La regressione lineare semplice

Il fondamento di base
La stima dei parametri del modello
La bontà di adattamento

Analisi della dipendenza
Regressione lineare

Regressione: studio di come varia in media un carattere DIPENDENTE al variare di quello INDIPENDENTE.

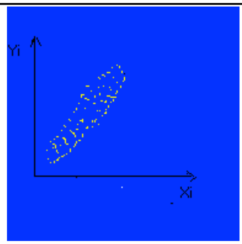
Y,X = VARIABILI NUMERICHE

Y= DIPENDENTE X= INDIPENDENTE

Fu introdotta da Francis Galton mostrando come alcuni fenomeni antropometrici si comportavano

E'importante la scelta ed il ruolo delle variabili

$(x_i, y_i) \qquad i = 1,2,\ldots\ldots,n$



Obiettivo: individuazione di una funzione

$$\hat{y} = f(x; c_0, c_1, \ldots c_n)$$

che spieghi il legame tra x ed y

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

La retta è la funzione più usata perché di più semplice interpretazione

La regressione

Obiettivo:

studiare la relazione funzionale che intercorre tra una variabile dipendente *(Y)* ed una variabile indipendente *(X)*.

Fasi:

1. Scelta del tipo di funzione
2. Determinazione dei parametri incogniti
3. Verifica della bontà di adattamento (adeguatezza del modello)

Scelta del tipo di funzione

Nel caso più semplice si ipotizza che esiste un legame di tipo lineare tra le due variabili, per cui la relazione che lega *Y* ad *X* è la seguente:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Determinazione dei parametri incogniti

Tecnica: Metodo dei minimi quadrati

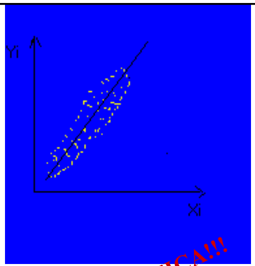
Obiettivo:

$$\min_{\beta_0,\beta_1} S = \sum_{i=1}^{n}\left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right]^2$$

Bisogna risolvere il sistema:

$$\begin{cases} \dfrac{\partial S}{\partial \hat{\beta}_0} = 0 \\ \dfrac{\partial S}{\partial \hat{\beta}_1} = 0 \end{cases}$$

$$\sum_i (\hat{y}_i - y_i)^2 = \min$$



Bisogna rendere minima la somma degli errori al quadrato

$$(\hat{y}_i - y_i)^2$$

ESISTE UNA SOLUZIONE ANALITICA!!!

Come si determinano i parametri?

$$\sum_i (\hat{y}_i - y_i)^2 = \underbrace{\sum_i \left(\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i\right)^2}_{S}$$

E' un problema di minimizzazione!

$$\min_{\hat{\beta}_0, \hat{\beta}_1} S$$

Determinazione dei parametri incogniti (Popolazione)

Le soluzioni sono date da:

$$\hat{\beta}_1 = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{Cod(XY)}{Dev(X)} = \frac{\sum x_i y_i - n\mu_x\mu_y}{\sum x_i^2 - n\mu_x^{\;2}}$$

$$\hat{\beta}_0 = \mu_y - \hat{\beta}_1\mu_x$$

Francesco Mola

Determinazione dei parametri incogniti (Campione)

Le soluzioni sono date da:

$$b_1 = \frac{Cov(X,Y)}{S_x^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\bar{x}^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Francesco Mola

Esempio1

Determinare i parametri della retta di regressione che studia la dipendenza del carattere Y dal carattere X

X	Y
46	54
65	52
30	9
85	91
99	60
56	43
57	21

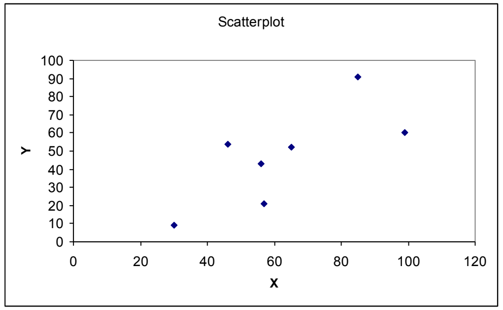
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\hat{\beta}_1 = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{Cod(XY)}{Dev(X)} = \frac{\sum x_i y_i - n\mu_x\mu_y}{\sum x_i^2 - n\mu_x^{\;2}}$$

$$\hat{\beta}_0 = \mu_y - \hat{\beta}_1\mu_x$$

Francesco Mola

Esempio1



Francesco Mola

Esempio1

	X	Y	xy	x_i^2
	46	54	2.484	2.116
	65	52	3.380	4.225
	30	9	270	900
	85	91	7.735	7.225
	99	60	5.940	9.801
	56	43	2.408	3.136
	57	21	1.197	3.249
totale	438	330	23.414	30.652

$$\mu_{x^2} = \frac{30.652}{7} = 4.378{,}86$$

$$Cod(XY) = 23.414 - 7 \cdot (62{,}57 \cdot 47{,}14) = $$
$$= 2.765{,}43$$

$$Dev(X) = 4.378{,}86 - 7 \cdot (62{,}57)^2 = 3.245{,}71$$

$$\hat{\beta}_1 = \frac{2.765{,}43}{3.245{,}71} = 0{,}852$$

$$\hat{\beta}_0 = 47{,}14 - 0{,}85 \cdot 62{,}57 = -6{,}171$$

$$\mu_x = \frac{438}{7} = 62{,}57$$

$$\mu_y = \frac{330}{7} = 47{,}14$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{y}_i = -6{,}171 + 0{,}852 x_i$$

Esempio1

X	Y	$\hat{Y}$	$Y - \hat{Y}$
46	54	33	21
65	52	49	3
30	9	19	-10
85	91	66	25
99	60	78	-18
56	43	42	1
57	21	42	-21
totale 438	330	330	0

$$\hat{y}_i = -6{,}171 + 0{,}852 x_i$$

$$33{,}021 = -6{,}171 + 0{,}852 * (46)$$

$$78{,}177 = -6{,}171 + 0{,}852 * (99)$$

$$\boxed{\sum (y_i - \hat{y}_i) = 0}$$

$$\boxed{\sum y_i = \sum \hat{y}_i}$$

Verifica della bontà di adattamento

Si consideri la devianza di *Y*

$$Dev(Y) = \sum_{i=1}^{n} (y_i - \mu_y)^2$$

Vale la seguente proprietà:

$$Dev(Y) = \sum_{i=1}^{n} (y_i - \mu_y)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \mu_y)^2 =$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \mu_y)^2 + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \mu_y) =$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \mu_y)^2 + 2\left[\left(\sum y_i - \sum \hat{y}_i\right)\left(\sum \hat{y}_i - n\mu_y\right)\right] =$$

Scomposizione della devianza

$$Dev(Y) = \sum_{i=1}^{n} (y_i - \mu_y)^2 =$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \mu_y)^2 + 2\left[\underbrace{\boxed{\left(\sum y_i - \sum \hat{y}_i\right)}}_{=0} \left(\sum \hat{y}_i - n\mu_y\right)\right]$$

Una proprietà del metodo dei minimi quadrati assicura che : $\sum y_i = \sum \hat{y}_i$

Per cui l'ultimo termine della scomposizione della devianza è nullo.

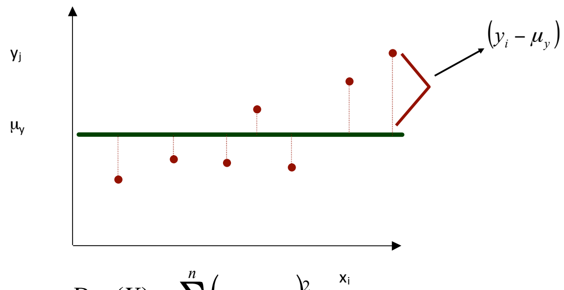
Scomposizione della devianza (cont.)

$$Dev(Y)=\sum_{i=1}^{n}(y_i-\mu_y)^2=\sum_{i=1}^{n}(\hat{y}_i-\mu_y)^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

$$Dev(Y) \qquad = Dev(R) \qquad + Dev(E)$$

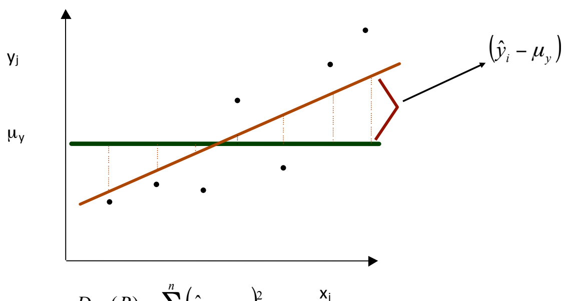
Devianza		Devianza		Devianza
Totale	=	di regressione	+	Residua
Dev(Y)		Dev(R)		Dev(E)

Devianza Totale



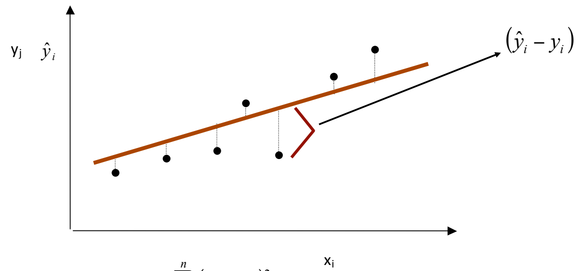
$$Dev(Y)=\sum_{i=1}^{n}(y_i-\mu_y)^2$$

Devianza di regressione



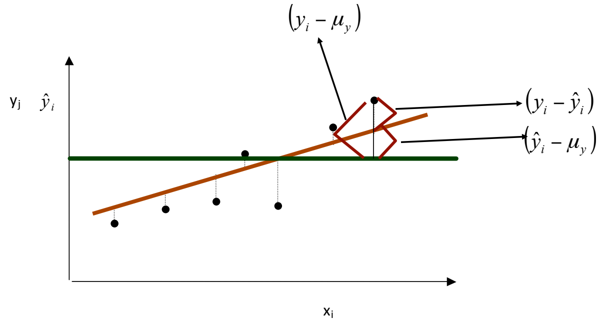
$$Dev(R)=\sum_{i=1}^{n}(\hat{y}_i-\mu_y)^2$$

Devianza Residua



$$Dev(E)=\sum_{i=1}^{n}(\hat{y}_i-y_i)^2$$

Scomposizione della Devianza … con un po' di immaginazione



Indice di determinazione lineare

Permette di misurare la bontà di adattamento:

$$R^2 = \frac{Dev(R)}{Dev(Y)} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \mu_y)^2}{\sum_{i=1}^{n} (y_i - \mu_y)^2}$$

oppure $$R^2 = 1 - \frac{Dev(E)}{Dev(Y)} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \mu_y)^2}$$

Riconsideriamo l'esempio 1 e calcoliamo la bontà di adattamento

X	Y	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$	$Y-\mu_y$	$(Y-\mu_y)^2$	$\hat{Y}-\mu_y$	$(\hat{Y}-\mu_y)^2$
46	54	33,02	20,98	440,08	6,86	47,02	-14,12	199,40
65	52	49,21	2,79	7,78	4,86	23,59	2,07	4,27
30	9	19,39	-10,39	107,94	-38,14	1.454,88	-27,75	770,25
85	91	66,25	24,75	612,52	43,86	1.923,45	19,11	365,12
99	60	78,18	-18,18	330,49	12,86	165,31	31,04	963,26
56	43	41,54	1,46	2,13	-4,14	17,16	-5,60	31,37
57	21	42,39	-21,39	457,71	-26,14	683,45	-4,75	22,55
438	330	330	0	1.958,64	0	4.314,86	0	2.356,22

$Dev(E)$ $Dev(Y)$ $Dev(R)$

$$R^2 = \frac{Dev(R)}{Dev(Y)} = \frac{2356,22}{4314,86} = 0,546 \qquad R^2 = 1 - \frac{Dev(E)}{Dev(Y)} = 1 - \frac{1958,64}{4314,86} = 0,546$$

Indice di determinazione lineare: caratteristiche

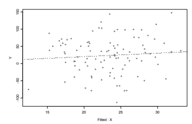
- Indica quanta parte della devianza di *Y* è spiegata dalla retta di regressione
- Solo nel caso della regressione lineare semplice vale che: l'indice di determinazione lineare è pari al quadrato del coefficiente di correlazione lineare!
- Dalla scomposizione di *Dev(Y)* si ricava che:

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

Casi:

R^2 prossimo a 0

scarso adattamento



R^2 prossimo a 1

adattamento quasi perfetto