Università degli Studi di Napoli "Parthenope"

Corso di Laurea in Statistica e Informatica per l'Azienda, la Finanza e le Assicurazioni (SIAFA)

STATISTICA
II MODULO

Sergio LONGOBARDI
longobardi@uniparthenope.it

Funzione di regressione

Attraverso la funzione di regressione otteniamo una ***relazione matematica*** che lega il comportamento di due caratteri quantitativi,

Tale relazione permette di analizzare e misurare l'influenza di una variabile (***indipendente*** o condizionante) su un'altra variabile (***dipendente*** o condizionata),

Mediante la regressione è possibile stimare:

- la ***relazione funzionale*** che lega il carattere dipendente a quello indipendente
- il valore della variabile dipendente sulla base del valore della variabile indipendente

Funzione di regressione

Una funzione di regressione fornisce una relazione matematica intercorrente tra due variabili, come ad esempio “Condizioni meteorologiche” e “Comportamento del sig, Rossi”,

Condizioni meteorologiche **Var, INDIPENDENTE (X)**		**Comportamento del sig, Rossi** **Var, DIPENDENTE (Y)**
Pioggia	→	Porta con sé l’ombrello grande
Nuvoloso	→	Porta con sé l’ombrello da borsa
Sereno	→	Indossa soltanto la giacca

Il Comportamento del sig, Rossi è *funzione* delle Condizioni meteorologiche e si può indicare nel modo seguente:

Comportamento del sig, Rossi = f (Condizioni meteorologiche)

In generale, date due variabili Y e X, si dice che Y = f(X) se l’andamento della variabile X influenza quello della variabile Y,

Funzione di regressione

La relazione matematica intercorrente tra due variabili che si stima mediante la regressione lineare semplice è una **RELAZIONE FUNZIONALE LINEARE** del tipo:

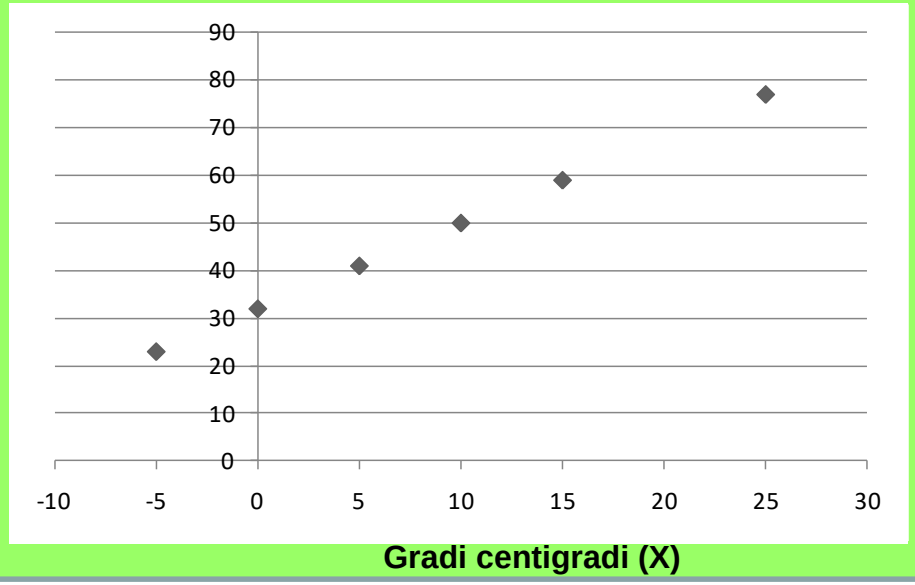
$$Y = \beta_0 + \beta_1 X$$

ESEMPIO

Conversione della temperatura da gradi centigradi (X) in Fahrenheit (Y):

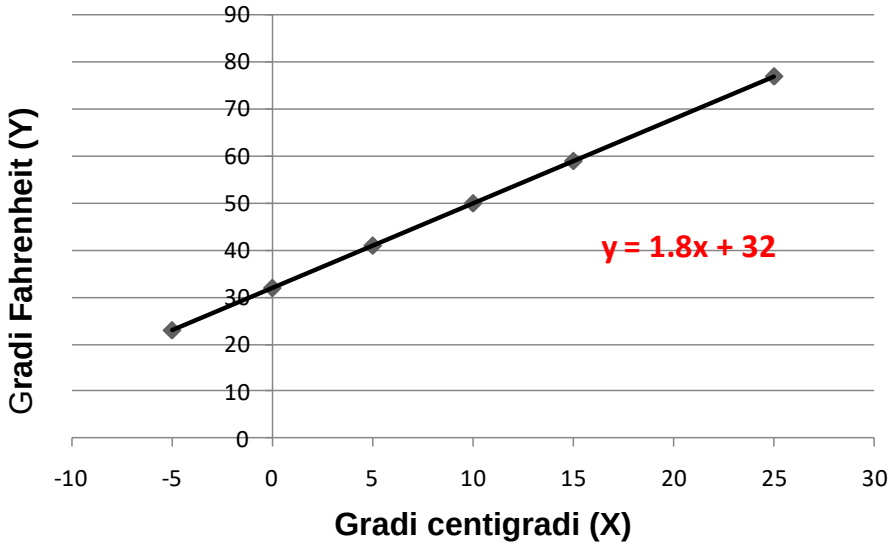
Y=32+1,8 X

Gradi Fahrenheit Y	*Gradi centigradi X*
23	-5
32	0
41	5
50	10
59	15
77	25



Funzione di regressione

INTERPRETAZIONE GEOMETRICA



Y=32+1,8 X

Intercetta

Rappresenta il punto in cui la retta interseca l'asse delle ordinate (Y) ed è il valore di Y in corrispondenza di X=0

Coefficiente angolare

Misura l'inclinazione della retta e rappresenta la variazione di Y in corrispondenza di una variazione unitaria di X

Funzione di regressione

Per descrivere e analizzare i fenomeni empirici si ricorre a delle relazioni più complesse di quelle funzionali definite

RELAZIONI STATISTICHE

Una relazione statistica tra una variabile dipendente (Y) ed una indipendente (X) si rappresenta mediante la seguente equazione:

$$Y = f(X) + \varepsilon$$

Componente deterministica

Componente stocastica

Considerando un relazione statistica di tipo lineare:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Funzione di regressione

Perché si considera anche una componente stocastica ε nel descrivere la relazione di dipendenza tra Y e X?

…perché empiricamente la relazione tra due caratteri non è mai perfetta ma è influenzata da una molteplicità di fattori che non sempre si riescono a considerare,

…,,di conseguenza la componente stocastica riassume l'influenza su Y di tutti i fattori diversi da X che non sono stati osservati o non sono osservabili

Esempio

Se si analizza la relazione tra reddito e consumi alimentari si osserverà una elevata dipendenza dei consumi dal reddito ma questa relazione non è perfetta in quanto i consumi sono influenzati non solo dal reddito ma anche da numerose altre variabili come lo stile di vita, il livello sociale, il livello culturale, etc,…,)

Funzione di regressione

Retta di regressione

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Parametri della retta di regressione

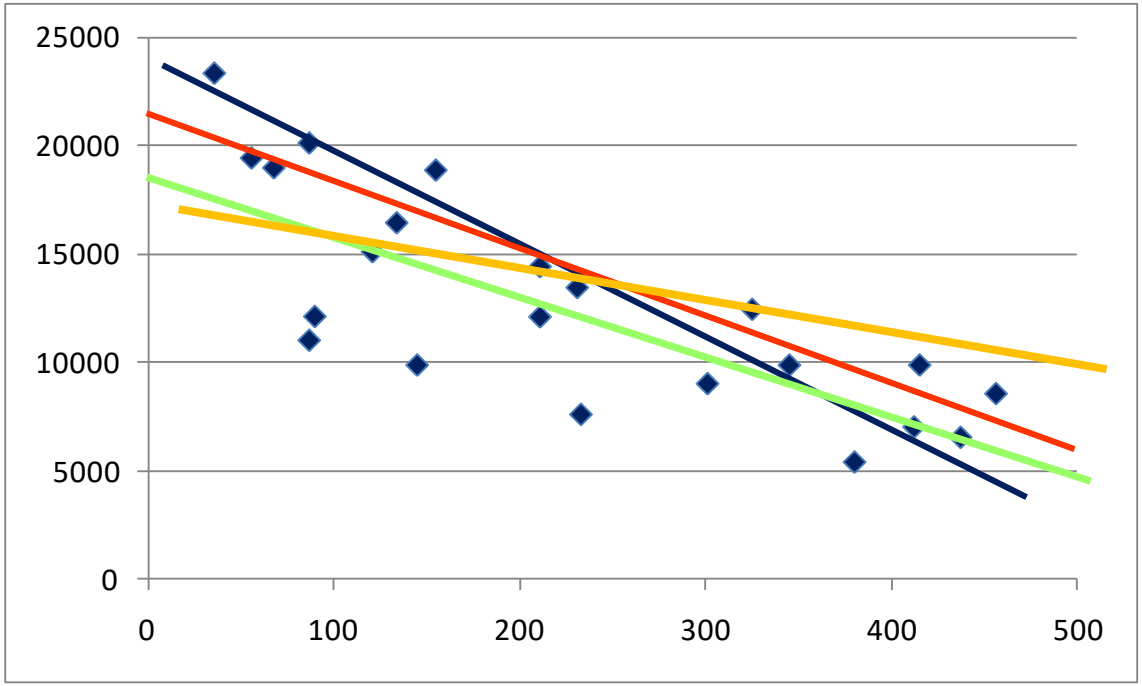
Errore

L'analisi della regressione permette di esprimere la relazione statistica tra Y e X stimando i parametri della retta di regressione (intercetta e coefficiente angolare) e la grandezza delle errore ε sulla base delle osservazioni campionarie,

Funzione di regressione

Anno	Precipitaz, (mm) X_i	Arrivi turistici Y_i
1978	415	9880
1979	211	12102
1980	121	15110
1981	36	23345
1982	134	16443
1983	456	8560
1984	87	20123
1985	345	9876
1986	211	14423
1987	325	12450
1988	56	19432
1989	68	18980
1990	155	18872
1991	231	13456
1992	301	9021
1993	412	7023
1994	437	6543
1995	90	12123
1996	380	5402
1997	145	9880
1998	87	11023
1999	233	7603

La regressione permette di stimare la retta che esprime la relazione tra le due variabili e che quindi si adatta nel modo migliore ai dati osservati



Come si stabilisce la retta che si adatta meglio ai dati?

Metodo dei Minimi quadrati

- La regressione lineare semplice consiste nel calcolare la retta che interpola meglio la nuvola di punti rappresentata nel diagramma di dispersione,

- Individuare una retta significa trovare la coppia di parametri β_0 e β_1 che la caratterizzano,

- La scelta di questi parametri non è casuale, ma al contrario sarà finalizzata a individuare, tra le infinite rette possibili, quella che può meglio rappresentare i punti del diagramma di dispersione,

Funzione di regressione

Si stimano i parametri della retta che interseca nel modo migliore i dati osservati

Il criterio è quello di considerare tra tutte le possibili rette quella che minimizza la differenza tra i valori stimati (punti della retta) e quelli osservati (punti empirici),

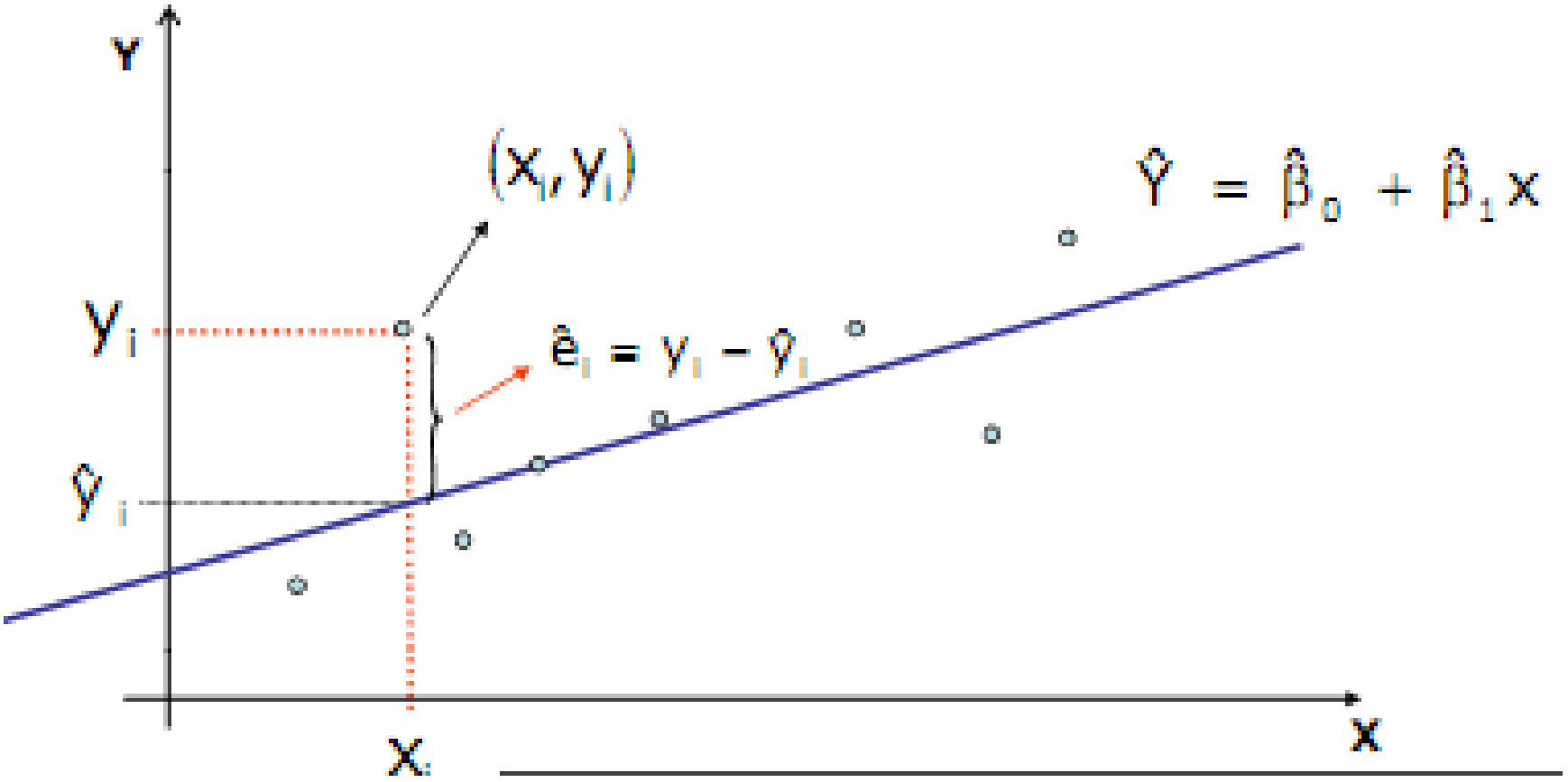
Si definisce quindi con il termine **RESIDUO** ***(e_i)*** la differenza tra il valore osservato ed il corrispondente valore fornito dalla retta di regressione (valore stimato o valore teorico)

$$\hat{e}_i = \mathrm{y}_i - \hat{\mathrm{y}}_i$$

Poiché i valori stimati sono a pari a: $\hat{\mathrm{y}}_i = \beta_o + \beta_1 x_i$

$$\hat{e}_i = y_i - (\beta_o + \beta_1 x_i)$$

Metodo dei Minimi quadrati



Metodo dei Minimi quadrati

Il criterio per la stima dei parametri è quello di individuare la retta (e quindi i parametri) che minimizza i residui al quadrato,

$$G(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \hat{e}_i^{\,2} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 =$$

$$= \sum_{i=1}^{n} \left[\; Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \;\right]^2 = \min$$

Metodo dei Minimi quadrati

$$\sum_{i=1}^{n}\left[\, Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \,\right]^2 = \min$$

Per trovare il valore dei parametri che minimizzano la funzione occorre porre le derivate parziali uguali a zero,

$$\frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_0} \Rightarrow \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1} \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Metodo dei Minimi quadrati

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Il coefficiente β_0 rappresenta il valore atteso di Y quando X è uguale a 0,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{Cod(X,Y)}{Dev(X)}$$

Il coefficiente β_1 rappresenta il cambiamento atteso in Y associato ad una variazione unitaria di X,

ESEMPIO

$$\bar{x} = \frac{\sum xi}{n} = 13{,}3; \bar{y} = \frac{\sum yi}{n} = 34{,}3;$$

	Anni studio (X_i)	Redd annuo (Y_i)
	5	22
	8	25
	11	28
	12	30
	13	32
	15	35
	16	38
	17	40
	17	45
	19	48
TOT.	133	343

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{Cod(X,Y)}{Dev(X)}$$

ESEMPIO

(REGRESSIONE)

	Anni studio (X_i)	Redd annuo (Y_i)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$	$(x_i - \bar{x})^2$
	5	22	-8,3	-12,3	102,09	68,89
	8	25	-5,3	-9,3	49,29	28,09
	11	28	-2,3	-6,3	14,49	5,29
	12	30	-1,3	-4,3	5,59	1,69
	13	32	-0,3	-2,3	0,69	0,09
	15	35	1,7	0,7	1,19	2,89
	16	38	2,7	3,7	9,99	7,29
	17	40	3,7	5,7	21,09	13,69
	17	45	3,7	10,7	39,59	13,69
	19	48	5,7	13,7	78,09	32,49
TOT.	133	343			322,1	174,1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{322}{174,2} = 1,85$$

ESEMPIO

(REGRESSIONE)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{322}{174{,}2} = 1{,}85$$

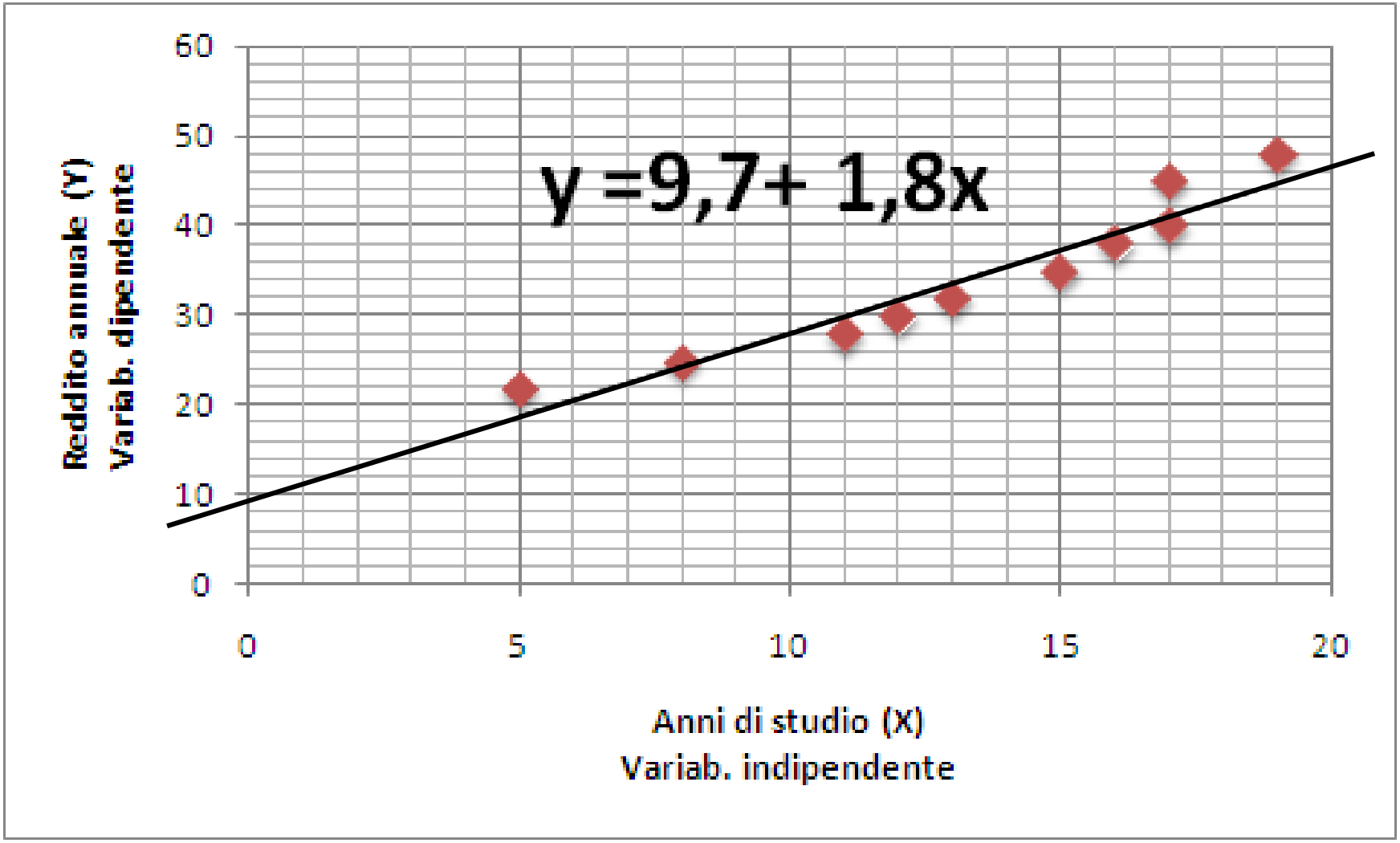
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 34{,}3 - (1{,}85 \times 13{,}3) = 34{,}3 - 24{,}6 = 9{,}7$$

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

$$y = 9{,}7 + 1{,}8x$$

ESEMPIO

(REGRESSIONE)



Utilizzo a fini previsivi del modello

Avendo stimato la seguente relazione:

$$y = 9{,}7 + 1{,}85x$$

Sarà possibile prevedere i valori della y in base ai valori della variabile indipendente:

Quanto sarà il reddito stimato nel caso in cui gli anni di studio sono pari a 11?

$$\hat{y} = 9{,}7 + 1{,}8 \cdot 11 = 29{,}5$$

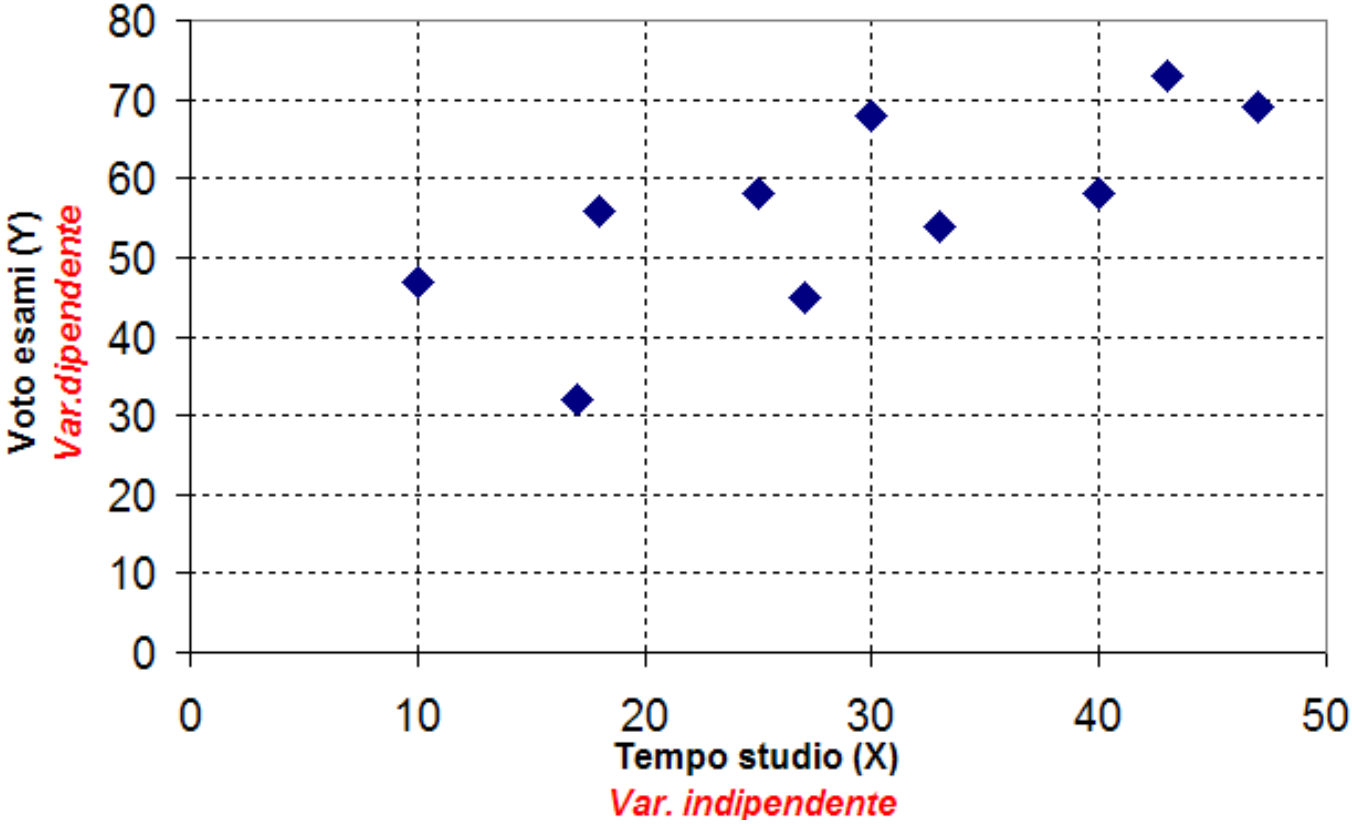
E se gli anni di studio sono 9?

$$\hat{y} = 9{,}7 + 1{,}8 \cdot 9 = 25{,}9$$

ESEMPIO 2

(REGRESSIONE)

TEMPO STUDIO (X)	VOTO ESAMI (Y)
40	58
43	73
18	56
10	47
25	58
33	54
27	45
17	32
30	68
47	69



ESEMPIO 2

$$\bar{x} = \frac{\sum xi}{n} = 29; \bar{y} = \frac{\sum yi}{n} = 56;$$

	TEMPO STUDIO (X)	VOTO ESAMI (Y)	$x_i - \bar{x}$	$y_i - \bar{y}$
	40	58	11	2
	43	73	14	17
	18	56	-11	0
	10	47	-19	-9
	25	58	-4	2
	33	54	4	-2
	27	45	-2	-11
	17	32	-12	-24
	30	68	1	12
	47	69	18	13
TOTALE	290	560		

ESEMPIO 2

(REGRESSIONE)

	TEMPO STUDIO (X)	VOTO ESAMI (Y)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	40	58	11	2	121	22
	43	73	14	17	196	238
	18	56	-11	0	121	0
	10	47	-19	-9	361	171
	25	58	-4	2	16	-8
	33	54	4	-2	16	-8
	27	45	-2	-11	4	22
	17	32	-12	-24	144	288
	30	68	1	12	1	12
	47	69	18	13	324	234
TOTALE	290	560			1304	971

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{971}{1304} = 0{,}74$$

ESEMPIO 2

(REGRESSIONE)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{971}{1304} = 0{,}74$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 56 - (0{,}74 \times 29) = 56 - 21{,}4 = 34{,}6$$

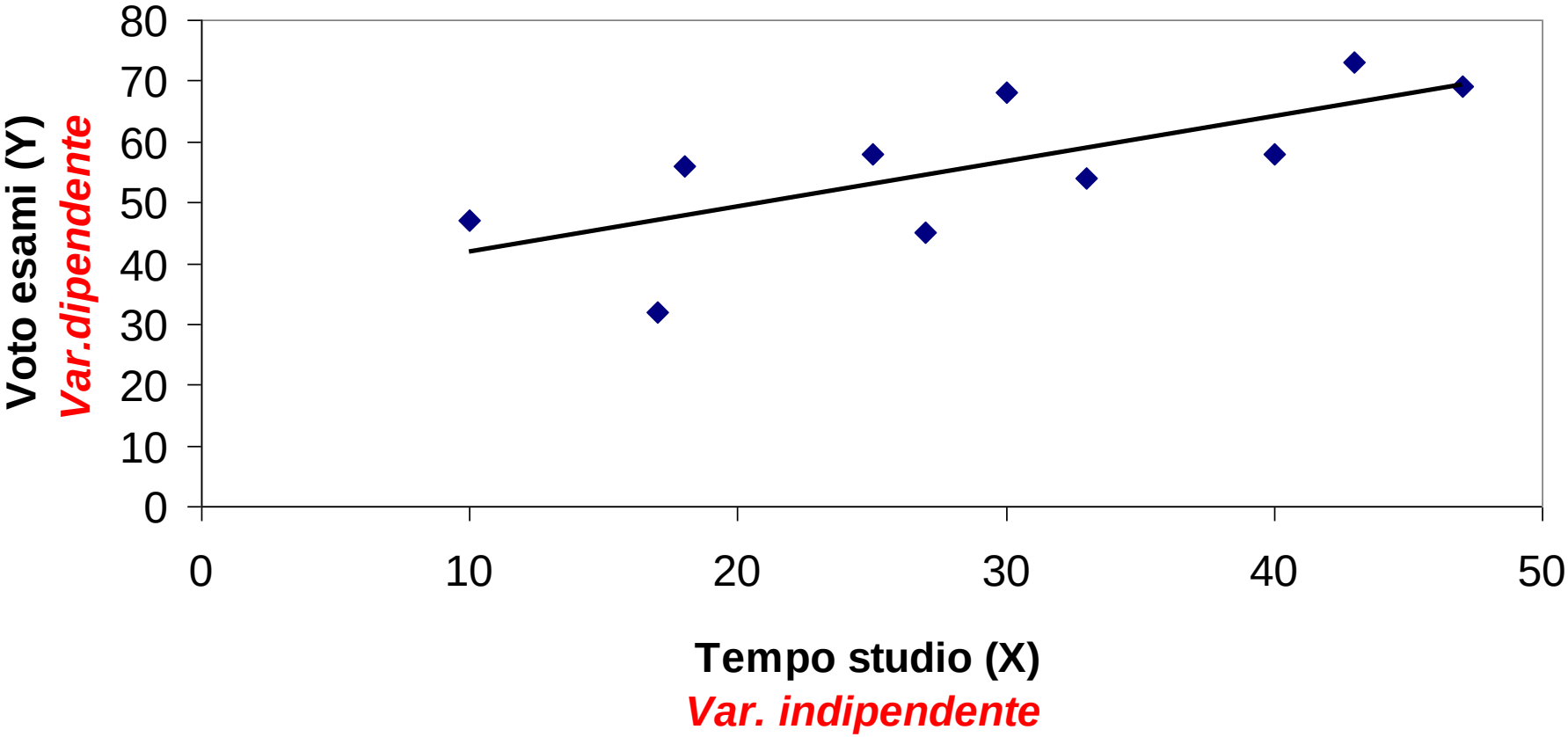
$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$y = 34{,}6 + 0{,}74x$$

ESEMPIO 2

(REGRESSIONE)

y=34,6+0,74x



Utilizzo a fini previsivi del modello

Avendo stimato la seguente relazione:

$$y = 34{,}6 + 0{,}74x$$

Sarà possibile prevedere i valori della y in base ai valori della variabile indipendente:

Quanto sarà il voto all'esame nel caso in cui le ore di studio sono pari a 37?

$$\hat{y} = 34{,}6 + 0{,}74 \cdot 37 = 61{,}98$$

E se le ore sono 29?

$$\hat{y} = 34{,}6 + 0{,}74 \cdot 29 = 48{,}66$$

Regressione lineare semplice: la valutazione del modello

La valutazione (bontà) del modello di regressione viene effettuata analizzando l' adeguatezza del modello nel descrivere la realtà,

A tale scopo si considera come misura dell'adattamento la quota di variabilità della Y "spiegata" dal modello

Regressione lineare semplice: bontà di adattamento

La retta di regressione rappresenta una interpolazione lineare del diagramma di dispersione e, pertanto, in essa è implicito un grado di approssimazione,

Un criterio si basa sulla scomposizione della varianza o della devianza se si considera solo il numeratore della varianza della variabile dipendente:

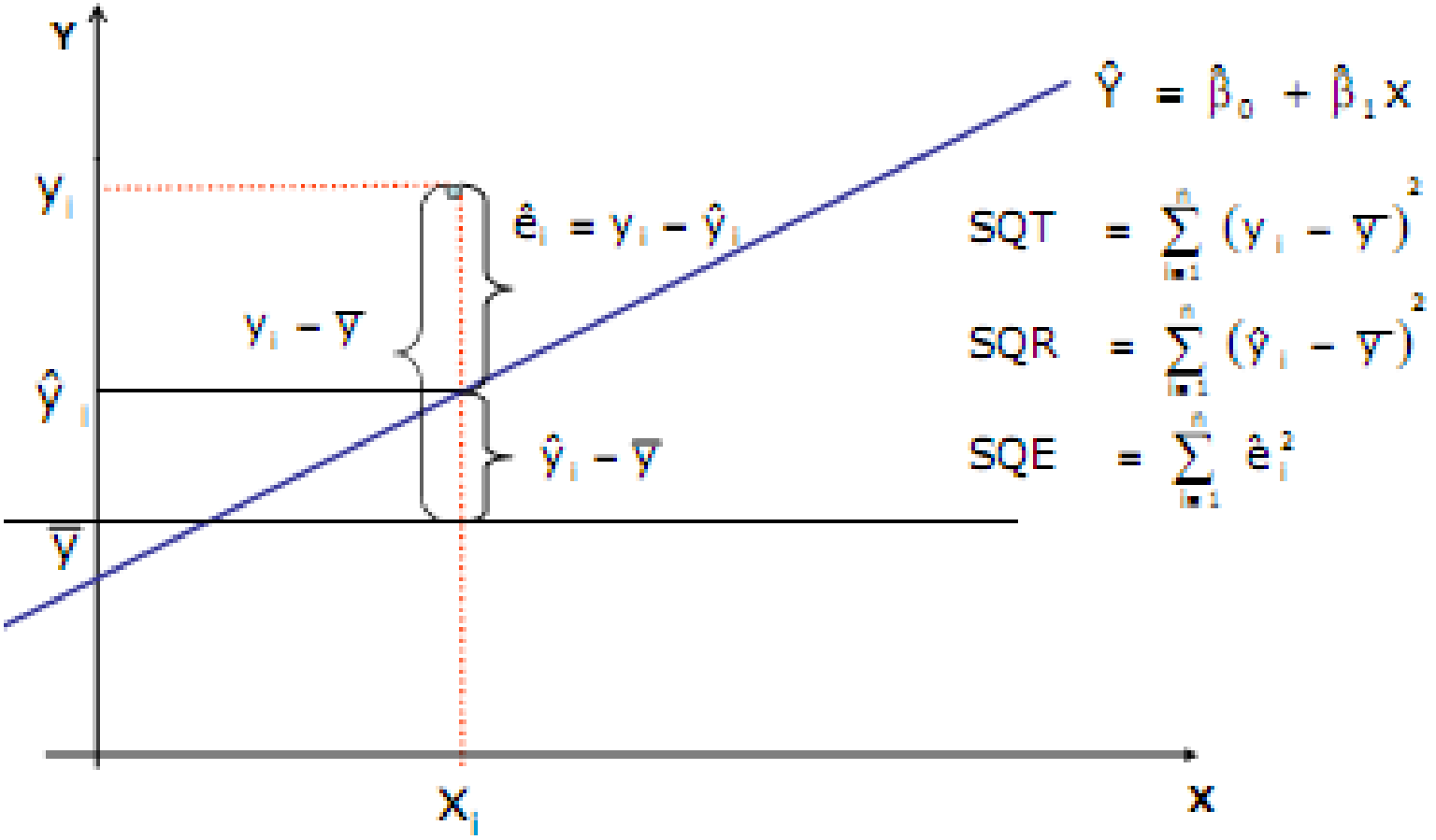
Devianza Totale

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

Devianza di regressione (somma dei quadrati spiegati dalla regressione)

Devianza dell'errore somma dei quadrati degli errori

RAPPRESENTAZIONE GRAFICA



Regressione lineare semplice: bontà di adattamento

Un modello di regressione si adatterà in modo adeguato ai dati quando la variabilità (misurata sulla base della devianza) della variabile dipendente è "spiegata" dal modello, cioè quando la quota di devianza di regressione sulla devianza totale tende all'unità

$$\frac{SQR}{SQT} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1$$

Regressione lineare semplice: bontà di adattamento

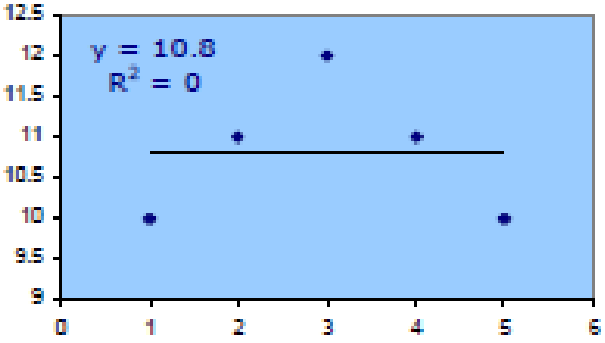
Sulla base della precedente relazione, si ottiene, quale principale indice di bontà della retta di regressione, l'indice di determinazione, definito da:

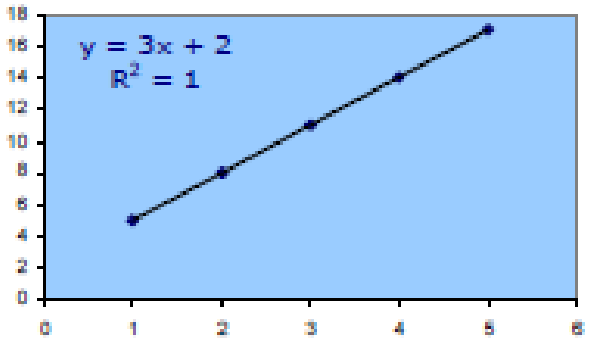
$$R^2 = \frac{SQR}{SQT} = \frac{SQT - SQE}{SQT} = 1 - \frac{SQE}{SQT} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

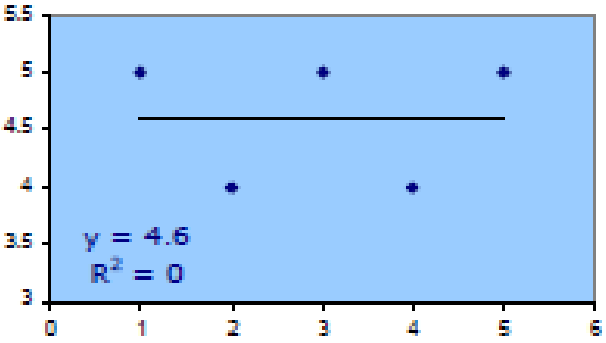
- $R^2=0$ in assenza di dipendenza lineare di Y da X (i punti osservati si dispongono casualmente sul piano (x,y) oppure evidenziano un legame non lineare [SQR=0]

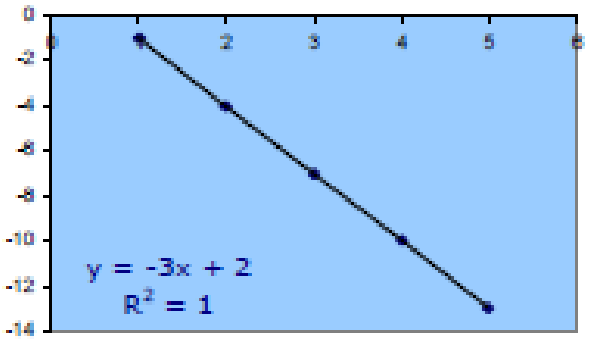
- $R^2=1$ nella situazione di perfetta dipendenza lineare (i punti osservati sono allineati su una retta) [SQE=0]

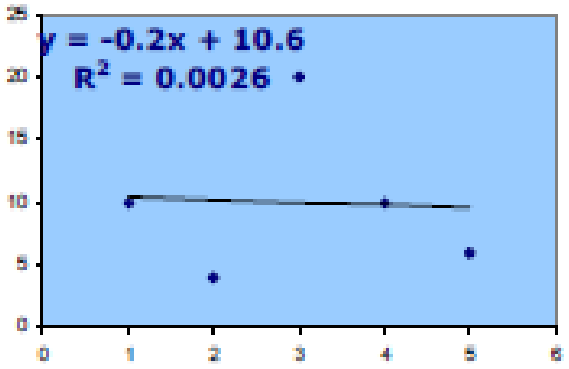
ESEMPI

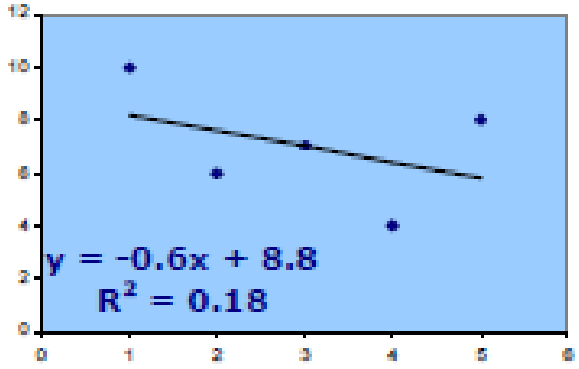

y = 10.8
R² = 0


y = 3x + 2
R² = 1


y = 4.6
R² = 0


y = -3x + 2
R² = 1


y = -0.2x + 10.6
R² = 0.0026


y = -0.6x + 8.8
R² = 0.18

Esempio di R^2

Nell'esempio precedente tra anni di studio e reddito era stata individuata la seguente relazione.

$$y = 9{,}7 + 1{,}85x$$

Anni studio (X_i)	Redd. annuo (Yi)
5,00	22,00
8,00	25,00
11,00	28,00
12,00	30,00
13,00	32,00
15,00	35,00
16,00	38,00
17,00	40,00
17,00	45,00
19,00	48,00

(SQT) Devianza totale

(SQE) Devianza residua

(SQR) Devianza di regressione

$$R^2 R^2 = \frac{SQR}{SQT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{595{,}86}{650{,}10} = 0{,}916\ 0{,}916$$

Esempio di R^2

Nell'esempio precedente tra anni di studio e reddito era stata individuata la seguente relazione.

$$y = 9{,}7 + 1{,}85x$$

Anni studio (X_i)	Redd. annuo (Yi)	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$\hat{y}_i$
5,00	22,00	-12,30	151,29	18,94
8,00	25,00	-9,30	86,49	24,49
11,00	28,00	-6,30	39,69	30,04
12,00	30,00	-4,30	18,49	31,89
13,00	32,00	-2,30	5,29	33,74
15,00	35,00	0,70	0,49	37,44
16,00	38,00	3,70	13,69	39,29
17,00	40,00	5,70	32,49	41,14
17,00	45,00	10,70	114,49	41,14
19,00	48,00	13,70	187,69	44,84
			650,10	

(SQT) Devianza totale

(SQE) Devianza residua

(SQR) Devianza di regressione

$$R^2 R^2 = \frac{SQR}{SQT} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{595{,}86}{650{,}10} = 0{,}916\ 0{,}916$$

Esempio di R^2

Nell'esempio precedente tra anni di studio e reddito era stata individuata la seguente relazione.

$$y = 9{,}7 + 1{,}85x$$

Anni studio (X_i)	Redd. annuo (Yi)	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$\hat{y}_i$	$y_i - \hat{y}$	$(y_i - \hat{y})^2$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$
5,00	22,00	-12,30	151,29	18,94	-3,06	9,35	-15,36	235,84
8,00	25,00	-9,30	86,49	24,49	-0,51	0,26	-9,81	96,18
11,00	28,00	-6,30	39,69	30,04	2,04	4,17	-4,26	18,12
12,00	30,00	-4,30	18,49	31,89	1,89	3,58	-2,41	5,79
13,00	32,00	-2,30	5,29	33,74	1,74	3,04	-0,56	0,31
15,00	35,00	0,70	0,49	37,44	2,44	5,97	3,14	9,88
16,00	38,00	3,70	13,69	39,29	1,29	1,67	4,99	24,93
17,00	40,00	5,70	32,49	41,14	1,14	1,31	6,84	46,83
17,00	45,00	10,70	114,49	41,14	-3,86	14,88	6,84	46,83
19,00	48,00	13,70	187,69	44,84	-3,16	9,97	10,54	111,15
			650,10			**54,19**		**595,86**

(SQT) Devianza totale

(SQE) Devianza residua

(SQR) Devianza di regressione

$$R^2 \; R^2 = \frac{SQR}{SQT} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{595{,}86}{650{,}10} = 0{,}916 \; 0{,}916$$