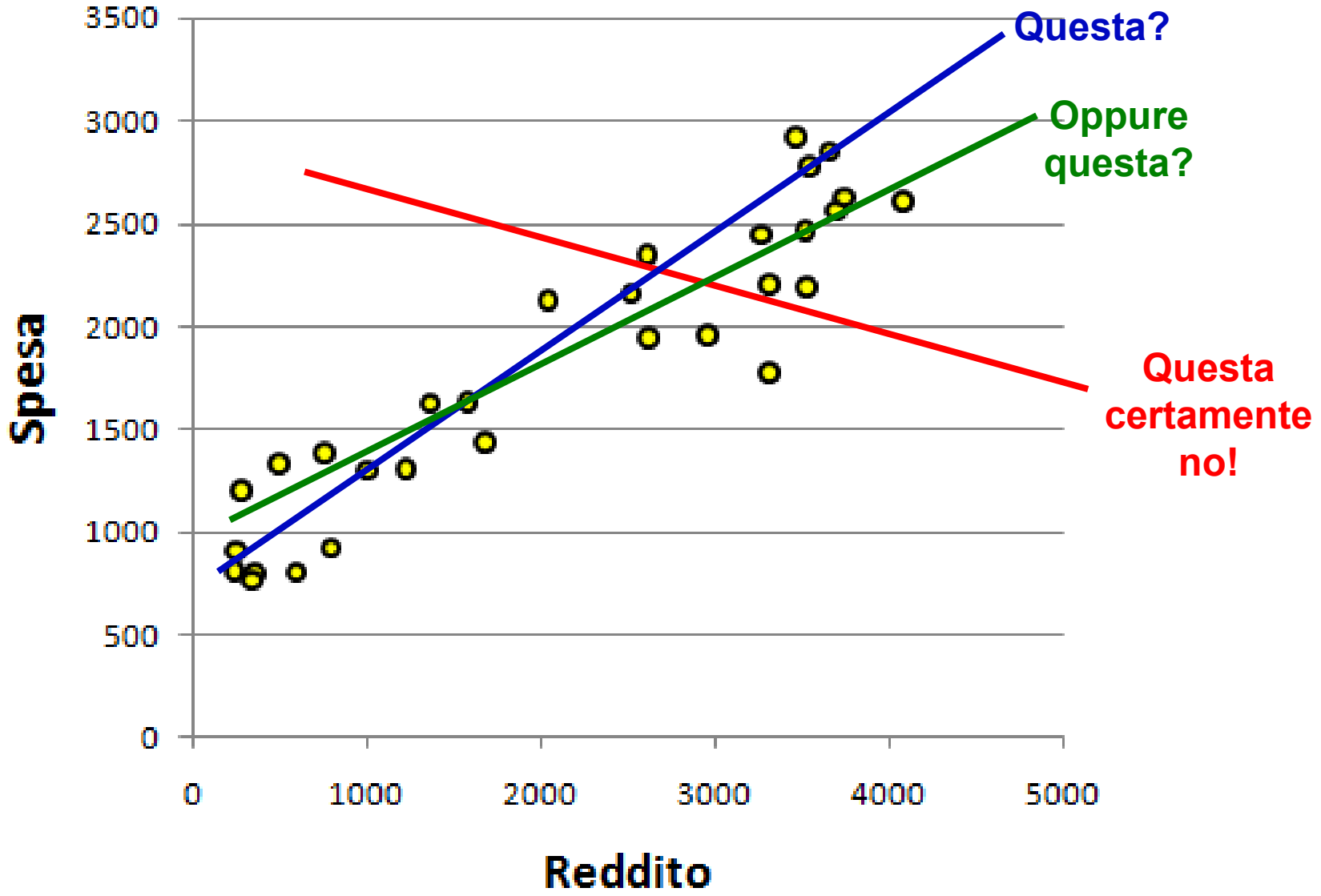
Quale retta ?



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

Metodo dei minimi quadrati

Retta stimata: $\hat{Y} = a + bX$

$\hat{Y}$ **ordinata teorica** corrispondente ad un dato valore di X

coefficiente a - **intercetta** – è l'**ordinata all'origine** della retta

coefficiente di regressione *b* è il **coefficiente angolare** della retta

retta di regressione stimata è tanto più adatta a descrivere la relazione studiata quanto più i **punti osservati** (Y_i) si trovano in prossimità di essa

Si definiscono **residui campionari**:

$$e_i = Y_i - \hat{Y}_i = Y_i - a - bX_i$$

Criterio dei minimi quadrati (**OLS**): *a* e *b* sono scelti in modo da **minimizzare la somma dei quadrati dei residui campionari**

$$f(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

Stima dei parametri con metodo dei minimi quadrati

$$f(a,b)=\sum_{i=1}^{n} e_i^2=\sum_{i=1}^{n}\left(Y_i-a-bX_i\right)^2$$ **= minimo**

$$\frac{\partial f(a,b)}{\partial a}=\frac{\partial f(a,b)}{\partial b}=0$$

Risolvendo le due derivate rispetto ad *a* e b, si ottiene:

$$\sum_{i=1}^{n} Y_i = na + b\sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2$$

$$a=\bar{Y}-b\bar{X}$$

$$b=\sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2$$

x_i e y_i indicano gli scarti delle osservazioni campionarie dal relativo valore medio

Proprietà dei minimi quadrati

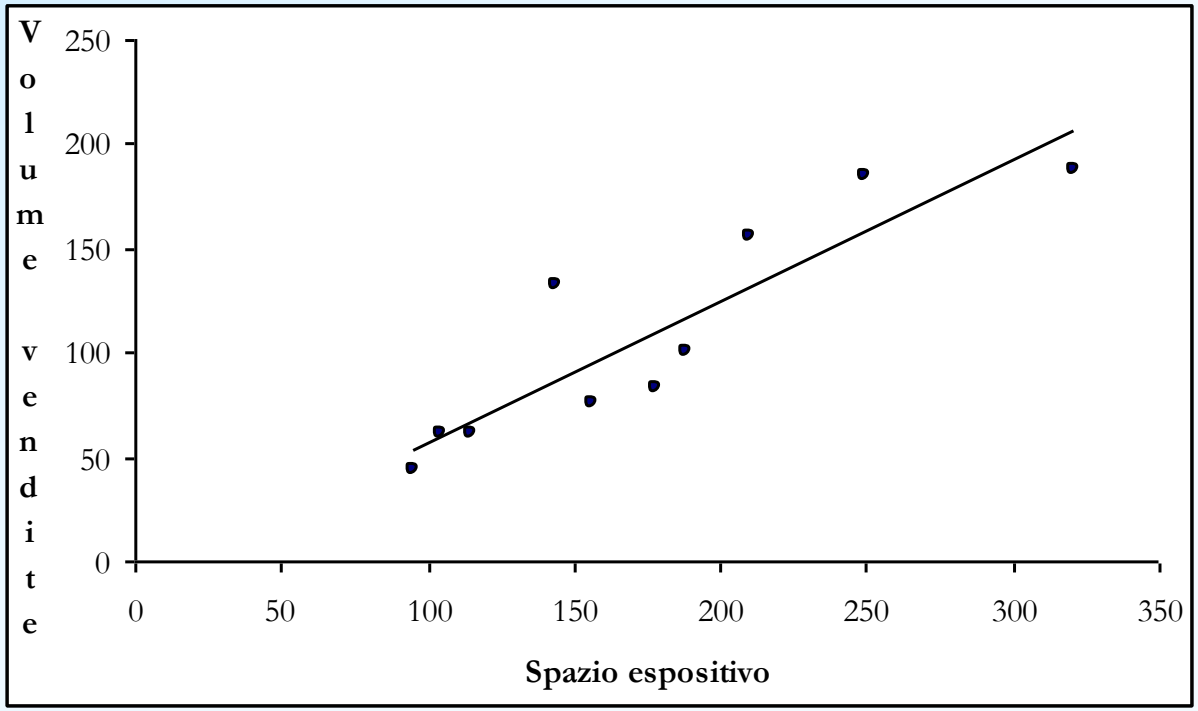
La somma dei valori teorici è uguale alla somma dei valori osservati: $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$

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:

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

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 $(\bar{x}, \bar{y})$

Stima retta di regressione esempio supermercati



$$\hat{Y} = -10{,}19 + 0{,}67 \cdot X$$

Che dire del valore dell'intercetta ?

***coefficiente di regressione**:*

ad ogni incremento unitario della variabile X la variabile Y subisce anch'essa un incremento, di intensità 0,67 – ovvero ad ogni incremento di un m^2 nella superficie del supermercato il volume delle vendite settimanali aumenta di 67 euro (0,67 x 100)

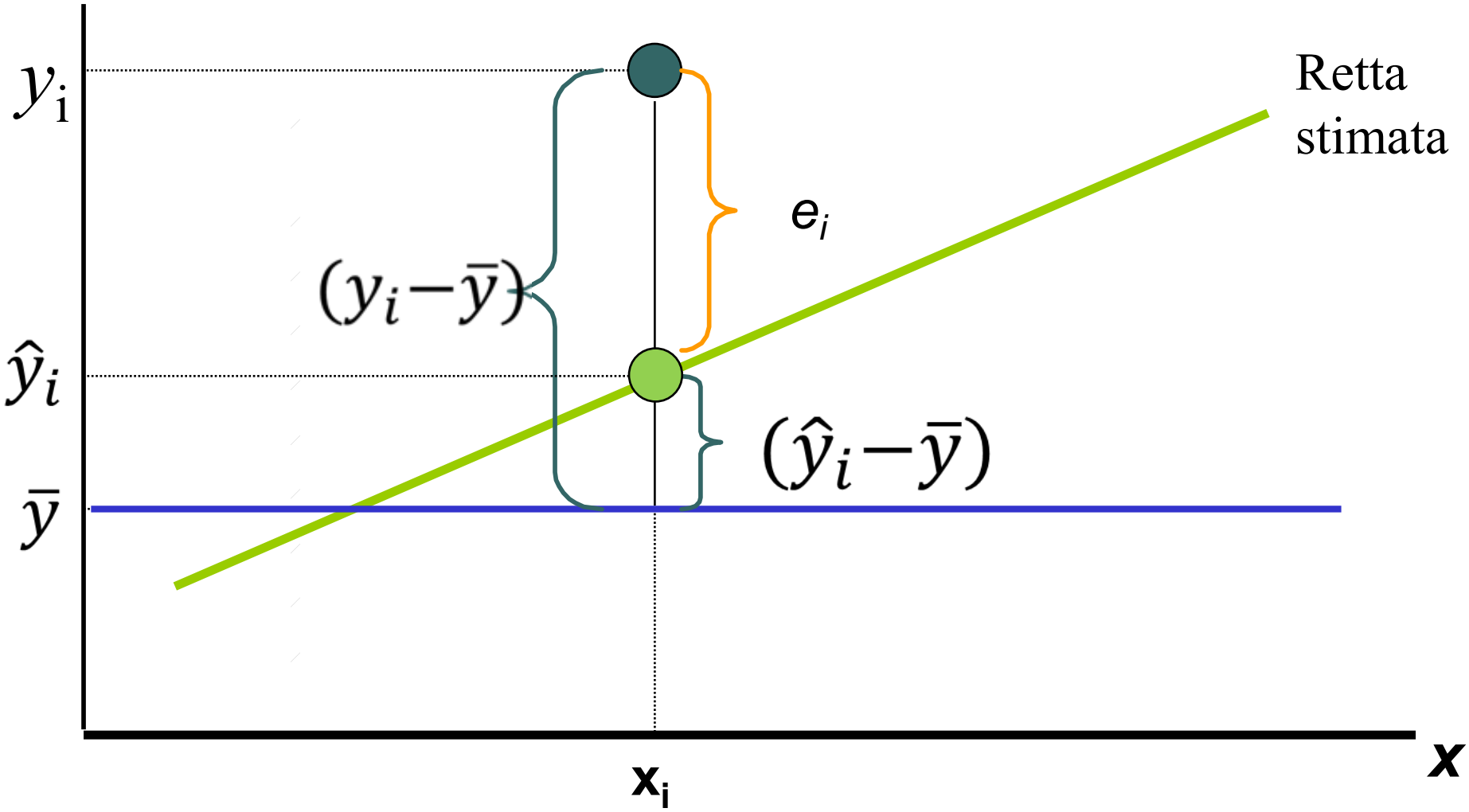
Bontà di adattamento

La retta (modello) stimata descrive **bene** i dati 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

Scomposizione dello scostamento dalla media di Y

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) = (\hat{y}_i - \bar{y}) + e_i$$



Scomposizione della devianza di Y

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + e_i$$

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

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

devianza **totale** dei valori della variabile dipendente

Misura la variazione dei valori di Y intorno alla loro media

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

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

Misura della bontà di adattamento

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:

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

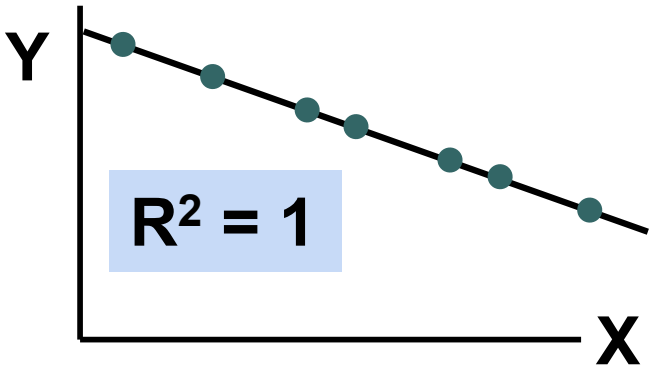
frazione di varianza della variabile risposta "spiegata" dal modello

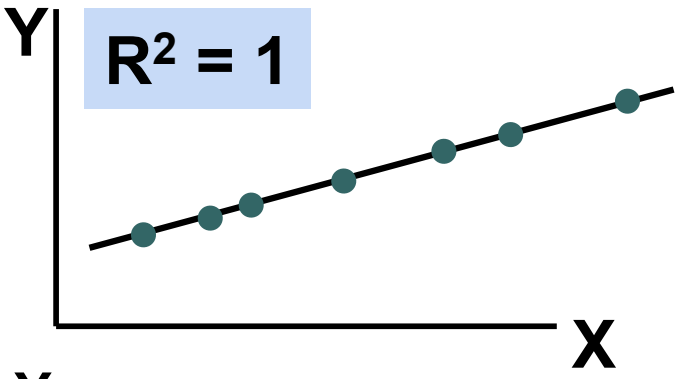
L'indice R^2, essendo un rapporto d'una parte al tutto, può assumere valori compresi tra 0 ed 1:

se $R^2 = 0$ l'adattamento è pessimo

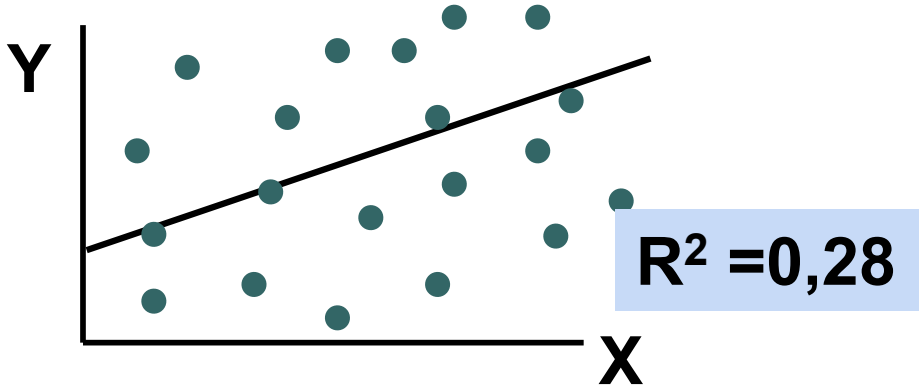
se $R^2 = 1$ l'adattamento è perfetto

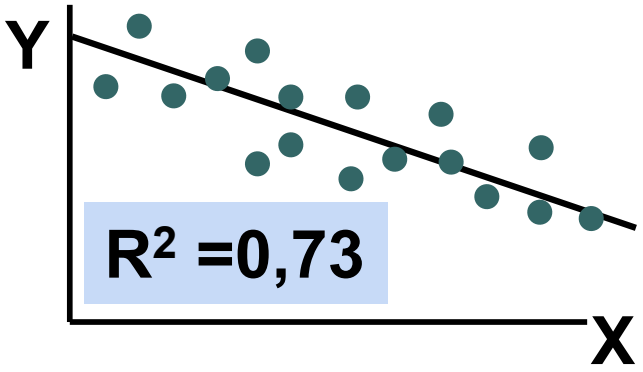
Possibili situazioni





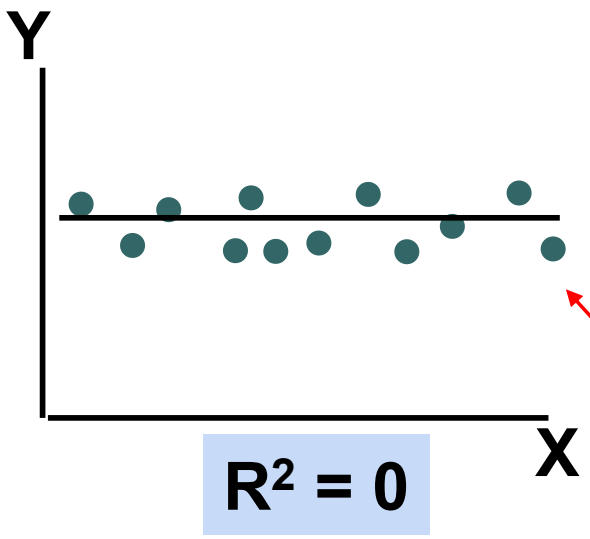
Relazione lineare perfetta fra X e Y:
Il 100% della variabilità di Y è spiegata dalla variabilità di X





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

Possibili situazioni



$R^2 = 0$

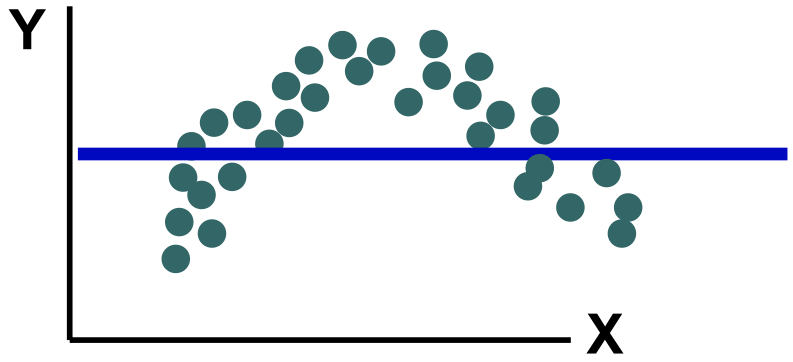
Nessuna relazione lineare fra X e Y:

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

se la situazione è quella indicata dal grafico !

Osservazioni su R^2 /1

1. $R^2=0$ e $R^2=1$ rappresentano dei casi limite che in pratica non si presentano mai
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

3. Non si utilizza il termini "**causata**" poiché un valore di R^2 prossimo ad 1 non implica necessariamente un nesso causale tra le due variabili.

Osservazioni su R^2 /2

R^2: indicazione ***globale*** su bontà di adattamento

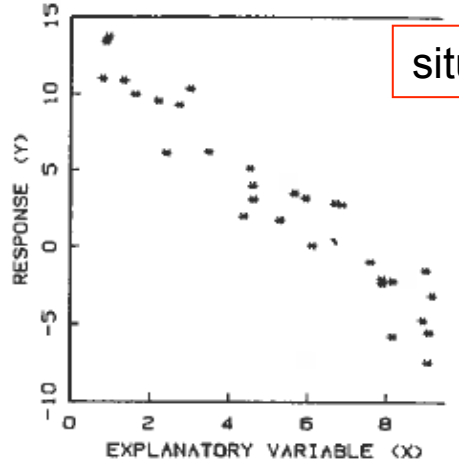
Analisi residui

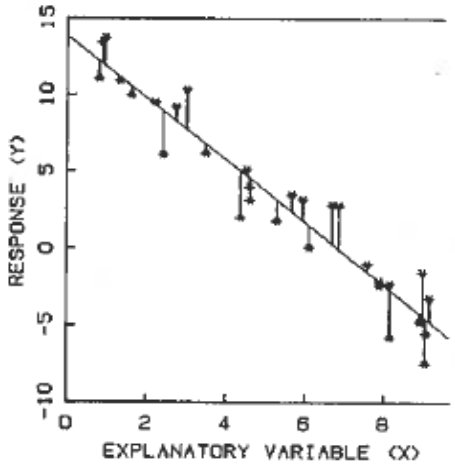
consente di identificare meglio eventuali cause di scarso/ non adattamento, dovute ad es. a relazioni non lineari (eventualmente linearizzabili) e/o presenza di outliers.

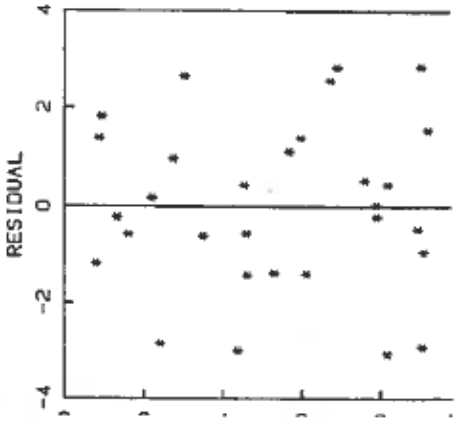
Grafico dei residui contro la variabile esplicativa (x_i, e_i): se l'analisi di regressione lineare è soddisfacente, i residui non mostrano 'regolarità' degne di interesse.

Possibili situazioni

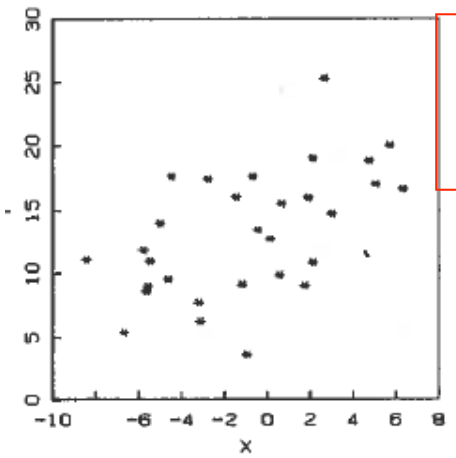
situazione ideale !

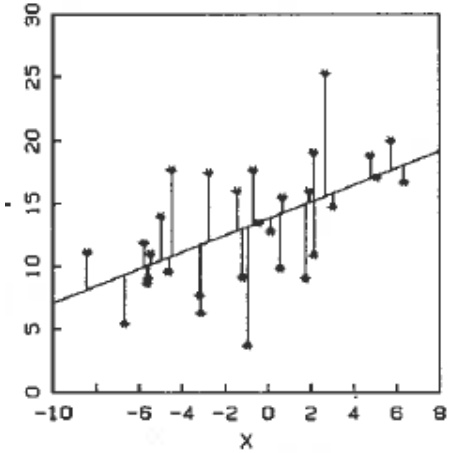


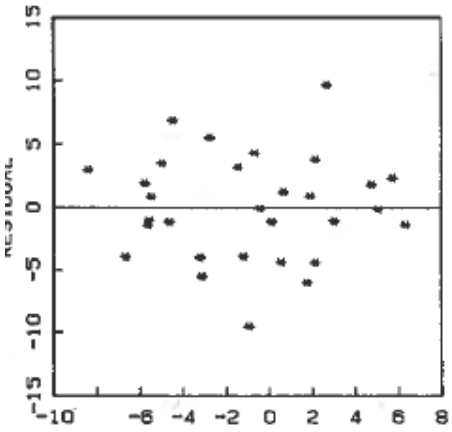




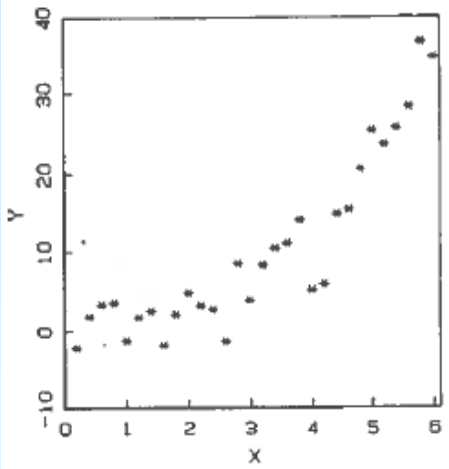
situazione simile ma relazione più debole: residui più dispersi



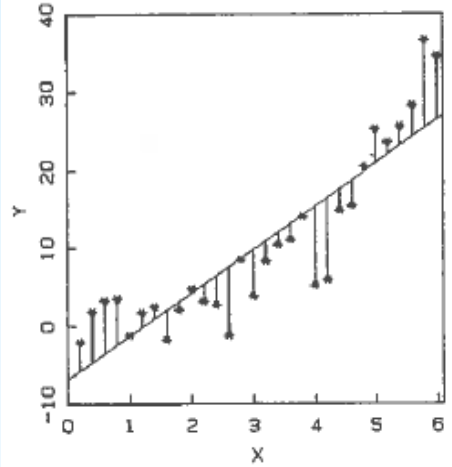


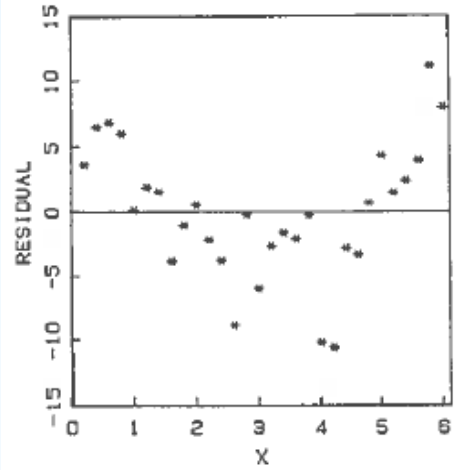


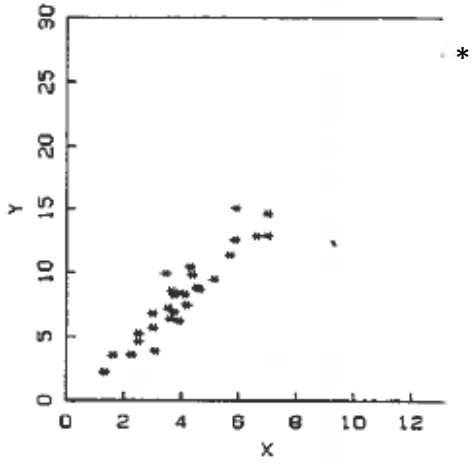
Fonte: Chambers et al., 1983



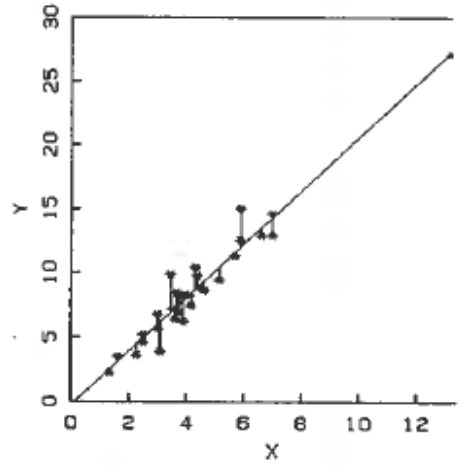
relazione curvilinea: curvatura ancora più evidente dei residui (alto, basso, alto o viceversa)

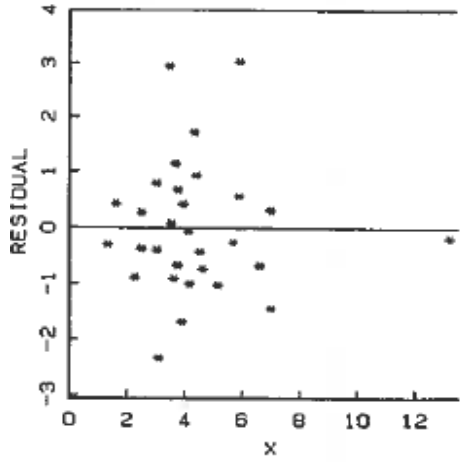






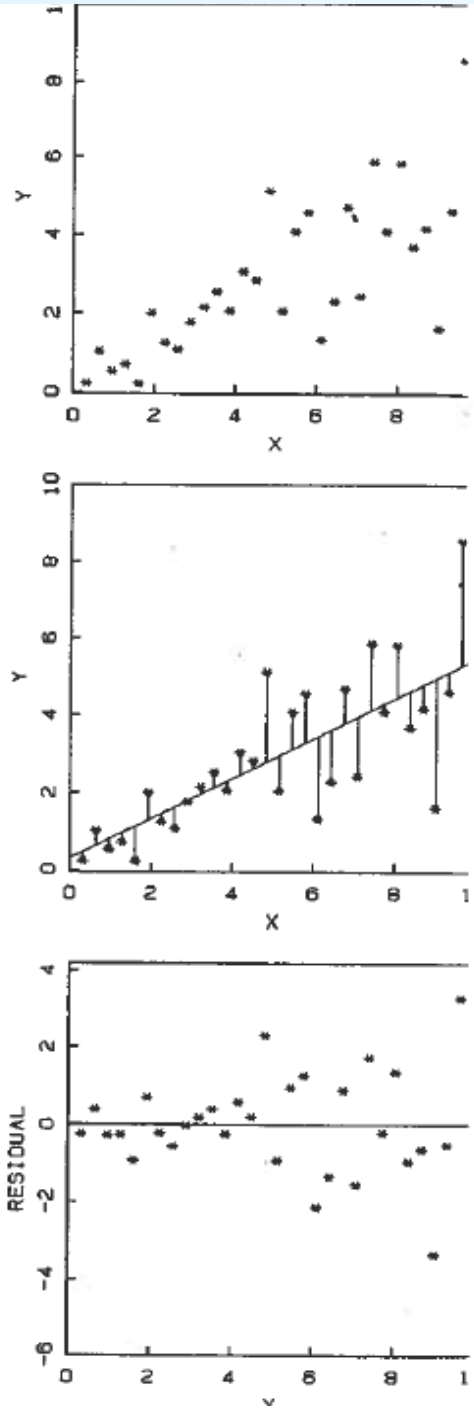
un punto presenta un valore x molto distante dagli altri ma consistente (in termini di y) rispetto alla configurazione lineare del resto dei punti (x = *leverage point*)





Fonte: Chambers et al., 1983

Variabilità dei punti non è costante al variare di x, ancora più evidente nei residui.
Possibile alternativa: trasformare le variabili



Fonte: Chambers et al., 1983

Relazioni “linearizzabili”

Relazioni non lineari ma “linearizzabili” mediante trasformazioni di variabili (a seguito dell’ispezione del diagramma di dispersione):

Trasformazioni più comuni per tener conto di un andamento curvilineo con y che cresce più velocemente di x

1. $Y = \alpha + \beta X^2$
2. $Log(Y) = \alpha + \beta X$
3. $Log(Y) = \alpha + \beta Log(X)$

(Metodo *minimi quadrati* può essere applicato a modelli generali $g(Y) = \alpha + \beta h(X)+(errore)$ con $g(Y)$ e $h(X)$ appropriate funzioni.

Importante è che *il modello sia lineare nei parametri*