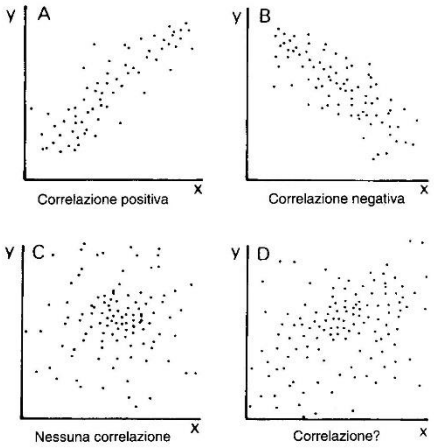
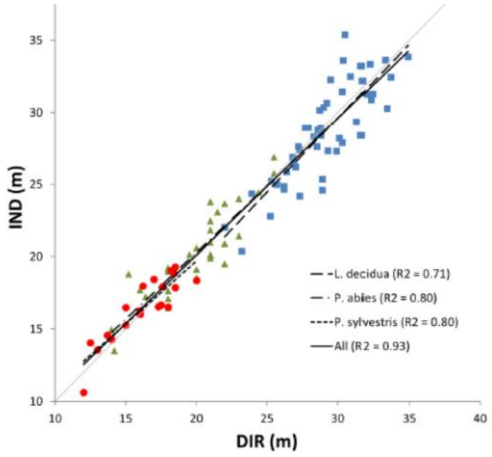
Corso di Laurea di Scienze Forestali ed Ambientali

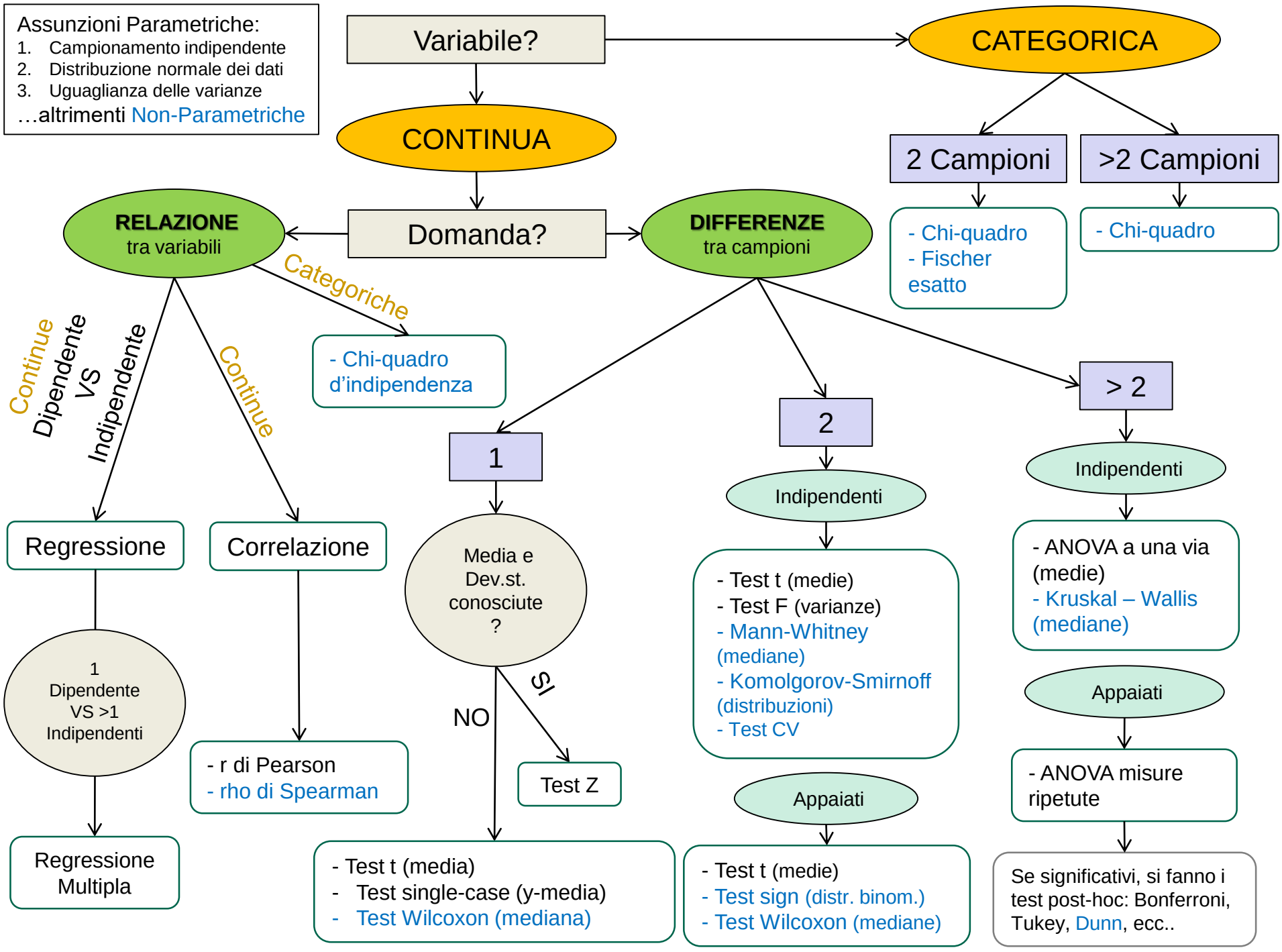
Ecologia e Statistica per l'ambiente

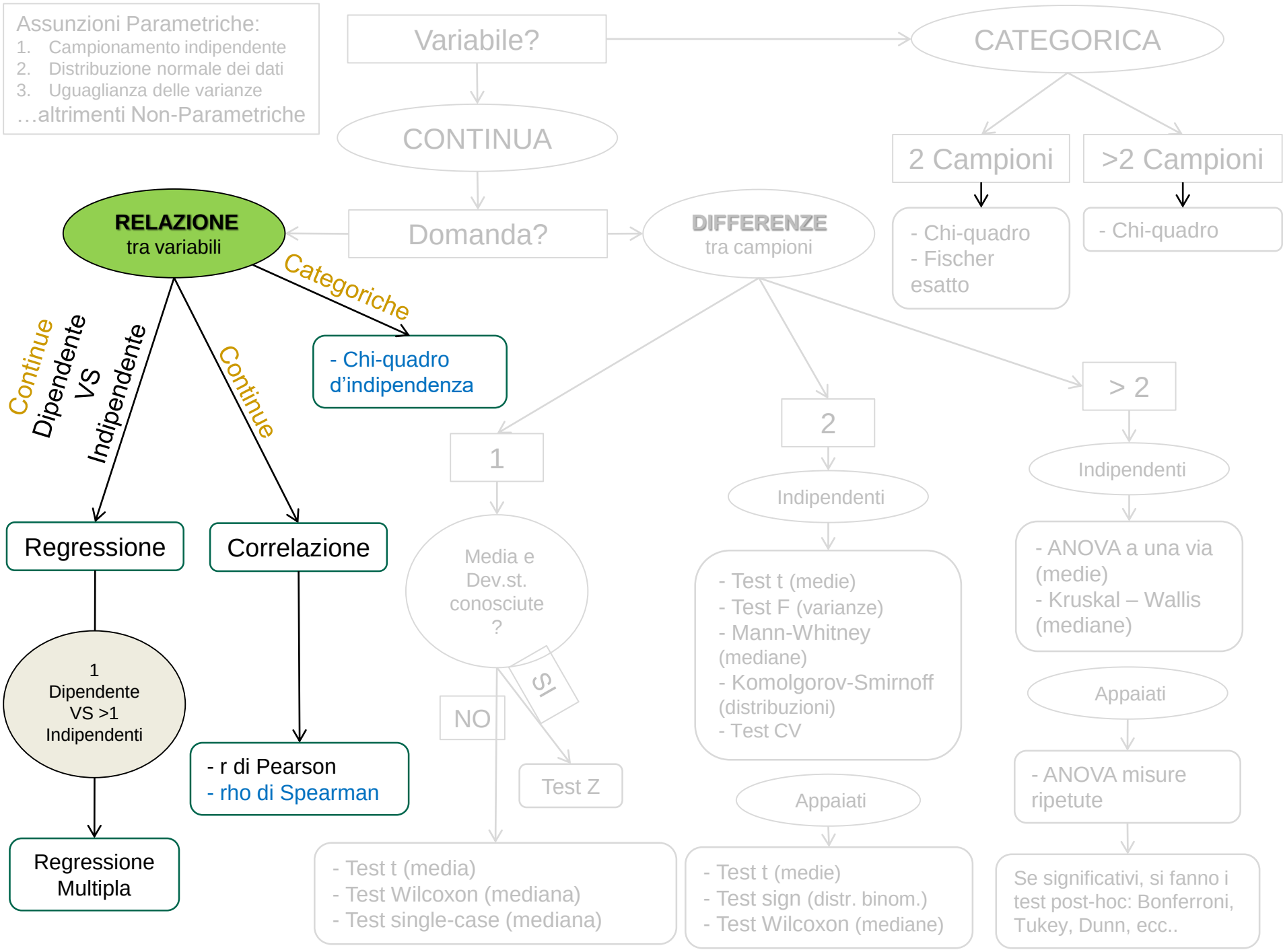
a.a. 2020-2021 - MATTEO GARBARINO - matteo.garbarino@unito.it

RELAZIONI TRA VARIABILI
Correlazione e Regressione



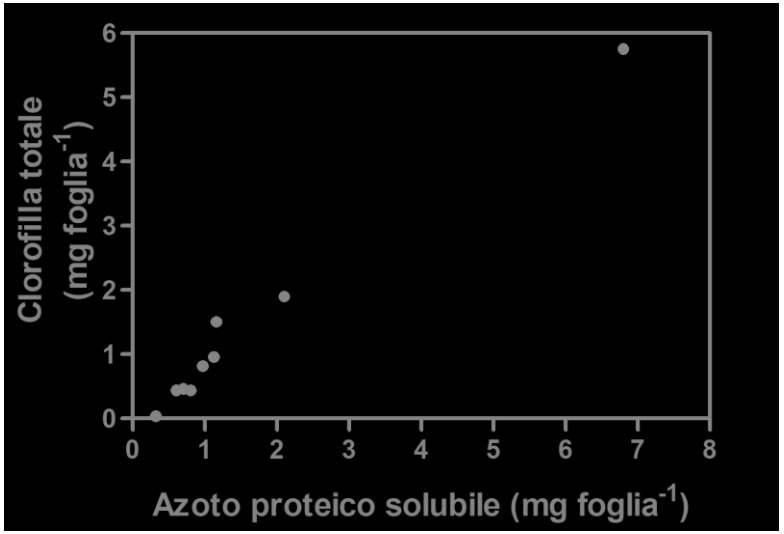


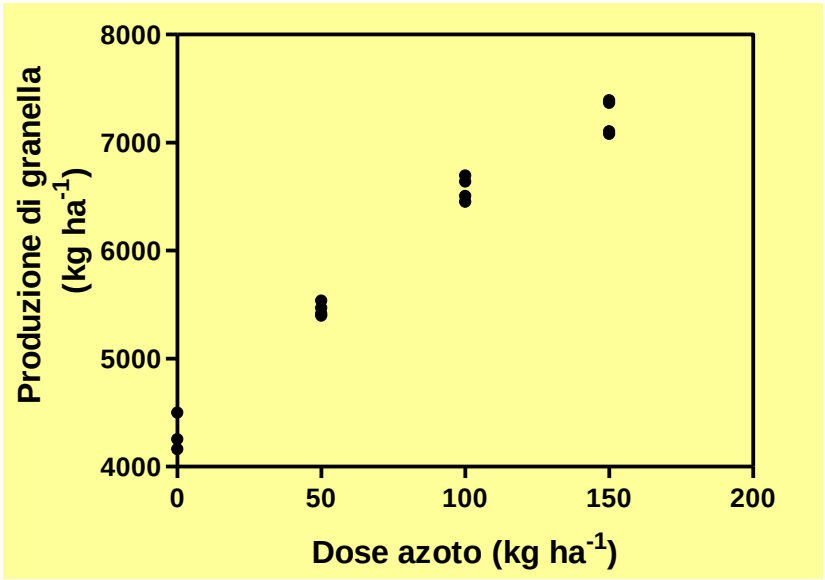

Assunzioni Parametriche:
1. Campionamento indipendente
2. Distribuzione normale dei dati
3. Uguaglianza delle varianze
…altrimenti Non-Parametriche
Variabile?
CATEGORICA
CONTINUA
2 Campioni
>2 Campioni
- Chi-quadro
- Fischer esatto
- Chi-quadro
Domanda?
RELAZIONE tra variabili
DIFFERENZE tra campioni
Categoriche
- Chi-quadro d'indipendenza
Continue Dipendente VS Indipendente
Continue
Regressione
Correlazione
1 Dipendente VS >1 Indipendenti
Regressione Multipla
- r di Pearson
- rho di Spearman
1
2
> 2
Media e Dev.st. conosciute ?
SI
NO
Test Z
- Test t (media)
- Test single-case (y-media)
- Test Wilcoxon (mediana)
Indipendenti
- Test t (medie)
- Test F (varianze)
- Mann-Whitney (mediane)
- Komolgorov-Smirnoff (distribuzioni)
- Test CV
Appaiati
- Test t (medie)
- Test sign (distr. binom.)
- Test Wilcoxon (mediane)
Indipendenti
- ANOVA a una via (medie)
- Kruskal – Wallis (mediane)
Appaiati
- ANOVA misure ripetute
Se significativi, si fanno i test post-hoc: Bonferroni, Tukey, Dunn, ecc..

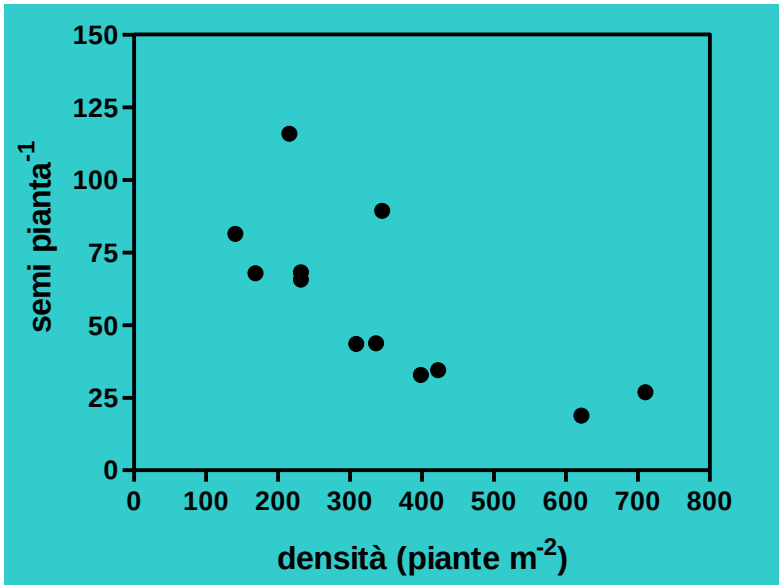

Assunzioni Parametriche:
1. Campionamento indipendente
2. Distribuzione normale dei dati
3. Uguaglianza delle varianze
…altrimenti Non-Parametriche
Variabile?
CATEGORICA
CONTINUA
2 Campioni
>2 Campioni
- Chi-quadro
- Fischer esatto
- Chi-quadro
RELAZIONE tra variabili
Domanda?
DIFFERENZE tra campioni
Categoriche
Continue Dipendente VS Indipendente
Continue
- Chi-quadro d'indipendenza
1
2
> 2
Indipendenti
Indipendenti
Regressione
Correlazione
Media e Dev.st. conosciute ?
- Test t (medie)
- Test F (varianze)
- Mann-Whitney (mediane)
- Komolgorov-Smirnoff (distribuzioni)
- Test CV
- ANOVA a una via (medie)
- Kruskal – Wallis (mediane)
1 Dipendente VS >1 Indipendenti
NO
SI
Test Z
Appaiati
Appaiati
- r di Pearson
- rho di Spearman
- Test t (media)
- Test Wilcoxon (mediana)
- Test single-case (mediana)
- Test t (medie)
- Test sign (distr. binom.)
- Test Wilcoxon (mediane)
- ANOVA misure ripetute
Regressione Multipla
Se significativi, si fanno i test post-hoc: Bonferroni, Tukey, Dunn, ecc..

Analisi di relazioni tra variabili

- Correlazione
- Regressione





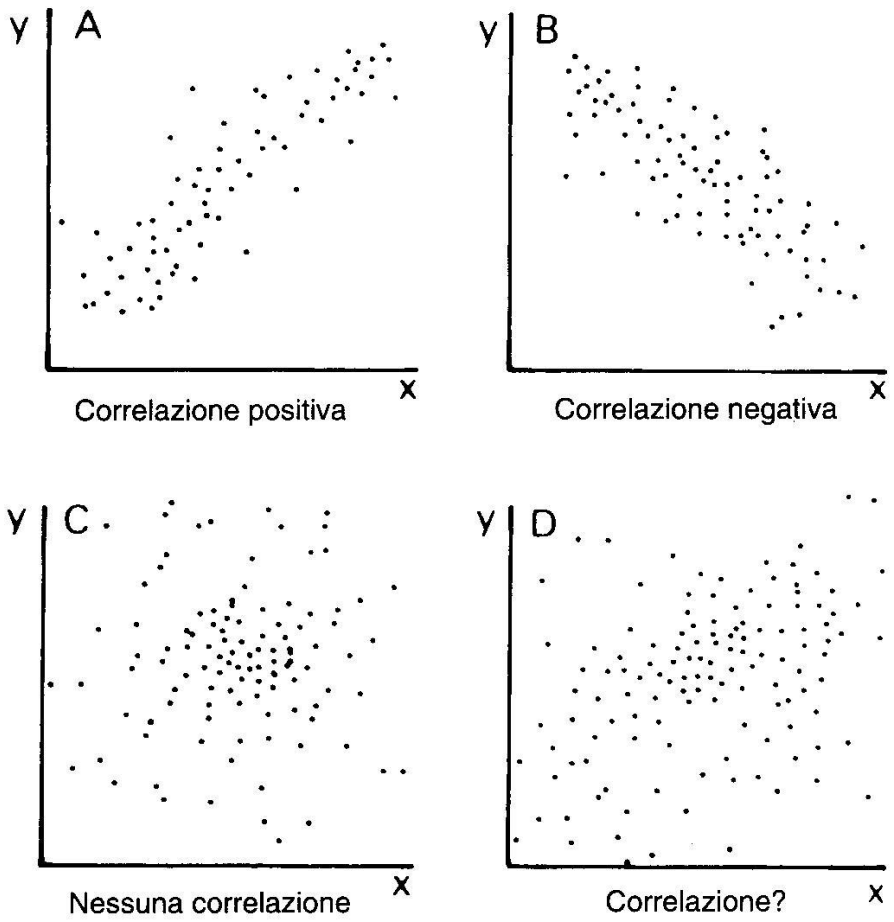


Analisi di relazioni tra variabili

- **Correlazione:** analizza se esiste una relazione tra due variabili (come [+/-] e quanto [-1←→1] due variabili variano insieme)

- **Regressione**: analizza la forma della relazione tra variabili

Correlazione di variabili



Analizzare la correlazione

2 coefficienti di correlazione:

- **Pearson** (r) *coefficiente di correlazione lineare* (parametrico)
- **Spearman** (rho – ρ - r_s) *coefficiente di correlazione per ranghi* (non parametrico)

Entrambi vanno da -1 (correl. negativa) a +1 (correl. positiva). 0 corrisponde ad assenza di correlazione

Coefficiente di correlazione di Pearson: r

CORRELAZIONE PARAMETRICA

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**

Tipologie di Variabili

- **Scala nominale**: categorie non ordinabili (es. categoria forestale: lariceto/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 (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:

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

Pearson's r: significatività

- 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
- P (t, n-2) *in excel =distrib.t(t;gdl)*

Pearson's r: assunzioni

OK: la correlazione è significativa ma….

- Le 2 variabili sono distribuite normalmente?
- La relazione tra le 2 variabili è lineare? (cf. trasformazione dei dati)
- Anche se c'è correlazione non vuol dire che ci sia nesso di causa-effetto; **se SI → regressione**
- Osservare la frazione di variabilità spiegata (r^2, *coefficiente di determinazione*)

Coefficiente di correlazione di Spearman: r_s

CORRELAZIONE NON PARAMETRICA:

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

Spearman's r_s

Procedura:

- Ordinare i dati dal più piccolo al più grande
- Calcolare r_s non sui dati ma sui ranghi (d=differenza tra ranghi)

$$\mathbf{r_s = 1 - 6*(d_1^2 + d_2^2 + \ldots + d_n^2)/(n(n2-1))}$$

- Valutare la significatività di $\mathbf{r_s}$ ricorrendo ad apposite tavole (i softwares lo fanno in automatico)

Interpretare i risultati della correlazione

Attenzione….

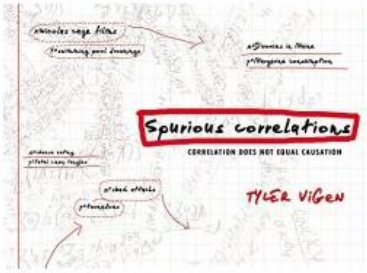
Anche se c'è correlazione non vuol dire che ci sia nesso di causa-effetto e altre variabili possono essere la causa delle variazioni.

Attenzione alle Correlazioni Spurie

tylervigen.com

about | twitter | email | subscribe

Spurious correlations



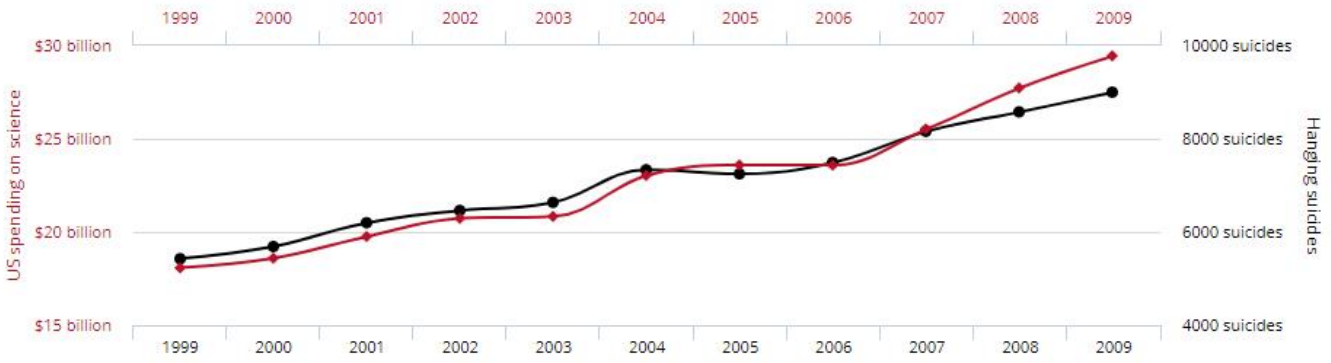
Now a ridiculous book!

- Spurious charts
- Fascinating factoids
- Commentary in the footnotes

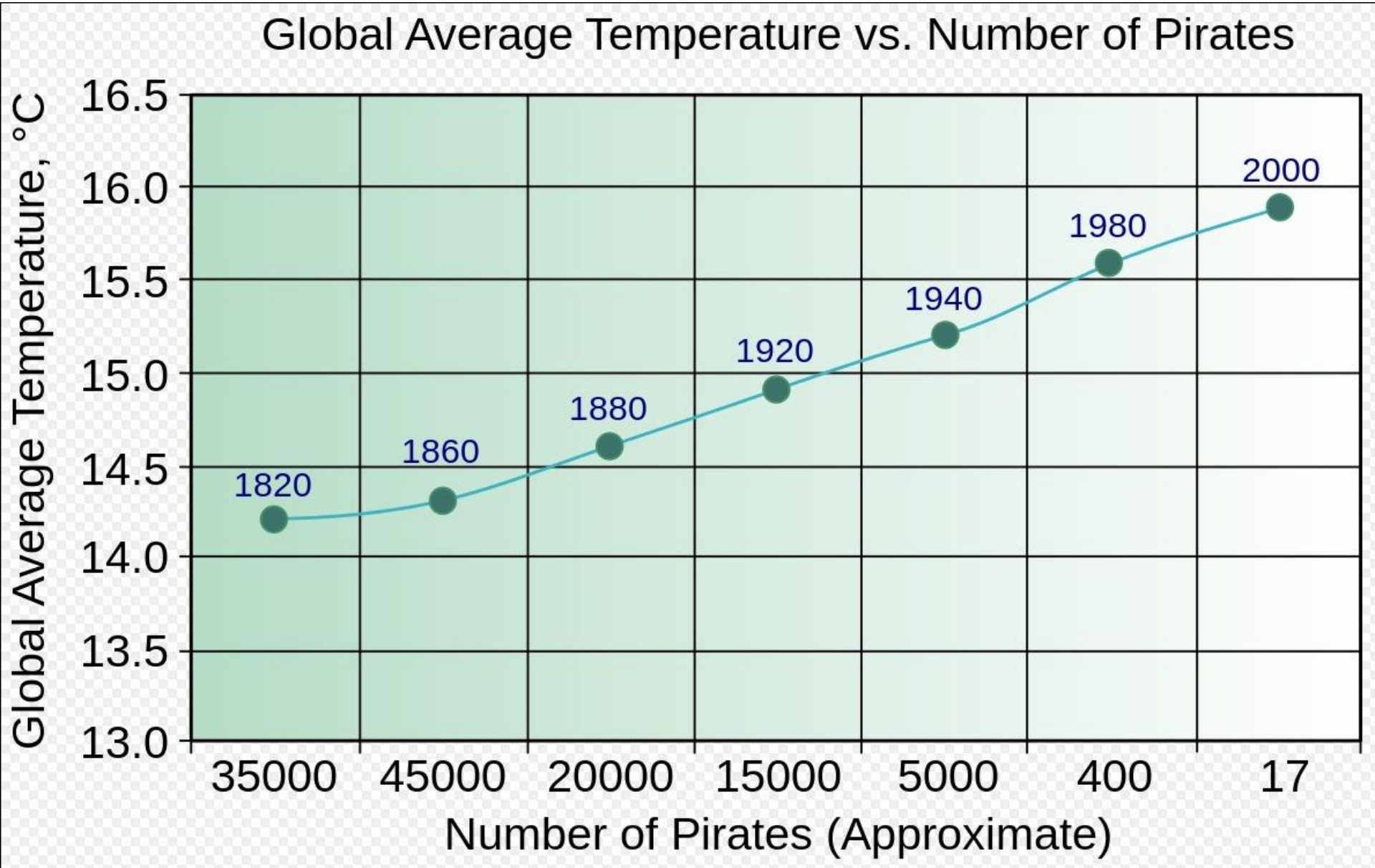
Amazon | Barnes & Noble | Indie Bound

US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation

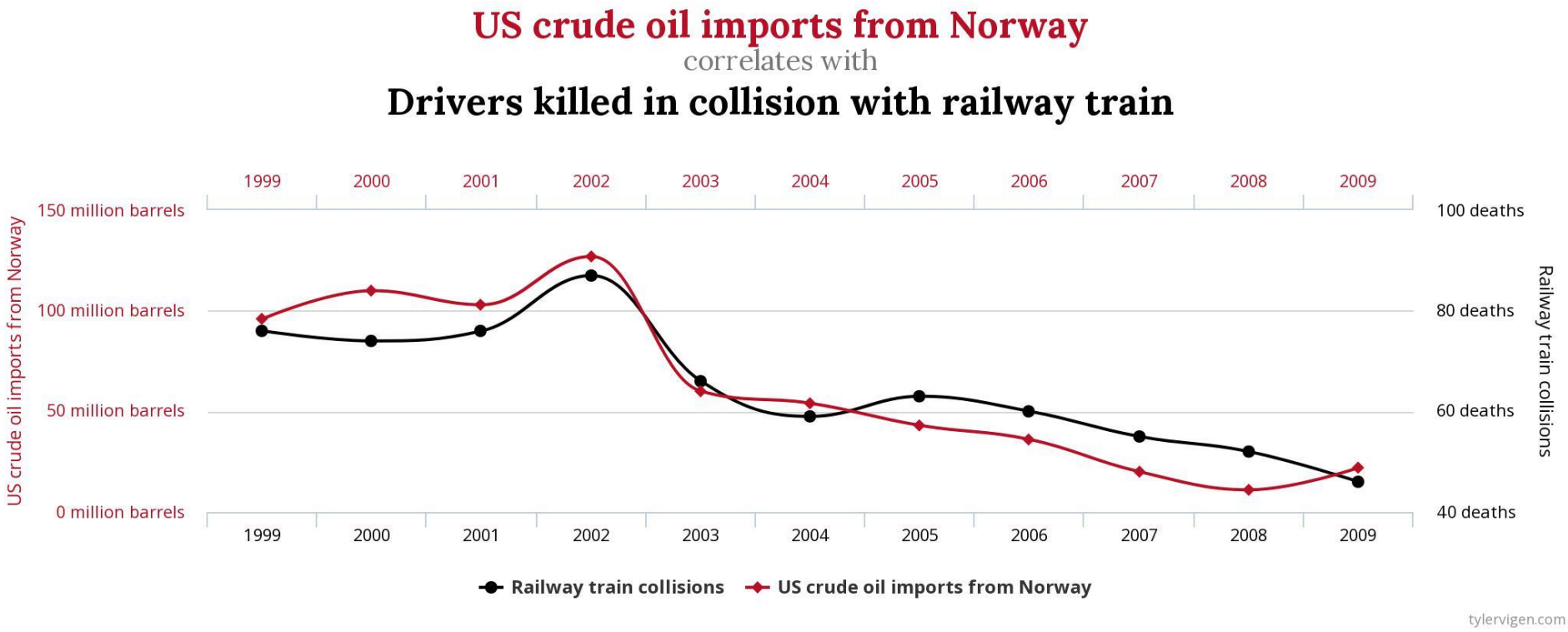
Correlation: 99.79% (r=0.99789126)



Interpretare i risultati della correlazione



Interpretare i risultati della correlazione

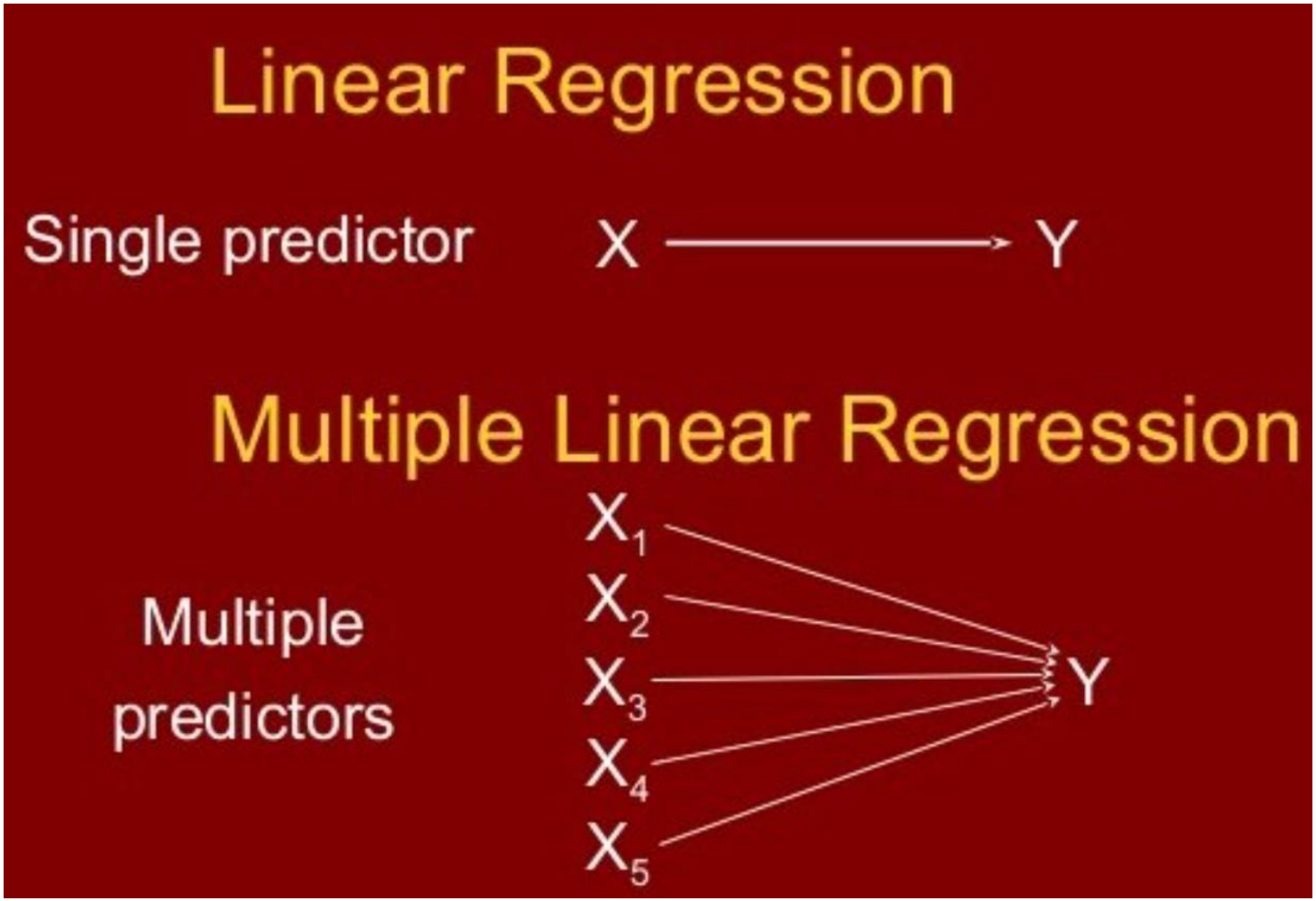


Analisi di regressione

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

- **Regressione semplice:** 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



Analisi di regressione

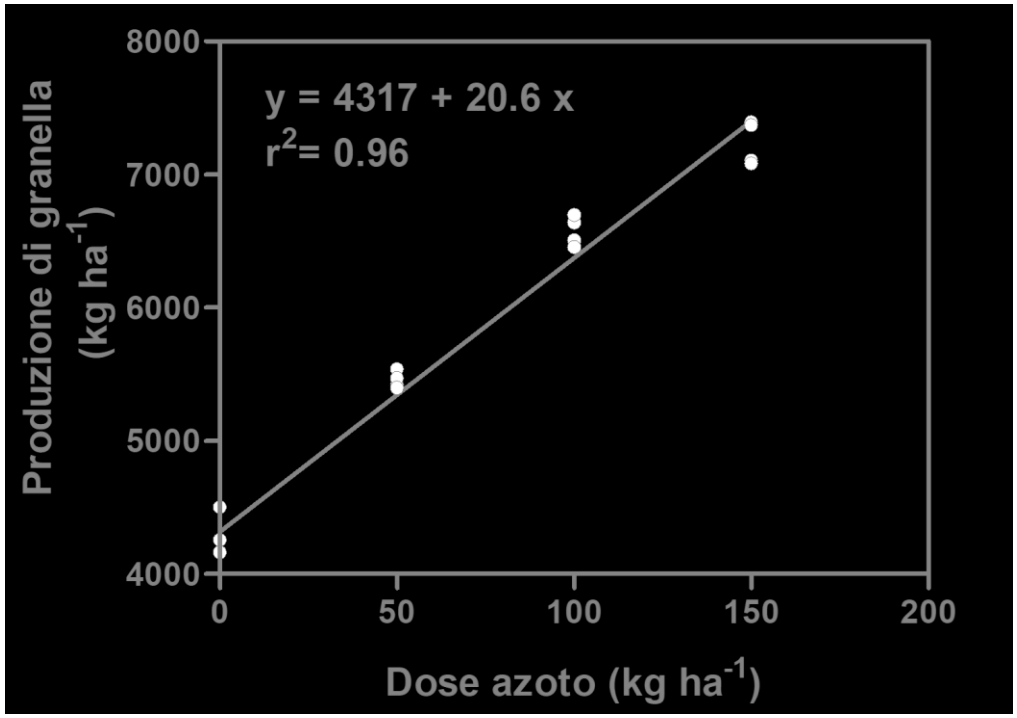
Perché è importante:

- Permette di costruire un modello funzionale della risposta di una variabile (effetto) rispetto 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) **nell'intervallo dei valori di X usato per la regressione**

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 (di quanto varia la Y per ogni variazione di una unità di X).

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

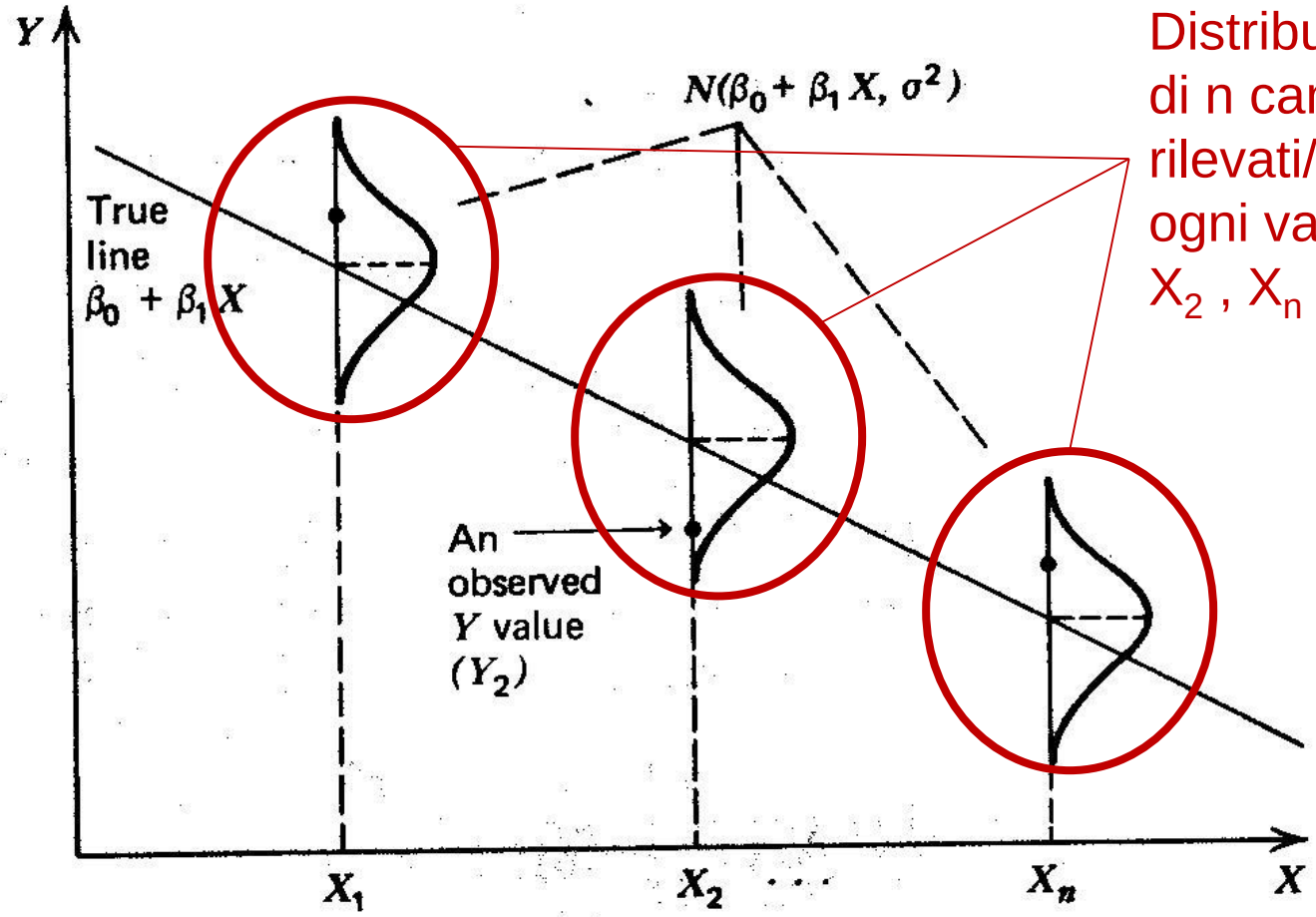
Regressione lineare

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 a ogni valore di X
- Ad ogni valore di X i dati Y seguono la distribuzione normale e hanno la stessa varianza

Regressione lineare



Distribuzione normale di n campioni di Y_n rilevati/osservati per ogni valore di X (X_1 , X_2 , X_n)

Regressione lineare

Procedura:
metodo dei minimi quadrati (*least squares*).

La linea è posizionata in modo tale da rendere la somma dei quadrati dei residui (linee verticali) più piccola possibile.

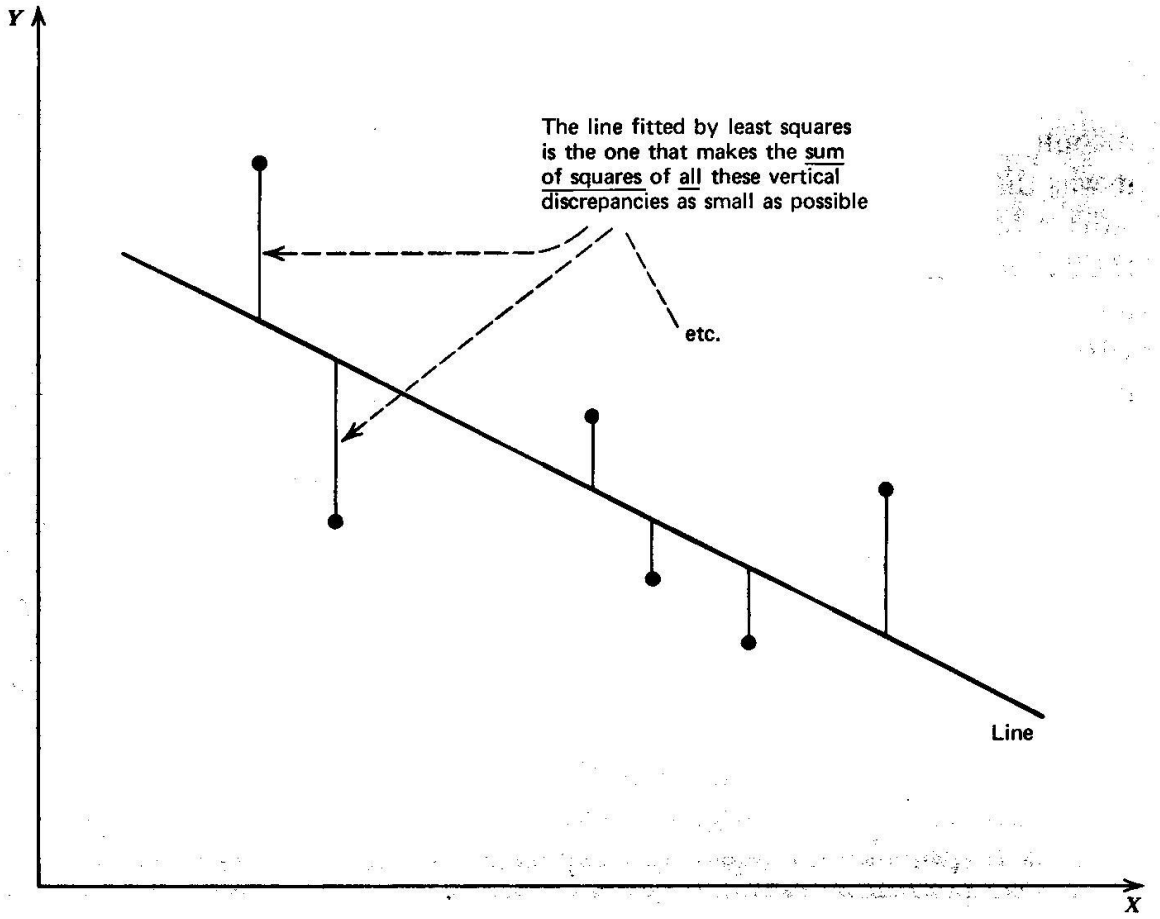


Figure 1.5 The vertical deviations whose sum of squares is minimized for the least squares procedure.

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

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

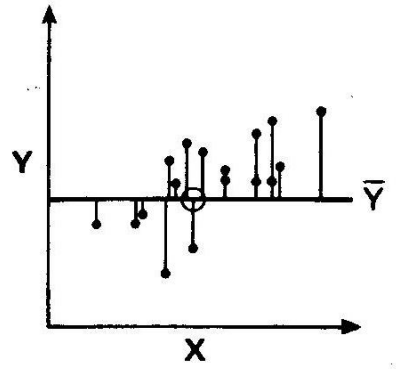
La variazione totale nei dati Y

in parte è spiegata dalla regressione

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

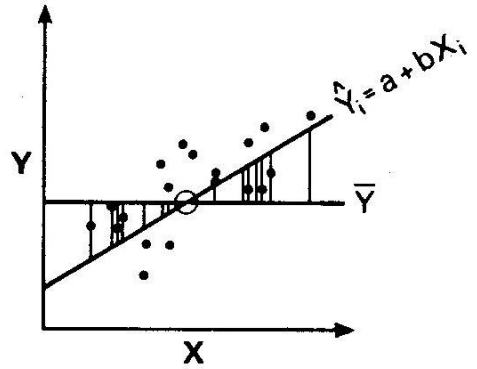
Y = valore osservato
$\hat{Y}$ = valore stimato (reg.)
$\overline{Y}$ = valore medio

Regressione lineare



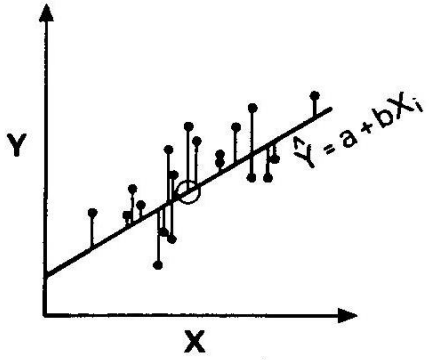
Devianza totale

a) Total sum of squares



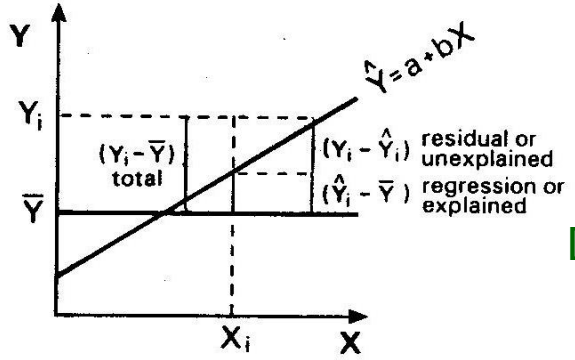
b) Regression sum of squares

Devianza spiegata



Devianza residua

c) Residual sum of squares

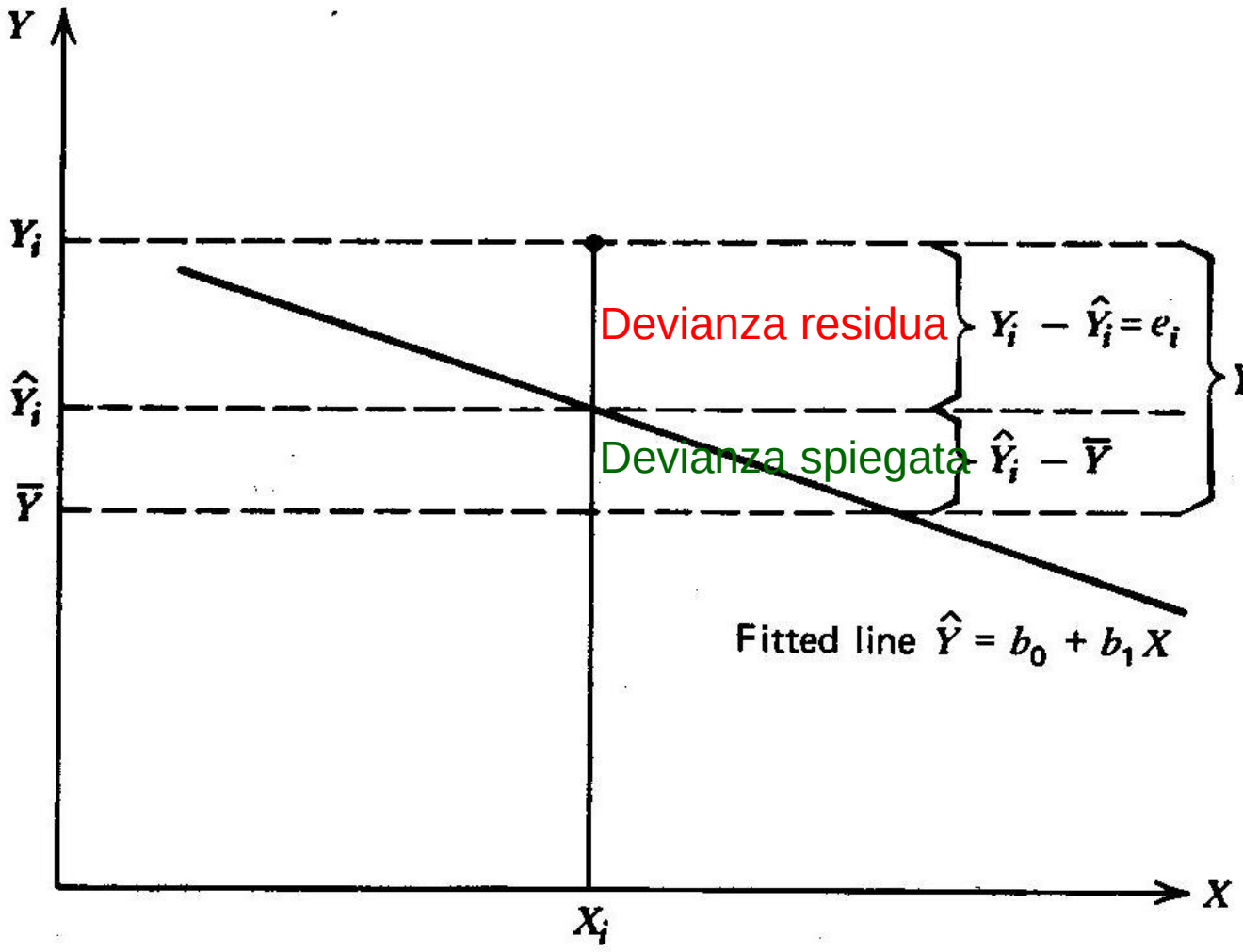


Devianza residua

Devianza spiegata

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'errore standard della pendenza b:

Ipotesi nulla = la pendenza è uguale a 0

2. Analisi della varianza: si esamina il rapporto tra varianza spiegata 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)	$\sum_{i=1}^{N} (\hat{Y}_i - Y)^2$	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

OK la regressione è significativa ma… assunzioni!

- La variabile dipendente (Y) è campionata indipendentemente ad ogni valore di X? **Es. analisi di crescita di individui**
- Ad ogni valore di X, i dati Y hanno la stessa varianza?

Es. varianza maggiore per 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)?

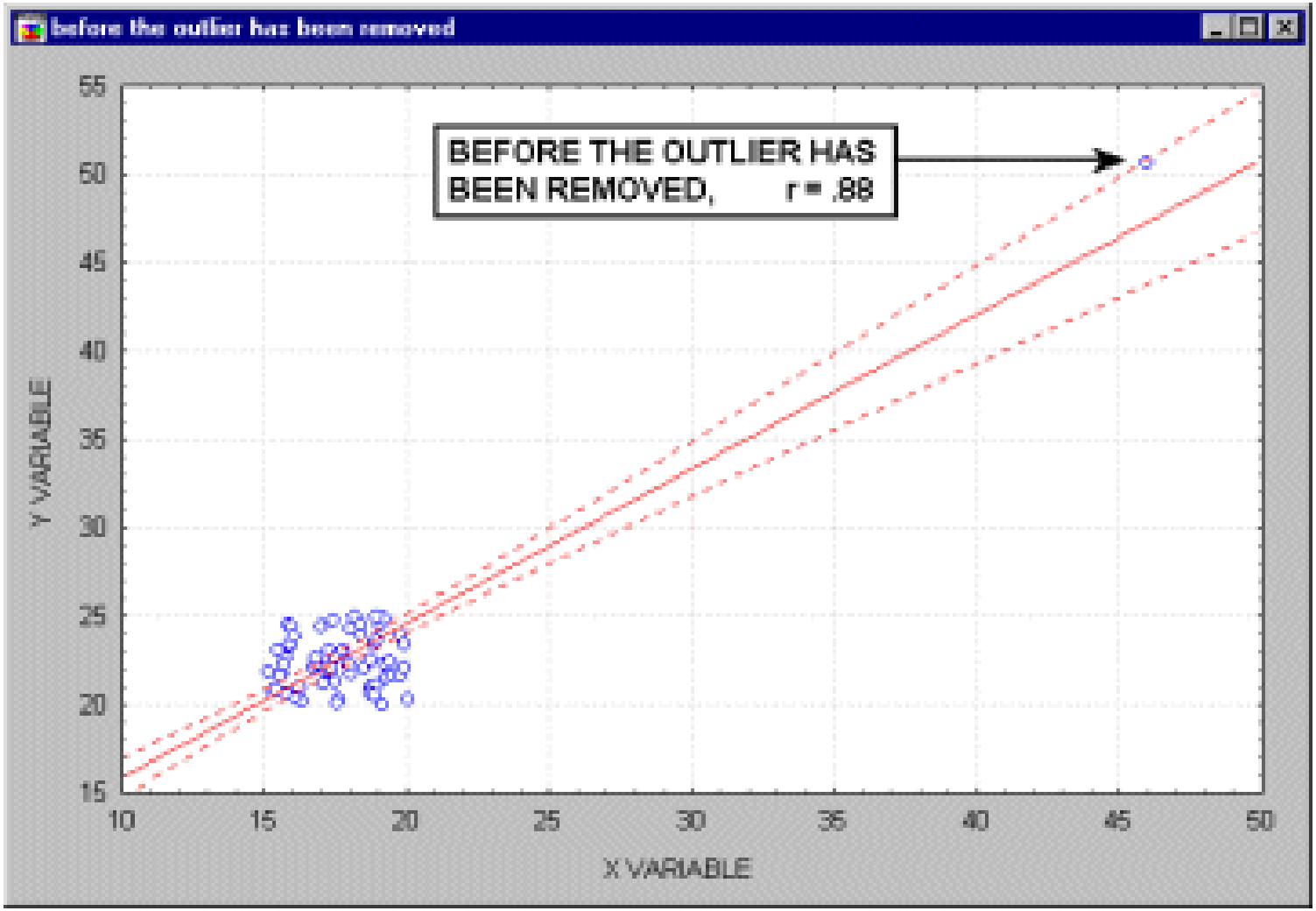
Dati anomali

- Violazioni significative dalle assunzioni possono essere rilevate esaminando i residui (differenze tra valori stimati e misurati della variabile di risposta)
- Valori anomali (*outlier*) possono "attrarre" la retta di regressione in una direzione particolare

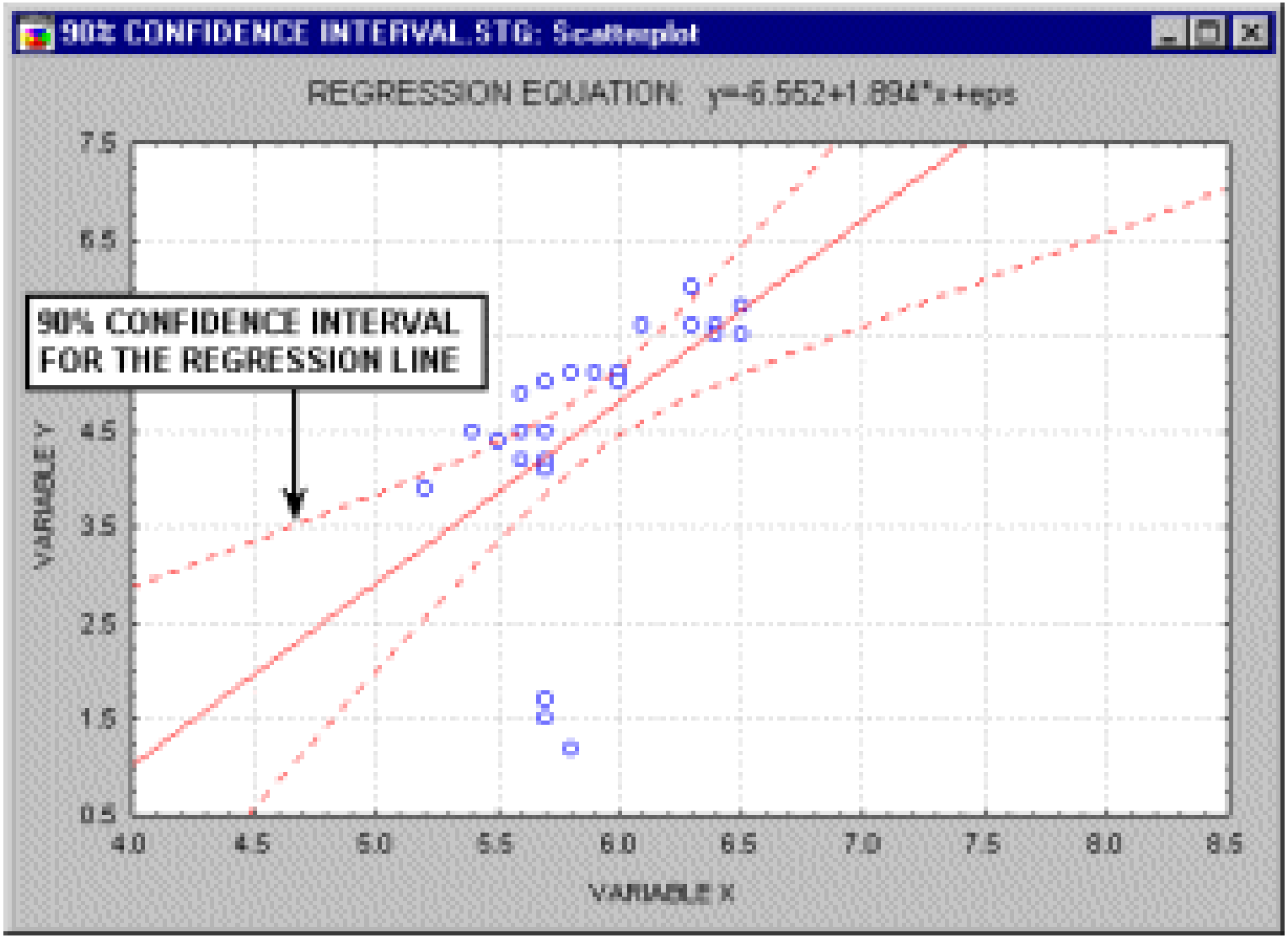
Dati anomali

- Anche se esistono strumenti statistici per evidenziare dati che possono essere esterni al campo di variabilità della variabile dipendente o indipendente, definire questi dati anomali è un problema del ricercatore
- Si deve cercare di risalire alle cause che possono aver determinato l'anomalia della misurazione giustificando quindi l'eliminazione del dato stesso

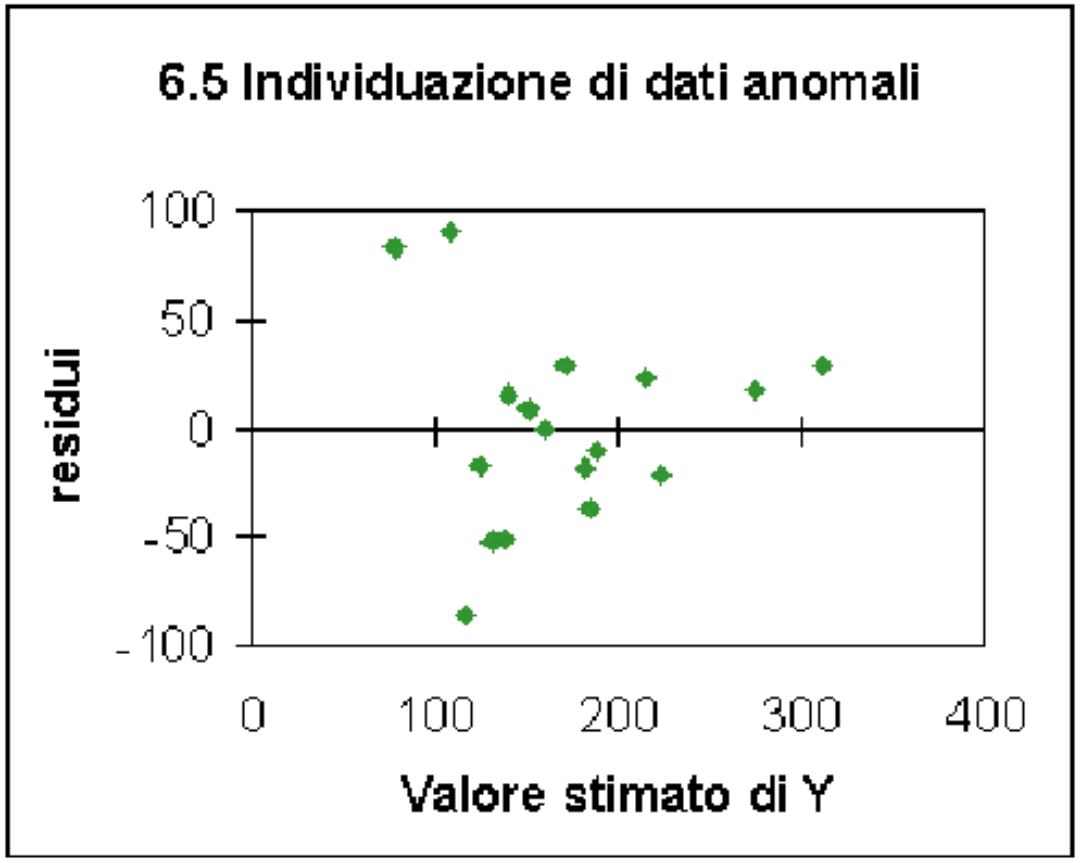
Un caso estremo



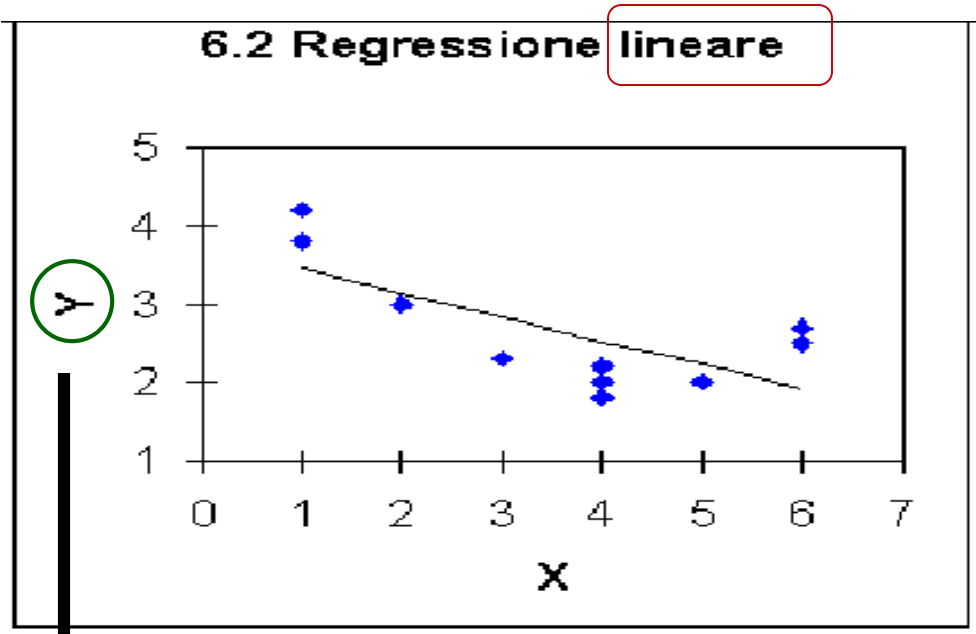
Intervalli di confidenza

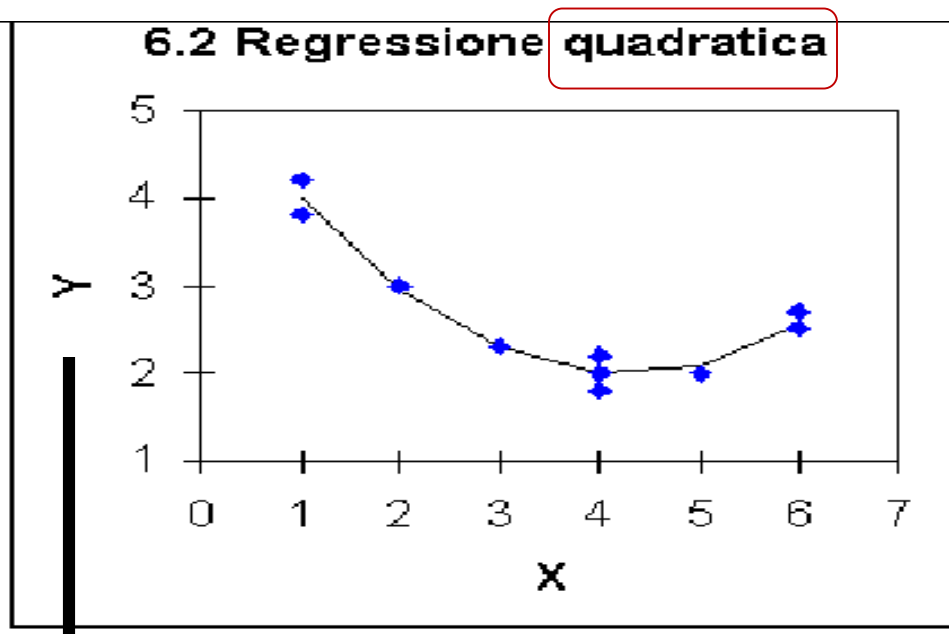


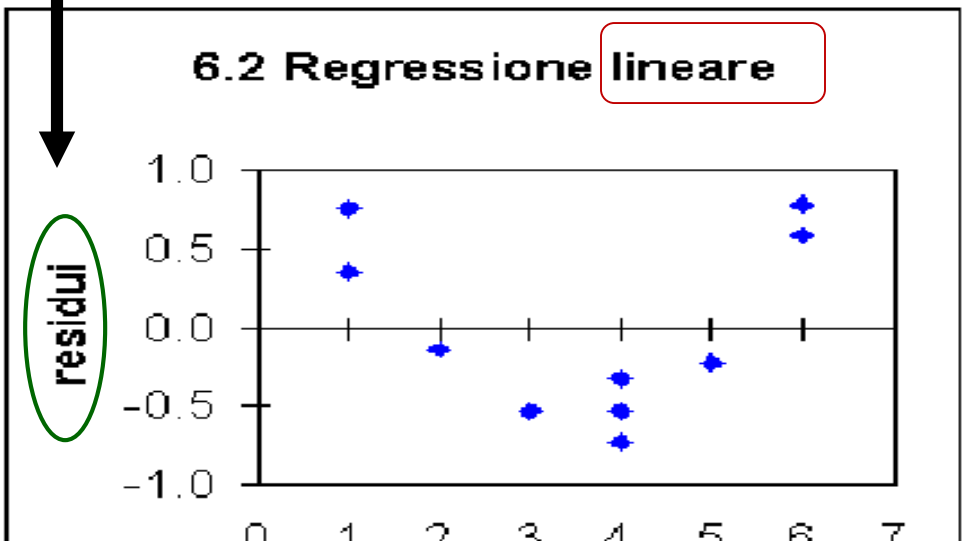
Analisi dei residui

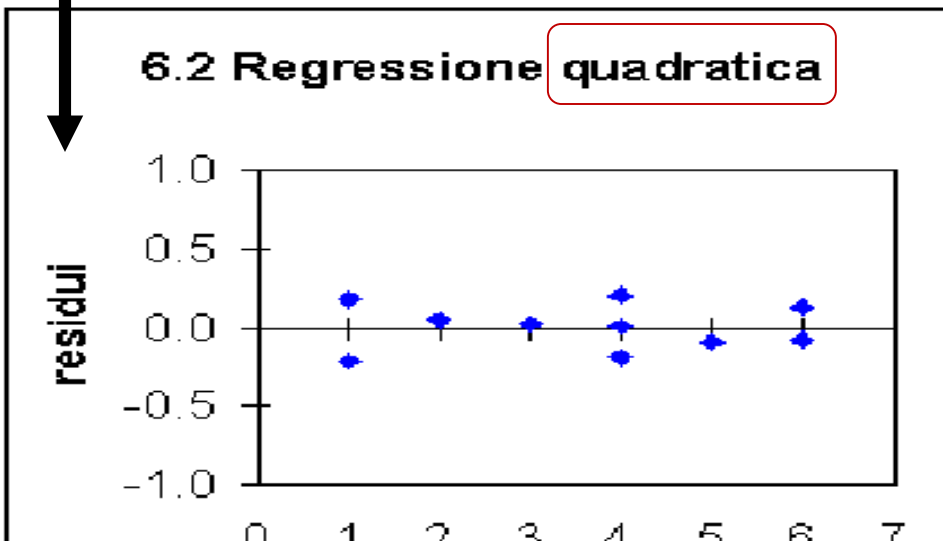


Deviazione dalla linearità









Regressione multipla

- Relazione tra una variabile dipendente e diverse variabili indipendenti
- La regressione non può essere visualizzata in un grafico bi-dimensionale (tante X)
- La procedura di regressione multipla stima una equazione lineare nella forma:

$$Y = a + b_1*X_1 + b_2*X_2 + ... + b_p*X_p$$

- Con più variabili indipendenti si ricorre a tecniche di **analisi multivariata** (correlazione canonica)

Regressione multipla

$$Y = a + b_1*X_1 + b_2*X_2 + \ldots + b_p*X_p$$

- a = valor medio di Y quando tutte le X sono pari a 0
- b_i = variazione media di Y associata a una variazione unitaria di X_1 quando tutte le altre X sono costanti

ATTENZIONE! Se b1>b2 allora X1 è più importante di X2? NO, perché cambiando la scala della variabile cambia il valore del coefficiente!

Coefficienti standardizzati

Procedura

- Standardizzare ciascuna variabile sottraendo ai valori la rispettiva media e dividendo per la rispettiva deviazione standard
- Stimare i parametri del modello usando le variabili standardizzate

Numero di variabili indipendenti

- La regressione multipla suggerisce una tecnica “seducente”: inserire quante più variabili indipendenti e selezionare quelle che risultano significative (*backward, forward, stepwise*)
- Si raccomanda un numero di osservazioni 10-20 volte superiore al numero delle variabili indipendenti

Multicollinearità

- Nel caso si abbiano a disposizione numerose variabili indipendenti, è opportuno verificare se i regressori risultano correlati tra loro.
- Ad esempio, in studi in pieno campo in una località, quando si consideri la risposta fenologica della pianta come funzione di temperatura e fotoperiodo, è frequente il caso in cui ci sia una elevata correlazione tra le due variabili meteorologiche.
- La stima dei parametri in queste condizioni è del tutto inutile ai fini previsionali.

Multicollinearità

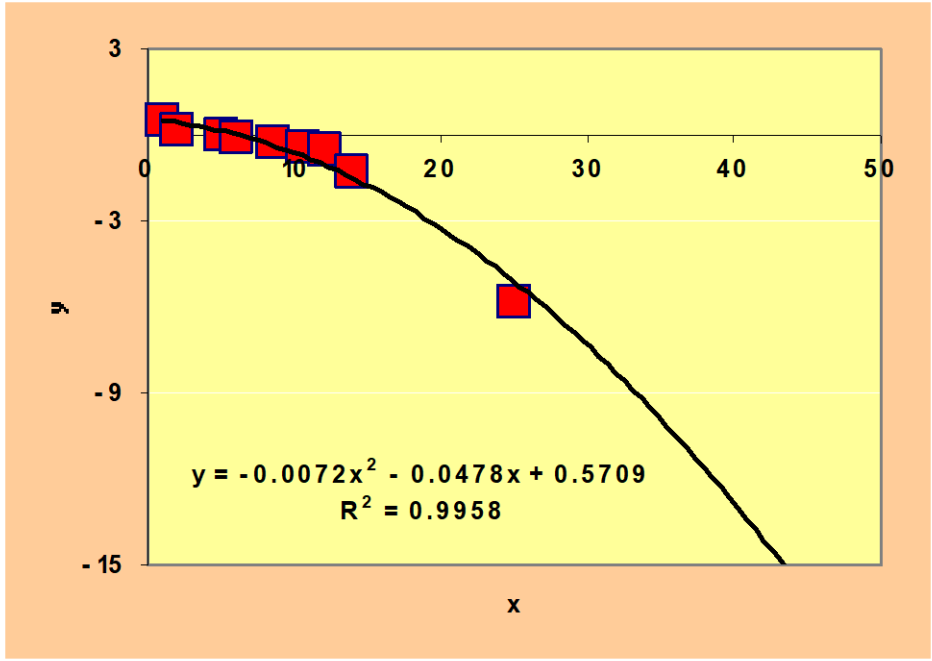
- Inserendo nel modello (regressione multipla) molte variabili «ridondanti» e quindi correlate tra di loro (es. quota e temperatura oppure diametro e area basimetrica, ecc…) ottengo un aumento di complessità del modello senza migliorarne la capacità previsionale.
- Più «rumore» che «segnale»
- È sconsigliato dunque effettuare tante misure della stessa categoria di variabile.

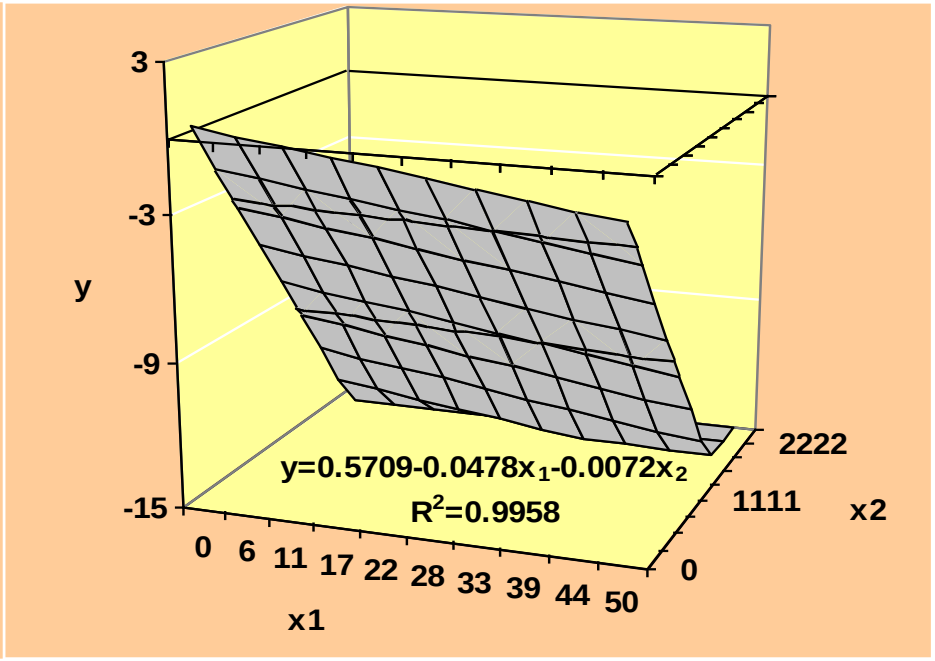
Regressione non lineare

- I modelli non lineari sono più difficili da specificare e stimare: definizione della funzione, dichiarazione e inizializzazione dei parametri
- La stima dei parametri è un processo iterativo (problemi di convergenza: valori iniziali, metodo iterativo, passo di iterazione)
- Output: significatività della regressione (test F), valori stimati dei parametri, errore standard asintotico, matrice di correlazione dei parametri
- Se possibile ricorrere alla linearizzazione

Modelli non lineari linearizzabili

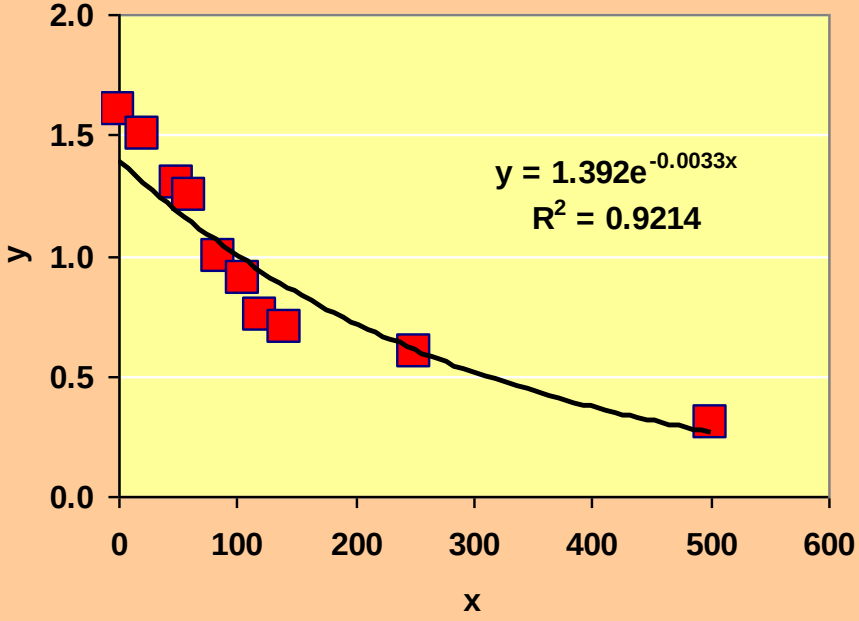
- **Modelli polinomiali**. Es. parabola: $y=a+bx+cx^2$ [come reg. multipla con due var. indipendenti]

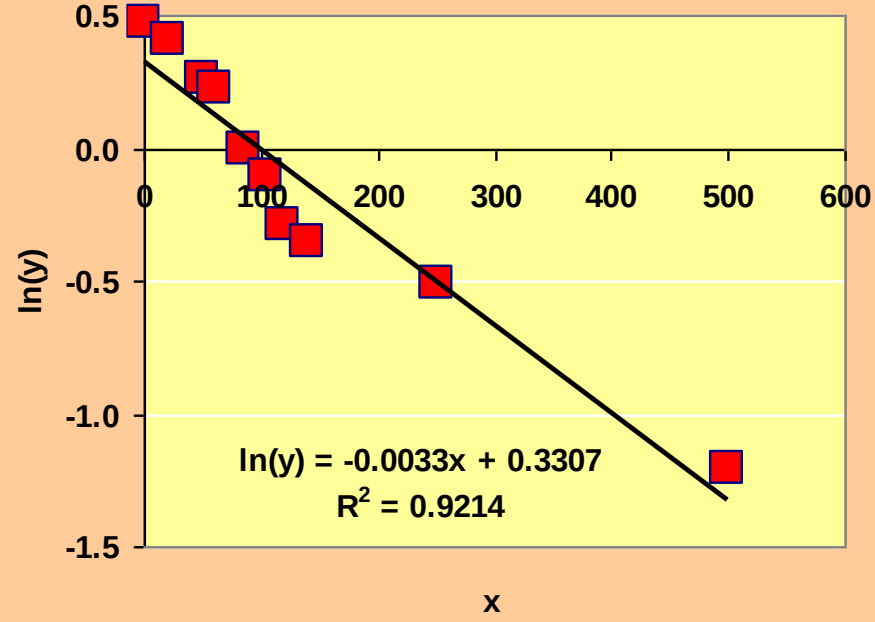




Modelli non lineari linearizzabili

Modelli esponenziali. Es. decadimento: $y=a \cdot e^{-k \cdot x}$
[trasformazione logaritmica: $ln(y)=ln(a)-k \cdot x$]

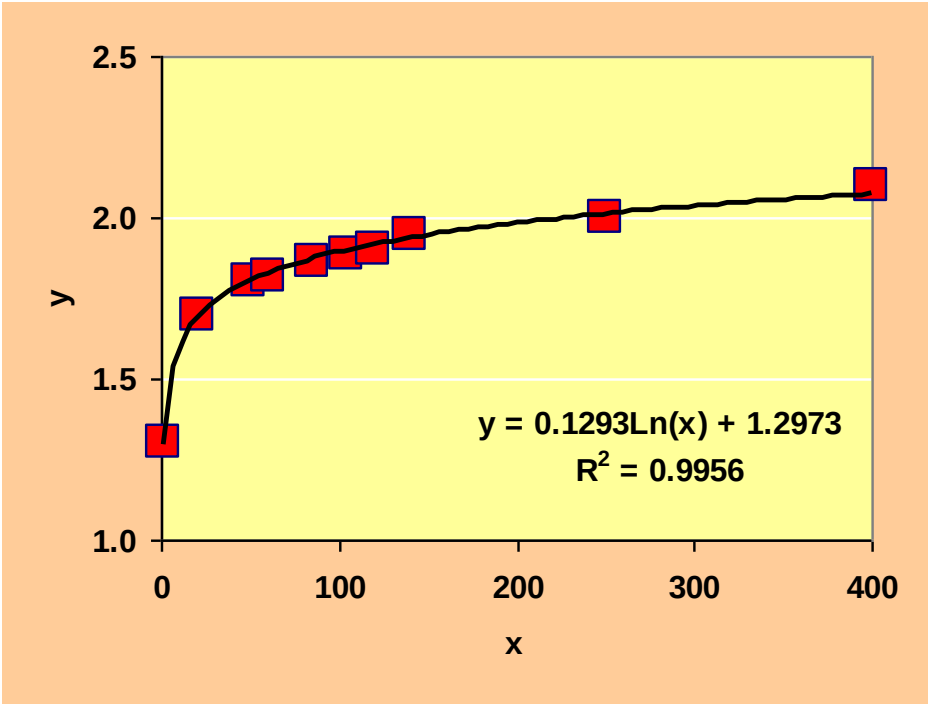


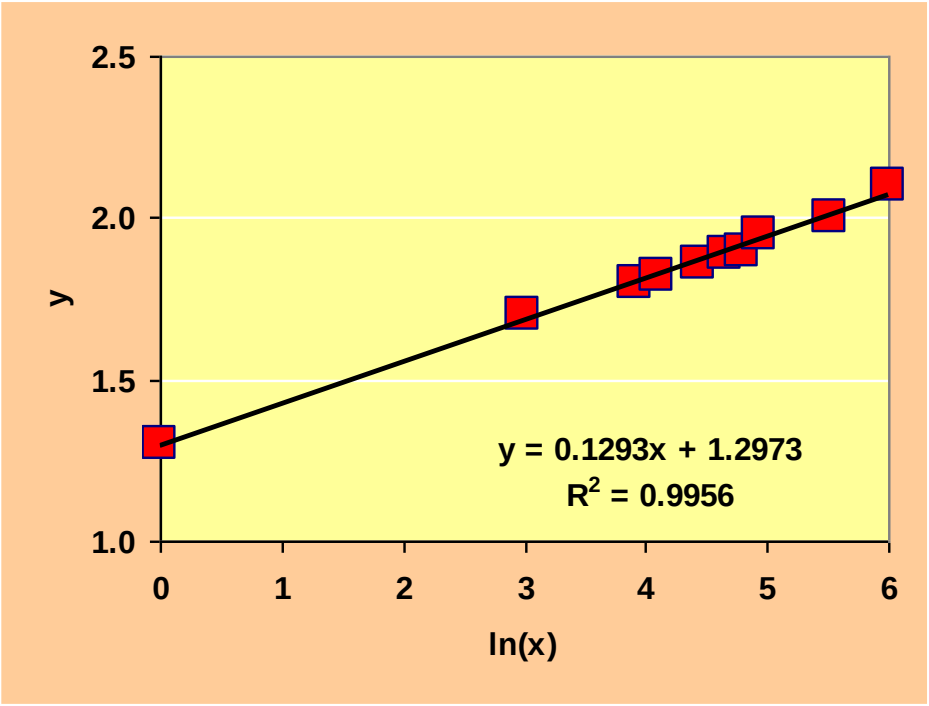


Modelli non lineari linearizzabili

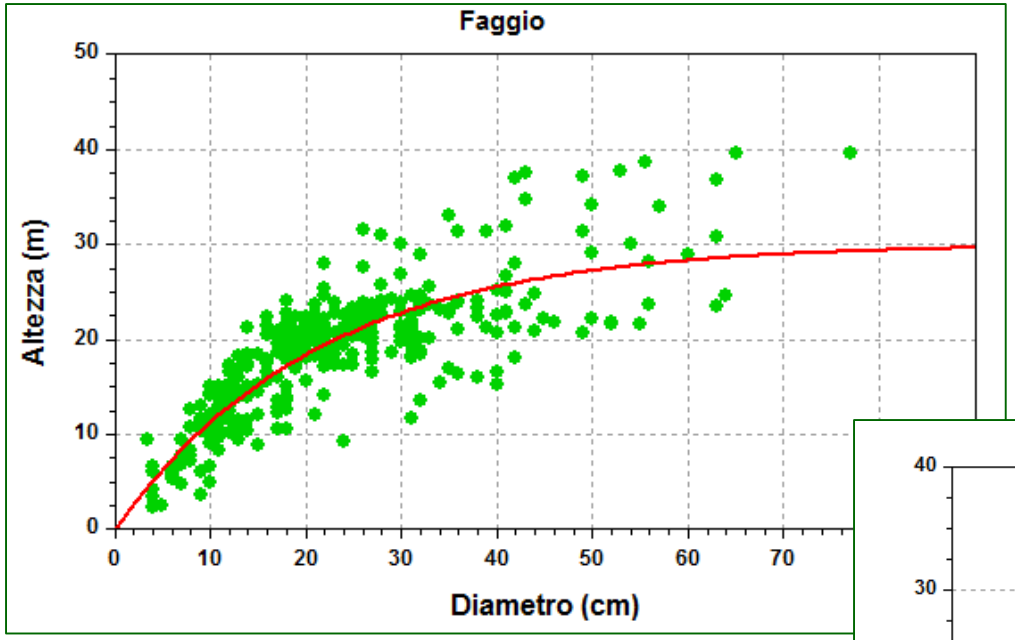
Modelli logaritmici. Es. : ***y=a+b·ln(x)***

[ln(x) variabile indipendente]

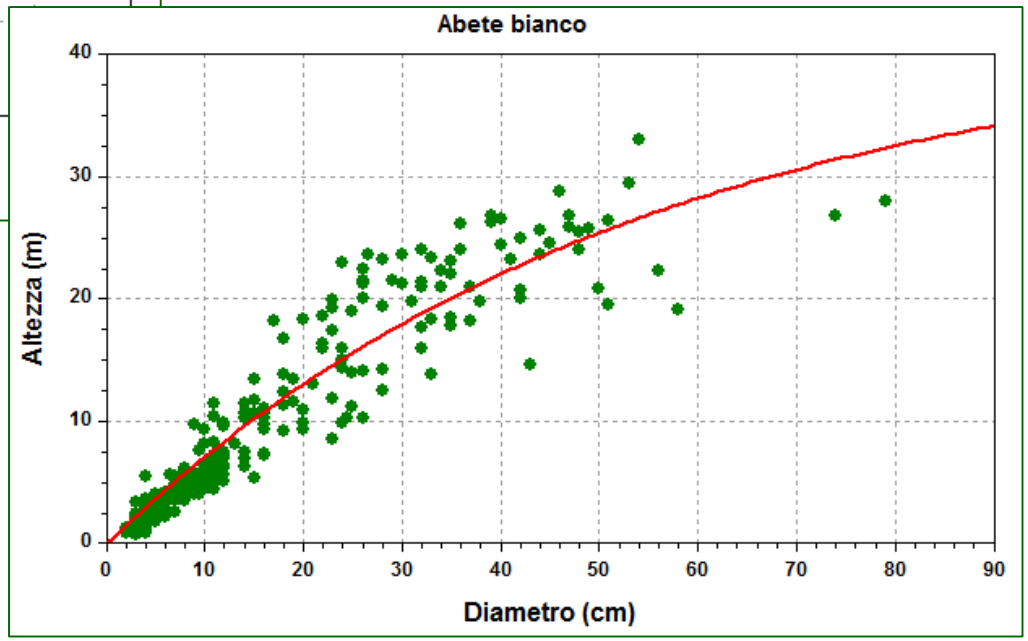




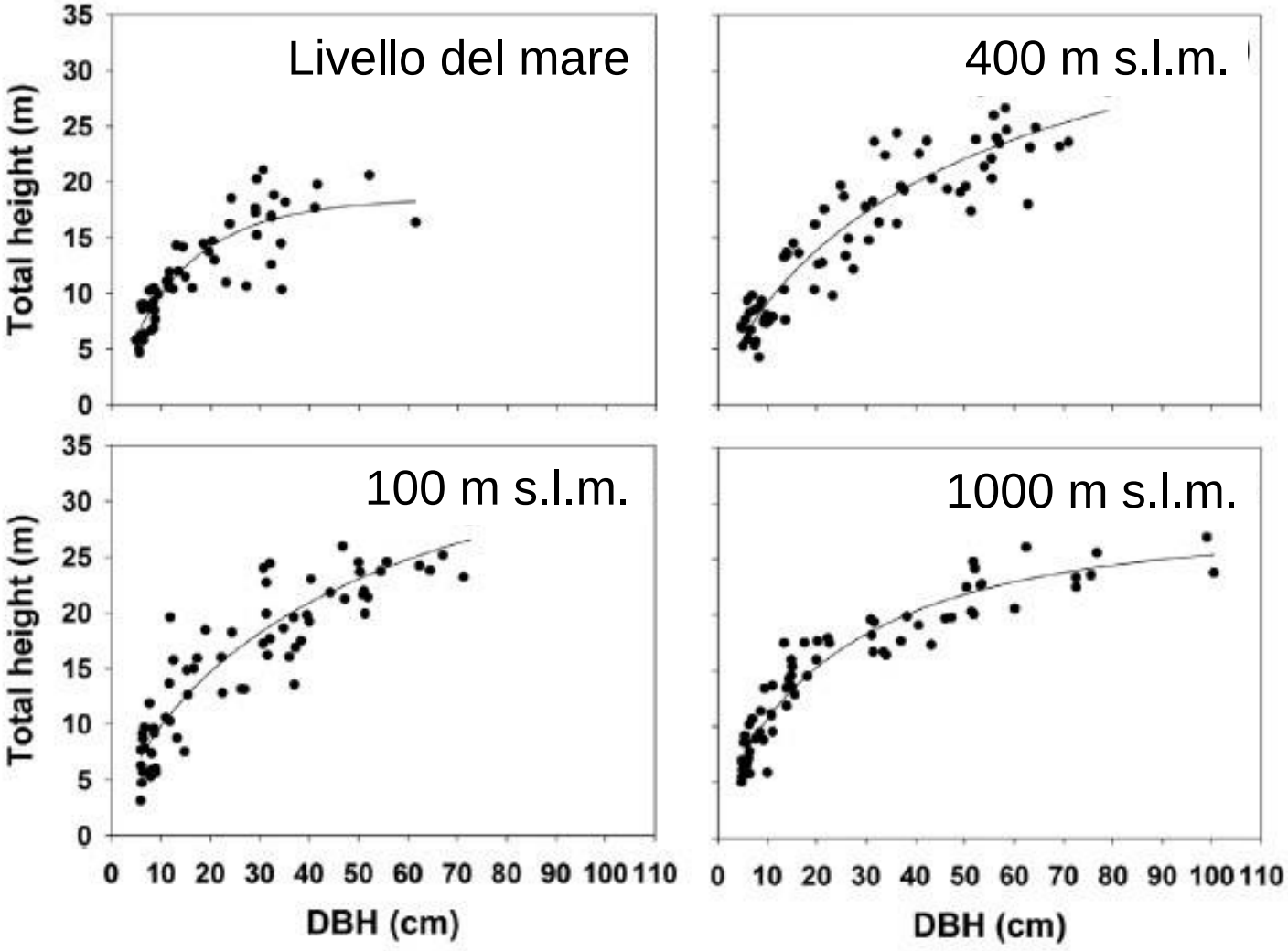
Modelli Logaritmici



Curve ipsometriche di faggio (R_2=0.93); e abete bianco (R_2=0.95) al Parco del Gran Sasso e Monti della Laga (Urbinati et al. 2014).



Modelli Logaritmici

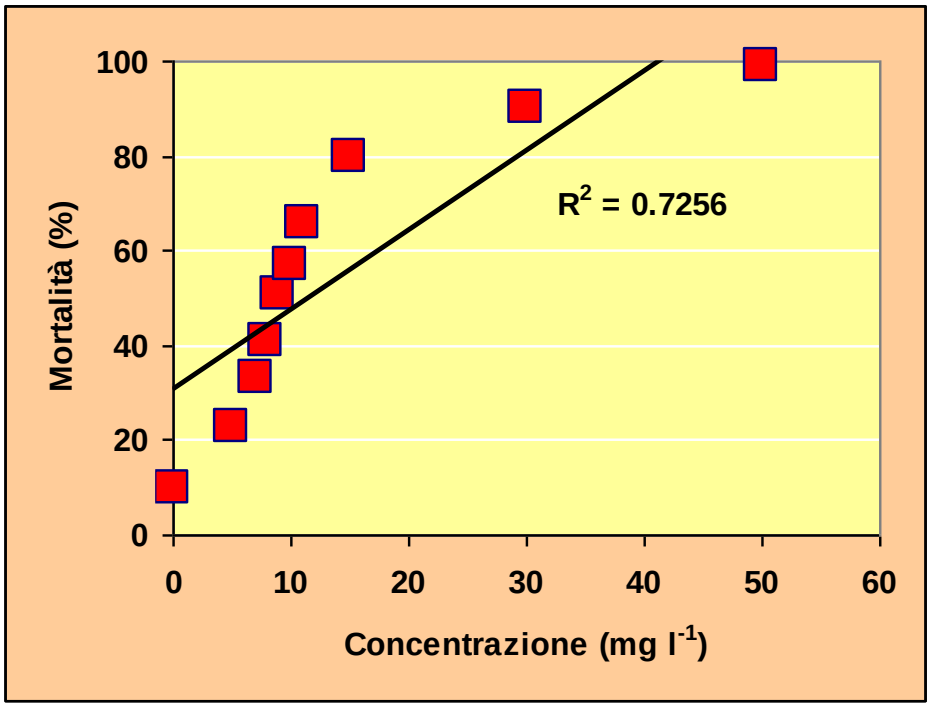


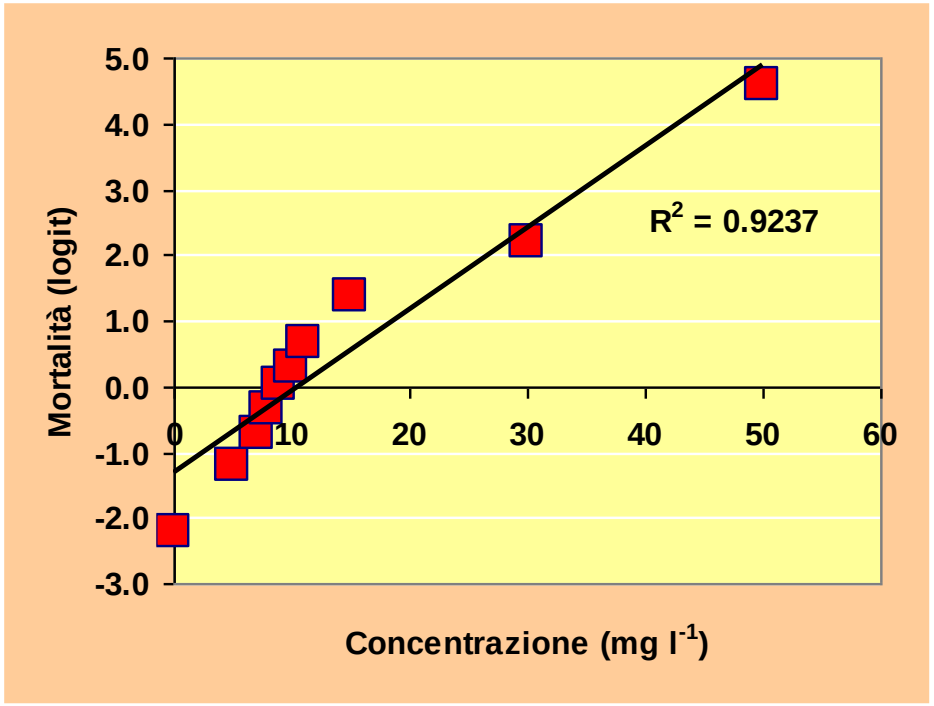
Modelli lineari per proporzioni trasformate: Logit

- **odds=p/(1-p)** (p: prob. evento favorevole)
- **logit=ln(odds)=y**
- Modello:

$$y=a+b_1 \cdot x_1+b_2 \cdot x_2+\ldots+b_n \cdot x_n$$

Relazioni Conc.-Mortalità



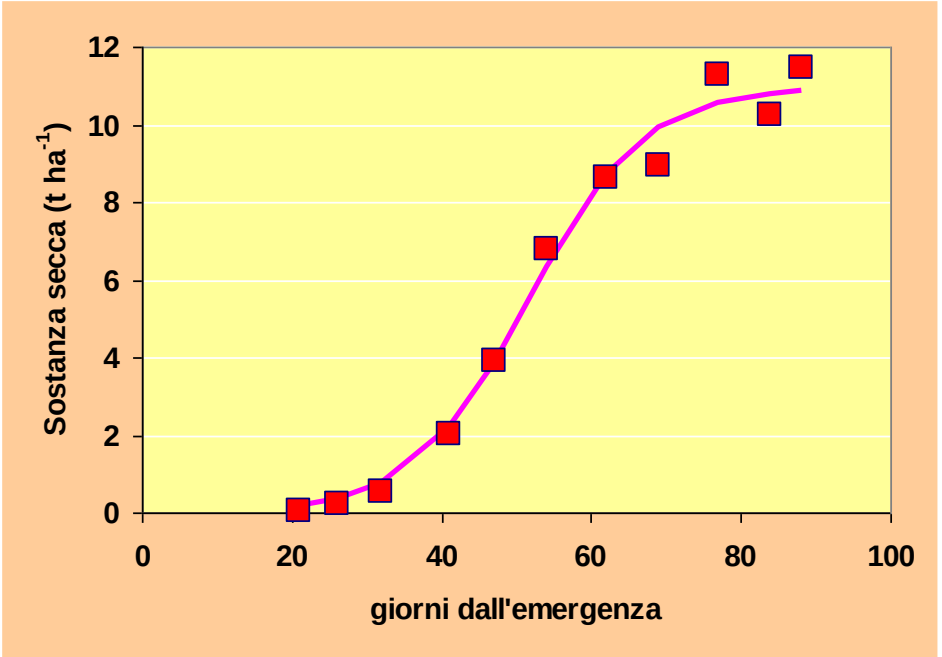


Accrescimento di piante erbacee

Equazione logistica

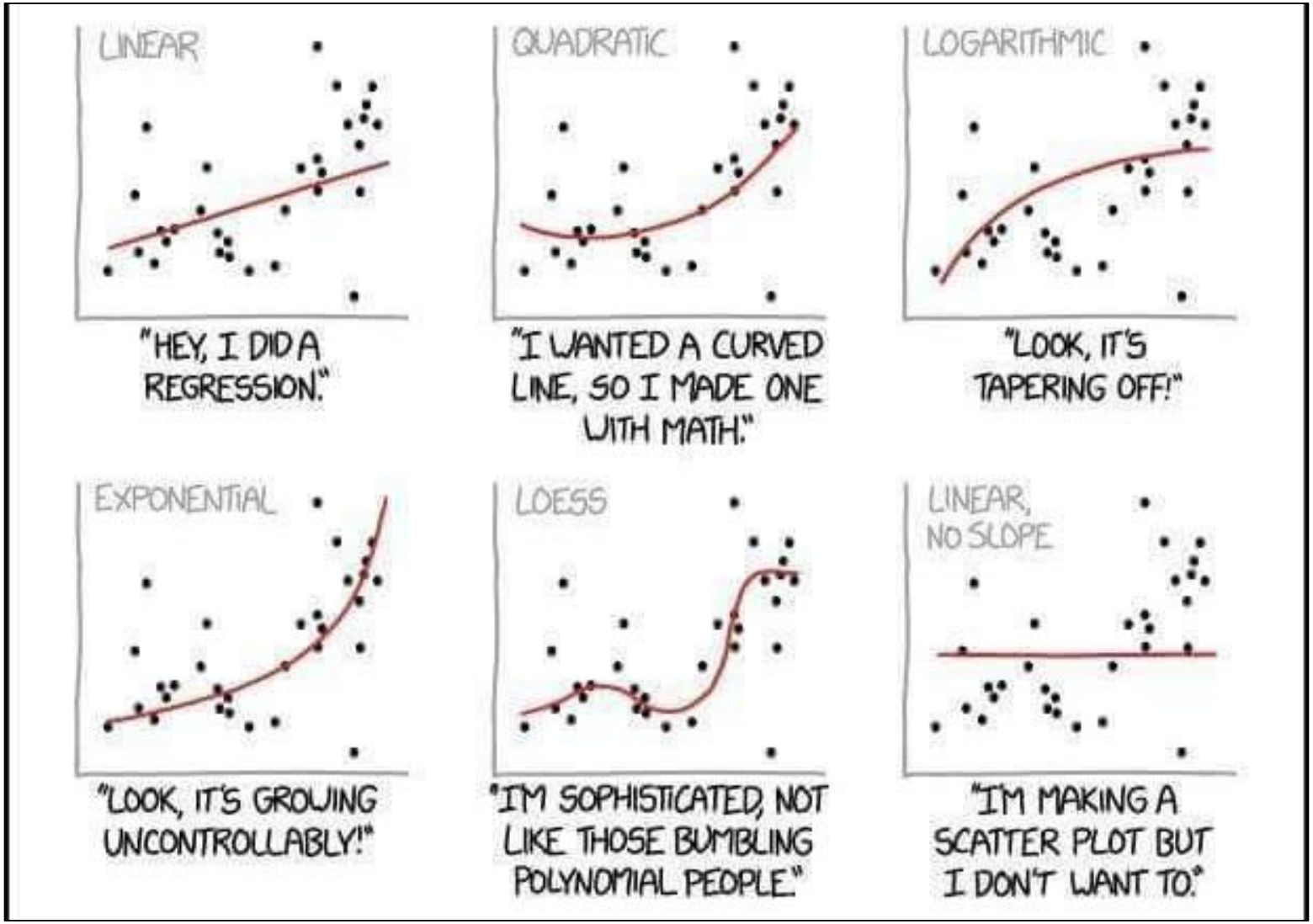
$$Y=a/[1+\exp(b+c\cdot t)]$$

t = giorni dalla emergenza



Metodi di fitting

e loro messaggio



Metodi di fitting

e loro messaggio