

Corso di Statistica Facoltà di Economia

a.a. 2010-2011

francesco mola

Lezione n° 10

Contenuti

- Regressione
- Determinazione della bontà di adattamento
- Rette di regressione
- Alcune relazioni importanti
- Esempi

Lez10_a.a. 2010-2011

statistica-francesco mola

2

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

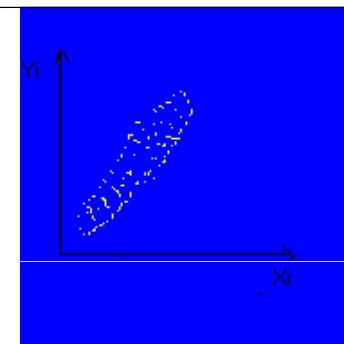
E'importante la scelta ed il ruolo delle variabili

Lez10_a.a. 2010-2011

statistica-francesco mola

3

$$(x_i, y_i) \quad i = 1, 2, \dots, n$$



Obiettivo: individuazione di una funzione

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

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

Lez10_a.a. 2010-2011

statistica-francesco mola

4

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:

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

Bisogna risolvere il sistema:

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

Determinazione dei parametri incogniti

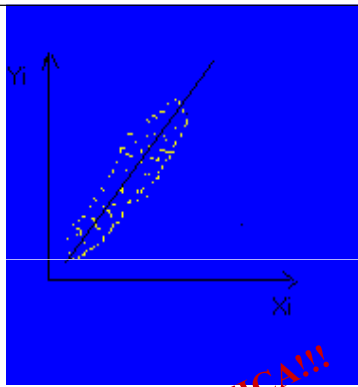
Le soluzioni sono date da:

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

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

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

Bisogna rendere minima la somma degli errori al quadrato



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

ESISTE UNA SOLUZIONE ANALITICA!!

Come si determinano i parametri?

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

S

E' un problema di minimizzazione!

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

$$\frac{\partial S}{\partial \hat{\beta}_0} = 2 \sum_i (\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i) = 0$$

$$\frac{\partial S}{\partial \hat{\beta}_1} = 2 \sum_i (\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i) x_i = 0$$

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i \\ \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i \end{cases}$$

Esprimiamo il sistema in funzione delle medie

$$\left(\mu_x = \frac{\sum x_i}{n} \Rightarrow n\mu_x = \sum x_i \right)$$

Ottenendo:

$$\begin{cases} n\hat{\beta}_0 + n\hat{\beta}_1\mu_x = n\mu_y \\ n\hat{\beta}_0\mu_x + \hat{\beta}_1\sum x_i^2 = \sum x_i y_i \end{cases}$$

Se sottraiamo alla 2^a la 1^a moltiplicato la media di x otteniamo:

$$n\cancel{\hat{\beta}_0\mu_x} - n\cancel{\hat{\beta}_0\mu_x} + \hat{\beta}_1(\sum x_i^2 - n\mu_x^2) =$$

$$= \sum x_i y_i - n\mu_x \mu_y$$

E ricaviamo i parametri incogniti:

$$\begin{cases} \hat{\beta}_0 = \mu_y - \hat{\beta}_1\mu_x \\ \hat{\beta}_1 = \frac{\sum x_i y_i - n\mu_x \mu_y}{\sum x_i^2 - n\mu_x^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \end{cases}$$

Esempio1

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

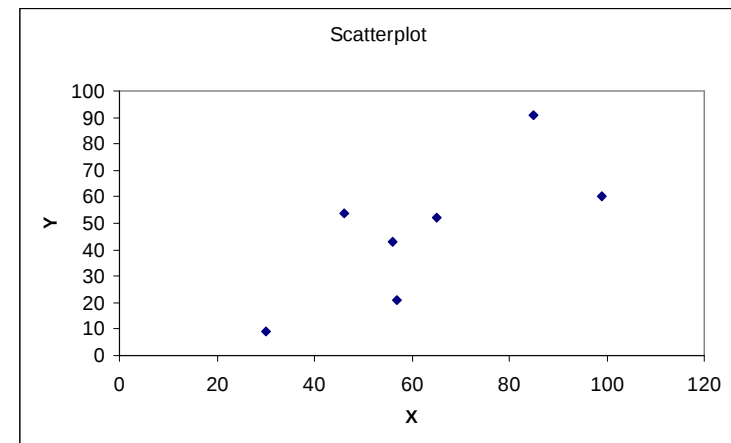
X	Y
46	54
65	52
30	9
85	91
99	60
56	43
57	21

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

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

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

Esempio1



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

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

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

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

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

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

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

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

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

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

Lez10_a.a. 2010-2011

statistica-francesco mola

17

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

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

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

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

Lez10_a.a. 2010-2011

statistica-francesco mola

18

Verifica della bontà di adattamento

Si consideri la devianza di Y

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

Vale la seguente proprietà:

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

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

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

Lez10_a.a. 2010-2011

statistica-francesco mola

19

Scomposizione della devianza

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

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

=0

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.

Lez10_a.a. 2010-2011

statistica-francesco mola

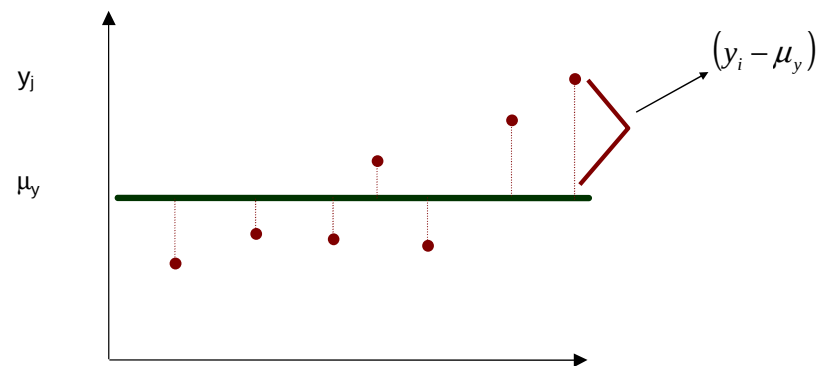
20

Scomposizione della devianza (cont.)

$$\begin{aligned} Dev(Y) &= \sum_{i=1}^n (y_i - \mu_y)^2 = \sum_{i=1}^n (\hat{y}_i - \mu_y)^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ Dev(Y) &= Dev(R) + Dev(E) \end{aligned}$$

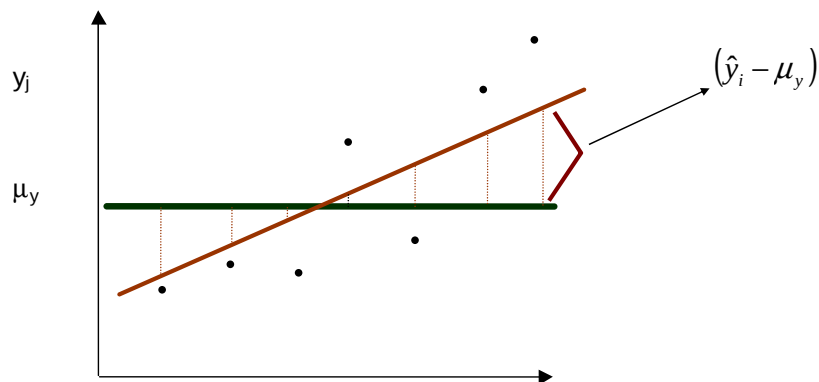
Devianza Totale	=	Devianza di regressione	+	Devianza Residua
Dev(Y)		Dev(R)		Dev(E)

Devianza Totale



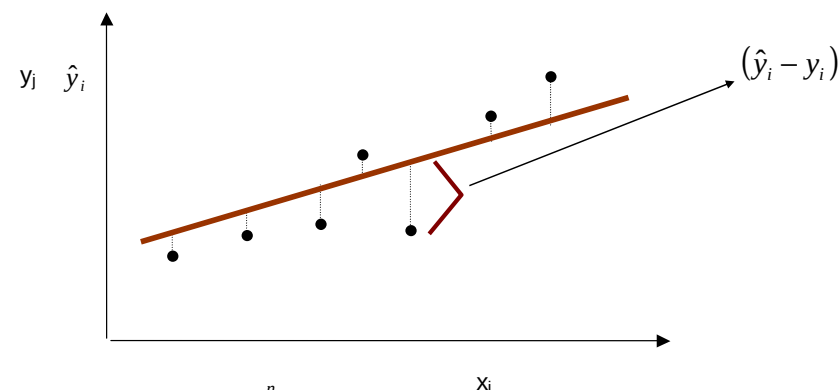
$$Dev(Y) = \sum_{i=1}^n (y_i - \mu_y)^2$$

Devianza di regressione



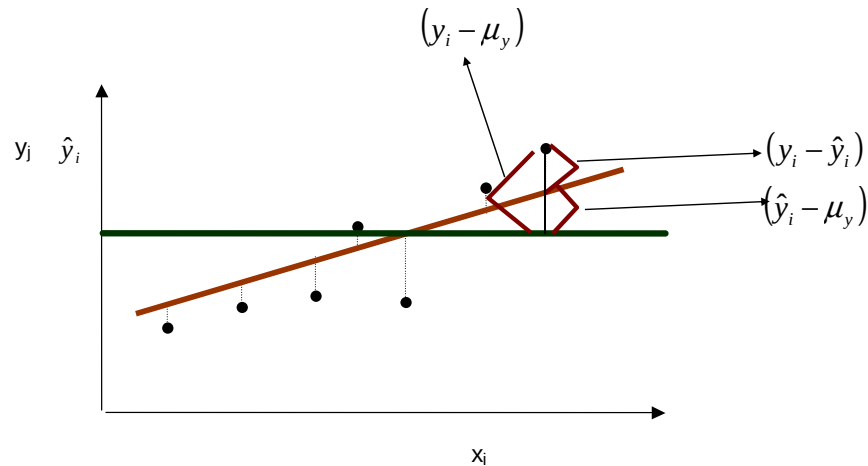
$$Dev(R) = \sum_{i=1}^n (\hat{y}_i - \mu_y)^2$$

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:

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

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

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

$Dev(E)$

$Dev(Y)$

$Dev(R)$

$$R^2 = \frac{Dev(R)}{Dev(Y)} = \frac{2356,22}{4314,86} = 0,546$$

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

Indice di determinazione lineare: caratteristiche

- Indica quanta parte della devianza di Y è spiegata dalla retta di regressione
- Dalla scomposizione di $Dev(Y)$ si ricava che:

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

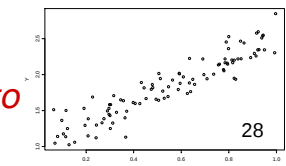
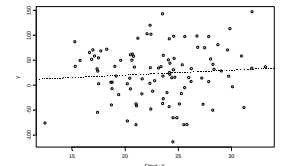
Casi:

R^2 prossimo a 0

scarso adattamento

R^2 prossimo a 1

adattamento quasi perfetto



Alcune relazioni importanti

Si dimostra che il coefficiente di correlazione lineare e l'indice di determinazione lineare sono legati dalla seguente relazione:

$$R^2 = (\rho_{xy})^2 \quad \rho_{xy} = \sqrt{R^2} \text{ a meno del segno}$$

Valgono inoltre le seguenti relazioni:

$$\rho_{xy} = \sqrt{\hat{\beta}'_1 \hat{\beta}_1}$$

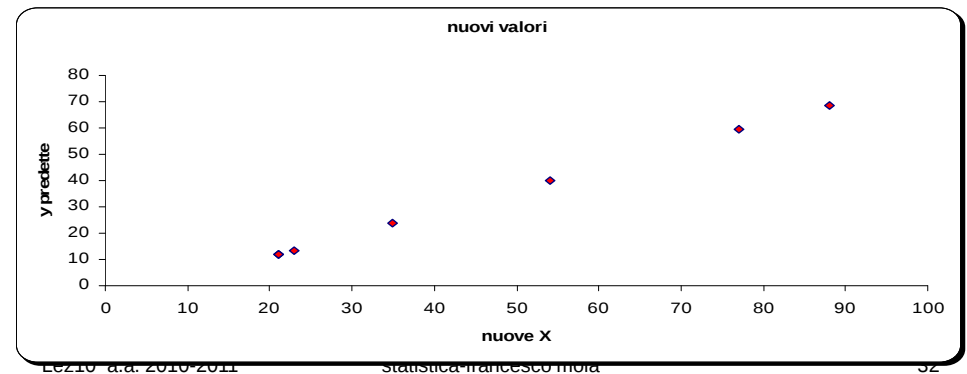
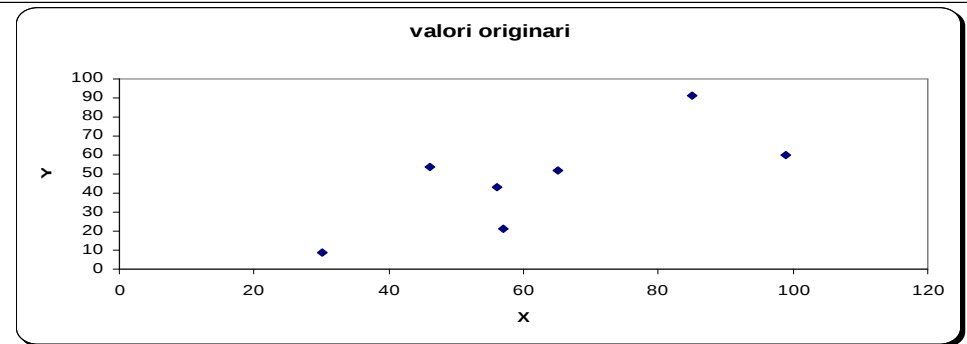
$$\hat{y}_i = \mu_y + \rho_{xy} \frac{\sigma_y}{\sigma_x} (x_i - \mu_x) \quad \hat{x}_i = \mu_x + \rho_{xy} \frac{\sigma_x}{\sigma_y} (y_i - \mu_y)$$

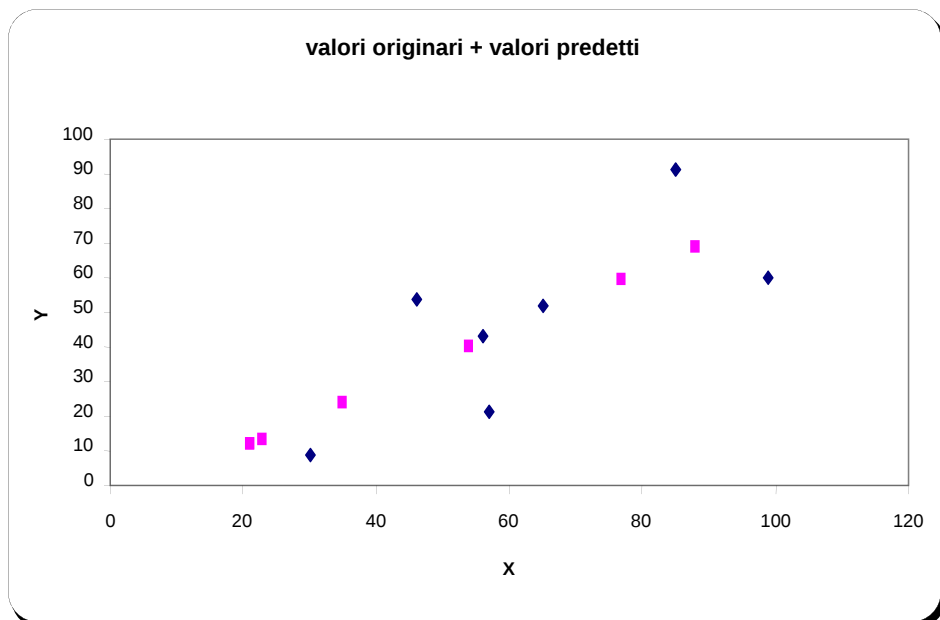
predizione

- Predizione (to predict) e previsione (to forecast): quali differenze?
- Quando la x è il tempo...
- Interpretazione del coefficiente di regressione
- Interpretazione della costante

Supponiamo di avere nuove osservazioni o di voler immaginare cosa accade se i valori della variabile dipendente mutano

xnuove	ypredette	
21	12	$12 \approx -6,171 + 0,852 * (21)$
77	59	
54	40	
35	24	
88	69	$69 \approx -6,171 + 0,852 * (88)$
21	12	
23	13	



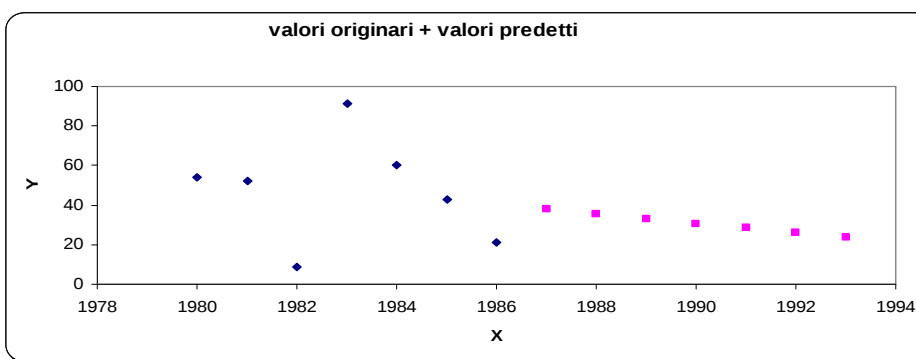
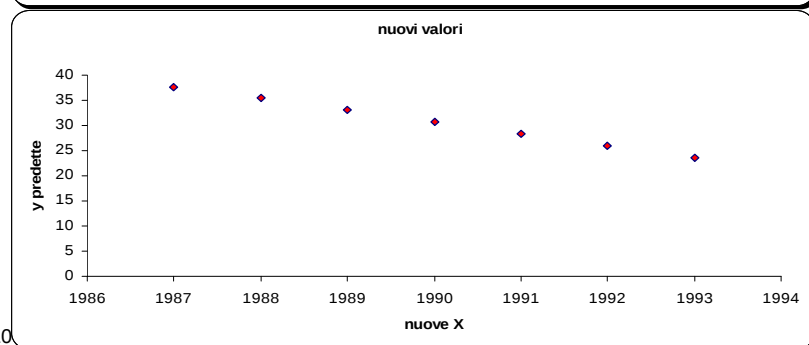
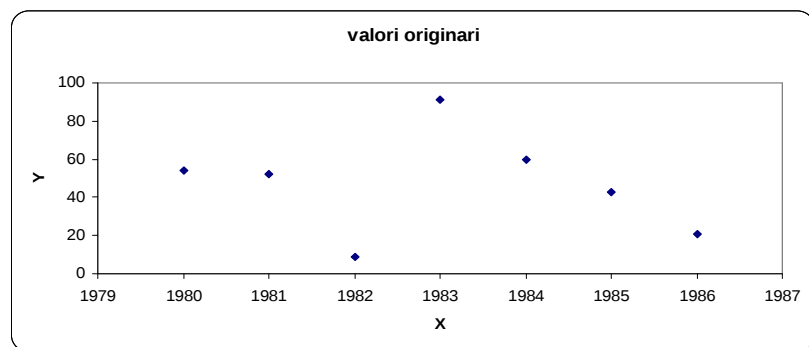


$$\hat{\beta}_1 = -2,358 \quad R^2 = \frac{Dev(R)}{Dev(Y)} = 0,036$$

$$\hat{\beta}_0 = 4721,357 \quad \rho = \sqrt{R^2} = -0,19$$

X	Y	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$	$Y - \mu_y$	$(Y - \mu_y)^2$	$\hat{Y} - \mu_y$	$(\hat{Y} - \mu_y)^2$	xnuove	ypredette
1980	54	54,21	0	0,046	7	47	7,071	50,005	1987	38
1981	52	51,86	0	0,02	5	24	4,714	22,224	1988	35
1982	9	49,5	-41	1640	-38	1.455	2,357	5,5561	1989	33
1983	91	47,14	44	1923	44	1.923	-0	7E-26	1990	31
1984	60	44,79	15	231,5	13	165	-2,36	5,5561	1991	28
1985	43	42,43	1	0,327	-4	17	-4,71	22,224	1992	26
1986	21	40,07	-19	363,7	-26	683	-7,07	50,005	1993	24
13881	330	330	0	4.159	0	4.315	0	156		

$$\hat{y}_i = 4721,357 - 2,358x_i$$



$$\hat{\beta}_1 = -2,358 \quad R^2 = \frac{Dev(R)}{Dev(Y)} = 0,036$$

$$\hat{\beta}_0 = 4721,357 \quad \rho = \sqrt{R^2} = -0,19$$

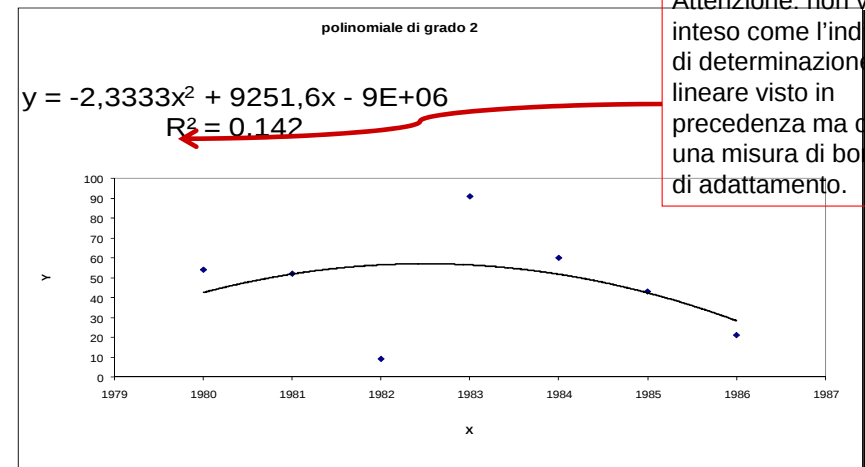
Che dire dell'intercetta?
 Che dire del valore basso dell'indice di determinazione lineare?
 Abbiamo sbagliato i calcoli o.....ad individuare il modello?

$$\hat{\beta}_1 = -2,358 \quad R^2 = \frac{Dev(R)}{Dev(Y)} = 0,036$$

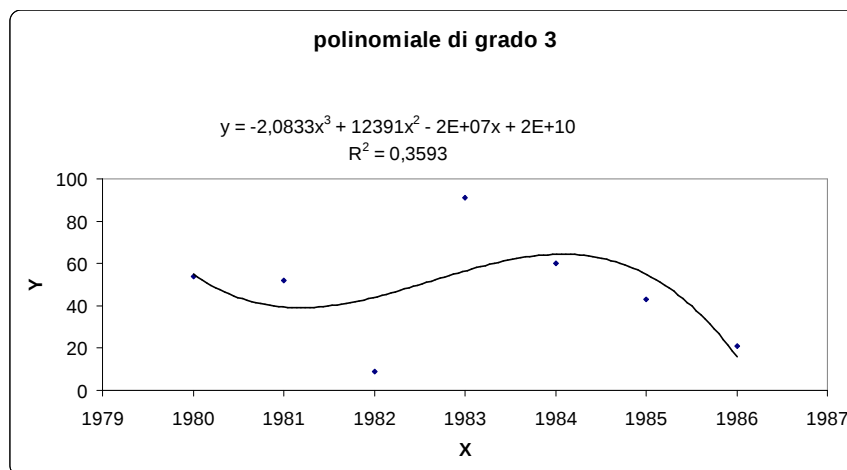
$$\hat{\beta}_0 = 4721,357 \quad \rho = \sqrt{R^2} = -0,19$$

Che dire dell'intercetta?
 Che dire del valore basso dell'indice di determinazione lineare?
 Abbiamo sbagliato i calcoli o.....ad individuare il modello?

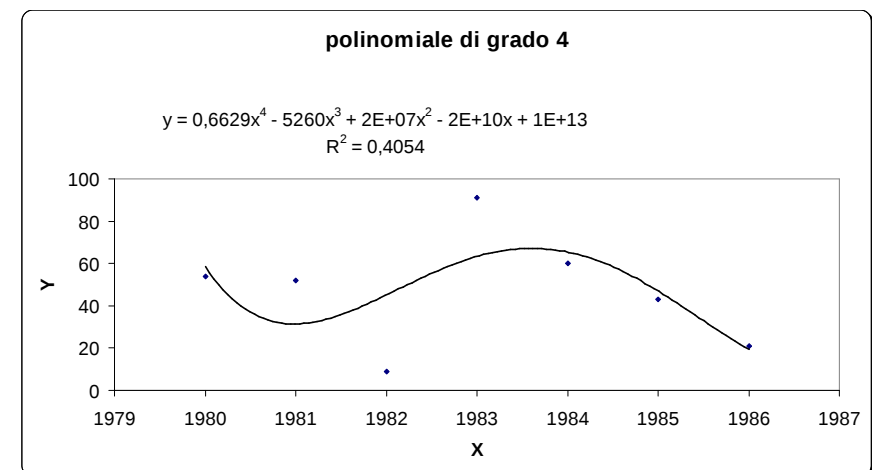
Proviamo con una polinomiale



Proviamo con una polinomiale



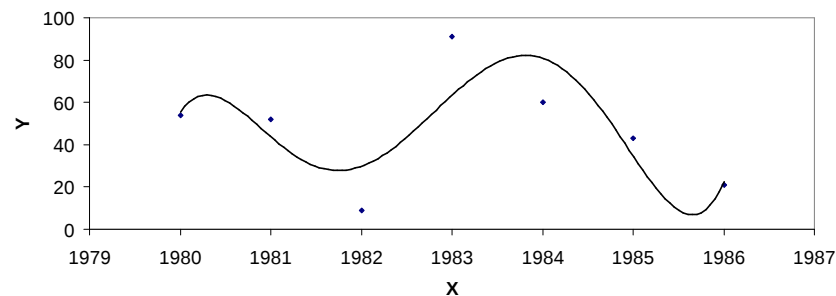
Proviamo con una polinomiale



Proviamo con una polinomiale

polinomiale di grado 5

$$y = 1,075x^5 - 10658x^4 + 4E+07x^3 - 8E+10x^2 + 8E+13x - 3E+16$$
$$R^2 = 0,5891$$



Proviamo con una polinomiale

polinomiale di grado 6

$$y = -1,7778x^6 + 21153x^5 - 1E+08x^4 + 3E+11x^3 - 4E+14x^2 + 3E+17x - 1E+20$$
$$R^2 = 1$$

