# Corso di Modelli per l'analisi statistica

Prof. G. Scandurra

a.a. 2020-2021

# Analisi di regressione

Con la regressione, si studia la dipendenza di una variabile Y da un'altra variabile X

Correlazione vs. Regressione

La correlazione studia l'associazione TRA due variabili attraverso misure "simmetriche" di interdipendenza (concordanza o discordanza)

La regressione individua una variabile come dipendente dall'altra e analizza la relazione di dipendenza della prima dalla seconda

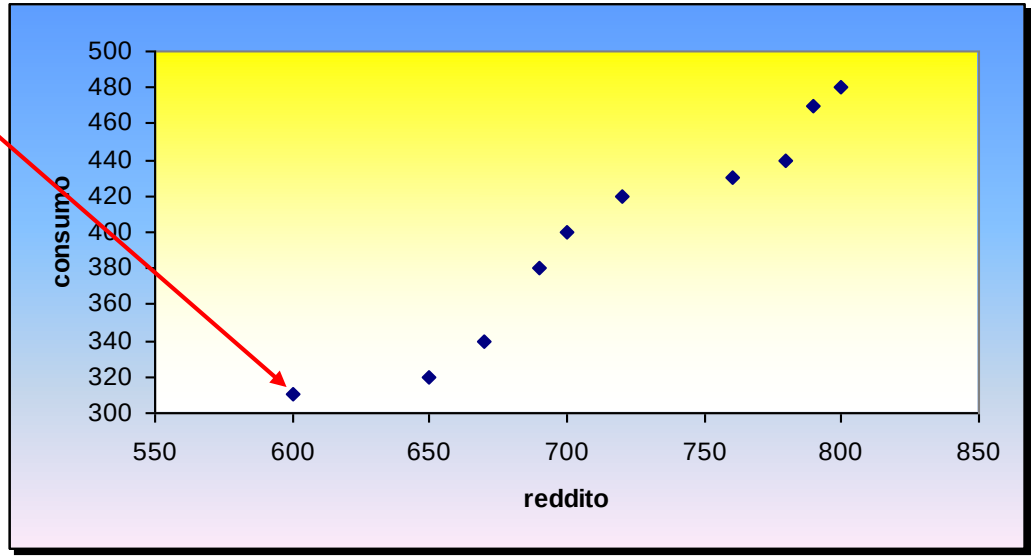
# Finalità della regressione

- Descrivere e interpretare la relazione di dipendenza di una variabile dall'altra
- Prevedere valori di una variabile utilizzando valori dell'altra variabile

In un'ottica descrittiva è un semplice metodo di interpolazione dei dati:
attraverso una funzione matematica (l'equazione di una retta) cerca di descrivere "al meglio" la nuvola di punti osservata nel grafico di dispersione

Esempio1: in un campione di 10 famiglie italiane si rilevano il reddito mensile (X) e il consumo mensile per generi alimentari (Y), in euro

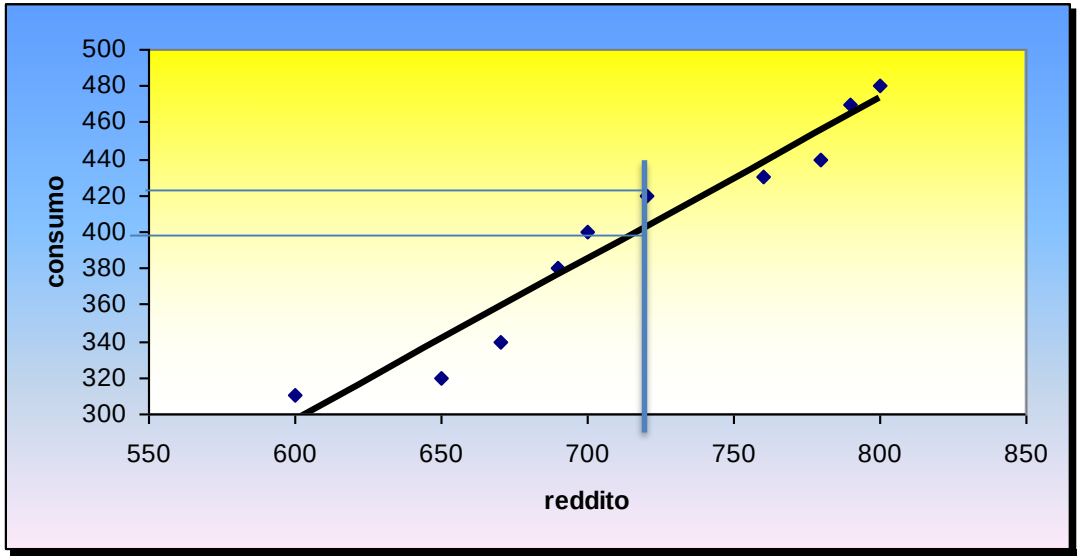
| | |
|---|---|
| 670 | 340 |
| 690 | 380 |
| 700 | 400 |
| 720 | 420 |
| 760 | 430 |
| 780 | 440 |
| 790 | 470 |
| 800 | 480 |
| | |
| | |
| | |




Il grafico di dispersione suggerisce la presenza di un legame lineare di tipo crescente

Con l’analisi di regressione lineare si deve stimare quella retta che descrive meglio la nuvola di punti evidenziata dal grafico

La retta è stimata quando conosciamo il valore dell’intercetta all’origine e del coefficiente angolare (pendenza)



# Caso di studio

Una catena di negozi deve prendere decisioni sull'apertura di un nuovo punto vendita

Può scegliere tra locali di diversa ampiezza

La decisione può essere basata sulla relazione di dipendenza delle vendite annuali dalla superficie dei punti vendita attualmente esistenti

Ci si aspetta che al crescere della superficie aumentino anche le vendite

Quale valore delle vendite mi posso aspettare per un negozio di 30 mq?

Se la differenza della superficie di due negozi è di 10 mq, quale differenza nelle vendite mi posso aspettare?

# Regressione lineare semplice

La relazione che esprime la dipendenza lineare di Y da X è:

$$Y = f(X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$$

Y è la variabile dipendente (o risposta)
X è la variabile esplicativa (o indipendente)
ε è il termine di errore (componente casuale)
$\beta_0$ e $\beta_1$ sono i parametri da stimare

# Coefficienti di regressione

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

$\beta_0$ (intercetta) è il valore che assume la variabile dipendente Y quando la variabile esplicativa X=0

$\beta_1$ (coefficiente angolare) indica di quanto varia in media la variabile dipendente Y per un incremento unitario di X

# Perché si introduce la componente di errore nel modello

- Negli studi empirici la relazione tra due variabili non è mai una relazione funzionale esatta del tipo $Y=f(X)$
- Comportamenti economici e sociali non sono descritti adeguatamente da relazioni che fanno corrispondere ad un dato valore di X un unico valore di Y

Esempio:

Nello studio della relazione di dipendenza del consumo familiare (Y) dal reddito familiare (X), è ragionevole ipotizzare che famiglie con lo stesso reddito abbiano comportamenti di consumo differenti

# Cosa rappresenta la componente di errore

Il termine di errore ε tiene conto di ogni altro fattore (non osservato o non osservabile) che, oltre alla variabile esplicativa, può influenzare la risposta Y

Esempio:

Il consumo delle famiglie può dipendere, oltre che dal reddito disponibile, anche dal numero di componenti, dalla loro età e dal livello di istruzione

# Regressione lineare semplice

Obiettivo:

stimare i coefficienti $\beta_0$ e $\beta_1$ (quindi stimare l'equazione di una retta del tipo $Y=\beta_0+\beta_1 X$) e la grandezza dell'errore $\varepsilon$

utilizzando le osservazioni sulle variabili X e Y sotto forma di n coppie di valori, del tipo

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

# Stima dei coefficienti di regressione

Una qualsiasi retta di equazione

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

sul piano (X,Y) è una stima della relazione lineare ipotizzata

Tra tutte le rette, vogliamo individuare quella che "si adatta meglio" ai dati, cioè quella che passa "più vicino" alla nuvola di punti del diagramma

# Metodo dei minimi quadrati

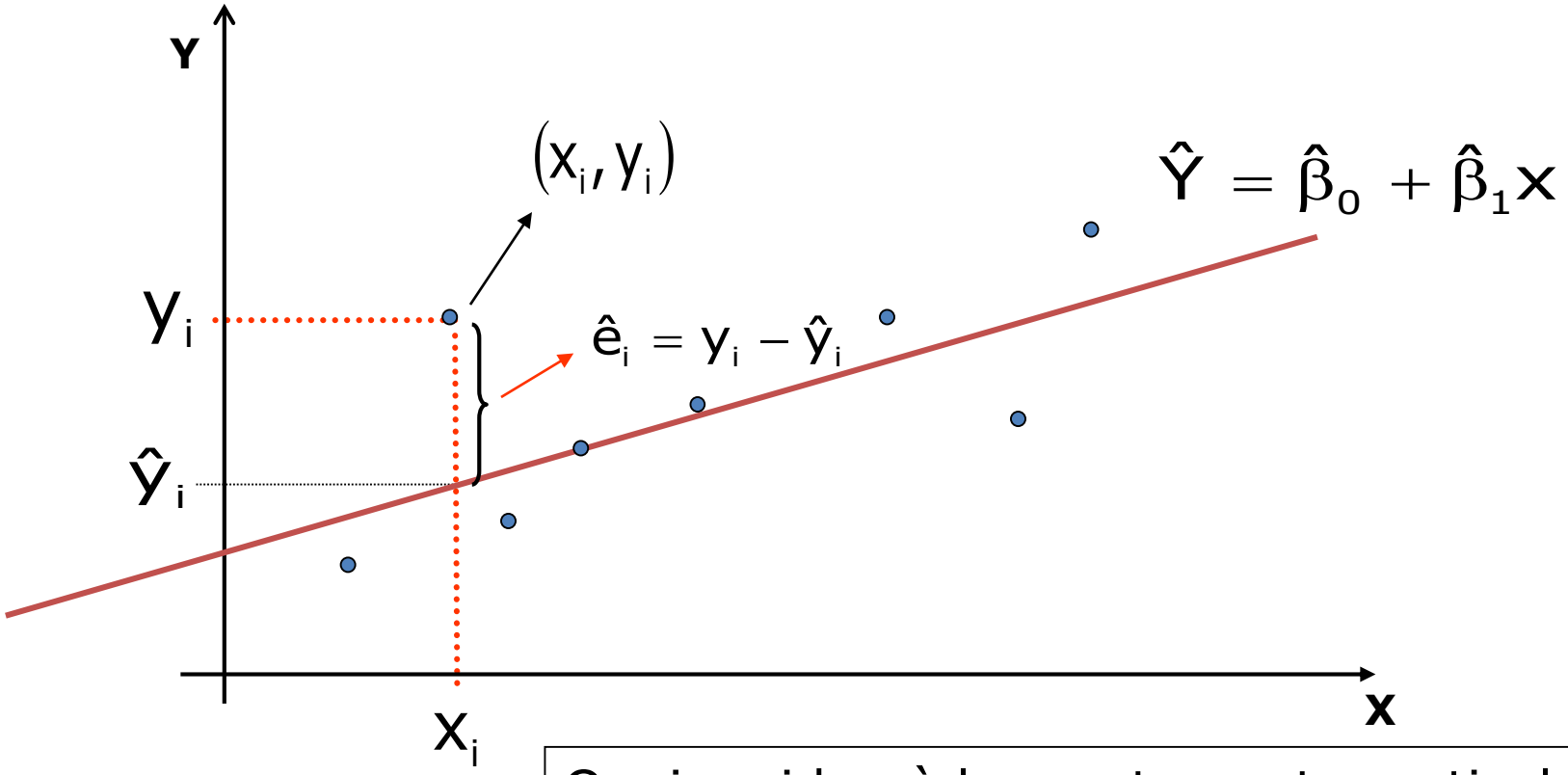
Per ogni *i=1…n* definiamo il residuo come la differenza tra il valore osservato di Y e il corrispondente valore stimato

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Si stimano $\beta_0$ e $\beta_1$ in modo tale da minimizzare la funzione G, ossia la somma dei quadrati dei residui

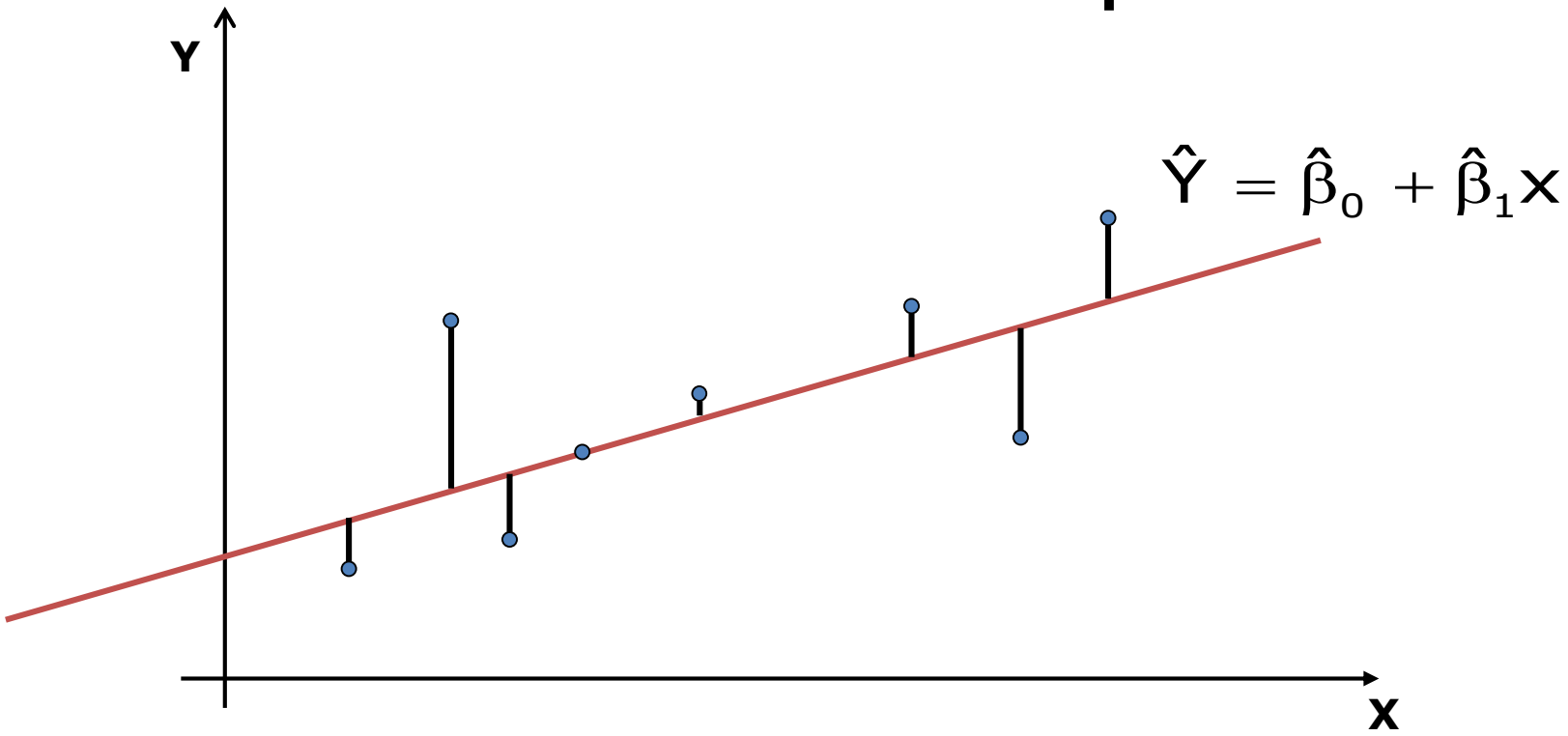
$$G(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$$

# Rappresentazione grafica del residuo




Ogni residuo è lo scostamento verticale tra il valore osservato e il corrispondente valore sulla retta

# Rappresentazione grafica del metodo dei minimi quadrati




La retta si individua minimizzando la funzione G, ossia la somma dei quadrati di tutti gli scostamenti verticali

# Stima dei parametri della retta

Si dimostra che i valori dei parametri $\beta_0$ e $\beta_1$ che minimizzano la funzione G sono:

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

È il coefficiente angolare della retta (pendenza)

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

È l'intercetta all'origine

# Riprendiamo l'esempio consumo-reddito

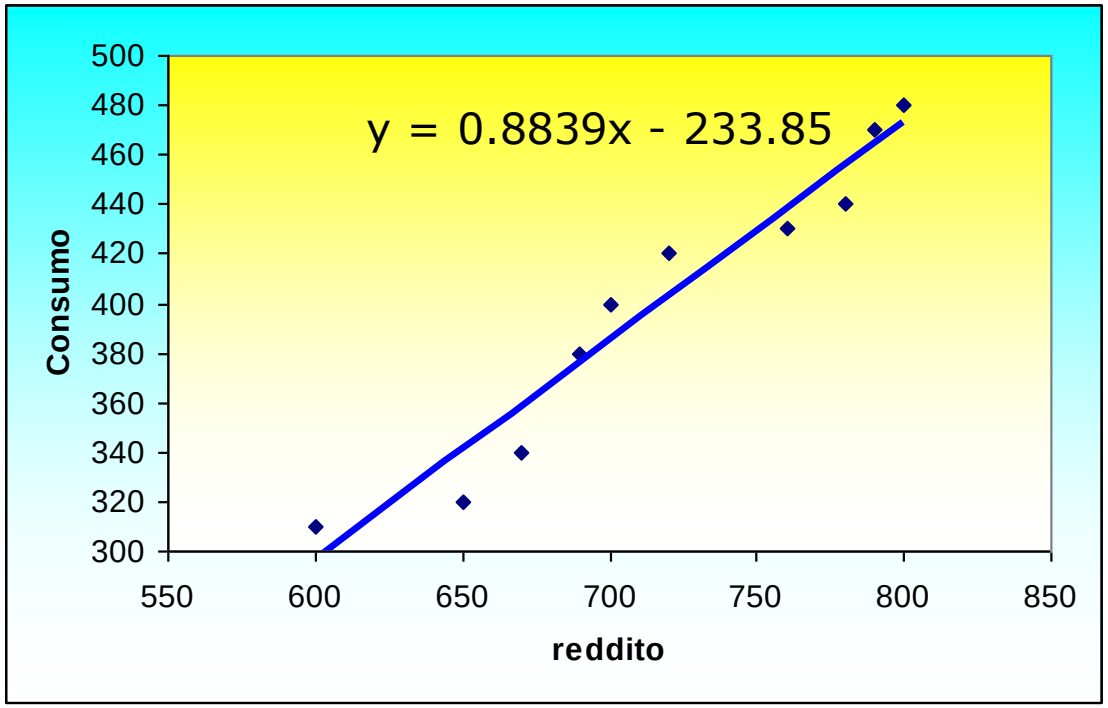
| X | Y | $y_i - \bar{y}$ | $x_i - \bar{x}$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|
| 600 | 310 | -89 | -116 | 10324 | 13456 |
| 650 | 320 | -79 | -66 | 5214 | 4356 |
| 670 | 340 | -59 | -46 | 2714 | 2116 |
| 690 | 380 | -19 | -26 | 494 | 676 |
| 700 | 400 | 1 | -16 | -16 | 256 |
| 720 | 420 | 21 | 4 | 84 | 16 |
| 760 | 430 | 31 | 44 | 1364 | 1936 |
| 780 | 440 | 41 | 64 | 2624 | 4096 |
| 790 | 470 | 71 | 74 | 5254 | 5476 |
| 800 | 480 | 81 | 84 | 6804 | 7056 |
| 716 | 399 | | | 34860 | 39440 |

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

$$\hat{\beta}_0 = 399 - 0{,}884 * 716 = -233{,}8$$

$$\hat{\beta}_1 = \frac{34860}{39440} = 0{,}884$$

# Rappresentazione grafica della retta




Per un incremento unitario di reddito il consumo aumenta in media di 0,8839

(se il reddito aumenta di 1000€, il consumo in media cresce di 884€)

$$\hat{\beta}_0 = -233{,}85$$
$$\hat{\beta}_1 = 0{,}8839$$

# Valori previsti e residui

Valori osservati di X

Valori osservati di Y

Valori stimati (previsti) di Y

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

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

| *X* | *Y* | *Y prevista* | *Residui* |
|---|---|---|---|
| 600 | 310 | 296,47 | 13,53 |
| 650 | 320 | 340,66 | -20,66 |
| 670 | 340 | 358,34 | -18,34 |
| 690 | 380 | 376,02 | 3,98 |
| 700 | 400 | 384,86 | 15,14 |
| 720 | 420 | 402,54 | 17,46 |
| 760 | 430 | 437,89 | -7,89 |
| 780 | 440 | 455,57 | -15,57 |
| 790 | 470 | 464,41 | 5,59 |
| 800 | 480 | 473,25 | 6,75 |

# Stima della risposta media

Le stime dei parametri della retta possono essere utilizzate per stimare il valore medio di Y per un dato valore di X

La stima è data da $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Ad es., in base alla retta $\hat{y}_i = -233{,}8 + 0{,}884 X_i$ possiamo stimare il valore medio del consumo Y quando il reddito X=750

$$\hat{y}_i = -233{,}8 + 0{,}884 \cdot 750 = 429{,}2$$

# Capacità esplicativa della retta

Una volta stimati i coefficienti del modello, serve una misura che permetta di valutare se il modello lineare specificato si adatti bene ai dati osservati

L'indice che misura la bontà di adattamento della retta di regressione ai dati è il coefficiente di determinazione $R^2$, che si definisce a partire dalla scomposizione della devianza totale della variabile risposta

# Scomposizione della devianza totale

Si può dimostrare che vale la seguente relazione:

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

SQT= Somma Quadrati Totale

SQR= Somma Quadrati Regressione

SQE= Somma Quadrati Errore

# Capacità esplicativa della retta

L’obiettivo della regressione è quello di studiare come varia la var. dipendente Y al variare della var. indipendente X.

Ci interessa capire a che cosa è dovuta la variabilità di Y.

Prima della regressione la variabilità di Y è misurata dalla devianza SQT(somma dei quadrati degli scarti delle osservazioni Y dalla loro media).

# Capacità esplicativa della retta

Dopo la regressione, avendo introdotto le osservazioni della variabile X e stimato il modello, la variabilità non spiegata di Y (da intendere come il grado di incertezza residua data la funzione di regressione stimata) è misurata da SQE

SQE è la variabilità dei valori Y rispetto alla retta stimata

# Capacità esplicativa della retta

SQE=0 quando tutti i punti osservati stanno sulla retta e quindi la retta si adatta perfettamente ai dati e il modello è buono per fare previsioni.

In questo caso SQR è massima e pari a SQT.

SQR rappresenta il vantaggio derivante dalla stima della funzione di regressione, in termini di riduzione dell'incertezza residua

Maggiore è SQR, maggiore è la riduzione della variabilità residua (maggiore è la variabilità della Y spiegata dalla relazione di Y con X)

# Capacità esplicativa della retta

Al crescere di SQE aumenta la dispersione dei valori di Y intorno alla retta stimata e peggiora quindi l'adattamento del modello ai dati

SQE raggiunge il valore massimo quando SQR=0 cioè quando i valori stimati di Y giacciono sulla retta parallela all'asse X (Y=y medio), cioè quando la retta "migliore" per rappresentare l'insieme dei punti è la retta parallela all'asse X (in media Y rimane costante al variare di X)

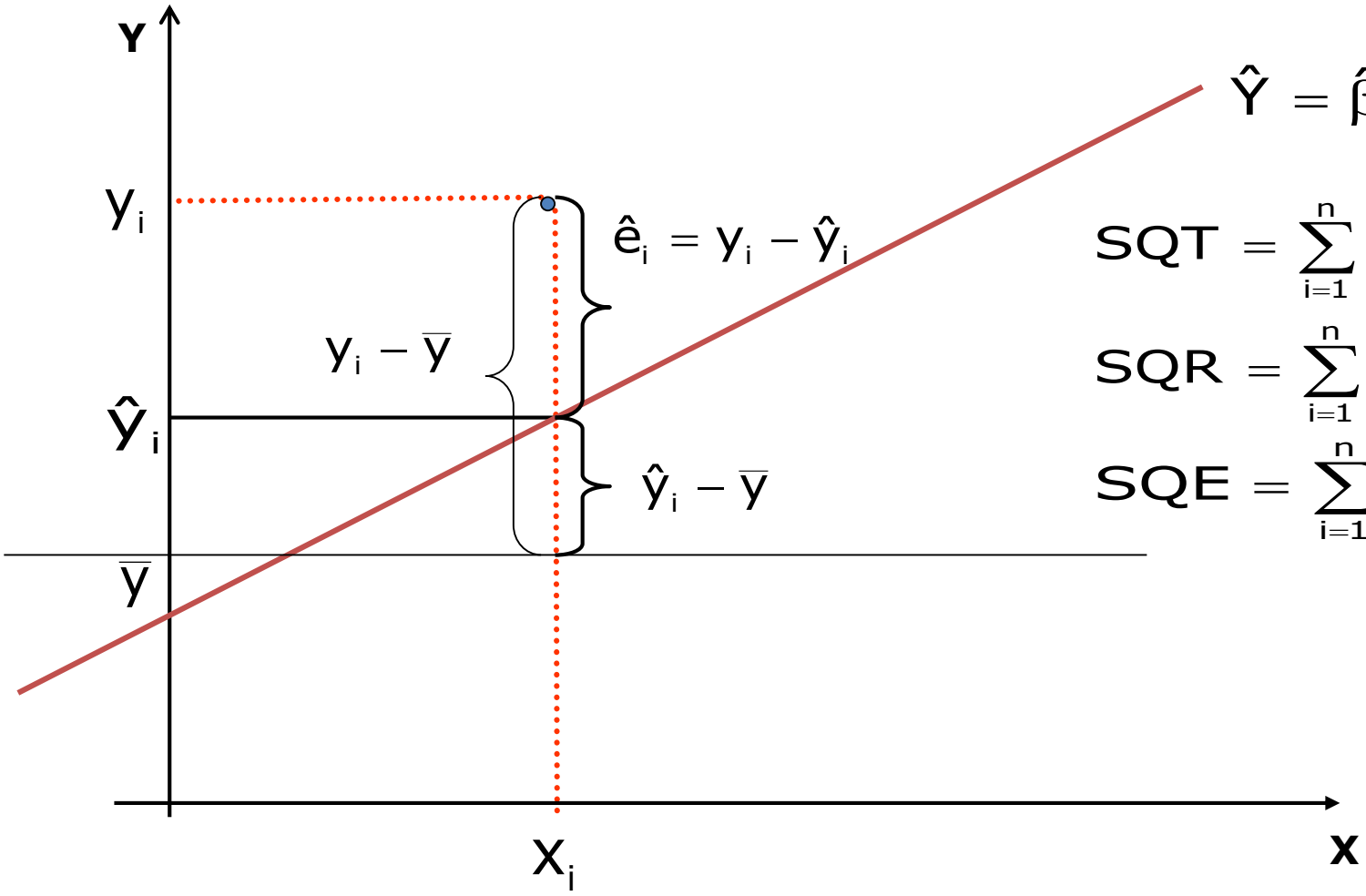
# Capacità esplicativa della retta

Si introduce un indice per misurare la bontà dell'adattamento, dato dal rapporto tra SQR e SQT

R2=0 quando SQR=0 (assenza di relazione lineare tra X e Y)

R2=1 quando SQE=0 (perfetta dipendenza lineare di Y da X)

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




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

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

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

$$SQE=\sum_{i=1}^{n}\hat{e}_i^2$$

# Coefficiente di determinazione

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

$R^2$ misura la quota di variabilità totale di Y spiegata dalla dipendenza lineare di Y da X

- $R^2=0$ in assenza di dipendenza lineare di Y da X (i punti osservati si dispongono casualmente sul piano (x,y) oppure evidenziano un legame non lineare) [SQR=0]
- $R^2=1$ nella situazione di perfetta dipendenza lineare (i punti osservati sono allineati su una retta) [SQE=0]

# Coefficiente di determinazione

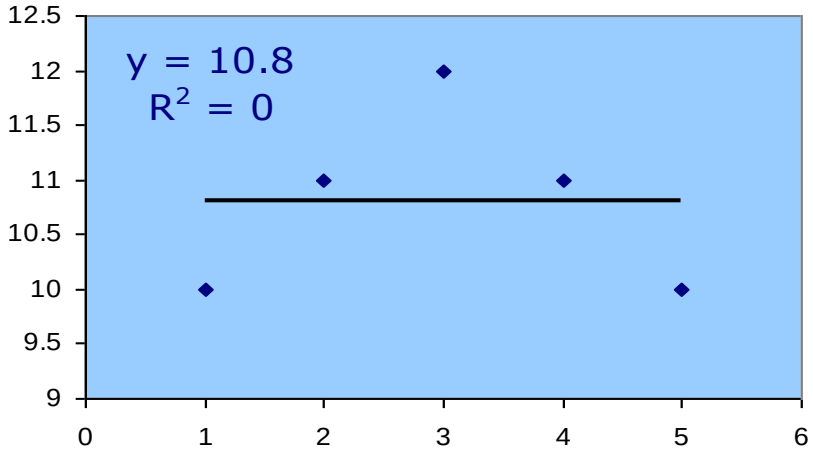
$$0 \leq R^2 \leq 1$$

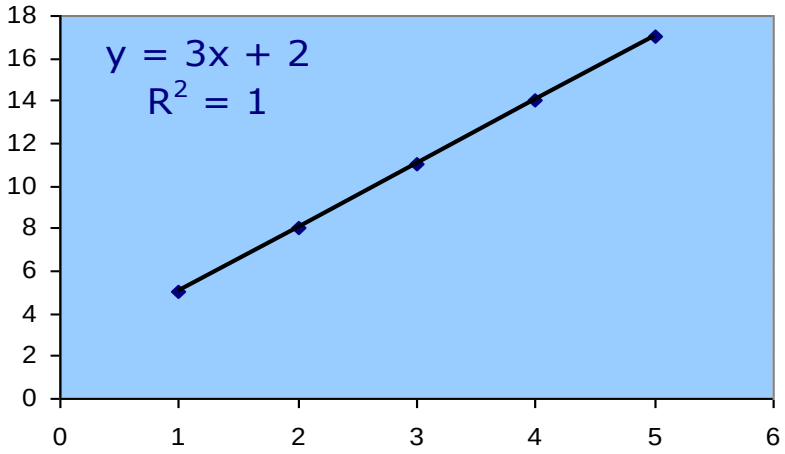
Più il valore di $R^2$ si avvicina a 1, minori sono le distanze dei punti osservati dalla retta e migliore quindi è l'adattamento della retta di regressione ai dati

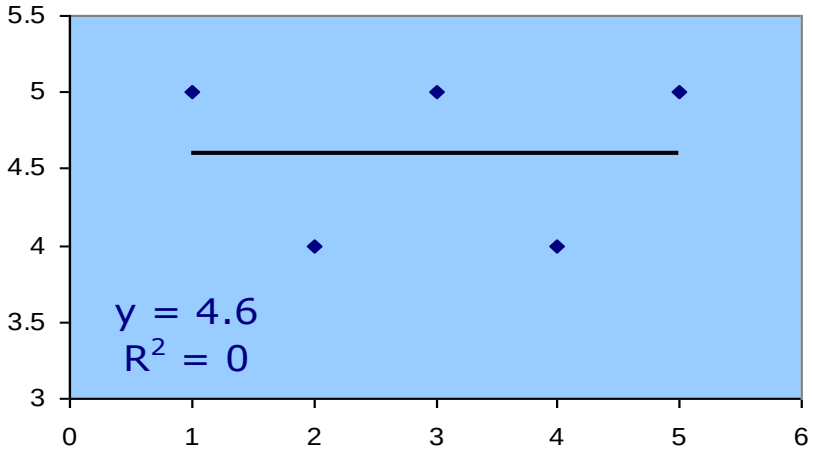
Si può dimostrare che

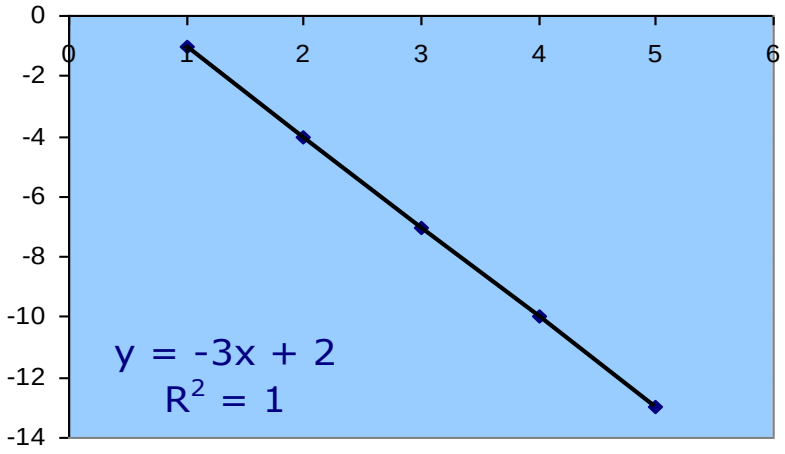
$$R^2 = (\rho_{xy})^2 = \left( \frac{\sigma_{xy}}{\sigma_x \sigma_y} \right)^2$$

# Esempi

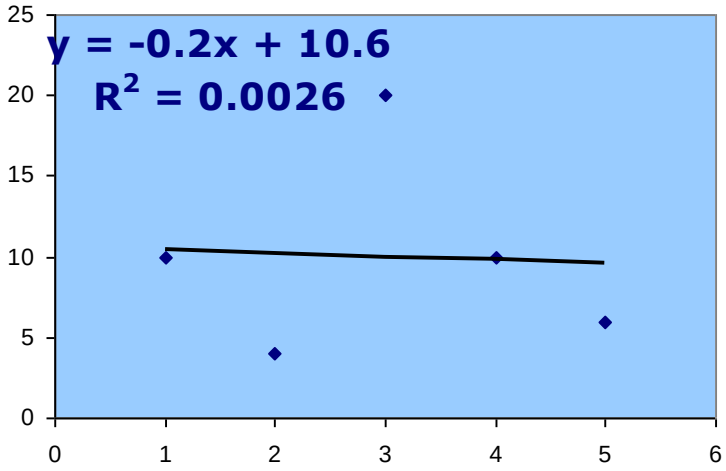

y = 10.8
R2 = 0
12.5
12
11.5
11
10.5
10
9.5
9
0
1
2
3
4
5
6

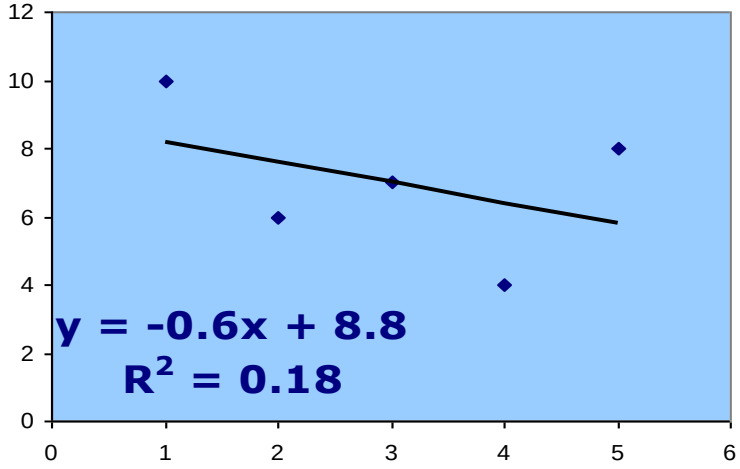

y = 3x + 2
R2 = 1
18
16
14
12
10
8
6
4
2
0
0
1
2
3
4
5
6

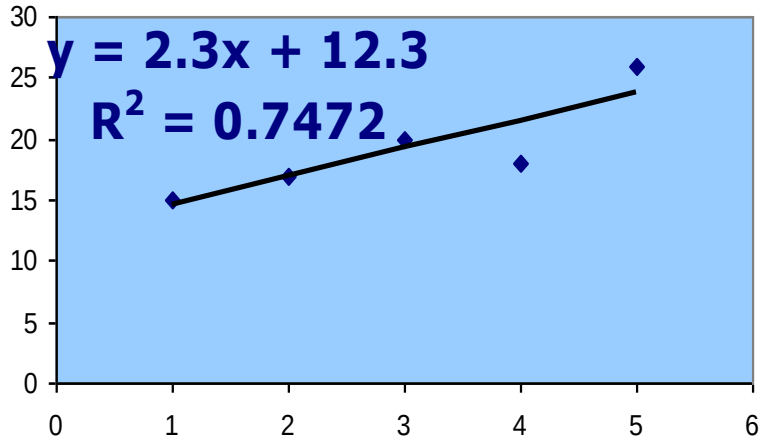

y = 4.6
R2 = 0
5.5
5
4.5
4
3.5
3
0
1
2
3
4
5
6

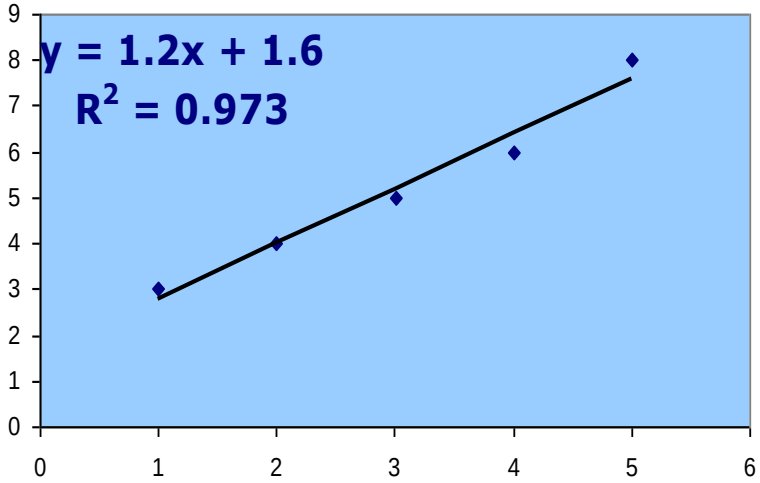

y = -3x + 2
R2 = 1
0
-2
-4
-6
-8
-10
-12
-14
0
1
2
3
4
5
6

# Esempi


y = -0.2x + 10.6
R² = 0.0026


y = -0.6x + 8.8
R² = 0.18


y = 2.3x + 12.3
R² = 0.7472


y = 1.2x + 1.6
R² = 0.973

ESERCIZIO

In un ipermercato di Napoli è stata svolta un'indagine per rilevare il prezzo del pane negli ultimi cinque mesi (in euro al Kg) e le quantità consumate in media in un giorno (in Kg)

| prezzo | quantità |
|---|---|
| 1,65 | 210 |
| 1,67 | 198 |
| 1,68 | 176 |
| 1,69 | 175 |
| 1,7 | 174 |

Stimare la retta di regressione che mette in relazione la quantità in funzione del prezzo

# Stima dei coefficienti di regressione - Output Excel

| | *Coefficienti* |
|---|---|
| Intercetta | 1529 |
| Variabile X 1 | -800 |

La retta stimata è Y=1529-800X
Un aumento di 1€ del prezzo al Kg del pane fa diminuire la quantità media di pane consumato giornalmente di 800 kg

| | |
|---|---|
| R al quadrato | 0,87 |

La dipendenza lineare è forte ($R^2$=0,87)

# Rappresentazione punti osservati e retta di regressione - Output Excel