

1

## La Regressione Lineare (Semplice)

- Relazione funzionale e statistica tra due variabili
- Modello di regressione lineare semplice
- Stima puntuale dei coefficienti di regressione
- Decomposizione della varianza
- Coefficiente di determinazione
- Proprietà degli stimatori dei coefficienti
- Proprietà dello stimatore della risposta media
- Errori standard

***Contenuti del Capitolo 16 del libro di testo***

## 2 **Introduzione**

Dall'analisi ed inferenza riguardante una singola variabile statistica passiamo alla **relazione tra (due) variabili statistiche.**

Le relazioni tra variabili importanti nell'analisi della realtà economico-aziendale possono essere matematicamente espresse come:

$$Y=f(X)$$

dove la funzione **f** può assumere varie forme, lineari o non lineari, e può non essere conosciuta in modo preciso.

Consideriamo il caso più semplice quello lineare:



**regressione lineare semplice**

### 3 *Esempi*

- Il presidente di una ditta di materiali da costruzione ritiene che la Quantità media annua di piastrelle,  $Q$ , venduta sia una funzione (lineare) del Valore complessivo dei permessi edilizi rilasciati,  $V$ , nell'anno passato:  $Q=f(V)$ .
- Un grossista di cereali vuole conoscere l'effetto della produzione annua Complessiva,  $C$ , sul prezzo di vendita a tonnellata,  $P$ :  $Q=f(P)$ .
- L'area marketing di un'azienda ha necessità di sapere come il prezzo della Benzina influenzi la quantità venduta: ricorrendo alla serie storica dei prezzi settimanali e dei dati di vendita intendono sviluppare un modello (lineare) che indichi di quanto variano le vendite al variare del prezzo:  $Q=f(P)$ .

4

## Relazione funzionale e statistica

### Obiettivo:

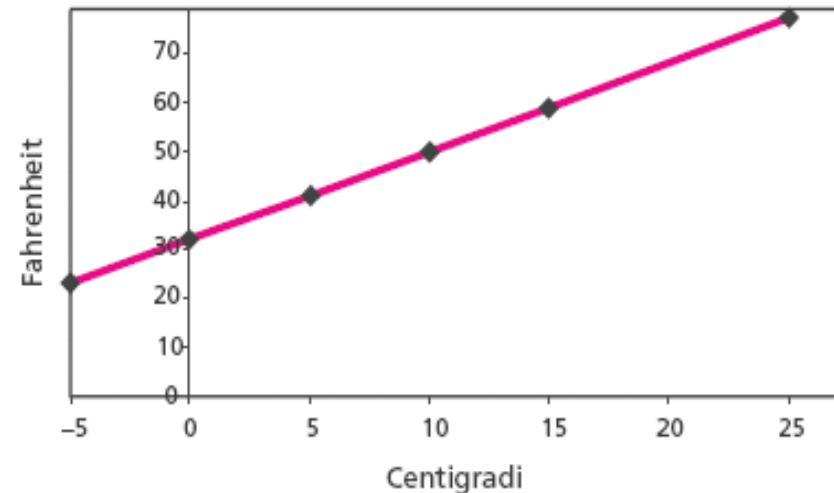
Date due variabili,  $X$  e  $Y$ , si è interessati a comprendere come la variabile  $Y$  (**dipendente** o **risposta**) sia influenzata dalla  $X$  (**esplicativa** o **indipendente**).

$Y$  è funzione di  $X$  se ad ogni valore di  $X$  corrisponde un solo valore di  $Y$ . La **relazione funzionale** è **lineare**, se possiamo scrivere:

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

$\beta_0$  = **intercetta**

$\beta_1$  = **coefficiente angolare**



5

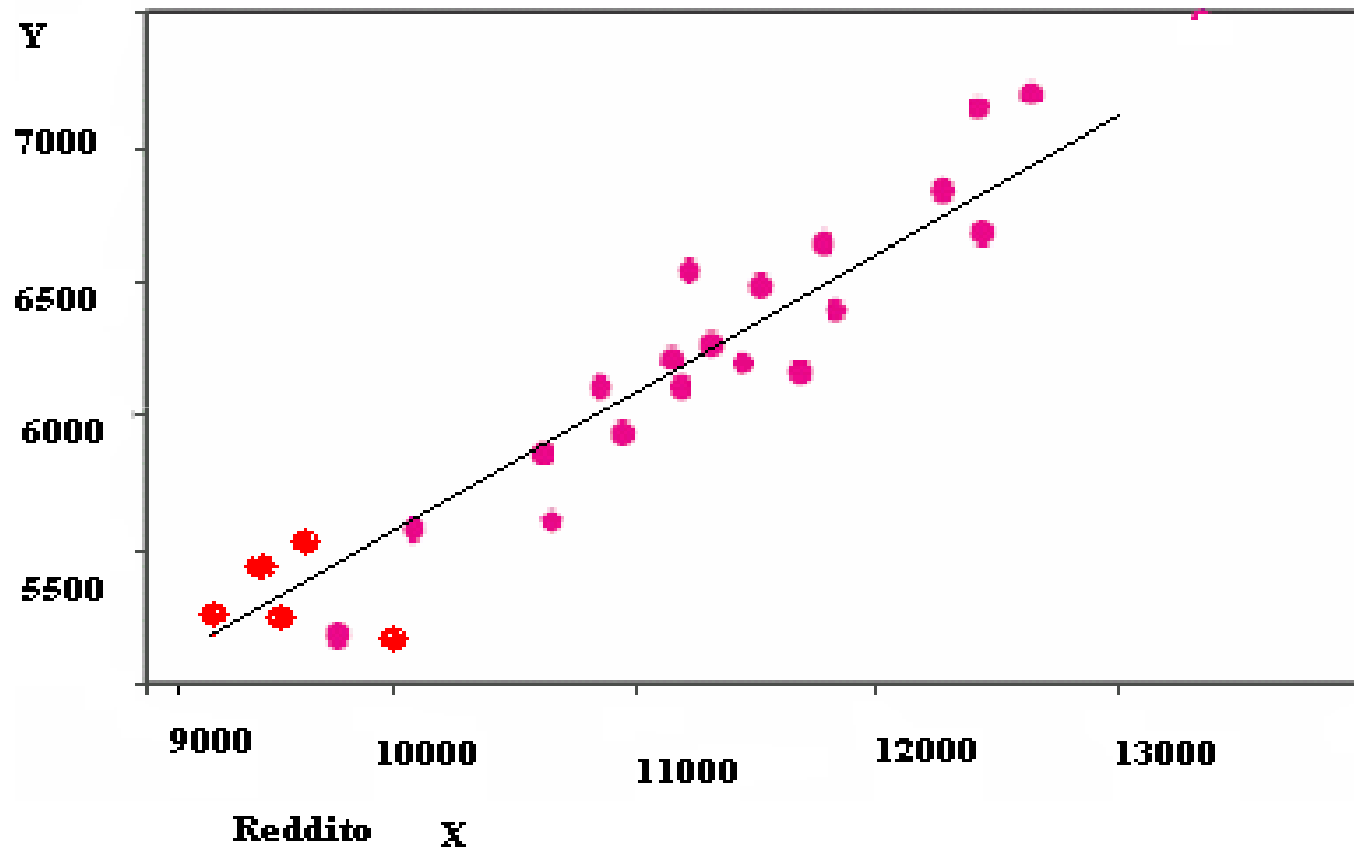
## Esempio

Per dislocare in maniera ottimale i punti vendita, un'azienda vuole stimare un modello lineare che prevede le vendite per nucleo familiare in funzione del reddito familiare disponibile sulla base dei dati provenienti da una indagine campionaria :

anno	Reddito (X)	Vendite (Y)	anno	Reddito (X)	Vendite (Y)
1	9098	5492	12	11307	5907
2	9138	5540	13	11432	6124
3	9094	5305	14	11449	6186
4	9282	5507	15	11697	6224
5	9229	5418	16	11871	6496
6	9347	5320	17	12018	6718
7	9525	5538	18	12523	6921
8	9756	5692	19	12053	6471
9	10282	5871	20	12088	6394
10	10662	6157	21	12215	6555
11	11019	6342	22	12494	6755

6

Il diagramma a dispersione indica una relazione lineare; all'aumentare del reddito disponibile aumentano le vendite:



L'analisi della regressione fornisce il modello:

$$Y=1922.39+0.381517X$$

7

Il modello riassume le informazioni dei dati campionari e non dimostra che un aumento del reddito determina un aumento delle vendite.

La teoria economica postula l'esistenza di un legame causa-effetto, i preliminari risultati precedenti, possono fornire l'evidenza empirica .

In generale una buona analisi statistica combinata alla teoria e all'esperienza può consentire di giungere a fondate conclusioni.

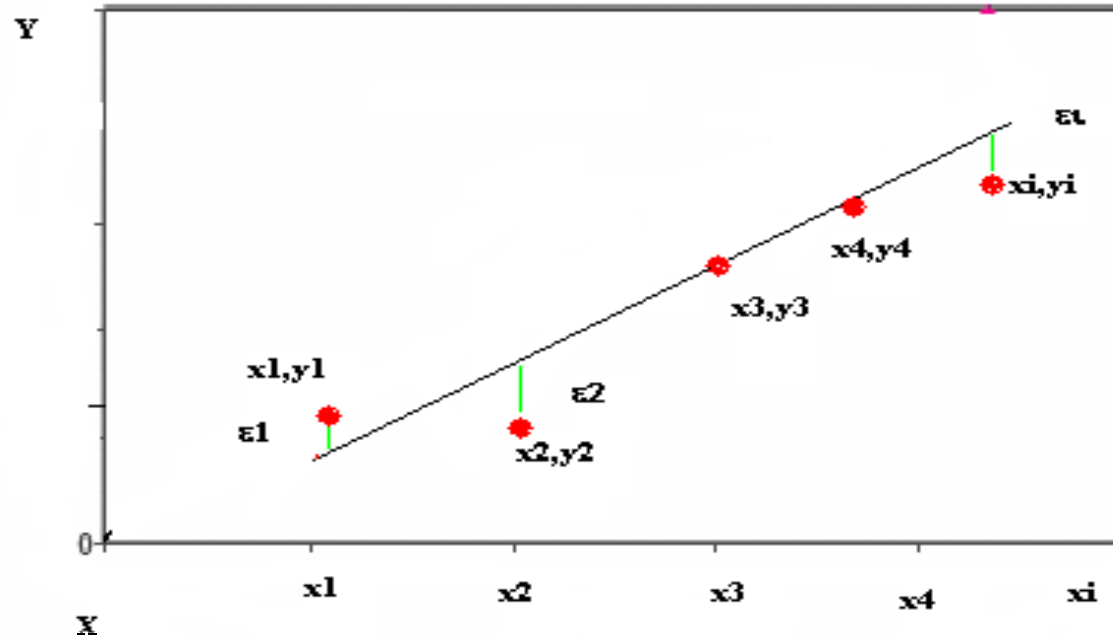
Nell'esempio, è noto dalla teoria che la quantità di beni acquistata in un certo mercato (Y) può essere modellizzata come funzione lineare del reddito disponibile (X): se il reddito disponibile è  $x_i$  la quantità acquistata sarà  $y_i$ .

Altri fattori tuttavia influenzano le quantità acquistate, alcuni sconosciuti (es. la diversa propensione al consumo delle famiglie), altri identificabili quali il prezzo del bene, e quello dei beni concorrenti, etc.



8

Ciò fa sì che:



Nel modello lineare semplice gli effetti di tutti i fattori diversi dal reddito, per spiegare la quantità acquistata vengono sintetizzati in una componente di errore:  $\varepsilon_i$ .

Per il generico valore sarà quindi  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

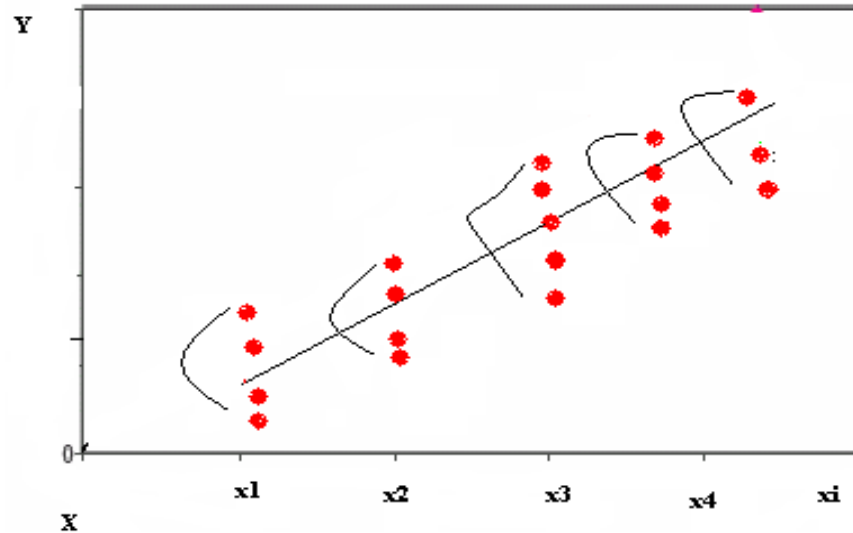
Al variare del campione inoltre avremo in corrispondenza di ogni dato valore  $x_i$  tutto un insieme di possibili valori (una distribuzione di valori) di Y 



9

Pertanto si assume che in P per ogni valore di X, sia il valore medio di Y funzione lineare di X:

$$Y = \beta_0 + \beta_1 x$$



Il modello di regressione lineare fornisce il valore atteso della variabile aleatoria Y (v. dipendente o risposta) quando X assume un particolare valore; in base all'ipotesi di linearità l'espressione per il valore atteso può essere scritta come:

$$E(Y/X=x) = \beta_0 + \beta_1 x$$

Il valore osservato di Y in corrispondenza ad un dato valore di X è invece pari al valore atteso (o media di P) più un errore aleatorio  $\varepsilon$

$$y_i = \beta_0 + \beta_1 x + \varepsilon_i$$

La variabile  $\varepsilon$ , errore aleatorio, rappresenta la variazione di Y non spiegata dalla relazione lineare.

In sintesi: negli studi empirici, la relazione tra  $Y$  e  $X$  non è mai funzionale (a un valore  $X$  corrispondono più valori di  $Y$ ).

Una **relazione statistica** tra la  $Y$  e la  $X$  può essere descritta da:

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

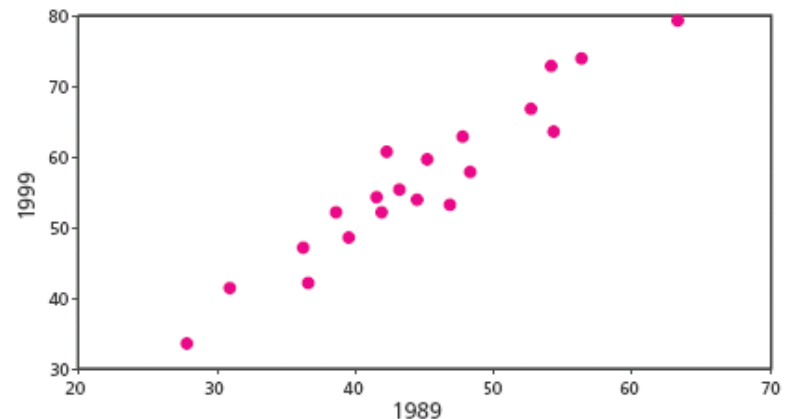
$f(X)$  definisce il contributo della  $X$

$\varepsilon$  rappresenta il contributo di tutti i fattori non osservati

⇒  $f(X)$  è una componente deterministica

⇒  $\varepsilon$  è una componente stocastica

⇒  $Y$  è una variabile casuale.





## Modello di regressione lineare semplice

Introducendo opportune assunzioni si ottiene il **modello di regressione lineare semplice**.

### Assunzione 1:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{per ogni osservazione } i=1, \dots, n$$

### Assunzione 2:

**Le  $\varepsilon_i$  sono variabili casuali indipendenti con valore atteso  $E(\varepsilon_i) = 0$  e varianza costante  $V(\varepsilon_i) = \sigma^2$  per ogni  $i=1, \dots, n$**

### Assunzione 3:

**I valori  $x_i$  della variabile esplicativa X sono noti senza errore**

# Modello di regressione lineare semplice

**Assunzione 1:** implica che la funzione  $f(X)$  è **lineare**.

**Assunzione 2:** implica che per ogni valore fissato di  $X$ , la  $Y$  possiede sempre lo stesso grado di variabilità (**ipotesi di omoschedasticità**). Inoltre, poiché la  $\varepsilon_i$  è una variabile casuale, anche  $Y$  è una variabile casuale.

Pertanto, le osservazioni  $y_i$  sono realizzazioni di variabili casuali

- indipendenti
- con valore atteso  $E(Y_i|X = x_i) = \beta_0 + \beta_1 x_i$
- con varianza  $V(Y_i|X = x_i) = \sigma^2$

## 13 Stima puntuale dei coefficienti di regressione

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

il valore di Y fornito dalla retta stimata dove  $\hat{\beta}_0$  e  $\hat{\beta}_1$  sono le stime dei coefficienti di regressione.

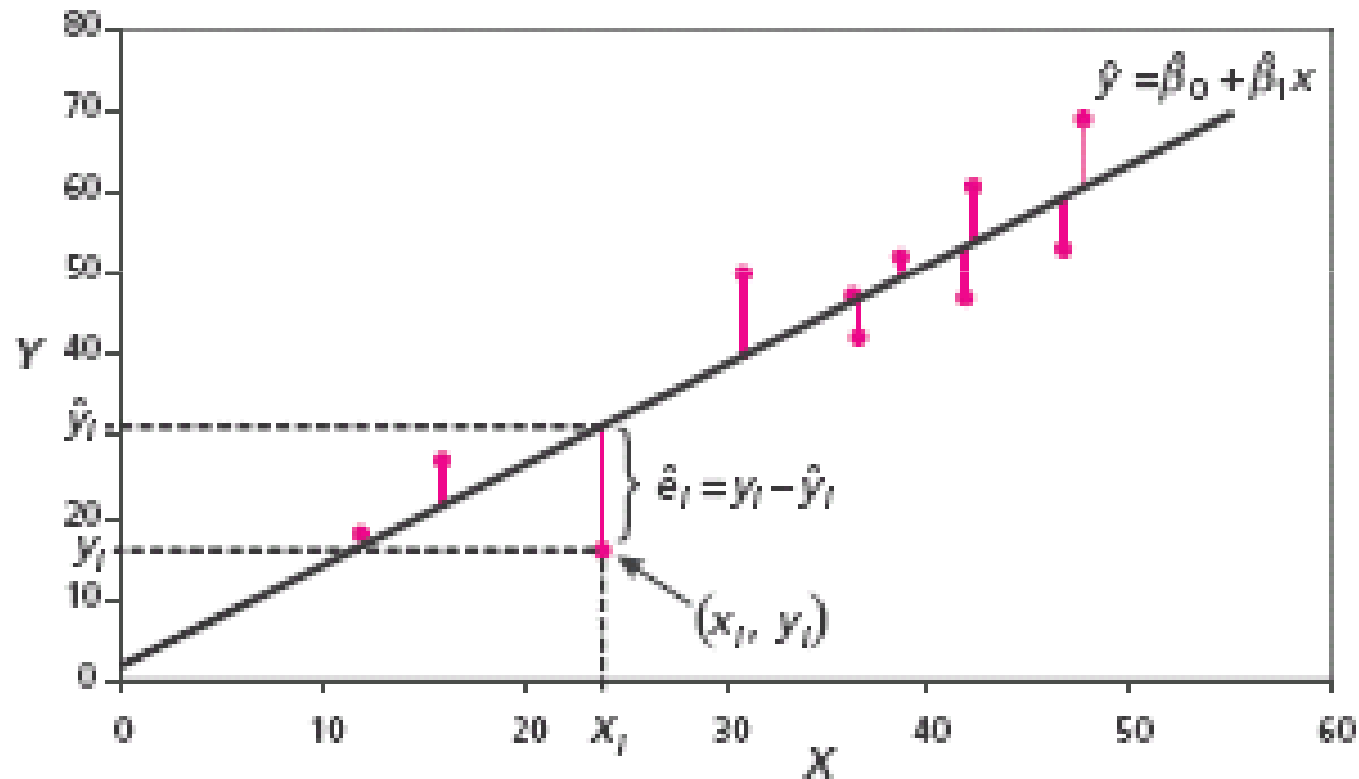
Metodo di stima  Metodo dei minimi quadrati

Consiste nel ricercare le stime di  $\beta_0$  e  $\beta_1$ , che rendono minima la funzione di perdita:

$$G(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

## 14 Stima puntuale dei coefficienti di regressione

Chiameremo **residuo**  $i$ -esimo la differenza tra il valore osservato  $y_i$  e quello fornito dalla retta stimata,  $\hat{y}_i$



## 15 Stima puntuale dei coefficienti di regressione

Procedimento:

1) Porre uguali a zero le derivate prime rispetto ai parametri

$$\begin{cases} \frac{\partial G(\beta_0, \beta_1)}{\partial \beta_0} = 0 \\ \frac{\partial G(\beta_0, \beta_1)}{\partial \beta_1} = 0 \end{cases}$$

2) Risolvendo il sistema si ottengono le stime dei minimi quadrati dei coefficienti di regressione

$$\hat{\beta}_1 = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

## Decomposizione della varianza

Le stime dei minimi quadrati possiedono un'importante proprietà, nota come **decomposizione della varianza totale**:

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

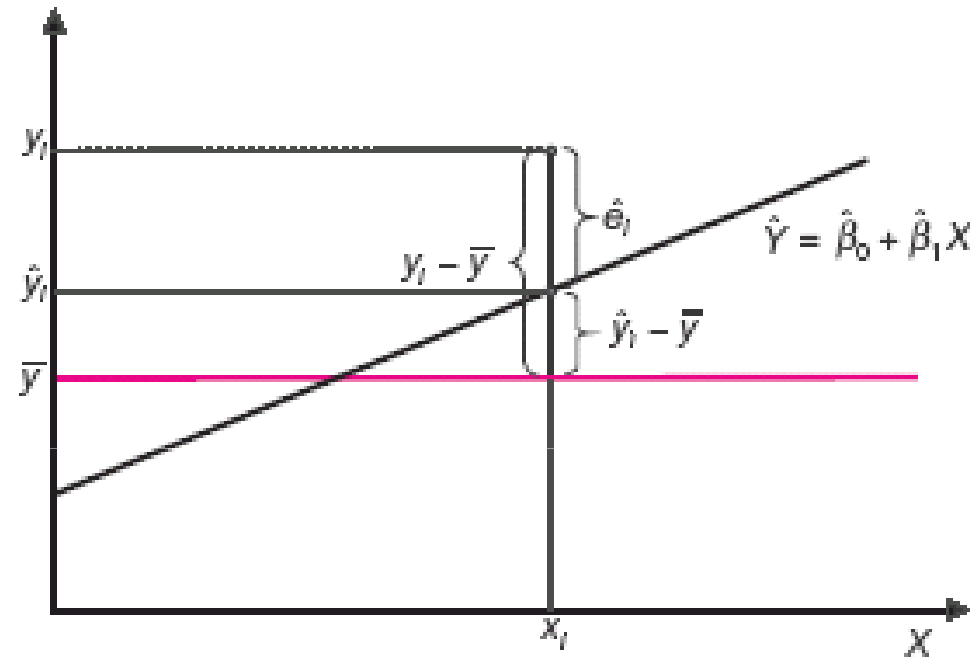
Somma totale dei quadrati (SQT)  $SQT = \sum_{i=1}^n (y_i - \bar{y})^2$

Somma dei quadrati della regressione (SQR)  $SQR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

Somma dei quadrati degli errori (SQE)  $SQE = \sum_{i=1}^n \hat{e}_i^2$



## Decomposizione della varianza



- **SQR=0**    **→** **SQE=SQT** e i valori stimati sono tutti uguali alla media campionaria  $\bar{y}$
- **SQR=SQT** **→** **SQE=0** e tutti i valori stimati sono uguali a quelli osservati.

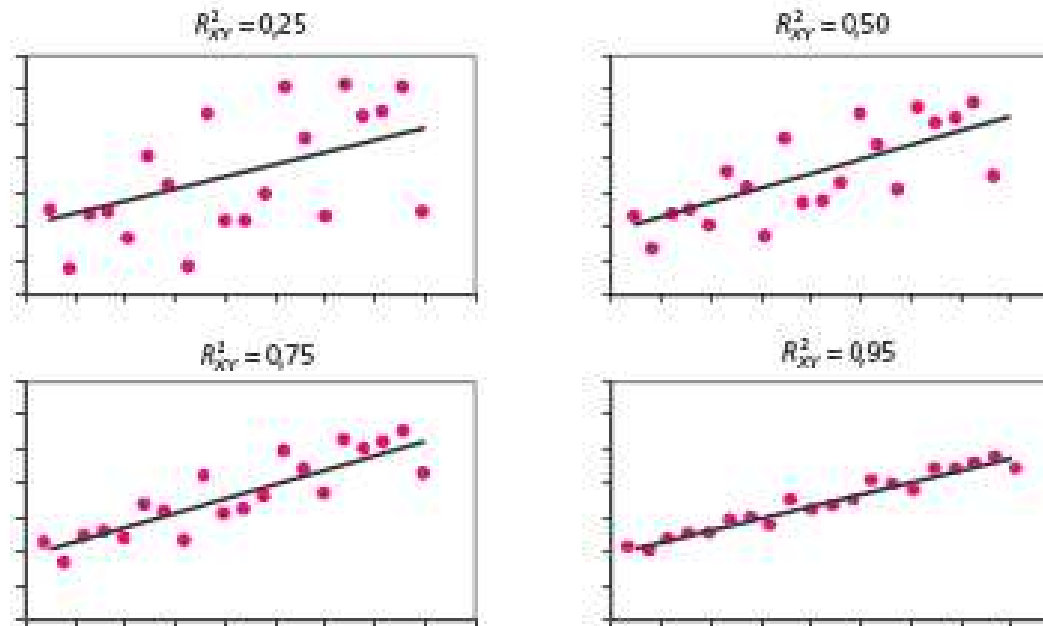
## Coefficiente di determinazione

Dalla relazione  $SQT=SQR+SQE$  si può definire un indice che misura la bontà di adattamento della retta di regressione.

Il rapporto

$$R_{XY}^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$

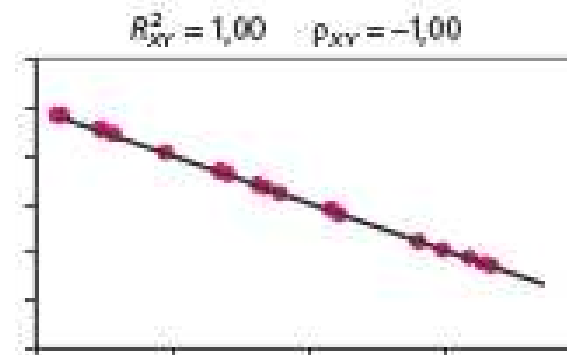
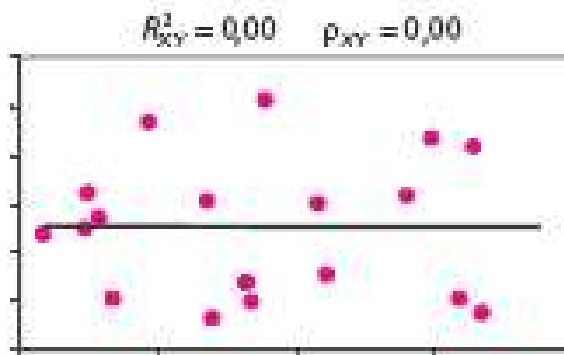
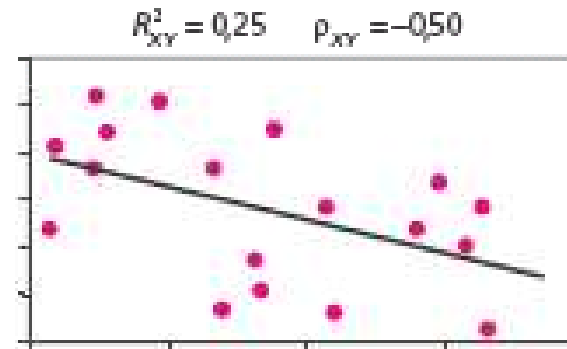
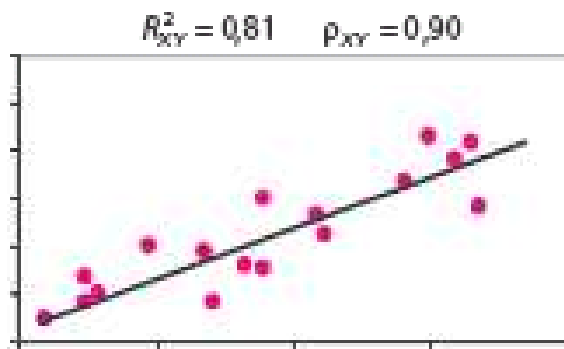
è detto **coefficiente di determinazione** e indica la proporzione di variabilità di Y spiegata dalla variabile esplicativa X, attraverso il modello di regressione.



# Coefficiente di determinazione

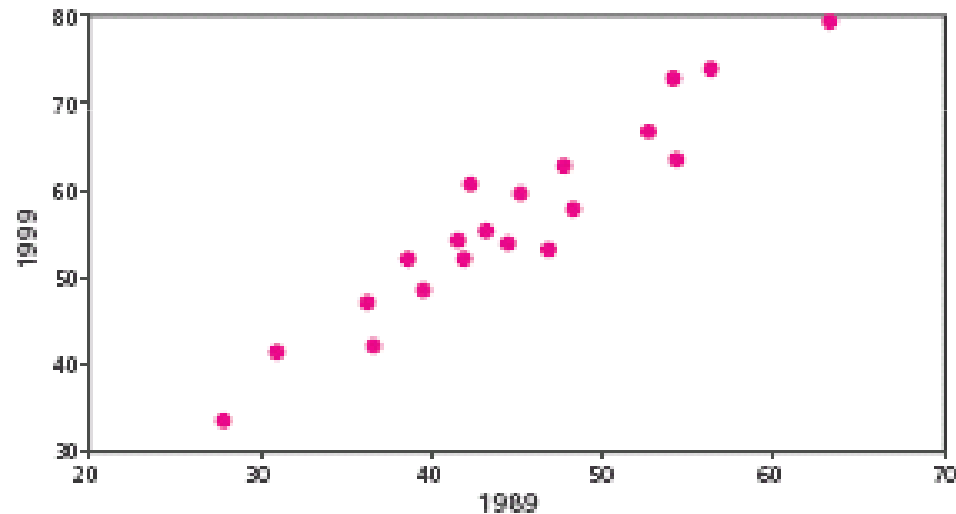
Si può dimostrare che il coefficiente di determinazione corrisponde al quadrato del coefficiente di correlazione lineare:

$$R_{XY}^2 = (\rho_{XY}^2) = \left( \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \right)^2$$



## Modello di regressione lineare - esempio

Su un campione di 20 aree amministrative si osserva il **reddito pro-capite** nel 1989 (X) e 1999 (Y).



Si ipotizza il seguente modello:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Area	X:1989	Y:1999
1	47,8	63,0
2	27,9	33,4
3	36,6	42,0
4	54,2	72,8
5	41,9	52,0
6	44,4	54,0
7	54,3	63,4
8	42,3	60,7
9	48,2	58,0
10	41,5	54,4
11	43,2	55,5
12	56,3	74,0
13	63,3	79,2
14	46,8	53,1
15	45,2	59,6
16	38,7	52,0
17	36,3	47,2
18	39,5	48,7
19	30,9	41,4
20	52,6	66,9

## Modello di regressione lineare - esempio

Si ottengono le seguenti stime dei coefficienti del modello:

$$\hat{\beta}_1 = 1,255 \quad \hat{\beta}_0 = 0,595$$

ossia la retta di regressione:  $\hat{y}_i = 0,595 + 1,255x_i$

Il coefficiente di correlazione è  $\rho_{XY} = 0,956$

SQT=2497,6 da cui:  $R_{XY}^2 = (0,956)^2 = 0,914$

ossia circa il 91% della variabilità totale di Y è spiegata dal modello di regressione.

## Proprietà degli stimatori dei coefficienti

### Proprietà degli stimatori dei minimi quadrati

1.  $B_0$  e  $B_1$  sono stimatori **corretti** di  $\beta_0$  e  $\beta_1$
2. Nella classe degli stimatori corretti di  $\beta_0$  e  $\beta_1$  che sono **funzioni lineari** delle  $y_i$ , gli stimatori dei minimi quadrati sono i **più efficienti**. (Gauss-Markov)
3. La varianza e covarianza degli stimatori dei minimi quadrati sono:

$$V(B_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad V(B_0) = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\text{Cov}(B_0, B_1) = -\sigma^2 \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

23

## Proprietà dello stimatore della risposta media

Per lo **stimatore della risposta media**  $\hat{Y}_i$  valgono le seguenti proprietà:

1. Lo stimatore  $\hat{Y}_i$  è **corretto**, ossia  $E(\hat{Y}_i) = \beta_0 + \beta_1 x_i$
2. La varianza è:

$$V(\hat{Y}_i) = \sigma^2 \left[ \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{h=1}^n (x_h - \bar{x})^2} \right]$$

Una misura della variabilità degli stimatori dei coefficienti di regressione e della risposta media è data dagli **errori standard**, ossia le radici quadrate delle varianze:

$$\sigma(B_0) = \sqrt{V(B_0)} \quad \sigma(B_1) = \sqrt{V(B_1)} \quad \sigma(\hat{Y}_i) = \sqrt{V(\hat{Y}_i)}$$

## 24 Errore standard

Ora sebbene il metodo M.Q. individua la retta che minimizza la differenza tra i valori osservati e quelli previsti, questa non conduce quasi mai a previsioni scovre da errori. E' quindi necessaria una statistica campionaria che misuri la variabilità degli scostamenti dei valori osservati dai previsti.

Inoltre, gli errori standard dipendono dalla quantità ignota:

$$\sigma^2 = V(Y_i) = V(\varepsilon_i)$$

pertanto la si sostituisce con una sua stima  $s^2$  ottenendo gli stimatori

$$s(B_0) \quad s(B_1) \quad s(\hat{Y}_i)$$

Lo stimatore che si utilizza per ottenere la stima della varianza è dato da:

$$s^2 = \frac{\sum_{i=1}^n \hat{e}_i^2}{n-2}$$

La radice quadrata è una misura della variabilità degli scostamenti dei valori osservati da quelli previsti dal modello e viene chiamato **errore standard della stima (di regressione)**.