

Covarianza

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Formula equivalente

$$\begin{aligned}\sigma_{xy} &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \\ &= \frac{1}{n} \sum_{i=1}^n (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}) = \\ &= \frac{1}{n} \sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \bar{y} - \bar{x} \frac{1}{n} \sum_{i=1}^n y_i + \frac{1}{n} \sum_{i=1}^n \bar{x} \bar{y} \\ \sigma_{xy} &= \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}\end{aligned}$$

The diagram illustrates the simplification of the covariance formula. In the third line, the terms $-\frac{1}{n} \sum_{i=1}^n x_i \bar{y}$ and $-\bar{x} \frac{1}{n} \sum_{i=1}^n y_i$ are circled in red. A blue arrow points from $\bar{y}$ to the first circle, and another blue arrow points from $\bar{x}$ to the second circle. The final line shows the simplified formula $\sigma_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}$.

Formula equivalente

$$\begin{aligned}\sigma_{xy} &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \\ &= \frac{1}{n} \sum_{i=1}^n (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}) = \\ &= \frac{1}{n} \sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \bar{y} - \bar{x} \frac{1}{n} \sum_{i=1}^n y_i + \frac{1}{n} \sum_{i=1}^n \bar{x} \bar{y} \\ \sigma_{xy} &= \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}\end{aligned}$$

The diagram illustrates the simplification of the covariance formula. In the third line, the terms $-\frac{1}{n} \sum_{i=1}^n x_i \bar{y}$ and $-\bar{x} \frac{1}{n} \sum_{i=1}^n y_i$ are circled in red. A blue arrow points from $\bar{y}$ to the first circle, and another blue arrow points from $\bar{x}$ to the second circle. The final line shows the simplified formula $\sigma_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}$.

Compute Correlation Coefficient

$$\rho_{xy} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Calcolo Covarianza e Correlazione

	X	Y	X ²	Y ²	XY
	1,6	10	2,56	100	16
	2	15	4	225	30
	3,5	20	12,25	400	70
	3	21	9	441	63
	3,2	24	10,24	576	76,8
	4	30	16	900	120
total	17,3	120	54,05	2642	375,8
mean	2,88333	20	9,00833	440,333	62,6333

Correlazione

- **Correlazioni positive** indica relazione lineari positive. Tanto più ρ si avvicina a 1 tanto più i punti si avvicinano a una linea retta. Se ρ si vicino a 0 indica maggiori distanza da una linea retta.
- **Correlazioni negative** indica relazione lineare decrescente, all'aumentare di un carattere l'altro tende a diminuire. Tanto più ρ si avvicina a -1 tanto più i punti si avvicinano a una linea retta (con inclinazione negativa) Se ρ si vicino a 0 indica maggiori distanza da una linea retta.

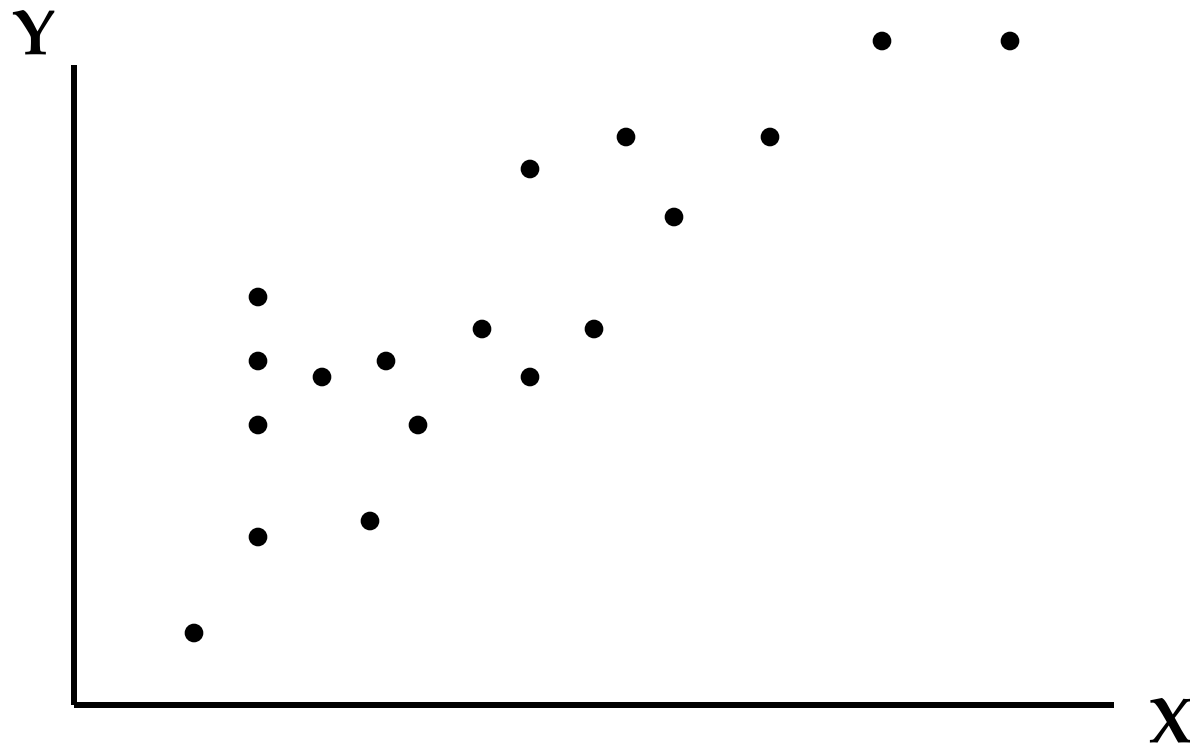
Correlazione

$\rho_{xy} = +1$ indica relazione lineare perfetta
positiva

$\rho_{xy} = 0$ indica assenza di relazione lineare

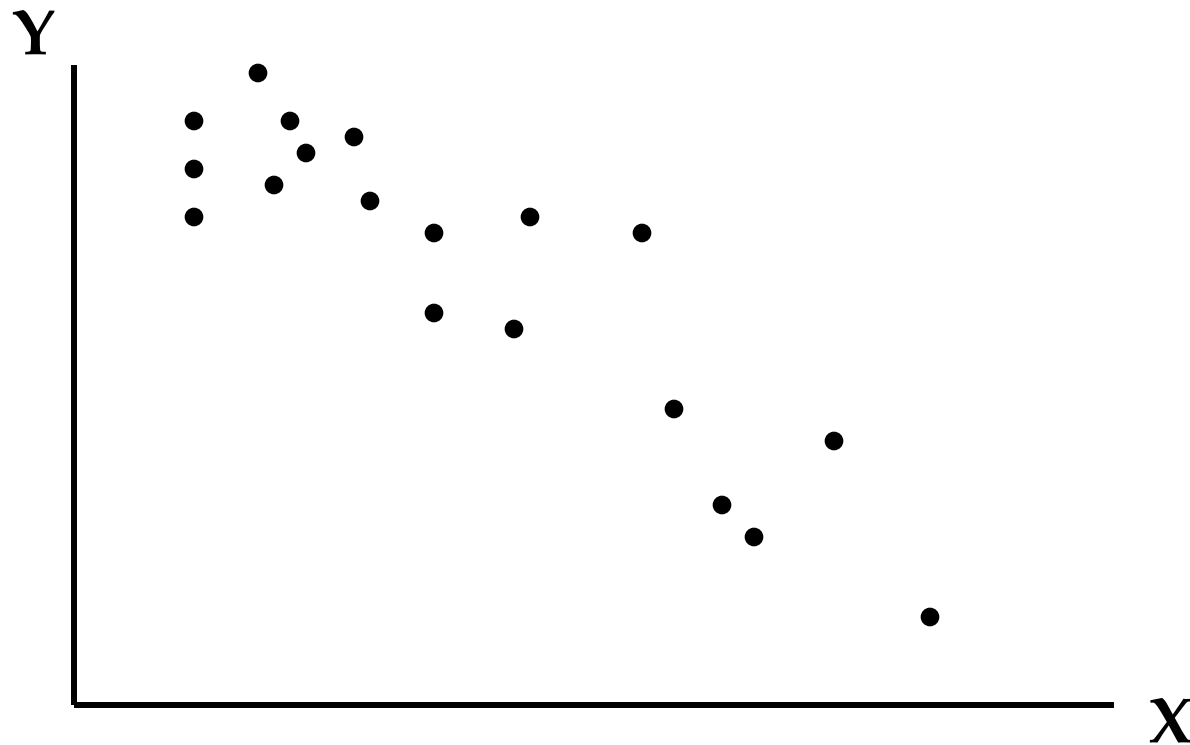
$\rho_{xy} = -1$ indica relazione lineare perfetta
negativa

Scatter Plots and Correlation



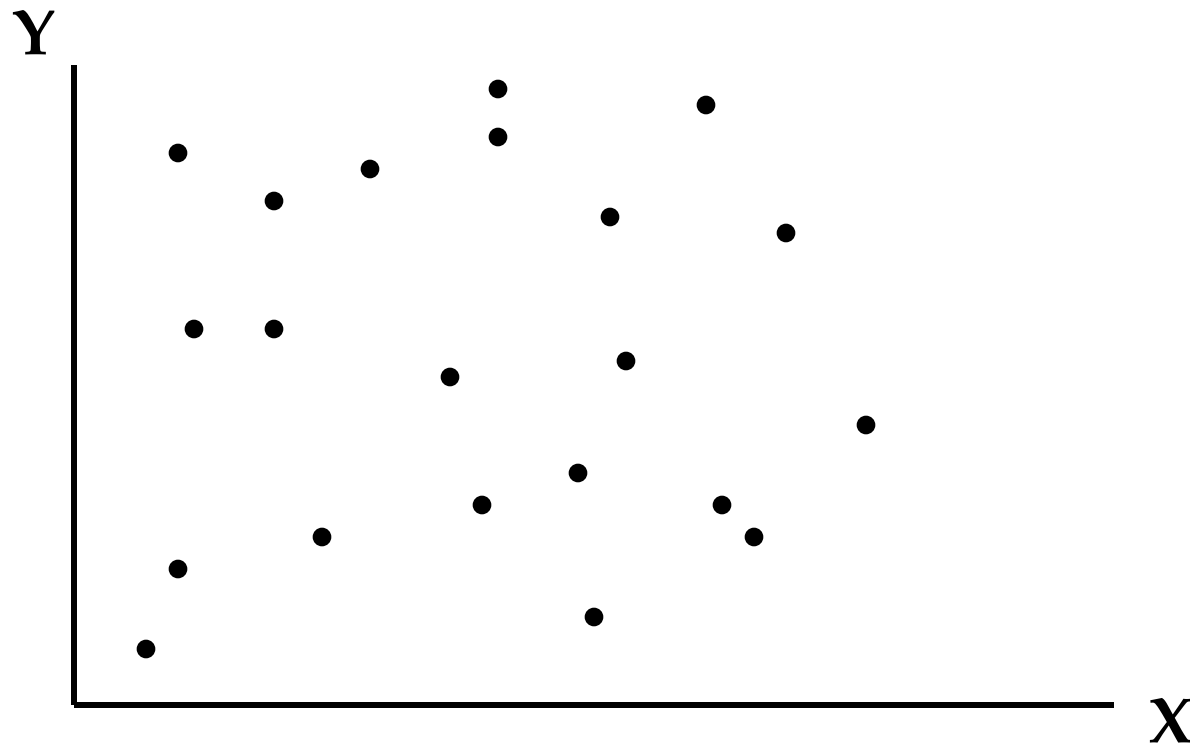
(a) $\rho = 0.8$

Scatter Plots and Correlation



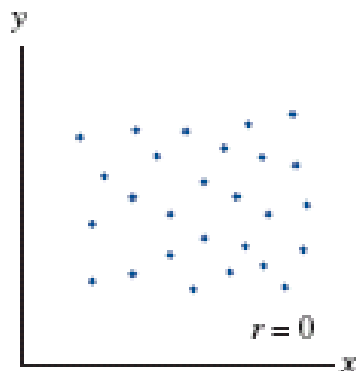
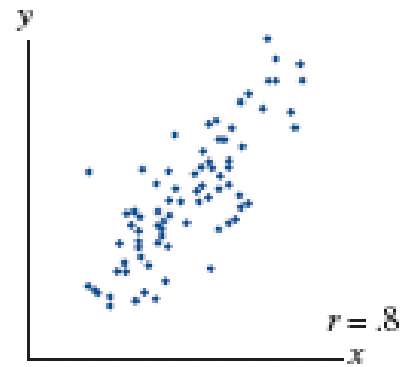
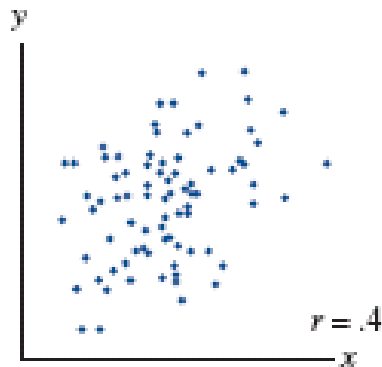
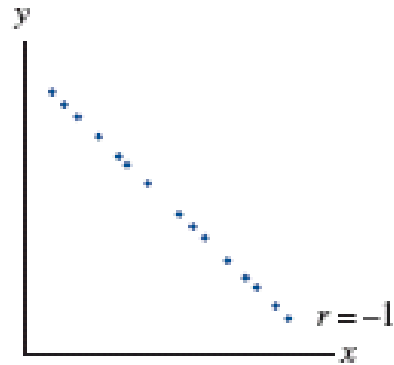
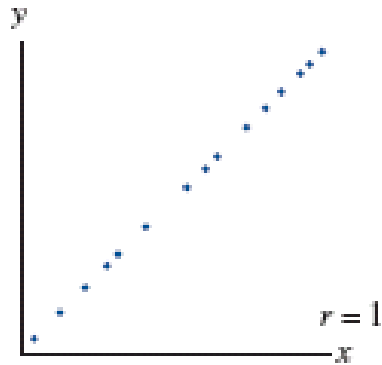
(b) $\rho = -0.8$

Scatter Plots and Correlation



(c) $\rho = 0$

Correlation coefficient:



Modello Lineare

Maura Mezzetti

`maura.mezzetti@uniroma2.it`

Introduzione

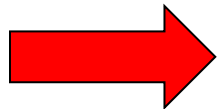
Dall'analisi ed inferenza riguardante una singola variabile statistica passiamo alla relazione tra (due) variabili statistiche. Le relazioni tra variabili importanti nell'analisi della realtà economico-aziendale possono essere matematicamente espresse come:

$$Y=f(X)$$

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



regressione lineare semplice

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 : $P=f(C)$.
- 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: $Q=f(P)$.

Obiettivo:

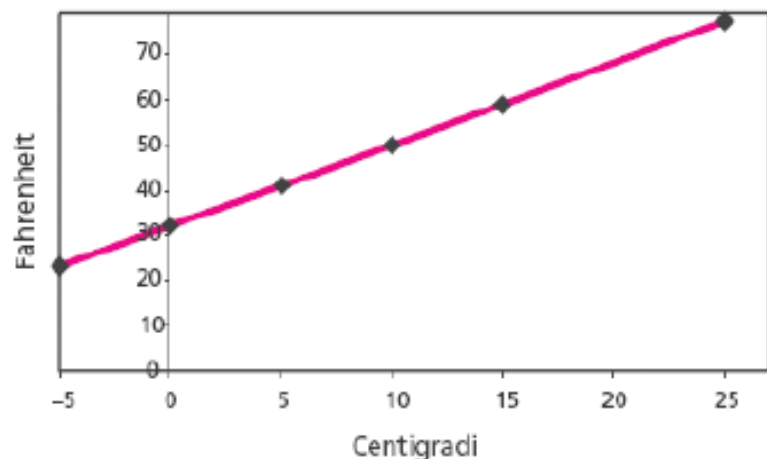
Date due variabili, X e Y , si è interessati a comprendere come la variabile Y (**dipendente** o **risposta**) sia influenzata dalla X (**esplicativa** o **indipendente**).

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:

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

β_0 = **intercetta**

β_1 = **coefficiente angolare**



Esempio: Utilizzo di Internet e Prodotto Interno Lordo (PIL)

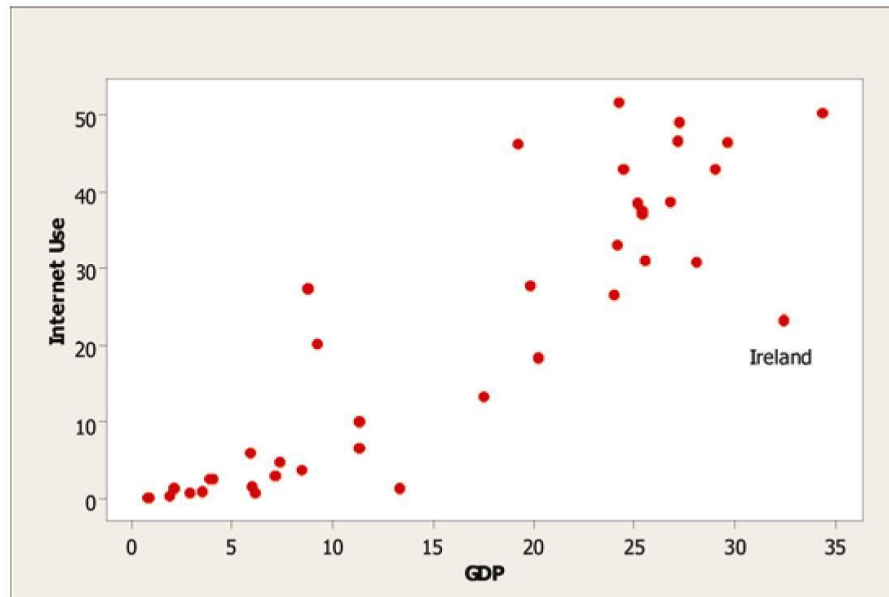
	INTERNET	GDP		INTERNET	GDP
Algeria	0.65	6.09	Japan	38.42	25.13
Argentina	10.08	11.32	Malaysia	27.31	8.75
Australia	37.14	25.37	Mexico	3.62	8.43
Austria	38.7	26.73	Netherlands	49.05	27.19
Belgium	31.04	25.52	New Zealand	46.12	19.16
Brazil	4.66	7.36	Nigeria	0.1	0.85
Canada	46.66	27.13	Norway	46.38	29.62
Chile	20.14	9.19	Pakistan	0.34	1.89
China	2.57	4.02	Philippines	2.56	3.84
Denmark	42.95	29	Russia	2.93	7.1
Egypt	0.93	3.52	Saudi Arabia	1.34	13.33
Finland	43.03	24.43	South Africa	6.49	11.29
France	26.38	23.99	Spain	18.27	20.15
Germany	37.36	25.35	Sweden	51.63	24.18
Greece	13.21	17.44	Switzerland	30.7	28.1
India	0.68	2.84	Turkey	6.04	5.89
Iran	1.56	6	United Kingdom	32.96	24.16
Ireland	23.31	32.41	United States	50.15	34.32
Israel	27.66	19.79	Vietnam	1.24	2.07
			Yemen	0.09	0.79

Scatterplot

Grafico di Dispersione

Rappresentazione grafica della relazione tra due variabili quantitative:

- Asse orizzontale: Variabile esplicativa, x
- Asse verticale: Variabile di risposta, y

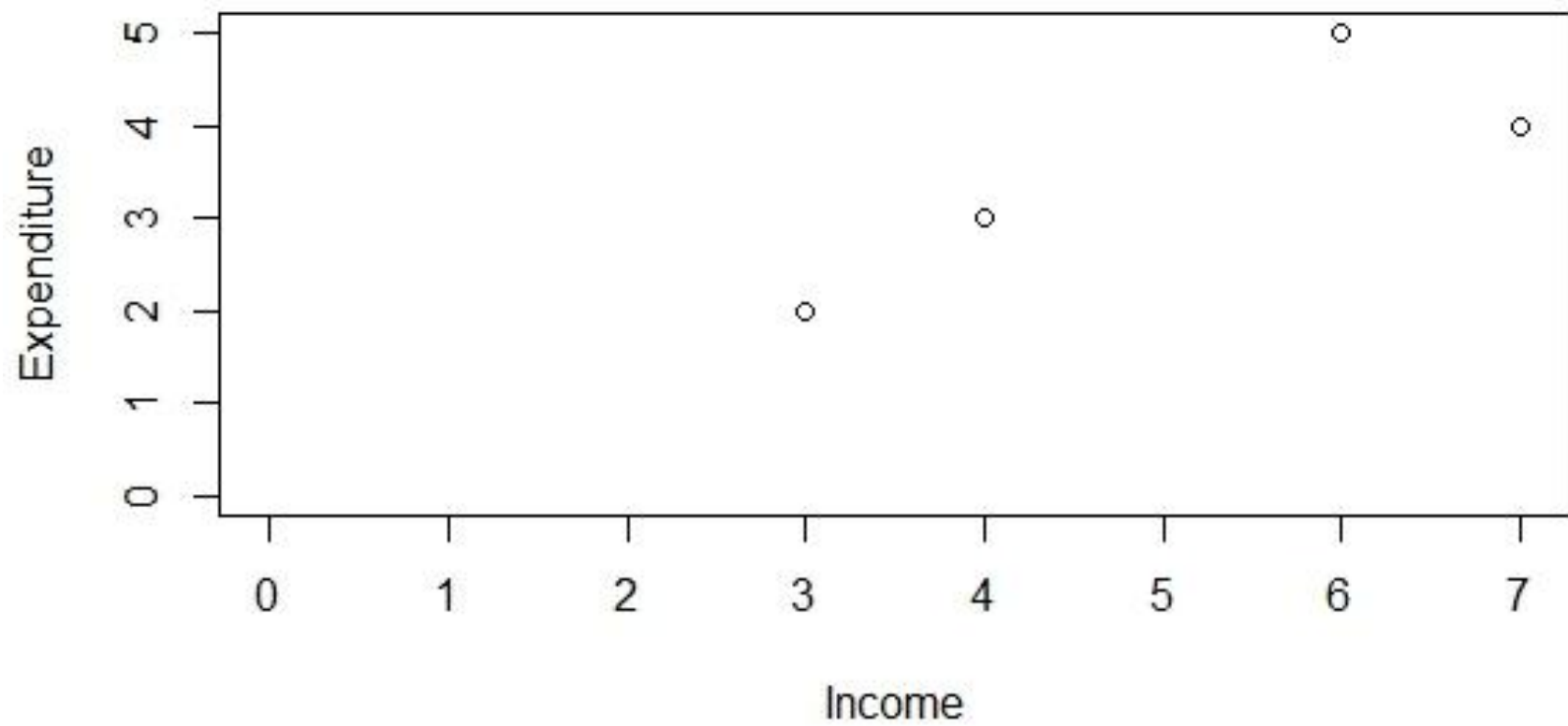


Esercizio

Si consideri la seguente distribuzione bivariata per spese e delle entrate per 5 famiglie:

- a) disegnare il grafico a dispersione dei dati osservati
- b) Valutare la relazione tra X e Y

Expenditures (1000 €)	Income (1000 €)
3	4
4	7
5	6
2	3
3	4



Y	X	Y ²	X ²	XY
3	4	9	16	12
4	7	16	49	28
5	6	25	36	30
2	3	4	9	6
3	4	9	16	12
17	24	63	126	88
3.4	4.8	12.6	25.2	17.6

$$Cov(X, Y) = 17.6 - 3.4 \times 4.8 = 1.28$$

$$Var(X) = 25.2 - 4.8^2 = 2.16 \quad Var(Y) = 12.6 - 3.4^2 = 1.04$$

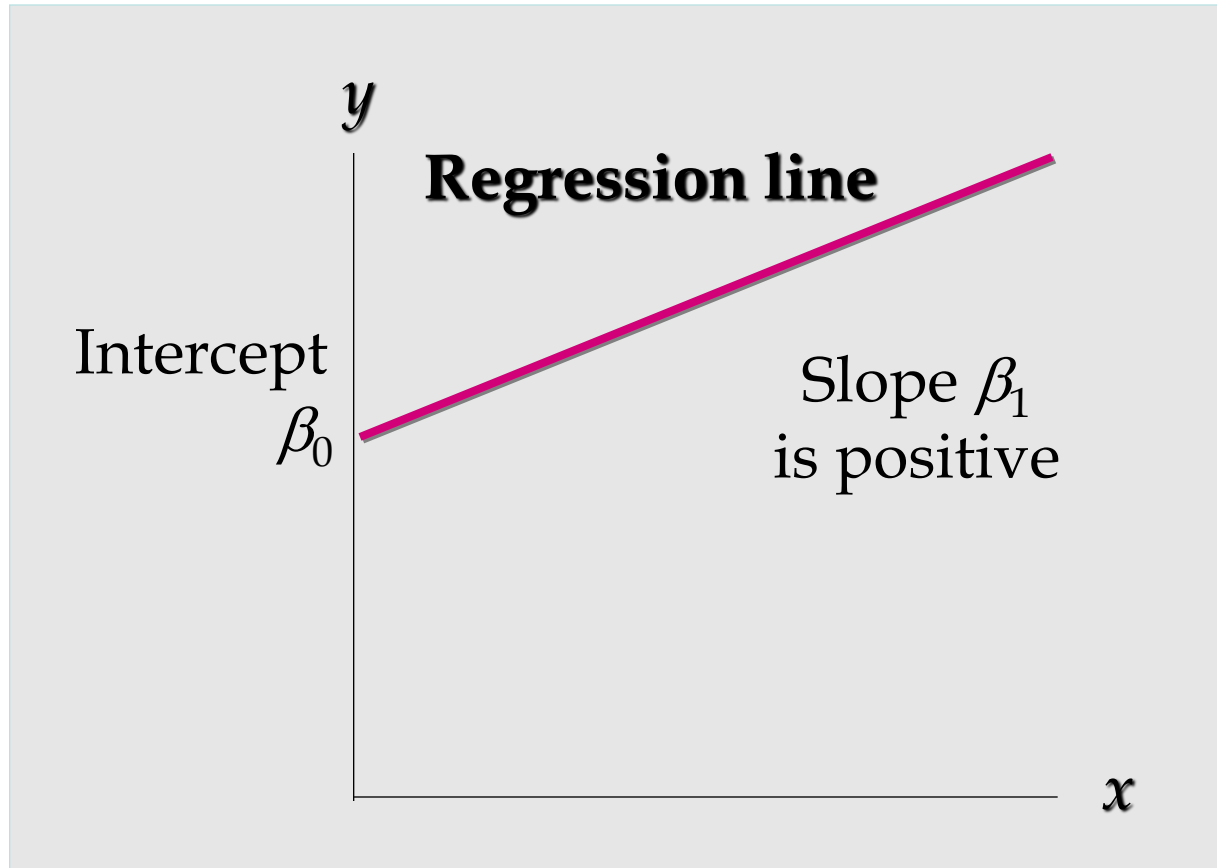
$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{1.28}{\sqrt{2.16 \times 1.04}} = 0.85 \quad \rho^2 = 0.7225$$

Regressione lineare semplice

- La regressione lineare semplice coinvolge una variabile indipendente e una variabile dipendente.
- Quando ci sono due o più variabili indipendenti si chiama regressione multipla.
- La relazione tra le due variabili è approssimata da una linea retta.

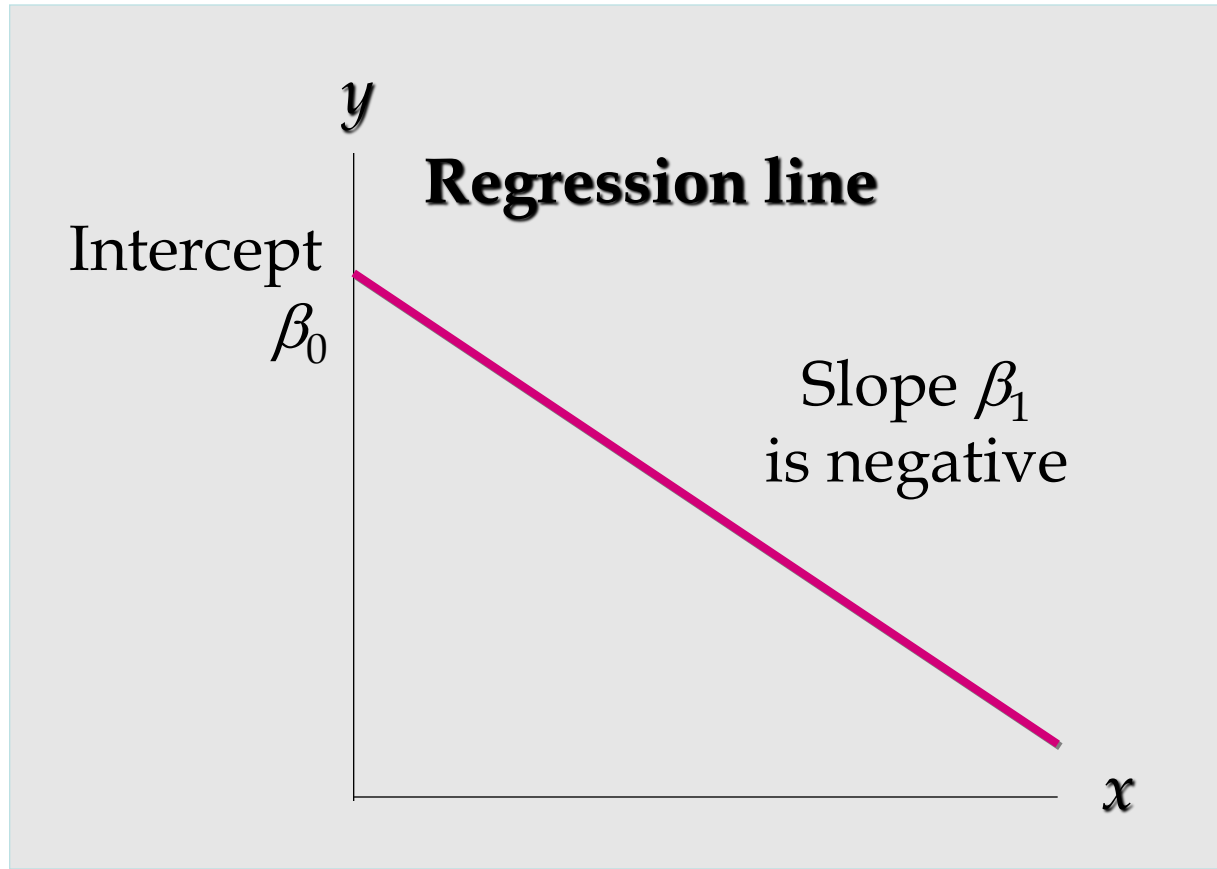
Equazione di Regressione Lineare Semplice

Relazione lineare positiva

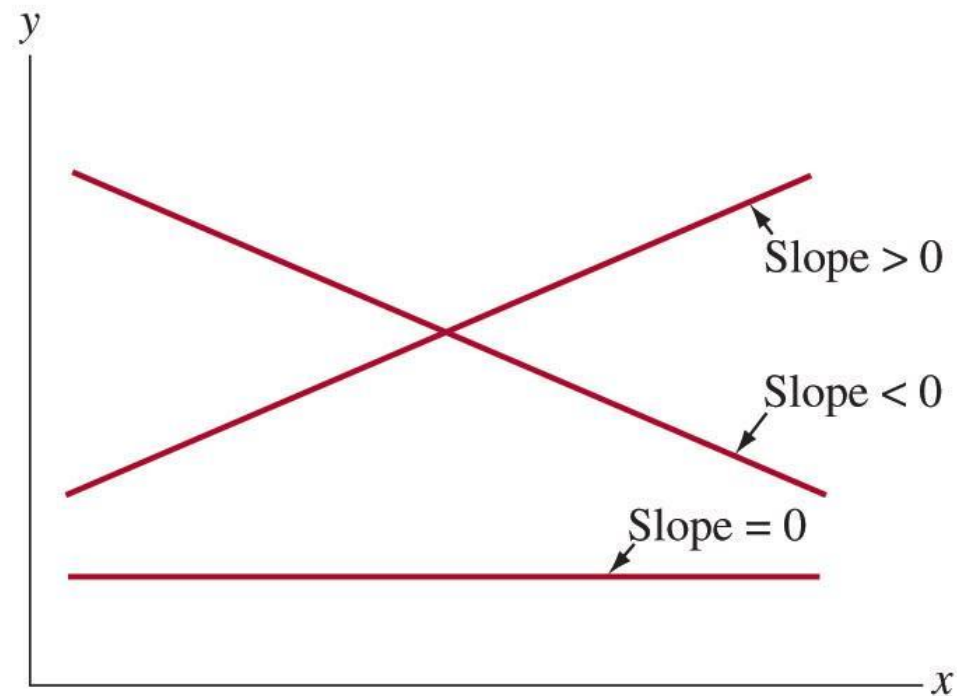


Equazione di Regressione Lineare Semplice

Relazione lineare negativa



Slope Values: Positivo, Negativo, Uguale a 0



Retta di regressione

- Ad ogni punto x , l'equazione:

$$\hat{y} = \beta_0 + \beta_1 x$$

- Fornisce un valore predittivo per y
- ... non tutti i soggetti con uguale valore di x avranno lo stesso valore per y



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 ε_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 ε_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 $V(Y_i|X = x_i) = \sigma^2$

Interpolazione

Come trovare un valore plausibile per i parametri incogniti?

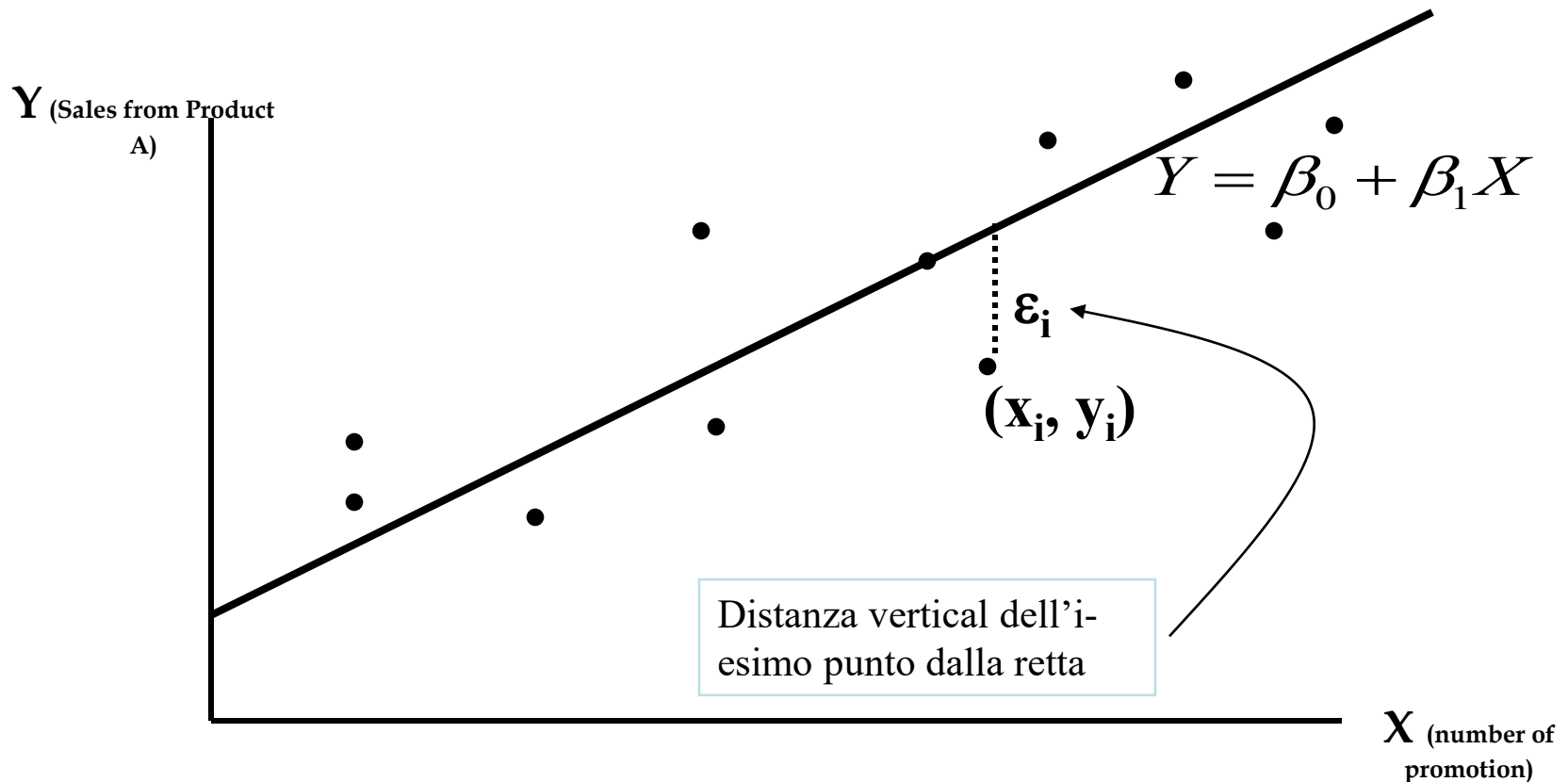
Minimizziamo la distanza tra i valori osservati e valori sulla retta

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

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

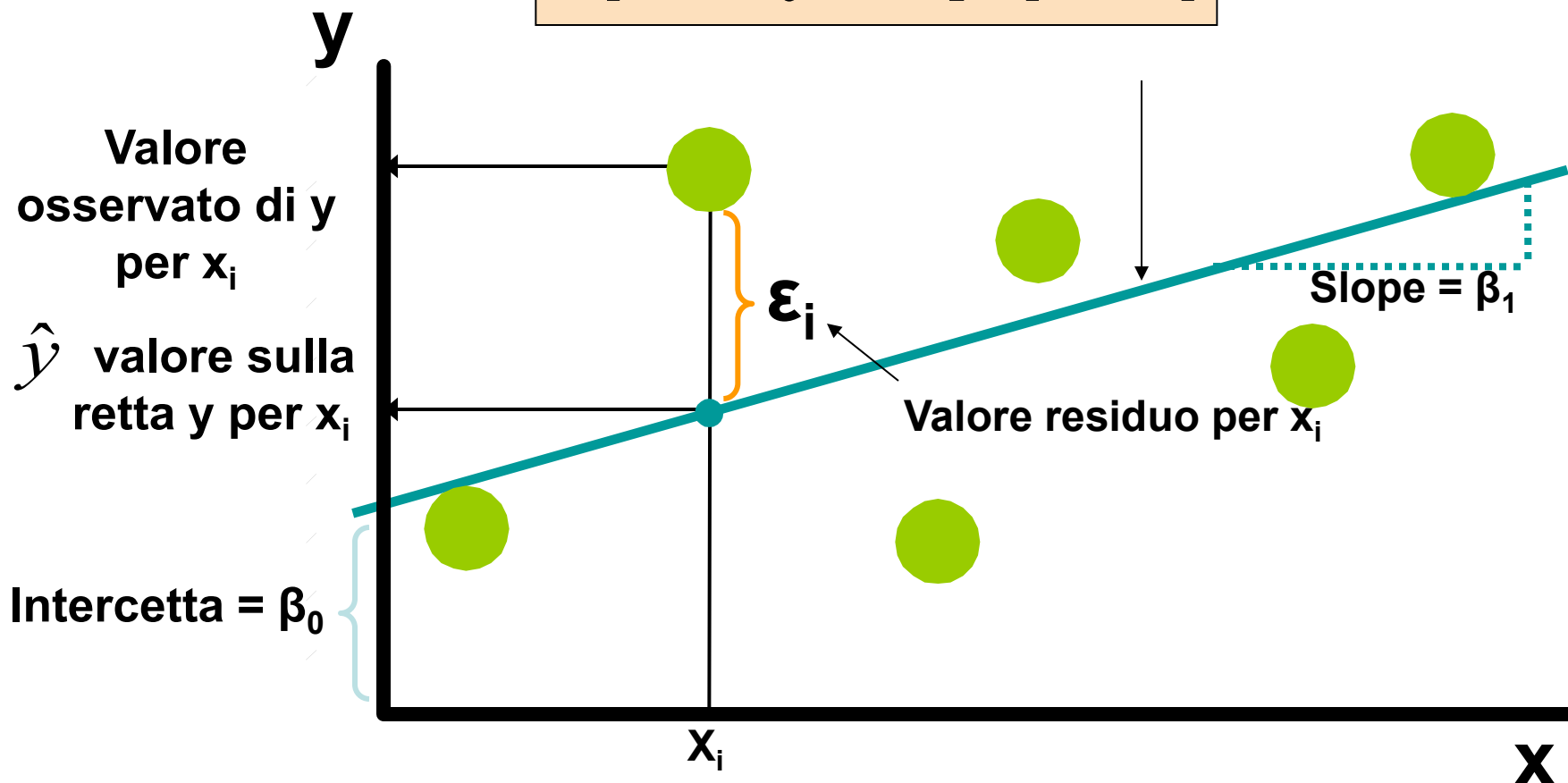
Interpolazione

Scegliamo come stime di β_0 e β_1 quei valori tale che sia minimizzata la distanza al quadrati tra i punti osservati e il corrispondente punto sulla retta.



Modello di regression lineare semplice

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



Stima della retta di regressione

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

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

Stimatori della retta di regressione

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Teorema di Gauss Markov

All'interno della classe degli stimatori lineari e corretti di β_0 e β_1 coefficienti del modello lineare,

gli stimatori dei minimi quadrati sono quelli a varianza minima

Stima di β_0, β_1 : il metodo dei minimi quadrati

Tra le infinite rette che si ottengono quando β_0 e β_1 variano nello spazio parametrico $\mathbb{R}^2$, cerchiamo la coppia di valori $(\hat{\beta}_0, \hat{\beta}_1)$ che minimizzano una particolare misura della discrepanza tra valori osservati, y_i , e quelli interpolati mediante la retta di regressione. In particolare, cerchiamo i valori tali che

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

Denotiamo con $G(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$ la nostra funzione obiettivo.

Derivando parzialmente rispetto ai parametri ed eguagliando a zero il risultato, si ottiene un sistema di due equazioni in due incognite:

$$\begin{aligned}\frac{\partial G}{\partial \beta_0} &= -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial G}{\partial \beta_1} &= -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) x_i = 0\end{aligned}$$

I valori dei coefficienti che minimizzano $G(\beta_0, \beta_1)$ devono soddisfare il sistema di due equazioni in due incognite (detto sistema di equazioni normali):

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_i x_i &= \sum_i y_i \\ \hat{\beta}_0 \sum_i x_i + \hat{\beta}_1 \sum_i x_i^2 &= \sum_i x_i y_i \end{aligned}$$

ovvero

$$\begin{aligned} \hat{\beta}_0 + \hat{\beta}_1 \bar{x} &= \bar{y} \\ \hat{\beta}_0 \bar{x} + \hat{\beta}_1 \frac{1}{n} \sum_i x_i^2 &= \frac{1}{n} \sum_i x_i y_i \end{aligned}$$

Risolvendo rispetto a $\hat{\beta}_0$ la prima equazione:

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

Sostituendo nella seconda equazione,

$$\left(\frac{1}{n} \sum x_i^2 - \bar{x}^2\right) \hat{\beta}_1 = \frac{1}{n} \sum_i x_i y_i - \bar{x} \bar{y}$$

pertanto,

$$\hat{\beta}_1 = \frac{\frac{1}{n} \sum_i x_i y_i - \bar{x} \bar{y}}{\frac{1}{n} \sum x_i^2 - \bar{x}^2} = \frac{\text{Cov}(x, y)}{V(x)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

Y	X	Y ²	X ²	XY
3	4	9	16	12
4	7	16	49	28
5	6	25	36	30
2	3	4	9	6
3	4	9	16	12
17	24	63	126	88
3.4	4.8	12.6	25.2	17.6

$$Cov(X, Y) = 17.6 - 3.4 \times 4.8 = 1.28$$

$$Var(X) = 25.2 - 4.8^2 = 2.16 \quad Var(Y) = 12.6 - 3.4^2 = 1.04$$

$$\beta_1 = \frac{1.28}{2.16} = 0.59 \quad \beta_0 = 3.4 - 0.59 \times 4.8 = 0.568$$

$$\hat{y} = 0.568 + 0.59x$$

Decomposizione della varianza

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

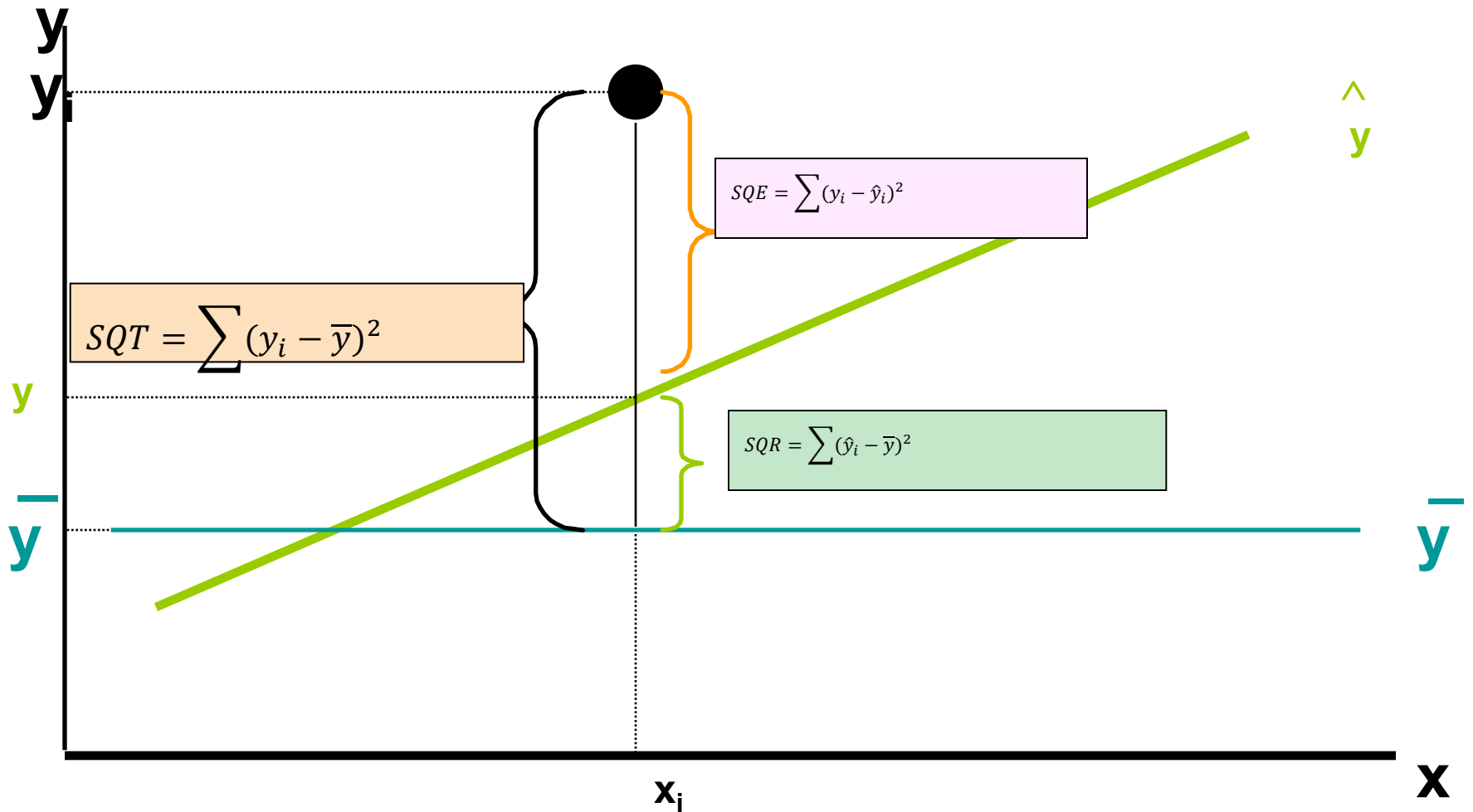
$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{e}_i^2$$

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

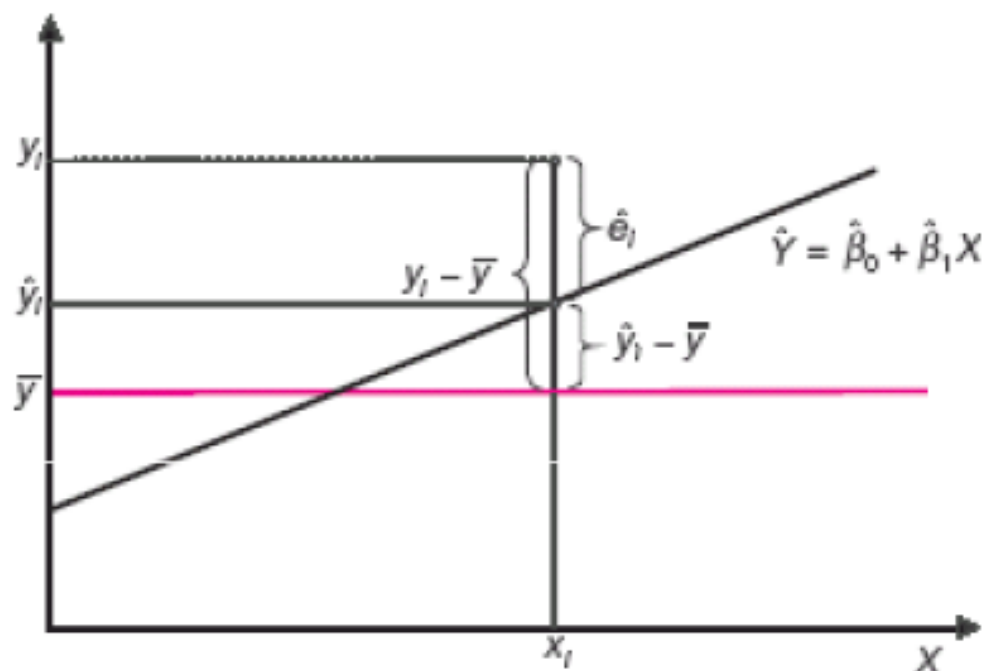
Somma dei quadrati della regressione (SQR) $SQR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

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

Misura della varianza



Decomposizione della varianza



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

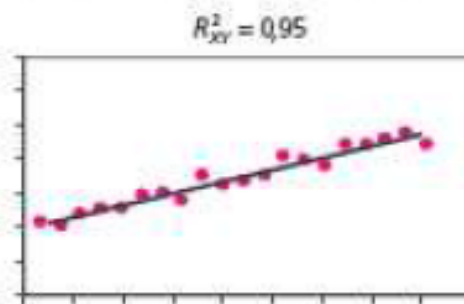
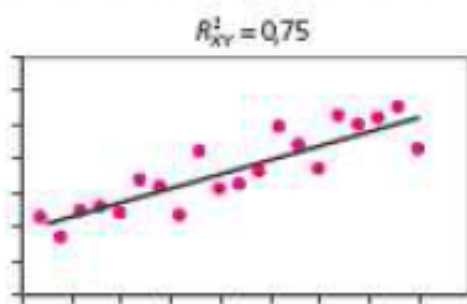
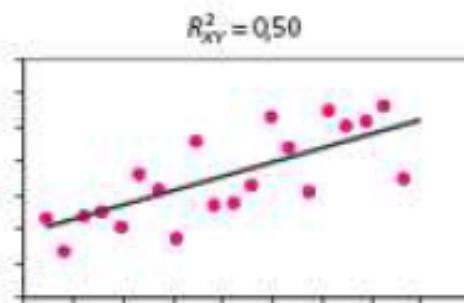
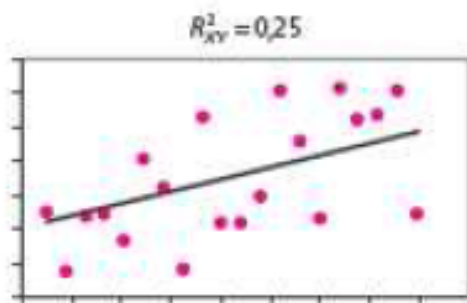
Coefficiente di determinazione

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

Il rapporto

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

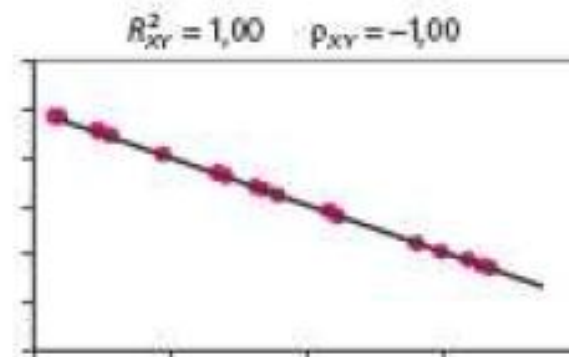
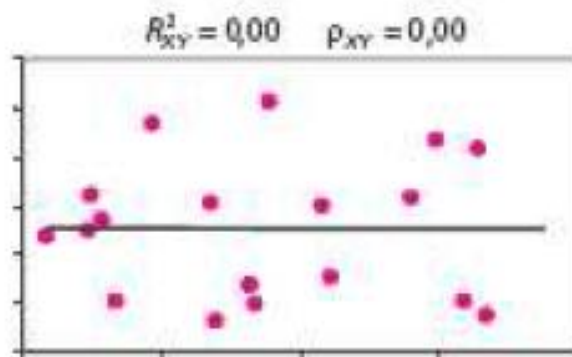
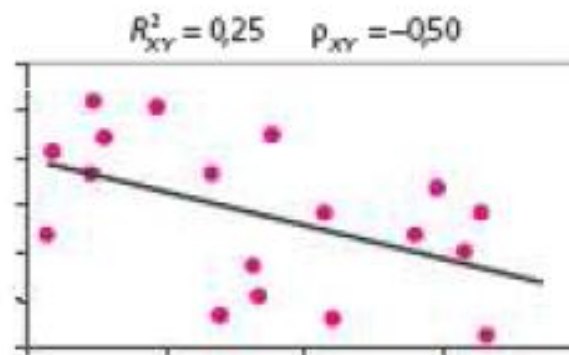
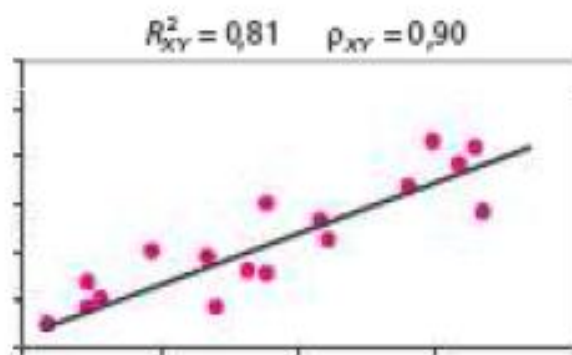
è detto **coefficiente di determinazione** e indica la proporzione di variabilità di Y spiegata dalla variabile esplicativa X, attraverso il modello di regressione.



Coefficiente di determinazione

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

$$R^2_{XY} = (\rho^2_{XY}) = \left(\frac{\sigma_{XY}}{\sigma_X \sigma_Y} \right)^2$$



Indice di determinazione

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

$$R^2 = \rho^2$$

Y	Y predicted	Y-Y pred	(Y-Y pred) ²	(Y-Y med) ²
3	2.928	0.072	0,005	0.16
4	4.698	-0.698	0,487	0.36
5	4.108	0.892	0,795	2.56
2	2.338	-0.338	0,114	1.96
3	2.928	0.072	0,005	0.16
			1,406	5,2

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 0.7225$$

Y	Y predicted	Y pred-Ymed	(Y pred-Ymed) ²	(Y-Y med) ²
3	2.928	-0.472	0.223	0.16
4	4.698	1.298	1.685	0.36
5	4.108	0.708	0.501	2.56
2	2.338	-1.062	1.128	1.96
3	2.928	-0.472	0.223	0.16
			3,76	5,2

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 0.7225$$

Y	Y predicted	Y pred-Ymed	(Y pred-Ymed) ²	(Y-Y med) ²
3	2.928	-0.472	0.223	0.16
4	4.698	1.298	1.685	0.36
5	4.108	0.708	0.501	2.56
2	2.338	-1.062	1.128	1.96
3	2.928	-0.472	0.223	0.16
$\sum_{i=1}^n (\hat{Y}_i - Y_i)^2$			3,76	5,2

$$\frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - R^2 = 0.2775$$

$$\sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = 0.2775 \times \sum_{i=1}^n (Y_i - \bar{Y})^2 = 1,443$$

Valori interpolati (predetti) e residui

- ▶ Valori interpolati

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

- ▶ Residui dei minimi quadrati

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

I residui sono importanti per effettuare una serie di diagnosi circa il modello di regressione stimato.

I residui sono utilizzati al fine di stimare σ^2 .

Proprietà dei residui e dei valori predetti

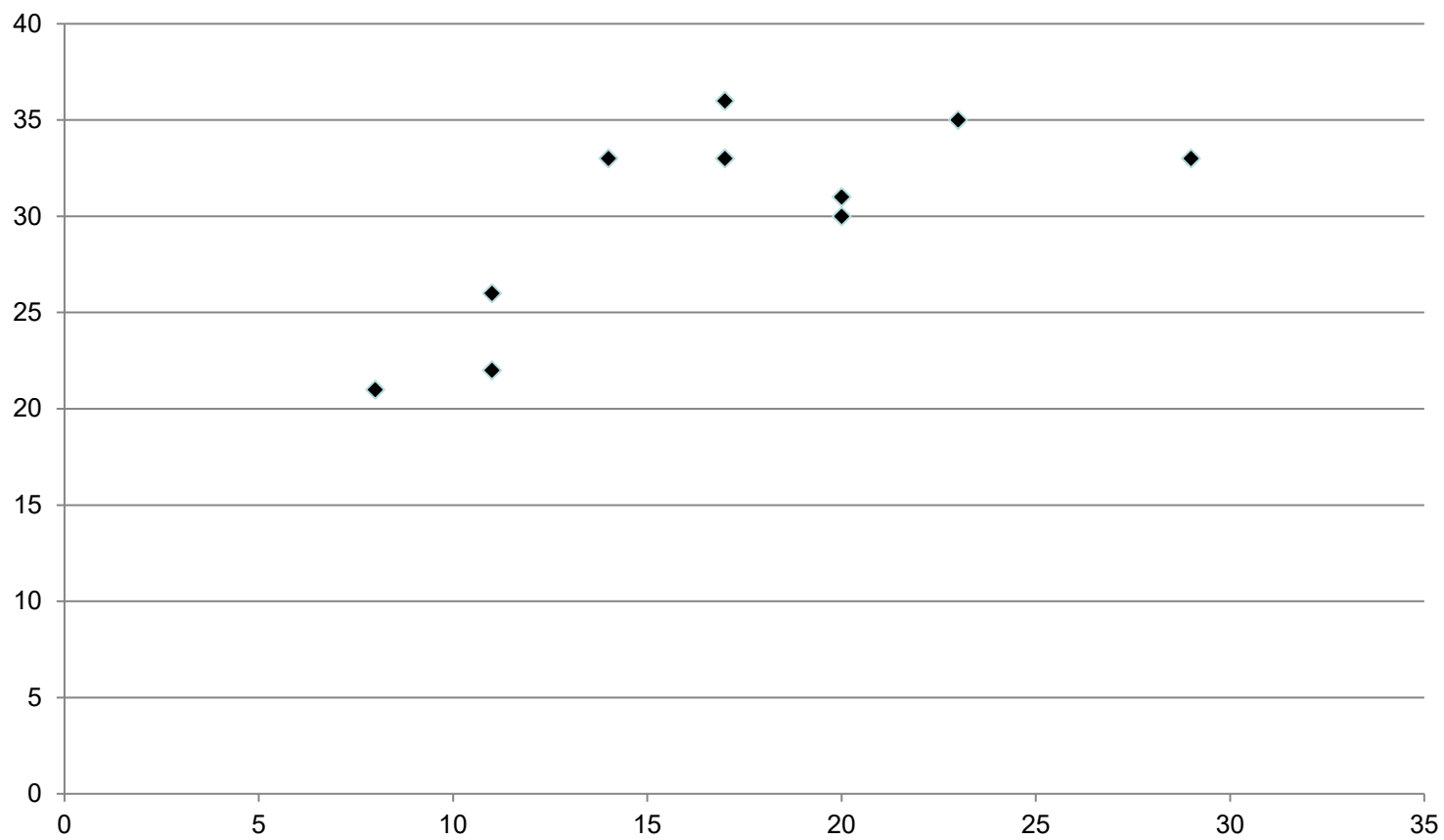
- ▶ I residui dei minimi quadrati hanno somma (media) nulla: $\sum_i \hat{e}_i = 0$. (le stime dei MQ soddisfano la prima delle equazioni normali: $\frac{\partial G}{\partial \beta_0} = 0$).
- ▶ I residui sono incorrelati con i valori x_i : $\sum_i x_i \hat{e}_i = 0$ (le stime dei MQ soddisfano la seconda delle equazioni normali: $\frac{\partial G}{\partial \beta_1} = 0$).
- ▶ I valori predetti, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ sono incorrelati con i residui $\hat{e}_i$: $\sum_i \hat{y}_i \hat{e}_i = 0$. Questa proprietà discende dalle due precedenti.
- ▶ La media dei valori predetti, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ coincide con la media dei valori osservati: $\frac{1}{n} \sum_i \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y}$ (questo consegue dalla prima proprietà).
- ▶ La retta di regressione passa per il baricentro $(\bar{x}, \bar{y})$: il valore sulla retta di regressione corrispondente a $\bar{x}$ è $\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y}$.

Esempio

I dati sottostanti mostrano i voti ottenuti da 10 studenti nei test di Matematica e Fisica

Pupil	A	B	C	D	E	F	G	H	L	M
Maths	20	23	8	29	14	11	11	20	17	17
Physics	30	35	21	33	33	26	22	31	33	36

	x	y	x²	y²	xy
	20	30	400	900	600
	23	35	529	1225	805
	8	21	64	441	168
	29	33	841	1089	957
	14	33	196	1089	462
	11	26	121	676	286
	11	22	121	484	242
	20	31	400	961	620
	17	33	289	1089	561
	17	36	289	1296	612
sum	170	300	3250	9250	5313
mean	17	30	325	925	531,3



$$Var(X) = 325 - 17^2 = 36$$

$$Var(Y) = 925 - 30^2 = 25$$

$$Cov(X, Y) = 531.3 - 17 \times 30 = 21.3$$

$$Corr(X, Y) = \frac{21.3}{\sqrt{36 \times 25}} = 0.71$$

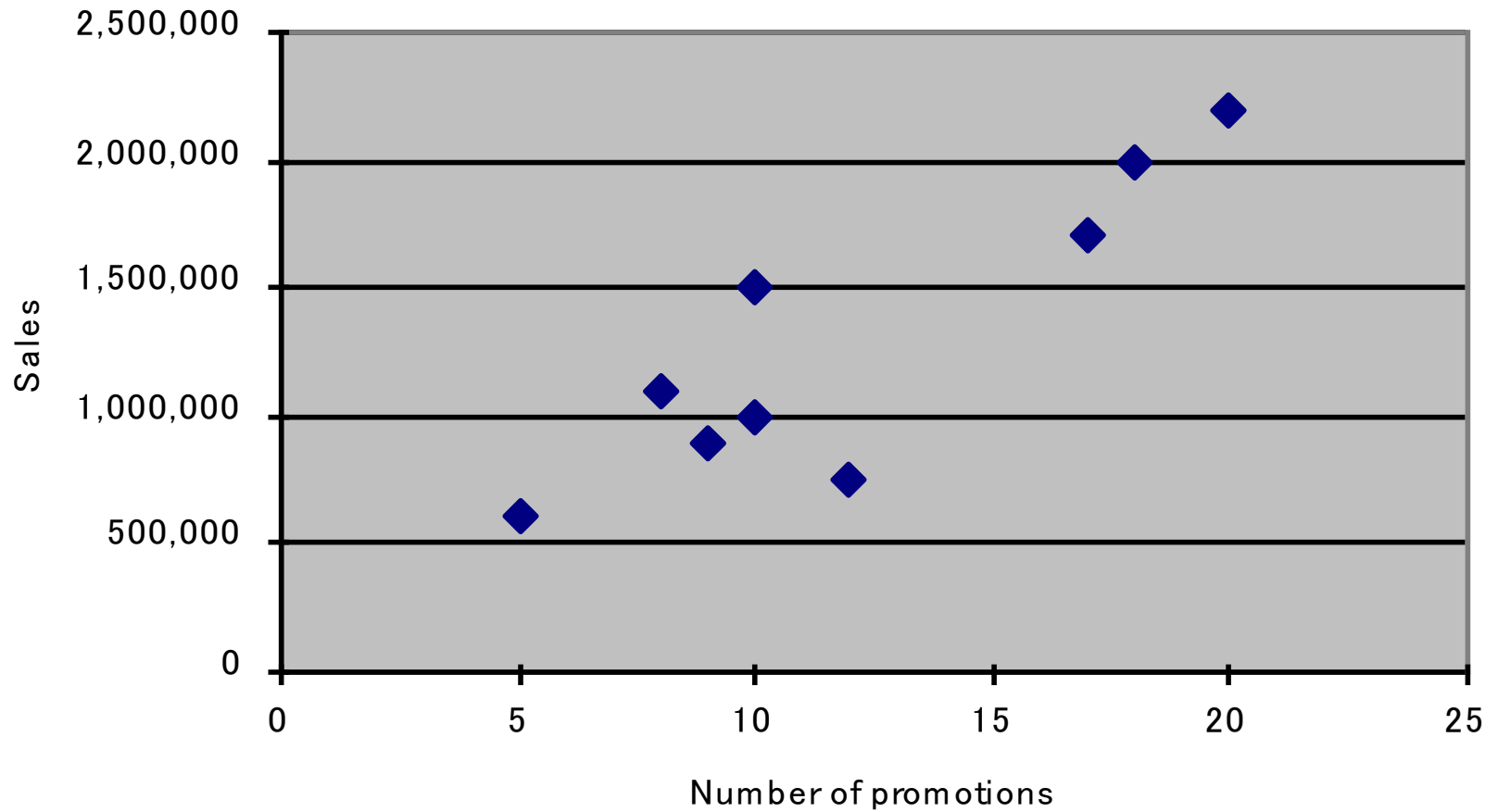
$$\beta_1 = \frac{21.3}{36} = 0.59$$

$$\beta_0 = 30 - 17 \times 0.59 = 19.94$$

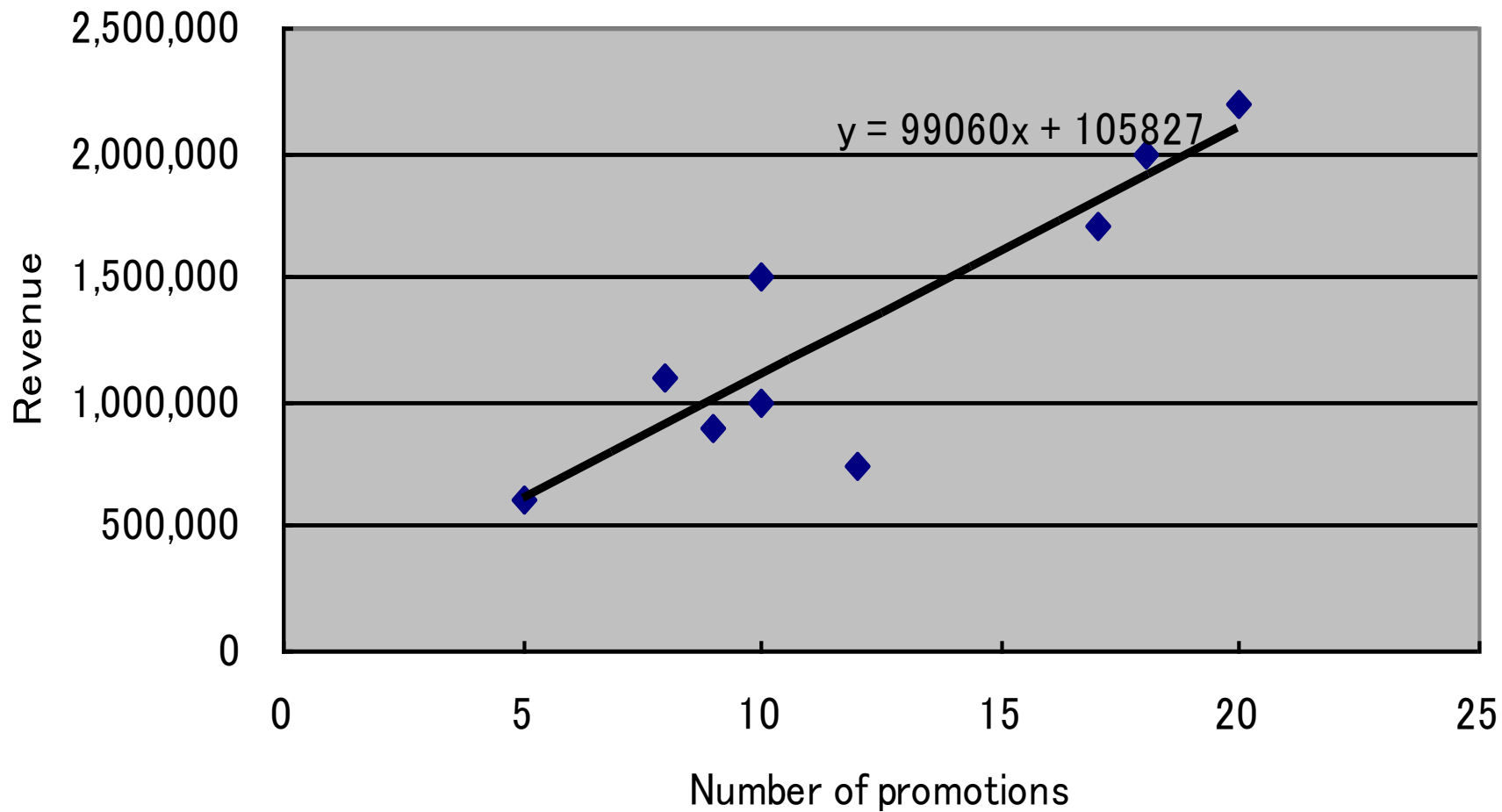
Come misurare la direzione della relazione: Esempio

Numero di promozioni	Ricavi dal prodotto A (in 1000 yen)
5	600
10	1000
8	1100
9	900
10	1500
12	750
20	2200
18	2000
17	1700

Product A: Promotion and sales



Product A: Promotion and sales



Inferenza sui coefficienti di regressione

Nel modello $Y = \beta_0 + \beta_1 X + \epsilon$ introduciamo l'assunzione

$$\epsilon \sim N(0, \sigma^2).$$

Le conseguenze dell'assunzione di normalità sono che, estratto un campione casuale,

- ▶ $Y_i | X_i = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$
- ▶ B_0 e B_1 sono combinazioni lineari di v.c. normali. Pertanto,

$$B_0 \sim N \left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right] \right),$$

$$B_1 \sim N \left(\beta_1, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \right)$$

Proprietà degli stimatori

$$\begin{aligned} E(\hat{\beta}_1) &= E\left(\frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) = \frac{\sum_{i=1}^n (X_i - \bar{X})E(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})(\beta_0 + \beta_1 X_i - \beta_0 - \beta_1 \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})(\beta_1 (X_i - \bar{X}))}{\sum_{i=1}^n (X_i - \bar{X})^2} = \beta_1 \end{aligned}$$

$$E(\hat{\beta}_0) = E(\bar{Y}) - E(\hat{\beta}_1)\bar{X} = \beta_0 + \beta_1 \bar{X} - \beta_1 \bar{X} = \beta_0$$

Proprietà degli stimatori

$$\begin{aligned}\text{Var}(\hat{\beta}_1) &= \text{Var}\left(\frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 \text{Var}(Y_i)}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} = \\ &= \sigma^2 \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} \\ \text{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\end{aligned}$$

$$\text{Var}(\hat{\beta}_0) = \text{Var}(\bar{Y}) + \text{Var}(\hat{\beta}_1) \bar{X}^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)$$

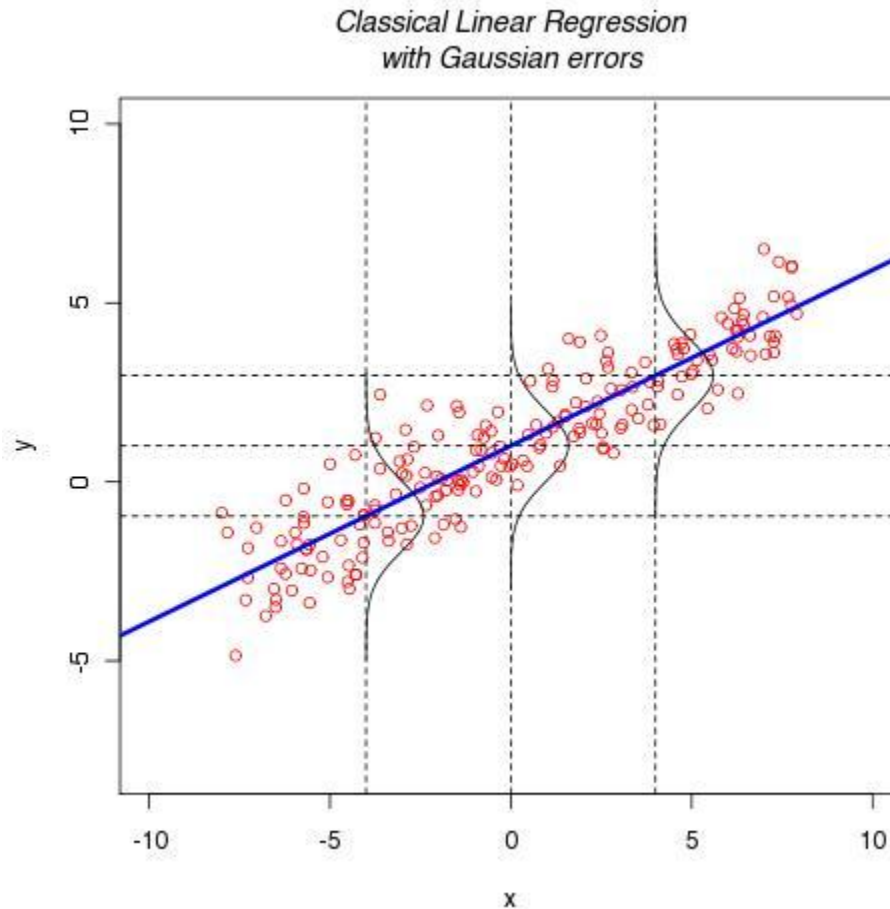
Proprietà degli stimatori dei minimi quadrati

1. $\widehat{\beta}_0$ e $\widehat{\beta}_1$ Sono stimatori corretti di β_0 e β_1
2. All'interno della classe degli stimatori lineari e corretti di β_0 e β_1 **gli stimatori dei minimi quadrati sono quelli a varianza minima** (Teorema di Gauss Markov)
3. Le varianze degli stimatori sono:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)$$

Ipotesi forti del modello lineare: errore Gaussiano



- ▶ B_0 e B_1 sono gli stimatori di massima verosimiglianza dei coefficienti di regressione.
- ▶ Denotando

$$S(B_0) = \sqrt{S^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right)}, S(B_1) = \sqrt{\frac{S^2}{\sum_i (x_i - \bar{x})^2}},$$

dove $S^2 = \sum_i (Y_i - B_0 - B_1 x_i)^2 / (n - 2)$, si ha

$$\frac{B_0 - \beta_0}{S(B_0)} \sim t_{n-2}, \quad \frac{B_1 - \beta_1}{S(B_1)} \sim t_{n-2}$$

Questi risultati vengono utilizzati per costruire intervalli di confidenza per i coefficienti di regressione e per testare ipotesi sui medesimi.

Ad esempio, se $t_{\alpha/2}$ rappresenta il valore della v.c. $T \sim t_{n-2}$ tale che $P(T < t_{\alpha/2}) = 1 - \alpha/2$, ovvero, $P(T < -t_{\alpha/2}) = \alpha/2$, allora gli IC al livello di confidenza del $100(1 - \alpha)\%$ per i due parametri sono rispettivamente:

$$B_0 \pm t_{\alpha/2} S(B_0), \quad B_1 \pm t_{\alpha/2} S(B_1).$$

Gli IC stimati sono

$$\hat{\beta}_0 \pm t_{\alpha/2} s(B_0), \quad \hat{\beta}_1 \pm t_{\alpha/2} s(B_1),$$

dove

$$s(B_0) = \sqrt{s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right)}, \quad s(B_1) = \sqrt{\frac{s^2}{\sum_i (x_i - \bar{x})^2}}.$$

Si supponga di voler sottoporre a verifica l'ipotesi: $H_0 : \beta_1 = b_1$ contro l'alternativa bidirezionale $H_1 : \beta_1 \neq b_1$. La regione critica del test è costituita dai valori t della statistica

$$T = \frac{B_1 - b_1}{S(B_1)} \sim t_{n-2}$$

tali che $|t| > t_{\alpha/2}$. Pertanto, si calcola sul campione osservato

$$t = \frac{\hat{\beta}_1 - b_1}{s(B_1)}$$

e si accetta H_0 se t è compreso tra $-t_{\alpha/2}$ e $t_{\alpha/2}$.

$$\hat{\beta} : N\left(\beta, \frac{\sigma_{\varepsilon}^2}{\sum x_i^2}\right)$$

$$\hat{\alpha} : N\left(\alpha, \sigma^2 \frac{\sum X_i^2}{N \sum x_i^2}\right)$$

In virtù del Teorema del Limite Centrale, anche se le y non fossero distribuite normalmente (sotto condizioni abbastanza generali) si avrebbe comunque una distribuzione asintoticamente normale per i suddetti parametri

ESERCIZIO

Con riferimento ad un campione di 40 aziende, si sono state osservate le seguenti statistiche campionarie relative alle vendite in migliaia di euro (Y) ed alle spese in marketing in migliaia di euro (X):

$$\sum_{i=1}^n y_i = 520 \quad \sum_{i=1}^n y_i^2 = 8654 \quad \sum_{i=1}^n x_i = 100.5 \quad \sum_{i=1}^n x_i^2 = 544.5 \quad \sum_{i=1}^n x_i y_i = 2020.8$$

Si calcolino le stime dei coefficienti β_0 e β_1 della regressione di Y su X



$$\bar{x} = \frac{\sum x_i}{n} = \frac{100.5}{40} = 2.5125$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{520}{40} = 13$$

$$\beta_1 = \frac{Cov(x, y)}{Var(x)} = \frac{\frac{\sum_{i=1}^n x_i y_i}{n} - \bar{x}\bar{y}}{\frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2} = \frac{\frac{2020.8}{40} - 2.5125 \times 13}{\frac{544.5}{40} - 2.5125^2}$$

$$\beta_1 = \frac{17.8575}{7.2998} = 2.4463$$

Si calcolino le stime dei coefficienti β_0 e β_1
della regressione di Y su X



$$\bar{x} = \frac{\sum x_i}{n} = \frac{100.5}{40} = 2.5125$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{520}{40} = 13$$

$$\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 13 - 2.4463 \times 2.5125 = 6.8537$$

$$\beta_1 = \frac{17.8575}{7.2998} = 2.4463$$

- Si calcoli il coefficiente di determinazione,

$$R^2 = \frac{Cov(x, y)^2}{Var(x)Var(y)} = \frac{17.8575^2}{7.2998 \times \left(\frac{\sum_{i=1}^n y_i^2}{n} - \bar{y}^2 \right)}$$

$$R^2 = \frac{17.8575^2}{7.2998 \times \left(\frac{8654}{40} - 13^2 \right)} = \frac{17.8575}{7.2998 \times 46.35} = 0.9226$$

Esercizio

Con riferimento alla seguente distribuzione di un collettivo di aziende secondo il fatturato e la superficie dello stabilimento:

- a) calcolare i parametri della retta di regressione del fatturato (Y) sulla superficie (X), commentare il risultato, rappresentare graficamente la nuvola di punti e la retta di regressione;
- b) calcolare un opportuno indice di bontà di adattamento del modello ai dati;
- c) secondo il modello, qual è il fatturato di un'azienda con superficie 3?
- d) calcolare la somma dei residui.

Esercizio

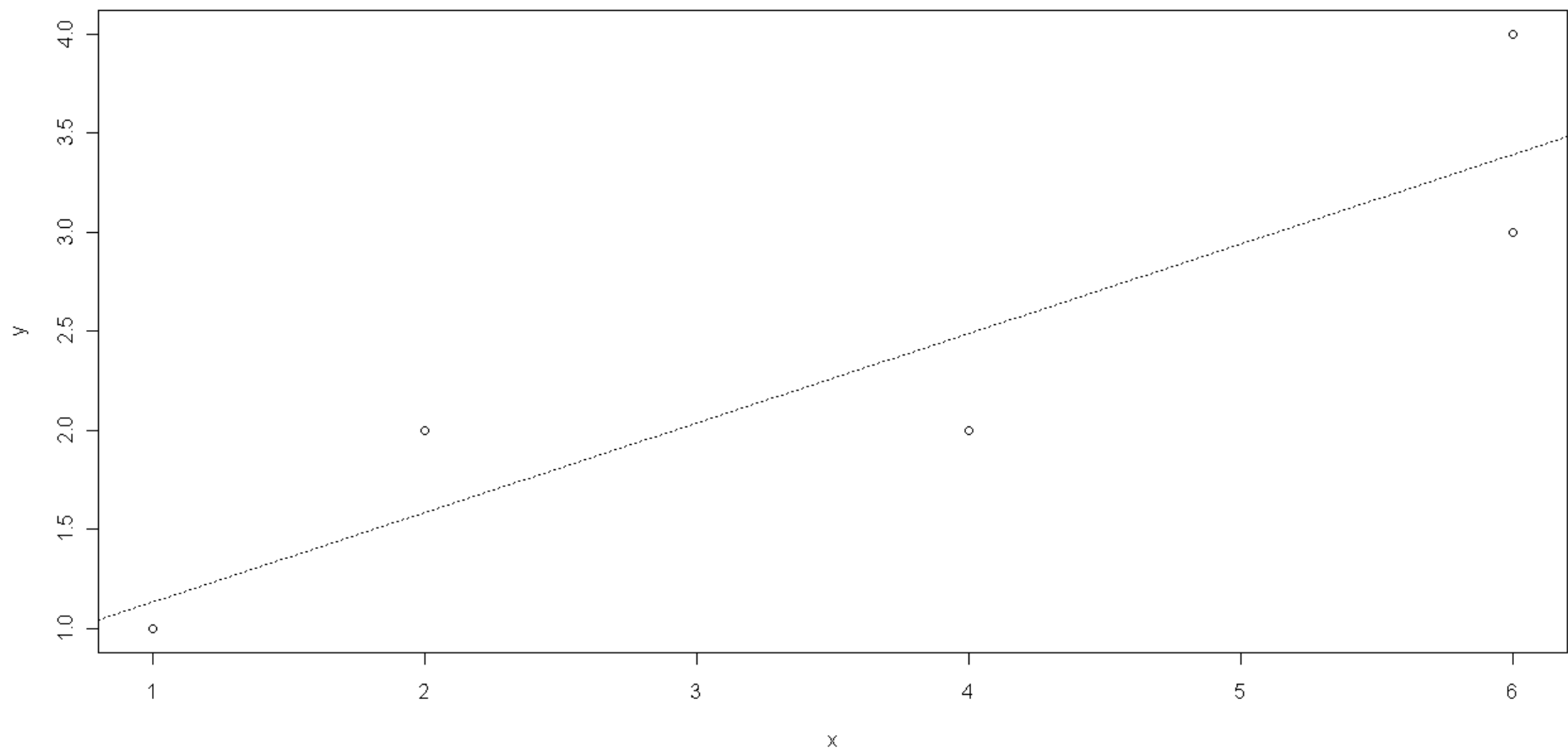
Fatturato (milioni di €)	Superficie (1000 mq)
3	6
2	4
4	6
2	2
1	1

	x	y	x²	xy
	6	3	36	18
	4	2	16	8
	6	4	36	24
	2	2	4	4
	1	1	1	1
somma	19	12	93	55
media	3,8	2,4	18,6	11

$$\beta_1 = \frac{Cov(x, y)}{Var(x)} = \frac{\frac{\sum_{i=1}^n x_i y_i}{n} - \bar{x}\bar{y}}{\frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2} = \frac{11 - 2.4 \times 3.8}{18.6 - 3.8^2}$$

$$\beta_1 = \frac{1.88}{4.16} = 0.45$$

$$\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 2.4 - 0.45 \times 3.8 = 0.69$$



	x	y	x²	xy	y²
	6	3	36	18	9
	4	2	16	8	4
	6	4	36	24	16
	2	2	4	4	4
	1	1	1	1	1
somma	19	12	93	55	34
media	3,8	2,4	18,6	11	6,8

$$R^2 = \frac{Cov(x, y)^2}{Var(x)Var(y)} = \frac{1.88^2}{4.16 \times (6.8 - 2.4^2)} = \frac{1.88^2}{4.16 \times 1.04} = 0.817$$

c) $\hat{y} = 0.69 + 0.45 \times 3 = 2.04$

d) somma dei residui sempre NULLA

c) $\hat{y} = 0.69 + 0.45 \times 3 = 2.04$

$$R^2 = \frac{Cov(x, y)^2}{Var(x)Var(y)} = \frac{1.88^2}{4.16 \times (6.8 - 2.4^2)} = \frac{1.88^2}{4.16 \times 1.04} = 0.817$$

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

$$\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - 0.817 = 0.183$$

$$\sum_{i=1}^n (\hat{y}_i - y_i)^2 = 0.183 \times \sum_{i=1}^n (y_i - \bar{y})^2 = 0.183 \times n \times \text{var}(y)$$

$$\sum_{i=1}^n (\hat{y}_i - y_i)^2 = 0.9516$$

X	Y	(X-mX)	(Y-mY)	(X-mX)(Y-mY)	(X-mX) ²	(Y-mY) ²	Y [^]	(Y-Y [^])	(Y-Y [^]) ²	(Y [^] -mY)	(Y [^] -mY) ²
6	3	2,2	0,6	1,32	4,84	0,36	3,39	0,39	0,16	0,99	0,99
4	2	0,2	-0,4	-0,08	0,04	0,16	2,49	0,49	0,24	0,09	0,01
6	4	2,2	1,6	3,52	4,84	2,56	3,39	-0,61	0,37	0,99	0,99
2	2	-1,8	-0,4	0,72	3,24	0,16	1,59	-0,41	0,17	-0,81	0,66
1	1	-2,8	-1,4	3,92	7,84	1,96	1,13	0,13	0,02	-1,27	1,60
				9,4	20,8	5,2		0,0	0,95		4,25

media 3,8 2,4
varianza 4,16 1,04

Esercizio 3

L'ufficio marketing della premiata Acciaieria JHK & Co. Sta raccogliendo informazioni sulle 20 imprese che costituiscono i suoi principali clienti. Nel seguito sono riportate alcune informazioni relative all'ammontare degli acquisti annui di ciascuna impresa cliente (indicato con Y ed espresso in migliaia di tonnellate) ed al numero di linee produttive possedute da ciascuna di queste (X):

$$\sum_{i=1}^{20} x_i = 46,6 \quad \sum_{i=1}^{20} y_i = 166,3 \quad \sum_{i=1}^{20} x_i^2 = 150,22 \quad \sum_{i=1}^{20} y_i^2 = 1449,31$$
$$\sum_{i=1}^{20} x_i y_i = 394,96$$

Nel suddetto ufficio si ritiene che vi sia una relazione di dipendenza lineare dell'ammontare di acciaio acquistato dalle aziende ed il numero di linee produttive da queste possedute. Supponendo vera tale ipotesi:

- a) Specificare il modello statistico di dipendenza dell'ammontare di acquisti dal numero di linee produttive.
- b) Determinare le stime dei coefficienti del modello mediante il metodo dei minimi quadrati.

Supponiamo l'errore sia Gaussiano l'ufficio vendite della suddetta acciaieria vuole verificare la plausibilità dell'ipotesi di dipendenza lineare dell'ammontare degli acquisti dal numero di linee produttive.

- a) Specificare il sistema di ipotesi da sottoporre a verifica
- b) Individuare un'opportuna statistica test ed indicare la sua distribuzione di probabilità.
- c) In base ai dati osservati, l'ipotesi dell'ufficio vendite è accettabile ad un livello di significatività di $\alpha=0,05$?

Svolgimento

- a) Un modello statistico adatto a descrivere una relazione di dipendenza lineare tra due variabili è il modello di regressione lineare semplice:

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

- b) Le stime dei coefficienti del modello mediante il metodo dei minimi quadrati si ottengono (utilizzando le formule calcolatorie) come:

$$\begin{aligned}\hat{\beta}_1 &= \frac{Cov(x, y)}{V(x)} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{394,96 - 20 \cdot 46,6 / 20 \cdot 166,3 / 20}{150,22 - 20 \cdot (46,6 / 20)^2} = \\ &= \frac{394,96 - 387,48}{150,22 - 108,58} = \frac{7,48}{41,64} = 0,18\end{aligned}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{166,3}{20} - 0,18 \frac{46,6}{20} = 7,896$$

c) Il coefficiente angolare è positivo: questo significa che tra le due variabili c'è una relazione di dipendenza diretta, anche se debole.

d) $y = 7,896 + 0,18 \cdot x$

Il coefficiente β_1 ci dice esattamente di quanto varia l'ammontare degli acquisti annui di un'impresa cliente a fronte di una variazione unitaria nel numero di linee produttive. Se $\beta_1=0,18$, significa che, per ogni variazione unitaria di x , y aumenta o diminuisce di 0,18. Se la variazione di x è pari a -4, allora y varia di $0,18 \cdot (-4) = -0,72$ migliaia di tonnellate: a fronte di una diminuzione di 4 unità delle linee produttive, l'ammontare degli acquisti annui diminuisce di 0,72 migliaia di tonnellate.

- a) Il sistema di ipotesi per un test sui parametri del modello di regressione lineare semplice è il seguente:

$$\begin{cases} H_0 : \beta_1 = 0 & \text{ipotesi di indipendenza lineare} \\ H_1 : \beta_1 \neq 0 & \text{ipotesi di dipendenza lineare} \end{cases}$$

- b) In mancanza di informazioni sulla varianza del termine di errore, la statistica più appropriata è:

$$T = \frac{B_1 - 0}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2}$$

c) Calcoliamo in primo luogo:

$$\hat{\sigma}_{\varepsilon}^2 = \frac{Dev_{disp}(Y)}{(n-2)} = \frac{65,19}{18} = 3,62$$

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 = 150,22 - 20 \cdot \left(\frac{46,6}{20}\right)^2 = \\ &= 150,22 - 20 \cdot (2,33)^2 = 150,22 - 108,578 = 41,642\end{aligned}$$

E poi:

$$t_{oss} = \frac{0,18}{\sqrt{\frac{3,62}{41,642}}} = \frac{0,18}{0,29} = 0,62$$

Infine identifichiamo la regione di rifiuto. Si tratta di un test bilaterale, per cui cerchiamo:

$$\pm t_{n-2; \alpha/2} = \pm t_{18; 0,025} = \pm 2,101$$

$-t_{18; \alpha/2} < t_{oss} < t_{18; \alpha/2}$ e dunque è compreso nella regione di non rifiuto dell'ipotesi nulla. Concludiamo quindi che non possiamo rifiutare, ad un livello di significatività del 5%, l'ipotesi nulla di indipendenza lineare.

```
. correlate area selling_price
(obs=28)
```

	area sellin~e	
area	1.0000	
selling_pr~e	0.9251	1.0000

```
. regress selling_price area
```

Source	SS	df	MS
Model	4630.76207	1	4630.76207
Residual	780.426849	26	30.0164173
Total	5411.18892	27	200.414404

```
Number of obs =      28
F( 1,      26) =   154.27
Prob > F       =    0.0000
R-squared      =    0.8558
Adj R-squared  =    0.8502
Root MSE      =    5.4787
```

selling_pr~e	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
area	31.25273	2.516177	12.42	0.000	26.08066	36.42481
_cons	-1.466856	3.353966	-0.44	0.665	-8.361033	5.427321