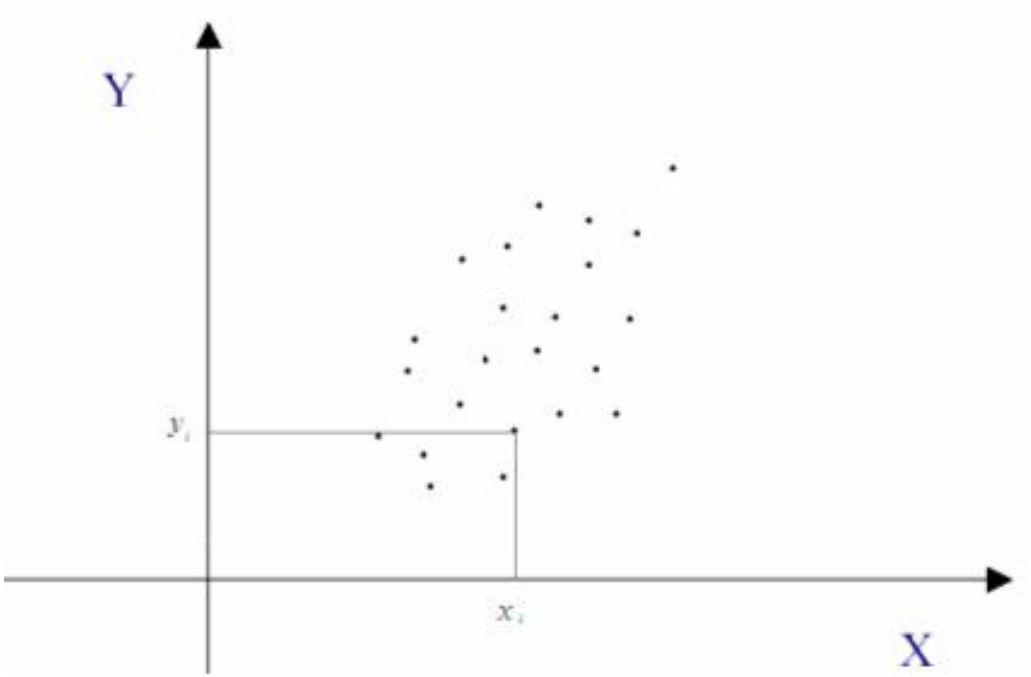
Assume sempre valori compresi tra 0 e 1.

Indipendenza correlativa

- Se due caratteri sono entrambi quantitativi è possibile studiare l'esistenza di un legame per cui a variazioni di un carattere corrispondono variazioni dell'altro carattere secondo una relazione lineare.
- I caratteri hanno lo stesso ruolo (simmetrico). Non è possibile stabilire un criterio di antecedenza logica per alcuno dei fenomeni.

Scatterplot

- Un primo modo per verificare l'esistenza di una correlazione lineare tra due caratteri quantitativi X e Y, e quello di rappresentare la distribuzione doppia (X,Y) attraverso un grafico a dispersione (o *scatterplot).*



Codevianza / Covarianza

Una misura assoluta della concordanza/discordanza è la **codevianza *(covarianza)***

Distribuzione doppia per unità

$$Codev(X,Y)=\sum_{i=1}^{r}\sum_{j=1}^{c}(x_i-M_x)(y_j-M_y)=$$

$$\sum_{i=1}^{r}\sum_{j=1}^{c}x_i y_j - nM_xM_y$$

$$\mathrm{Cov(X,Y)}=\frac{\mathrm{Codev(X,Y)}}{\mathrm{n}}$$

$$Codev(X,Y)=\sum_{i=1}^{r}\sum_{j=1}^{c}(x_i-M_x)(y_j-M_y)n_{ij}=$$

$$\sum_{i=1}^{r}\sum_{j=1}^{c}x_i y_j n_{ij} - nM_xM_y$$

Distribuzione doppia per frequenze

Indipendenza/dipendenza correlativa

Se $\mathrm{Codev}(X,Y)=0$ Indipendenza correlativa

Se $\mathrm{Codev}(X,Y)>0$ concordanza

Ai valori più piccoli (grandi) di uno dei due caratteri corrispondono in media i valori più piccoli (grandi) della seconda variabile.

Se $\mathrm{Codev}(X,Y)<0$ discordanza

Ai valori più piccoli (grandi) di uno dei due caratteri corrispondono in media i valori più grandi (piccoli) della seconda variabile.

La covarianza è una misura simmetrica:

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X).$$

Codevianza / Covarianza

Se X e Y sono statisticamente indipendenti, implica che: Cov(X, Y) = 0

Non vale il viceversa:
se la Cov(X, Y) = 0, questo non implica che X e Y siano indipendenti.

La covarianza si può annulla anche se i prodotti degli scostamenti dalla media si compensano.

Data la seguente distribuzione:

X	-2	-1	0	1	2
Y	4	1	0	1	4

Si ha: $\bar{x} = 0$, $\bar{y} = 2$ e $\text{Cov}(X, Y) = \frac{1}{4}(-8-1+1+8)-0 \bullet 2 = 0$

Vi è indipendenza correlativa, ma non vi è indipendenza statistica in quanto le due variabili sono legate dalla relazione funzionale:

$$y = x^2.$$

Coefficiente di correlazione lineare

Una misura relativa della concordanza è data dal coefficiente di correlazione r

$$r = \frac{codev(X,Y)}{\sqrt{dev(X) * dev(Y)}} \qquad -1 \leq r \leq 1$$

Il coefficiente r è un numero puro, che varia tra –1 a 1 ed ha il segno algebrico della codevianza.

Coefficiente di correlazione lineare

Se r =-1	$\Rightarrow$	vi è perfetta relazione lineare tra X e Y e vi è discordanza.
Se -1 < r <0	$\Rightarrow$	vi è discordanza.
Se r = 0	$\Rightarrow$	non vi è relazione lineare tra X e Y, le variabili sono incorrelate, non vi è né concordanza, né discordanza.
Se 0 < r <1	$\Rightarrow$	vi è concordanza.
Se r = 1	$\Rightarrow$	vi è perfetta relazione lineare tra X e Y e vi è concordanza.

Coefficiente di correlazione lineare

$r = 0 \Rightarrow$ l'indipendenza statistica

l'indipendenza statistica $\Rightarrow r = 0$

Il coefficiente di correlazione nullo non implica l'indipendenza statistica, ma solo l'indipendenza lineare.

Esempio

Automobili	Distanza percorsa (km) X	Tempo di consegna (in giorni) Y
1	60	20
2	156	24
3	148	32
4	168	28
5	180	43
6	300	27
7	235	45
8	195	38

Le due medie sono: $\bar{x} = 180{,}25$ e $\bar{y} = 32{,}125$, quindi:

Esempio

x	y	x^2	y^2	xy
60	20	3600	400	1200
156	24	24336	576	3744
148	32	21904	1024	4736
168	28	28224	784	4704
180	43	32400	1849	7740
300	27	90000	729	8100
235	45	55225	2025	10575
195	38	38025	1444	7410
1442	257	293714	8831	48209

Il coefficiente di correlazione è:

$$r = \frac{48209 - 8(180{,}25)(32{,}125)}{\sqrt{(293714 - 8 \cdot 32490{,}06)(8831 - 8 \cdot 1032{,}016)}} = 0{,}43.$$

Indipendenza in media

- Si supponga di aver una distribuzione doppia di una variabile Y quantitativa e di una variabile X che può essere sia quantitativa che qualitativa e di voler misurare quanto Y dipenda in media da X.

- **Y è indipendente in media da X se ogni distribuzione parziale della Y ha la stessa media aritmetica**

Rapporto di correlazione: eta quadro

Data una matrice a doppia entrata è possibile calcolare le medie parziali ognuna delle distribuzioni parziali

Indipendenza in media

X \ Y	y_1	y_2	…	y_j	…	y_c	Totale	Medie Parziali
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1c}	$n_{1\bullet}$	$\bar{y}_1$
⋮	⋮			⋮		⋮	⋮	⋮
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ic}	$n_{i\bullet}$	$\bar{y}_i$
⋮	⋮			⋮		⋮	⋮	⋮
x_r	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rc}	$n_{r\bullet}$	$\bar{y}_r$
Totale	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet c}$	n	
Medie Parziali	$\bar{x}_1$	$\bar{x}_2$	…	$\bar{x}_j$	…	$\bar{x}_c$		

$$\bar{y} = \frac{\bar{y}_1 n_{1\bullet} + \bar{y}_2 n_{2\bullet} + \cdots + \bar{y}_r n_{r\bullet}}{n}$$

$$\bar{y}_i = \frac{\sum_{j=1}^{c} y_j n_{ij}}{n_{i\bullet}}$$

Rapporto di Correlazione

Esiste indipendenza in media di una variabile Y dalla variabile X se le medie parziali sono tutte uguali tra loro al variare delle modalità dell'altra variabile

Indipendenza in media

$$\bar{y}_1 = \bar{y}_2 = \cdots \bar{y}_i = \cdots \bar{y}_r$$

Poiché

$$\bar{y} = \frac{\bar{y}_1 n_{1\bullet} + \bar{y}_2 n_{2\bullet} + \cdots + \bar{y}_r n_{r\bullet}}{n}$$

Abbiamo

$$\bar{y} = \bar{y}_1 = \bar{y}_2 = \cdots \bar{y}_i = \cdots \bar{y}_r$$

Rapporto di Correlazione

$$\eta^2_{YX} = \frac{\sum (\bar{y}_i - \bar{y})^2 n_{i.}}{\sum (y_j - \bar{y})^2 n_{.j}} = \frac{Dev(B)}{Dev(Y)} = 1 - \frac{Dev(W)}{Dev(Y)}$$

$$0 \le \eta^2_{YX} \le 1$$

Indipendenza in media

Se entrambe le variabili X e Y sono quantitative, è possibile calcolare

$$\eta^2_{XY} = \frac{\sum (\bar{x}_j - \bar{x})^2 n_{.j}}{\sum (x_i - \bar{x})^2 n_{i.}}$$

$$\eta^2_{YX} \neq \eta^2_{XY}$$

Esempio

Si consideri la distribuzione doppia del reddito Y e del titolo di studio X di 50 impiegati di una azienda.

Reddito (milioni) Titolo di studio	10 - 16	17 - 25	26 - 36	37 - 49	Totale
Diploma	4	6	5	1	16
Laurea	0	2	8	2	12
Altro	16	4	2	0	22
Totale	20	12	15	3	50

Le medie parziali sono:

$$\bar{y}_1 = \frac{13\cdot 4 + 21\cdot 6 + 31\cdot 5 + 43\cdot 1}{16} = 23{,}5$$

$$\bar{y}_2 = \frac{13\cdot 0 + 21\cdot 2 + 31\cdot 8 + 43\cdot 2}{12} = 31{,}3$$

$$\bar{y}_3 = \frac{13\cdot 16 + 21\cdot 4 + 31\cdot 2 + 43\cdot 0}{22} = 16{,}1.$$

La media generale è:

$$\bar{y} = \frac{13\cdot 20 + 21\cdot 12 + 31\cdot 15 + 43\cdot 3}{50} = 22{,}12.$$

...continua

$\bar{y}_i$	$n_{i\bullet}$	$(\bar{y}_i - \bar{y})^2 n_{i\bullet}$
23,5	16	30,4704
31,3	12	1011,269
16,1	22	797,2888
		1839,028

y_i	$n_{\bullet j}$	$(y_i - \bar{y})^2 n_{\bullet j}$
13	20	1663,488
21	12	15,0528
31	15	1182,816
43	3	1307,923
		4169,28

$$\eta^2_{YX} = \frac{1839,028}{4169,28} = 0,44$$

Indipendenza

È possibile stabilire la seguente gerarchia tra i tre concetti di indipendenza:

Indipendenza statistica $\Rightarrow$ Indipendenza in media $\Rightarrow$ Indipendenza correlativa

Non vale il viceversa

Indipendenza correlativa $\not\Rightarrow$ Indipendenza in media $\not\Rightarrow$ Indipendenza statistica

Il Modello di Regressione Lineare Semplice

Ipotesi alla base del modello di regressione

- Modello lineare:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

  - Ipotesi

$$E(\varepsilon_i) = 0$$

$$Var(\varepsilon_i) = \sigma^2$$

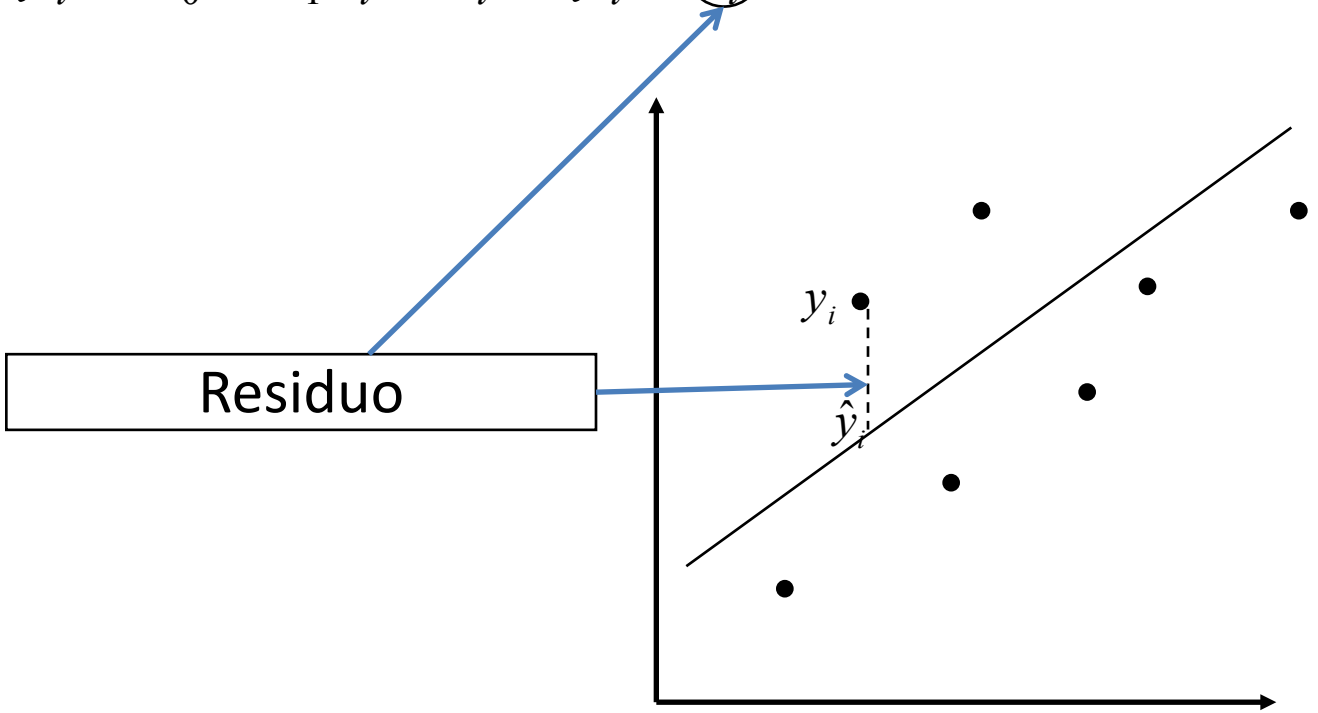
$$Cov(\varepsilon_i, \varepsilon_j) = 0$$

X è deterministica

Stima dei parametri

Vogliamo stimare i coefficienti β_0, β_1 sulla base di *n* osservazioni campionarie (x_i, y_i). Estratto un campione casuale si consideri il modello lineare :

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



Metodo dei minimi quadrati

Consiste nel minimizzare la somma dei quadrati dei residui rispetto a b_0, b_1

$$S(b_0, b_1) = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

Metodo dei minimi quadrati

Derivando $S(b_0, b_1)$ rispetto al coefficiente b_0 si ha:

uguagliando a "0" si ottiene:

$$\frac{\partial S}{\partial b_0} = -2\sum[y_i - b_0 - b_1 x_i]$$

$$\sum y_i - nb_0 - b_1 \sum x_i = 0$$

$$\sum y_i = nb_0 + b_1 \sum x_i \Rightarrow n\bar{y} = nb_0 + b_1 n\bar{x}$$

Metodo dei minimi quadrati

Dividendo il tutto per *n* si ha:

$$\bar{y} = b_0 + b_1 \bar{x}$$

da cui si evidenzia che la retta dei minimi quadrati passa per il punto

$$(\bar{x}; \bar{y})$$

Da l'ultima espressione si ottiene:

$$b_0 = \bar{y} - b_1 \bar{x}$$

Sostituendo questa espressione nella funzione $S(b_0, b_1)$ si ha:

$$\sum [(y_i - \bar{y}) - b_1 (x_1 - \bar{x})]^2$$

Metodo dei minimi quadrati

Derivando rispetto a b_1 e uguagliando a "0" si ottiene:

$$\frac{\partial S}{\partial b_1} = -2\sum\left[(y_i - \bar{y}) - b_1(x_i - \bar{x})\right](x_i - \bar{x}) = 0$$

$$b_1 \sum (x_i - \bar{x})^2 = \sum (y_i - \bar{y})(x_i - \bar{x})$$

Da ciò deriva:

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

Proprietà della retta dei minimi quadrati

La retta dei minimi quadrati è l'unica retta che minimizza la somma dei quadrati dei residui, non è possibile trovare nessuna altra retta che abbia la stessa o la più piccola somma dei quadrati dei residui.

La retta dei minimi quadrati passa per il punto di coordinate $(\bar{x};\bar{y})$ indicato come centro di gravita della nube dei punti.

La somma dei residui della retta dei minimi quadrati è uguale a zero

$$\sum e_i = 0$$

Se $b_1 > 0\ (<0)$ la retta di regressione ha pendenza positiva (negativa).
Se $b_1 = 0$ la retta è parallela all'asse delle ascisse.

Misura della bontà di accostamento

Per verificare la validità del modello di regressione prescelto si può determinare una misura di quanto i valori teorici siano vicini ai valori osservati.

Una misura della bontà di accostamento la si ottiene dalla decomposizione della devianza totale della variabile dipendente Y.

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

S_{yy} — SSE — SSR

Devianza totale

Devianza residua

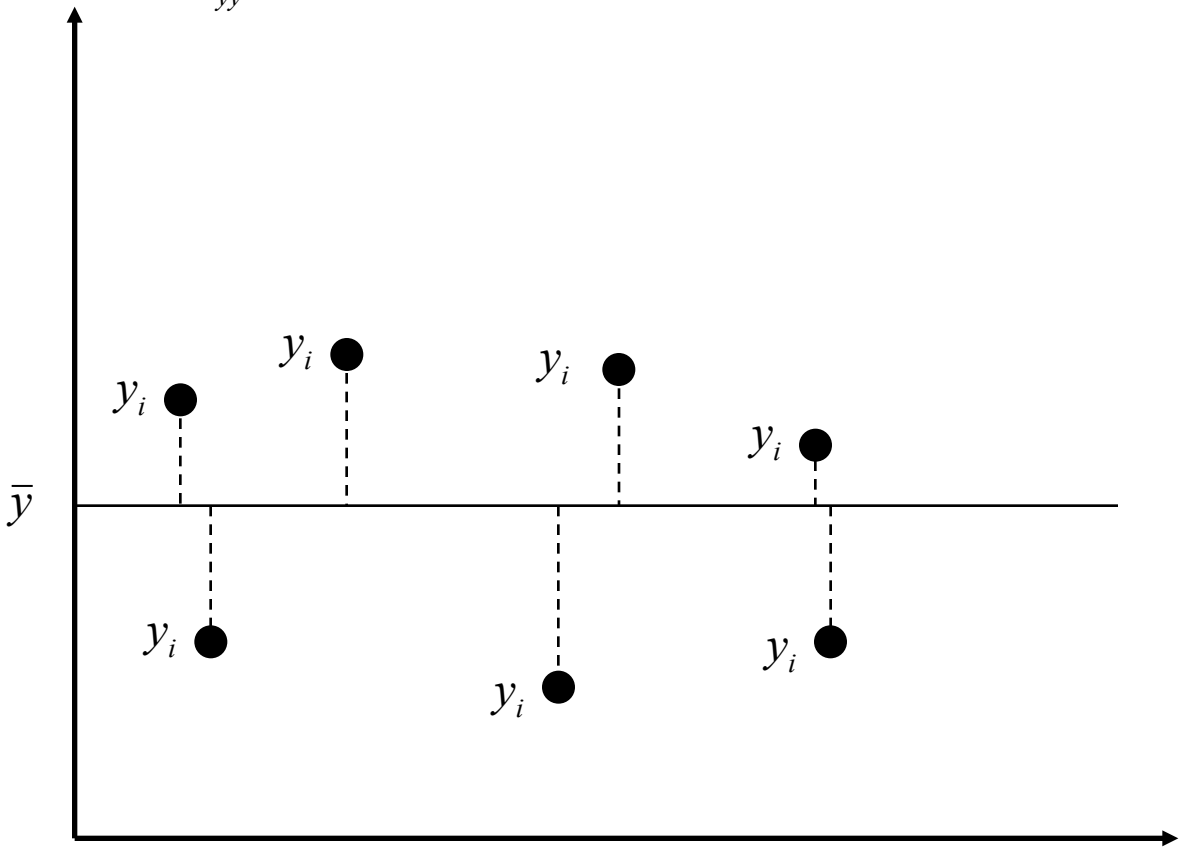
Devianza di regressione

Misura della bontà di accostamento

$$\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$$

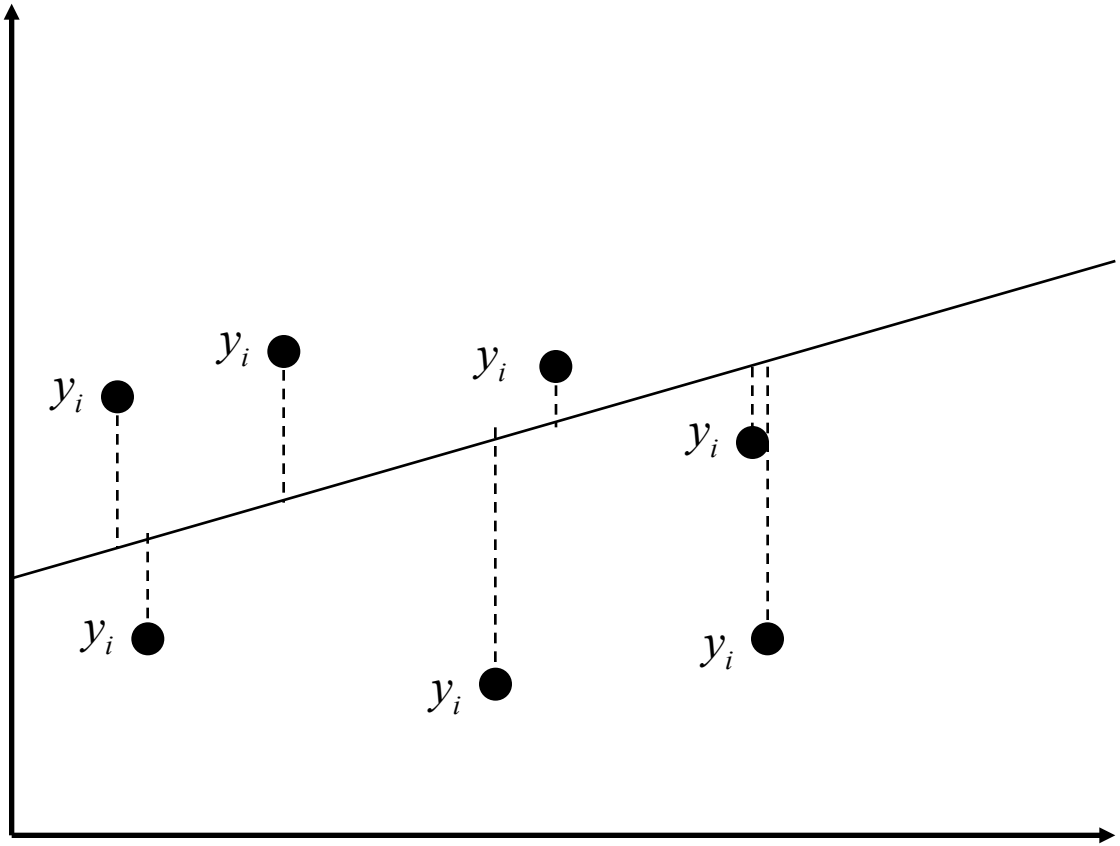
Devianza totale

S_{yy}



Misura della bontà di accostamento

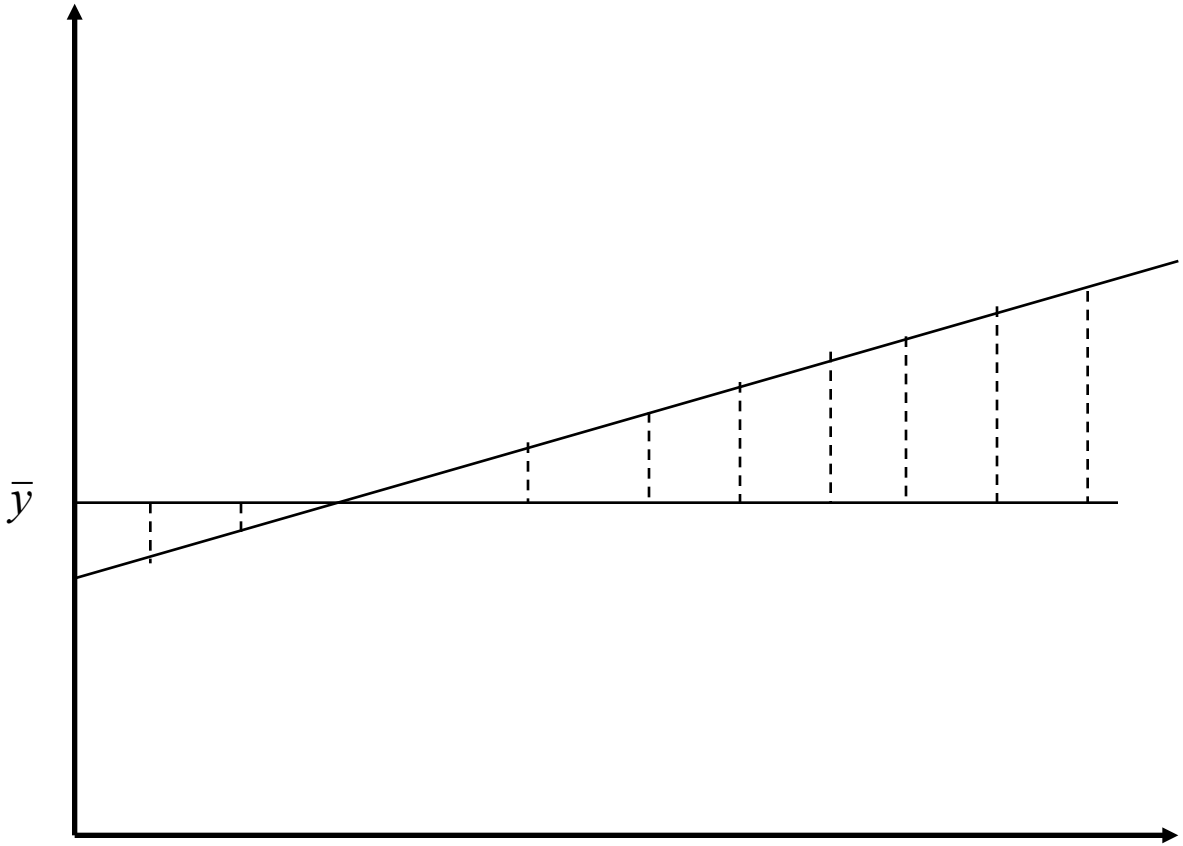
$$\underbrace{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}_{SSE}$$ Devianza residua



Misura della bontà di accostamento

$$\underbrace{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}_i\right)^2}_{SSR}$$

Devianza di regressione



Misura della bontà di accostamento

Dalla decomposizione possiamo ricavare un indice della bontà di accostamento tra i valori osservati e la retta di regressione (valori stimati) che può essere definito come:

$$R^2 = \frac{SSR}{S_{yy}} = \frac{devianza\ \ di\ \ regressione}{devianza\ \ totale}$$

Coefficiente di determinazione lineare

Misura della bontà di accostamento

Proprietà del coefficiente di determinazione:

$$0 \leq R^2 \leq 1$$

Il coefficiente di determinazione è uguale a “0” se e solo se:

$$\sum (y_i - \bar{y})^2 \neq 0 \quad \text{e} \quad \sum (\hat{y}_i - \bar{y})^2 = 0$$

Si dimostra che l’indice R^2 non è altro che il quadrato del coefficiente di correlazione lineare.

$$r = \frac{Cod(XY)}{\sqrt{Dev(X) \cdot Dev(Y)}} \qquad R^2 = r^2$$

Misura della bontà di accostamento

A partire dalla decomposizione della devianza totale si costruisce la seguente tabella:

Sorgente di Variazione	Somma dei Quadrati (Devianza)	Gradi di Libertà	Varianza	Rapporto F
Regressione	SSR	1	$MSR = SSR/1$	MSR/MSE
Residuo	SSE	$n-2$	$MSE = SSE/(n-2)$	
Totale	S_{yy}	$n-1$	$MST = S_{yy}/(n-1)$	

Predizione di nuove osservazioni e per un singolo valore di risposta

Un importante applicazione del modello di regressione è quello della previsione di future osservazioni corrispondenti ad un fissato valore della variabile indipendente.

Se x_0 è il valore della variabile indipendente allora:

$$\hat{y}_0 = b_0 + b_1 x_0$$

rappresenta la stima di un futuro valore della variabile di risposta.