CAPITOLO 1 - CORRELAZIONE E REGRESSIONE

COVARIANZA $\quad \boxed{Cov(X,Y) = E[(X-\mu_X)\cdot(Y-\mu_Y)]}$

VARIANZA $\quad \boxed{\sigma_X = Cov(X,X) = E[(X-\mu_X)^2]}$

COEFFICIENTE DI CORRELAZIONE $\quad \boxed{\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}}$

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

COVARIANZA E C. DI CORRELAZIONE EMPIRICI

$$\widehat{Cov} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

X e Y stessa numerosità

→ sono buone stime di quelli teorici, ad esempio grazie alla legge dei grandi numeri.

$$\boxed{-1 \le \hat{\rho} \le 1}$$

$\hat{\rho} \to 0 \Rightarrow$ sintomo di indipendenza

$\hat{\rho}$ non è affetto da variazione per scala, è ok.

$Cov \to 0 \Rightarrow$ non ci dice nulla, perché dipende dalla scala. Usare $\hat{\rho}$.

REGRESSIONE LINEARE SEMPLICE

Hyp: $\quad \boxed{Y = aX + b + \varepsilon}$ $\quad X, Y, \varepsilon$ v.a. ; $a, b \in \mathbb{R}$

X = predittore (input)

Y = output

ε = errore casuale ($E[\varepsilon] = 0$, se fosse $\neq 0$ sarebbe inglobato in b) $\quad$ H_{yp}: $Cov(X,\varepsilon) = 0$

SCOPI

1) Trovare a, b e σ_ε (approssimati) partendo da dati sperimentali.

2) Interpretare il concetto di ρ.

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

Sotto tale ipotesi abbiamo: (linearità di E)

$$E[Y] = E[aX + b + \varepsilon] \overset{\downarrow}{=} aE[X] + E[b] + E[\varepsilon] =$$

$$= aE[X] + E[b] \quad \text{, da cui}$$

con $E[b] = b$

$$Var[Y] = a^2 Var[X] + Var[\varepsilon] \qquad (Var[b] = 0)$$

e $Cov(Y,X) = Cov(aX + b + \varepsilon, X) = a\,Cov(X,X) + Cov(\varepsilon, X) =$ ($Cov(b,\ell) = 0$; perché ε e X sono scorrelate)

$$= a\,Var[X]$$

$$\Rightarrow \boxed{a = \frac{Cov(Y,X)}{Var[X]}} \Rightarrow \frac{a}{\sigma_Y} = \frac{Cov(Y,X)}{\sigma_Y \cdot \sigma_X^2} = \frac{\rho_{XY}}{\sigma_X}$$

$$\Rightarrow a = \frac{\sigma_Y}{\sigma_X} \cdot \rho_{XY}$$

$$\boxed{b = \mu_Y - a\mu_X}$$

$$\boxed{\sigma_\varepsilon^2 = \sigma_Y^2 - a^2\sigma_X^2}$$

$a > 0 \iff \rho_{XY} > 0$ legame DIRETTO

$a < 0 \iff \rho_{XY} < 0$ legame INVERSO

$$\sigma_\varepsilon^2 = \sigma_Y^2 - \frac{\sigma_Y^2 \cdot \rho_{XY}^2 \cdot \sigma_X^2}{\sigma_X^2} = \sigma_Y^2(1 - \rho_{XY}^2) \Rightarrow \boxed{\rho_{XY} \to \pm 1 \Rightarrow \sigma_\varepsilon^2 \to 0}\ !$$

ρ_{XY} non è tanto legato all'inclinazione della retta quanto non è legato alla precisione con cui la retta di regressione (il modello) descrivono il legame tra variabili. Se $\rho > 0$, allora la retta è inclinata positivamente, e viceversa.

È opportuno standardizzare l'errore: $\frac{\sigma_\varepsilon^2}{\sigma_Y^2} = 1 - \rho_{XY}^2$

Standardizzare una v.a. X significa fare

$$\tilde{X} = \frac{X - \mu_X}{\sigma_X}$$

→ sottrae la media (centrare)

→ divide per dev. std. (portare tra -1 e 1)

Usando la semplice retta di regressione $\hat{y} = \hat{a}x + \hat{b}$ non riusciremo a spiegare completamente Y, il modello spiega solo parte di Y, in particolare NON SPIEGA $Y - aX + b$, la cui varianza è $\frac{\sigma_\varepsilon^2}{\sigma_Y^2}$ (std.)

mentre la varianza spiegata è $\boxed{1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2}}$, ovvero ρ_{XY}^2

Quindi ρ^2 è indice anche di bontà del modello.

MATRICE DI COVARIANZA

$$Q = \begin{pmatrix} \sigma_x^2 & Cov(X,Y) \\ Cov(X,Y) & \sigma_Y^2 \end{pmatrix}$$

Se invece di avere il vettore (X,Y) avessi avuto il vettore $X = (X_1, \dots, X_m)$, la matrice sarebbe stata $m \times m$ con $(Cov(X_i, X_i) = \sigma_i^2)$

$Q_{ij} = Cov(X_i, X_j)$ (definita per $i = j$,

MATRICE DI CORRELAZIONE

$$\rho = \begin{pmatrix} 1 & \rho(X,Y) \\ \rho(X,Y) & 1 \end{pmatrix}$$

Stesse considerazioni di sopra

Le proprietà di Q sono:
- SIMMETRIA $(Q_{ij} = Q_{ji})$
- NON-NEGATIVITÀ $(x^T Q x \geq 0)$

TEOREMA SPETTRALE

Se Σ è una matrice simmetrica $n \times n$ (come Q), allora esiste una base ortonormale $\{e_1, \dots, e_m\}$ (con $e_1 \perp e_2 \perp \dots \perp e_m$, $|e_i| = 1$, $i = 1, \dots, m$) di $\mathbb{R}^m$ i cui elementi sono autovettori di Σ: per opportuni numeri $\lambda_1, \dots, \lambda_m$ vale

$$\Sigma e_i = \lambda_i e_i \qquad i = 1, \dots, m$$

con $\lambda_1 \geq \dots \geq \lambda_m$ (ricordo)

Sia $U = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$ (non è importante), e sia $\Sigma_e = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \lambda_m \end{pmatrix}$ (autovalori di Q)

$|e_i| = 1 = \sqrt{e_{i1}^2 + \dots + e_{im}^2} = 1$

allora $\Sigma = U \cdot \Sigma_e \cdot U^T$, ovvero Σ (che può essere Q) è DIAGONALIZZABILE

$$\boxed{Q = U Q_e U^T}$$

Q_e è la matrice diagonale che ha i valori degli autovalori di Q.

$Q_{e_i} = \lambda_i e_i$

$\exists$ una matrice simmetrica definita non negativa $\sqrt{Q}$: $(\sqrt{Q})^2 = Q$

MATRICE COV. EMPIRICA

	X_1	.	X_p
1	$x_{1,1}$		$x_{1,p}$
2	$x_{2,1}$		$x_{2,p}$
⋮			
n	$x_{n,1}$		$x_{n,p}$

(colonne)

$X_1, \dots, X_p$ possono essere variabili come PIL, spese per istruzione, per la sanità ecc.

$1, 2, \dots, n$ possono essere righe (individui) di un'indagine, come le nazioni europee.

$$\hat{Q}_{ij} = \frac{1}{n} \sum_{k=1}^{n} (x_{k,i} - \bar{x}_i)(x_{k,j} - \bar{x}_j)$$

(colonne)

Con $\bar{x}_i = \frac{1}{n} \sum_{k=1}^{n} x_{k,i}$

Le matrici Q vengono "piccoline" rispetto alla tabella iniziale!

$$\hat{\rho}_{ij} = \frac{\sum_{k=1}^{n} (x_{k,i} - \bar{x}_i)(x_{k,j} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n} (x_{k,i} - \bar{x}_i)^2 \sum_{k=1}^{n} (x_{k,j} - \bar{x}_j)^2}}$$

→ misura il legame lineare tra la colonna i e la colonna j

TRASFORMAZIONE LINEARE (AFFINE) DEI MOMENTI

Sia $X = (X_1, \dots, X_m)$ vettore casuale,
A = matrice $n \times d$ ($A : \mathbb{R}^m \to \mathbb{R}^d$)
$Y = (Y_1, \dots, Y_d)$ vettore casuale : se $Y = AX + b$

Allora $\mu_Y = A\mu_X + b$ e se Q^X e Q^Y rappresentano le matrici di covarianza, si ha:

$$Q^Y = A Q^X A^T$$

Il plot di una tabella mostra le varie correlazioni tra le colonne

plot(V2, V1)

V1			
	V2		
		V3	
			V4

questa casella è la con (V_1, V_2)

(è matrice simmetrica quindi uguale a (V_2, V_1) ; i plot sono invertiti)

REGRESSIONE LINEARE MULTIPLA

Abbiamo una tabella di questo tipo:

	X_1	...	X_p	Y
1	$x_{1,1}$	...	$x_{1,p}$	y_1
2	$x_{2,1}$	...	$x_{2,p}$	y_2
...	...	...	...	...
n	$x_{n,1}$	...	$x_{n,p}$	y_n

$X_1, \dots, X_p, Y$ sono delle variabili (PIL, PLLC, spese ....)

$1, 2, \dots, n$ sono unità sperimentali (persone, offerte città, o regioni).

$X_1, \dots, X_p$ sono input, fattori o predittori

Y è l'output. Vogliamo, partendo da un modello, predire il valore di Y in corrispondenza di input mai sperimentati.

Il nostro modello è

$$Y = a_1 X_1 + \dots + a_p X_p + b + \varepsilon \qquad (1)$$

Questa NON è un'ipotesi, è sempre vera, il problema è quanto sono grandi i vari ε_i $i = 1, \dots, n$. Infatti più grandi sono i residui, più impreciso risulta il modello. Calcoleremo degli $\hat{a}_1, \dots, \hat{a}_p, \hat{b}$ ottimali.

Allora gli ε_i saranno determinati dai dati sperimentali come:

$$\hat{\varepsilon}_i = y_i - (\hat{a}_1 x_1 + \dots + \hat{a}_p x_p + \hat{b})$$

La (percentuale di) varianza spiegata sarà

$$R^2 = 1 - \frac{\sigma^2_\varepsilon}{\sigma^2_Y}$$

Ci sono 3 metodi per calcolarsi i coefficienti del modello ← 1 TEORICO, 2 MINIMI QUADRATI, MACH, 3 " " : GEOMETRICO

Per l'AA 2014/2015, ci interessa solo il teorico.

METODO TEORICO

TEO Supponiamo che $Y, X_1, \dots, X_p$ ~~ed~~ ε siano legate dalla relazione (1).

Supponiamo ε scorrelata da ciascuna delle X_i $i = 1, \dots, p$, e che $E[\varepsilon] = 0$. Supponiamo Q_X invertibile, dove Q_X è la matrice di covarianza del vettore $X = X_1, \dots, X_p$.

Allora (vedi pag. seguente)