

Studio della dipendenza di una variabile quantitativa da un'altra variabile quantitativa (regressione)

Task Description

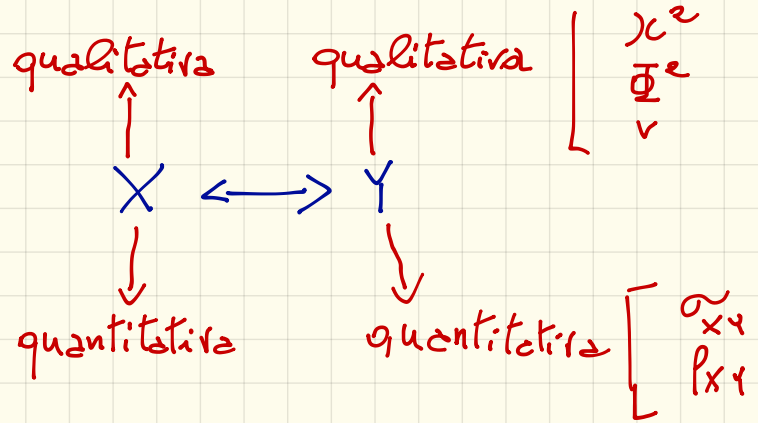
Due Date

Status

☐	L'analisi di regressione per misurare la dipendenza in media di una variabile quantitativa da una variabile quantitativa. La retta di regressione come interpolazione del diagramma di dispersione.		
☐	Criterio di interpolazione dei minimi quadrati: sistema dei minimi quadrati e corrispondenti soluzioni.		
☐	Valutazione della bontà di adattamento (adattamento): scomposizione della devianza complessiva ed indice R^2 (bontà di adattamento).		
☐	Formule alternative per il calcolo della devianza di regressione (e quindi anche della devianza dell'errore)		
☐	Esempio su serie bivariate grezza e su tabella a doppia entrata		
☐			
☐			
☐			
☐			

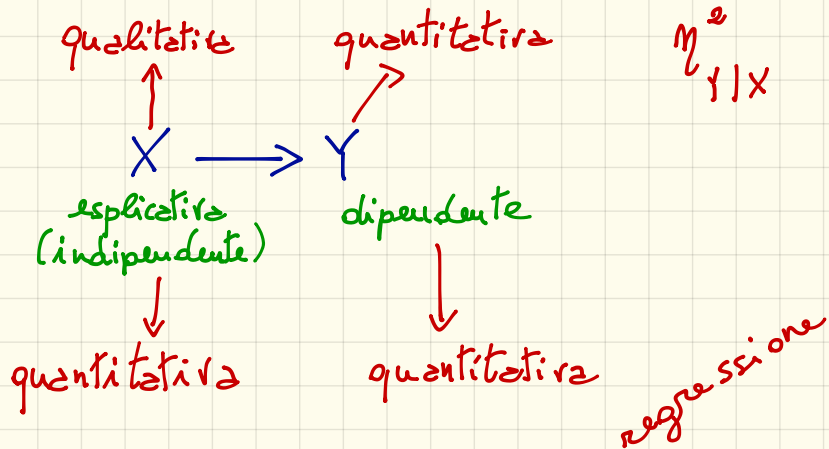
Analisi
simmetrico

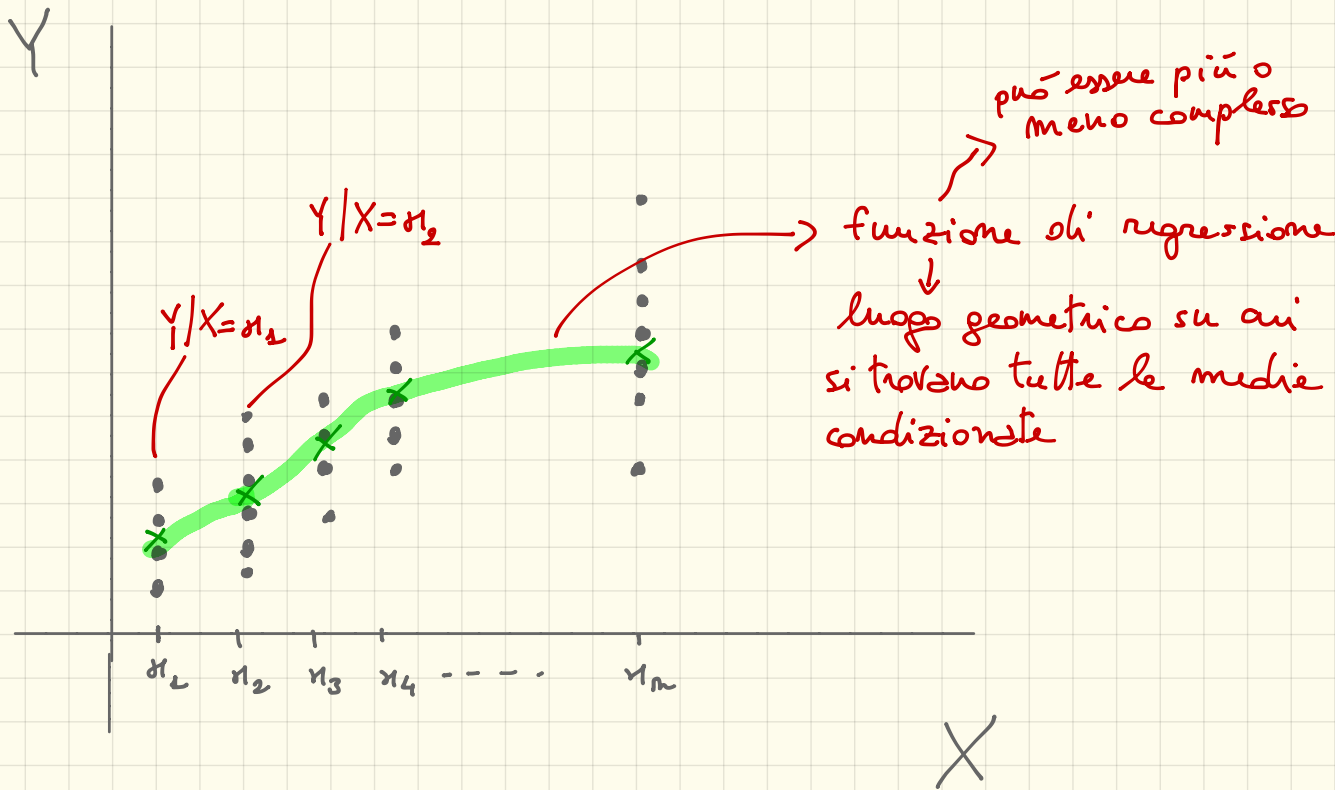
associazione
connessione
correlazione
interdipendenza



Analisi
asimmetrico

dipendenza





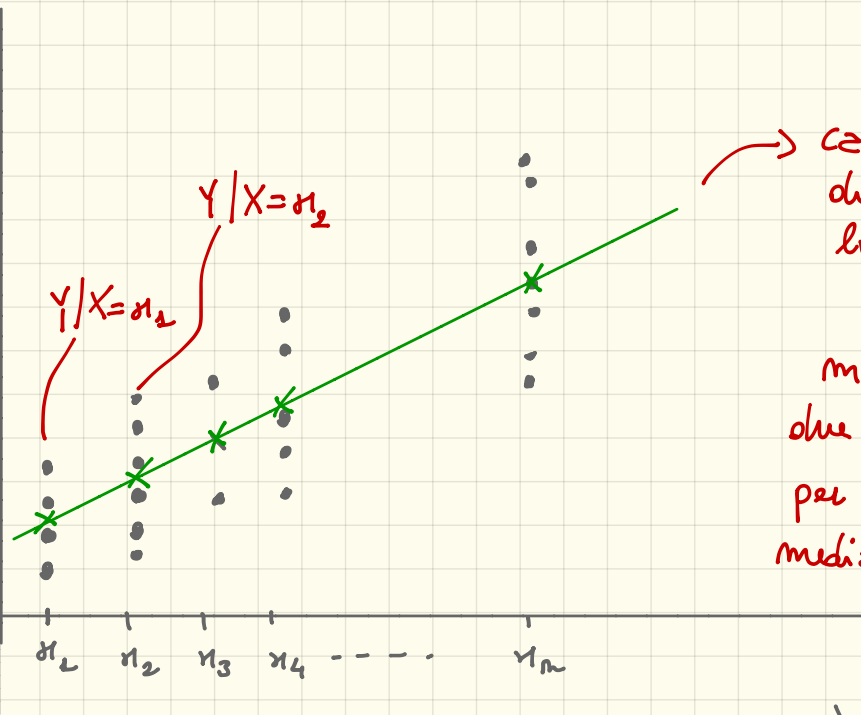
X = reddito, Y = consumo

X = voto diploma, Y = voto laurea

X = voto laurea, Y = stipendio

X = spese pubbliche, Y = rendite

Y

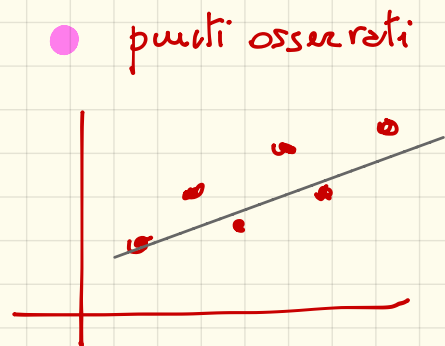
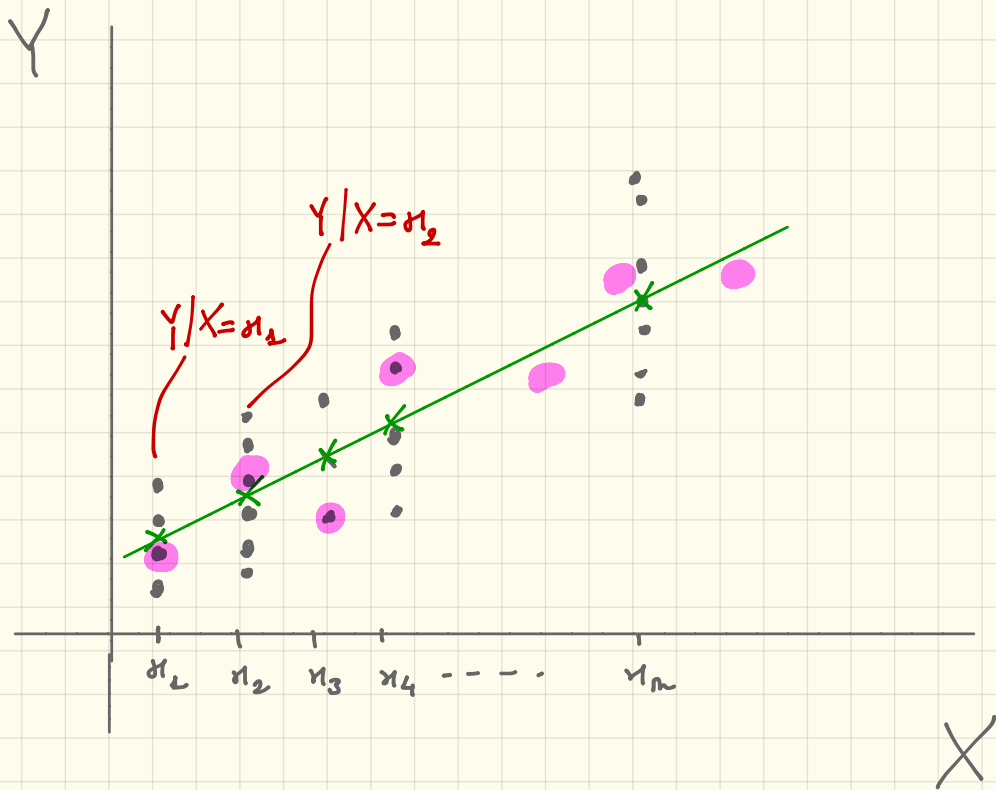


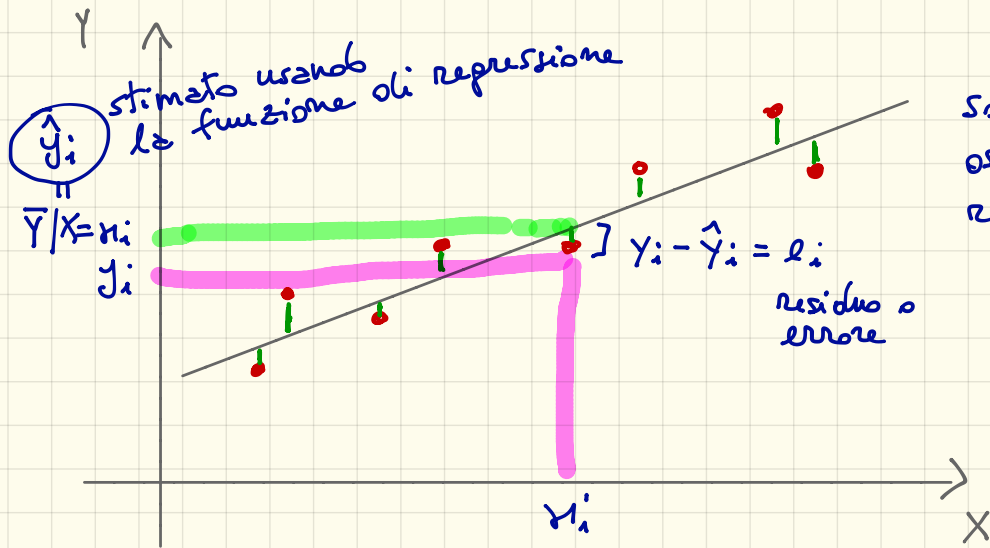
→ caso in cui la funzione di regressione è di tipo lineare

↓
mi basta "conoscere" i due coefficienti della retta per avere informazioni sulle medie condizionate

$$\bar{Y} | X$$

X





sostituendo ai punti osservati la "migliore" retta che passa per i punti

la ricerca della retta viene fatta cercando di minimizzare gli errori

$d(y_i, \hat{y}_i)$ la distanza tra il valore osservato e il valore suggerito dalla retta

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

$$d(y_i, \hat{y}_i) \quad |y_i - \hat{y}_i| \rightsquigarrow \min \sum_{i=1}^n |y_i - \hat{y}_i|$$

criterio degli scostamenti
medianti



retta meno sensibile alle
presenze di valori anomali
(più robusta)

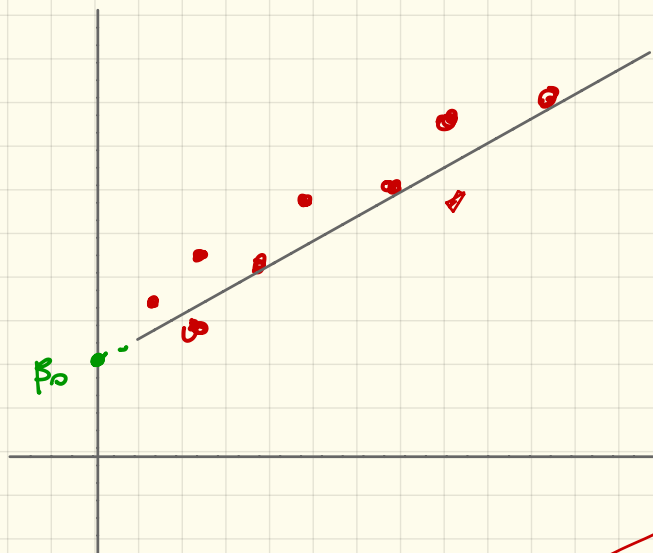
REGRESSIONE
MEDIANA

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

criterio degli scostamenti
quadratici o dei minimi
quadrati

REGRESSIONE
MEDIA

criterio più agevole da
trattare da un punto di
vista analitico



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

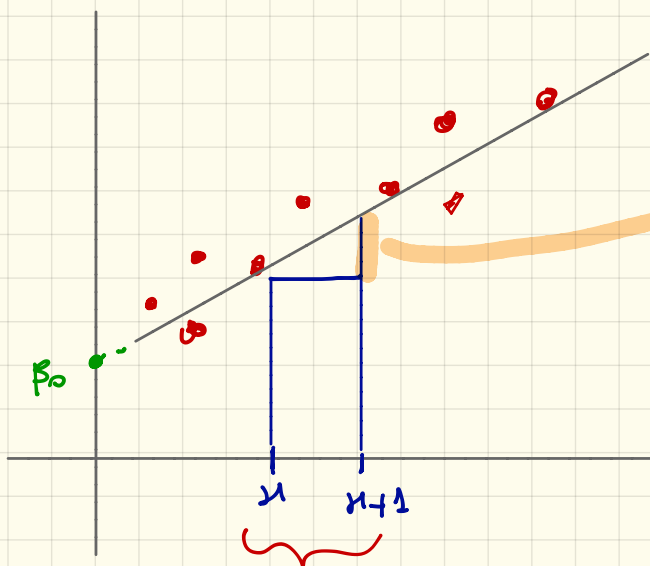


termina moto
o intercalte

β_0 è il valore
di $\hat{y}$ quando $X=0$

$$\begin{aligned}\hat{y}(0) &= \beta_0 + \beta_1 \times 0 \\ &= 0\end{aligned}$$

la possibilità di interpretarlo
è legato alla particolare
applicazione



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

pendenza o
coefficiente angolare

$$\begin{aligned}\hat{y}(x+1) &= \beta_0 + \beta_1(x+1) = \\ &= \beta_0 + \beta_1 x + \beta_1\end{aligned}$$

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

incremento unitario
sulle X

$$\hat{y}(x+1) - \hat{y}(x) = \beta_1$$

variazione che mi aspetto
sulla $\hat{y}$ per un incremento
unitario sulle X

sulla media
della $\hat{y}$

$$\min \sum_i (y_i - \hat{y}_i)^2 = \min \sum_i (y_i - [\beta_0 + \beta_1 x_i])^2 =$$

$$= \min_{\beta_0, \beta_1} \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\left[\begin{array}{l} \frac{\partial \sum_i (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_0} = 0 \\ \frac{\partial \sum_i (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1} = 0 \end{array} \right]$$

funzione
quadratica

posso lavorare
solo con le derivate
prime poichè sono sicuro che
si tratta di punti di
minimo

$$\left\{ \begin{array}{l} -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) = 0 \rightarrow \sum_i y_i - \beta_0 \sum_i 1 - \beta_1 \sum_i x_i = \sum_i \beta_0 = n \beta_0 \\ -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{array} \right.$$

$$\beta_0 = \frac{\sum y_i}{n} - \beta_1 \frac{\sum x_i}{n} = \mu_y - \beta_1 \mu_x$$

Sostituendo il valore ottenuto per β_0 nella seconda equazione, si ottiene:

$$\beta_1 = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{\sum (x_i - \mu_x)^2} = \frac{\text{CODEV}(X, Y) / n}{\text{DEV}(X) / n} = \frac{\sigma_{xy}}{\sigma_x^2}$$

Inoltre, sia per la serie bivariata grezza che per la tabella doppia, è possibile sfruttare le formule alternative di calcolo per la covarianza (covarianza) e la devianza (varianza)

Serie bivariata grezza:

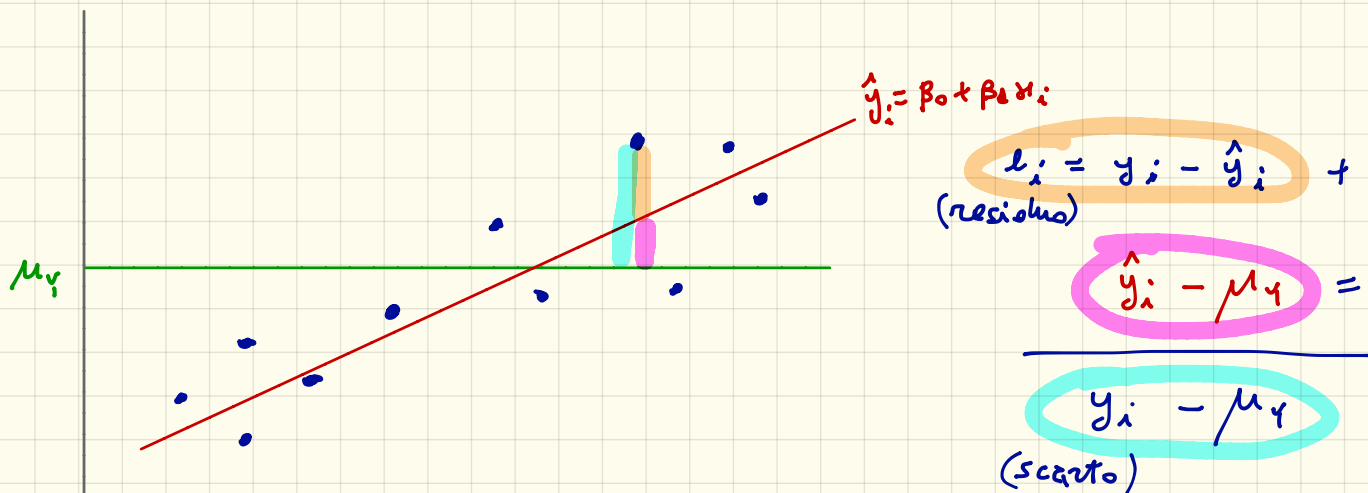
$$\beta_1 = \frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{\sum_i (x_i - \mu_x)^2} = \frac{\sum_i x_i y_i - n \mu_x \mu_y}{\sum_i x_i^2 - n \mu_x^2}$$

formule "ridotte"

Tabella a doppia entrata:

$$\beta_1 = \frac{\sum_{i=1}^n \sum_{j=1}^c (x_i - \mu_x)(y_j - \mu_y) m_{ij}}{\sum_{i=1}^n (x_i - \mu_x)^2 m_{i\bullet}} = \frac{\sum_{i=1}^n \sum_{j=1}^c x_i y_j m_{ij} - n \mu_x \mu_y}{\sum_{i=1}^n x_i^2 m_{i\bullet} - n \mu_x^2}$$

VALUTAZIONE DELLA «VALIDITA'» DEL MODELLO

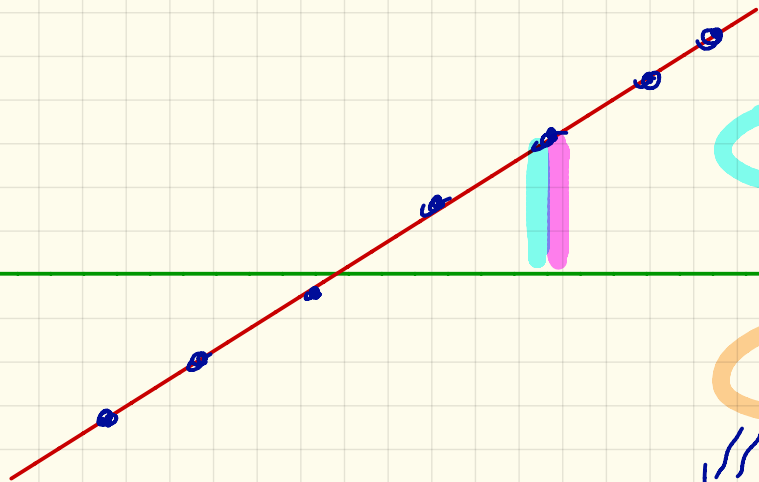


$$Dev(Y) = \sum_{i=1}^n (y_i - \mu_y)^2 = \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{Dev(E)} + \underbrace{\sum_{i=1}^n (\hat{y}_i - \mu_y)^2}_{Dev(R)}$$

$$\sigma_{xy} = \sigma_x \sigma_y$$

$$\rho_{xy} = +1$$

$$\hat{\beta}_1 = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{\cancel{\sigma_x} \sigma_y}{\cancel{\sigma_x}^2} = \frac{\sigma_y}{\sigma_x}$$



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

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

$$y_i - \hat{y}_i = 0$$

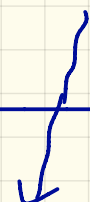
$$\sum_i (y_i - \hat{y}_i)^2 = \text{Der}(E) = 0$$

$$\text{Der}(Y) = \text{Der}(R)$$

in queste situazioni le cose "invisibili"
per predire la Y si considerano sempre
 μ_Y (per qualunque X)


$$\mu_Y = \hat{y}$$

$$(y_i - \mu_Y) = (y_i - \hat{y}_i) + (\hat{y}_i - \mu_Y)$$
$$= 0$$


$$\text{Dev}(Y) = \sum_i (y_i - \mu_Y)^2 = \sum_i (y_i - \hat{y}_i)^2 = \text{Dev}(E)$$

$$\text{Dev}(R) = \sum_i (\hat{y}_i - \mu_Y)^2 = 0$$

$$\text{Dev}(R) = \sum_i (\hat{y}_i - \mu_Y)^2$$

parte della variabilità
di Y spiegata dalla regressione
(ovvero dalle X)

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

parte di variabilità di Y non
spiegata

→ $\text{Dev}(R)$ è un indice della "bontà" del modello

$$0 \leq \text{Dev}(R) \leq \text{Dev}(Y)$$

[quando $\text{Dev}(E) = 0$]

Posso ottenere una misura normalizzata sfruttando le formule di normalizzazione:

$$0 \leq \frac{\overset{=0}{\text{Der}(R)} - \underset{=0}{\min}}{\underset{=0}{\text{Der}(Y)} - \underset{=0}{\min}} \leq 1$$

$\rightarrow R^2 = \pm 1$
 $\rightarrow \sigma_{xy} = \pm \sigma_x \sigma_y$
 $= 1$ quando $\text{Der}(R) = \text{Der}(Y)$
 ovvero quando $\text{Der}(E) = 0$

$$R^2 = \frac{\text{Der}(R)}{\text{Der}(Y)}$$

indice di bontà di adattamento
del modello di regressione
(indice di determinazione)
lineare

$= 0$ quando $\text{Der}(R) = 0$

$\downarrow$
 $\text{Der}(E) = \text{Der}(Y)$

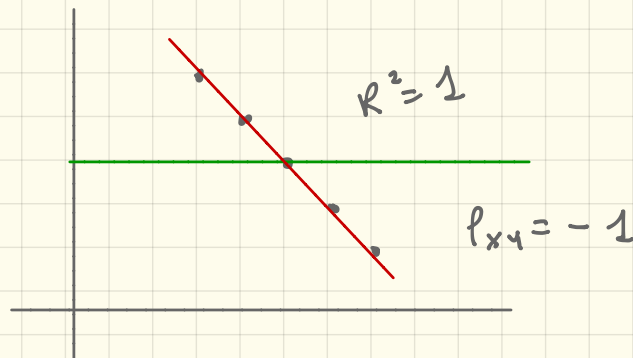
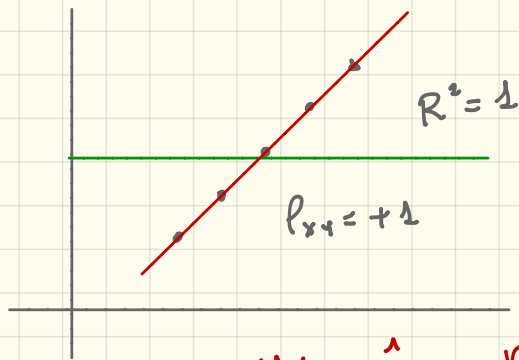
$\xrightarrow{\quad \quad \quad} \beta_1 = 0 \rightarrow \sigma_{xy} = \rho_{xy} = 0$

$$\text{Der}(Y) = \text{Der}(R) + \text{Der}(E)$$


$$R^2 = \frac{\text{Der}(R)}{\text{Der}(Y)} = \frac{\text{Der}(Y) - \text{Der}(E)}{\text{Der}(Y)}$$

$$= 1 - \frac{\text{Der}(E)}{\text{Der}(Y)}$$

Der(E) $\leadsto$ devianza errore: è una misura della "bontà" di adattamento del modello, ottenuta attraverso l'errore di previsione



$$y_i = \hat{y}_i \quad \forall i = 1, n$$

$$\hookrightarrow y_i - \hat{y}_i = 0 \rightarrow \begin{cases} \text{Der}(E) = 0 \\ \text{Der}(Y) = \text{Der}(R) \end{cases}$$

obiettivo:

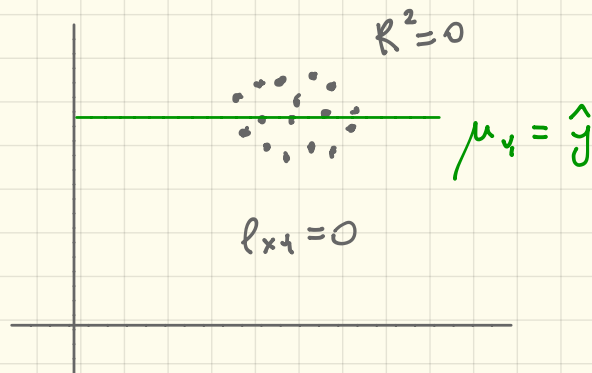
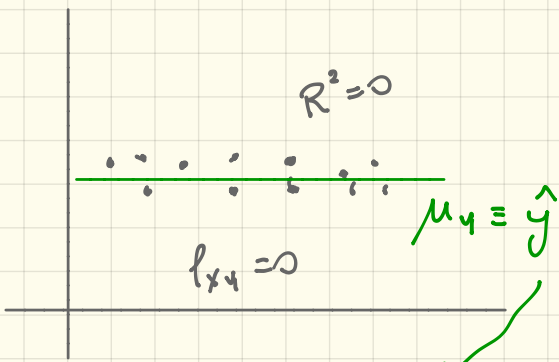
indicatore che assume il valore più basso nel caso di modello poco utile e il valore più alto nel caso opposto

$$\frac{\text{Der}(R) - 0}{\text{Der}(Y) - 0} = \frac{\text{Der}(R)}{\text{Der}(Y)} = R^2 \leftarrow \text{Der}(R)$$

min = 0

max = Der(Y)

indice normalizzato in $[0, 1]$



il miglior modo per predire l_z è di usare la sua media

Esiste un legame tra R^2 e $l_{xy} \leadsto l_{xy}^2 = R^2$

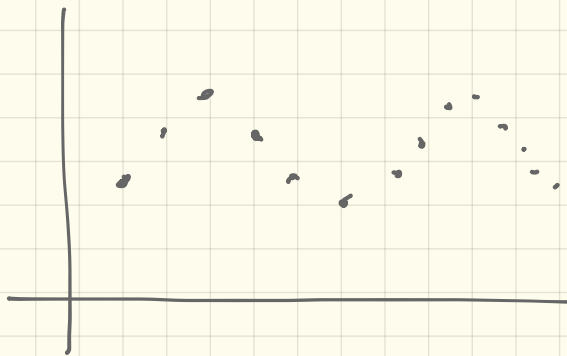
formula utile ai
fini del calcolo

$$R^2 = \frac{\text{Der}(R)}{\text{Der}(Y)} = \frac{\text{Der}(Y) - \text{Der}(E)}{\text{Der}(Y)} = 1 - \frac{\text{Der}(E)}{\text{Der}(Y)}$$

$\text{min} = 0$
 $\text{max} = \text{Der}(Y)$

NOTA: $\text{Der}(Y) = \text{Der}(R) + \text{Der}(E) \Rightarrow \text{Der}(R) = \text{Der}(Y) - \text{Der}(E)$

NOTA: attenzione al caso di legame non lineare tra X e Y



$R^2 = 0 \not\Rightarrow$ assenza legame

$\Rightarrow$ assenza legame lineare (come per il P_{xy})

DIMOSTRAZIONE DELLA PROPRIETA' DI DECOMPOSIZIONE DI $Dev(Y)$

$$Dev(Y) = \sum_{i=1}^n (y_i - \mu_Y)^2 =$$

$$= \sum_i ([y_i - \hat{y}_i] + [\hat{y}_i - \mu_Y])^2 =$$

0

$$= \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{Dev(E)} + \underbrace{\sum_i (\hat{y}_i - \mu_Y)^2}_{Dev(R)} + 2 \sum_i (y_i - \hat{y}_i)(\hat{y}_i - \mu_Y)$$

$$\rightarrow \sum_i \{ (y_i - \hat{y}_i) \hat{y}_i + (y_i - \hat{y}_i) \mu_Y \} =$$

$$= \sum_i \{ (y_i - \hat{y}_i) (\beta_0 + \beta_1 x_i) + (y_i - \hat{y}_i) \mu_Y \} =$$

$$= \beta_0 \underbrace{\sum_i (y_i - \hat{y}_i)}_{\text{(*)} = 0} + \beta_1 \underbrace{\sum_i (y_i - \hat{y}_i) x_i}_{\text{(\bullet)} = 0} + \mu_Y \underbrace{\sum_i (y_i - \hat{y}_i)}_{\text{(*)} = 0}$$

⊛ condizione minimo
I equazione sistema
minimi quadrati

⊙ condizione minimo
II equazione sistema
minimi quadrati

Il calcolo di R^2 è partire da $\text{Dev}(R)$ e/o $\text{Dev}(E)$
 (normalizzate usando $\text{Dev}(Y)$ e abbinanza laborioso)

X	Y	$\hat{Y} = \beta_0 + \beta_1 X$	$(y_i - \hat{y}_i)^2$	$(\hat{y}_i - \mu_Y)^2$
x_1	y_1	$\hat{y}_1 = \beta_0 + \beta_1 x_1$	$(y_1 - \hat{y}_1)^2$	$(\hat{y}_1 - \mu_Y)^2$
x_2	y_2	$\hat{y}_2 = \beta_0 + \beta_1 x_2$	$(y_2 - \hat{y}_2)^2$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_i	y_i	$\hat{y}_i = \beta_0 + \beta_1 x_i$	$(y_i - \hat{y}_i)^2$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_m	y_m	$\hat{y}_m = \beta_0 + \beta_1 x_m$	$(y_m - \hat{y}_m)^2$	
			$\sum (y_i - \hat{y}_i)^2$	$\sum (\hat{y}_i - \mu_Y)^2$
			" $\text{Dev}(E)$	" $\text{Dev}(R)$

NOTA: se avete $\text{Dev}(Y)$ potete
 calcolare una sola delle due
 e ottenere l'altra per differenza

FORMULE ALTERNATIVE CALCOLO R^2

$$R^2 = \frac{\text{Dev}(R)}{\text{Dev}(Y)} = 1 - \frac{\text{Dev}(E)}{\text{Dev}(Y)} = \rho_{xy}^2$$

Si ha anche che:

$$R^2 = \rho_{xy}^2 = \left[\frac{\text{CODEV}(X, Y)}{\sqrt{\text{Dev}(X)} \sqrt{\text{Dev}(Y)}} \right]^2 =$$

$$= \frac{\text{Cov}(X, Y)^2}{\text{Dev}(X) \text{Dev}(Y)} \frac{\text{Dev}(X)}{\text{Dev}(X)} = \left[\frac{\text{CODEV}(X, Y)}{\text{Dev}(X)} \right]^2 \frac{\text{Dev}(X)}{\text{Dev}(Y)} =$$

$$= \beta_1^2 \frac{\text{Dev}(X)}{\text{Dev}(Y)}$$

Dalle relazioni sopra esposte ci ha quindi:

$$R^e = \frac{\text{Dev}(R)}{\text{Dev}(Y)} = \rho_{xy}^2 \implies \text{Dev}(R) = \rho_{xy}^2 \text{Dev}(Y)$$

$$R^e = \frac{\text{Dev}(R)}{\text{Dev}(Y)} = \beta_1^2 \frac{\text{Dev}(X)}{\text{Dev}(Y)} \implies \text{Dev}(R) = \beta_1^2 \text{Dev}(X)$$

E' possibile cioè calcolare la $\text{Dev}(R)$, e quindi anche la $\text{Dev}(E) = \text{Dev}(Y) - \text{Dev}(R)$, a partire da ρ_{xy} e $\text{Dev}(Y)$ oppure a partire da β_1 e $\text{Dev}(Y)$

↳ questo permette di evitare il calcolo di $\sum (y_i - \hat{y}_i)^2$ e $\sum (\hat{y}_i - \mu_Y)^2$ che può essere laborioso

Esempio su serie bivariate grezze:

- $X \rightarrow$ superficie appartamenti in mq
- $Y \rightarrow$ prezzo dell'appartamento (in migliaia di €)

App.	mq (X)	Prezzo (Y)	$X-M(X)$	$Y-M(Y)$	$[X-M(X)]*[Y-M(Y)]$	$X * Y$
1	80	212	-51,0	-202,3	10.319	16.960
2	200	313	69,0	-101,3	-6.992	62.600
3	185	717	54,0	302,7	16.344	132.645
4	140	431	9,0	16,7	150	60.340
5	95	270	-36,0	-144,3	5.196	25.650
6	60	261	-71,0	-153,3	10.887	15.660
7	210	431	79,0	16,7	1.317	90.510
8	65	140	-66,0	-274,3	18.106	9.100
9	70	282	-61,0	-132,3	8.072	19.740
10	120	600	-11,0	185,7	-2.042	72.000
11	100	303	-31,0	-111,3	3.451	30.300
12	90	220	-41,0	-194,3	7.968	19.800
13	180	749	49,0	334,7	16.399	134.820
14	220	663	89,0	248,7	22.131	145.860
15	150	623	19,0	208,7	3.965	93.450
	1.965	6.215	0,0	0,0	115.270	929.435

	X	Y	XY
Media	131,00	414,33	
Sqm	54,44	197,06	
Var	2.964,00		
Cov			7.684,6667

Beta ₁	2,59
Beta ₀	74,69

App.	mq (X)	Prezzo (Y)	$\hat{Y}$	$Y - \bar{Y}$	$\hat{Y} - \bar{Y}$	$Y - \hat{Y}$	$(Y - \bar{Y})^2$	$(\hat{Y} - \bar{Y})^2$	$(Y - \hat{Y})^2$
1	80	212	282	-202	-132	-70	40.939	17.484	4.915
2	200	313	593	-101	179	-280	10.268	32.003	78.527
3	185	717	554	303	140	163	91.607	19.601	26.459
4	140	431	438	17	23	-7	278	544	44
5	95	270	321	-144	-93	-51	20.832	8.712	2.601
6	60	261	230	-153	-184	31	23.511	33.885	945
7	210	431	619	17	205	-188	278	41.952	35.402
8	65	140	243	-274	-171	-103	75.259	29.281	10.654
9	70	282	256	-132	-158	26	17.512	25.012	667
10	120	600	386	186	-29	214	34.472	813	45.876
11	100	303	334	-111	-80	-31	12.395	6.460	959
12	90	220	308	-194	-106	-88	37.765	11.300	7.750
13	180	749	541	335	127	208	112.002	16.139	43.109
14	220	663	645	249	231	18	61.835	53.244	321
15	150	623	464	209	49	159	43.542	2.427	25.410
1.965		6.215	6.215	0	0	0	582.495	298.857	283.639
							DEV(Y)	DEV(R)	DEV(E)

R quadro	DEV(R) / DEV(Y)	0,51
	1 - DEV(E) / DEV(Y)	0,51
	RHO(X, Y) ^ 2	0,51

Esempio su tabella a doppia entrata (X → voto diploma; Y → voto laurea)

① Tabella di input

		VOTO LAUREA (Y)		
		80 - 105	105 - 110	
VOTO DIPLOMA (X)	60 - 80	35	20	55
	80 - 90	20	25	45
	90 - 100	10	55	65
		65	100	165

② sostituisco alle classi i corrispondenti centri

		VOTO LAUREA (Y)		
		92,5	107,5	
VOTO DIPLOMA (X)	70	35	20	55
	85	20	25	45
	95	10	55	65
		65	100	165
Y ^ 2		8556,25	11556,25	

tabella dei prodotti incrociati $x_i y_j m_{ij}$ per il calcolo della covarianza usando la formula ridotta

		VOTO LAUREA (Y)	
		92,5	107,5
VOTO DIPLOMA (X)	70	226625	150500
	85	157250	228437,5
	95	87875	561687,5

1412375

quadrati di X e Y, utili per il calcolo delle corrispondenti varianze usando le formule ridotte

$$\sum_i \sum_j x_i y_j m_{ij}$$

tabella dei prodotti degli scarti: $(x_i - \mu_x)(y_j - \mu_y)m_{ij}$, per il calcolo della covarianza usando la formula originaria

		VOTO LAUREA (Y)		
		-9,09	5,91	
VOTO DIPLOMA (X)	-13,94	4435,261708	-1647,38292	194,31
	1,06	-192,8374656	156,6804408	1,12
	11,06	-1005,509642	3594,69697	122,34

5340,91

(Y - M(Y))^2	82,64	34,92
--------------	-------	-------

$$\sum_i \sum_j (x_i - \mu_x)(y_j - \mu_y)m_{ij}$$

scarti al quadrato di X e Y, utili per il calcolo delle varianze con le formule originarie

(beta0 + beta1 * xi - yj)^2 * nij		VOTO LAUREA (Y)	
		92,5	107,5
$\hat{y}_i$ (attenzione: una per ogni x_i) beta0 + beta1 * xi	97,61	913,00	1957,26
	101,89	1764,95	785,68
	104,75	1501,05	415,42
		7337,35	
		DEV(E)	

$\sum_i \sum_j (\hat{y}_i - y_i)^2 m_{ij}$

M(X)	83,94
M(Y)	101,59
M(X^2)	7159,09
M(Y^2)	10374,43
VAR(X)	113,27
DEV(Y)	8863,64
VAR(Y)	53,72
SQM(X)	10,64
SQM(Y)	7,33
M(XY)	8559,85
COV(X,Y)	32,37
CORR(X,Y)	0,41
BETA0	77,60
BETA1	0,29
R^2	0,17
DEV(E)	7337,35
DEV(R)	1526,28