Regressione

1

Il modello di regressione

- Il modello di regressione descrive la dipendenza di una variabile quantitativa (variabile dipendente) da una o più variabile quantitative (variabili esplicative).
- Il modello ha diversi scopi: descrittivo, interpretativo e previsivo.
- Si distingue:
- Modello di regressione semplice: 1 variabile esplicativa
- Modello di regressione multipla: più variabili esplicative

2

Modello di regressione semplice

- Nel modello di regressione semplice, si indica
  Y = variabile dipendente (o risposta)
  X = variabile esplicativa (o indipendente)

- Ad esempio, si assume un modello di regressione semplice con
  Y = fatturato
  X = spese in mktg

3

Modello di regressione semplice

- Il modello di regressione semplice è, in generale,

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

- Y è la variabile dipendente
- $f(X)$ è il contributo della variabile X al valore di Y
- ϵ è il contributo di altri fattori non considerati

4

Modello di regressione lineare semplice

- Il modello di regressione lineare semplice è

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

- Dunque $f(X) = \beta_0 + \beta_1 X$
- $f(X)$ è una funzione lineare

5

- Quindi

$$Y \neq \beta_0 + \beta_1 X$$

perché

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

6

- Se fosse

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

si avrebbe una relazione deterministica tra X e Y.

- Relazione deterministica tra X e Y: a specifici valori delle X è associato un solo valore di Y.
- Es: $Y = 10 + 3X$
- Se $X = 2$, $Y = 10 + 6 = 16$
- Se $X = 18$, $Y = 10 + 24 = 34$
- …

7

L'errore ϵ

- ϵ è il cosiddetto errore che evita che tra X e Y ci sia una relazione deterministica.
- In pratica l'errore ϵ include altre variabili non incluse nel modello ed eventi imprevedibili.
- Il valore medio (atteso) dell'errore ϵ è 0,

$$E(\epsilon) = 0$$

- Ne consegue che il valore medio (atteso) di Y (che ci si aspetta di osservare senza considerare l'effetto di altre variabili e in assenza di eventi «particolari») è

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

- La relazione tra $E(Y)$ e X è lineare, ovvero rappresentabile da una retta, la retta di regressione.

8

I parametri del modello

- β_0 e β_1 sono i parametri (o coefficienti) del modello.
- Geometricamente β_0 e β_1 sono i coefficienti della retta di regressione.

9

Il parametro β_0

- Il parametro β_0 è l'intercetta della retta di regressione.
- Interpretazione: è il valore atteso di Y se X = 0.

- Esempio
- $Y = 10 + 3X + \epsilon$
- $\mathrm{E}(Y) = 10 + 3X$
- Se $X = 0$, $\mathrm{E}(Y) = 10$

10

Il parametro β_1

- Il parametro β_1 è la pendenza della retta di regressione.
- Interpretazione: è la variazione attesa di Y quando X aumenta di un'unità.

- Esempio
- $Y = 10 + 3X + \epsilon$
- Se $X = 4,\ \mathrm{E}(Y) = 10 + 12 = 22$
- Se $X = 5,\ \mathrm{E}(Y) = 10 + 15 = 25$
- La differenza è $25 - 22 = 3$.

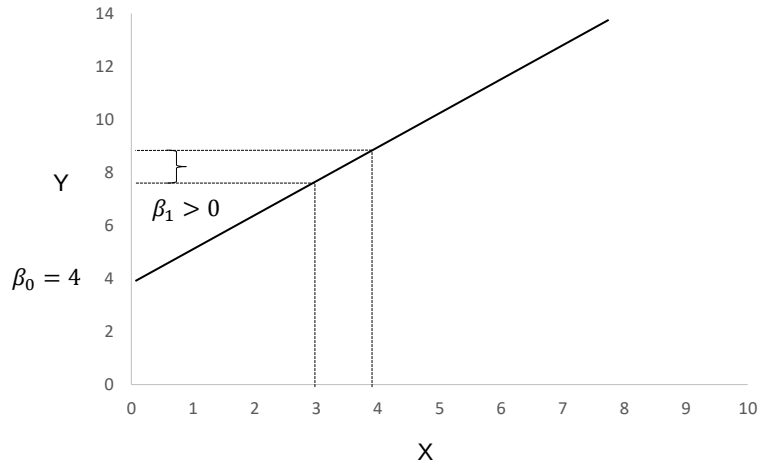
11

Il parametro β_1

- Se $\beta_1 > 0$, c'è concordanza (relazione diretta) tra X e Y.
- Es: $Y = 10 + 3X + \epsilon$
- Quando X aumenta di 1 unità, il valore atteso di Y varia di 3 (aumenta di 3).
- Se $\beta_1 < 0$, c'è discordanza (relazione inversa) tra X e Y.
- Es: $Y = 10 - 2X + \epsilon$
- Quando X aumenta di 1 unità, il valore atteso di Y varia di -2 (diminuisce di 2).

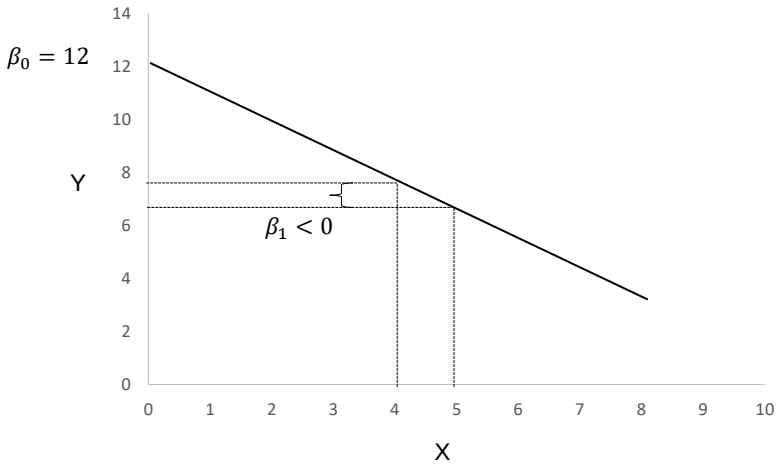
12

Figura 1



13

Figura 2



14

I dati e le assunzioni

- Avendo a disposizione *n* osservazioni, cioè *n* coppie di dati (x_i, y_i), le assunzioni del modello di regressione lineare semplice sono:

1. $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ per ogni osservazione $i = 1, \dots, n$.
2. Il valore medio di ε_i è 0.
3. I valori x_i sono noti senza errore.

- Ne consegue che

$$E(Y_i) = \beta_0 + \beta_1 x_i$$

per ogni osservazione $i = 1, \dots, n$.

15

Stima dei parametri

- In pratica, β_0 e β_1 non sono noti. Essi sono **stimati** a partire da un insieme di *n* osservazioni.

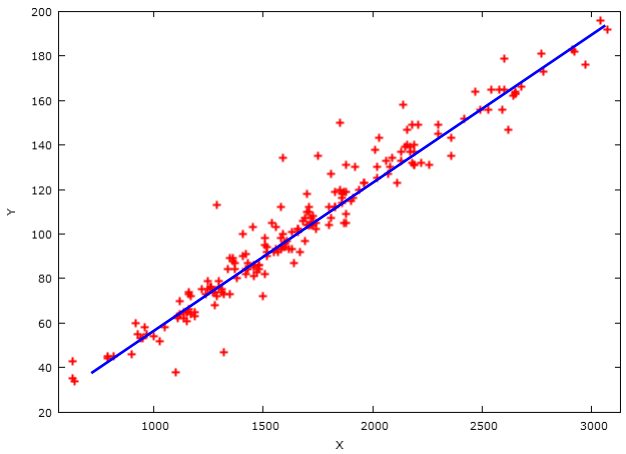
16

Esempio (n = 210)
REDDITO MENSILE (X),
SPESA PER MUSEI E MOSTRE (Y)

Reddito (X)	*Spesa (Y)*
1340	84
820	45
2360	135
1700	118
1210	68
1590	100
2130	133
2490	156
…	…
…	…

17

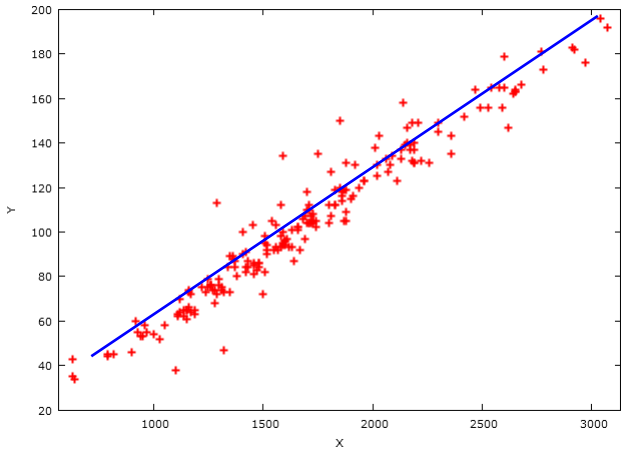
Quale retta?



18

18

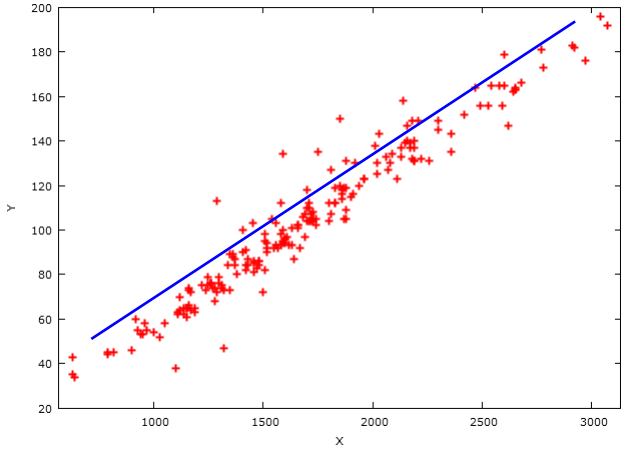
Quale retta?



19

19

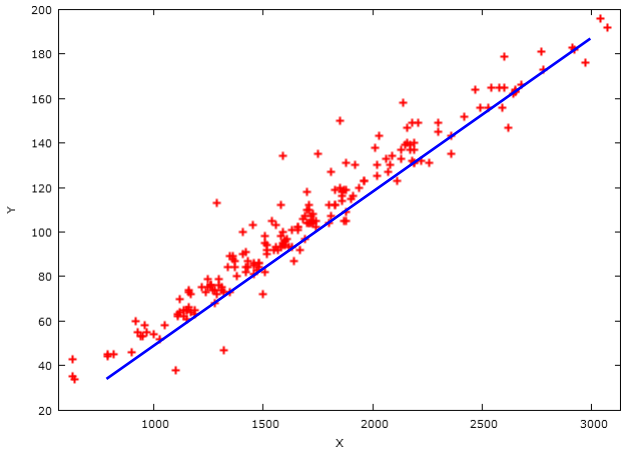
Quale retta?



20

20

Quale retta?



21

21

Il metodo dei minimi quadrati

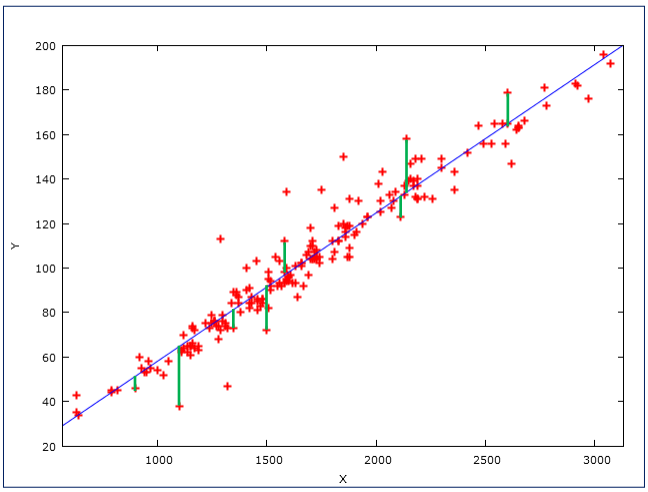
- Il metodo di stima utilizzato è il metodo dei **MINIMI QUADRATI** che assicura che la retta sia il più vicino possibile ai punti delle osservazioni.
- Ratio: individuare quei valori di β_0 e β_1 che minimizzano la somma dei quadrati delle differenze tra dati osservati e dati teorici,

$$\min\left[\left(Y_1-\hat{Y}_1\right)^2+\left(Y_2-\hat{Y}_2\right)^2+\cdots+\left(Y_n-\hat{Y}_n\right)^2\right]$$

- I dati teorici sono quelli previsti dalla retta di regressione.
- Dunque

$$\min[(Y_1-\beta_0-\beta_1x_1)^2+\cdots+(Y_n-\beta_0-\beta_1x_n)^2]$$

22



23

23

Formula di β_1

- La stima del parametro β_1 è data da

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

dove

$\bar{x}$ è la media della variabile X

$\bar{y}$ è la media della variabile Y

- Dividendo per n, si ottiene

$$\hat{\beta}_1 = \frac{\sigma_{XY}}{\sigma_X^2}$$

24

Formula di β_0

- La stima del parametro β_0 è data da

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

25

Retta di regressione stimata

- Una volte ottenute le stime dei parametri

  $\hat{\beta}_0 = -8{,}719$

  $\hat{\beta}_1 = 0{,}067$

  si definisce l'equazione della retta di regressione stimata dai dati

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

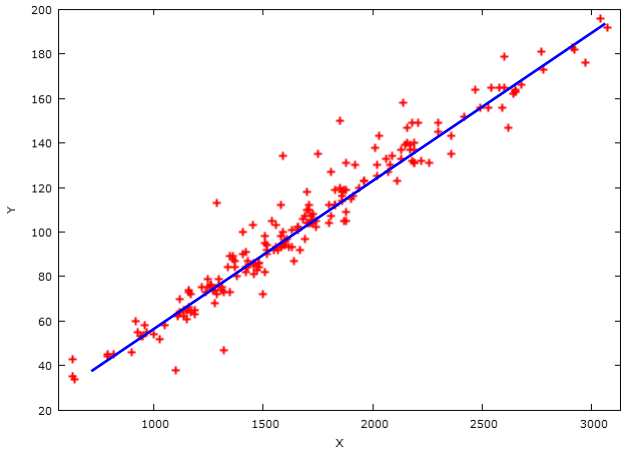
ovvero

$$\hat{Y} = -8{,}719 + 0{,}067X$$

26

26

Retta di regressione



27

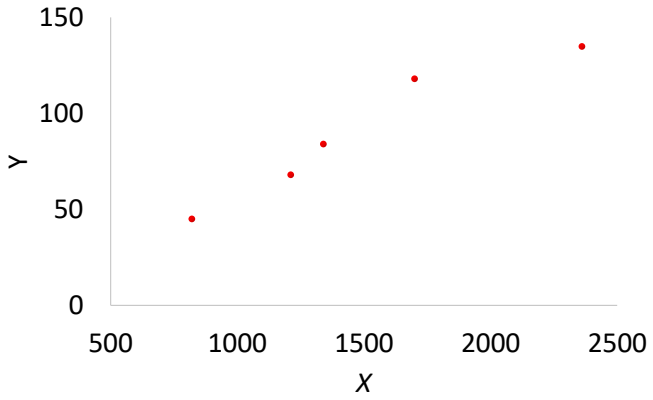
27

Esempio (n = 5)
REDDITO MENSILE (X),
SPESA PER MUSEI E MOSTRE (Y)

Reddito (X)	*Spesa (Y)*
1340	**84**
820	**45**
2360	**135**
1700	**118**
1210	**68**

28

Diagramma di dispersione



29

29

Esempio

X	Y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1340	84	-146	-6	876
820	45	-666	-45	29970
2360	135	874	45	39330
1700	118	214	28	5992
1210	68	-276	-22	6072
TOT				82240

$\bar{x} = 1486$

$\bar{y} = 90$

30

Esempio

X	Y	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1340	84	-146	21316
820	45	-666	443556
2360	135	874	763876
1700	118	214	45796
1210	68	-276	76176
TOT			1350720

31

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

$$\hat{\beta}_1 = \frac{82240}{1350720} = 0{,}061$$

32

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

$$\hat{\beta}_0 = 90 - 0{,}061 \cdot 1486 = -0{,}477$$

33

La retta di regressione

- L'equazione della retta di regressione è

$$\hat{Y} = -0{,}477 + 0{,}061X$$

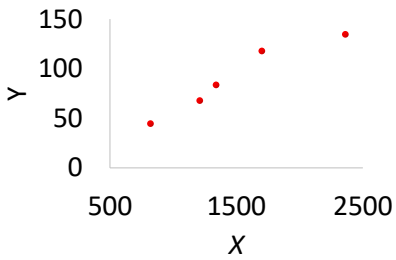
34

Calcolo dei valori teorici: $\hat{Y} = -0{,}477 + 0{,}061X$

X	Y	$\hat{y}_i$
1340	84	81,111
820	45	49,450
2360	135	143,214
1700	118	103,030
1210	68	73,195
TOT		

35

Rappresentazione della retta



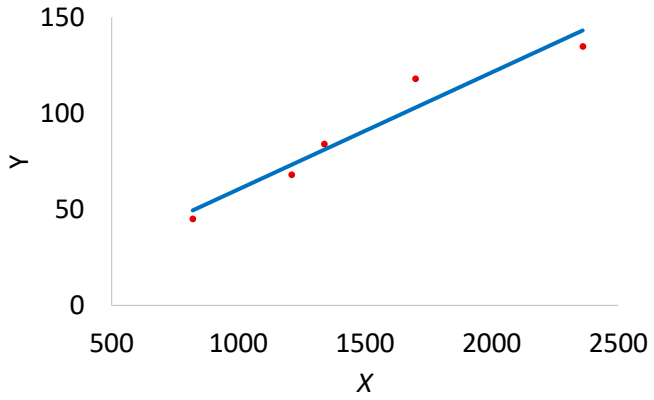
Si individuano 2 punti (sufficienti per far passare una retta).

- Si può scegliere $\mathrm{X} = \min(X)$ e $\mathrm{X} = \max(\mathrm{X})$
- Con $X = 820$ si ha $\hat{Y} = 49{,}450$
- Con $X = 2360$ si ha $\hat{Y} = 143{,}214$

36

36

Retta di regressione



37

37

I residui

- I residui sono definiti come la differenza tra i valori osservati e i valori teorici

$$e_i = y_i - \hat{y}_i$$

- Per costruzione la somma dei residui è pari a 0.

38

38

Calcolo dei residui: $e = Y - \hat{Y}$

X	*Y*	$\hat{y}_i$	e_i
1340	84	81,111	2,889
820	45	49,450	-4,450
2360	135	143,214	-8,214
1700	118	103,030	14,970
1210	68	73,195	-5,195
TOT			0

39

Proprietà della retta di regressione

1. La retta di regressione passa sempre per il punto $(\bar{x}, \bar{y})$

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

2. La somma dei residui è nulla

$$\sum_{i=1}^{n} e_i = 0$$

40

Segno di ρ_{XY} e segno di $\widehat{\beta}_1$

- Se $\rho_{XY} > 0$, la retta di regressione ha una pendenza positiva ($\hat{\beta}_1 > 0$)

- Se $\rho_{XY} < 0$, la retta di regressione ha una pendenza negativa ($\hat{\beta}_1 > 0$)

41

Bontà di adattamento di una retta di regressione

- La bontà di adattamento (o goodness-of-fit) di un modello di regressione semplice è misurata dall'indice di determinazione R^2_{XY}.
- L'indice R^2_{XY} assume valori compresi tra 0 e 1,

$$0 \leq R^2_{XY} \leq 1$$

- L'indice R^2_{XY} si basa sulla distanza tra i valori osservati (i punti del diagramma di dispersione) e la retta di regressione.
- Più i valori osservati sono vicini alla retta, migliore è l'adattamento, maggiore è R^2_{XY}.

42

42

Scomposizione della devianza di Y

- Si definisce devianza (SQT) di Y, il numeratore della varianza di Y:

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

- SQT = somma dei quadrati dei totali
- Si può dimostrare che

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

$$SQT \quad = \quad SQR \quad + \quad SQE$$

- SQR = somma dei quadrati della regressione
- SQE = somma dei quadrati degli errori

43

43

Formula

- L'indice R^2_{XY} è dato da

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

- ovvero

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

- Poiché $SQT = SQR + SQE$

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

44

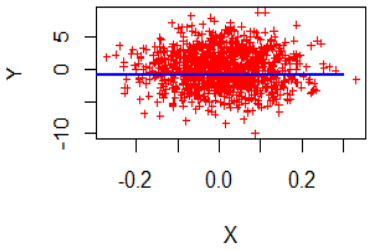
44

Interpretazione

$R^2_{XY} = 0$

Non esiste associazione.

I punti del diagramma di dispersione sono intorno ad una retta con pendenza nulla.

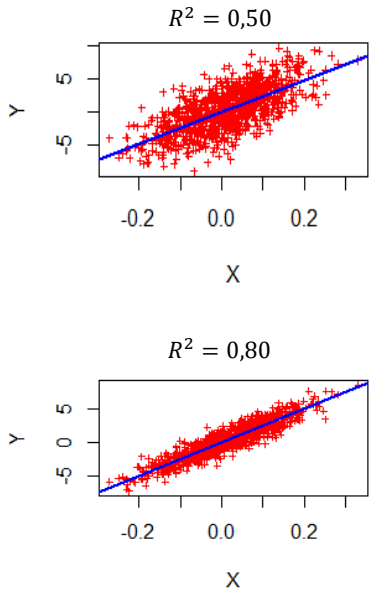


45

$0 \leq R^2_{XY} \leq 1$

Esiste concordanza (associazione positiva).

I punti del diagramma di dispersione sono intorno ad una retta con pendenza POSITIVA.

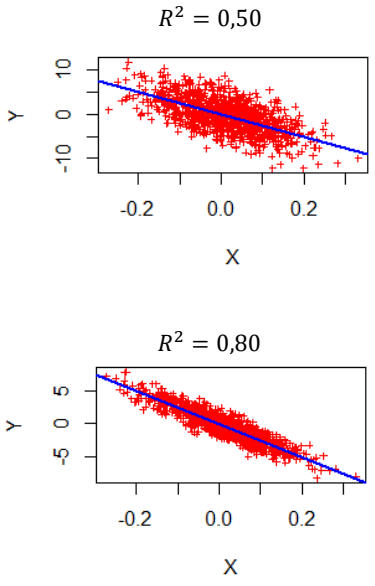


46

$0 \leq R^2_{XY} \leq 1$

Esiste discordanza (associazione negativa).

I punti del diagramma di dispersione sono intorno ad una retta con pendenza NEGATIVA.

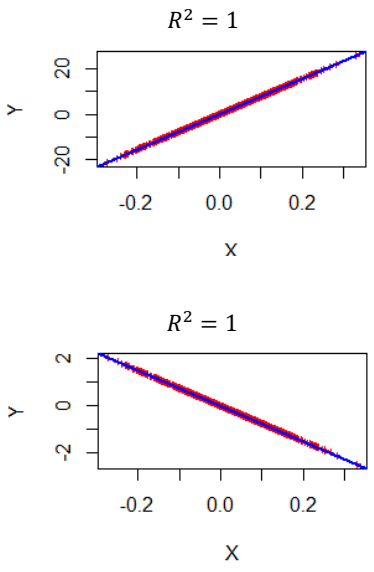


47

$R^2_{XY} = 1$

Esiste concordanza o discordanza perfetta.

I punti del diagramma di dispersione sono su una retta con pendenza NON nulla.



48

Interpretazione dell'indice R^2_{XY}

- L'indice moltiplicato per 100 esprime la percentuale di variabilità totale della Y spiegata dalla variabile X (è dunque una misura dell'«importanza» della variabile esplicativa).

49

49

Formula alternativa

- È facile dimostrare che

$$R^2_{XY} = \rho^2_{XY}$$

50

50

Esempio di calcolo di R^2_{XY}

X	Y	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1340	84	-6	36
820	45	-45	2025
2360	135	45	2025
1700	118	28	784
1210	68	-22	484
TOT			5354

51

Esempio di calcolo di R^2_{XY}

X	Y	e_i	e_i^2
1340	84	2,889	8,346
820	45	-4,450	19,802
2360	135	-8,214	67,470
1700	118	14,970	224,101
1210	68	-5,195	26,988
TOT		0	346,707

52

Esempio di calcolo di R^2_{XY}

$$R^2_{XY} = 1 - \frac{SQE}{SQT} = 1 - \frac{346{,}707}{5354} = 0{,}935$$

- La bontà di adattamento è elevata.
- La percentuale della variabilità della variabile *Spesa per musei e mostre* spiegata dalla variabile *Reddito mensile* è pari al 93,5%.

53

53

Previsioni

- La previsione della variabile Y in corrispondenza di un valore di X è ottenuta dall'equazione della retta di regressione.
- Definito x_f il valore per il quale si intende fare la previsione, si ha

$$\hat{Y}(x_f) = \hat{\beta}_0 + \hat{\beta}_1 x_f$$

54

54

Esempio di previsione

- La previsione in corrispondenza di $X = 1000$ è data da

$$\hat{Y}(1000) = -0{,}477 + 0{,}061 \cdot 1000 = 60{,}523$$

55

55