

Corso di

Statistica

La regressione

Lezione otto

Cristina Davino

cdavino@unimc.it

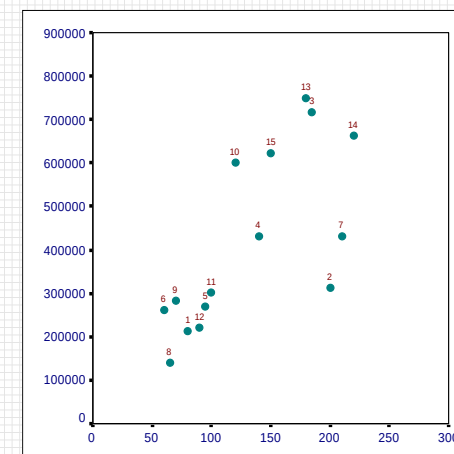


La Regressione

App.	mq (X)	Prezzo in € (Y)
1	80	212000
2	200	313000
3	185	717000
4	140	431000
5	95	270000
6	60	261000
7	210	431000
8	65	140000
9	70	282000
10	120	600000
11	100	303000
12	90	220000
13	180	749000
14	220	663000
15	150	623000
1.965		6.215.000

X → Variabile indipendente o esplicativa

Y → Variabile dipendente



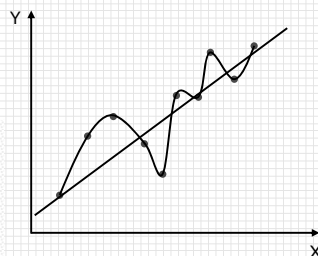
La Regressione

X → Variabile indipendente

Y → Variabile dipendente

Dipendenza funzionale (o deterministica): $Y = f(X; \theta)$

- Da un punto di vista analitico, i valori della **Y** possono essere determinati senza errore a partire dai soli valori della **X**;
- Da un punto di vista grafico, la dipendenza funzionale implica la definizione di una funzione che passi per tutti i punti, e che quindi richiede la determinazione di tanti parametri quanti sono i punti.



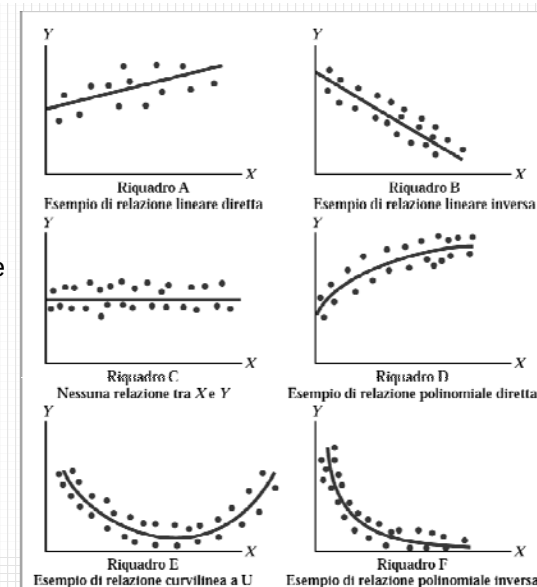
Dipendenza statistica: $Y = f(X; \theta) + e$

- Il valore della variabile dipendente non è univocamente determinato a partire dal solo valore della variabile esplicativa, potendosi osservare, per ciascun di **X**, più valori di **Y**;
- Da un punto di vista grafico, la dipendenza statistica implica una funzione che passi fra i punti osservati. Il numero di parametri da determinare dipende, in questo caso, dal tipo di funzione scelta e non dal numero di punti osservati.

La Regressione

X → Variabile indipendente

Y → Variabile dipendente



La Regressione $Y = f(X; \theta) + e$

$X \rightarrow$ Variabile indipendente

$Y \rightarrow$ Variabile dipendente

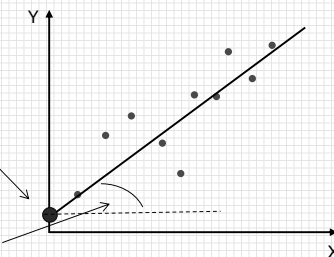
La scelta del tipo di funzione: $\rightarrow$ **Lineare**

Equazione di una retta $\rightarrow y = b_0 + b_1 x$

b_0 è l'**intercetta** sull'asse delle ordinate. Può essere interpretato come il valore di Y per $X=0$ (quando ciò ha senso).

b_1 è il **coefficiente angolare** della retta di regressione in quanto funzione dell'angolo che la retta forma con l'asse delle ascisse. Esprime dunque la **pendenza** (positiva, negativa o nulla) della retta.

Esprime anche quanto varia la variabile Y al variare unitario della variabile X.



La Regressione $Y = f(X; \theta) + e$

$X \rightarrow$ Variabile indipendente

$Y \rightarrow$ Variabile dipendente

La scelta del tipo di funzione: $\rightarrow$ **Lineare**

La determinazione dei vincoli:

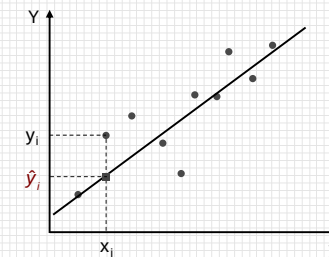
$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \min$$

$$\hat{y}_i = b_0 + b_1 x_i$$

$$\sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 = \min$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\text{Cov}(XY)}{\text{Var}(X)}$$



La Regressione $Y = f(X; \theta) + e$

$X \rightarrow$ Variabile indipendente

$Y \rightarrow$ Variabile dipendente

La scelta del tipo di funzione: $\rightarrow$ **Lineare**

$$b_0 = \bar{y} - b_1 \bar{x}$$



$$\bar{y} = b_0 + b_1 \bar{x}$$

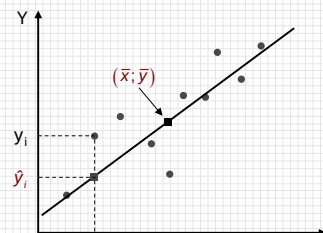
E' l'**intercetta** sull'asse delle ordinate. Può essere interpretato come il valore di Y per $X=0$ (quando ciò ha senso).

Il punto di coordinate $(\bar{x}; \bar{y})$ è un punto della retta di regressione. La retta di regressione passa, dunque, sempre per il baricentro della nube.

E' il **coefficiente angolare** della retta di regressione in quanto funzione dell'angolo che la retta forma con l'asse delle ascisse. Esprime dunque la **pendenza** (positiva, negativa o nulla) della retta.

$$b_1 = \frac{\text{Cov}(XY)}{\text{Var}(X)}$$

Esprime anche quanto varia la variabile Y al variare unitario della variabile X.



La Regressione $Y = f(X; \theta) + e$

$X \rightarrow$ Variabile indipendente

$Y \rightarrow$ Variabile dipendente

La scelta del tipo di funzione: $\rightarrow$ **Lineare**

$$b_0 = \bar{y} - b_1 \bar{x}$$



$$\bar{y} = b_0 + b_1 \bar{x}$$

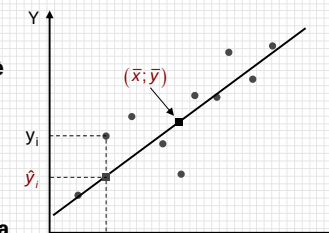
E' l'**intercetta** sull'asse delle ordinate. Può essere interpretato come il valore di Y per $X=0$ (quando ciò ha senso).

Il punto di coordinate $(\bar{x}; \bar{y})$ è un punto della retta di regressione. La retta di regressione passa, dunque, sempre per il baricentro della nube.

E' il **coefficiente angolare** della retta di regressione in quanto funzione dell'angolo che la retta forma con l'asse delle ascisse. Esprime dunque la **pendenza** (positiva, negativa o nulla) della retta.

$$b_1 = \frac{\text{Cov}(XY)}{\text{Var}(X)}$$

Esprime anche quanto varia la variabile Y al variare unitario della variabile X.

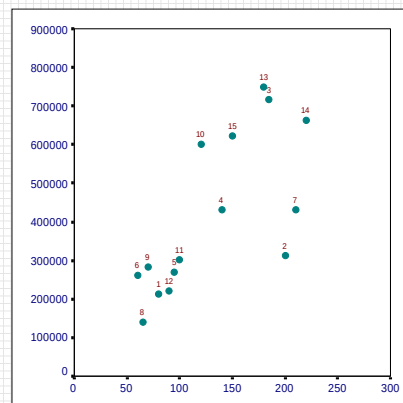


La Regressione $Y = f(X; \theta) + e$

X → Variabile indipendente

Y → Variabile dipendente

App.	mq (X)	Prezzo in € (Y)
1	80	212000
2	200	313000
3	185	717000
4	140	431000
5	95	270000
6	60	261000
7	210	431000
8	65	140000
9	70	282000
10	120	600000
11	100	303000
12	90	220000
13	180	749000
14	220	663000
15	150	623000
1.965		6.215.000



Media_X = 131,0 Media_Y = 414.333,3
 Dev_X = 54,44 Dev_Y = 197.060,96

La Regressione $Y = f(X; \theta) + e$

X → Variabile indipendente

Y → Variabile dipendente

App.	mq (X)	Prezzo in € (Y)	X-M(X)	Y-M(Y)	[X-M(X)]*[Y-M(Y)]
1	80	212000	-51,0	-202333,3	10.319.000
2	200	313000	69,0	-101333,3	-6.992.000
3	185	717000	54,0	302666,7	16.344.000
4	140	431000	9,0	16666,7	150.000
5	95	270000	-36,0	-144333,3	5.196.000
6	60	261000	-71,0	-153333,3	10.886.667
7	210	431000	79,0	16666,7	1.316.667
8	65	140000	-66,0	-274333,3	18.106.000
9	70	282000	-61,0	-132333,3	8.072.333
10	120	600000	-11,0	185666,7	-2.042.333
11	100	303000	-31,0	-111333,3	3.451.333
12	90	220000	-41,0	-194333,3	7.967.667
13	180	749000	49,0	334666,7	16.398.667
14	220	663000	89,0	248666,7	22.131.333
15	150	623000	19,0	208666,7	3.964.667
1.965		6.215.000	0,0	0,0	115.270.000

Media_X = 131,0 Media_Y = 414.333,3

Dev_X = 54,44 Dev_Y = 197.060,96

$$Cov(XY) = \frac{\sum (x_i - \mu_x) \cdot (y_i - \mu_y)}{n} = \frac{115.270.000}{15} = 7.684.666,7$$

$$b_1 = \frac{Cov(XY)}{Var(X)} = \frac{7.684.666,7}{2964,00} = 2.592,67$$

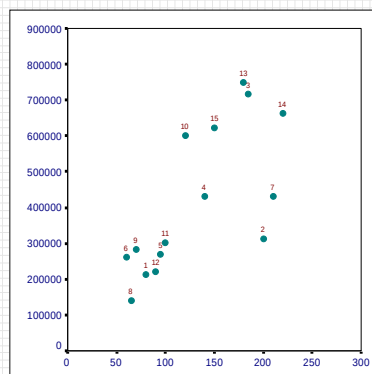
$$b_0 = \bar{y} - b_1 \bar{x} = 414.333,3 - 2.592,67 \times 131 = 74.693,5$$

La Regressione $Y = f(X; \theta) + e$

$$\hat{y} = b_0 + b_1 x$$

$$\hat{y} = 74.693,88 + 2.592,67x$$

App.	mq (X)	Prezzo in € (Y)	$\hat{Y}$	$Y - \hat{Y}$
1	80	212.000	282.107	-70.107
2	200	313.000	593.227	-280.227
3	185	717.000	554.337	162.663
4	140	431.000	437.667	-6.667
5	95	270.000	320.997	-50.997
6	60	261.000	230.254	30.746
7	210	431.000	619.154	-188.154
8	65	140.000	243.217	-103.217
9	70	282.000	256.181	25.819
10	120	600.000	385.814	214.186
11	100	303.000	333.961	-30.961
12	90	220.000	308.034	-88.034
13	180	749.000	541.374	207.626
14	220	663.000	645.081	17.919
15	150	623.000	463.594	159.406
1.965		6.215.000	6.215.000	0



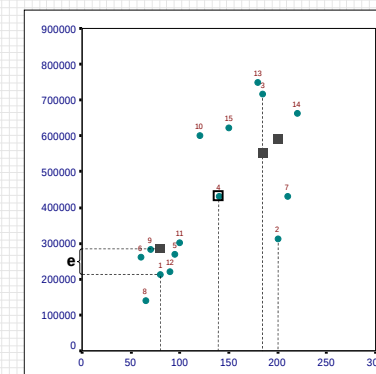
Media_X = 131,0 Media_Y = 414.333,3
 Dev_X = 54,44 Dev_Y = 197.060,96

La Regressione $Y = f(X; \theta) + e$

$$\hat{y} = b_0 + b_1 x$$

$$\hat{y} = 74.693,88 + 2.592,67x$$

App.	mq (X)	Prezzo in € (Y)	$\hat{Y}$	$Y - \hat{Y}$
1	80	212.000	282.107	-70.107
2	200	313.000	593.227	-280.227
3	185	717.000	554.337	162.663
4	140	431.000	437.667	-6.667
5	95	270.000	320.997	-50.997
6	60	261.000	230.254	30.746
7	210	431.000	619.154	-188.154
8	65	140.000	243.217	-103.217
9	70	282.000	256.181	25.819
10	120	600.000	385.814	214.186
11	100	303.000	333.961	-30.961
12	90	220.000	308.034	-88.034
13	180	749.000	541.374	207.626
14	220	663.000	645.081	17.919
15	150	623.000	463.594	159.406
1.965		6.215.000	6.215.000	0



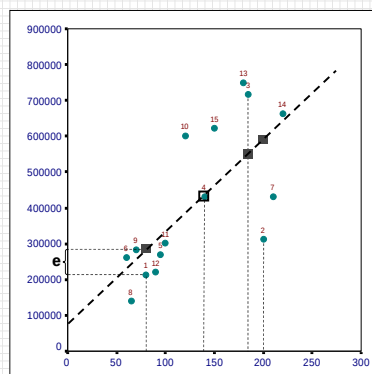
Media_X = 131,0 Media_Y = 414.333,3
 Dev_X = 54,44 Dev_Y = 197.060,96

La Regressione $Y = f(X; \theta) + e$

App.	mq (X)	Prezzo in € (Y)	$\hat{Y}$	$Y - \hat{Y}$
1	80	212.000	282.107	-70.107
2	200	313.000	593.227	-280.227
3	185	717.000	554.337	162.663
4	140	431.000	437.667	-6.667
5	95	270.000	320.997	-50.997
6	60	261.000	230.254	30.746
7	210	431.000	619.154	-188.154
8	65	140.000	243.217	-103.217
9	70	282.000	256.181	25.819
10	120	600.000	385.814	214.186
11	100	303.000	333.961	-30.961
12	90	220.000	308.034	-88.034
13	180	749.000	541.374	207.626
14	220	663.000	645.081	17.919
15	150	623.000	463.594	159.406
1.965		6.215.000	6.215.000	0

$$\hat{Y} = b_0 + b_1 X$$

$$\hat{Y} = 74.693,88 + 2.592,67x$$



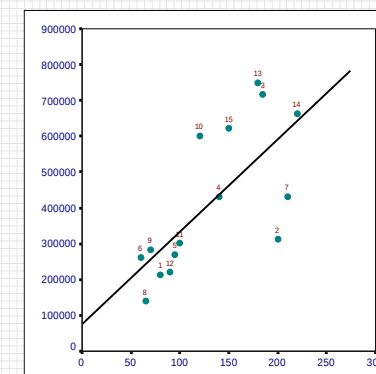
Media_X = 131,0 Media_Y = 414.333,3
 Dev_X = 54,44 Dev_Y = 197.060,96

La Regressione $Y = f(X; \theta) + e$

App.	mq (X)	Prezzo in € (Y)	$\hat{Y}$	$Y - \hat{Y}$
1	80	212.000	282.107	-70.107
2	200	313.000	593.227	-280.227
3	185	717.000	554.337	162.663
4	140	431.000	437.667	-6.667
5	95	270.000	320.997	-50.997
6	60	261.000	230.254	30.746
7	210	431.000	619.154	-188.154
8	65	140.000	243.217	-103.217
9	70	282.000	256.181	25.819
10	120	600.000	385.814	214.186
11	100	303.000	333.961	-30.961
12	90	220.000	308.034	-88.034
13	180	749.000	541.374	207.626
14	220	663.000	645.081	17.919
15	150	623.000	463.594	159.406
1.965		6.215.000	6.215.000	0

$$\hat{Y} = b_0 + b_1 X$$

$$\hat{Y} = 74.693,88 + 2.592,67x$$



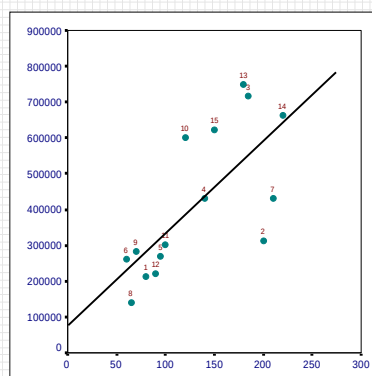
Media_X = 131,0 Media_Y = 414.333,3
 Dev_X = 54,44 Dev_Y = 197.060,96

La Regressione: Interpolazione ed estrapolazione

App.	mq (X)	Prezzo in € (Y)	$\hat{Y}$	$Y - \hat{Y}$
1	80	212.000	282.107	-70.107
2	200	313.000	593.227	-280.227
3	185	717.000	554.337	162.663
4	140	431.000	437.667	-6.667
5	95	270.000	320.997	-50.997
6	60	261.000	230.254	30.746
7	210	431.000	619.154	-188.154
8	65	140.000	243.217	-103.217
9	70	282.000	256.181	25.819
10	120	600.000	385.814	214.186
11	100	303.000	333.961	-30.961
12	90	220.000	308.034	-88.034
13	180	749.000	541.374	207.626
14	220	663.000	645.081	17.919
15	150	623.000	463.594	159.406
1.965		6.215.000	6.215.000	0

$$\hat{Y} = b_0 + b_1 X$$

$$= 74.693,88 + 2.592,67x$$



Qual è il prezzo previsto di un appartamento di 160 mq?



$$\hat{Y} = 74.693,88 + 2.592,67 \cdot 160$$

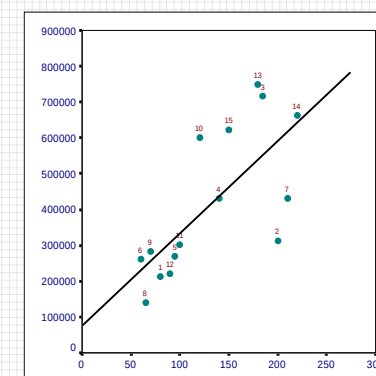
$$= 489.520,7$$

La Regressione: Interpolazione ed estrapolazione

App.	mq (X)	Prezzo in € (Y)	$\hat{Y}$	$Y - \hat{Y}$
1	80	212.000	282.107	-70.107
2	200	313.000	593.227	-280.227
3	185	717.000	554.337	162.663
4	140	431.000	437.667	-6.667
5	95	270.000	320.997	-50.997
6	60	261.000	230.254	30.746
7	210	431.000	619.154	-188.154
8	65	140.000	243.217	-103.217
9	70	282.000	256.181	25.819
10	120	600.000	385.814	214.186
11	100	303.000	333.961	-30.961
12	90	220.000	308.034	-88.034
13	180	749.000	541.374	207.626
14	220	663.000	645.081	17.919
15	150	623.000	463.594	159.406
1.965		6.215.000	6.215.000	0

$$\hat{Y} = b_0 + b_1 X$$

$$\hat{Y} = 74.693,88 + 2.592,67x$$

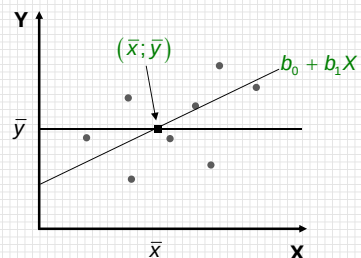


Qual è il prezzo previsto di un appartamento di 260 mq?

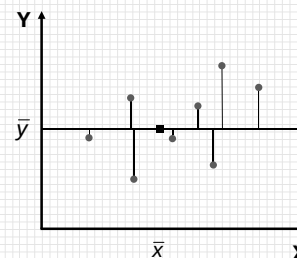


?

La Regressione: Valutazione dell'adattamento

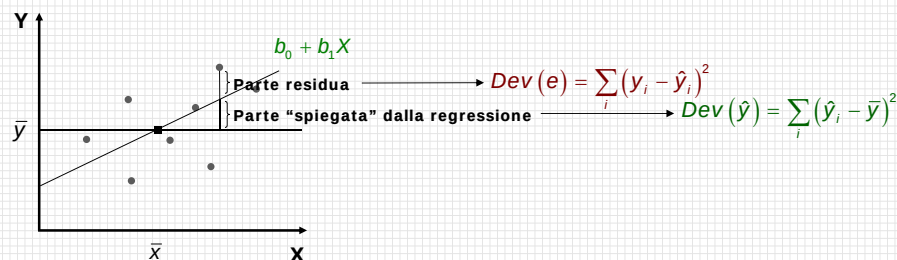


La Regressione: Valutazione dell'adattamento



$$Dev(Y) = \sum_i (y_i - \bar{y})^2$$

La Regressione: Valutazione dell'adattamento



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

$$Dev(\hat{y}) = \sum_i (\hat{y}_i - \bar{y})^2$$

$$Dev(Y) = \sum_i (y_i - \bar{y})^2$$

$$Dev(\hat{y}) = \sum_i (\hat{y}_i - \bar{y})^2$$

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

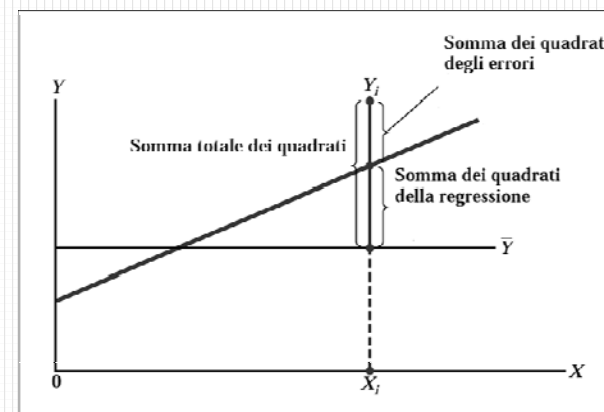
$$Dev(Y) = Dev(\hat{y}) + Dev(e)$$

R²
Indice di
determinazione

$$R^2 = \frac{Dev(\hat{y})}{Dev(Y)} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

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

La Regressione: Valutazione dell'adattamento



L'indice di determinazione serve a valutare la capacità
 previsiva del modello statistico ossia il grado di
 accostamento tra valori empirici e valori teorici

La Regressione

$$\hat{y} = 74.693,88 + 2.592,67x$$

App.	m q (X)	Prezzo in € (Y)	$\hat{Y}$	$Y - \hat{Y}$	$\hat{Y} - \bar{Y}$	$Y - \hat{Y}$
1	80	212.000	282.107	-202.333	-132.226	-70.107
2	200	313.000	593.227	-101.333	178.894	-280.227
3	185	717.000	554.337	302.667	140.004	162.663
4	140	431.000	437.667	16.667	23.334	-6.667
5	95	270.000	320.997	-144.333	-93.336	-50.997
6	60	261.000	230.254	-153.333	-184.079	30.746
7	210	431.000	619.154	16.667	204.821	-188.154
8	65	140.000	243.217	-274.333	-171.116	-103.217
9	70	282.000	256.181	-132.333	-158.153	25.819
10	120	600.000	385.814	185.667	-28.519	214.186
11	100	303.000	333.961	-111.333	-80.373	-30.961
12	90	220.000	308.034	-194.333	-106.299	-88.034
13	180	749.000	541.374	334.667	127.041	207.626
14	220	663.000	645.081	248.667	230.747	17.919
15	150	623.000	463.594	208.667	49.261	159.406
1.965	6.215.000	6.215.000	0	0	0	0

La Regressione

$$\hat{y} = 74.693,88 + 2.592,67x$$

App.	m q (X)	Prezzo in € (Y)	$\hat{Y}$	$(Y - \hat{Y})^2$	$(\hat{Y} - \bar{Y})^2$	$(Y - \hat{Y})^2$
1	80	212.000	282.107	40.938.777.778	17.483.727.210	4.915.031.754
2	200	313.000	593.227	10.268.444.444	32.003.085.447	78.527.393.139
3	185	717.000	554.337	91.607.111.111	19.601.133.620	26.459.127.322
4	140	431.000	437.667	277.777.778	544.475.934	44.453.442
5	95	270.000	320.997	20.832.111.111	8.711.614.942	2.600.724.704
6	60	261.000	230.254	23.511.111.111	33.885.224.478	945.320.444
7	210	431.000	619.154	277.777.778	41.951.534.609	35.401.954.462
8	65	140.000	243.217	75.258.777.778	29.280.705.777	10.653.805.643
9	70	282.000	256.181	17.512.111.111	25.012.283.333	666.640.808
10	120	600.000	385.814	34.472.111.111	813.352.938	45.875.646.835
11	100	303.000	333.961	12.395.111.111	6.459.770.030	958.561.153
12	90	220.000	308.034	37.765.444.444	11.299.556.109	7.749.978.661
13	180	749.000	541.374	112.001.777.778	16.139.342.188	43.108.537.570
14	220	663.000	645.081	61.835.111.111	53.244.368.793	321.099.637
15	150	623.000	463.594	43.541.777.778	2.426.614.965	25.410.267.386
1.965	6.215.000	6.215.000	582.495.333.333	298.856.790.373	283.638.542.960	

$$Dev(Y) = \sum_i (y_i - \bar{y})^2 = 582.495.333.333$$

$$Dev(\hat{Y}) = \sum_i (\hat{y}_i - \bar{y})^2 = 298.856.790.373$$

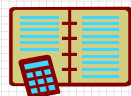
$$Dev(e) = \sum_i (y_i - \hat{y}_i)^2 = 283.638.542.960$$

$$R^2 = \frac{Dev(\hat{Y})}{Dev(Y)} = \frac{298.856.790.373}{582.495.333.333} = 0,513$$

Dove e come studiare



- S. Borra, A. Di Ciaccio (2008) – *Statistica – Metodologie per le scienze economiche e sociali* – McGraw-Hill. Cap. 16 (escluso paragrafo 16.6).
- D. Piccolo (2004) – *Statistica per le decisioni* – Il Mulino. Cap. 18 (escluso paragrafi 18.6, 18.7).



File “esercizi statistiche bivariate.pdf”

- Esercizio n. 3
- Esercizio n. 5
- Esercizio n. 8
- Esercizio n. 9

Riepilogo



Associazione tra caratteri quantitativi

- ✓ Studio della interdipendenza
 - ✓ Diagramma di dispersione
 - ✓ Covarianza
 - ✓ Coefficiente di correlazione
- ✓ Studio della dipendenza (Regressione)
 - ✓ Dipendenza funzionale (o deterministica) e dipendenza statistica
 - ✓ Scelta del tipo di funzione
 - ✓ Scelta del vincolo (Metodo dei minimi quadrati)
 - ✓ Calcolo dei parametri
 - ✓ Interpolazione ed estrapolazione
 - ✓ Analisi della bontà dell'adattamento
 - ✓ Devianza totale
 - ✓ Devianza spiegata (o di regressione)
 - ✓ Devianza residua
 - ✓ Indice di determinazione
 - ✓ Analisi dei residui