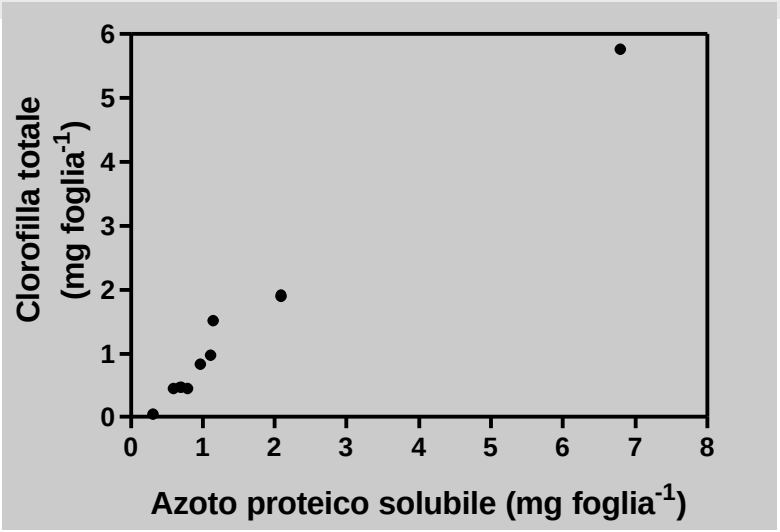
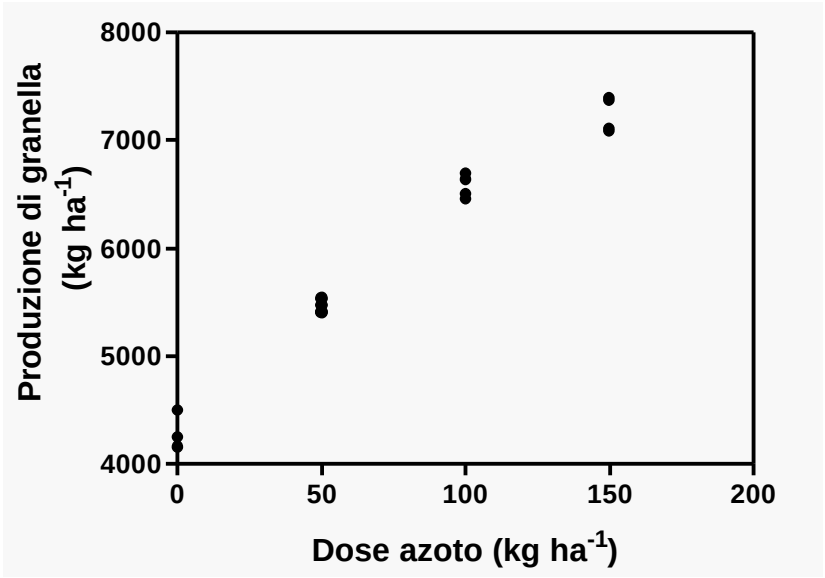
Dott. Raffaele Casa - Dipartimento di Produzione Vegetale
Modulo di Metodologia Sperimentale
Febbraio 2003

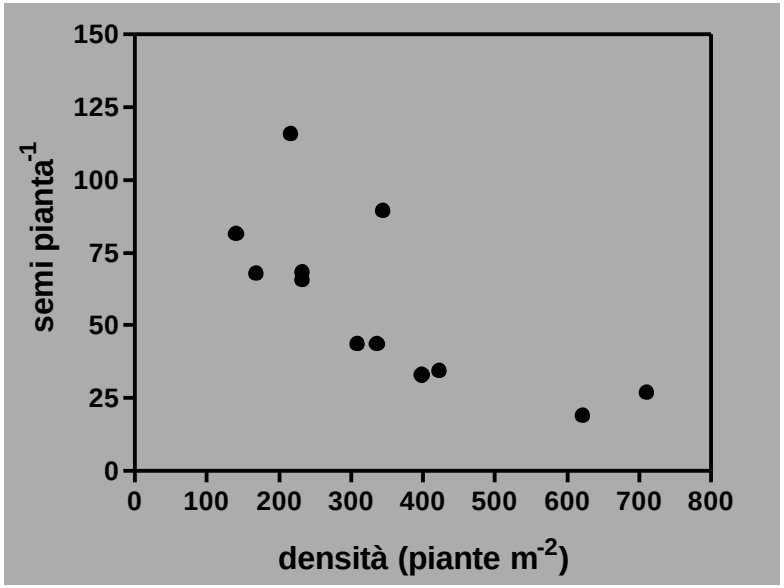
Relazioni tra variabili: *Correlazione e regressione lineare*

Analisi di relazioni tra variabili

- Correlazione
- Regressione



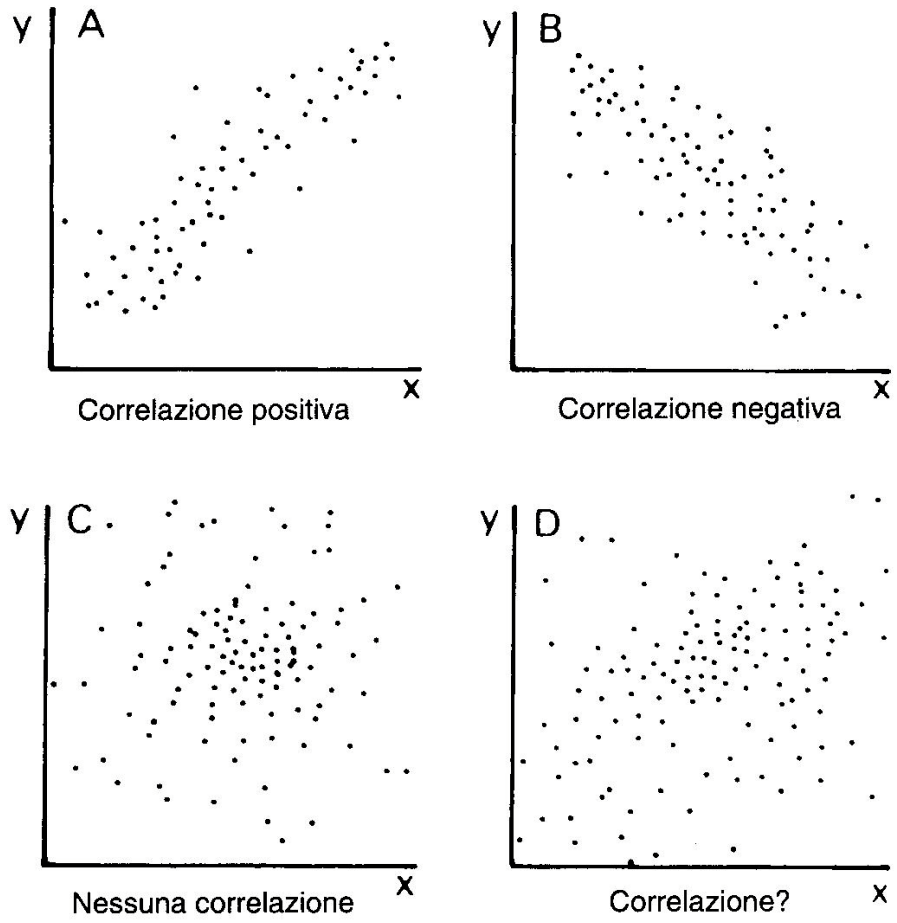




Analisi di relazioni tra variabili

- **Correlazione:** analizza se esiste una relazione tra due variabili (come e quanto due variabili variano insieme)
- **Regressione**: analizza la forma della relazione tra variabili

Covariazione di variabili



Analizzare la correlazione

- 2 coefficienti di correlazione:
- **Pearson** *product-moment* (parametrico)
- **Spearman** *rank* correlation (non parametrico)
- Entrambi vanno da -1 (correl.negativa) a +1 (correl.positiva). 0 corrisponde ad assenza di correlazione

Coefficiente di correlazione di Pearson: r

PARAMETRICO

Assunzioni:

- entrambe le variabili devono essere continue
- i dati devono essere secondo una scala a intervalli o razionale
- entrambe le variabili devono seguire una distribuzione normale
- la relazione tra le variabili è lineare

Tipo di dati

- **Scala nominale**: categorie non ordinabili (es. ambiente:macchia/pineta/faggeta; forma foglia:ellittica/lanceolata...)
- **Scala ordinale**: categorie ordinabili (es. alto/medio/basso; raro/comune/abbondante)
- **Scala per intervalli**: distanza quantificabile tra categorie, è possibile sottrarre ma non sommare (es. date, temperature)
- **Scala razionale**: possibile tutte le operazioni (+ - * ÷), variabili quantitative (es. lunghezza)

Coefficiente di correlazione di Pearson: r

- Procedura:
- Calcolo di r tra le variabili X e Y:

$$r = \frac{\sum_{i=1}^{N} X_i Y_i - \frac{\sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{N}}{\sqrt{\sum_{i=1}^{N} X_i^2 - \frac{\sum_{i=1}^{N} (X_i)^2}{N}} \sqrt{\sum_{i=1}^{N} Y_i^2 - \frac{\sum_{i=1}^{N} (Y_i)^2}{N}}}$$

Esempio: come calcolare il coefficiente di correlazione di Pearson

- Esempio: funzione “Pearson” o “Correlazione”

- Calcolo matrice di correlazione in Excel:

Strumenti - >Analisi dati -> Correlazione

Coefficiente di correlazione di Pearson: r

La correlazione è significativa?

- Il valore di **r** è stato calcolato da un campione e non dalla popolazione (r)
- Il valore calcolato indica una correlazione significativa?

Coefficiente di correlazione di Pearson: r

La correlazione è significativa?

- Ipotesi nulla: $\rho = 0$ (ρ è il coefficiente di correlazione della popolazione, **r** del campione).
- Calcolare t: $$t = r\sqrt{\frac{n-2}{1-r^2}}$$
- Valutare significatività di t per GDL = N-2

Coefficiente di correlazione di Pearson: r

OK: la correlazione è significativa ma….

- Le 2 variabili sono distribuite normalmente?
- La relazione tra le 2 variabili è lineare? (cf. trasformazione dei dati)
- Ricordarsi che anche se c'e' correlazione non vuol dire che c'e' nesso di causa-effetto …
- osservare la frazione di variabilità spiegata r^2 (*coefficiente di determinazione*)

Coefficiente di correlazione di Spearman: r_s

NON PARAMETRICO :

- i dati non devono avere distribuzione normale.
- Si possono usare dati da scala ordinale
- Si possono utilizzare anche campioni piccoli (da 7 a 30 coppie di dati)

Coefficiente di correlazione di Spearman: r_s

Procedura:

- Ordinare i dati dal più piccolo al più grande.
- Calcolare $\mathbf{r_s}$ come per $\mathbf{r}$ (Pearson) non sui dati ma sui ranghi (cioe' i numeri d'ordine)
- N.B. se più dati hanno lo stesso rango usare la media dei ranghi.
- Valutare la significatività di $\mathbf{r_s}$ calcolando il valore di t con la stessa formula usata per $\mathbf{r}$

Esempio: come calcolare il coefficiente di correlazione di Spearman

- Esempio:
- calcolo r Spearman in Excel

Interpretare i risultati della correlazione

Attenzione….

- Anche se c’e’ correlazione non vuol dire che ci sia nesso di causa-effetto …ed altre variabili possono essere la causa delle variazioni

Analisi di regressione

Lo scopo dell'analisi di regressione è di determinare la forma della relazione funzionale tra variabili (*relazione causa-effetto*)

Regressione semplice (lineare o non lineare): determinare la forma della relazione tra 2 variabili (una indipendente ed una dipendente)

•**Regressione multipla:** determinare la forma della relazione tra più variabili (più indipendenti ed una dipendente)

Analisi di regressione

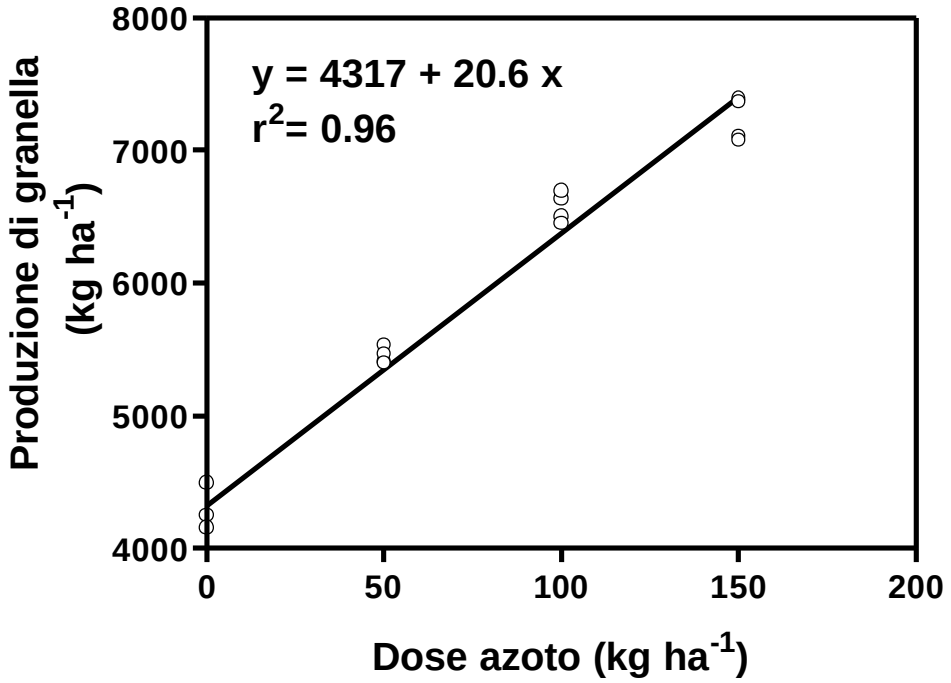
Perché è importante:

- Ci permette di costruire un modello funzionale della risposta di una variabile (effetto) ad un'altra (causa)
- Conoscendo la forma della relazione funzionale tra variabile indipendente e dipendente è possibile **stimare** il valore della variabile dipendente conoscendo quello della variabile indipendente (interpolazione) **solo nel range di dati X usato per la regressione (non è corretto estrapolare)**

Regressione lineare (semplice)

Nella regressione lineare la relazione tra variabili (*causa-effetto*) è rappresentata da una linea retta

N.B: se siamo indecisi su quale delle nostre variabili è dipendente e quale indipendente, allora l'analisi di regressione non è adatta!



Regressione lineare

La relazione tra variabili è espressa dall'equazione:

$Y = a+bX$

dove X è la variabile indipendente, Y la variabile dipendente, a è l'intercetta (il valore di Y quando $X=0$) e b è la pendenza (quanto aumenta Y per ogni aumento di un'unità di X).

N.B: La retta passa per il punto delle medie delle due variabili $\left(\overline{X}, \overline{Y}\right)$

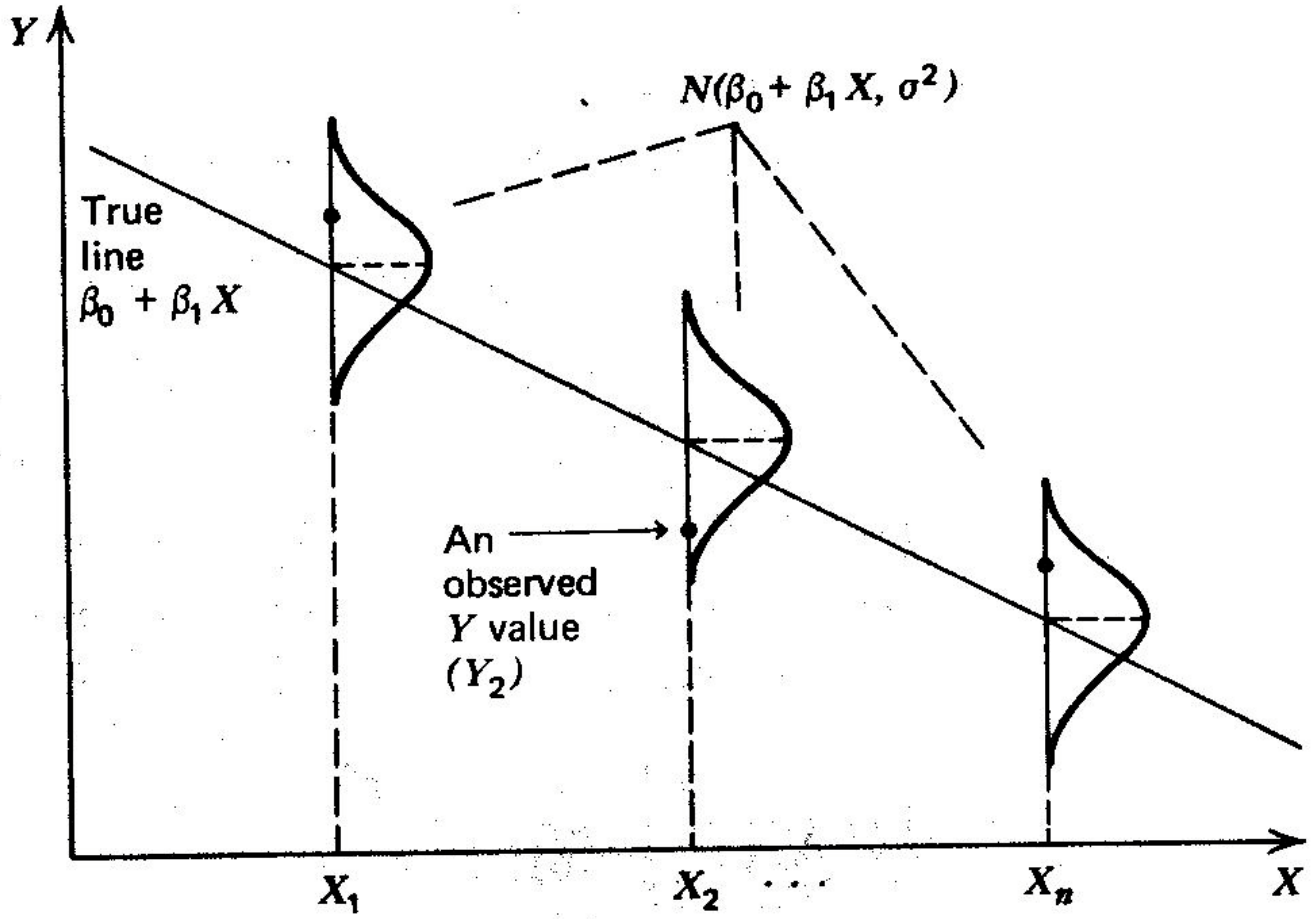
Regressione lineare

PARAMETRICO :

Assunzioni:

- Dati da scala per intervalli o scala razionale
- La variabile indipendente (X) è misurata senza errore (è fissata dallo sperimentatore)
- La variabile dipendente (Y) è campionata indipendentemente ad ogni valore di X
- Ad ogni valore di X i dati Y seguono la distribuzione normale ed hanno la stessa varianza

Regressione lineare



Regressione lineare

Procedura: metodo dei minimi quadrati (least squares)

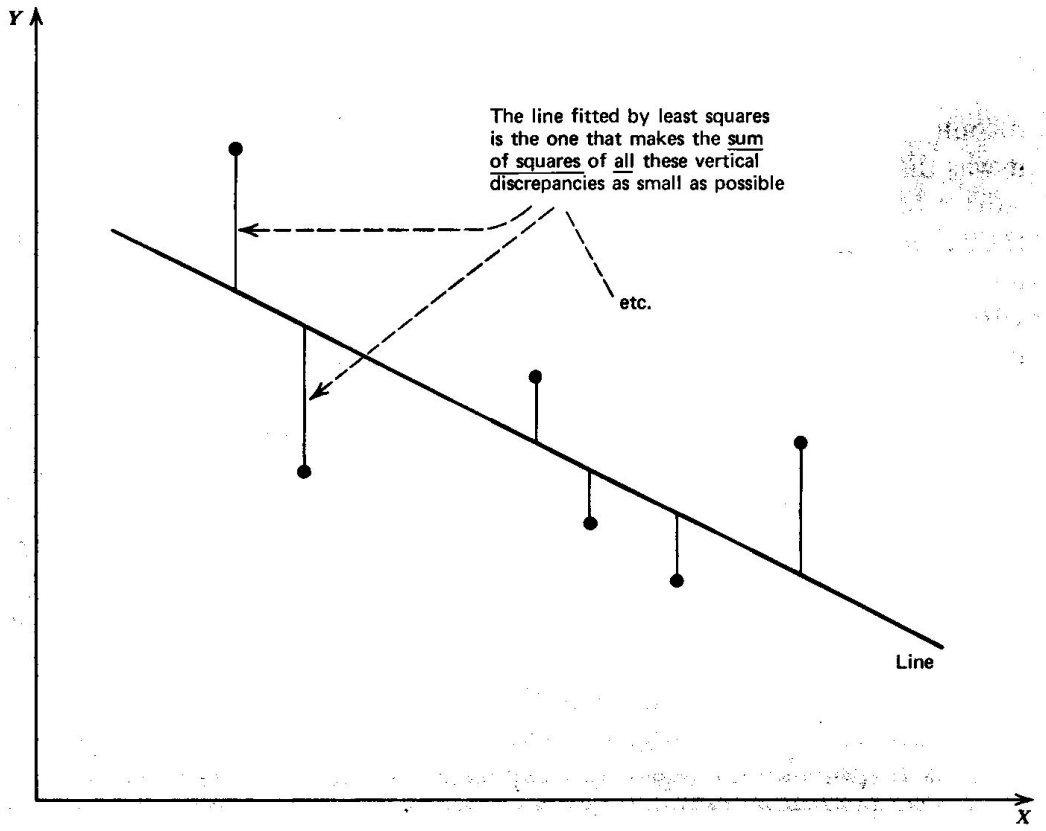


Figure 1.5 The vertical deviations whose sum of squares is minimized for the least squares procedure.

Regressione lineare

Procedura:

1. Stima della pendenza b

$$b = \frac{\sum_{i=1}^{N} X_i Y_i - \frac{\sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{N}}{\sum_{i=1}^{N} X_i^2 - \frac{\sum_{i=1}^{N} (X_i)^2}{N}}$$

2. Stima dell'intercetta a

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

Regressione lineare

Variazione (devianza) spiegata / non spiegata dalla regressione nei dati Y

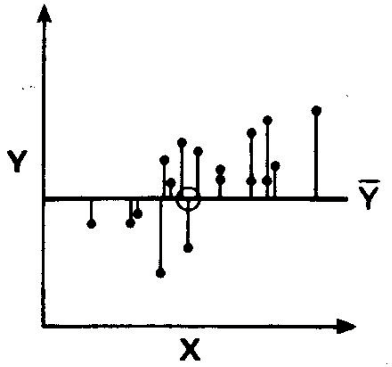
$$\sum_{i=1}^{N}(Y_i-\overline{Y})^2=\sum_{i=1}^{N}(\hat{Y}_i-\overline{Y})^2+\sum_{i=1}^{N}(\hat{Y}_i-Y_i)^2$$

La variazione totale nei dati Y

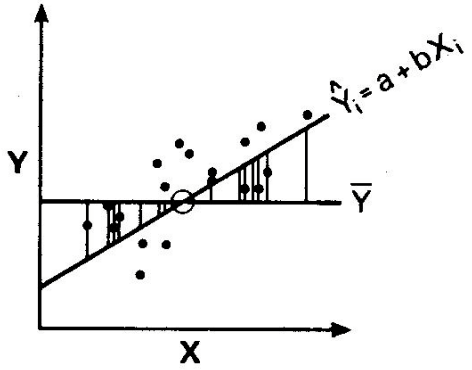
in parte è spiegata dalla regressione

ed in parte non è spiegata dalla regressione (variazione residua)

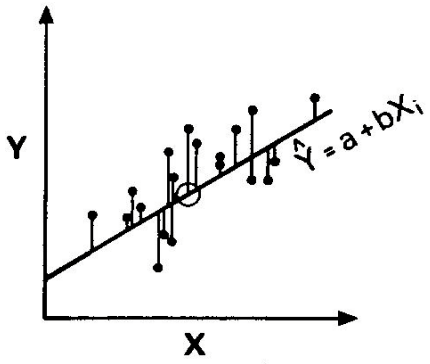
Regressione lineare



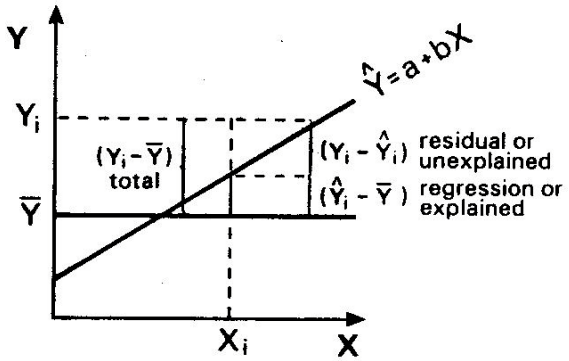
a) Total sum of squares



b) Regression sum of squares

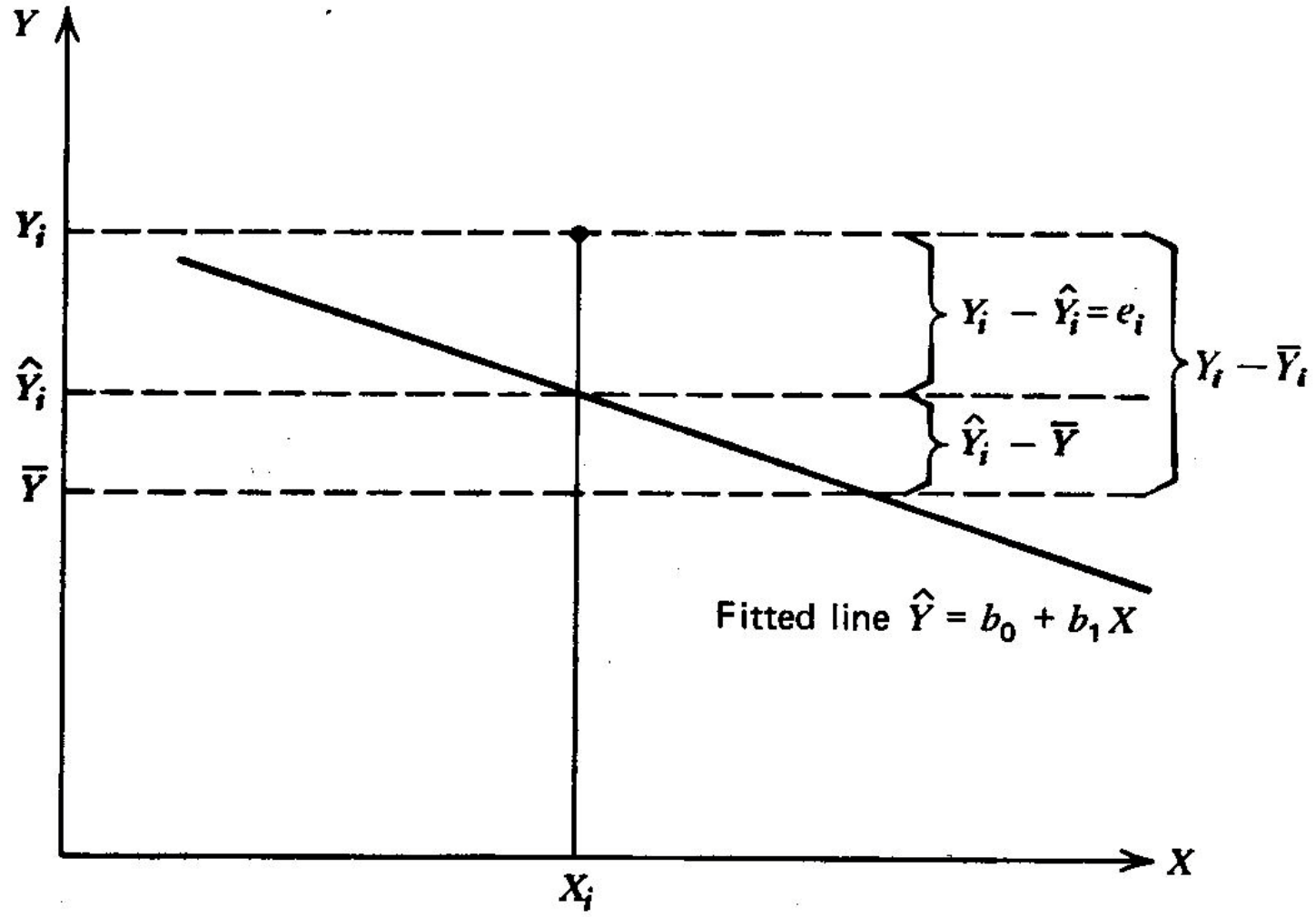


c) Residual sum of squares



d) Splitting the total sum of squares into regression (explained) and residual (unexplained) components

Regressione lineare



Regressione lineare

Come quantificare la bontà della regressione?

Il *coefficiente di determinazione (va da 0 a 1)*

$$r^2 = \frac{devianza\_spiegata}{devianza\_tot} = \frac{\sum_{i=1}^{N} (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}$$

Regressione lineare

La regressione è significativa?

- L’equazione è stata ricavata da un campione e non dalla popolazione

1. Test t sull’err.standard della pendenza b:

Ipotesi nulla=la pendenza è uguale a 0

2. Analisi della varianza: si esamina il rapporto tra varianza spigata dalla regressione e varianza residua.

Regressione lineare

La regressione è significativa?

1. Test t sull'errore standard della pendenza **b** (con n-2 GDL):

$$t = \frac{b - H_o}{Err.St_b}$$

H_o =ipotesi nulla;

Regressione lineare

Errore standard della pendenza **b** :

$$Err.St_b = \sqrt{\frac{\left(\sum_{i=1}^{N} (Y_i - \overline{Y})^2 - \frac{\sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{N} (X_i - \overline{X})^2} \right)}{(n-2)\sum_{i=1}^{N} (X_i - \overline{X})^2}}$$

Regressione lineare

2. Analisi della varianza: test F del rapporto tra varianza spiegata dalla regressione e varianza residua.

Fonti di variazione	Devianze	Descrizione	Gradi di libertà
Spiegata dalla regressione	$\sum_{i=1}^{N} (\hat{Y}_i - \overline{Y})^2$	Somma dei quadrati delle deviazioni dei valori stimati di Y rispetto alla media di Y	**k**
Non spiegata dalla regressione (residua)	$\sum_{i=1}^{N} (\hat{Y}_i - Y_i)^2$	Somma dei quadrati delle differenze tra i valori stimati ed osservati di Y	**n-k-1**
Totale	$\sum_{i=1}^{N} (Y_i - \overline{Y})^2$	Somma dei quadrati delle devaiazioni tra i valori osservati di Y e la media di Y	**n-1**

dove:
n = numero di osservazioni
k= sempre 1 per la regressione lineare

Regressione lineare

- **Errore standard e limiti di confidenza**
- L'errore standard dei valori stimati di Y è uguale alla deviazione standard dei residui:

$$S_{XY} = \sqrt{\frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{n}}$$

Per piccoli campioni si usa:

$$S_{XY} = \sqrt{\frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{n-2}}$$

- **Analisi dei residui**
- Standardizzazione (divisione per S_{XY})
- Distribuzione casuale sopra e sotto la linea (+/-)?

Regressione lineare

Esempio: dati *granella-azoto*

- calcolo regressione lineare in Excel

Regressione lineare

OK la regressione è significativa ma… **assunzioni**!

•La variabile dipendente (Y) è campionata indipendentemente ad ogni valore di X ? **Cf. es. analisi di crescita di individui**

•Ad ogni valore di X, i dati Y hanno la stessa varianza?

Cf. aumento di varianza tra individui di maggiori dimensioni

•Ad ogni valore di X, i dati Y seguono la distribuzione normale?

•La variabile indipendente (X) è misurata senza errore (è fissata dallo sperimentatore)?