01/10/2013

LEZIONE 1 — REGRESSIONE LINEARE SEMPLICE E PRELIMINARI

Definizione 1 — Covarianza tra X e Y

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

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

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

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 valore di numero reale, la correlazione no. La covarianza risente del sistema di unità utilizzato: "cov = 57" non vuol dire nulla, la correlazione invece non cambia al variare delle unità utilizzate. 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

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

↳ la costante sparisce

$\mathrm{Cov}(Z, \alpha X + \beta Y + \gamma)$ = stessa formula "rovesciata" → la covarianza è SIMMETRICA

$[\mathrm{Var}[aX+b] = \mathrm{Var}[aX] = a^2\,\mathrm{Var}[X]$ (b sparisce)]

↳ la varianza è invariante per traslazioni

↳ analogamente la covarianza (vuol facciamo la dim.)

TEOREMA

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

DIM.

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

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

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

Sui dati numerici ci sono apposite "formule empiriche", con queste vediamo se è $= 0$ per dire che sono indipendenti (quelle viste finora sono "formule teoriche")

La covarianza è totalmente inadatta per l'uso pratico poiché o è $\emptyset$ stesso 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 sui dati concreti risulta pari a $0{,}1$ si può approssimare a $\emptyset$ (ha senso farlo) → quindi usiamo 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)

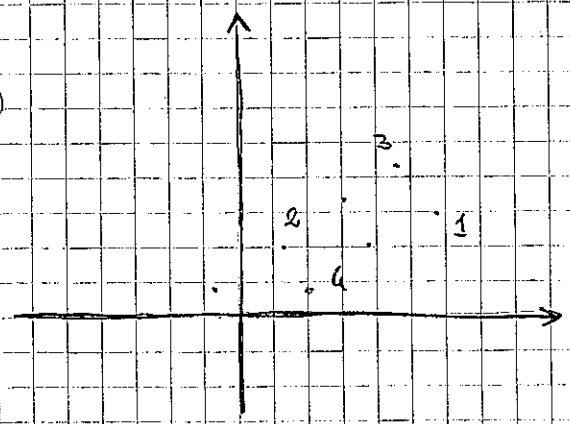
$$\hat{\mathrm{Cov}} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}_x\right)\left(y_i - \hat{\mu}_y\right)$$

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

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

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

"plot (x, y)"



Oss.

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

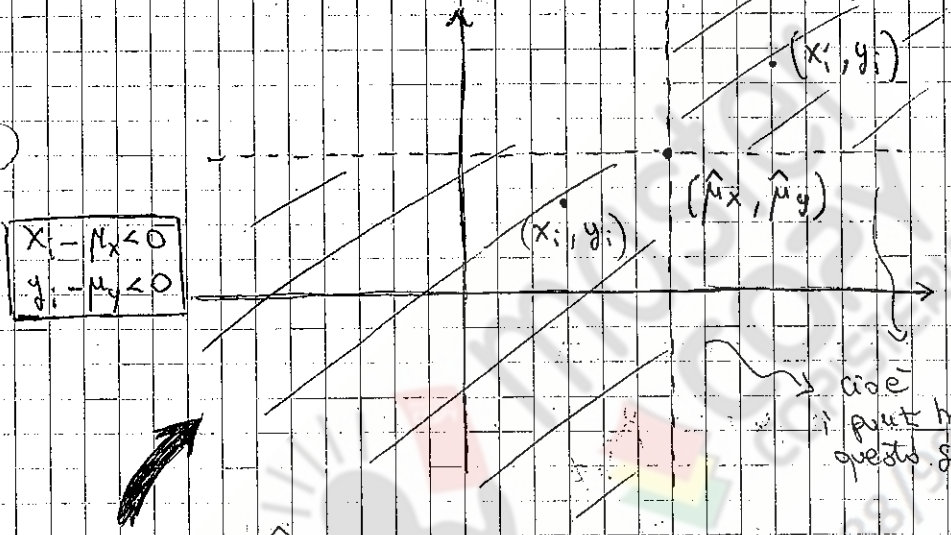
$\hat{Cov} > 0$ punti allineati in questo senso

$\hat{Cov} < 0$

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

Spiegazione intuitiva, non è un teorema

Come "leggo" 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" $\hat{Cov} > 0$ → 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 → vi ricordo l'ip iniziale (non tutti i termini da sommare saranno positivi), quindi i punti saranno in realtà allineati PREVALENTEMENTE in quel senso



REGRESSIONE LINEARE SEMPLICE

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

È UN MODELLO

$X \rightarrow$ input, fattore, predittore

$Y \rightarrow$ output

$\varepsilon \rightarrow$ errore

ricavare a e b dai dati (?)

C'è un modo teorico:

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

Ipotesi:

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

$\exists$ solo 2 numeri a e b per cui siano valide queste ipotesi

dalla seconda ipotesi $\Rightarrow \mu_Y = a\mu_X + b + \mu_\varepsilon$ (ricavata)

$\mu_\varepsilon = 0$

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

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

$Cov(\varepsilon, X) \rightarrow 0$ 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 aggiusterei con la costante b), $\sigma_X > 0$

$\Rightarrow$ $a, b, \sigma_\varepsilon^2$ 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 $\sigma_\varepsilon^2 = \sigma_Y^2 - a^2\sigma_X^2$) □

Tabella **dati**:

	X	Y
1	x_1	y_1
⋮	⋮	⋮
n	x_n	y_n

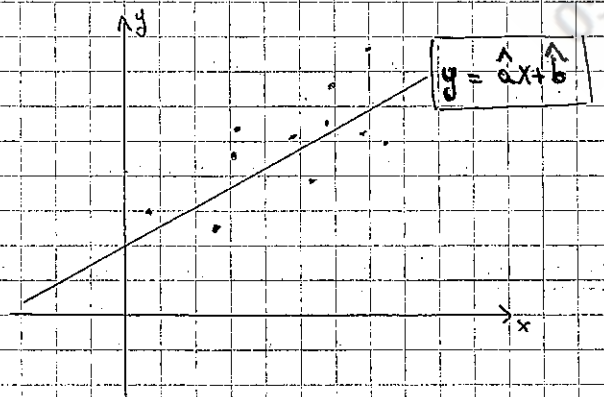
$$\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

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

giustificazione teorica?

La vedremo formalmente per la regressione lineare multipla



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

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.