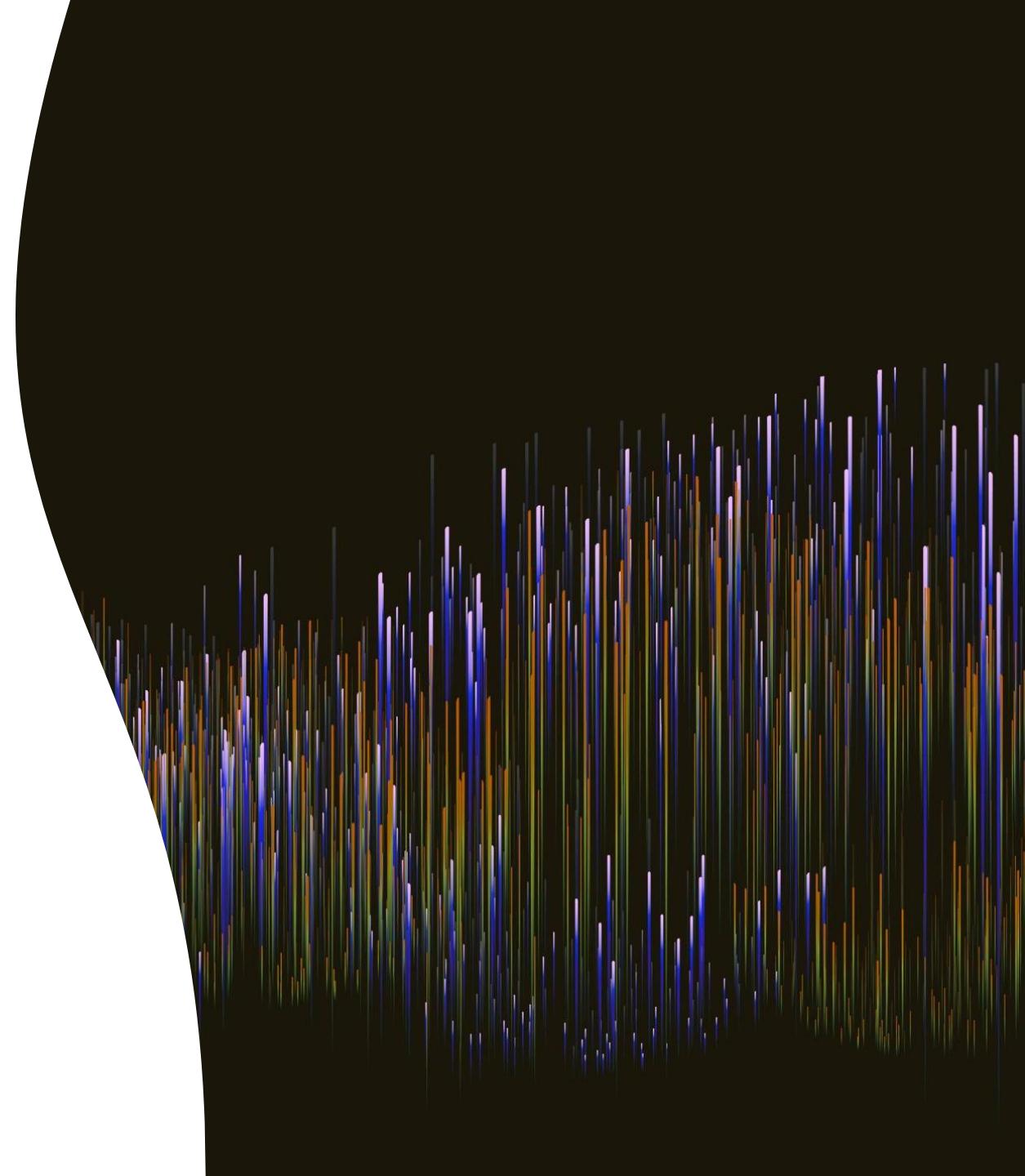


Regressione Lineare



Tipici problemi nel business



Quanto le mie vendite (Y) sono influenzate del prezzo unitario (x) praticato? (ricerca **esplicativa**)

Quanti pezzi mi venderò se pratico un prezzo unitario di 10 Euro?
(ricerca **predittiva**)

Quanto le mie vendite (Y) sono influenzate della promozione pubblicitaria (x_1) e del prezzo unitario (x_2) praticato? (ricerca **esplicativa**)

Quanti pezzi venderò se pratico un prezzo unitario di 10 Euro e spendo 3000 Euro in promozione pubblicitaria? (ricerca **predittiva**).



Si può usare la regressione lineare.

Introduzione



Si può essere interessati ad capire se e come varia una variabile (**variabile dipendente**, in genere rappresentata con y) al variare di una o più (**variabili indipendenti o esplicative**, generalmente rappresentate con x oppure x_1, x_2 , ecc.), individuando un'opportuna **funzione analitica** che rappresenti tale relazione.

Nel caso di una sola variabile indipendente si parla di **regressione semplice**.

In presenza di due o più variabili indipendenti si parla di **regressione multipla**.

La scelta del ruolo delle variabili (quale è la dipendente e quali le esplicative) è ovviamente extra-statistica.

Notazioni

Indichiamo con y la variabile *dipendente* (o *variabile risposta*) e con x la *variabile esplicativa oppure: indipendente, regressore, predittore*.

Quando ci riferiamo ai dati usiamo:

x_i e y_i per indicare i valori osservati sulla generica unità i osservata con $i = 1, 2, \dots, n$, dove n è il numero delle unità osservate.

Un esempio di funzione lineare

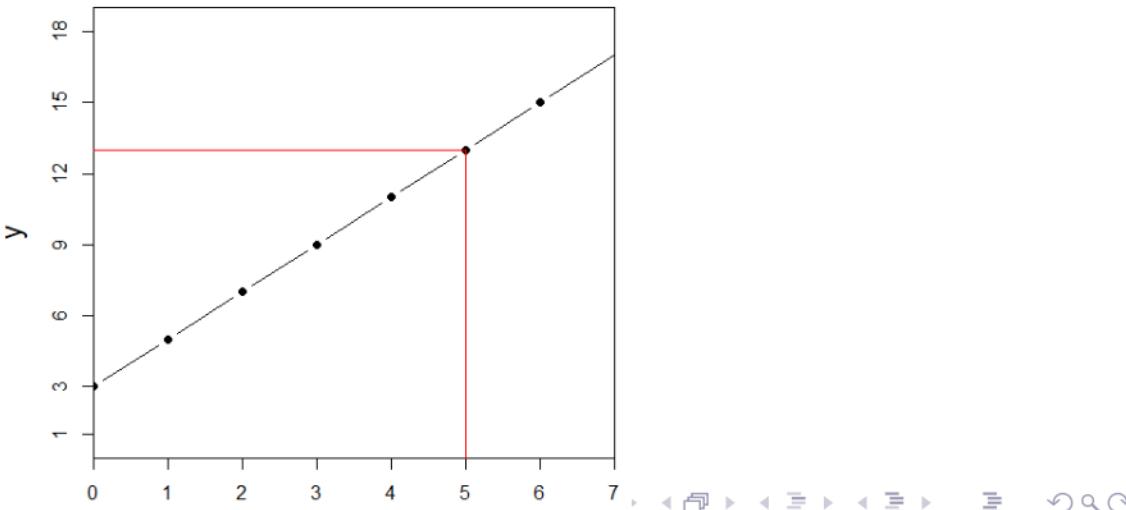


La funzione matematica $y = \beta_0 + \beta_1 x$ esprime le osservazioni di y come una **funzione lineare** delle osservazioni di x .

Tale formula definisce una retta con **pendenza** (inclinazione o coefficiente angolare) β_1 e **intercetta sull'asse y** , β_0 .

La formula $y = 3 + 2x$ è una funzione lineare del tipo $y = \beta_0 + \beta_1 x$ dove $\beta_0 = 3$ e $\beta_1 = 2$. Ciascun numero x sostituito nella formula produce un determinato valore di y . Se $x = 5 \rightarrow y = 3 + 2 \times 5 = 13$.

L'asse orizzontale x contiene tutti i possibili valori di x e l'asse verticale y tutti i possibili valori di y . I due assi si intersecano nel punto $x = 0, y = 0$, chiamato *origine*.



Interpretazione dell'intercetta e della pendenza 1/2

- Se $x = 0 \Rightarrow y = \beta_0 + (\beta_1 \times 0) = \beta_0$. Allora β_0 è il valore di y quando $x = 0$. Per questo motivo è detto intercetta: la retta intercetta (interseca) l'asse y nel punto di coordinate $(0, \beta_0)$.
- La pendenza β_1 esprime la variazione di y per incrementi unitari di x cioè per due valori di x che si differenziano di 1 (ad esempio $x=0$ e $x=1$, $x=6$ e $x=7$, ecc.). Infatti poniamo:

$$x+1 \Rightarrow y_{(1)} = \beta_0 + \beta_1(x+1) = \beta_0 + \beta_1 x + \beta_1$$

$$x \Rightarrow y_{(0)} = \beta_0 + \beta_1 x$$

$$\text{da cui } y_{(1)} - y_{(0)} = \beta_0 + \beta_1 x + \beta_1 - (\beta_0 + \beta_1 x) = \beta_1$$

Per due valori di x che differiscono di 10 unità (es. $x = 4, x = 14$) abbiamo:

$$x = 14 \Rightarrow \beta_0 + 14\beta_1$$

$$x = 4 \Rightarrow \beta_0 + 4\beta_1$$

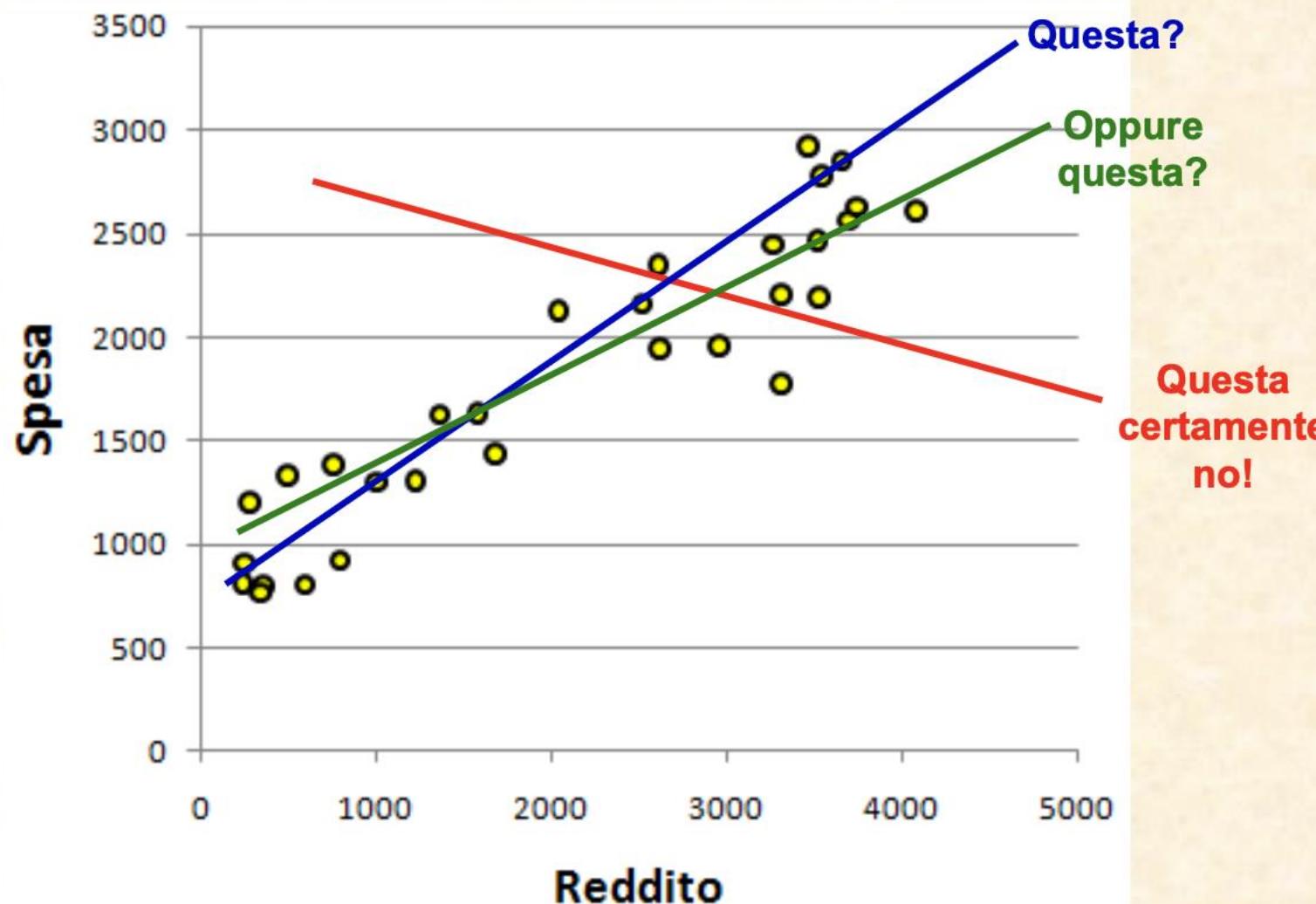
da cui

$$\beta_0 + 14\beta_1 - (\beta_0 + 4\beta_1) = (14 - 4)\beta_1 = 10\beta_1$$

Interpretazione dell'intercetta e della pendenza 2/2

- ▶ Se il caso $x = 0$ è illogico allora l'intercetta β_0 ha un significato puramente geometrico . Ad esempio: se x è la superficie di un appartamento e y è il suo valore commerciale, il caso $x = 0$ è illogico.
- ▶ Se $\beta_1 > 0$, quando x aumenta, y aumenta: la retta va verso l'alto e si dice che la relazione tra le due variabili è positiva o concordante.
Se $\beta_1 < 0$, quando x aumenta, y diminuisce (ovvero quando x diminuisce, y aumenta): la retta va verso il basso e si dice che la relazione tra le due variabili è negativa o discordante.
Quando $\beta_1 = 0$, il grafico è una retta orizzontale in corrispondenza $y = \beta_0$.

Quale retta?

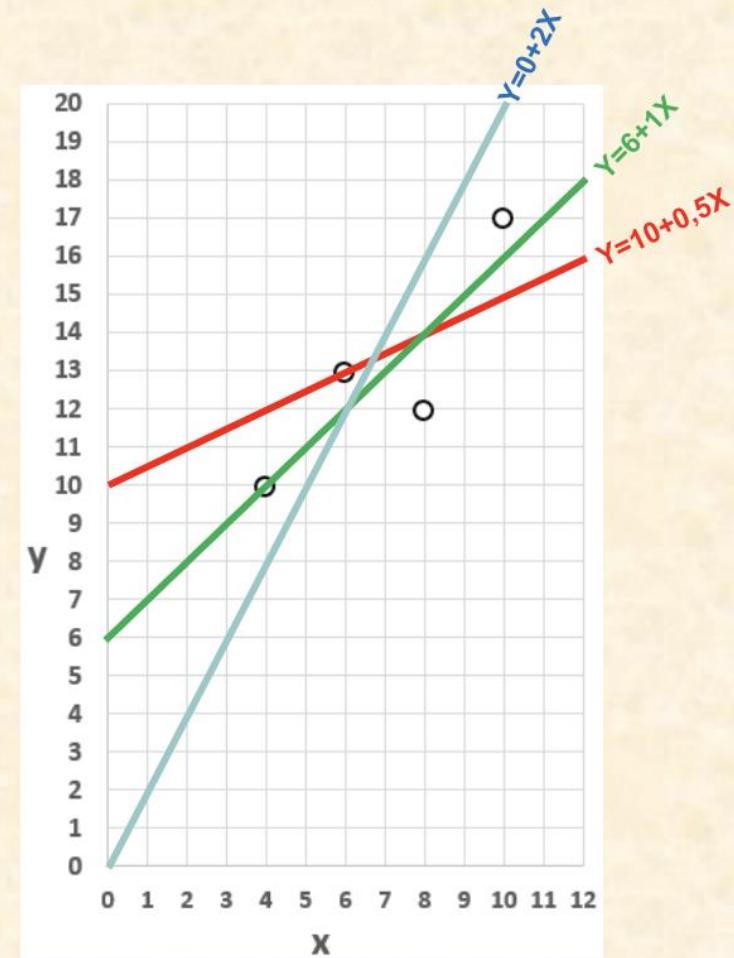


x	y
4	10
6	13
8	12
10	17

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...

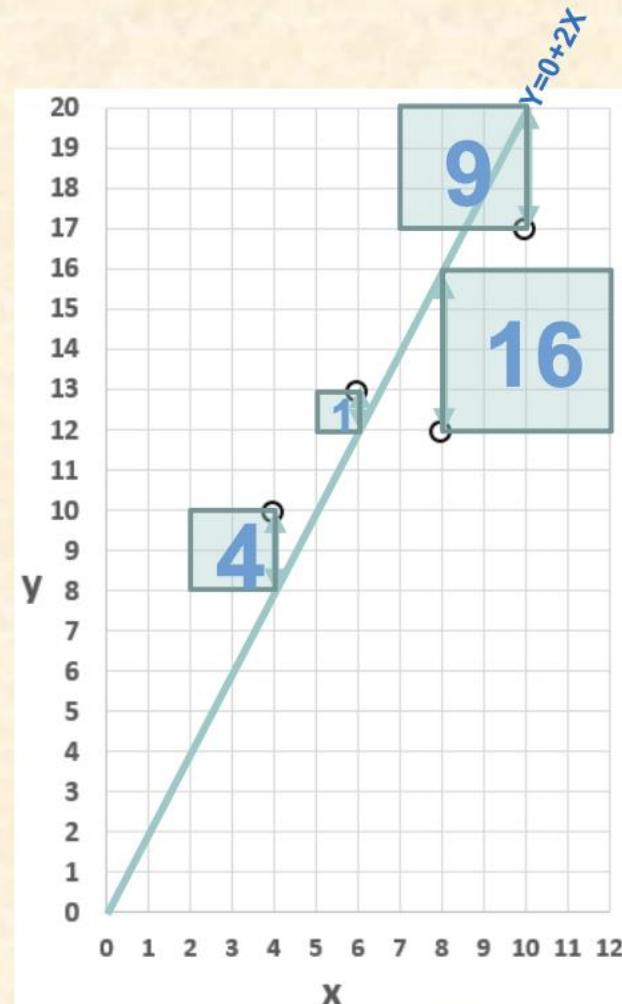
x	y
4	10
6	13
8	12
10	17

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

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

$$\text{SSQ}_{Y=0+2X} \quad 4+1+9+16=30$$

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

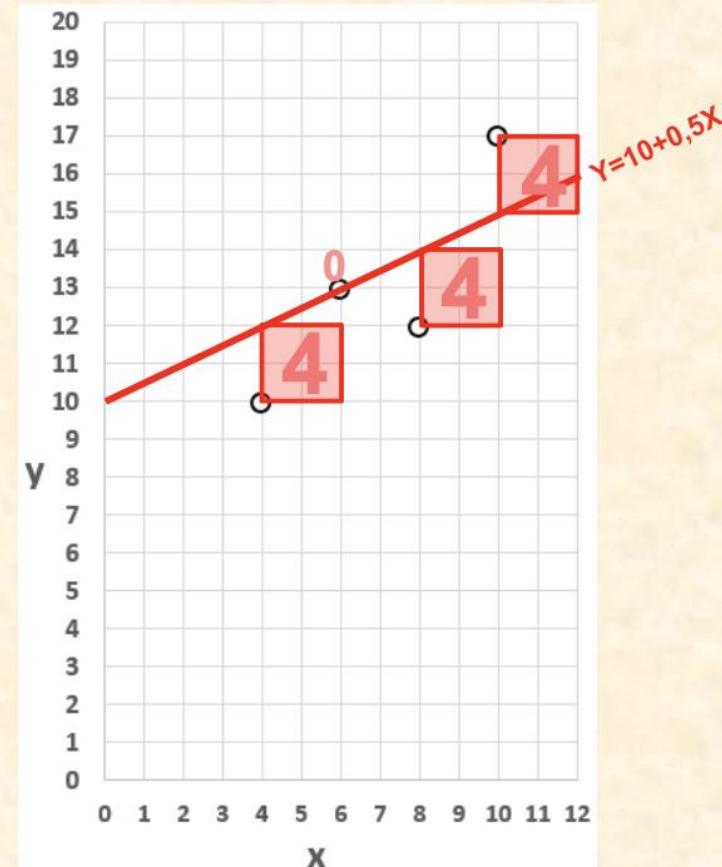


x	y
4	10
6	13
8	12
10	17

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**

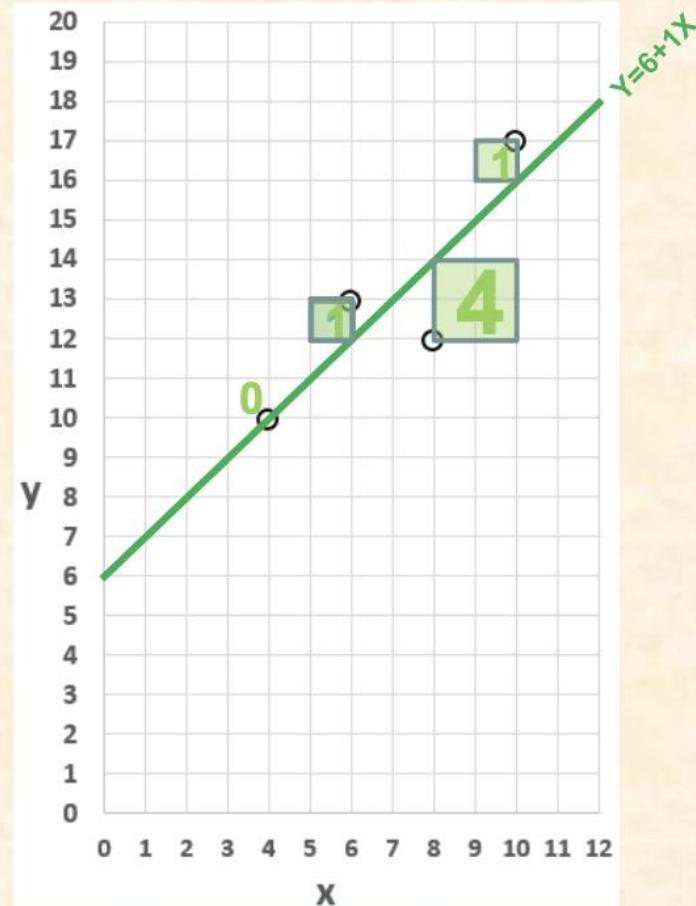
Sì, la rossa ha complessivamente una somma dei quadrati delle distanze minore, ma come si comporta la verde?



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**)



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:

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

Il coefficiente b_1 è detto anche **coefficiente di regressione**.

La formula per calcolare b_1 è la seguente:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Il denominatore è la **devianza** dei valori di x e ha ovviamente segno positivo (somma valori al quadrato).

Il numeratore è la **codevianza** di x e y e il suo segno può essere negativo (relazione discordante tra x e y) o positivo (relazione concordante tra x e y). Misura i movimenti **congiunti** delle due variabili.

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

La predizione (previsione) della spesa mensile di energia elettrica impiegando il numero di stanze dell'abitazione

Impiegando il metodo dei **minimi quadrati** (v. oltre) sull'esempio della spesa di energia elettrica, otteniamo la retta stimata:

$$\hat{y} = 42.09 + 7.72 \times$$

dove: $\hat{y} = \widehat{\text{Spesa_en_el}}$ è il valore predetto della variabile dipendente $x = \text{N_stanze}$.

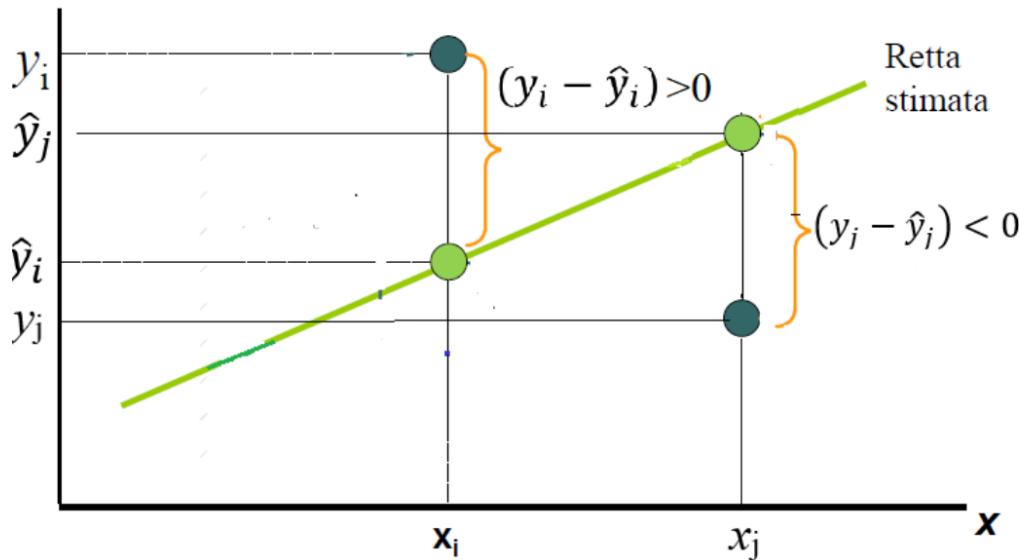
b_1 è positivo: più stanze ci sono nell' abitazione e maggiore è la spesa (mensile) predetta di energia elettrica.

Il valore $b_1 = 7.72$ ci dice che *una stanza in più comporta un incremento di spesa mensile di energia elettrica pari a 7 Euro e 72 centesimi*.

Analogamente: *2 stanze in più comportano un incremento di spesa mensile di energia elettrica pari a 15 Euro e 44 centesimi (7.72×2)*.

Gli errori di predizione (previsione): i residui o residui di regressione o di interpolazione (nel libro sono detti scarti)

Per una osservazione i la differenza $(y_i - \hat{y}_i)$ tra il valore osservato y_i e il valore predetto \hat{y}_i , è detto **residuo**.
Esso può essere positivo o negativo.



RSS: Residual Sum of Squares (talvolta detta anche SSE: Sum of Squared Errors) o **DEVIANZA RESIDUA**

Ogni osservazione ha un residuo. Se la retta stimata cade vicino ai punti, i residui sono piccoli. Sintetizziamo la grandezza dei residui con la somma dei loro quadrati:

$$\text{DEVIANZA RESIDUA} \text{ o } RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n r_i^2$$

Il quadrato garantisce che i termini della sommatoria sono tutti positivi: nessun bilanciamento tra residui negativi e positivi.

Ogni retta stimata ha i suoi residui e un determinato valore di RSS.

RSS sintetizza gli errori di predizione della retta stimata.

Le stime dei **minimi quadrati** b_0 e b_1 sono quei valori dei parametri che individuano la retta per la quale **RSS è minima**.

La retta dei minimi quadrati (MQ): proprietà

La retta stimata col metodo dei MQ è detta anche *retta dei MQ* ed è quella per la quale la somma dei quadrati dei residui (RSS) è la più piccola possibile. I punti sono più vicini a questa retta che non a qualsiasi altra retta usando come misura di distanza la RSS. Oltre a garantire questo, la retta dei MQ ha le seguenti proprietà.

- ▶ Ha qualche residuo positivo e qualche residuo negativo ma la loro somma (e quindi la loro media) è pari a **zero**. Le predizioni alte si bilanciano con quelle basse.
- ▶ La retta passa attraverso il punto di coordinate (\bar{x}, \bar{y}) ; passa cioè per il *centro* dei dati (già visto).
- ▶ La media dei valori predetti \hat{y}_i è uguale alla media \bar{y} dei valori osservati y_i .

La retta dei minimi quadrati (MQ): valutazione della bontà di adattamento ai punti: la riduzione di RSS [1/5]

Valutare quanto bene si adatta la retta ai punti equivale a rispondere alla seguente domanda: "Quanto mi aiuta la retta $b_0 + b_1 x$ a predire la y ?"

Se con la retta è possibile predire la y molto meglio di quanto si può fare **senza la x** allora la retta è utile a predire la y ovvero si adatta bene a rappresentare i dati.

Allora per capire la bontà della retta stimata, confrontiamo due situazioni.

1. Predire la y senza la x e calcolare RSS.
2. Predire la y utilizzando la retta stimata e calcolare RSS.
3. Confronto di RSS ottenuto nel caso 1 (RSS_1) con quello del caso 2 (RSS_2) calcolando la riduzione di RSS ottenuta impiegando la x .

La retta dei minimi quadrati (MQ) e riduzione di RSS [2/5]: predire la y senza la x

Predire la y senza la x equivale a utilizzare la forma analitica:

$y = \beta$ (retta priva della x e quindi una retta orizzontale).

La stima dei MQ di β è \bar{y} , la media dei valori di y .

In tal caso il singolo residuo è $(y_i - \bar{y})$ e allora RSS sarà::

$$RSS_1 = \sum_{i=1}^n (y_i - \bar{y})^2 = TSS \text{ o } \mathbf{DEVIANZA TOTALE di } y$$

Tale somma è detta TSS: Total Sum of Squares o anche
DEVIANZA TOTALE di y : variabilità dei valori della y intorno alla loro media.

Nell'esempio sulla spesa mensile di energia elettrica abbiamo:

$RSS_1 = TSS = 152880.9$ (Euro al quadrato).

La retta dei minimi quadrati (MQ) e riduzione di RSS [3/5]: predire la y con la x

Considerando l'esempio della spesa mensile di energia elettrica (y) e del numero di stanze dell'abitazione (x) abbiamo la retta stimata:

$$\hat{y}_i = 42.09 + 7.72 x_i$$

da cui

$$RSS_2 = \sum_{i=1}^{529} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{529} (y_i - (42.09 + 7.72 x_i))^2 = 108871.5$$

La retta dei minimi quadrati (MQ) e riduzione di RSS [5/5]: l'indice R^2 (R-quadro)

Riepilogando, la riduzione relativa di RSS che si ottiene impiegando la x e la retta dei MQ è:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i(y_i - \hat{y}_i)^2}{\sum_i(y_i - \bar{y})^2}$$

R^2 è detto indice (coefficiente) di **determinazione lineare**. Poiché $RSS \leq TSS$ sarà $0 \leq R^2 \leq 1$.

ESEMPIO. Nel caso della retta stimata per prevedere la spesa di energia elettrica col numero di stanze dell'abitazione abbiamo:

$TSS = 152880.9$ $RSS = 108871.5$ da cui:

$$R^2 = 1 - \frac{108871.5}{152880.9} = 0.2879$$

La riduzione di RSS è circa del 28.8%. Non molto soddisfacente!

Altro modo di derivazione e interpretazione di R^2 . La **ESS**: Explained Sum of Squares o **DEVIANZA DI REGRESSIONE** [2/4]

$$\begin{aligned} (y_i - \bar{y}) &= (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \\ \downarrow &\quad \downarrow &\quad \downarrow \\ \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ \downarrow &\quad \downarrow &\quad \downarrow \\ \text{TSS} &= \text{ESS} + \text{RSS} \end{aligned}$$

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\text{Devianza di regressione}}{\text{Devianza totale di } y}$$

ESS è la variabilità dei valori predetti \hat{y} intorno alla loro media \bar{y} .
Poiché $\text{ESS} \leq \text{RSS}$ allora ritroviamo ancora che: $0 \leq R^2 \leq 1$.

Esempio della spesa mensile di energia elettrica [4/4]

Devianza di regressione (ESS)	44009.4
Devianza residua (RSS)	108871.5
Devianza totale (TSS)	152880.9

$$R^2 = \frac{\text{Devianza di regressione (ESS)}}{\text{Devianza totale (TSS)}} = \frac{44009.4}{152880.9} = 0.2879$$

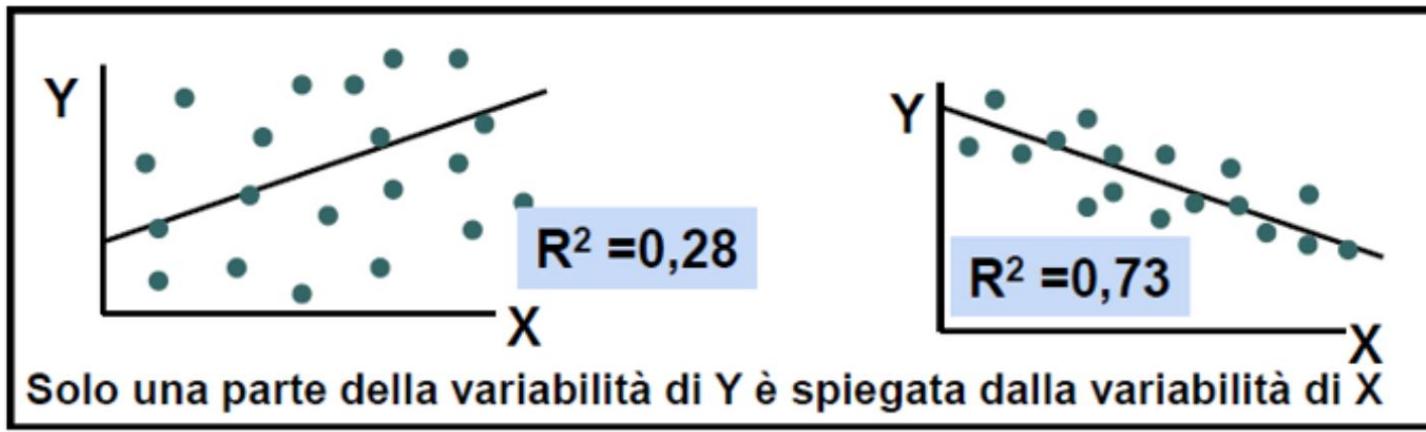
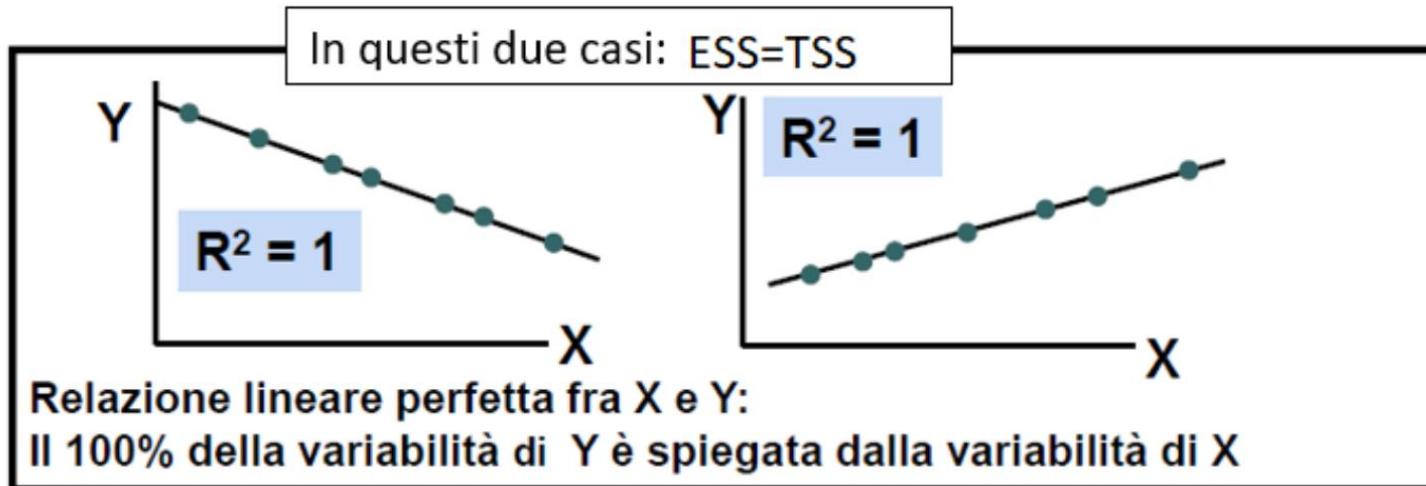
Interpretazione del valore di R^2 .

Dal calcolo sopra vediamo che:

$$44009.4 = 0.2879 \times 152880.9$$

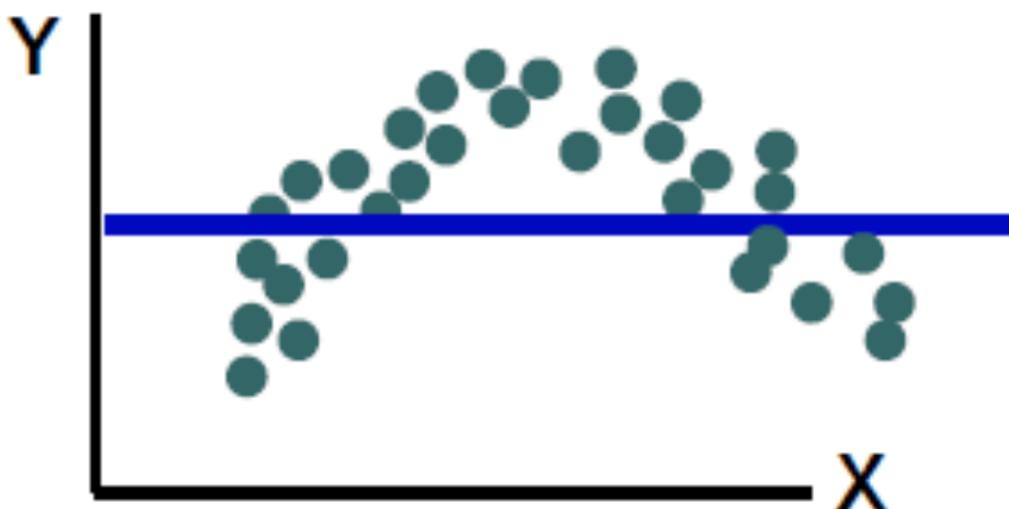
Possiamo allora dire che: la devianza di regressione ESS è il 28.8% della devianza totale di y (TSS). Ovvero: la retta stimata spiega circa il 28.8% della variabilità totale della y .

Esempi di valori di R^2

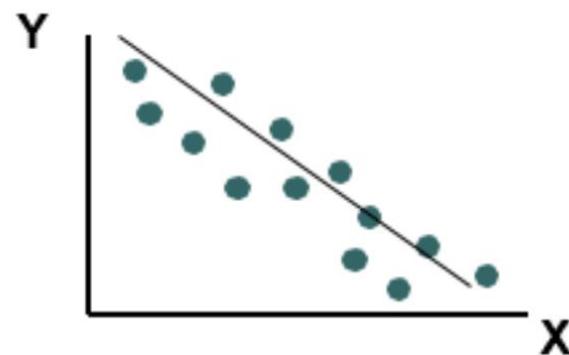
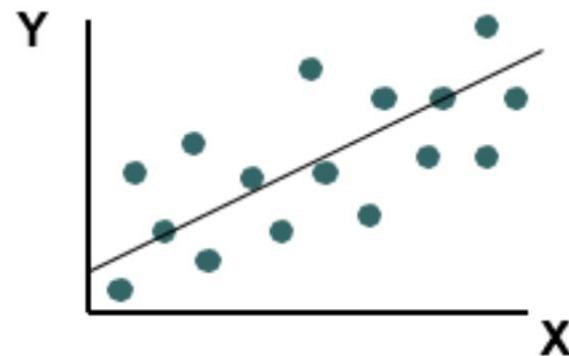


R^2 descrive la forza del legame **lineare** fra x e y.

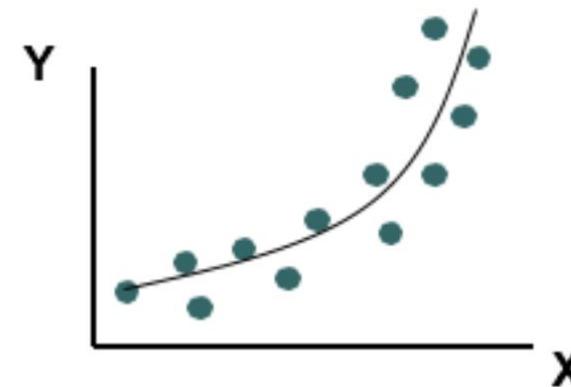
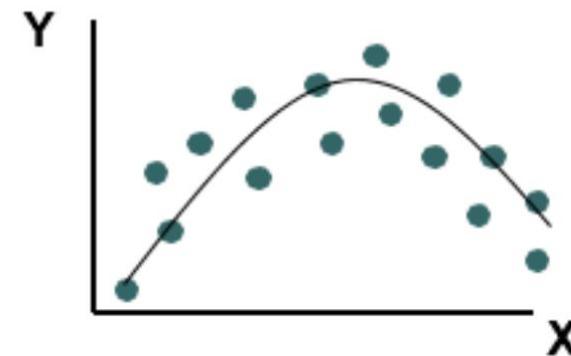
Nel caso esemplificato in figura, il valore di R^2 sarà praticamente 0 ma c'è una relazione (non lineare) fra x e y.



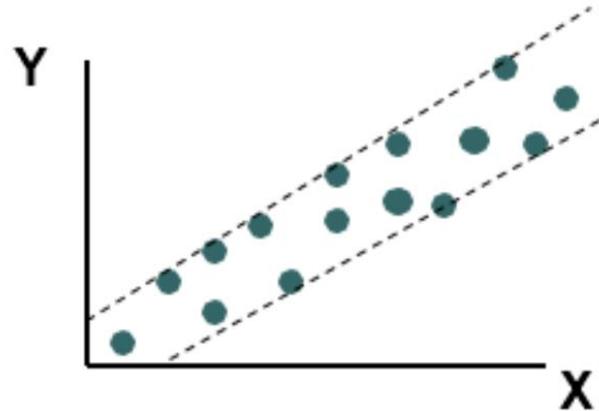
Relazione Lineare



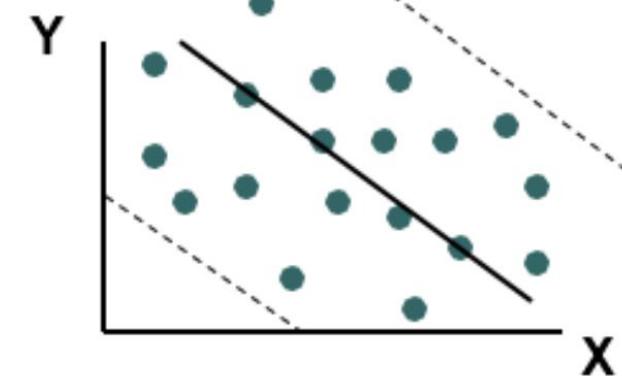
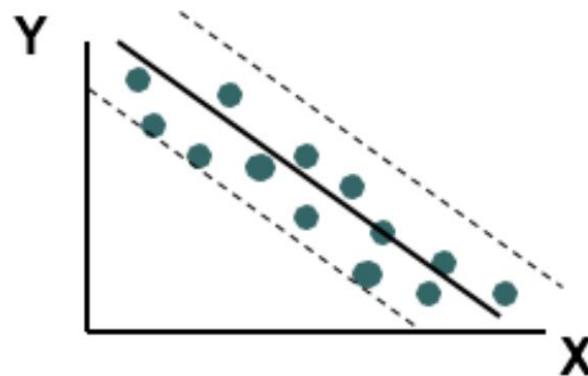
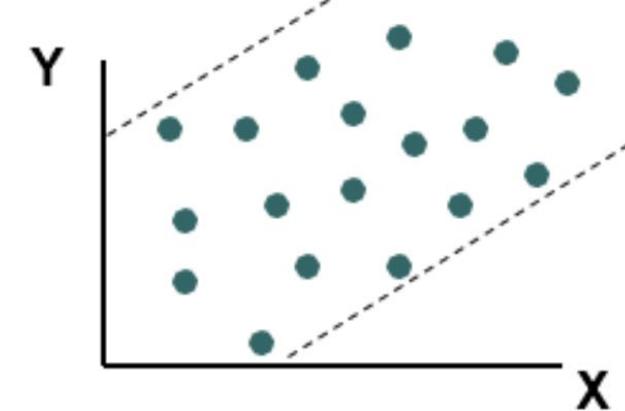
Relazione non lineare



Relazione forte



Relazione debole



REGRESSIONE LINEARE (CAPITOLO 6)

PARTE II: INFERNZA SU β_1

Torniamo ai nostri dati sul numero di stanze dell'abitazione (x) e la spesa di energia elettrica (y).

La retta stimata coi minimi quadrati è: $\hat{y} = 42.09 + 7.72x$

Il valore di b_1 è diverso da zero ... ma la stima è stata ricavata da un campione e allora ... i casi potrebbero essere due:

- 1) nella popolazione $\beta_1 = 0$ (non esiste relazione tra le due variabili), ma il valore b_1 è diverso da zero solo a causa della variabilità campionaria delle stime;
- 2) nella popolazione $\beta_1 \neq 0$ (esiste una relazione lineare tra le due variabili).

Come facciamo a discriminare tra le due situazioni?

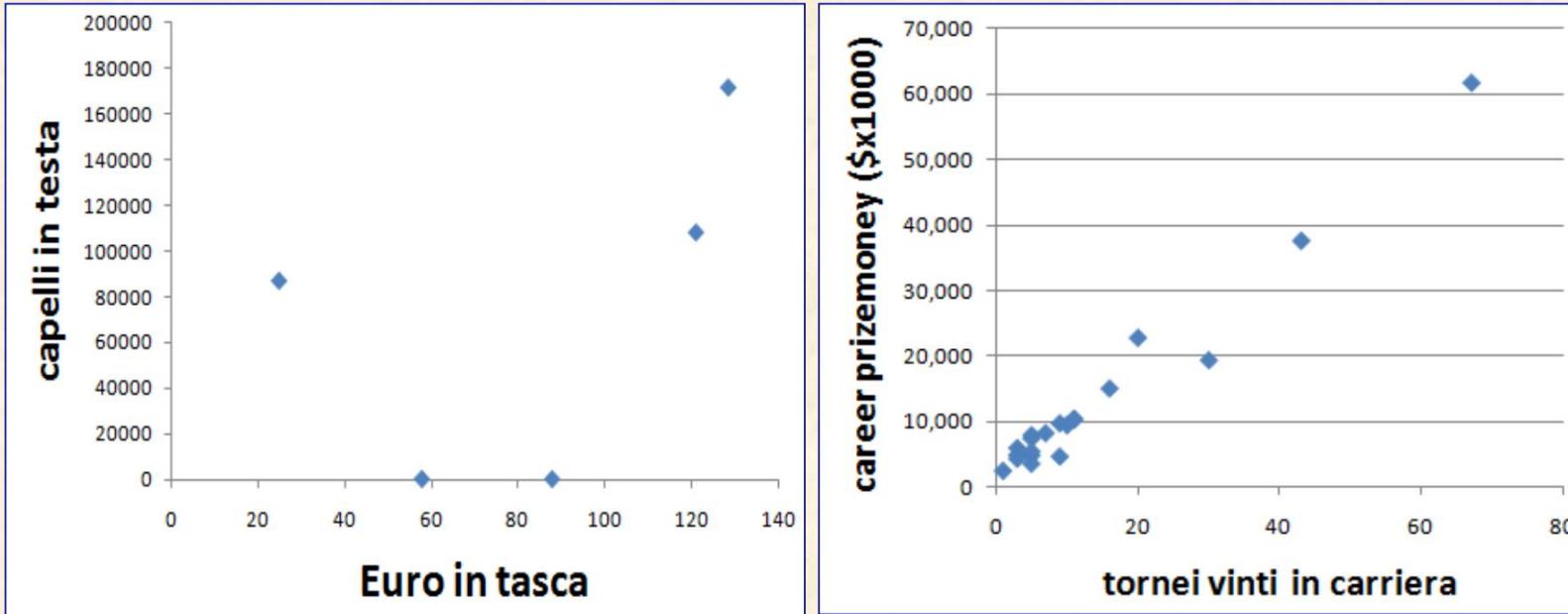
Bisogna condurre un test delle ipotesi impostando l'ipotesi nulla $H_0 : \beta_1 = 0$.

INFERENZA SUL COEFFICIENTE ANGOLARE

Prendiamo i seguenti esempi:

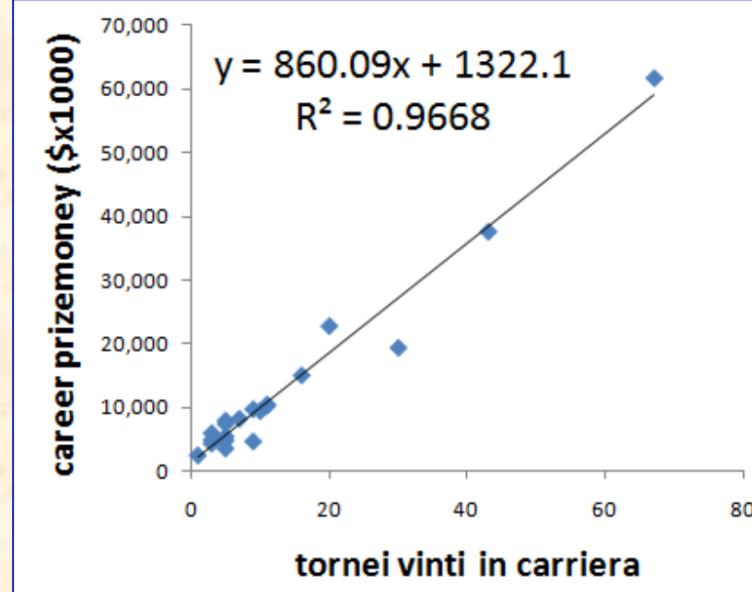
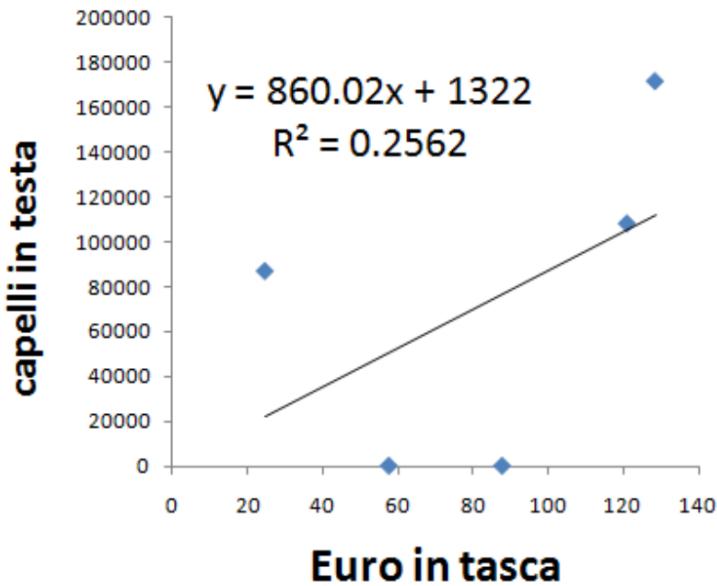
1. conto i soldi in tasca di 5 miei amici e il numero di capelli di ciascuno
2. Conto il numero di tornei vinti e il career prizemoney (in migliaia di \$) dei primi 20 tennisti al mondo (dati 16/3/2011)

Diamo un'occhiata agli scatter:

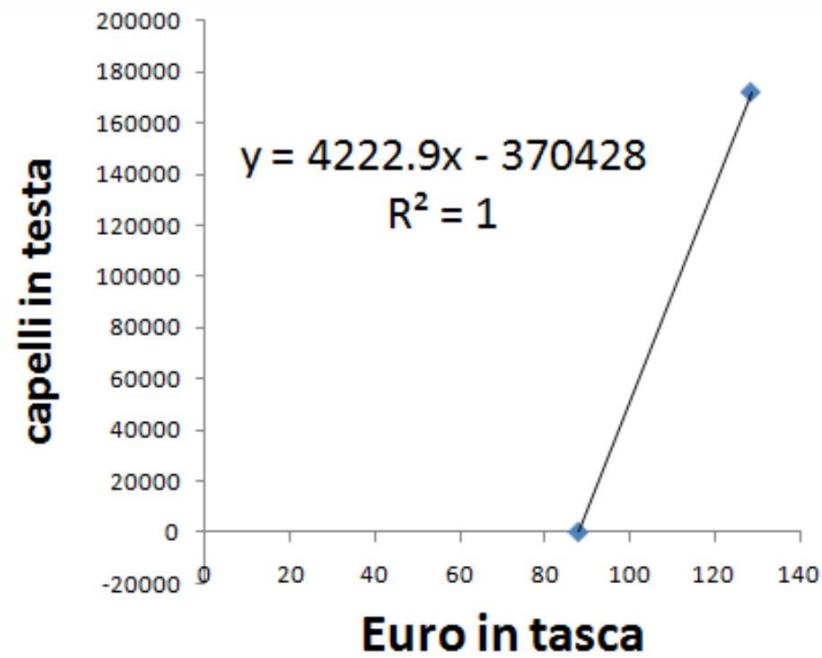


Cosa vediamo? Quello che ci aspettavamo: non “si vede” niente nel primo grafico, mentre nel secondo è evidente una disposizione dello scatter attorno a una retta immaginaria, però...

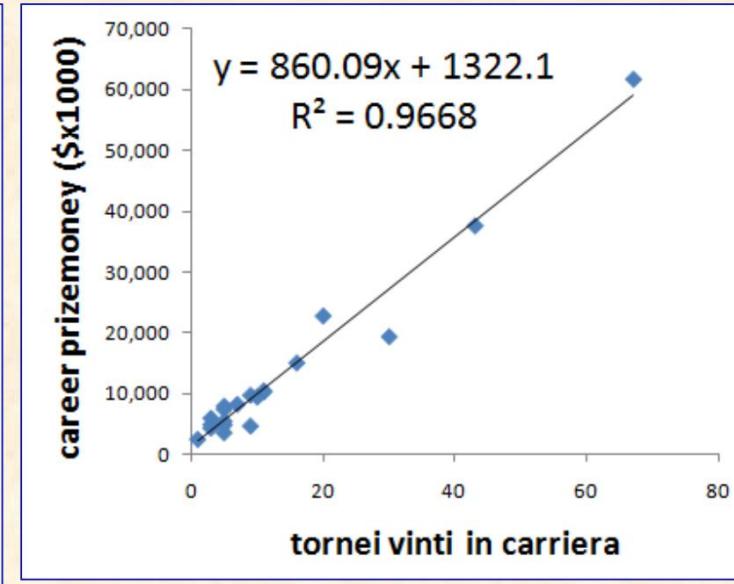
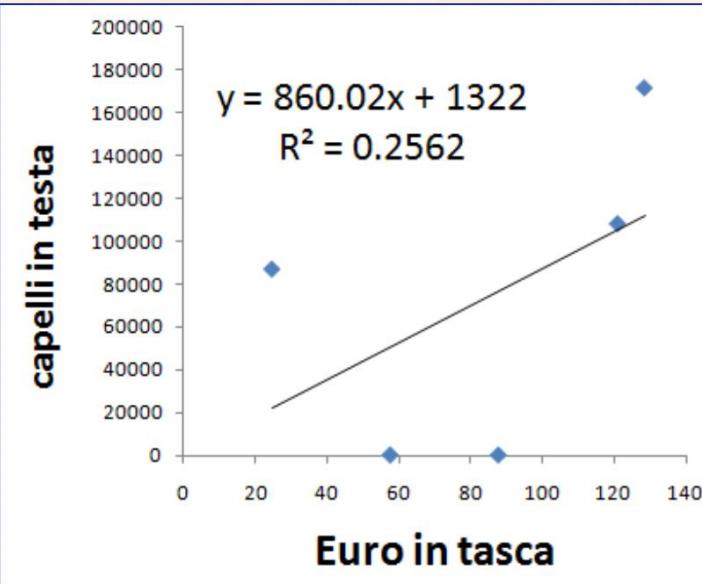
Però... se provo a stimare le rette dei minimi quadrati...



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



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 probabilmente una retta dei m.q. profondamente differente

$$Y = E(Y|x) + \epsilon = (\beta_0 + \beta_1 x) + \epsilon$$

Vediamo che Y è scomposto nella somma di due elementi:

- la **componente sistematica** $E(Y|x) = (\beta_0 + \beta_1 x)$ che esprime la relazione deterministica tra x e $E(Y|x)$;
- la **componente di disturbo** ϵ

Ipotizziamo che: $E(Y|x) = \beta_0 + \beta_1 x$ e che:

1) $\epsilon \sim N(0, \sigma^2)$ da cui:

$$E(\epsilon) = 0, \text{Var}(\epsilon) = \sigma^2$$

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

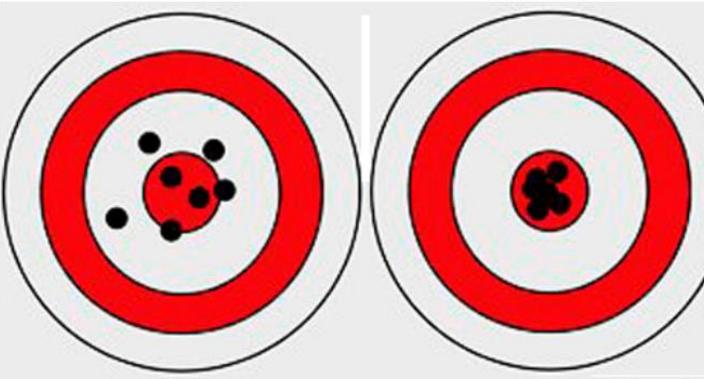
2) le x sono valori dati

3) le osservazioni sono indipendenti (campione casuale semplice)

La varianza è costante per ogni x (ipotesi di *omoschedasticità*): infatti la curva normale ha la stessa forma al variare di x .

Il centro del bersaglio è il valore β_1 (qualunque esso sia).

I punti rappresentano possibili valori b_1 generati da due ipotetici stimatori di β_1 . Tutti e due gli stimatori sono corretti ovvero sono tarati sul centro del bersaglio. Quello di destra è più preciso.



La precisione di B_1 stimatore dei MQ di β_1 è data dalla radice della sua varianza ovvero dalla sua deviazione standard che è:

$$DS(B_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Se è vera H_0 , mi aspetto di avere stime b_1 interne o vicine alla circonferenza azzurra.

Se invece H_0 è falsa, essendo B_1 stimatore corretto di β_1 , esso tenderà a generare stime b_1 centrate sul valore vero di β_1 e quindi lontane dal centro dove $\beta_1 = 0$.

La variabilità di B_1 dipende da:

- σ^2 : la variabilità del disturbo ϵ (più grande σ^2 minore la precisione di B_1)
- dimensione campionaria n : più grande è n e più preciso è B_1
- $MSD(x)$: la variabilità dei valori di x intorno alla loro media: più grande è e più preciso è B_1 (v. lucido seguente).



Tuttavia noi non conosciamo $DS(B_1)$ perché non conosciamo il valore σ^2 . Possiamo però stimarlo dai residui di regressione.

La stima corretta di σ^2 è data da:

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

Sostituendo $s = \sqrt{s^2}$ nella formula di $DS(B_1)$ ricaviamo:

$$ES(B_1) = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$ES(B_1)$ è l'**errore standard** di B_1 ed è la stima di $DS(B_1)$.

Test delle ipotesi su β_1



Test delle ipotesi su β_1 per indagare sull'esistenza della relazione lineare con x

$H_0 : \beta_1 = 0$ (assenza di relazione)

$H_1 : \beta_1 \neq 0$ (presenza di relazione)

Per applicare il ragionamento visto con l'esempio dei bersagli, viene costruito il seguente rapporto (statistica test)

$$t = \frac{b_1 - 0}{ES(B_1)} = \frac{b_1}{ES(B_1)}$$

$$t = \frac{b_1 - 0}{ES(B_1)}$$

Numeratore di t : distanza tra b_1 e $\beta_1 = 0 \Leftarrow H_0$

Denominatore: in media, le stime di B_1 distano $ES(B_1)$ dal vero valore β_1 .

Se b_1 dista da 0 più di $ES(B_1)$ allora riterremo che ciò accade perché il vero β_1 è diverso da zero e rifiuteremo H_0 .

Come regola empirica, se $n > 60$ possiamo rifiutare H_0 se t è minore di -2 (se b_1 è negativo) o maggiore di 2 (se b_1 è positivo).

In sostanza, si rifiuta H_0 , se la distanza di b_1 da 0 è il doppio di $ES(B_1)$. Questo corrisponde approssimativamente ad una probabilità di rifiutare H_0 quando è vera pari a 0.05.

Test delle ipotesi su β_1 : il *p*-value

Un altro modo per decidere è quello di affidarci al *p*-value associato al valore di t .

Il *p*-value è la probabilità **se H_0 è vera** di osservare un valore di t uguale o più estremo di quello ottenuto dal campione.

Il *p*-value misura l'evidenza fornita dai dati contro l'ipotesi nulla: **minore** è il valore del *p*-value, e più è forte l'evidenza **contro** l'ipotesi nulla.

Posto α la probabilità di rifiutare H_0 quando è vera (probabilità dell'errore di I tipo), rifiuteremo H_0 quando $p\text{-value} < \alpha$.

In modo non del tutto corretto ma efficace, possiamo dire che il *p*-value ci mostra quanto è verosimile il valore b_1 se H_0 fosse vera. Se è poco verosimile, significa che l'evidenza empirica è in contrasto con H_0 e quindi si rifiuta H_0 .

Per dire che b_1 è poco verosimile se $H_0 : \beta_1 = 0$ è vera, confrontiamo il *p*-value con α e...

se p -value < α \Rightarrow rifiutiamo H_0
altrimenti, accettiamo H_0

Supponiamo di avere scelta la probabilità $\alpha = 0.05$ di rifiutare l'ipotesi nulla quanto è vera (probabilità dell'errore di I tipo).

- 1) Supponiamo di avere ottenuto, per il consueto test delle ipotesi $H_0 : \beta_1 = 0$ un $p - value < \alpha$. Allora si rifiuta H_0 e si dice **il coefficiente β_1 è significativamente diverso da 0 a livello $\alpha = 0.05$.**
- 2) Altrimenti se $p - value > \alpha$ si accetta H_0 e si dice **il coefficiente β_1 non è significativamente diverso da 0 a livello $\alpha = 0.05$.**

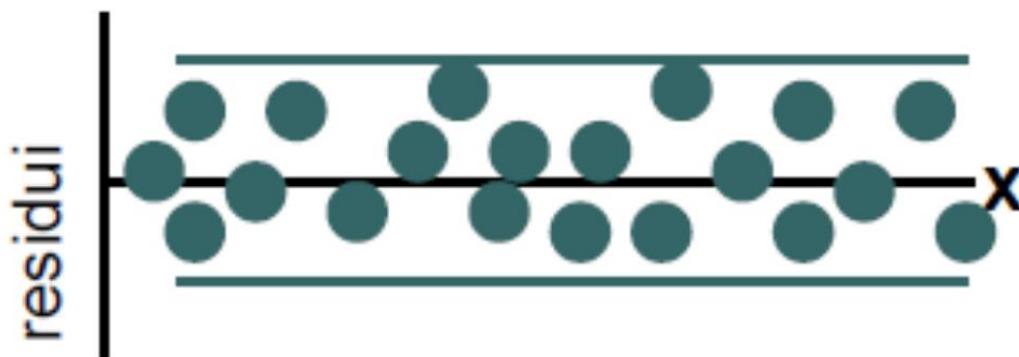
Analisi dei residui

Se la forma analitica stimata si avvicina a quella vera, i residui di regressione dovrebbero riflettere il comportamento di ϵ .

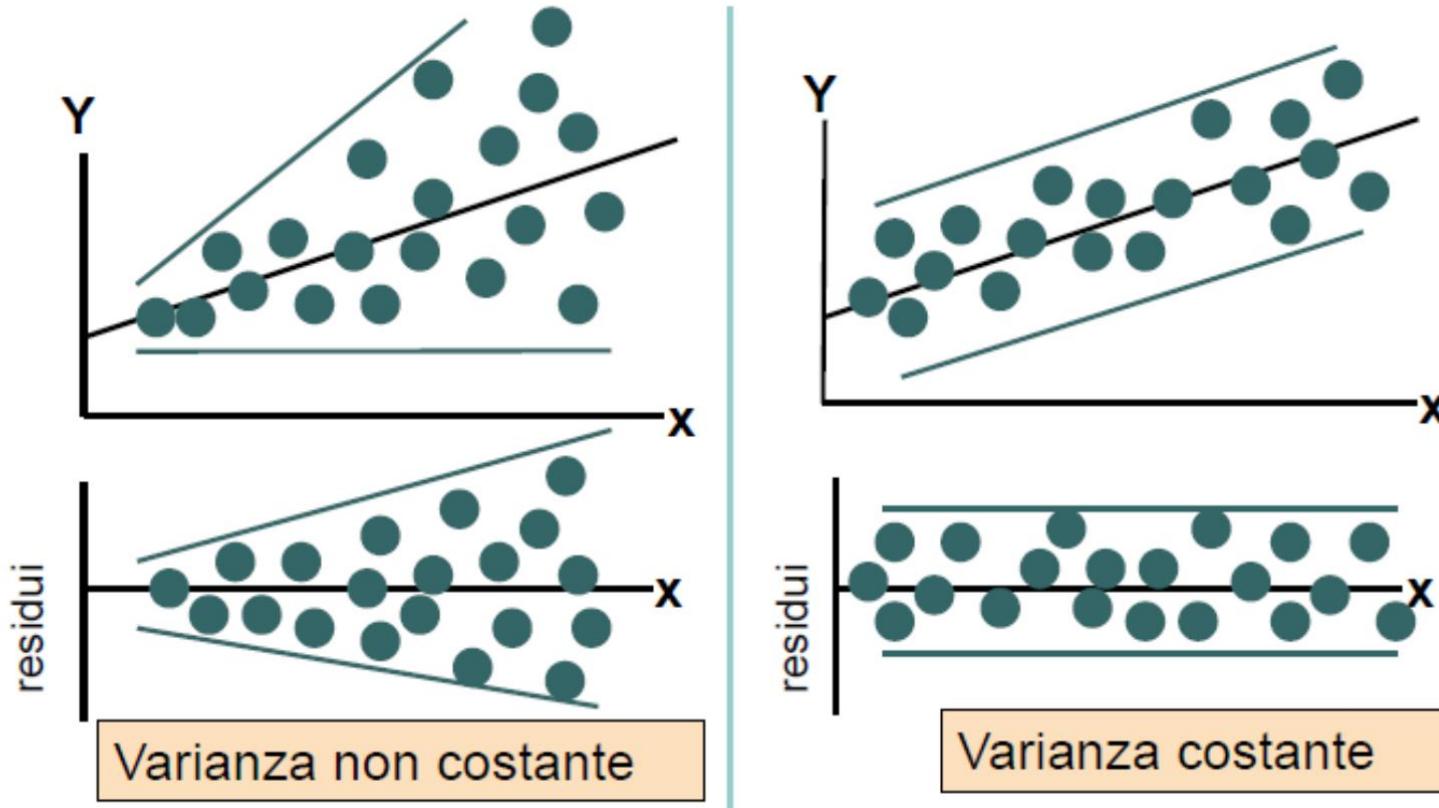
Allora è importante esaminare il comportamento dei residui. Si tratta di condurre alcune semplici analisi grafiche per indagare su:

- casualità dei residui
- ipotesi di omoschedasticità (varianza costante)
- indipendenza.

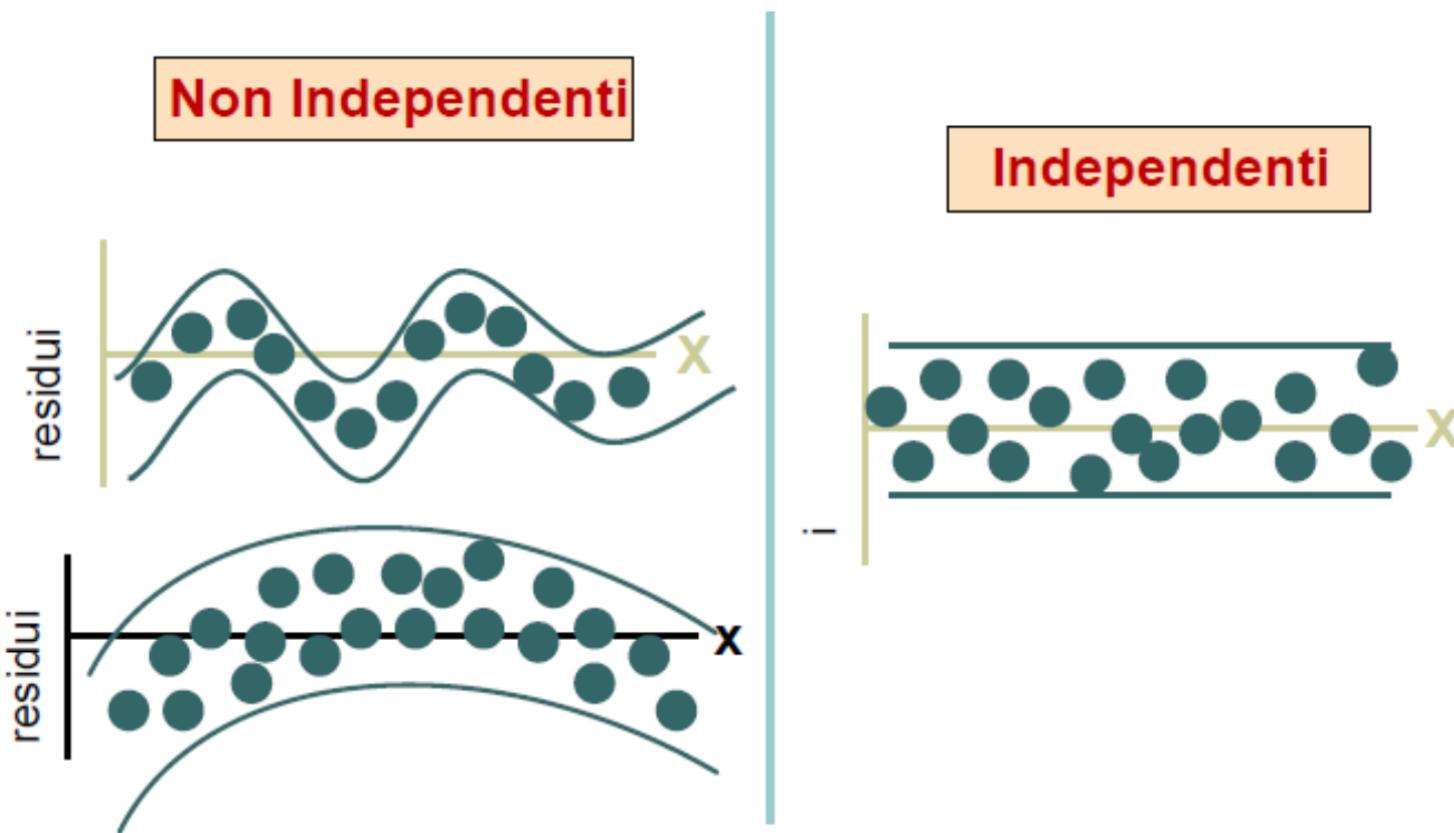
I residui non devono presentare alcun andamento sistematico.
L'ideale è che si dispongano come nella figura (e cioè casualmente
intorno allo 0).



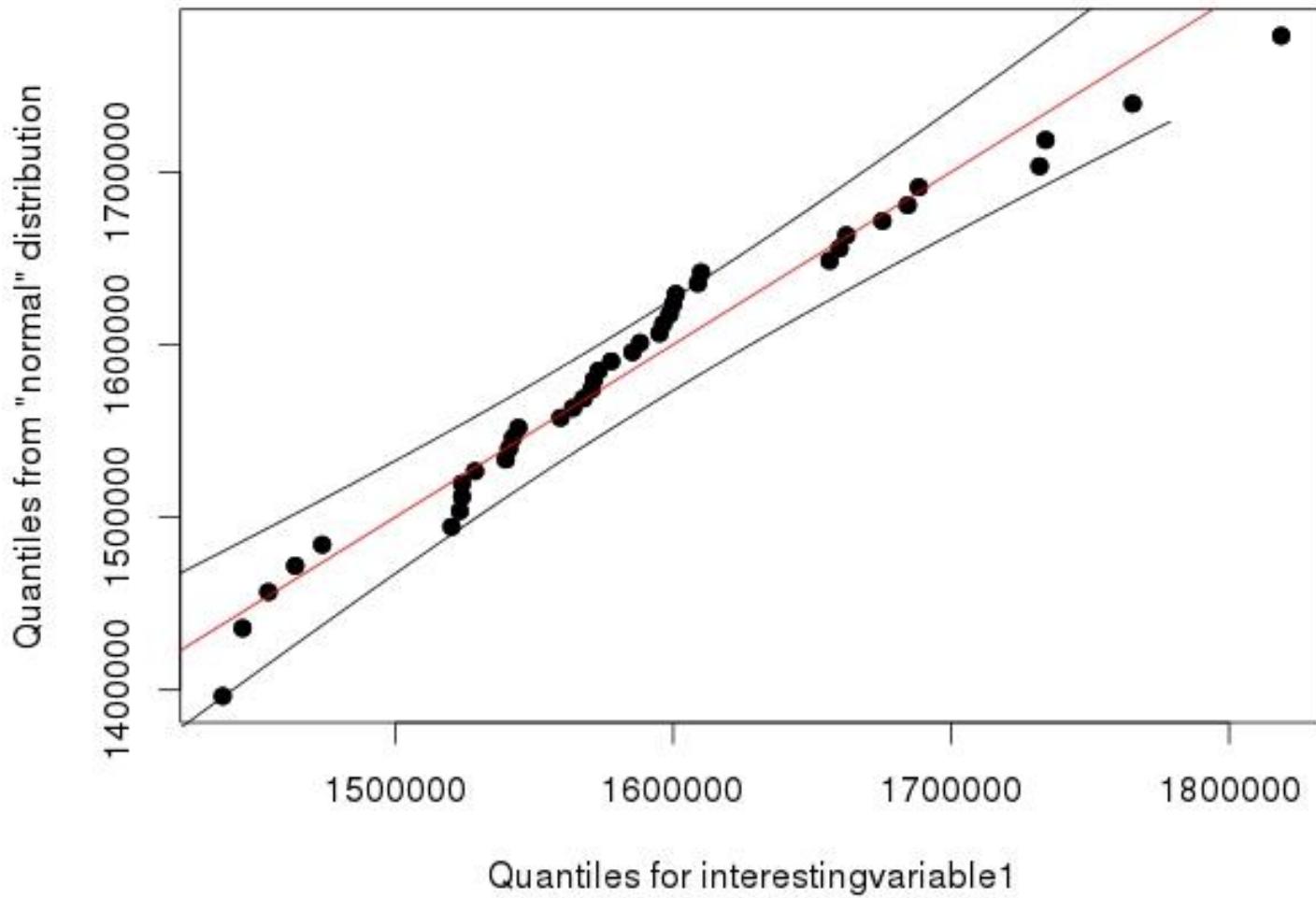
Analisi dei residui: omoschedasticità



Analisi dei residui: indipendenza



Q-Q Plot for "normal" distribution



La variabilità di B_1 dipende da:

- σ^2 : la variabilità del disturbo ϵ (più grande σ^2 minore la precisione di B_1)
- dimensione campionaria n : più grande è n e più preciso è B_1
- $MSD(x)$: la variabilità dei valori di x intorno alla loro media: più grande è e più preciso è B_1 (v. lucido seguente).



Tuttavia noi non conosciamo $DS(B_1)$ perché non conosciamo il valore σ^2 . Possiamo però stimarlo dai residui di regressione.

La stima corretta di σ^2 è data da:

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

Sostituendo $s = \sqrt{s^2}$ nella formula di $DS(B_1)$ ricaviamo:

$$ES(B_1) = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$ES(B_1)$ è l'**errore standard** di B_1 ed è la stima di $DS(B_1)$.

Test delle ipotesi su β_1



Test delle ipotesi su β_1 per indagare sull'esistenza della relazione lineare con x

$H_0 : \beta_1 = 0$ (assenza di relazione)

$H_1 : \beta_1 \neq 0$ (presenza di relazione)

Per applicare il ragionamento visto con l'esempio dei bersagli, viene costruito il seguente rapporto (statistica test)

$$t = \frac{b_1 - 0}{ES(B_1)} = \frac{b_1}{ES(B_1)}$$

STATISTICA PER LE APPLICAZIONI AZIENDALI 2024-25

REGRESSIONE LINEARE (CAPITOLO 6).
PARTE III: REGRESSIONE LINEARE
MULTIPLA

Laura Grassini

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)
Università degli Studi di Firenze

Scuola di Economia e Management
Corso di Laurea in Economia aziendale

Premessa



Quando abbiamo una sola x : regressione lineare **semplice**

Quando abbiamo due o più x : regressione lineare **multipla**

Esempio: le vendite (Euro) sono influenzate dal numero di dipendenti e dalla spesa pubblicitaria (Euro).

Struttura dei dati.

Negozio	Vendite (Y)	N. dipendenti (x1)	Spesa pubblicitaria (x2)
1	5100	3	950
2	5800	6	1700
3	5500	4	1100
4	7000	8	1300
5	4700	5	800
6	7700	7	1500
7	6300	3	2000
8	5200	6	1050
9	6800	4	1350
10	4900	8	950

Premessa (fine)



Specificazione del modello di regressione multipla con due predittori. Le ipotesi sono analoghe a quelle del modello di regressione semplice.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

dove:

1) $\epsilon \sim N(0, \sigma^2)$ da cui:

$$E(\epsilon) = 0, \text{var}(\epsilon) = \sigma^2$$

$$Y \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2; \sigma^2)$$

2) x_1 e x_2 sono valori dati (non sono variabili casuali)

3) le osservazioni sono indipendenti (campione casuale semplice)

La varianza di ϵ è costante al variare di x_1 e x_2 (ipotesi di omoschedasticità).

Stima coi minimi quadrati (MQ) dei coefficienti



Applicando il metodo dei minimi quadrati ricaviamo:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

dove b_0, b_1, b_2 sono le stime dei MQ di $\beta_0, \beta_1, \beta_2$.

I coefficienti b_1, b_2 sono detti "coefficienti **parziali** di regressione" perché indicano l'effetto di una variabile x condizionatamente alla presenza dell'altra nell'equazione.

Se le due variabili x_1 e x_2 sono **correlate**, i valori di b_1 e b_2 sono **differenti** da quelli dei corrispondenti coefficienti delle due rette dei MQ (regressione semplice): quella di y vs. x_1 e quella di y vs. x_2 . **Vedere slide 8-12.**

Interpretazione dei coefficienti



- ▶ b_0 è il valore di \hat{y} quando $x_1 = 0$ e $x_2 = 0$;
- ▶ b_1 è la variazione di \hat{y} quando x_1 aumenta di 1 unità **tenuta costante** x_2 ;

$$\hat{y}_{(1)} = (b_0 + b_1 (x_1 + 1) + b_2 x_2)$$

$$\hat{y}_{(1)} - \hat{y} = (b_0 + b_1 (x_1 + 1) + b_2 x_2) - (b_0 + b_1 x_1 + b_2 x_2) = b_1$$

- ▶ b_2 è la variazione di \hat{y} quando x_2 aumenta di 1 unità **tenuta costante** x_1 ;

$$\hat{y}_{(2)} = b_0 + b_1 x_1 + b_2 (x_2 + 1)$$

$$\hat{y}_{(2)} - \hat{y} = b_0 + b_1 x_1 + b_2 (x_2 + 1) - (b_0 + b_1 x_1 + b_2 x_2) = b_2$$

- ▶ non possiamo confrontare direttamente i valori dei coefficienti di regressione b_1 e b_2 perché sono quasi sempre espressi in una differente unità di misura o differente scala di valori (differenti ampiezze del range di valori).

Esempio: stima MQ vendite (y ; Euro) in funzione del n. dipendenti (x_1) e della spesa pubblicitaria (x_2 ; Euro)



L'equazione stimata è (uso del punto decimale):

$$\widehat{Vendite} = 2802.59 + 155.91 \text{ } N. \text{ dipendenti} + 1.776 \text{ } Spesa \text{ pubblicitaria}$$

Interpretazione dei coefficienti.

- ▶ Se non ci sono dipendenti (lavorano solo i titolari o proprietari del negozio e quindi se $N.dipendenti=0$) e non si spende in spesa pubblicitaria (Spesa pubblicitaria=0), le vendite mensili predette (o attese) ammontano a 2802.59 Euro.
- ▶ Mantenendo costante il livello di spesa pubblicitaria, con un dipendente in più le vendite mensili predette (o attese) aumentano di 155.91 Euro.
- ▶ Mantenendo costante il numero di dipendenti, con 1 Euro in più di spesa pubblicitaria, le vendite mensili predette (o attese) aumentano di 1.776 Euro (forse meglio dire: ogni 100 Euro in più di spesa pubblicitaria le vendite aumentano di 177.6 Euro).

Esempio: stima MQ vendite (Euro) in funzione del n. dipendenti e della spesa pubblicitaria

Anche i coefficienti del modello di regressione multipla sono soggetti al test delle ipotesi dove H_0 concerne l' **assenza di effetto sulla variabile dipendente** e cioè $H_0 : \beta_j = 0$, ($j = 1, \dots, k$), dove j indica la generica variabile x_j .

Lasciando perdere l'intercetta e posto $\alpha = 0.05$, vediamo che il coefficiente associato alle spese pubblicitarie è significativamente diverso da zero ($p\text{-value} < \alpha$) mentre il coefficiente del N. di dipendenti ha un $p\text{-value}$ superiore a 0.3 e cioè maggiore di α per cui si deve accettare l'ipotesi di assenza di effetto sulla y .

Possiamo allora eliminare la variabile Numero di dipendenti dal modello che diventa un modello di regressione lineare semplice con la sola x : *Spesa pubblicitaria*.

Bontà di adattamento: indice R^2 corretto



Non è appropriato confrontare l'indice R^2 di due modelli aventi differente numero di predittori.

Occorre confrontare l'indice detto R^2 **corretto** (*adjusted R²*). Con riferimento ad un modello di regressione lineare con k predittori (variabili x) esso è indicato con \bar{R}^2 dove:

$$\bar{R}^2 = 1 - \frac{\text{RSS}/(n - k - 1)}{\text{TSS}/(n - 1)}$$

Nella formula, RSS e TSS sono divisi per i rispettivi gdl.

Si noti che $(n - k - 1) = n - (k + 1)$ dove $(k + 1)$ è il numero di parametri nel modello (i k beta associati alle k variabili x , e l'intercetta).

Stima di σ^2



Così come avviene per il modello di regressione semplice, anche nel modello di regressione multipla la stima di σ^2 avviene mediante la somma dei quadrati dei residui RSS.

Nel caso di k variabili esplicative (variabili x), la stima di σ^2 è:

$$s^2 = \frac{RSS}{n - (k + 1)} = \frac{RSS}{n - k - 1}$$

dove $(n - k - 1)$ sono i gradi di libertà associati a RSS.

Predittore categorico (continua): la variabile dummy



Per gestire un predittore categorico, usiamo la tecnica della variabile dummy. La variabile dummy è una variabile numerica che assume solo due valori: 1 o 0.

Decidiamo di definire la variabile dummy z in funzione di x_2 nel seguente modo.

$$z_i = \begin{cases} 0 & \text{se } x_{2i} = \text{No (nessuna modifica alla confezione)} \\ 1 & \text{se } x_{2i} = \text{Sì (c'è la modifica alla confezione)} \end{cases} \quad (1)$$

Una variabile dummy **non ha un'unità di misura** tipo metri, secondi, euro, ecc. . Essa è adimensionale, perché indica solo la presenza o assenza di una certa caratteristica. Quindi i valori numerici 1 e 0 sono da considerarsi come numeri puri.

Predittore categorico (continua): la variabile dummy (continua)



Il modello di regressione lineare è:

$$E(Y|x_1, z) = \beta_0 + \beta_1 x_1 + \beta_2 z$$

E la stima dei MQ sarà:

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 z_i$$

da cui:

$$z_i = 0 \Rightarrow \hat{y}_i = b_0 + b_1 x_{1i} + b_2 \times 0 = b_0 + b_1 x_{1i}$$

$$z_i = 1 \Rightarrow \hat{y}_i = b_0 + b_1 x_{1i} + b_2 \times 1 = b_0 + b_1 x_{1i} + b_2$$

Ovvero:

$$z_i = 1 \Rightarrow \hat{y}_i = (b_0 + b_2) + b_1 x_{1i}$$

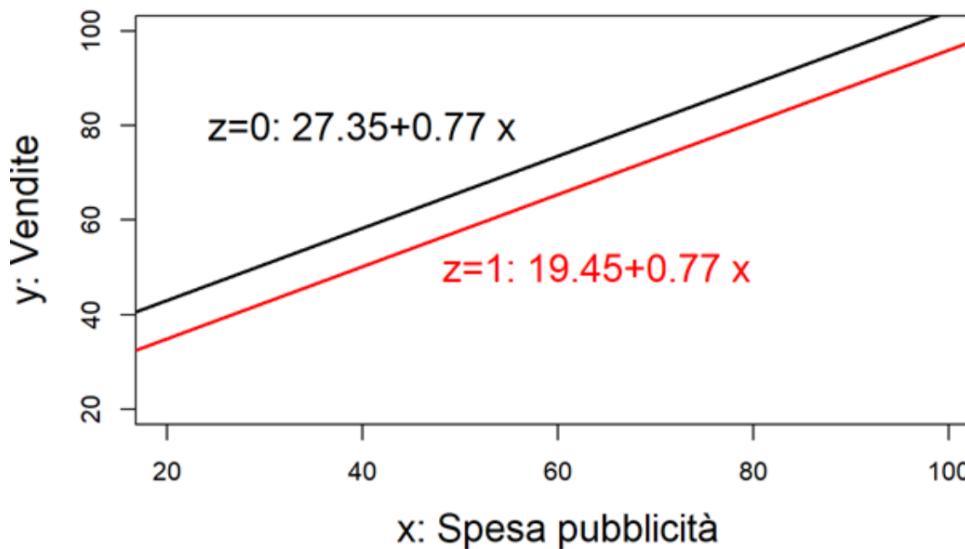
In pratica, se $b_2 \neq 0$, abbiamo due rette in funzione di x , aventi stessa pendenza (rette parallele) ma differente intercetta: una retta per $z = 1$ e una per $z = 0$.

Predittore categorico (fine): La variabile dummy (fine)



Le due rette in funzione di x sono parallele.

Poiché $b_2 < 0$, la retta per $z = 1$ sta sotto quella per $z = 0$.



Collinearità e multicollinearità (Paragrafo 6.3.4)



Quando due predittori (variabili x) sono molto correlati tra loro siamo in presenza del fenomeno della **collinearità** che può essere più o meno grave. Si parla di **multicollinearità** se siamo in presenza di un modello con più di due predittori fortemente correlati.

Considerando un modello con due predittori, in presenza di collinearità potrà accadere:

1. una delle due variabili è significativa e l'altra non lo è anche se lo sarebbe nella regressione lineare semplice;
2. l'effetto di ciascuna delle due variabili risulta non significativo; questo accade perché, la presenza di collinearità può inflazionare talmente l'errore standard dello stimatore del coefficiente da risultare un valore t vicino a zero, e quindi un p -value elevato che fa accettare l'ipotesi nulla di assenza di effetto sulla y ; si noti che i coefficienti possono risultare non significativi anche in presenza di un valore elevato dell' R^2 .

Predittori correlati. CASO 1



Vediamo l'esempio della spesa in energia elettrica in funzione del numero di stanze e della superficie dell'abitazione.

La correlazione tra numero di stanze e superficie dell'abitazione è pari a 0.77.

Regressione semplice Spesa_en_el (y) vs. N_stanze (x)

	Stima	Errore standard	statistica t	p-value
Intercetta	42.0946	1.9827	21.23	<2e-16
N_stanze	7.7197	0.5289	14.60	<2e-16

$$\widehat{\text{Spesa_en_el}} = 42.1 + 7.7 \text{ N_stanze}$$

Il p -value è molto basso, l'effetto di N_stanze è significativo.

Si noti che $< 2e-16 = 2 \times 10^{-16} \simeq 0$.

Preditori correlati. CASO 1 continua



Regressione semplice Spesa_en_el (y) vs. Superficie (x)

	Stima	Errore standard	statistica t	p-value
Intercetta	40.119	1.513	26.52	<2e-16
Superficie	0.298	0.015	20.88	<2e-16

$$\widehat{\text{Spesa_en_el}} = 40.1 + 0.298 \text{ Superficie}$$

Il p -value è molto basso, l'effetto di Superficie è significativo.

Predittori correlati. CASO 1 fine



Regressione multipla Spesa_en_el (y) vs. N_stanze,
Superficie (x)

	Stima	Errore standard	statistica t	p-value
Intercetta	39.292	1.752	22.422	<2e-16
N_stanze	0.676	0.725	0.932	0.352
Superficie	0.282	0.022	12.633	<2e-16

$$\widehat{\text{Spesa_en_el}} = 39.3 + 0.676 \text{ N_stanze} + 0.282 \text{ Superficie}$$

L'effetto di N_stanze sulla spesa mensile di energia elettrica
non è significativo.



CAPITOLO 4 - Controllo statistico della qualità dei prodotti e dei processi produttivi

Paragrafo 4.2 - Metodi off line e ANOVA

4



Miglioramento della qualità coi metodi off-line

E' importante trovare in **fase di progettazione** del processo le condizioni operative (tipo di materiali, taratura delle macchine, ecc.) che consentono di ottenere un **processo capace**.

Un modo per procedere è quello di effettuare degli **esperimenti programmati** per valutare **se** e in **quale misura** certi fattori del processo influenzano la caratteristica di qualità.



metodi off-line



Obiettivo dell'esperimento

L'esperimento è una serie di prove in cui vengono **deliberatamente** fatti variare i **livelli** dei **fattori controllabili** di processo in modo da poter osservare le corrispondenti variazioni sulla **variabile risposta** (la misura di qualità).

L'esperimento dovrà rispondere ai seguenti quesiti:

1. quali fattori influenzano la Y ?
2. quali livelli devono avere questi fattori in modo che la Y soddisfi il più possibile le specifiche ?



Concetti base della pianificazione sperimentale (2/3)

Trattamento: qualifica la prova a cui si sottopone l'unità.

Con **1 solo** fattore sperimentale, si hanno tanti trattamenti quanti sono i *livelli* del fattore.

Esempio. Fattore *temperatura*; 2 livelli (2 trattamenti): 180°C, 185°C

Con **2 o più fattori** sperimentali si hanno tanti trattamenti quanti sono le *combinazioni dei livelli* dei fattori.

Esempio. 2 fattori: *temperatura* (livelli: 180°C, 185°C) e *numero tazze di farina* (livelli: 2 tazze, 3 tazze).

4 trattamenti: (180°C, 2 tazze), (185°C, 2 tazze), (180°C, 3 tazze), (185°C, 3 tazze).

Valutazione dell'effetto del fattore sperimentale mediante l'ANOVA (ANalysis Of VAriance)

$$Y_{ij} \sim N(\mu_i; \sigma^2) \quad i=1, \dots, K; \quad j=1, \dots, n$$

i : indica il trattamento; j : indica l'unità sottoposta al trattamento i

L'ANOVA consiste nel condurre il seguente test

$$\begin{cases} H_0: \mu_1 = \mu_2 = \dots = \mu_K = \mu \\ H_1: \text{almeno una media è diversa dalle altre} \end{cases}$$

che equivale al test

$$\begin{cases} H_0: \mu_i - \mu = 0 \text{ per ogni } i=1, \dots, K \\ H_1: \mu_i - \mu \neq 0 \text{ per almeno un } i \end{cases}$$

$$\mu = \frac{\mu_1 + \mu_2 + \dots + \mu_K}{K}$$



Riepilogo: la variabilità intorno a μ

H₀ vera

$$E(Y_{ij} - \mu)^2 = \sigma^2$$

La variabilità intorno a μ è dovuta solo a cause accidentali

H₀ falsa (per almeno un livello i si ha: $\mu_i \neq \mu$)

$$E(Y_{ij} - \mu)^2 = E(Y_{ij} - \mu_i + \mu_i - \mu)^2 = E(Y_{ij} - \mu_i)^2 + (\mu_i - \mu)^2 = \sigma^2 + (\mu_i - \mu)^2$$

La variabilità intorno a μ è dovuta a cause accidentali e anche all'effetto del livello i del fattore



Esperimento con 1 fattore (4 livelli e dati bilanciati)

Materiale			
L1 (Fornitore 1)	L2 (Fornitore 2)	L3 (Fornitore 3)	L4 (Fornitore 4)
$y_{11}, y_{12}, \dots, y_{1n}$	$y_{21}, y_{22}, \dots, y_{2n}$	$y_{31}, y_{32}, \dots, y_{3n}$	$y_{41}, y_{42}, \dots, y_{4n}$

$$\bar{y}_i = \frac{1}{n} \sum_{j=1}^n y_{ij} \quad \text{media per il trattamento } i: \text{stima di } \mu_i$$

$$\bar{y} = \frac{1}{K} \sum_{i=1}^K \bar{y}_i \quad \text{media generale: stima di } \mu$$



Interpretazione della scomposizione

$$y_{ij} - \bar{y} = (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

Contiene componente
accidentale e, se H_0 è **falsa**,
anche componente
sistematica. Indica la
variabilità **fra** le medie dei
trattamenti (*between*)

Contiene solo la
componente accidentale.
Indica la variabilità
interna ai gruppi (*within*)
individuati dai trattamenti

Il procedimento del test confronta gli scostamenti *between*
(che potrebbero contenere anche gli effetti sistematici) con
quelli *within* che sono dovuti **solo** al caso.



La scomposizione della variabilità intorno alla media generale

Per i dati del trattamento i

$$\sum_{j=1}^n (y_{ij} - \bar{y})^2 = \sum_{j=1}^n [(\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)]^2 = n(\bar{y}_i - \bar{y})^2 + \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2$$

Per tutti i dati

$$\underbrace{\sum_{i=1}^K \sum_{j=1}^n (y_{ij} - \bar{y})^2}_{\text{sst}} = \underbrace{\sum_{i=1}^K n(\bar{y}_i - \bar{y})^2}_{\text{ssb}} + \underbrace{\sum_{i=1}^K \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2}_{\text{ssw}}$$

sum of squares
total

sum of squares
between

sum of squares
within

I residui del modello e la stima della varianza σ^2

I residui del modello sono:

$$r_{ij} = y_{ij} - \bar{y}_i$$

La stima (da stimatore **non distorto** anche se H_0 è falsa) di σ^2

$$msw = \frac{\sum_{i=1}^K \sum_{j=1}^n r_{ij}^2}{Kn - K} = \frac{\sum_{i=1}^K \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2}{Kn - K} = \frac{ssw}{Kn - K}$$

msw: mean squares within

Kn-K: gradi di libertà associati alla devianza within

La devianza between

Se H_0 è vera, possiamo ricavare un'altra stima di σ^2 da stimatore **non distorto**:

$$msb = \frac{\sum_{i=1}^K \sum_{j=1}^n (\bar{y}_i - \bar{y})^2}{K-1} = \frac{\sum_{i=1}^K n(\bar{y}_i - \bar{y})^2}{K-1} = \frac{ssb}{K-1}$$

msb: mean squares between

$K-1$: gradi di libertà associati alla devianza between (fra gruppi o fra medie)

Il test dell'ANOVA

$$\begin{cases} H_0: \mu_1 = \mu_2 = \dots = \mu_K = \mu \\ H_1: \text{almeno una media è diversa dalle altre} \end{cases}$$

Statistica test:

$$F_{oss} = \frac{msb}{msw} = \frac{ssb/(K-1)}{ssw/(Kn-K)}$$

Si rifiuta H_0 se $F_{oss} > F_\alpha$ dove:

$P(F_{(K-1),(Kn-K)} > F_\alpha) = \alpha$ (probabilità di rifiutare H_0 quando è vera)

$F_{(K-1),(Kn-K)}$ è la v.c. di Fisher con $(K-1), (Kn-K)$ gradi di libertà.

Tavola dell'ANOVA (ANOVA table)

Effetti	Devianza (sum of squares)	Gradi di libertà	Varianza (mean squares)	F di Fisher (F_{oss})	p -level
Trattamenti	ssb	$K-1$	msb	msb/msw	$P(F_{K-1,Kn-K} > F_{oss})$
Errore	ssw	$Kn-K$	msw		
Totale	sst	$Kn-1$			

stima di σ^2

NOTA BENE. Posto F_α tale che $P(F_{(K-1),(Kn-K)} > F_\alpha) = \alpha$,
 si rifiuta H_0 se $F_{oss} > F_\alpha$
 ovvero
 si rifiuta H_0 se $p\text{-level} < \alpha$

} è equivalente

I gradi di libertà (gdi)

$$sst = \sum_{i=1}^K \sum_{j=1}^n (y_{ij} - \bar{y})^2 \quad \rightarrow \quad \boxed{gdl = Kn - 1}$$

Kn: nr. totale unità

$$ssb = \sum_{i=1}^K \sum_{j=1}^n (\bar{y}_i - \bar{y})^2 \quad \rightarrow \quad \boxed{gdl = K - 1}$$

$$ssw = \sum_{i=1}^K \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2 \quad \rightarrow \quad \boxed{gdl = Kn - K}$$

K: n. trattamenti, n: nr. replicazioni per trattamento

Esempio 1 – ANOVA e regola del *p-value*

$\alpha=0.05$

<i>Effetti</i>	<i>SS</i>	<i>gdl</i>	<i>Varianza</i>	<i>F-value</i>	<i>p-value</i>
Trattamenti	0.0529	2	0.0264	0.1190	0.8892
Errore	1.9994	9	0.2222		
Totale	2.0523	11			

$p\text{-value} > \alpha \rightarrow$ si accetta H_0 e si conclude che le medie sono uguali

N.B. *F-value*: *F* osservato

Esempio 2 – ANOVA e regola del *p-value*

$\alpha=0.05$

<i>Effetti</i>	<i>SS</i>	<i>gdl</i>	<i>Varianza</i>	<i>F-value</i>	<i>p-value</i>
Trattamenti (between)	8.1105	2	4.0552	74.8966	2.5E-06
Errore (within)	0.4873	9	0.0541		
Totale	8.5978	11			

p-value < $\alpha \rightarrow$ si rifiuta H_0 e si conclude che **almeno 2 medie** sono diverse fra loro ovvero che il fattore influenza lo spessore del gres.

Quali medie sono diverse fra loro ?

Quando rifiutiamo H_0 concludiamo che **almeno 2 medie sono diverse** (H_1) e **NON** che *tutte le medie sono diverse fra loro*.

E' necessario condurre l'analisi ***post-hoc*** (confronti multipli).

Analisi post-hoc: confronto a coppie delle medie mediante ***t-test*** per verificare l'uguaglianza delle medie a 2 a 2.

Con K livelli del fattore, abbiamo $K(K-1)/2$ coppie di medie da confrontare e quindi con $K=3$ abbiamo 3 *t-test* da condurre

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

$$H_0: \mu_1 = \mu_3$$

$$H_1: \mu_1 \neq \mu_3$$

$$H_0: \mu_2 = \mu_3$$

$$H_1: \mu_2 \neq \mu_3$$

Fattore: temperatura.

Livello 1: bassa ; Livello 2: media; Livello 3: alta

t-test e analisi post hoc

$$\left\{ \begin{array}{l} H_0: \mu_i - \mu_h = 0 \quad i \neq h, \quad i, h = 1, \dots, K \\ H_1: \mu_i - \mu_h \neq 0 \end{array} \right.$$

Dal corso di Statistica 1 sappiamo che, in caso di varianze note, la distribuzione di riferimento e la statistica test sono:

$$\bar{Y}_i - \bar{Y}_h \sim N\left(0, \frac{\sigma_i^2}{n_i} + \frac{\sigma_h^2}{n_h}\right) \quad Z = \frac{\bar{y}_i - \bar{y}_h}{\sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_h^2}{n_h}}}$$

Se non conosciamo σ^2 , ma possiamo ipotizzare che sia la stessa per le due popolazioni, possiamo, come sappiamo, usare *msw* per stimarla e ricorrere a una T di Student (con i gdl di *msw*).

Pertanto le formule qui sopra diventano

$$\bar{Y}_i - \bar{Y}_h \sim N\left(0; 2 \frac{\sigma^2}{n}\right)$$

$$t_{oss} \sim \frac{\bar{y}_i - \bar{y}_h}{\sqrt{\frac{2 msw}{n}}}$$

sotto H_0 è la determinazione di una v.c t_{nK-K}

Fasi per condurre un'ANOVA con 1 fattore sperimentale

1. Effettuare il test F

1.1 Se l'ipotesi nulla (uguaglianza di tutte le medie) viene rifiutata → analisi post-hoc

1.2 Se l'ipotesi nulla non viene rifiutata ci sono 2 possibilità:

i) le medie sono uguali fra loro (H_0 è vera)

ii) le medie sono diverse fra loro ma il test non è stato in grado di dimostrarlo.

1.2 Si accetta H_0 : caso i)

Se si sospetta che *le medie siano uguali fra loro (caso i)*, ci sono due possibilità.

- a) Si ritiene che il fattore non abbia effetti sulla variabile risposta e quindi tale fattore non incide sulle prestazioni del processo;
- b) Le medie associate ai livelli scelti nell'esperimento sono uguali fra loro ma **si ritiene che il fattore abbia effetti**.

In questo caso si conduce un **nuovo esperimento** scegliendo altri livelli del fattore.

1.2 Si accetta H_0 : caso ii)

Se si sospetta che le medie associate ai livelli scelti del fattore sono diverse fra loro ma il test non è stato in grado di dimostrarlo (**caso ii**) ci due possibilità.

a) Il test F ha fornito un valore F_{oss} inferiore a F_α perché si è ottenuta una stima troppo elevata della varianza di errore (denominatore di F_{oss}) a causa di una bassa efficienza dello stimatore.

In questo caso si conduce un **nuovo esperimento** con un **numero di replicazioni** maggiore.

b) Il test F ha fornito un valore F_{oss} inferiore a F_α perché si è ottenuta una stima troppo elevata della varianza di errore (denominatore di F_{oss}) a causa dell'effetto di **fattori intervenienti** che non sono stati considerati o controllati nell'esperimento.

In questo caso si passerà all'**ANOVA e più vie**.