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, …, X_r:
$Y=f(X_1, X_2, \ldots, X_r)$ (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:

	Y						
X	Y_1	Y_2	…	Y_j	…	Y_h	
X_1	f_{11}	f_{12}	…	f_{1j}	…	f_{1h}	$f_{1\bullet}$
X_2	f_{21}	f_{22}	…	f_{2j}	…	f_{2h}	$f_{2\bullet}$
:	:	:		:		:	:
:	:	.		:		:	:
X_i	f_{i1}	f_{i2}	…	f_{ij}	…	f_{ih}	$f_{i\bullet}$
:	:	:		:		:	:
:	:	:		:		:	:
X_k	f_{k1}	f_{k2}	…	f_{kj}	…	f_{kh}	$f_{k\bullet}$
	$f_{\bullet 1}$	$f_{\bullet 2}$	…	$f_{\bullet j}$	…	$f_{\bullet h}$	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:

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

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

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

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

$$Y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + ... + a_s x^s$$

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

$$\min S = \sum_{i=1}^{n} \left(y_i - Y_i \right)^2 = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - a_2 x_i^{\,2} - a_3 x_i^{\,3} - ... - a_s x_i^{\,s} \right)^2$$

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.

ESERCIZIO 1

Sui seguenti valori osservati:

x	y
0	5
1	10
2	20
3	42
4	8
5	2

stimare con il metodo dei minimi quadrati i parametri della funzione teorica $Y = a + bx^2$.

$$\begin{cases} na + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i \end{cases}$$

x	y	x2	x4	x^2*y
0	5	0	0	0
1	10	1	1	10
2	20	4	16	80
3	42	9	81	378
4	8	16	256	128
5	2	25	625	50
	87	55	979	646

$$\begin{cases} 6a + 55b = 87 \\ 55a + 979b = 646 \end{cases}$$

$$\begin{cases} a = 17{,}4247 \\ b = -0{,}3191 \end{cases}$$

Da cui il modello teorico risulta il seguente: $Y=17{,}4247-0{,}3191x^2$

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***:

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

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

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:

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

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

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

ESERCIZIO 2

Sui seguenti valori osservati:

x	3	0	2	1	5	2	0	4	2	3
y	-2	0.5	-1	0	-4	-2	1	-3	-1.5	-1

stimare con il metodo dei minimi quadrati i parametri della retta di regressione $Y=a+bx$

x	y	x^2	xy
3	-2	9	-6
0	0,5	0	0
2	-1	4	-2
1	0	1	0
5	-4	25	-20
2	-2	4	-4
0	1	0	0
4	-3	16	-12
2	-1,5	4	-3
3	-1	9	-3
22	-13	72	-50

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

$$\begin{cases} 10a + 22b = -13 \\ 22a + 72b = -50 \end{cases}$$

$$\begin{cases} a = 0{,}695 \\ b = -0{,}907 \end{cases}$$

Oppure:

$$\begin{cases} b = \dfrac{Cov(x,y)}{\sigma_x^2} = \dfrac{M(X \cdot Y) - m_x m_y}{\sigma_x^2} \\ a = m_y - b m_x \end{cases}$$

Dove:

$$M(X \cdot Y) = \frac{(-50)}{10} = -5$$

$$m_x = \frac{22}{10} = 2{,}2$$

$$m_y = \frac{(-13)}{10} = -1{,}3$$

$$\sigma_x^2 = \frac{72}{10} - (2{,}2)^2 = 2{,}36$$

Quindi sostituendo si ottiene:

$$\begin{cases} b = \dfrac{-5 - (2{,}2)\cdot(-1{,}3)}{2{,}36} = \dfrac{(-2{,}14)}{2{,}36} = -0{,}907 \\ a = -1{,}3 - b \cdot 2{,}2 = 0{,}695 \end{cases}$$

Quindi la retta di regressione è:

$$Y = 0{,}695 - 0{,}907 \cdot x$$

ESERCIZIO 3

Su un campione di 100 consumatori è stata svolta un'indagine sul consumo mensile di un certo prodotto, rilevando la variabile ***X*** *quantità totale di prodotto consumata durante il mese* e la variabile ***Y*** *numero di volte in cui è stato acquistato il prodotto durante il mese*. I risultato ottenuti sono riportati nella seguente tabella a doppia entrata:

	Y			
X	0	2	5	9
0 - 2	14	20	6	0
2 - 4	6	9	7	8
4 - 8	0	6	12	12

Su tali valori osservati stimare con il metodo dei minimi quadrati i parametri della retta di regressione $Y=a+bx$.

$$\begin{cases} Na + b\sum_{i=1}^{3} x_i^c f(x_i^c) = \sum_{j=1}^{4} y_j f(y_j) \\ a\sum_{i=1}^{3} x_i^c f(x_i^c) + b\sum_{i=1}^{3} \left(x_i^c\right)^2 f(x_i^c) = \sum_{j=1}^{4} \sum_{i=1}^{3} x_i^c y_j f_{ij} \end{cases}$$

$$\begin{cases} 100a + 310b = 375 \\ 310a + 1390b = 1525 \end{cases}$$

$$\begin{cases} a = 1{,}13054 \\ b = 0{,}845 \end{cases}$$

x	y	f_{xy}	xyf_{xy}
1	0	14	0
1	2	20	40
1	5	6	30
1	9	0	0
3	0	6	0
3	2	9	54
3	5	7	105
3	9	8	216
6	0	0	0
6	2	6	72
6	5	12	360
6	9	12	648
		100	1525

x^c	f(x)	$x^cf(x)$	$(x^c)^2$	$(x^c)^2f(x)$
1	40	40	1	40
3	30	90	9	270
6	30	180	36	1080
	100	310		1390

y	f(y)	yf(y)	y^2	$y^2f(y)$
0	20	0	0	0
2	35	70	4	140
5	25	125	25	625
9	20	180	81	1620
	100	375	110	2385

Oppure:

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

Dove:

$$M(X \cdot Y) = \frac{1525}{100} = 15{,}25$$

$$m_x = \frac{310}{100} = 3{,}1$$

$$m_y = \frac{375}{100} = 3{,}75$$

$$\sigma_x^2 = \frac{1390}{100} - (3{,}1)^2 = 4{,}29$$

Quindi sostituendo si ottiene:

$$\begin{cases} b = \dfrac{15{,}25 - (3{,}1)\cdot(3{,}75)}{4{,}29} = \dfrac{3{,}625}{4{,}29} \\ a = 3{,}75 - b \cdot 3{,}1 \end{cases}$$

$$\begin{cases} b = 0{,}845 \\ a = 1{,}13054 \end{cases}$$

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:

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

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

Coefficiente di determinazione

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

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

ESERCIZIO (*v. sopra dati ESERCIZIO 1*)

Calcolo del *Coefficiente di determinazione*:

y	Y	$(Y-m)^2$	$(y-m)^2$
5	17,4247	8,55393	90,25
10	17,1057	6,78942	20,25
20	16,1485	2,71716	30,25
42	14,5532	0,00283	756,25
8	12,3198	4,75344	42,25
2	9,44823	25,52041	156,25
87		48,33719	1095,5

$$R^2 = \frac{Dev.regressione}{Dev.totale} = \frac{\sum_{i=1}^{n}\left(Y_i - m_y\right)^2}{\sum_{i=1}^{n}\left(y_i - m_y\right)^2} = \frac{48{,}33719}{1095{,}5} = 0{,}04412$$

Dove $m_y = \frac{87}{6} = 14{,}5$

Il valore vicino allo *zero* del *Coefficiente di determinazione* indica che il modello teorico scelto non è molto buono per interpretare la realtà osservata.

ESERCIZIO (*v. sopra dati ESERCIZIO 2*)

Calcolo del *Coefficiente di determinazione*:

x	y	Y	$(Y+1,3)^2$	$(y+1,3)^2$
3	-2	-2,026	0,527076	0,49
0	0,5	0,695	3,980025	3,24
2	-1	-1,119	0,032761	0,09
1	0	-0,212	1,183744	1,69
5	-4	-3,84	6,451600	7,29
2	-2	-1,119	0,032761	0,49
0	1	0,695	3,980025	5,29
4	-3	-2,933	2,666689	2,89
2	-1,5	-1,119	0,032761	0,04
3	-1	-2,026	0,527076	0,09
22	-13		19,414518	21,6

$$R^2 = \frac{Dev.regressione}{Dev.totale} = \frac{\sum_{i=1}^{n}\left(Y_i - m_y\right)^2}{\sum_{i=1}^{n}\left(y_i - m_y\right)^2} = \frac{19{,}414518}{21{,}6} = 0{,}8988$$

Il che esprime che il grado d'accostamento del modello teorico, la retta, ai valori osservati è molto buono.

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:

$$-1 \le r \le +1$$

- ***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***:

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

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

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***:

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

ESERCIZIO (*v. sopra dati ESERCIZIO 2*)

Calcolo del *Coefficiente di correlazione lineare*:

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

Dove

$Cov(X,Y) = -2{,}14$

$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{2{,}36} = 1{,}536$

$\sigma_y = \sqrt{\sigma_y^2} = \sqrt{3{,}85 - (-1{,}3)^2} = \sqrt{2{,}16} = 1{,}47$

Quindi sostituendo si ottiene:

$$r = \frac{(-2{,}14)}{(1{,}536) \cdot (1{,}47)} = -0{,}95$$

Tale valore esprime una forte correlazione lineare inversa fra le due variabili.

ESERCIZIO (*v. sopra dati ESERCIZIO 3*)

Calcolo del *Coefficiente di correlazione lineare*:

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

Dove:

$Cov(X,Y) = 3{,}625$

$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{4{,}29} = 2{,}07123$

$\sigma_y = \sqrt{\sigma_y^2} = \sqrt{\frac{2385}{100} - (3{,}75)^2} = \sqrt{9{,}7875} = 3{,}1285$

Quindi sostituendo si ottiene: $r = \frac{3{,}625}{2{,}07123 \cdot 3{,}1285} = +0{,}55943$