

## ***ANALISI DELLA REGRESSIONE***

L'**Analisi della Regressione** riguarda lo studio delle relazioni esistenti fra 2 o più caratteri quantitativi o variabili. La ricerca dei legami esistenti fra più variabili si pone come ricerca delle relazioni funzionali che pongono  $Y$  come grandezza dipendente da una serie di variabili indipendenti  $X_1, X_2, \dots, X_r$ :

$Y=f(X_1, X_2, \dots, X_r)$  (Analisi Multivariata della Regressione)

$Y=f(X)$  (Analisi Bivariata della Regressione)

## ***ANALISI BIVARIATA DELLA REGRESSIONE***

**I FASE:** rilevazione congiunta di *valori osservati* relativi alle due variabili  $X$  e  $Y$ .

**II FASE:** individuazione del tipo di funzione matematica (*funzione teorica*) adatta ad interpretare la realtà osservata.

**III FASE:** stima dei parametri incogniti della funzione teorica.

**IV FASE:** valutazione della bontà d'accostamento del modello teorico alla realtà osservata.

**I FASE:** consiste nella rilevazione congiunta delle due variabili  $X$  e  $Y$  su un campione di unità statistiche. Il campione deve essere di dimensione  $n$  sufficientemente elevata e rappresentativo.

Dal punto di vista operativo, questa I FASE porta alla costruzione di una Tabella di *valori osservati*:

- Nel caso di dati semplici:

X	$x_1$	$x_2$	...	$x_i$	...	$x_n$
Y	$y_1$	$y_2$	...	$y_i$	...	$y_n$

- Nel caso di dati ponderati:

X	Y						
	Y <sub>1</sub>	Y <sub>2</sub>	...	Y <sub>j</sub>	...	Y <sub>h</sub>	
X <sub>1</sub>	f <sub>11</sub>	f <sub>12</sub>	...	f <sub>1j</sub>	...	f <sub>1h</sub>	f <sub>1•</sub>
X <sub>2</sub>	f <sub>21</sub>	f <sub>22</sub>	...	f <sub>2j</sub>	...	f <sub>2h</sub>	f <sub>2•</sub>
:	:	:		:		:	:
:	:	.		:		:	:
X <sub>i</sub>	f <sub>i1</sub>	f <sub>i2</sub>	...	f <sub>ij</sub>	...	f <sub>ih</sub>	f <sub>i•</sub>
:	:	:		:		:	:
:	:	:		:		:	:
X <sub>k</sub>	f <sub>k1</sub>	f <sub>k2</sub>	...	f <sub>kj</sub>	...	f <sub>kh</sub>	f <sub>k•</sub>
	f <sub>•1</sub>	f <sub>•2</sub>	...	f <sub>•j</sub>	...	f <sub>•h</sub>	N

**II FASE:** consiste nella scelta della funzione matematica (*funzione teorica* o *funzione interpolante*) che sia in grado di interpretare la realtà osservata.

La scelta viene fatta sulla base delle conoscenze specifiche del fenomeno in esame oppure sulla base dell'esame della rappresentazione grafica dei valori osservati.

**III FASE:** consiste nella stima dei **parametri** incogniti della funzione teorica  $Y=f(X)$ . A tal fine esistono 3 metodi: il *Metodo dei Momenti*, il *Metodo della Massima Verosimiglianza* e il *Metodo dei Minimi Quadrati*.

### ***Metodo dei minimi quadrati***

Si applica solamente a *funzioni teoriche* che siano lineari nei parametri.

Si supponga di rilevare  $n$  coppie di valori osservati  $(x_i, y_i)$  e di proporre una *funzione matematica teorica*  $Y$ , definita da  $(s+1)$  parametri incogniti, quale interpretazione della relazione che collega le due variabili:

$$Y = f(x; a_0, a_1, \dots, a_s)$$

Il *Metodo dei minimi quadrati* permette di stimare i parametri del modello teorico  $Y$  e consiste nel minimizzare la quantità  $S$ :

$$\min S = \sum_{i=1}^n (y_i - Y_i)^2 = \sum_{i=1}^n [y_i - f(x_i; a_0, a_1, \dots, a_s)]^2$$

Come *funzione teorica* si consideri il polinomio di ordine  $s$ :

$$Y = a_0 + a_1x + a_2x^2 + a_3x^3 + \dots + a_sx^s$$

Per stimare gli  $(s+1)$  parametri incogniti, si applichi il *Metodo dei minimi quadrati*:

$$\min S = \sum_{i=1}^n (y_i - Y_i)^2 = \sum_{i=1}^n (y_i - a_0 - a_1x_i - a_2x_i^2 - a_3x_i^3 - \dots - a_sx_i^s)^2$$

Allora il minimo di  $S$  si ottiene ponendo uguali a *zero* le derivate parziali di  $S$  rispetto ai parametri incogniti; si ricava quindi un sistema di  $(s+1)$  equazioni in  $(s+1)$  incognite, appunto i parametri.

Nel caso particolare in cui il modello teorico scelto sia rappresentato da una **RETTA**:

$$y' = a + bx$$

Allora per stimare i due parametri **a** e **b** con il *Metodo dei minimi quadrati* si deve minimizzare **S** facendo le derivate parziali di **S** rispetto ad **a** e rispetto a **b**:

$$\min S = \sum_{i=1}^n (y_i - Y_i)^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases}$$

Per dati SEMPLICI:

$$\begin{cases} na + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \end{cases}$$

Per dati PONDERATI:

$$\begin{cases} Na + b \sum_{i=1}^n x_i f(x_i) = \sum_{j=1}^m y_j f(y_j) \\ a \sum_{i=1}^n x_i f(x_i) + b \sum_{i=1}^n x_i^2 f(x_i) = \sum_{j=1}^m \sum_{i=1}^n x_i y_i f_{ij} \end{cases}$$

In generale, sempre se il modello teorico è una **RETTA**, valgono anche le seguenti formule per la determinazione dei parametri incogniti **a** e **b**:

$$\begin{cases} b = \frac{M(X \cdot Y) - m_x m_y}{\sigma_x^2} \\ a = m_y - b m_x \end{cases}$$

**IV FASE:** consiste nella valutazione della bontà d'accostamento del modello teorico alla realtà osservata. Si basa sul confronto fra i valori osservati  $y_i$  e i valori teorici  $Y_i$ . Esistono vari Indici di accostamento:

$$\text{Errore medio di interpolazione} \quad s = \sqrt{\frac{\sum_{i=1}^n (y_i - Y_i)^2}{n-2}}$$

$$\text{Media aritmetica degli scarti assoluti} \quad \xi = \frac{\sum_{i=1}^n |y_i - Y_i|}{n}$$

**Coefficiente di determinazione**

$$R^2 = \frac{\text{Dev.regressione}}{\text{Dev.totale}} = \frac{\sum_{i=1}^n (Y_i - m_y)^2}{\sum_{i=1}^n (y_i - m_y)^2}$$

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

## ***ANALISI DELLA CORRELAZIONE***

L'**Analisi della Correlazione** studia il legame esistente fra due caratteri quantitativi o variabili.

### ***Il Coefficiente di Correlazione Lineare***

Per valutare la correlazione esistente fra le due variabili X e Y si utilizza l'indice chiamato "***Coefficiente di Correlazione Lineare***" (o "***Covarianza normalizzata***") e indicato con la lettera *erre minuscola*:

$$r = \frac{Cov(X, Y)}{\sigma_x \cdot \sigma_y}$$

Tale indice esprime la relazione lineare esistente fra le due variabili ed è un numero sempre compreso fra -1 e +1:

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

- *r* assume valore **-1** se fra X e Y esiste una perfetta relazione lineare inversa;
- *r* assume valore **0** se fra X e Y non esiste una relazione di tipo lineare (potrebbe esistere una relazione di tipo diverso);
- *r* assume valore **+1** se fra X e Y esiste una perfetta relazione lineare diretta.

***Covarianza*** fra X e Y:

$$Cov(X, Y) = M \left[ (X - m_x)(Y - m_y) \right] = M(X \cdot Y) - m_x m_y$$

Esistono altri procedimenti per il **calcolo** di *r*:

$$r = b \frac{\sigma_x}{\sigma_y}$$

$$r = \sqrt{b \cdot b'} \text{ dove } b' \text{ è il coefficiente angolare della retta } X = a' + b' y \text{ (} X = f(Y)\text{)}$$

Osservazione: se la funzione teorica  $Y = f(X)$  scelta per l'interpretazione della realtà è la RETTA, allora il quadrato del ***Coefficiente di Correlazione Lineare*** coincide con il ***Coefficiente di determinazione***:

$$(r)^2 = R^2$$

