Piano Nazionale Lauree Scientifiche (PLS 2016-2017)

Statistica Laboratorio di Statistica

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

DIPARTIMENTO DI SCIENZE ECONOMICHE E METODI MATEMATICI

Le relazioni tra variabili

prof.ssa Angela Maria D'Uggento angelamaria.duggento@uniba.it

materiale didattico I incontro

Dall'analisi statistica esplorativa ai modelli

A seconda del numero di variabili considerate, è possibile effettuare un'analisi

- UNIVARIATA
- BIVARIATA
- MULTIVARIATA

Analisi statistica univariata

Considera una sola variabile quantitativa/qualitativa per volta, ad esempio età, voto, numero iscritti, altezza, residenza, titolo studio, genere, ecc.

Gli strumenti di analisi:

1. Medie e rappresentazioni grafiche
2. Variabilità
3. Numeri indici

Analisi statistica bivariata

Studia due variabili quantitative/qualitative contemporaneamente, ad esempio peso/altezza, voto materia X/voto materia Y, numero iscritti/sesso, titolo di studio/voto di laurea, risparmio/consumo, ecc.

Gli strumenti:

1. Analisi della dipendenza (Regressione semplice)
2. Analisi della interdipendenza (Correlazione semplice)
3. Analisi dell'indipendenza (χ^2)

Analisi statistica multivariata

Studia l'influenza contemporanea di tre o più variabili

Tecniche statistiche:

1. Analisi fattoriale
2. Cluster analysis
3. Scaling Multidimensionale
4. Regressione multipla
5. Correlazione canonica
6. Analisi corrispondenze

Regressione, Correlazione, Indipendenza

REGRESSIONE → Analisi della **DIPENDENZA** tra Y (dipendente) in funzione di due o più variabili statistiche indipendenti $X_1, X_2, \ldots X_n$ (regressori)

CORRELAZIONE → Analisi della **INTERDIPENDENZA ASSOCIAZIONE** tra due o più variabili statistiche $X_1, X_2, \ldots X_n$

INDIPENDENZA → Il variare di una delle due variabili statistiche X_1 e X_2 **NON INFLUENZA,** non produce alcun effetto sull'altra

Regressione

Supponiamo di essere interessati alla relazione tra cifra investita in pubblicità (X) e il ritorno in termini di vendite (Y).

In realtà ci sono altri fattori che influenzano il valore di Y per un dato valore di X quali tipo prodotto, fattori economici, luogo di vendita, ecc.

Nella raccolta dati si rilevano le coppie di osservazioni (x_i, y_i) per ottenere una serie di *n* coppie di dati:

$$(x_1, y_1)\ (x_2, y_2), \ldots (x_i, y_i) \ldots (x_n, y_n).$$

Il grafico che le rappresenta è il grafico a **dispersione o scatter** plot

Altri esempi:

X= dimensione della casa (in m^2); Y=prezzo di vendita

X = n.ro clienti contattati; Y = n.ro contratti conclusi

X= reddito; Y=consumo

Regressione

Esempio: un'agenzia immobiliare rileva i seguenti dati:

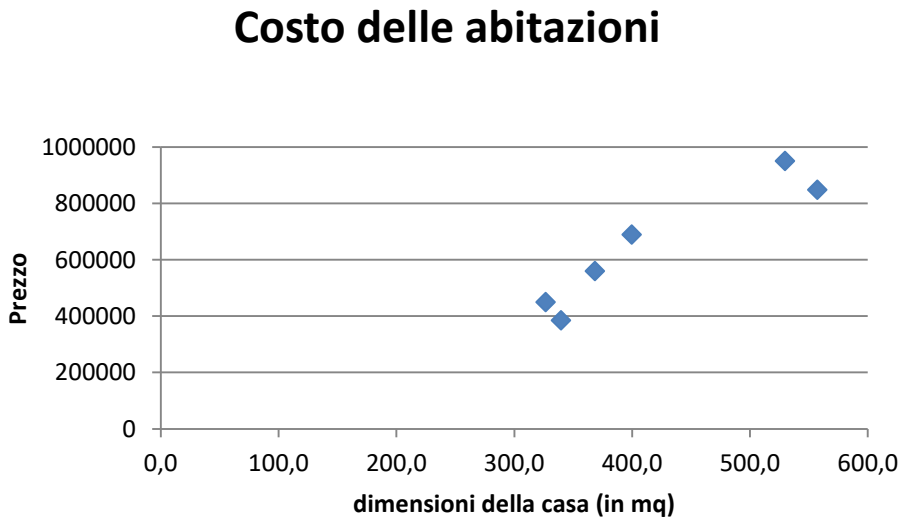
abitazione	prezzo (Euro)	dimensione (mq)	N.ro camere letto	N.ro servizi
1	689.500	400,0	4	3,5
2	385.000	340,0	5	3,0
3	449.900	326,9	4	2,5
4	949.900	530,0	5	4,0
5	848.000	557,5	4	3,5
6	559.900	368,7	4	3,5

Lo scopo della regressione lineare semplice è quello di individuare la eventuale relazione tra le due variabili (rappresentata graficamente dal grafico a dispersione o scatter) esplicitandola attraverso una funzione matematica. Tale modello matematico predice il valore di Y come una funzione lineare della variabile indipendente X.

I modelli di regressione multipla predicono il valore di Y come funzione di una serie di variabili indipendenti X_1, X_2 …. X_n

Regressione

Per verificare la presenza di una relazione è opportuno fare il grafico e, quindi, individuare il modello ed i suoi parametri



La variabile Y viene detta **variabile risposta** (o **variabile dipendente**), la variabile X viene detta **variabile esplicativa** (o **variabile indipendente** o **regressore**).

Regressione lineare semplice

La relazione tra le due variabili è, per semplicità, supposta lineare, pertanto sarà studiata attraverso la funzione della retta di regressione che, con dati campionari, è:

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

Variazione casuale dovuta ad altri fattori non misurabili

Intercetta.
Valore di Y per x=o

Coefficiente angolare. Varia tra $+\infty$ e $-\infty$.
Indica come varia in media Y al variare unitario di X.

Regressione lineare

La stima dei parametri incogniti della funzione di regressione avviene con il **metodo dei minimi quadrati.** Tale metodo consiste nel rendere minima la differenza al quadrato tra valori teorici (valori da modello) e valori empirici (dati osservati)

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

Regressione lineare

Uguagliando a zero le derivate parziali rispetto ai due parametri b_0 e b_1, si giunge al sistema risolutivo con le formule :

$$\begin{cases} b_0 = \bar{y} - b_1 \bar{x} \\ b_1 = \dfrac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} = \dfrac{Codev(X,Y)}{Dev(X)} = \dfrac{Co\,\mathrm{var}(X,Y)}{Var(X)} \end{cases}$$

I parametri della retta di regressione

$\mathbf{b_0}$ è l'intercetta o termine noto ed esprime il valore di Y quando x=0;

$\mathbf{b_1}$ è il coefficiente angolare della retta e quindi

$$-\infty \leq b \leq +\infty$$

E' anche detto coefficiente di regressione ed esprime la variazione media del carattere Y al variare **unitario** del carattere X.

Se $\mathbf{b_1}$ >0 c'è dipendenza diretta tra X e Y, cioè Y aumenta in media all'aumentare di X;

se $\mathbf{b_1}$ <0 c'è dipendenza inversa tra X e Y, cioè Y diminuisce in media all'aumentare di X;

se $\mathbf{b_1}$ =0 vi è indipendenza di Y da X .

Interpolazione ed estrapolazione

L'interpolazione si usa per predire i valori di Y servendosi dei valori di X che si trovano all'interno dell'intervallo dei dati.

L'estrapolazione si usa per predire i valori di Y servendosi dei valori di X che si trovano all'esterno dell'intervallo dei dati.

La prima è più attendibile perché non abbiamo garanzia dell'invarianza del comportamento del fenomeno al di fuori dell'intervallo.

Esempio di calcolo dei parametri della retta di regressione

X n.ro contatti	Y n.ro iscritti	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
20	14				
25	18				
30	26				
40	32				
50	38				
				Codev (X,Y)	**Dev (X)**

Per lo svolgimento si veda file Excel presente nel materiale didattico

Usare Excel

Per stimare i parametri

Intercetta (per calcolare b_0)

Pendenza (per calcolare b_1)

Per visualizzare equazione ed R^2:

Selezionare una delle coppie di punti

Premendo il tasto destro del mouse /aggiungi linea di tendenza/visualizza equazione/visualizza R^2

Correlazione semplice

La correlazione misura l'interdipendenza tra due caratteri X_1 ed X_2. in termini di concordanza o discordanza.

In tal caso non è possibile distinguere il carattere dipendente da quello indipendente.

Una misura assoluta della concordanza/discordanza è la codevianza.

$$Codev(X,Y) = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

La correlazione si misura con il coefficiente di correlazione r di Bravais Pearson

Correlazione semplice

Codevianza

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

=+1 max concordanza

=0 indifferenza

=-1 max discordanza

Devianza di X e Devianza di Y

$$-1 \leq r \leq +1$$

Il coefficiente di correlazione è anche denotato con ρ ed ha la stessa interpretazione

Relazione tra retta di regressione e r (coefficiente di correlazione)

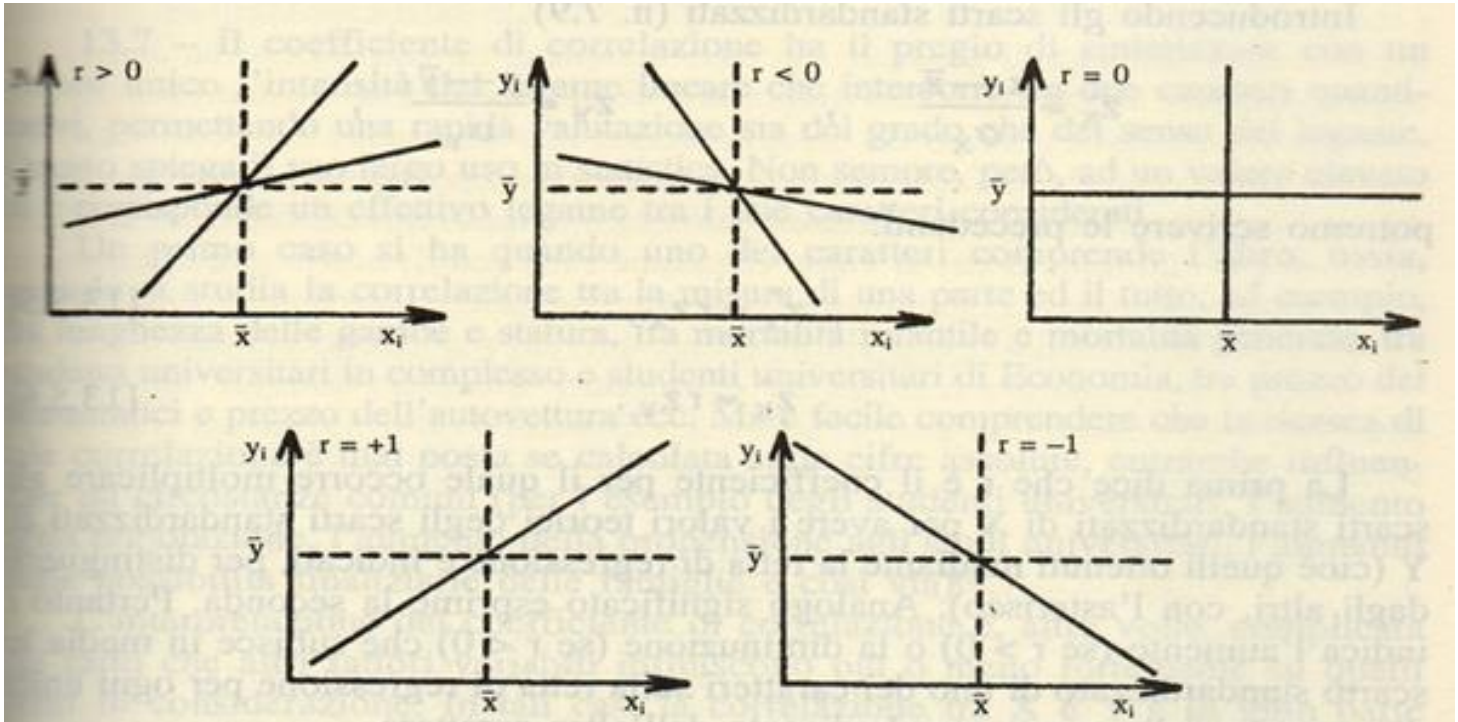
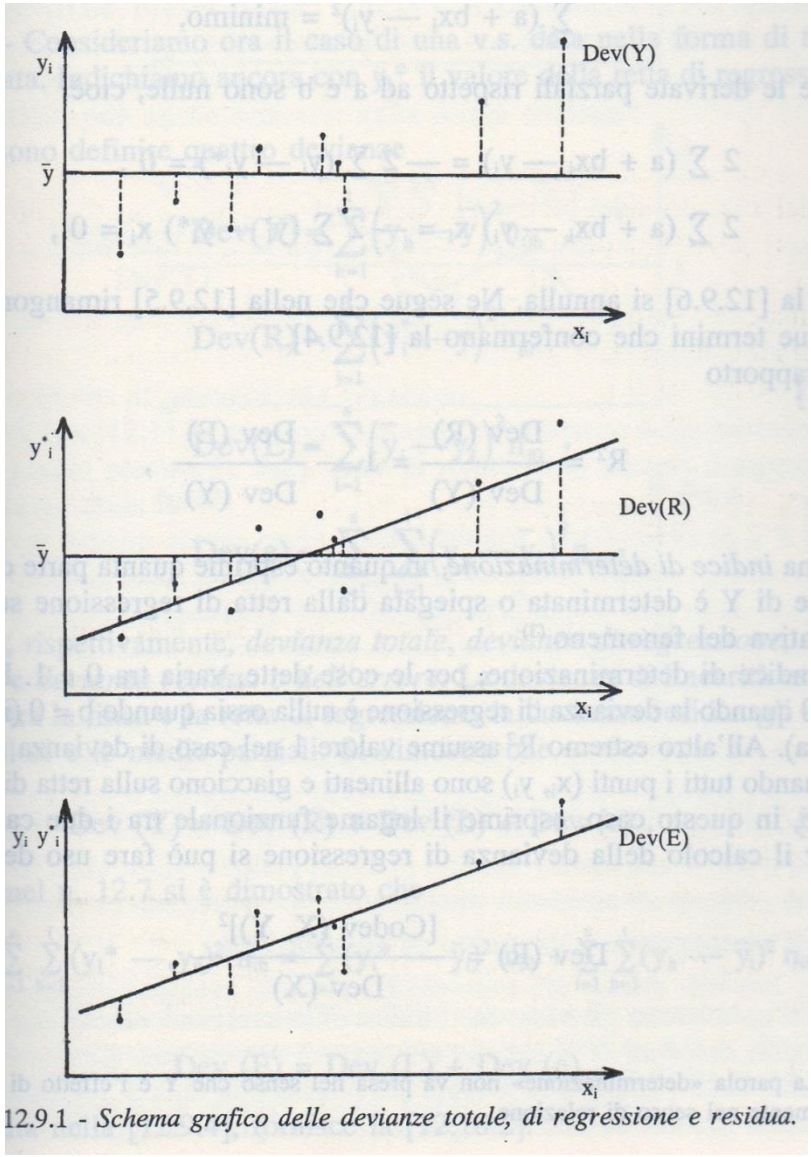


Fig. 13.4.1 - *Posizioni schematiche delle rette di regressione per vari valori del coefficiente di correlazione.*

Varianza di regressione



12.9.1 - *Schema grafico delle devianze totale, di regressione e residua.*

Analogamente a quanto detto per la media, è possibile studiare la dispersione dei valori osservati di Y intorno alla retta di regressione.
Dispersione elevata significa limitata attendibilità delle previsioni fatte con il modello scelto (retta di regressione); il contrario se la dispersione è bassa.
L'adattamento si misura con R^2 .
Le tre devianze sono date rispettivamente dalla somma dei quadrati dei segmenti verticali tratteggiati.
Si dimostra che
Dev(Y)=Dev(R)+Dev(E)

Indice di determinazione R^2

$$R^2 = \frac{Dev(R)}{Dev(Y)} = \frac{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}$$

R^2 esprime la bontà di adattamento del modello di regressione cioè quanta parte della devianza totale di Y è spiegata dalla retta scelta

$$R^2 = \frac{Dev(R)}{Dev(Y)} = 1 - \frac{Dev(E)}{Dev(Y)} = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}$$

$$0 \le R^2 \le 1$$

R^2=+1 quando Dev(E)=0 cioè tutti i punti sono perfettamente allineati su retta regressione

R^2 =0 quando Dev(R)=0 cioè b_1=0 cioè indipendenza in media