LA REGRESSIONE LINEARE SEMPLICE

Introduzione all'argomento

Nello studio delle relazioni tra due (o più) variabili, oltre a misurare **l'intensità** del legame esistente, si è anche interessati ad accertare **come** varia una di esse (dipendente) al variare dell'altra (indipendente, o delle altre, variabili indipendenti), individuando un'opportuna funzione analitica che sintetizzi tale relazione.

- Nel caso di una sola variabile indipendente si parla di *regressione semplice*;
- In presenza di due o più variabili indipendenti siamo nel campo della *regressione multipla*.

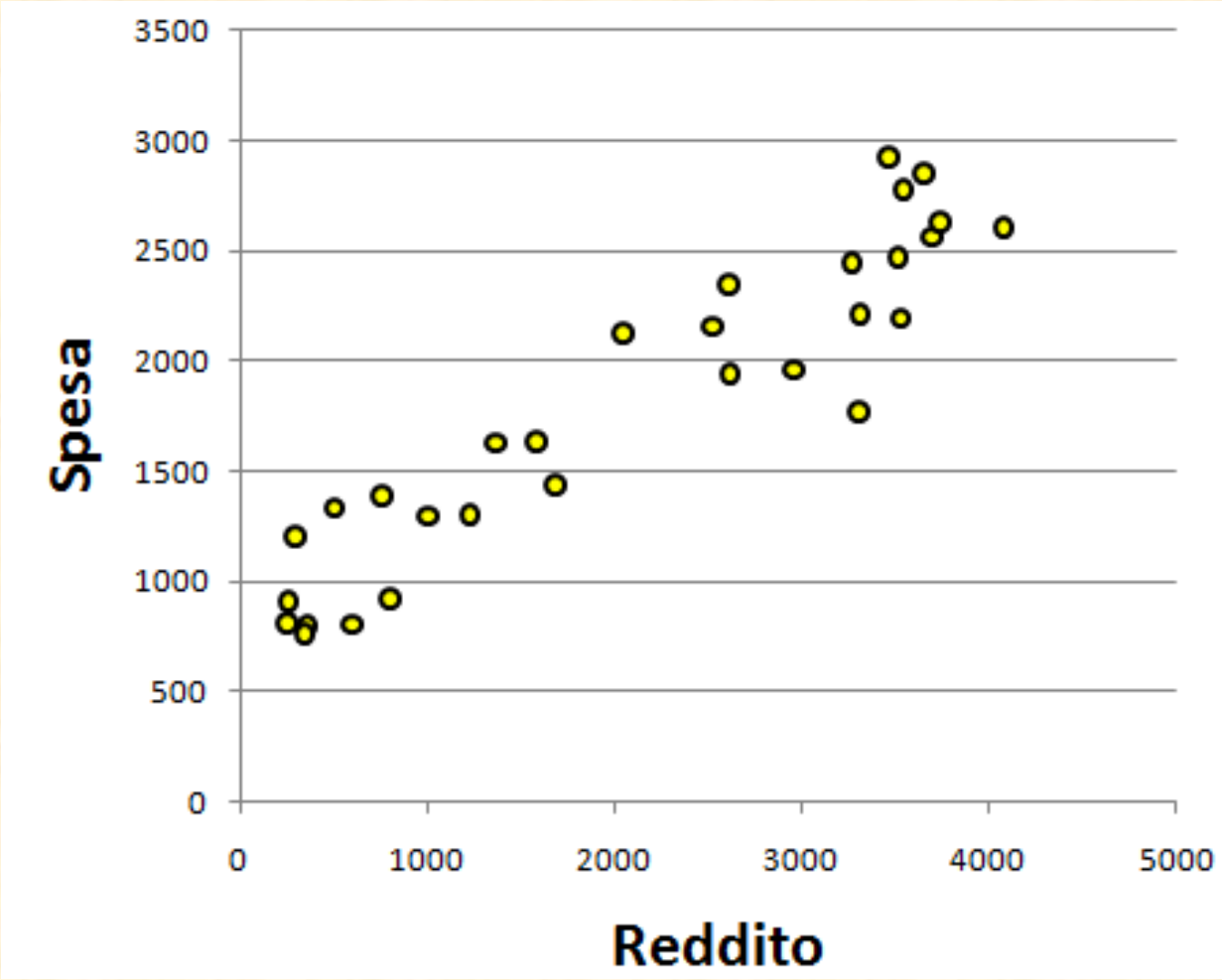
OBIETTIVI DELLA REGRESSIONE

***Descrizione**:* si vuole rappresentare tramite una funzione l'andamento dei valori d'una variabile al variare dell'altra.

***Interpretazione**:* si cerca di mettere in evidenza i nessi causali fra le variabili, per confermare (o smentire) una teoria (economica nel ns. caso).

***Previsione**:* si tenta di valutare in maniera attendibile il valore che assumerà la variabile dipendente in corrispondenza d'un valore noto della variabile esplicativa (o delle variabili esplicative, nel caso di regressione multipla).

Esempio: osservo reddito e spesa su 30 famiglie



Relazione tra due variabili (Regressione semplice)

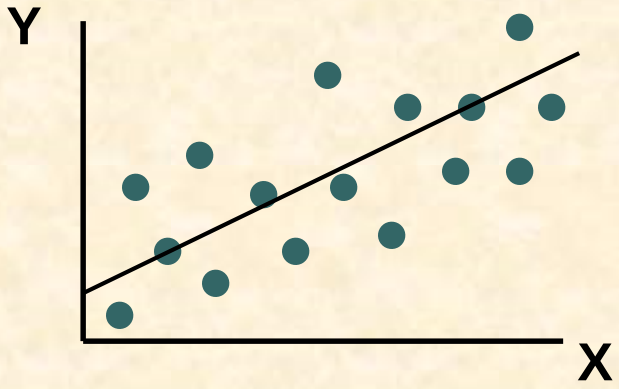
Dopo aver rappresentato graficamente i dati a mezzo dello *scatter-plot* se notiamo una regolarità di tipo lineare (i punti si dispongono grossomodo attorno ad una retta immaginaria) possiamo voler "sintetizzare" tale "regolarità" mediante una funzione analitica "ragionevolmente semplice"

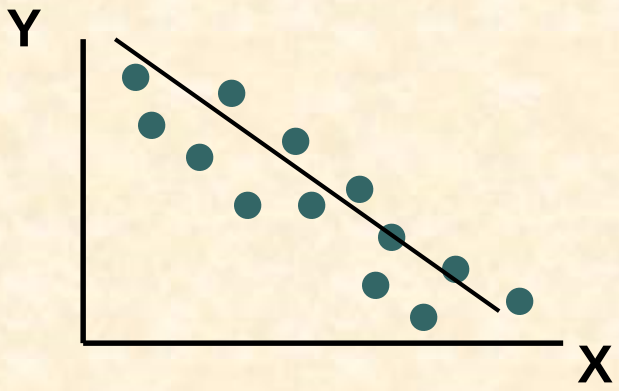
Il presupposto è che esista una variabile (la "X" detta indipendente o esogena) che è causa o che comunque agisce sull'altra (la "Y" detta dipendente o endogena).

La scelta del ruolo delle due variabili è una scelta extra-statistica

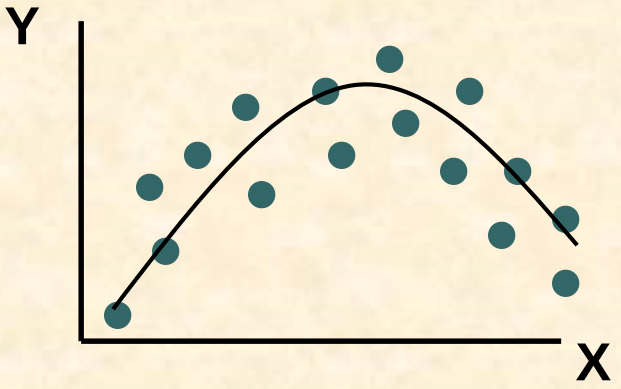
Tipologia di relazioni (una sola variabile indipendente)

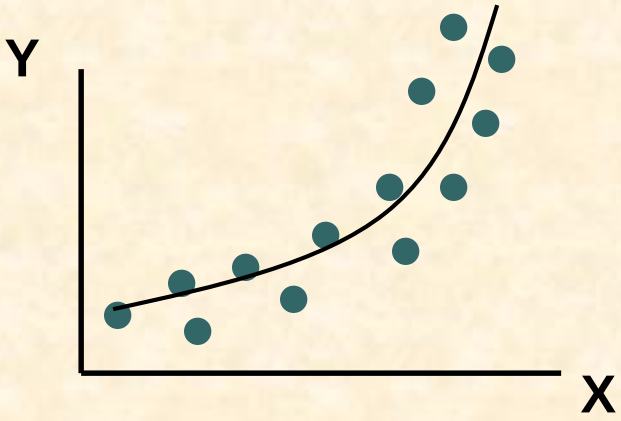
Relazione Lineare





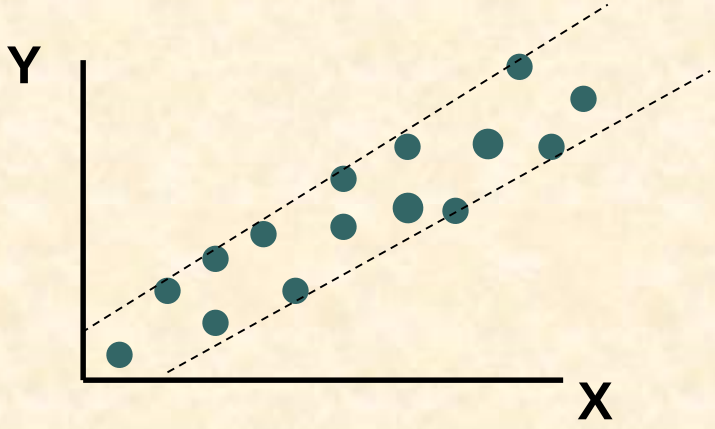
Relazione non lineare

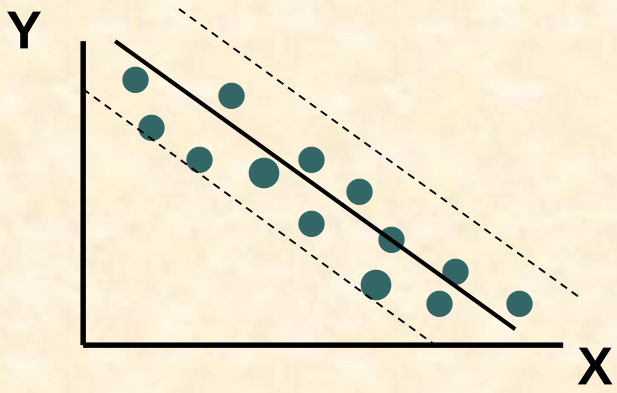




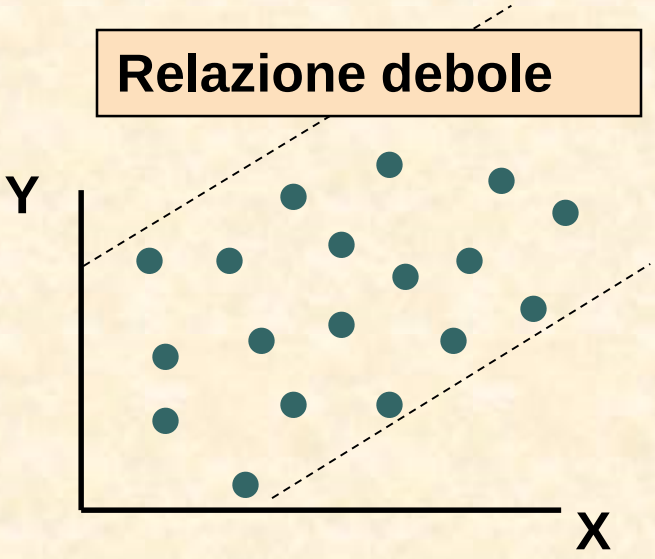
Tipologia di relazioni

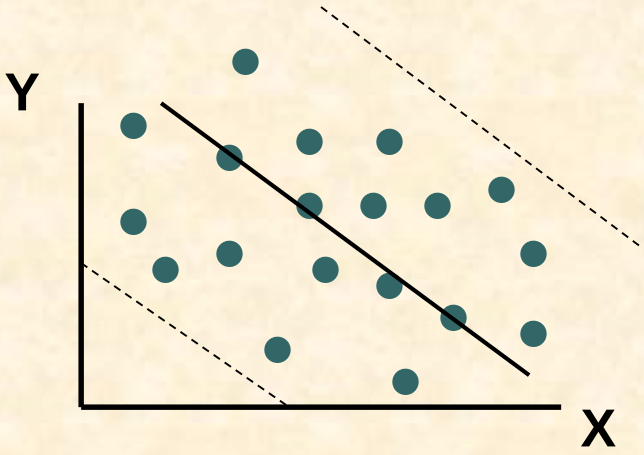
Relazione forte





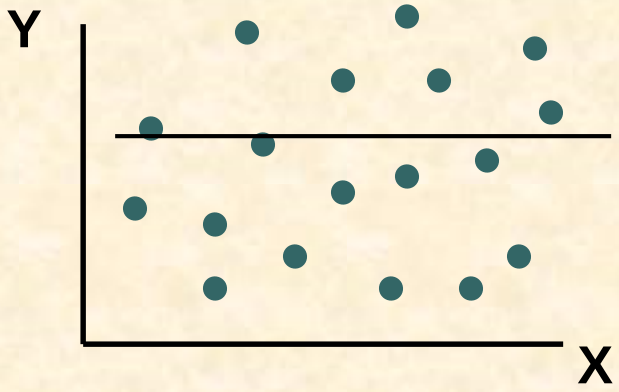
Relazione debole

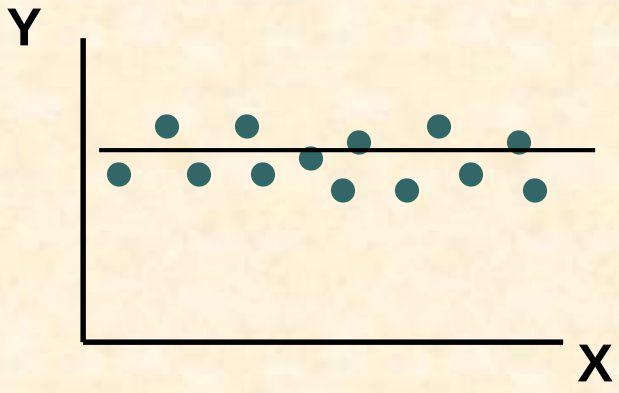




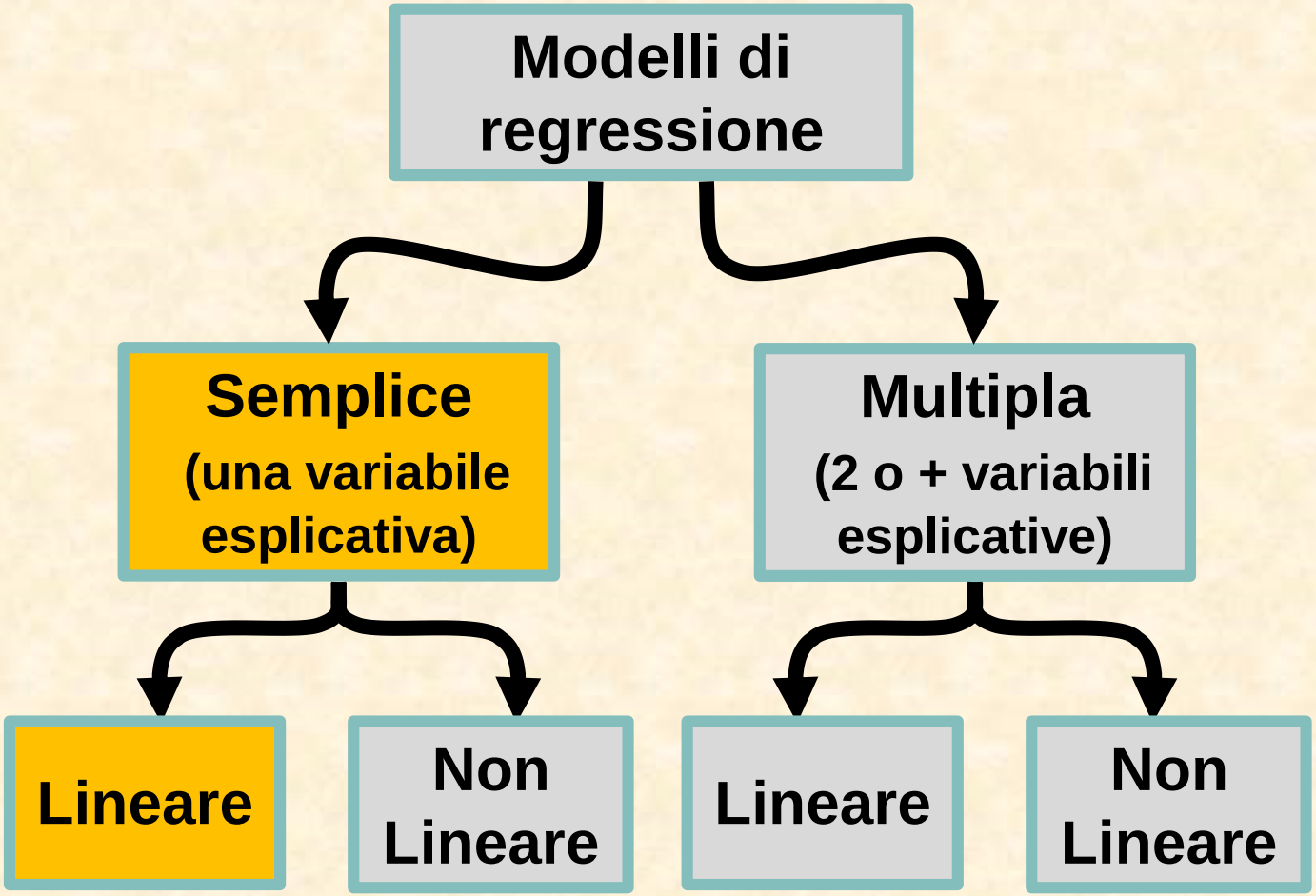
Tipologia di relazioni

Nessuna relazione





Tipologia di modelli di regressione



APPROCCIO DESCRITTIVO

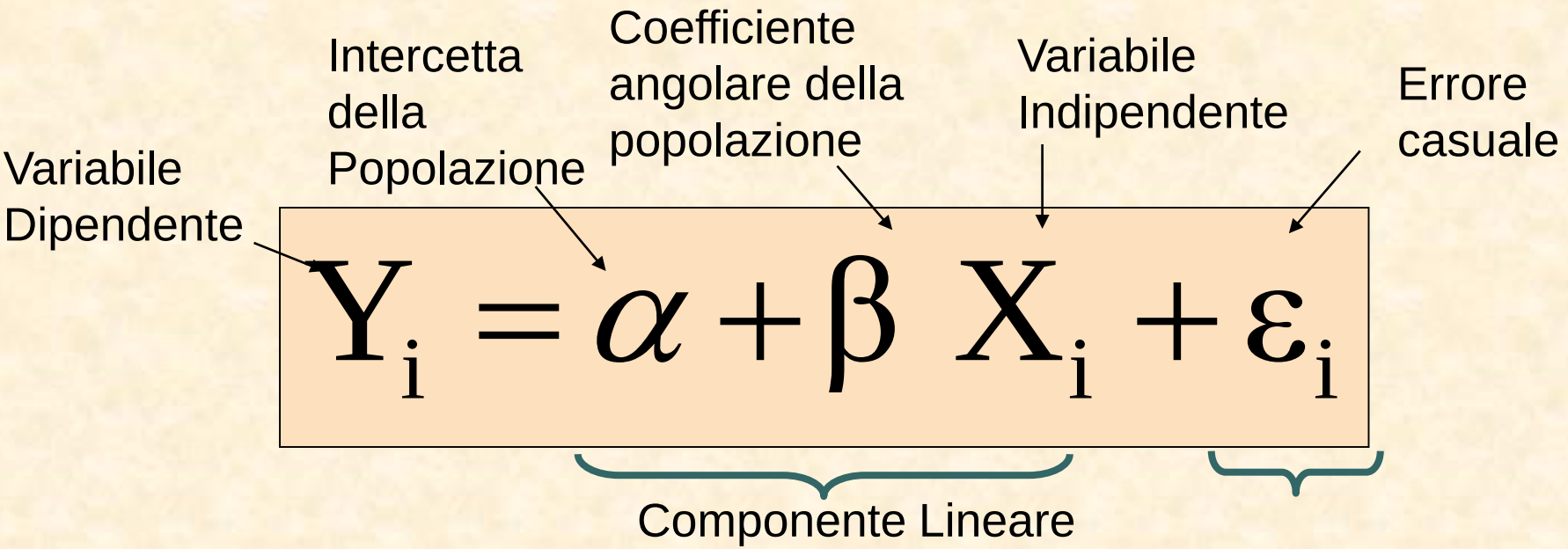
In un approccio descrittivo si considera la regressione come un problema di *interpolazione*, cioè di adattamento d'una funzione (in questo caso la retta) alla "nuvola" dei punti del diagramma di dispersione, in base a sole considerazioni di natura geometrica.

Regressione Lineare

Vi sono molti casi in pratica in cui la teoria di un fenomeno può essere sintetizzata da un modello espresso da una equazione lineare.

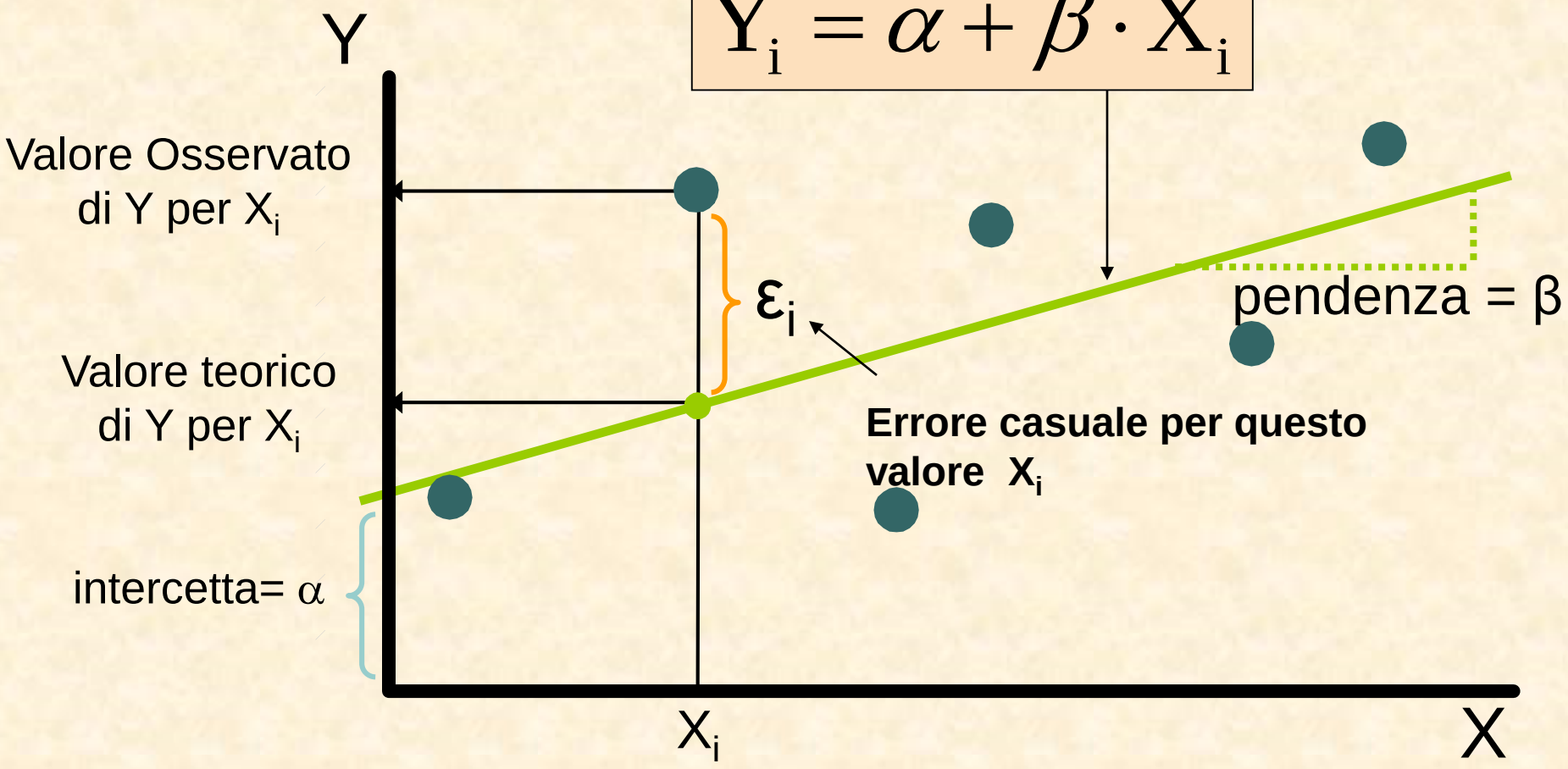
Ad esempio, ”Y" la spesa per consumo delle famiglie e sia ”X" il reddito disponibile.

L'idea che il consumo aumenti all’aumentare del reddito disponibile può essere espressa dalla relazione funzionale:



Il modello lineare

$$Y_i = \alpha + \beta \cdot X_i$$



DALLA TEORIA ALLA SIMULAZIONE

$$Y_i = \alpha + \beta\ X_i + \varepsilon_i$$

Il modello qui sopra mi dice che tutti i possibili Yi sono pari a una funzione lineare degli Xi, secondo un certo α e un certo β, a cui si somma un numero casuale εi.
Di fatto, noi possiamo osservare solo alcune coppie (campione) di Xi-Yi, in base alle quali stimare i parametri α e β e il tipo di errore.
Partiamo però da alcuni punti fermi (assiomi di partenza):

1) Fra X eY c'è una relazione lineare (non perfetta, in quanto "sporcata" dall'errore)
2) Gli errori hanno tutti (qualsiasi i) lo stesso valore atteso paria 0 e la stessa varianza, pari a σ2
3) I valori della X sono noti senza errore

REGRESSIONE LINEARE

Il “successo” del modello **lineare** dovuto a:

1. Ragioni di **Semplicità**: la retta è la più semplice funzione che lega due variabili, è facile da interpretare ed il suo significato è di agevole comprensione.
2. Esigenze di **sintesi**
3. Approssimazione funzionale (**effettiva linearità**): molte relazioni sono lineari o assai vicine alla linearità.
4. **Trasformazioni***:* spesso è possibile ottenere una relazione approssimativamente lineare trasformando una o entrambe le variabili in modo opportuno (ad esempio, considerando i logaritmi di *X* anziché i valori).
5. **Limitatezza dell’intervallo***:* anche se la relazione tra due variabili non è lineare, considerando un intervallo limitato dei valori di *X* e di *Y*, la retta fornisce spesso un’approssimazione soddisfacente

INTERPRETAZIONE

Nel modello di regressione lineare si assume che ciascun valore osservato della variabile dipendente sia esprimibile come funzione lineare del corrispondente valore della variabile esplicativa, più un termine residuo che traduce l'incapacità del modello di riprodurre con esattezza la realtà osservata.

il termine "ε" è il risultato di:

1. Errori e carenze nella misurazione e nella rilevazione di "Y" e di "X"
2. Inadeguatezza della "semplice" relazione lineare
3. Insufficienza del solo fattore X a "spiegare" da solo la Y

DALLA TEORIA ALLA SIMULAZIONE: IL FILE REGRESSIONE.XLS

Supponiamo di essere "onniscienti" cioè **di conoscere α, β e come si distribuisce ciascun εi** ; ad esempio, **supponiamo che ciascun εi si distribuisca normalmente, con varianza σ2 (uguale per qualsiasi i)**

Scriviamo in un foglio Excel (Simulazione) in **B1**, **B2** e **B4** i "parametri significativi" (il valore atteso di ciascun εi è pari a 0 per ipotesi)

	A	B	C
1	alfa	500	
2	beta	0.6	
3	media degli epsilon	0	
4	sigma2 degli epsilon	100000	
5			
6	decidere alfa, beta e		
7	sigma immettendo i valori		
8	nelle celle gialle		
9			

DALLA TEORIA ALLA SIMULAZIONE: IL FILE REGRESSIONE.XLS

Supponiamo di osservare un valore della variabile X (reddito) pari a 3529.
Il corrispondente valore Y (spesa) sarà pari a

$\alpha+\beta \cdot 3529$

+

un numero estratto casualmente da una normale con μ=**B3** e $\sigma 2$=**B4**

	D	E	F	G	H
1	id	xi	alfa+beta x	ei	Yi osserv.
2	1	3529	2617.4	-422.6	2194.80

DALLA TEORIA ALLA SIMULAZIONE: IL FILE REGRESSIONE.XLS

Vediamo le formule:

	D	E
1	id	xi
2	1	=ARROTONDA(100+CASUALE()*4000,0)

Supponiamo che la prima famiglia estratta abbia un reddito compreso fra 100 e 4100…

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Stiamo estraendo un Xi a caso

DALLA TEORIA ALLA SIMULAZIONE: IL FILE REGRESSIONE.XLS

Vediamo le formule:

	F
1	alfa+beta x
2	=B$1+B$2*E2

Nella cella F2 calcoliamo la "parte deterministica" del modello relativa a Xi

$$Y_i = \alpha + \beta\ X_i + \varepsilon_i$$

DALLA TEORIA ALLA SIMULAZIONE: IL FILE REGRESSIONE.XLS

Vediamo le formule:

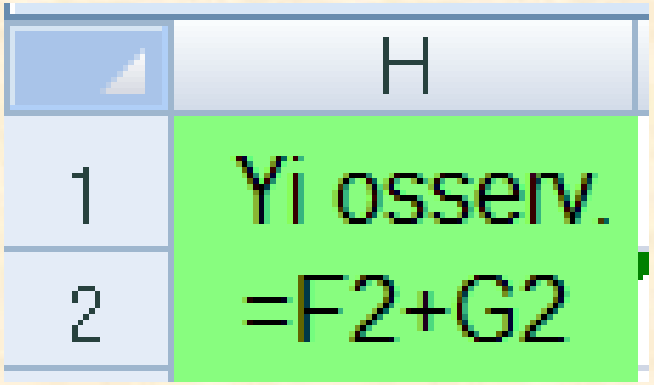
	G
1	ei
2	=INV.NORM(CASUALE(),B$3,B$4^0.5)

Nella cella G2 estraiamo un numero casuale da una normale con μ=B3 e σ2=B4.
Per convincersi che la formula sopra "fa proprio questo", guardare il foglio: "errori", in cui estraiamo 10000 numeri nello stesso modo

$$Y_i = \alpha + \beta\ X_i + \varepsilon_i$$

Generiamo questo secondo la "regola" ipotizzata

DALLA TEORIA ALLA SIMULAZIONE: IL FILE REGRESSIONE.XLS



Nella cella H2 sommo la "parte deterministica" e la "parte stocastica" (una specifica realizzazione della variabile casuale)

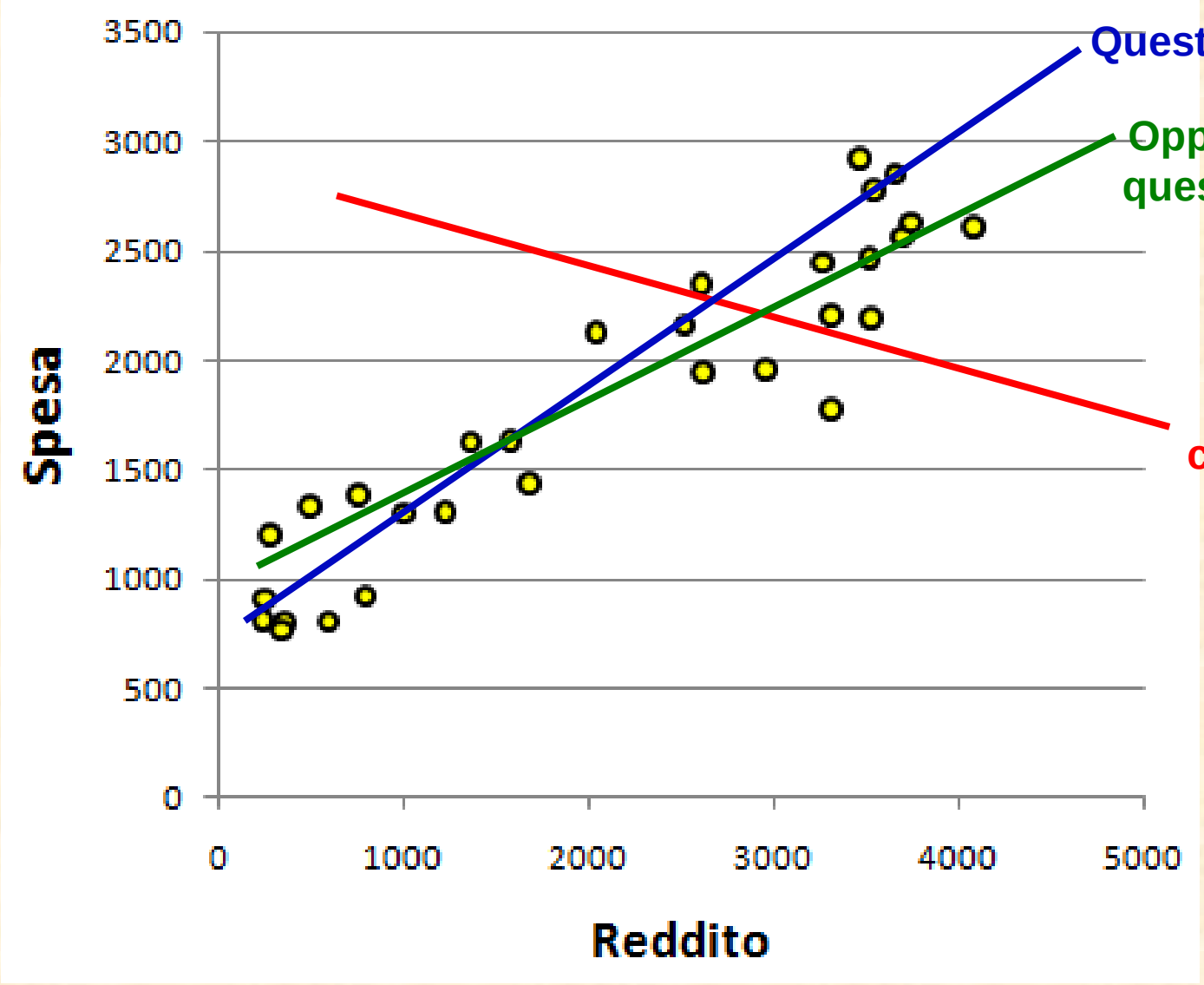
$$Y_i = \alpha + \beta\ X_i + \varepsilon_i$$

DALLA TEORIA ALLA SIMULAZIONE: IL FILE REGRESSIONE.XLS

Copio il contenuto delle celle della riga 2 (colonne dalla E alla H) fino alla riga 31. Ho estratto un campione casuale di Xi e Yi, in cui gli Yi seguono le ipotesi del modello lineare. Se adesso *mi tolgo le vesti del folletto onnisciente* (che conosce α, β e $\sigma 2$) e *mi metto i panni dello statistico*, devo provare a stimare α, β e σ^2 (o σ che è lo stesso) conoscendo le sole "celle verdi"

	D	E	F	G	H
1	id	xi	ei	alfa+beta x	Yi osserv.
2	1	3529	-422.6	2617.4	2194.80
3	2	288	534.1	672.8	1206.86
4	3	503	534.7	801.8	1336.53
5	4	598	-51.9	858.8	806.90
6	5	3545	150.1	2627.0	2777.09
7	6	357	85.7	714.2	799.89
8	7	253	257.9	651.8	909.72
9	8	3655	159.7	2693.0	2852.74
10	9	3698	-155.3	2718.8	2563.51
11	10	2609	287.2	2065.4	2352.61
12	11	755	434.6	953.0	1387.61
13	12	244	164.3	646.4	810.70
14	13	3740	-114.2	2744.0	2629.82
15	14	4079	-339.1	2947.4	2608.28
16	15	3269	-12.8	2461.4	2448.59
17	16	1681	-72.0	1508.6	1436.60
18	17	2042	401.0	1725.2	2126.16
19	18	2618	-124.8	2070.8	1946.02
20	19	1584	184.9	1450.4	1635.30
21	20	1367	308.0	1320.2	1628.19
22	21	3309	-710.7	2485.4	1774.66
23	22	3310	-275.5	2486.0	2210.54
24	23	3517	-142.7	2610.2	2467.46
25	24	340	62.3	704.0	766.30
26	25	2519	148.4	2011.4	2159.77
27	26	1227	71.5	1236.2	1307.74
28	27	1002	197.7	1101.2	1298.86
29	28	2953	-309.2	2271.8	1962.65
30	29	3461	349.4	2576.6	2925.96
31	30	796	-55.4	977.6	922.22

Quale retta?

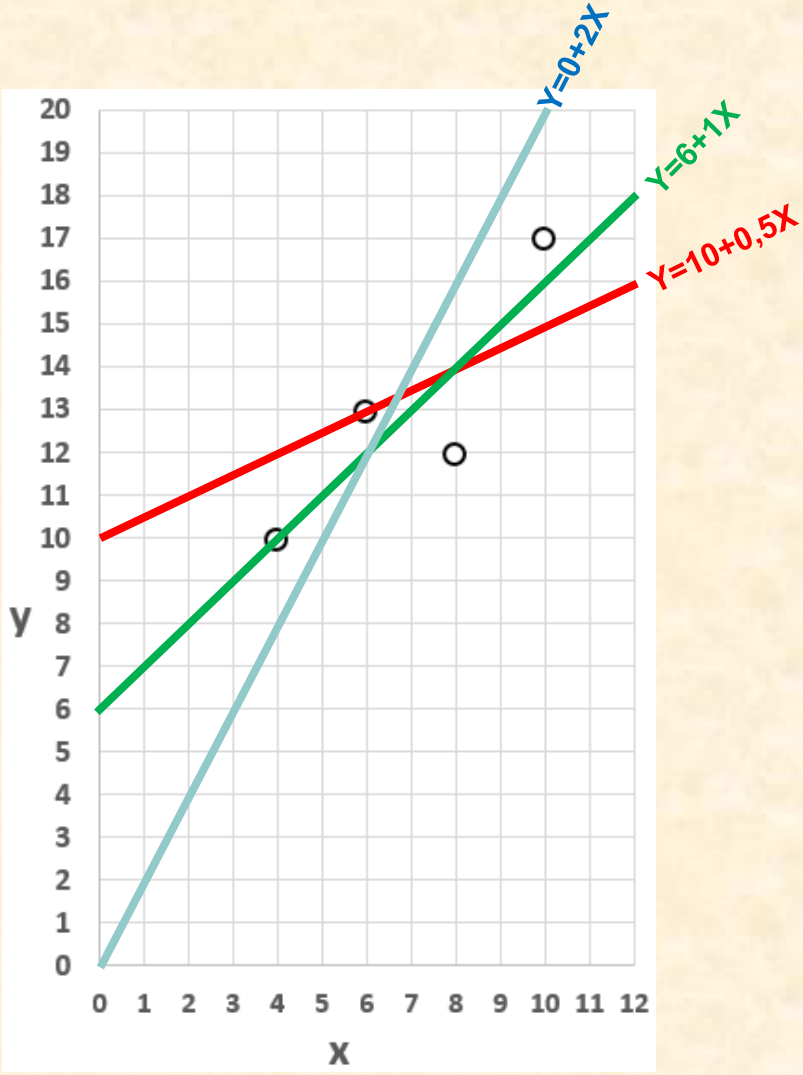

Questa?
Oppure questa?
Questa certamente no!
Spesa
Reddito
3500
3000
2500
2000
1500
1000
500
0
0
1000
2000
3000
4000
5000

x	y
4	10
6	13
8	12
10	17

Abbiamo 4 osservazioni su cui misuriamo due variabili: la X (var.indip.) e la Y (var. dip.)

Vogliamo sintetizzare la relazione tra X e Y mediante una retta. Vediamo tre possibili «candidate»

Quale è la migliore? La blu la rossa o la verde?



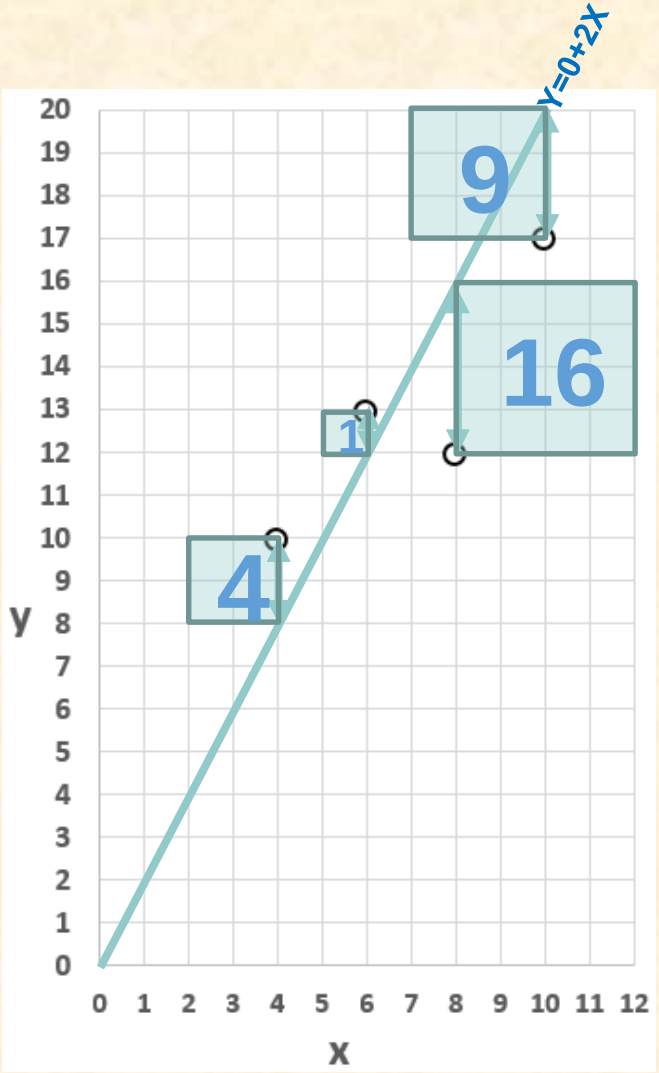
Proviamo a giudicare quanto è «buona» la blu…

x	y
4	10
6	13
8	12
10	17

Non ci interessano tanto le «distanze verticali» dei punti dalla retta…

Ma i QUADRATI di queste distanze (metodo dei minimi quadrati)

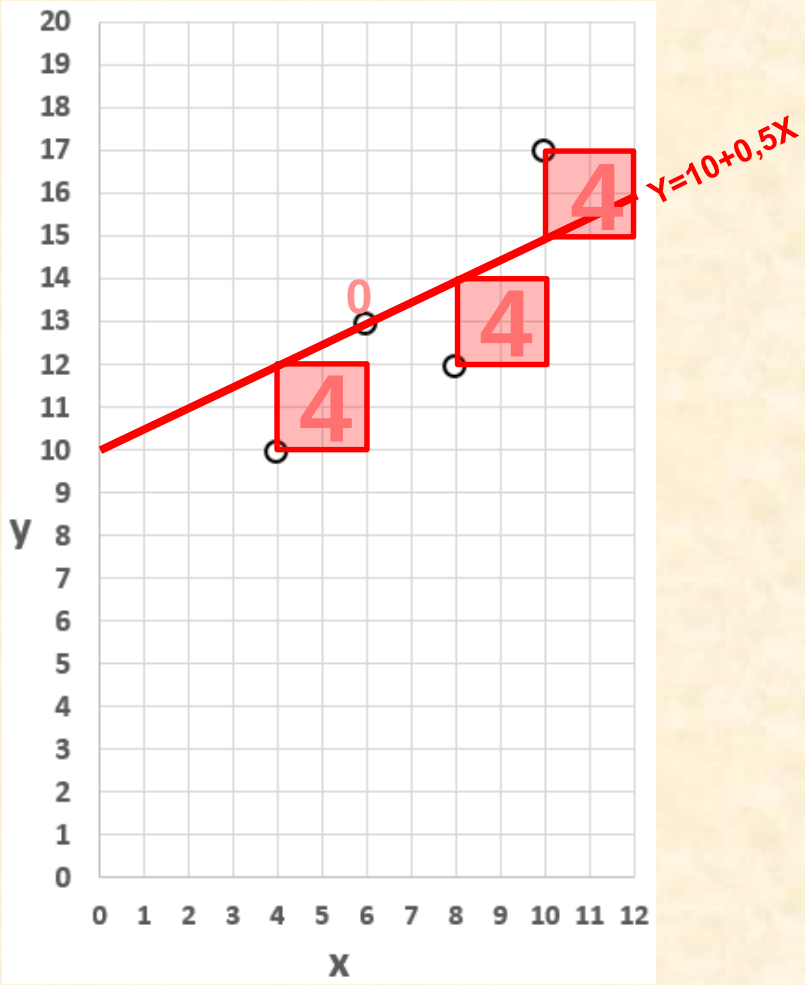
$SSQ_{Y=0+2X}$ **4+1+9+16=30**



Ok, la «distanza complessiva» della blu è 30. Forse la rossa è migliore?

x	y
4	10
6	13
8	12
10	17

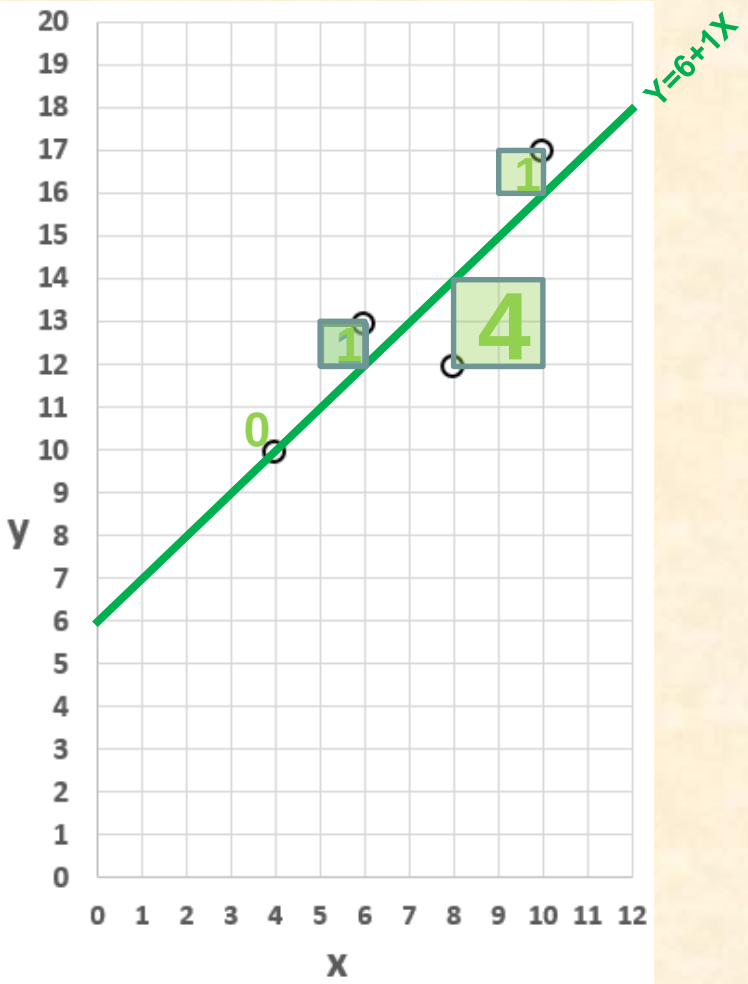
Abbiamo 4 osservazioni su cui misuriamo due variabili: la X (var.indip.) e la Y (var. dip.)



$SSQ_{Y=0+0,5X}$ **4+0+4+4=12**

Sì, la rossa ha complessivamente una somma dei quadrati delle distanze minore, ma come si comporta la verde?

x	y
4	10
6	13
8	12
10	17



$SSQ_{Y=6+1X}$ **0+1+4+1=6**

La verde è la migliore delle tre! Ha infatti una SSQ più bassa (in realtà potremmo dimostrare che la sua SSQ è **la minima assoluta**, ovvero è la **RETTA DEI MINIMI QUADRATI**)

Individuazione della “retta migliore” (ovvero dei parametri)

Occorre stabilire un criterio che ci permetta di scegliere quella che “passa più vicino ai punti” ovvero “si adatta meglio” allo *scatter-plot osservato*

Ogni scelta determina degli errori dovuti alla sostituzione di un valore presunto o teorico ad un valore osservato

Metodo dei minimi quadrati

La retta “migliore” è quella che più si avvicina all’insieme dei punti corrispondenti alle coppie di valori (x_i, y_i).

Per la stima dei parametri α e β si impiega abitualmente il metodo dei *minimi quadrati*, che consiste nella scelta della retta che rende minima la somma dei quadrati dei residui:

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

CALCOLO DEI PARAMETRI

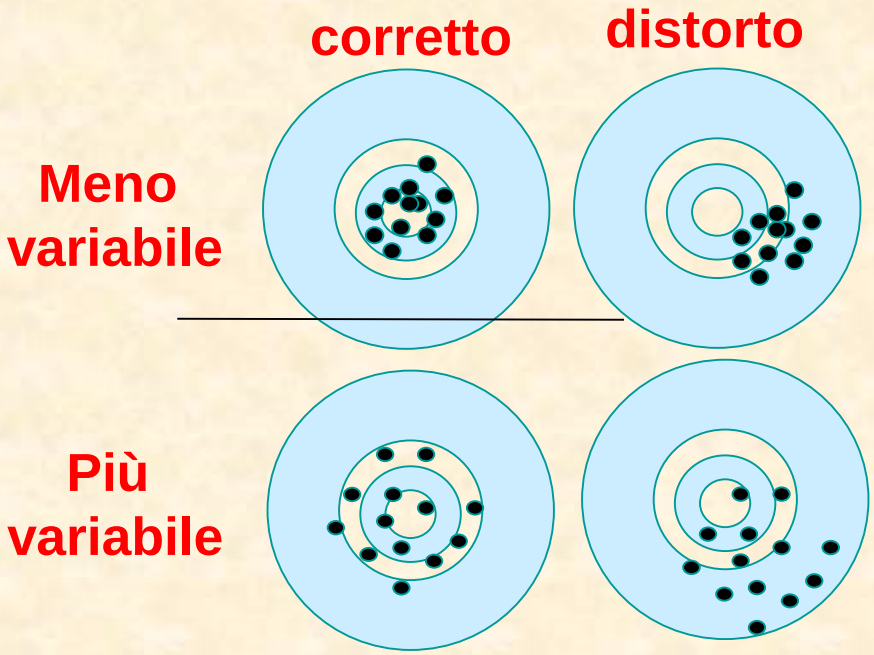
Si può dimostrare che la soluzione del problema di ottimo visto in precedenza è data da:

$$a = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

PERCHÉ IL METODO DEI MINIMI QUADRATI È IL MIGLIORE?

Si dimostra (Teorema di Gauss-Markov) che fra gli "stimatori corretti (e lineari)" di α e β, gli stimatori più efficienti sono quelli trovati col metodo dei minimi quadrati.
Stimatori corretti: non commettono un "errore sistematico"
Stimatori più efficienti: sono i "meno ballerini"



PROPRIETA' I

La somma dei valori teorici è uguale alla somma dei valori osservati: $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$

Da ciò consegue che anche la media dei valori teorici e la media dei valori osservati sono uguali e, inoltre, che la somma dei residui dei minimi quadrati è identicamente nulla:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

PROPRIETA' II

Nel diagramma di dispersione la retta di regressione passa sempre per il punto avente per coordinate la media di X e la media di Y, cioè nel punto (M_x, M_y)

I PARAMETRI DELLA RETTA DEI MINIMI QUADRATI CON EXCEL

	N	O
1	a	796.6347
2	b	0.482991

Vediamo le formule

	N	O
1	a	=INTERCETTA(H2:H31,E2:E31)
2	b	=PENDENZA(H2:H31,E2:E31)

Ricordiamo che in H2:H31 abbiamo i valori osservati (campione di 30 unità) della variabile Y e in E2:E31 i valori osservati della variabile X

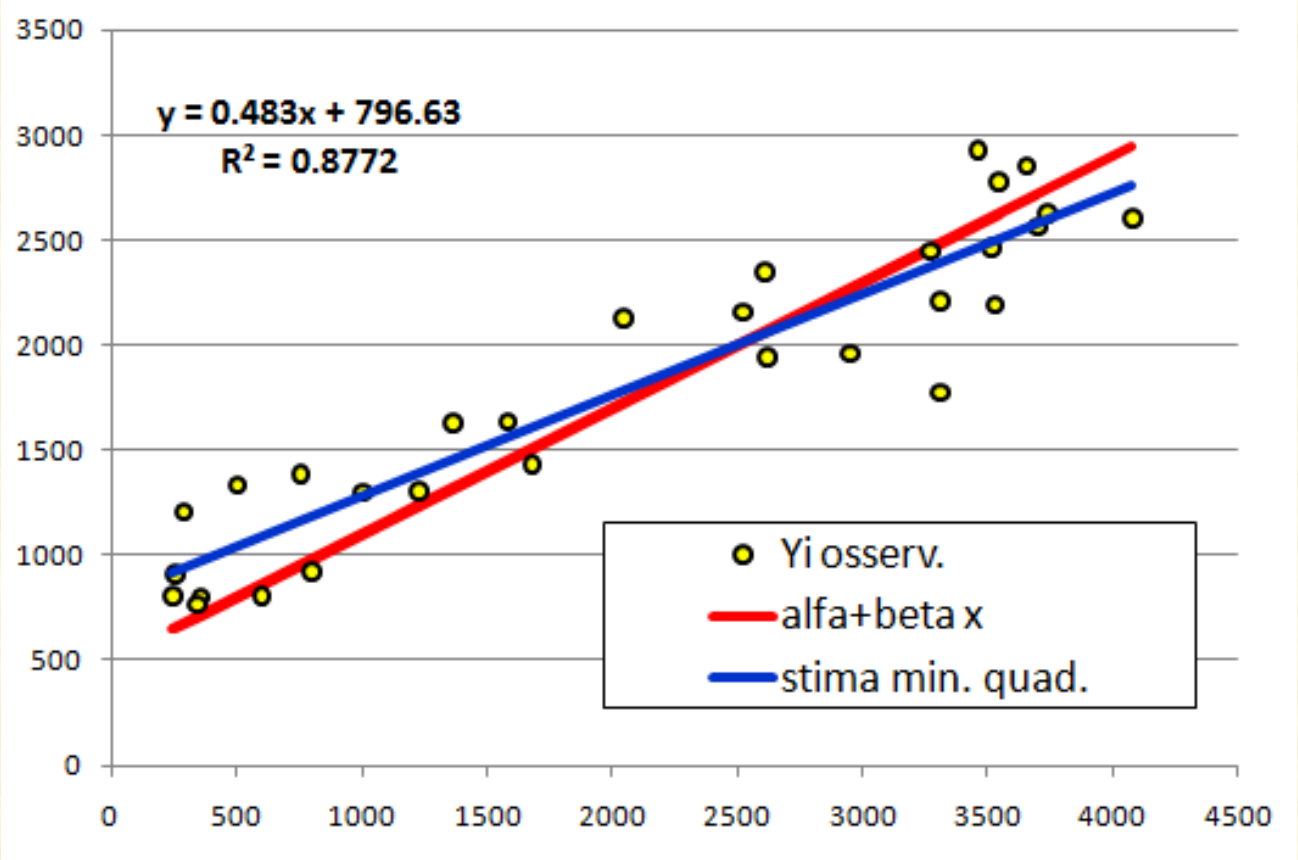
l'Output completo dell'analisi di regressione in Excel (si veda strumento_regressione_excel.pdf)

I parametri della retta dei m.q.

	A	B	C	D	E	F	G
1							
2	OUTPUT RIEPILOGO						
3							
4	*Statistica della regressione*						
5	R multiplo	0.936604					
6	R al quadrato	0.877228					
7	R al quadrato corretto	0.872843					
8	Errore standard	248.1987					
9	Osservazioni	30					
10							
11	ANALISI VARIANZA						
12		*gdl*	*SQ*	*MQ*	*F*	*Significatività F*	
13	Regressione	1	12324510.3	12324510.3	200.0647312	2.80751E-14	
14	Residuo	28	1724873.176	61602.61344			
15	Totale	29	14049383.47				
16							
17		*Coefficienti*	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*
18	Intercetta	796.6347	84.68062667	9.407520655	3.64146E-10	623.1743463	970.0951
19	Variabile X 1	0.482991	0.034147111	14.14442403	2.80751E-14	0.413044036	0.552938
20							

I PARAMETRI DELLA RETTA DEI MINIMI QUADRATI CON EXCEL

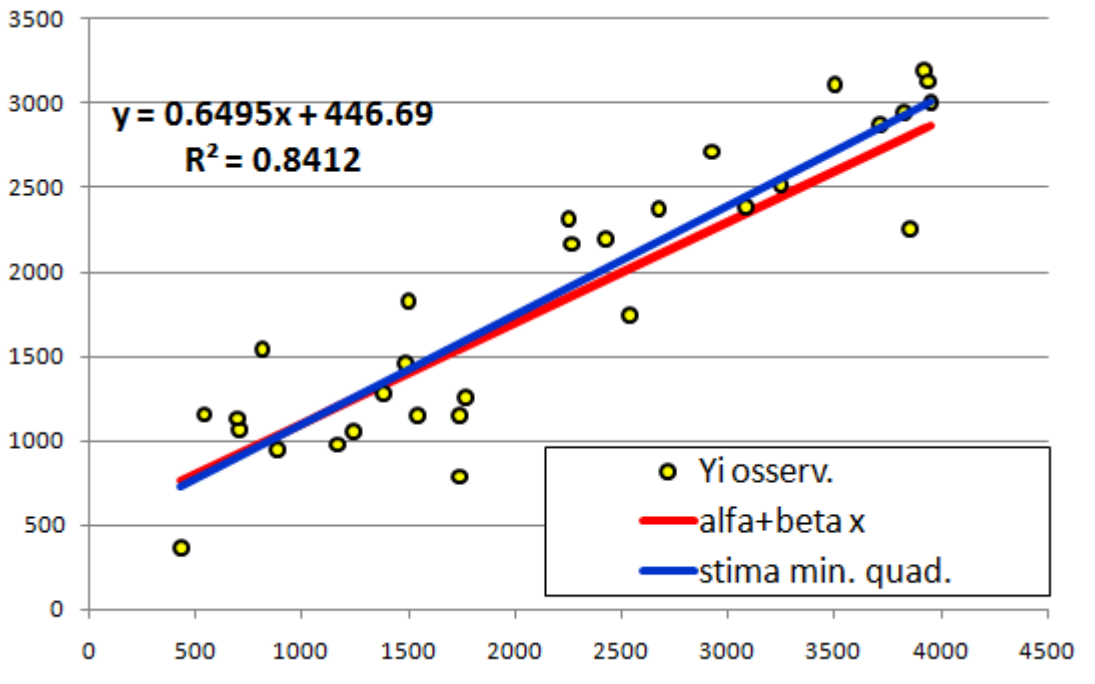
Come è andata questa volta (con questo campione)?



La retta "vera" (ma nota solo al folletto onnisciente) è quella rossa, la blu è la nostra stima

I PARAMETRI DELLA RETTA DEI MINIMI QUADRATI CON EXCEL

Come poteva andare (con altri campioni)?



Cliccando il tasto F9 nel foglio simulazione possiamo generare infiniti campioni con cui stimare α e β (la retta rossa sempre uguale)

VALORI TEORICI

La funzione lineare dei valori x_i (secondo i parametri a e b calcolati in base ai valori X e Y osservati) rappresenta la retta dei minimi quadrati:

$$\hat{y}_i = a + b\,x_i \qquad (i = 1, 2, \ldots, n)$$

ove con $\hat{y}_i$ si indicano i valori teorici, o valori stimati, della variabile dipendente.

Modello di regressione lineare semplice

L'equazione della retta dei m.q.
fornisce una *stima* della retta di regressione

Stima (previsione) del valore di Y per l'osservazione *i*

Stima dell'intercetta α

Stima del coefficiente angolare β (pendenza)

Valore di X per l'osservazione *i*

$$\hat{Y}_i = a + b X_i$$

RESIDUI: DEFINIZIONE

I residui sono definiti come la differenza tra i valori osservati y_i ed i corrispondenti valori teorici che si collocano sulla retta di regressione:

$$e_i = y_i - \hat{y}_i \qquad (i = 1, 2, \ldots, n)$$

RESIDUI: INTERPRETAZIONE

Ciascun residuo è dunque il valore numerico, riferito all'unità *i*-esima, in eccesso o in difetto, rispetto al corrispondente valore osservato, che non è "spiegato" dalla relazione lineare con la variabile indipendente.

ESEMPIO

Determiniamo la retta di regressione della spesa mensile per alimenti in funzione del reddito mensile

$$a = 796{,}63 \qquad b = 0{,}483$$

$$\hat{y}_i = 796{,}63 + 0{,}483 x_i$$

La conoscenza della retta dei minimi quadrati consente di stimare i valori della spesa in corrispondenza di ciascun valore del reddito. Ad esempio per la prima famiglia, con reddito di 3529 euro, si ottiene:

796,63+ 0,483 * 3529 = 2195

INTERPRETAZIONE

- In questo caso, il valore della costante *a* ha semplicemente un significato geometrico (l'ordinata all'origine); esso indicherebbe la spesa media (teorica) d'una famiglia con un reddito nullo.
- Il coefficiente *b* indica che, all'aumentare del reddito di 1000 euro, la spesa annua aumenta in media di circa 483 euro.

Bontà di adattamento

I minimi quadrati ci garantiscono il miglior adattamento possibile, ma siamo interessati a quantificare il grado di scostamento tra valori stimati e valori osservati

La verifica della validità o bontà di adattamento della retta di regressione è diretta a controllare che la retta di regressione sia realmente in grado di spiegare l'andamento delle osservazioni, in quanto **si può sempre adattare una retta con il metodo dei minimi quadrati anche nei casi in cui i punti non seguono una relazione lineare** ed in queste circostanze la retta di regressione ha una capacità minima, o nulla, di riassumere la relazione tra le variabili.

SCOMPOSIZIONE DELLA DEVIANZA

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

$$DEV(Y) = DEV(\hat{Y}) + DEV(E)$$

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

devianza totale dei valori della variabile dipendente

Misura la variazione dei valori di Y intorno alla loro media

$$DEV(\hat{Y}) = \sum_{i=1}^{n} (\hat{y}_i - M_y)^2$$

devianza dei valori stimati: devianza di regressione

Variazione spiegata attribuibile alla relazione fra la X e Y

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

devianza dei residui:

devianza residua

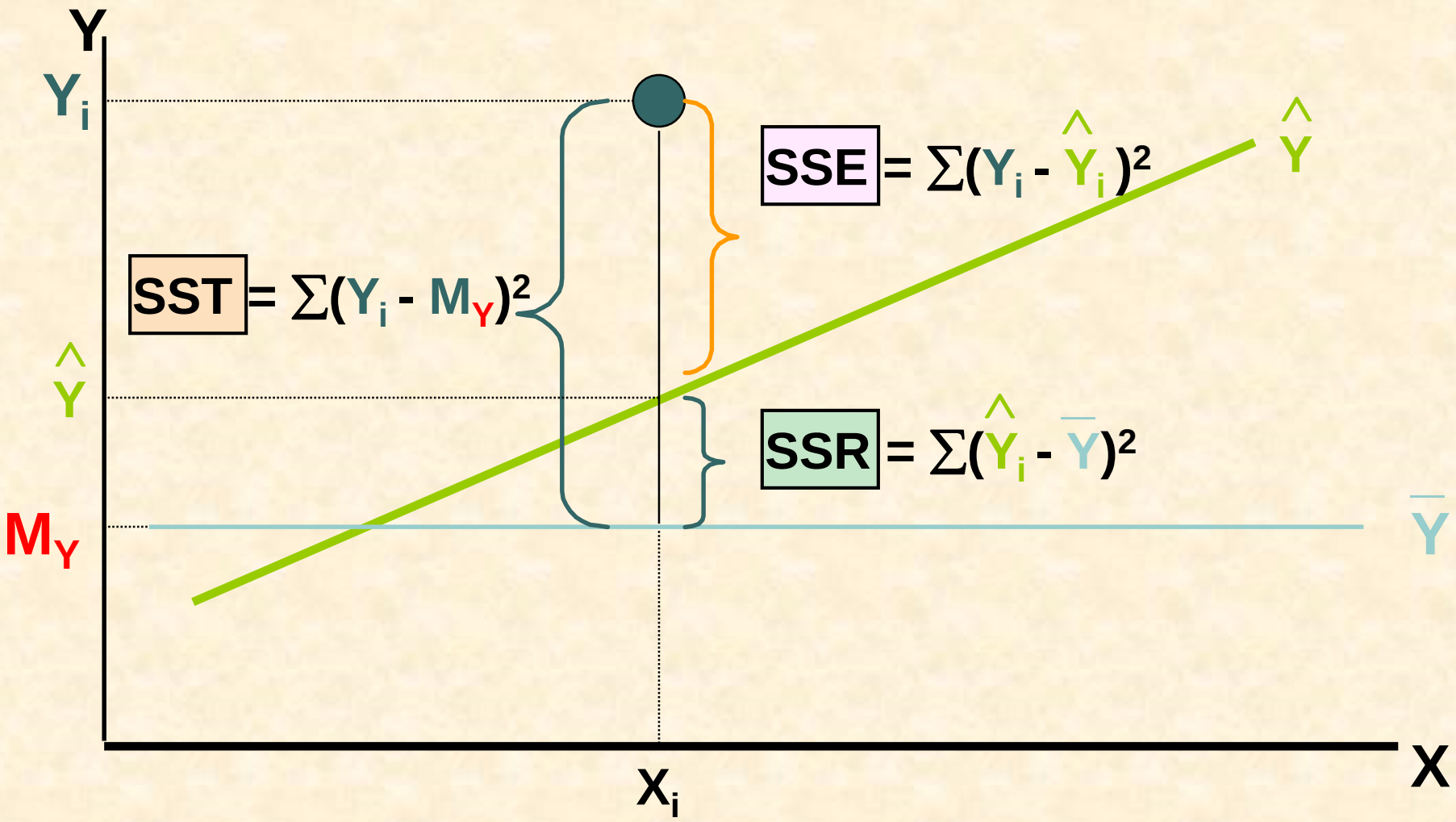
Variazione attribuibile a fattori estranei alla relazione fra la X e Y

OSSERVAZIONE

La devianza DEV(Y) dei valori osservati della variabile Y (che misura la variabilità degli scarti dei valori osservati dalla media) è il risultato del contributo di due componenti:

1. **la prima $DEV(\hat{Y})$ la devianza di regressione che misura la variabilità degli scarti tra i valori stimati (sulla retta dei minimi quadrati) e la media;**
2. **la seconda $DEV(E)$ è la devianza residua, che misura la variabilità dei residui, ovvero degli scarti tra i valori osservati (scatter di punti) e i corrispondenti valori teorici (sulla retta dei minimi quadrati)**

SCOMPOSIZIONE DELLA DEVIANZA



MISURA DELLA BONTA' DI ADATTAMENTO

La devianza di regressione è quella parte della devianza totale che è "spiegata" dalla relazione lineare con la variabile indipendente. Per misurare la bontà di adattamento si deve rapportare alla devianza totale, in quanto il suo valore numerico è influenzato dall'ordine di grandezza e dall'unità di misura della variabile dipendente e dal numero di osservazioni

Una misura relativa (e normalizzata) è l' indice di determinazione lineare che si indica con R^2 (*r-squared*) ed è il rapporto tra la devianza di regressione e la devianza totale:

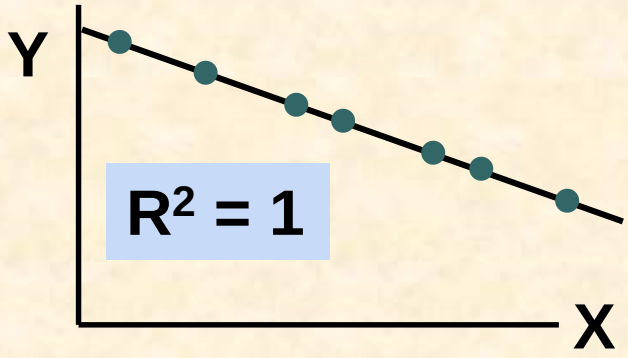
$$R^2 = \frac{DEV(\hat{Y})}{DEV(Y)} = 1 - \frac{DEV(E)}{DEV(Y)}$$

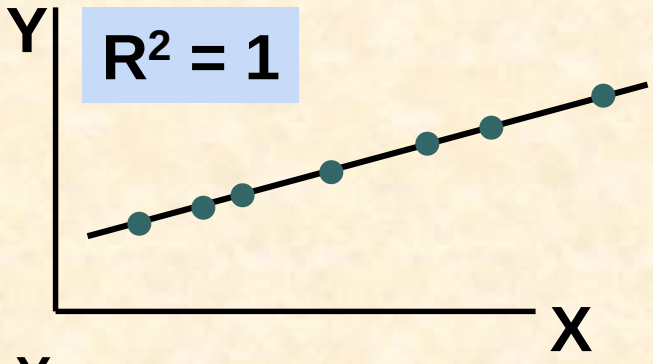
L'indice R^2, essendo un rapporto d'una parte al tutto, può assumere valori compresi tra 0 ed 1:

se $R^2 = 0$ l'adattamento è pessimo

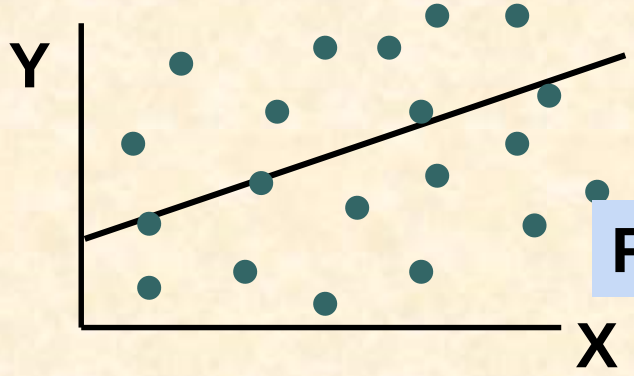
se $R^2 = 1$ l'adattamento è perfetto

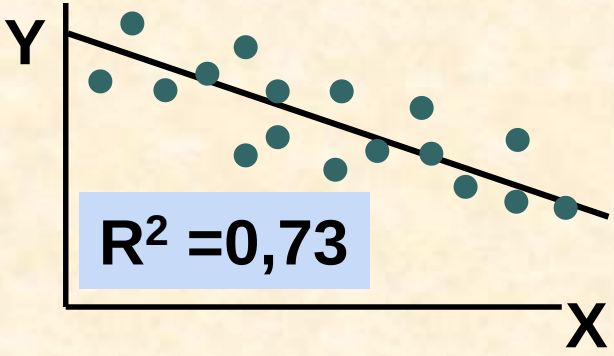
Esempi di R^2





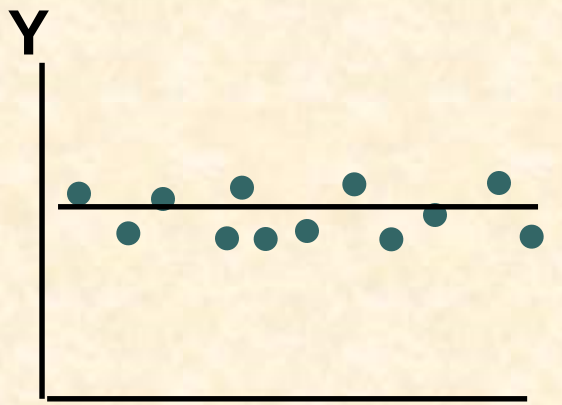
Relazione lineare perfetta fra X e Y:
Il 100% della variabilità di Y è spiegata dalla variabilità di X





Solo una parte della variabilità di Y è spiegata dalla variabilità di X

Esempi di R^2



$R^2 = 0$

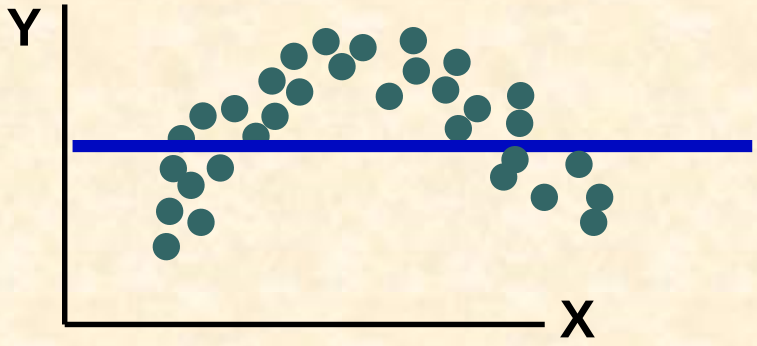
$R^2 = 0$

Nessuna relazione lineare fra X e Y:

Il valore di Y non dipende da X. (Nessuna variazione di Y è spiegata da X)

OSSERVAZIONI

1. $R^2=0$ e $R^2=1$ rappresentano dei casi limite. In pratica, si ha un indice di determinazione lineare interno all'intervallo [0, 1]
2. L'indice R^2 non misura *se* c'è una relazione tra le 2 variabili, ma solo quanto i dati osservati possano essere approssimati **da una retta:** se l'indice di determinazione lineare si rivela prossimo ad 1, si può dire che la variabilità di *Y* è "spiegata" in misura notevole dalla retta di regressione.



Fra X e Y sussiste una relazione, ma non è di tipo lineare: R^2 prossimo allo 0

l'Output completo dell'analisi di regressione in Excel (si veda strumento_regressione_excel.pdf)

La scomposizione della devianza e l'indice R^2

$$R^2 = \frac{SSR}{SST} = \frac{12324510}{14049383} = 0{,}87728$$

	A	B	C	D			
1							
2	OUTPUT RIEPILOGO						
3							
4	*Statistica della regressione*						
5	R multiplo						
6	R al quadrato	0.877228					
7	R al quadrato corretto						
8	Errore standard	248.1987					
9	Osservazioni	30					
10							
11	ANALISI VARIANZA						
12		*gdl*	*SQ*	*MQ*	*F*	*Significatività F*	
13	Regressione	1	12324510.3	12324510.3	200.0647312	2.80751E-14	
14	Residuo	28	1724873.176	1602.61344			
15	Totale	29	14049383.47				
16							
17		*Coefficienti*	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*
18	Intercetta	796.6347	84.68062667	9.407520655	3.64146E-10	623.1743463	970.0951
19	Variabile X 1	0.482991	0.034147111	14.14442403	2.80751E-14	0.413044036	0.552938
20							

L'indice R2

Dev. Regressione
Dev. Residua
Dev. totale

L'87,73% della variabilità della variabile dipendente è spiegata dalla variazione della variabile indipendente

Errore standard della stima

Ciascun residuo e_i può essere considerato una stima dell'errore ε_i

Si può dimostrare che lo scarto quadratico medio dei residui (corretto, usando il denominatore n-2) è una stima corretta della deviazione standard degli ε_i

$$S_{YX} = \sqrt{\frac{\sum_{i=1}^{n} e_i^{\,2}}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{DEV(E)}{n-2}}$$

l'Output completo dell'analisi di regressione in Excel (si veda strumento_regressione_excel.pdf)

L'errore standard della stima

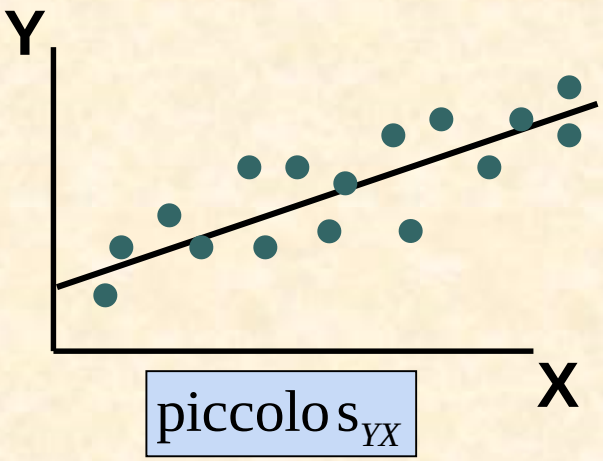
	A	B	C	D	E	F	G
1							
2	OUTPUT RIEPILOGO						
3							
4	*Statistica della regressione*						
5	R multiplo	0.936604					
6	R al quadrato	0.877228					
7	R al quadrato corretto						
8	Errore standard	248.1987					
9	Osservazioni	30					
10							
11	ANALISI VARIANZA						
12		*gdl*	*SQ*	*MQ*	*F*	*Significatività F*	
13	Regressione	1	[illegible]	12324510.3	200.0647312	2.80751E-14	
14	Residuo	28	1724873.176	61602.61344			
15	Totale	29	14049383.47				
16							
17		*Coefficienti*	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*
18	Intercetta	796.6347	84.68062667	9.407520655	3.64146E-10	623.1743463	970.0951
19	Variabile X 1	0.482991	0.034147111	14.14442403	2.80751E-14	0.413044036	0.552938
20							

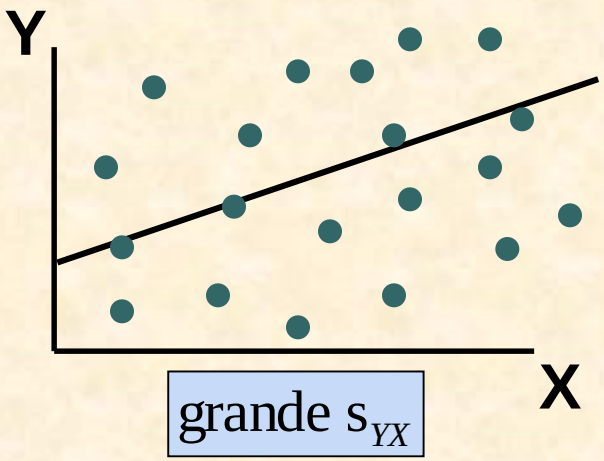
Syx

$$\sqrt{\frac{1724873}{30-2}} = 248{,}2$$

Un confronto fra Errori Standard

S_{YX} misura la variabilità dei valori osservati di Y intorno alla retta di regressione





L'ordine di grandezza di S_{YX} dovrebbe sempre essere giudicato in relazione all'ordine di grandezza della variabile dipendente Y

i.e., S_{YX} = 248,2 (euro) è moderatamente piccolo in relazione ai valorri della spesa che oscillano fra 500 e 3000

Inferenza sul coefficiente angolare :

La prima e più importante verifica riguarda l'esistenza o meno di una relazione tra la "Y" e la "X“ : la variabile Y varia linearmente al variare di X?

Inferenza sul coefficiente angolare :

Equazione di regressione stimata:

$$\text{spesa} = 797 + 0{,}483 \cdot \text{reddito}$$

La stima della pendenza è 0,483

Il reddito di una famiglia influenza la spesa?
(ovvero: posso dire dai risultati campionari – con un ragionevole margine di certezza – che β non sia 0?)

Sarei portato a rispondere: “certo che la influenza! la mia stima di β è 0,483 (non è 0), ovvero stimo che spendo circa 48 centesimi ogni euro guadagnato”.

Ma sarebbe un’argomentazione errata…

INFERENZA SUL COEFFICIENTE ANGOLARE

Prendiamo i seguenti esempi:

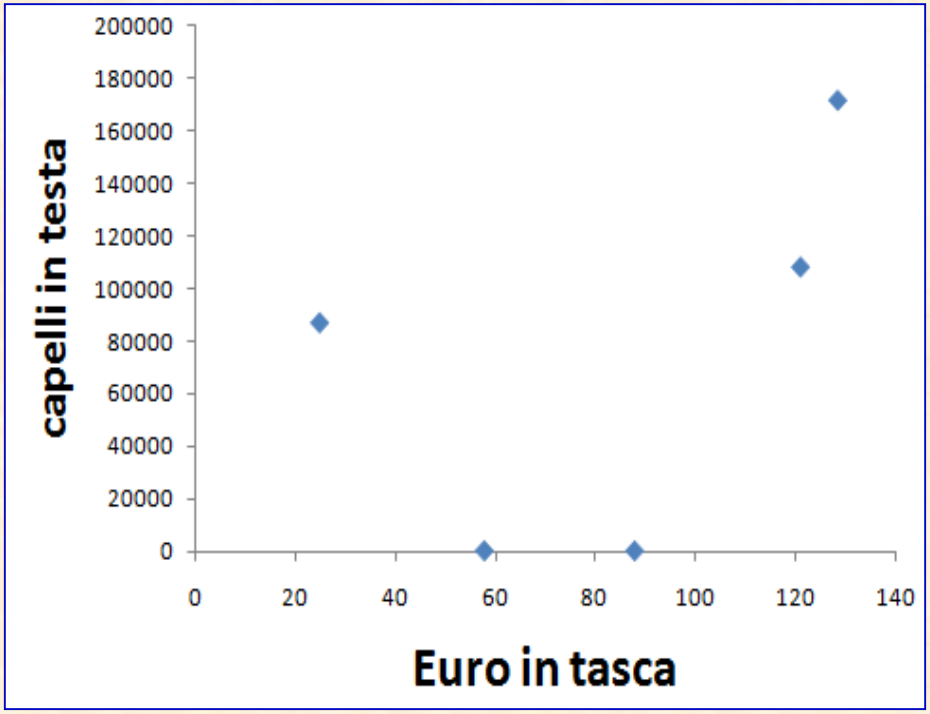
1. conto i soldi in tasca di 5 miei amici e il numero di capelli di ciascuno
2. Conto il numero di tornei vinti e il career prizemoney (in migliaia di $) dei primi 20 tennisti al mondo (dati 16/3/2011)

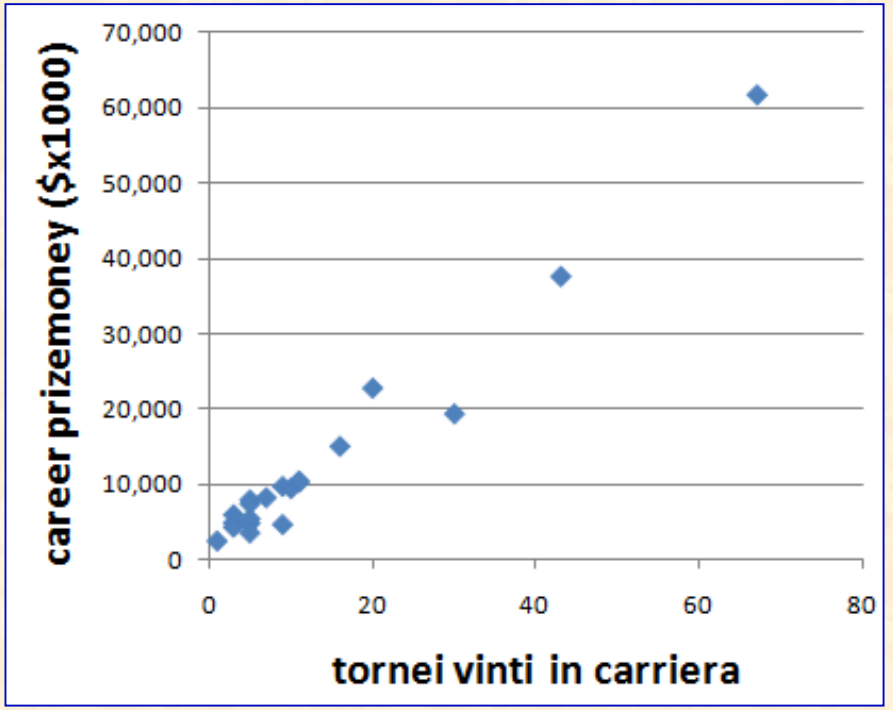
	euro in tasca	numero di capelli
Ugo	120.96	108293
Calvino	57.6	2
Piero	24.6	87052
Irsutino	128.42	171870
Pelatino	87.72	0

		titles	prizemoney
1	Nadal	43	37,685
2	Federer	67	61,839
3	Djokovic	20	22,851
4	soderling	9	9,773
5	Murray	16	15,090
6	Ferrer	11	10,294
7	Berdych	5	7,967
8	Roddick	30	19,427
9	Verdasco	5	7,455
10	Melzer	3	6,008
11	Monfils	3	4,924
12	Almagro	9	4,688
13	Youzhny	7	8,275
14	Wawrinka	3	4,350
15	Fish	5	4,846
16	Ljubicic	10	9,477
17	Tsonga	5	5,441
18	troick	1	2,493
19	Nalbandia	11	10,481
20	Cilic	5	3,583

INFERENZA SUL COEFFICIENTE ANGOLARE

Diamo un'occhiata agli scatter:

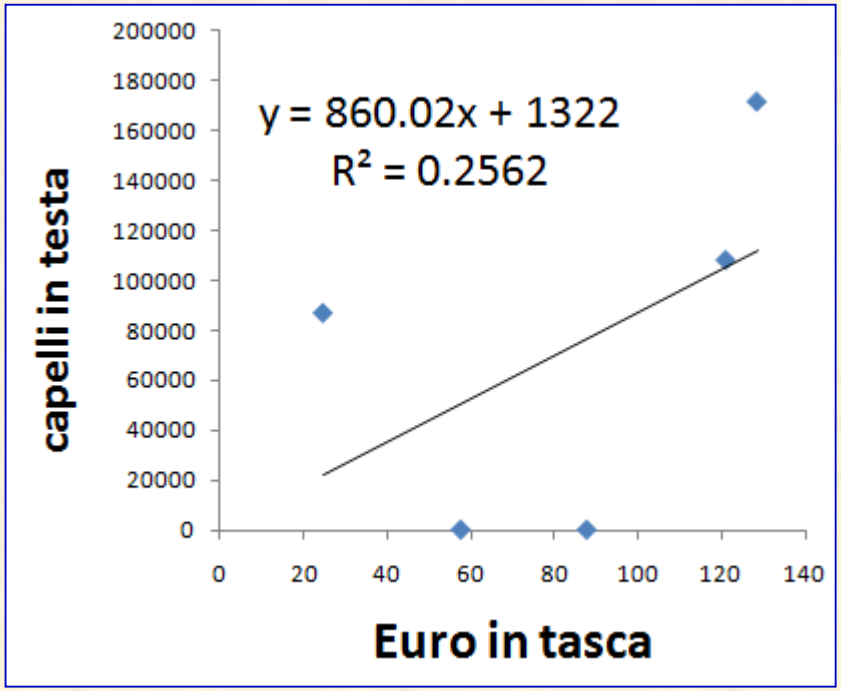


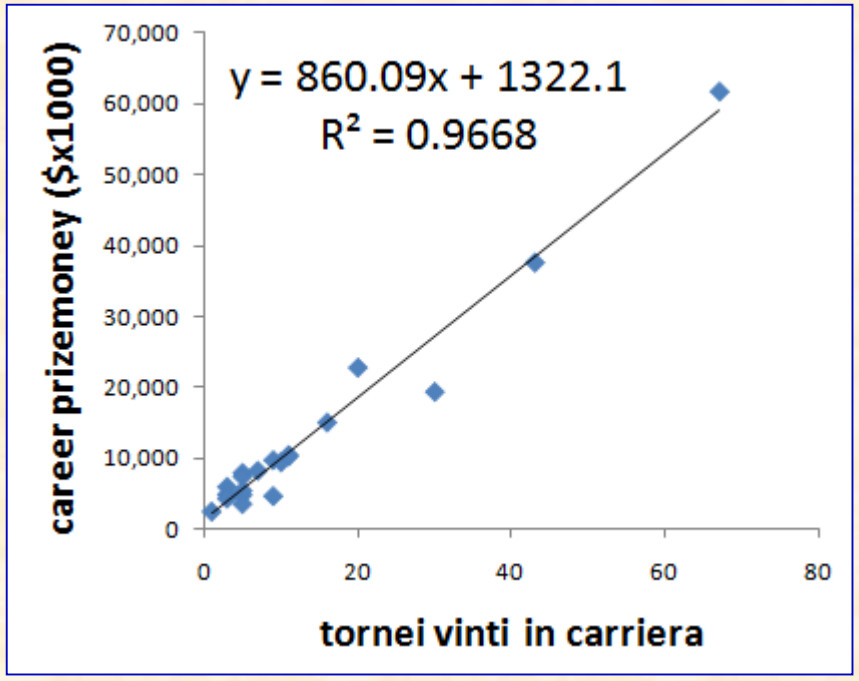


Cosa vediamo? Quello che ci aspettavamo: non "si vede" niente nel primo grafico, mentre nel secondo è evidente una disposizione dello scatter attorno a una retta immaginaria, però…

INFERENZA SUL COEFFICIENTE ANGOLARE

Però… se provo a stimare le rette dei minimi quadrati…

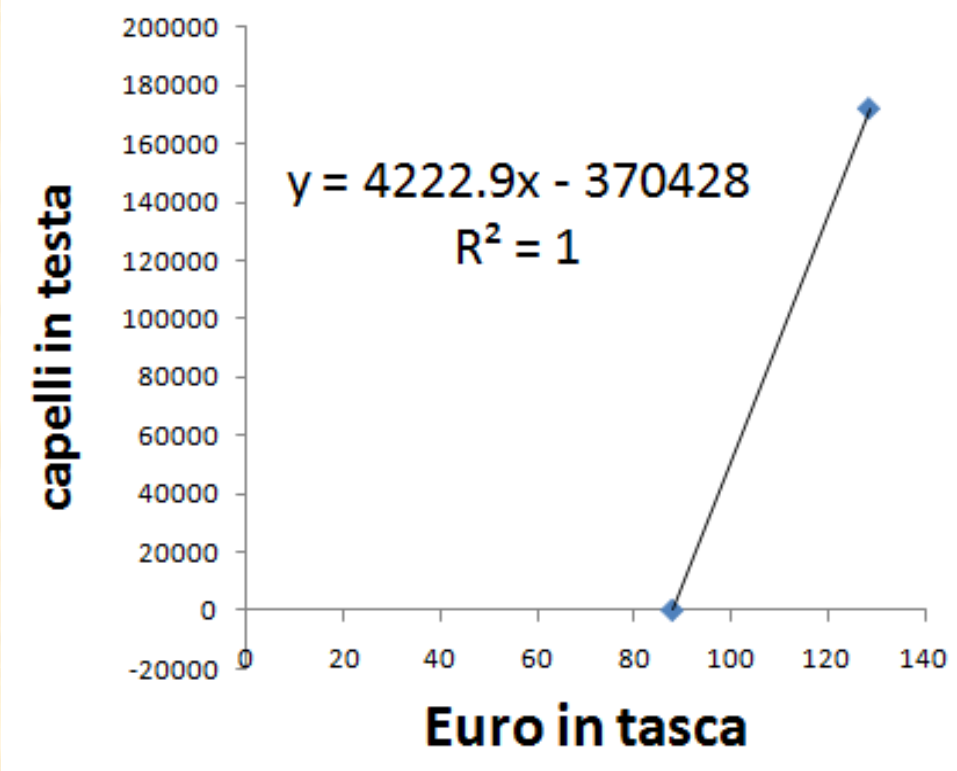




Orrore! È la stessa retta! All'aumentare degli euro in tasca aumenta come per magia il numero dei capelli! Abbiamo forse trovato la cura (recarsi al Bancomat) per la calvizie? Come mai Cesare Ragazzi non ci aveva pensato? Ovviamente le cose non stanno così. L'indice R2 mi dice già qualcosa ma basterà?

INFERENZA SUL COEFFICIENTE ANGOLARE

No, R^2 non basta, se infatti conto Euro e capelli solo a Irsutino e Pelatino…

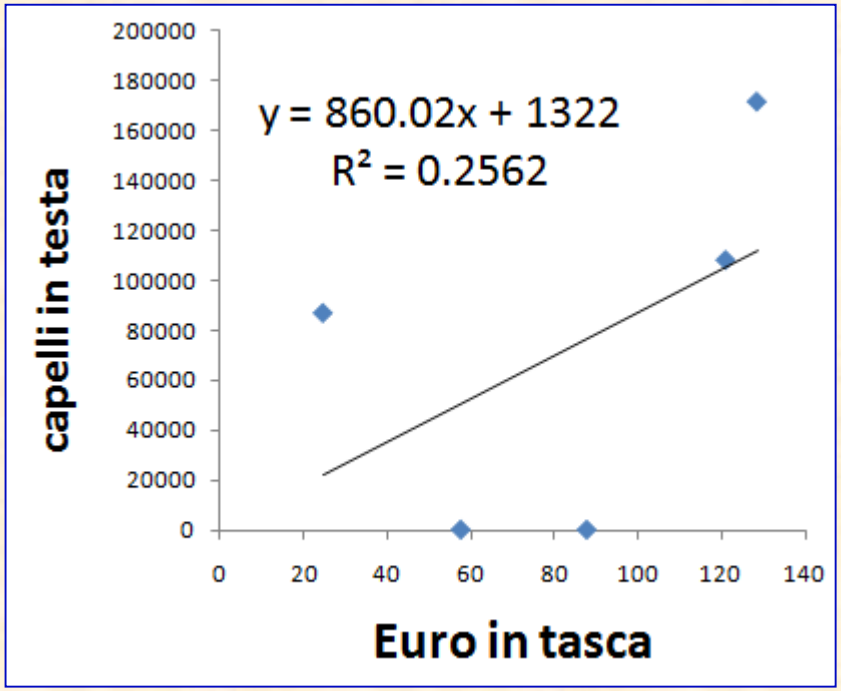


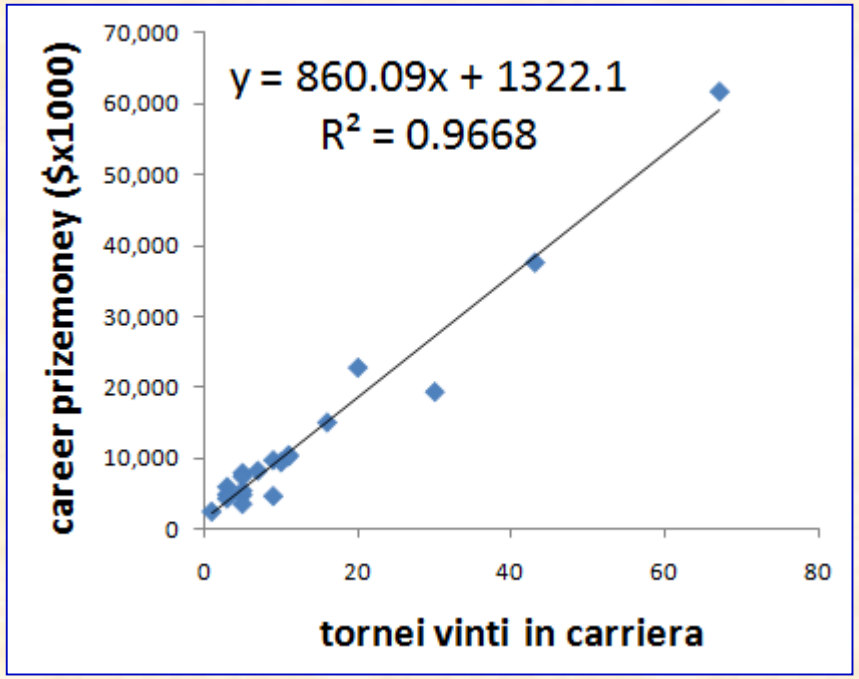
In questo caso addirittura R2=1 (con una diversa retta dei minimi quadrati)! , Infatti, con solo due punti la retta dei minimi quadrati ha (ovviamente) un adattamento perfetto (passa proprio da quei due punti).

Conclusione: **R2 non è sufficiente**!

INFERENZA SUL COEFFICIENTE ANGOLARE

Dov'è dunque la differenza dei due esempi?





Risiede nella maggiore "stabilità" del secondo esempio: i parametri stimati del primo esempio sono gli stessi, ma danno l'impressione di essere più "ballerini", più variabili: basterebbe aggiungere un'osservazione (rilevare euro in tasca e capelli in testa di un altro tizio) e avremmo probailmente una retta dei m.q. profondamente differente

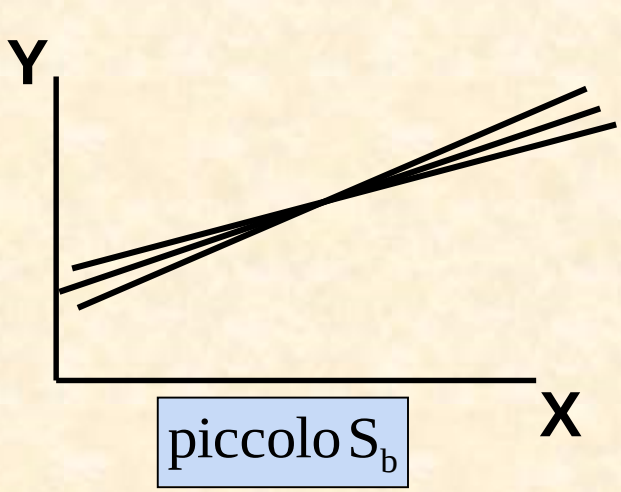
Test sul coefficiente angolare

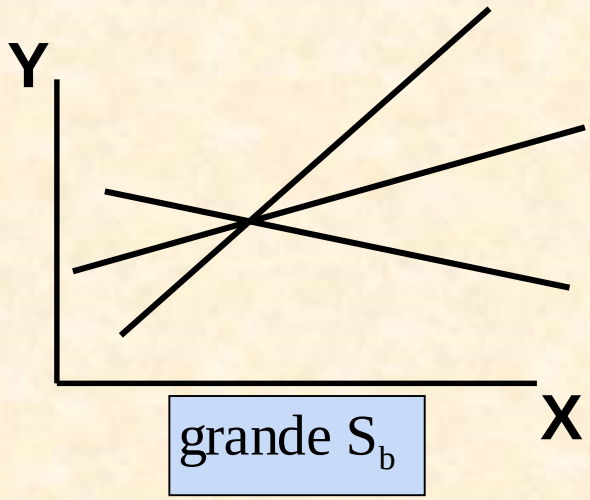
L'Errore Standard della stima del coefficiente angolare (b) è stimato da:

$$S_b = \frac{S_{YX}}{\sqrt{DEV(X)}} = \frac{S_{YX}}{\sqrt{\sum (X_i - \overline{X})^2}}$$

Significato dell'errore standard del coefficiente angolare

S_{b_1} è una misura della variazione nella pendenza di rette di regressione da possibili differenti campioni





l'Output completo dell'analisi di regressione in Excel (si veda strumento_regressione_excel.pdf)

L'errore standard del coefficiente angolare

	A	B	C	D	E	F	G
1							
2	OUTPUT RIEPILOGO						
3							
4	*Statistica della regressione*						
5	R multiplo	0.936604					
6	R al quadrato	0.877228					
7	R al quadrato corretto	0.872843					
8	Errore standard	248.1987					
9	Osservazioni	30					
10							
11	ANALISI VARIANZA						
12		*gdl*	*SQ*	*MQ*	*F*	*Significatività F*	
13	Regressione	1	12324510.3	12324510.3	200.0647312	2.80751E-14	
14	Residuo	28	1724873.176	61602.61344			
15	Totale	29	14049383.47				
16							
17		*Coefficienti*	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*
18	Intercetta	796.6347		9.407520655	3.64146E-10	623.1743463	970.0951
19	Variabile X 1	0.482991	0.034147111	14.14442403	2.80751E-14	0.413044036	0.552938
20							

$S_{yx}=0,0341$

Test sul coefficiente angolare : Test t

Test t per il coefficiente angolare della popolazione

ESISTE O NON ESISTE UNA RELAZIONE LINEARE TRA Y ed X?

Ipotesi nulla e alternativa

- H_0: $\beta = 0$ (non c'è relazione lineare)
- H_1: $\beta \neq 0$ (c'è una relazione lineare)

Test statistico

$$t = \frac{b - \beta}{S_b}$$

$$d.f. = n - 2$$

dove:

b = stima della pendenza

β= valore del coefficiente ipotizzato in H_0

S_b = Errore standard del coefficiente di regressione

Test sul coefficiente angolare : Test t: Esempio

$H_0: \beta = 0$

$H_1: \beta \neq 0$

Dall'output di Excel:

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*
Intercetta	796.634744	84.6806	9.4075	3.6415E-10
Variabile X 1	0.482991	0.0341	14.1444	2.8075E-14

$$t = \frac{b - \beta}{S_b} = \frac{0{,}483 - 0}{0{,}0341} = 14{,}144$$

Test sul coefficiente angolare : Test t: Esempio

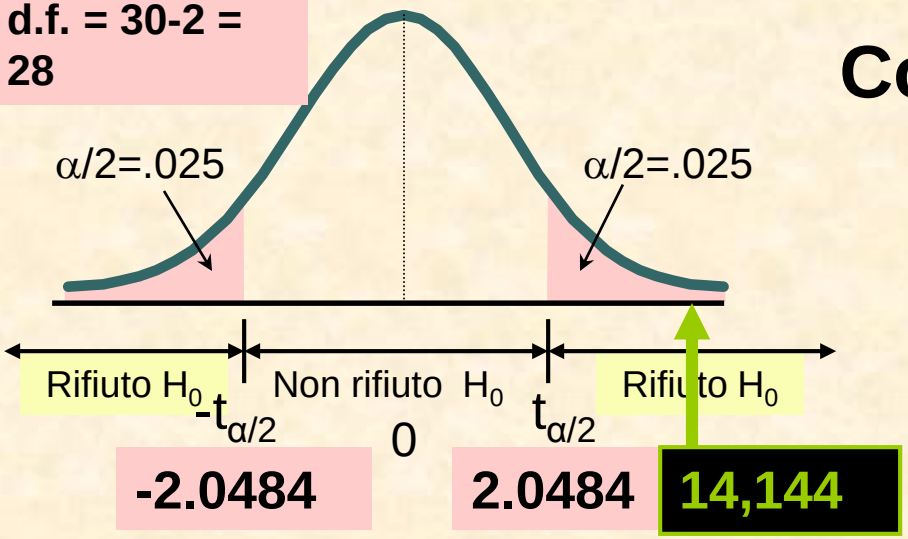
Test Statistic: **t = 14,144**

H_0: $\beta = 0$

H_1: $\beta \neq 0$

Decisione: Rifiuto H_0

d.f. = 30-2 = 28



Conclusioni:

Non ci sono elementi per ritenere che il reddito non influenzi il consumo: rifiuto H0, in quanto alla luce del campione è "troppo inverosimile"