01/10/2013

LEZIONE 1 REGRESSIONE LINEARE SEMPLICE E CORRELAZIONE

Definizione 1 Covarianza tra X e Y

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

$\mu_X = E[X]$, $\mu_Y = E[Y]$

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

proprietà:

- $Cov(X,X) = E[(X-\mu_X)^2] = Var[X] = E[X^2] - \mu_X^2$
- $Cov(X,Y) = E[XY] - \mu_X \mu_Y$
- $Cov(X,Y) \in (-\infty, +\infty)$ ⇝ la varianza non può essere negativa, mentre la covarianza sì

Definizione 2 Coefficiente di CORRELAZIONE

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$\sigma_X = \sqrt{Var[X]}$

Proposizione

① $-1 \le \rho(X,Y) \le 1$

② $Cov(\alpha X, \beta Y) = \alpha\beta\, Cov(X,Y)$

③ $\rho(\alpha X, \beta Y) = \rho(X,Y)$ $\quad \alpha, \beta > 0$

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

dim.

① → facoltativa

oss. preliminari

$E[\alpha X] = \alpha E[X] \longrightarrow \boxed{\mu_{\alpha X} = \alpha \mu_X}$ (il val. atteso è lineare)

$Var[\alpha X] = \alpha^2 Var[X] \longrightarrow \sqrt{Var[\alpha X]} = |\alpha| \sqrt{Var[X]} \longrightarrow \boxed{\sigma_{\alpha X} = |\alpha|\, \sigma_X}$

dim ②

$$Cov(\alpha X, \beta Y) = E[(\alpha X - \alpha\mu_X)(\beta Y - \beta \mu_Y)] = \alpha\beta E[(X-\mu_X)(Y-\mu_Y)] = \alpha\beta\, Cov(X,Y)$$

dim. ③

$$\rho(\alpha X, \beta Y) = \frac{Cov(\alpha X, \beta Y)}{\sigma_{\alpha X}\sigma_{\beta Y}} = \frac{\alpha\beta\, Cov(X,Y)}{|\alpha\beta|\,\sigma_X \sigma_Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \rho(X,Y)$$

la covarianza può assumere qualsiasi numero reale, la correlazione no. la covarianza risente dell'unità di misura: "cov = 57" non vuol dire nulla, la correlazione invece non cambia al variare dell'unità di misura. Quindi se dico "la correlazione è 0,8" ho significato (universale). Per scopi applicativi ρ è più importante di Cov e σ è più importante della varianza.

Proposizione

$$Cov(\alpha X + \beta Y + \gamma, Z) = \alpha\, Cov(X,Z) + \beta\, Cov(Y,Z)$$

↳ la costante sparisce

$Cov(Z, \alpha X + \beta Y + \gamma)$ = stessa formula "nove skiata" → la covarianza è SIMMETRICA

$[Var[\alpha X + b] = Var[\alpha X] = \alpha^2 Var[X]$ (b sparisce)$]$

↳ la varianza è invariante per traslazioni

↳ analogamente la covarianza (non facciamo la dim.)

TEOREMA

Se X e Y indipendenti $\Rightarrow$ $Cov(X,Y) = 0 = \rho(X,Y)$

DIM.

Indipendenza $\Rightarrow$ $E[XY] = \mu_X \mu_Y$ e siccome

$Cov(X,Y) = E[XY] - \mu_X \mu_Y$ è ovvio che sia $= \emptyset$

Nella pratica noi useremo questo risultato per dire che X e Y sono indipendenti (tenendo presente che non è effettivamente così, perché il teo enuncia la relazione inversa $\Rightarrow$)

Si dota numerici di dirò apposite "formule empiriche", con queste vedremo se è $= 0$ per dire che sono indipendenti (quelle usate finora sono "formule teoriche")

La covarianza è totalmente inadatta per l'uso pratico poiché o è $\emptyset$ fesso o non lo è (cioè non posso fare approssimazione a $\emptyset$ per valori come $0,1; \ldots$) perché la Cov dipende dall'unità utilizzata)

Se invece ρ calcolato su dati concreti risulta pari a $0,1$ si può approssimare a $\emptyset$ (ha senso farlo) → quindi userò la ρ (empirica) per valutare l'indipendenza.

DATI E VERSIONE EMPIRICA DI COV E ρ

	X	Y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
⋮	⋮	⋮
n	x_n	y_n

↓↓ "individui" (unità sperimentali)

$$\hat{Cov} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}_x\right)\left(y_i - \hat{\mu}_y\right)$$

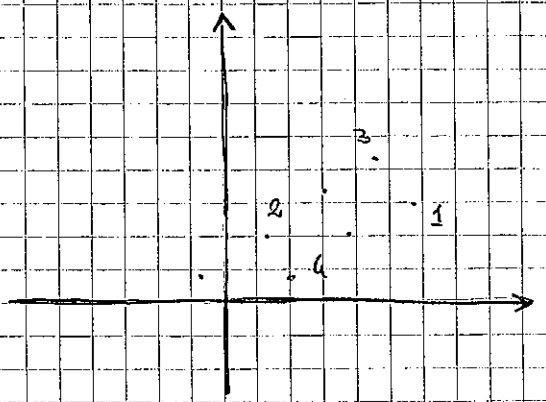
$$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2_x = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}_x\right)^2$$

$n-1$ (per avere uno stimatore non distorto della varianza, cioè con $n-1$ ho mediamente il vol. esatto)

$$\hat{\rho} = \frac{\sum_{i=1}^{n}\left(x_i - \hat{\mu}_x\right)\left(y_i - \hat{\mu}_y\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \hat{\mu}_x\right)^2 \sum_{i=1}^{n}\left(y_i - \hat{\mu}_y\right)^2}}$$

MASTER COPY
Tel. 050-8312126
Cell. 388 9837745

"plot (x, y)"



Oss

$\hat{Cov} > 0 \iff \hat{\rho} > 0$ (il segno è lo stesso)

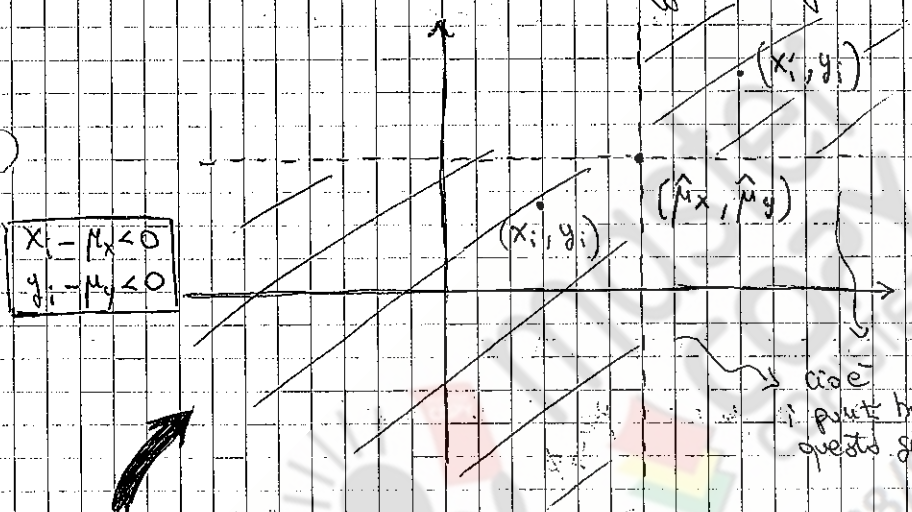
$\hat{Cov} > 0$ punti allineati in questo senso

$\hat{Cov} < 0$

$\hat{\rho} \sim 0$ (non posso dire $\hat{Cov} \sim 0$)

Spiegazione intuitiva, non è un teorema

Come "si legge" dalla formula?



$x_i - \mu_x > 0$
$y_i - \mu_y > 0$

$x_i - \mu_x < 0$
$y_i - \mu_y < 0$

cioè i punti hanno questo senso

(non vuol dire che sia la diagonale)

"Spieghiamo" $\hat{Cov} > 0 \rightarrow$ vuol dire che la somma è positiva.

Supponiamo inizialmente che TUTTI i termini da sommare siano positivi:

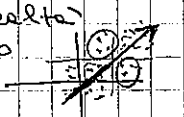
$(x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y) > 0$

Se un prodotto è positivo vuol dire che i fattori sono CONCORDI

2 casi → entrambi segno ⊕
entrambi segno ⊖

Se hanno segno ⊕ gli x_i, y_i sono nel I quadrante altrimenti nel terzo.

Avendo detto che il caso è quello → vogliamo l'ipotesi (non tutti i termini da sommare saranno positivi), quindi i punti saranno in realtà allineati PREVALENTEMENTE in quel senso



MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

REGRESSIONE LINEARE SEMPLICE

$Y = aX + b + \varepsilon$ $\quad Y, X, \varepsilon$ (3 variabili aleatorie)

È UN MODELLO

$X \rightarrow$ input, feature, predittore

$y \rightarrow$ output

$\varepsilon \rightarrow$ errore

ricavare (a) e (b) dai dati (?)

Ci è un modello teorico!

↳ Ipotesi teoriche su X, Y, ε ; hp fondamentale è che X e ε indipendenti

Ipotesi:

- X e ε indipendenti
- $Y = aX + b + \varepsilon$
- $\mu_\varepsilon = 0$ (convenzione)

$\exists$ solo 1 coppia a e b per cui siano valide queste hp

dalla seconda ipotesi $\Rightarrow \mu_Y = a\mu_X + b + \cancel{\mu_\varepsilon}^{\,\to 0}$ (linearità)

↓ ho trovato un legame tra tutti gli oggetti; a e b sono le incognite

$$Cov(Y,X) = Cov(aX+b+\varepsilon, X) = a\underbrace{Cov(X,X)}_{Var[X]} + \cancel{Cov(\varepsilon, X)}^{\,\to 0}$$

perché ho detto che X e ε sono indipendenti

$$a = \frac{Cov(Y,X)}{\sigma_X^2}$$

$$b = \mu_Y - a\mu_X$$

ho ricavato i coefficienti a e b univocamente

04/10/2013

Regressione Lineare Semplice

Tesi Se X, y, ε sono legate dalla relazione: $Y = aX + b + \varepsilon$ e $Cov(X, \varepsilon) = 0$, cioè X ed ε sono SCORRELATE

$E[\varepsilon] = 0$ (se non fosse ϕ la aggiungerei con la costante b), $\sigma_X > 0$

$\Rightarrow$ a, b, σ_ε^2 sono UNIVOCAMENTE DETERMINATE

$$a = \frac{Cov(y,X)}{\sigma_X^2}, \quad b = \mu_y - a\mu_x, \quad \sigma_\varepsilon^2 = \sigma_y^2 - a^2\sigma_x^2$$

Dim.

- $Cov(y, X) = a\sigma_x^2$ (da cui si ricava a)
- $\mu_y = a\mu_x + b$ (da cui si ricava b)
- $Var[y] = Var[aX + b + \varepsilon]$
  $= Var[aX + \varepsilon]$
  $= Var[aX] + Var[\varepsilon]$
  $= a^2 Var[X] + Var[\varepsilon]$ (da cui ricavo $\sigma_\varepsilon^2 = \sigma_y^2 - a^2\sigma_x^2$) □

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

Tabella dati:

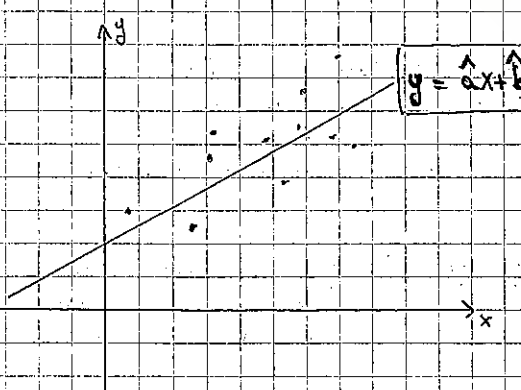
	X	Y
1	x_1	y_1
⋮		
n	x_n	y_n

$$\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}_X\right)^2$$

$$\hat{a} = \frac{\hat{Cov}(y,X)}{\hat{\sigma}_X^2}, \quad \hat{b} = \hat{\mu}_y - \hat{a}\hat{\mu}_x, \quad \hat{\sigma}_\varepsilon^2 = \hat{\sigma}_y^2 - \hat{a}^2\hat{\sigma}_x^2$$

giustificazione teorica? la vedremo formalmente per la regressione lineare multipla



$\hat{y} = \hat{a}x + \hat{b}$ RETTA DI REGRESSIONE

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

Ora voglio collegare il coefficiente di correlazione ρ con la regressione

Un collegamento è immediato: $a = \frac{Cov(X,y)}{\sigma_x^2} \to \frac{Cov(X,y)}{\sigma_x \cdot \sigma_y} \cdot \frac{\sigma_y}{\sigma_x} = \rho\frac{\sigma_y}{\sigma_x}$

Ho scoperto che c'è un legame tra ρ e il coefficiente angolare della retta di regres.