MODELLI DI REGRESSIONE

Il termine regressione è nato durante uno studio condotto da Pearson e Lee (1903) sull'altezza dei membri di gruppi familiari, analizzando 1078 coppie di valori padre-figlio per studiare la legge di Galton:

"ogni caratteristica di un uomo è condivisa con i parenti, ma in media ad un grado minore"

REGRESSIONE

- E’ la tecnica più adatta quando l’obiettivo principale consiste nello sviluppare un “modello predittivo”, cioè uno strumento che consenta di predire il livello di Y per un dato valore di X.

- studia una relazione di causa ed effetto, come varia una variabile
  - ☞ detta dipendente Y

  al variare di un'altra variabile
  - ☞ detta indipendente X

- La variabile X:
  - ☞ è una variabile non casuale
  - ☞ è affetta da errore trascurabile

per ogni valore di X esiste una sottopopolazione di valori di Y che seguono una distribuzione di Gauss.

Il più semplice modello di relazione che si può indagare è rappresentato da una retta la cui espressione matematica è:

$$y = a + b\,x$$

dato che per ogni x_i abbiamo una popolazione di valori y_i in statistica l'espressione della retta diventa:

$$y_i = a + b\,x_i + e_i$$

$\mathbf{e_i}$ è **l'errore** di misura legata alla variabilità dei soggetti sotto osservazione

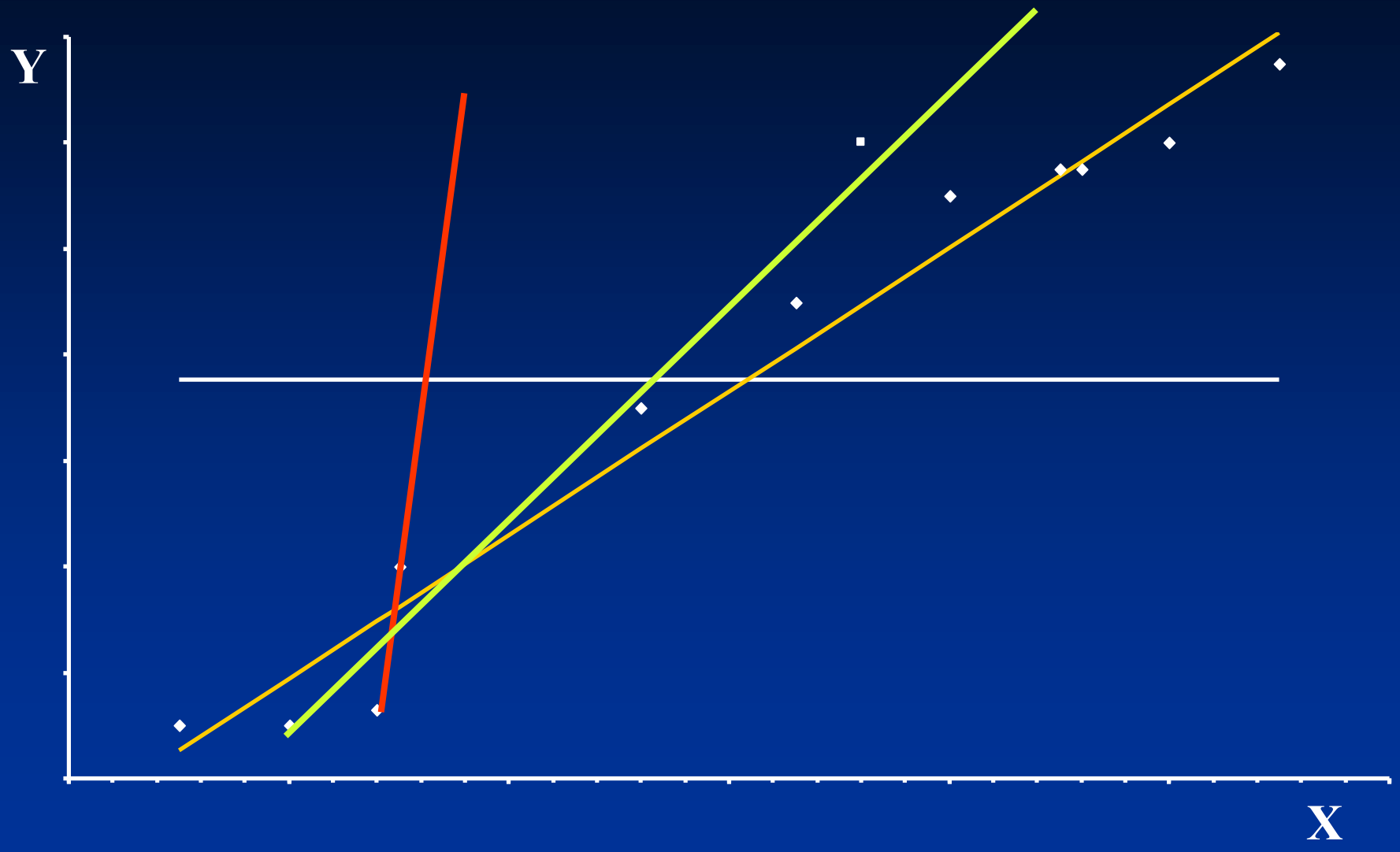
a è l'intercetta

b è il **coefficiente di regressione** (coefficiente angolare o pendenza della retta) ed esprime **quanto aumenta la variabile dipendente al variare unitario della variabile indipendente**.

IL PROBLEMA

Si dispone dei valori relativi alla statura in centimetri (x) ed al peso corporeo in Kg (y) di un campione di individui maschi adulti, della stessa classe di eta'

PESO (y)	60	57	71	66	65	78	82	78	62	70
STATURA(X)	171	169	181	173	178	180	185	183	170	174


Y
X

Le stime dei parametri a e b si ottengono con il :

METODO DEI MININI QUADRATI

che consiste nell'individuare la retta che renda minima la somma delle distanze al quadrato di ciascun punto y_i dai punti della retta di regressione stessa.

$$\sum [y_i - (a + bx_i)]^2 = \sum e_i^2$$

La stima dei due parametri della regressione sono:

$$\hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

l'equazione della retta diventa pertanto:

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

Per effettuare i conti è conveniente utilizzare le seguenti formula:

$$\hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} =$$

$$= \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\left[\sum x_i^2 - \frac{\left(\sum x_i \right)^2}{n} \right]}$$

Per risolvere il problema conviene preparare la seguente tabella dei dati

i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	171	60	10260	29241	3600
2	169	57	9633	28561	3249
3	181	71	12851	32761	5041
4	173	66	11418	29929	4356
5	178	65	11570	31684	4225
6	180	78	14040	32400	6084
7	185	82	15170	34225	6724
8	183	78	14274	33489	6084
9	170	62	10540	28900	3844
10	174	70	12180	30276	4900
Tot	**1764**	**689**	**121936**	**311466**	**48107**

effettuando i conti si ha:

$\bar{x} = 1764/10 = 176.4$
$\bar{y} = 689/10 = 68.9$

$$\hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]} =$$

$$= \frac{121936 - (1764 \times 689)/10}{311466 - (1764)^2/10} = \frac{396.4}{296.4} = 1.3374$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 68.9 - 1.3374 \times 176.4 = -167.0136$$

VERIFICA DELLE IPOTESI SU b

ANALISI DELLA VARIANZA DELLA REGRESSIONE.

$$H_0: b = 0$$

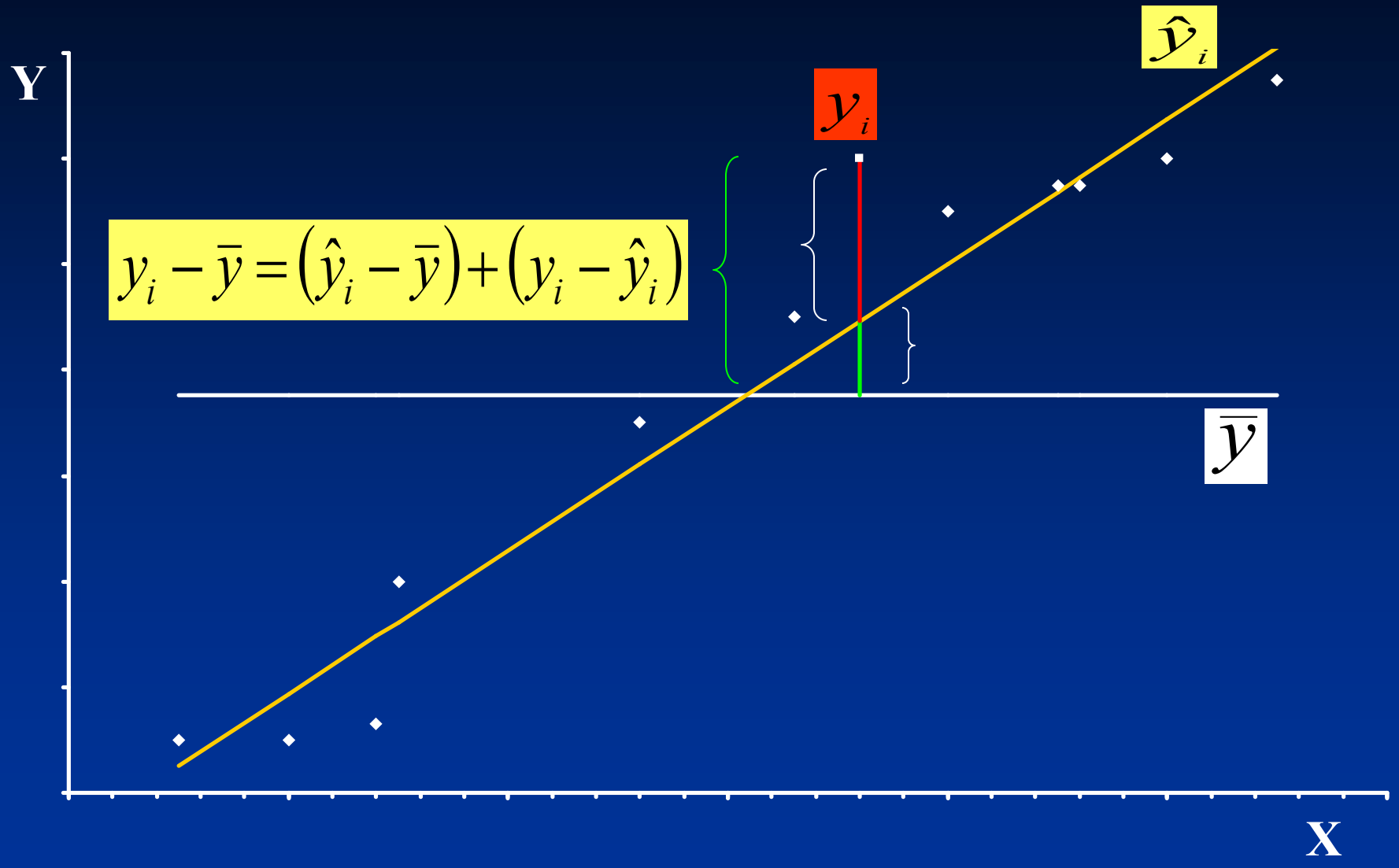
$$H_1: b \neq 0$$

L'obiettivo è di analizzare se le variazioni della Y associate alla X siano maggiori di quelle dovute al caso.

Se non rifiutiamo H_0 :

- ☞ forse c'è una relazione tra X ed Y ma non è possibile stimare Y conoscendo i valori di X;
- ☞ forse c'è un legame tra X ed Y ma non è descritto da una retta.

ANALISI DELLA VARIANZA



Si osservi che

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

Per tutti gli scostamenti si ottiene

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

Devianza Totale = Devianza di Regressione + Devianza residua

G.L.	N-1	1	N-2

STATISTICA TEST

La statistica test per la verifica dell'ipotesi nulla b = 0 mirerà a valutare quanto è più grande la varianza di regressione rispetto alla varianza residua:

$$F = \frac{dev.regr. / g.l.(1)}{dev.res. / g.l.(N-2)} =$$

$$= \frac{\frac{\sum (\hat{y}_i - \bar{y})^2}{1}}{\frac{\sum (y_i - \hat{y}_i)^2}{N-2}}$$

REGOLA DI DECISIONE

Fissato $\alpha = 0,05$
Individuata la distribuzione della statistica test (F-Fisher)

se Fcalc > Ftab allora rifiuterò H_0

$$DEV.TOTALE = \sum \left(y_i - \bar{y} \right)^2 = \sum y_i^2 - \left(\sum y_i \right)^2 / n$$

$$DEV.REGRESSIONE = \sum \left(\hat{y}_i - \bar{y} \right)^2 = b^2 \cdot \sum \left(x_i - \bar{x} \right)^2$$

$$= \frac{\left[\sum \left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right) \right]^2}{\sum \left(x_i - \bar{x} \right)^2} = \frac{num.\ \ b^2}{dev.x}$$

$$DEV.RESIDUA = DEV.TOT. - DEV.REG.$$

$$\hat{y}_i = \hat{a} + \hat{b}x_i = \bar{y} - \hat{b}\bar{x} + \hat{b}x_i =$$

$$= \bar{y} + \hat{b}(x_i - \bar{x})$$

$$\hat{y}_i - \bar{y} = \hat{b}(x_i - \bar{x})$$

$$\sum(\hat{y}_i - \bar{y})^2 = \hat{b}^2 \sum(x_i - \bar{x})^2$$

Per il nostro esempio

DEV. TOTALE = $\sum\left(y_i-\bar{y}\right)^2=\sum y_i^2-\left(\sum y_i\right)^2/n=$

$=48107-689^2/10=634.9$

DEV. REGRESSIONE= $\sum\left(\hat{y}_i-\bar{y}\right)^2=b^2\cdot\sum\left(x_i-\bar{x}\right)^2=$

$$=\frac{\left[\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)\right]^2}{\sum\left(x_i-\bar{x}\right)^2}=\frac{396.4^2}{296.4}=530.14$$

DEV. RESIDUA = $\text{TOTALE}-\text{REGRES}.=634.9-530.14=104.76$

Nell'esempio

A N O V A

Sorgenti di variazione	DEVIANZE	G.L.	VARIANZE	F_{CAL}
REGRESSIONE	530.14	1	530.14	40.48
RESIDUA	104.76	8	13.095	
TOTALE	634.9	9		

Essendo $F_{tab} = F_{1,8} = 5.32 < F_{cal} = 40.48$ rifiuto H_0

Dove F_{cal}= Varianza Regressione/Varianza residua = 530.14/13.095 = 40.48

VERIFICA DELLE IPOTESI SU b

$H_0: b = b_0$

$H_1: b \neq b_0$

$$T = \frac{b - b_0}{ES(b)}$$

$$ES(b) = \sqrt{\frac{\text{var}\, res}{devx}} =$$

$$= \sqrt{\frac{\sum (y_i - \hat{y}_i)^2 \Big/ N-2}{\sum (x_i - \bar{x})^2}}$$

La distribuzione della statistica test è t-Student con gradi di libertà pari ai gradi di libertà della varianza residua N - 2.

Fissato $\alpha = 0{,}05$ se $t_{calc} > t_{tab}$ rifiuterò H0

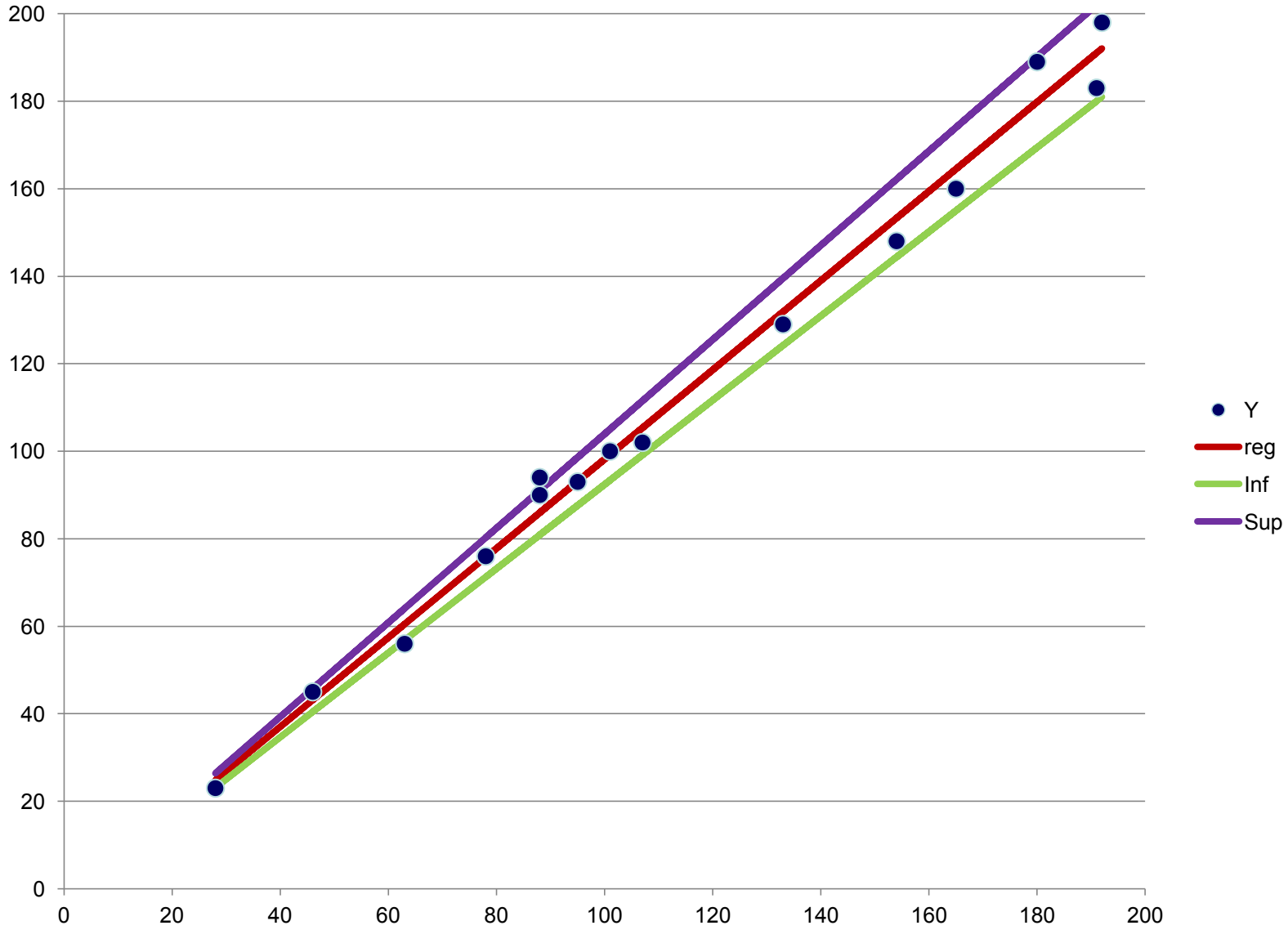
INTERVALLO DI CONFIDENZA per b

Può risultare utile determinare **l'intervallo di confidenza per b**

il metodo è lo stesso seguito per gli altri intervalli di confidenza:

stima ± (fattore di correzione × errore della stima)

$$\hat{b} \pm t_{\alpha,n-2} ES(b)$$


200
180
160
140
120
100
80
60
40
20
0
0
20
40
60
80
100
120
140
160
180
200
Y
reg
Inf
Sup

$H_0: b = b_0 = 0$

$H_1: b \neq b_0 \neq 0$

$$T = 1.3374/0.2102 = 6.3625$$

Essendo $T_{cal} = 6.3625 > T_{8,0.05} = 2.306$

Si rifiuta H_0, cioè b è significativamente diverso da zero

L'Intervallo di confidenza per b è : $1.3374 \pm (2.306 \times 0.2102)$

$0.8527 \leq b \leq 1.8221$

COEFFICIENTE DI DETERMINAZIONE

Il coefficiente di determinazione indica quanta parte delle osservazioni sono spiegate dal modello, cioè quanti dati cadono sulla retta stimata.

$$R^2 = \frac{devregr}{devtot} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

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

$R^2 = 0$ la retta di regressione coincide con la retta della media della variabile Y.

$R^2 = 1$ la retta di regressione spiega una alta percentuale di dati.

Nel caso del nostro esempio $R^2 = 0.835$

cioè il modello spiega 83.5% dei dati

$$\hat{y}_i = \hat{a} + \hat{b}x_i = \bar{y} - \hat{b}\bar{x} + \hat{b}x_i =$$

$$= \bar{y} + \hat{b}(x_i - \bar{x})$$

$$V(\hat{y}_i) = V\left(\bar{y} + \hat{b}(x_i - \bar{x})\right) = V(\bar{y}) + (x_i - \bar{x})^2 \times V(\hat{b})$$

$$V(\hat{y}_i) = Vres / N + (x_i - \bar{x})^2 \times Vres / \sum (x_i - \bar{x})^2$$

La retta di regressione risulta utile per stimare dei valori y_i in corrispondenza di valori teorici di x_i.
Anche per y_i è utile costruire l'intervallo di confidenza:

Stima ± errore della stima × fattore di correzione

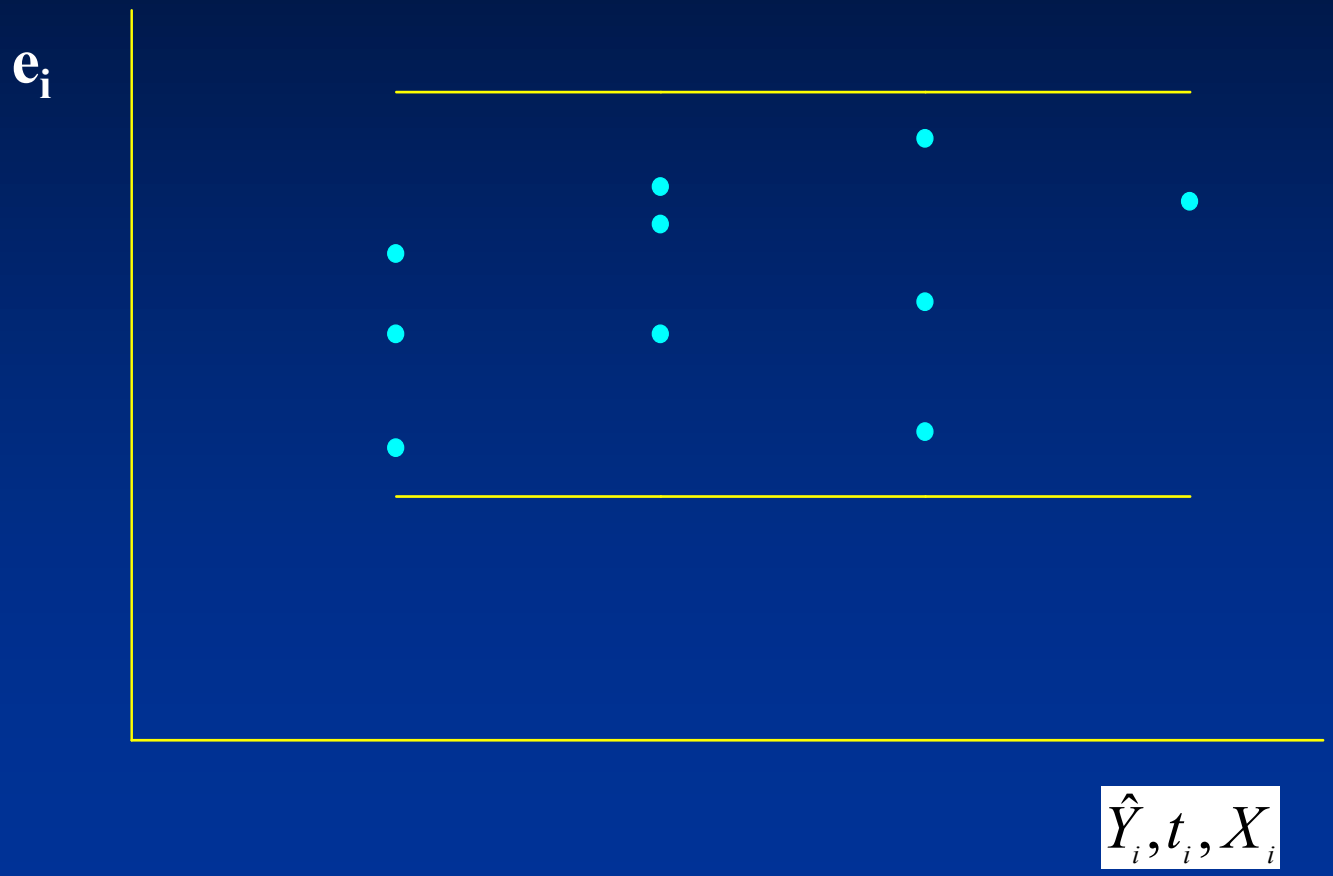
$$\hat{y}_i \pm t_{\alpha,n-2} ES(\hat{y}_i)$$

dove:

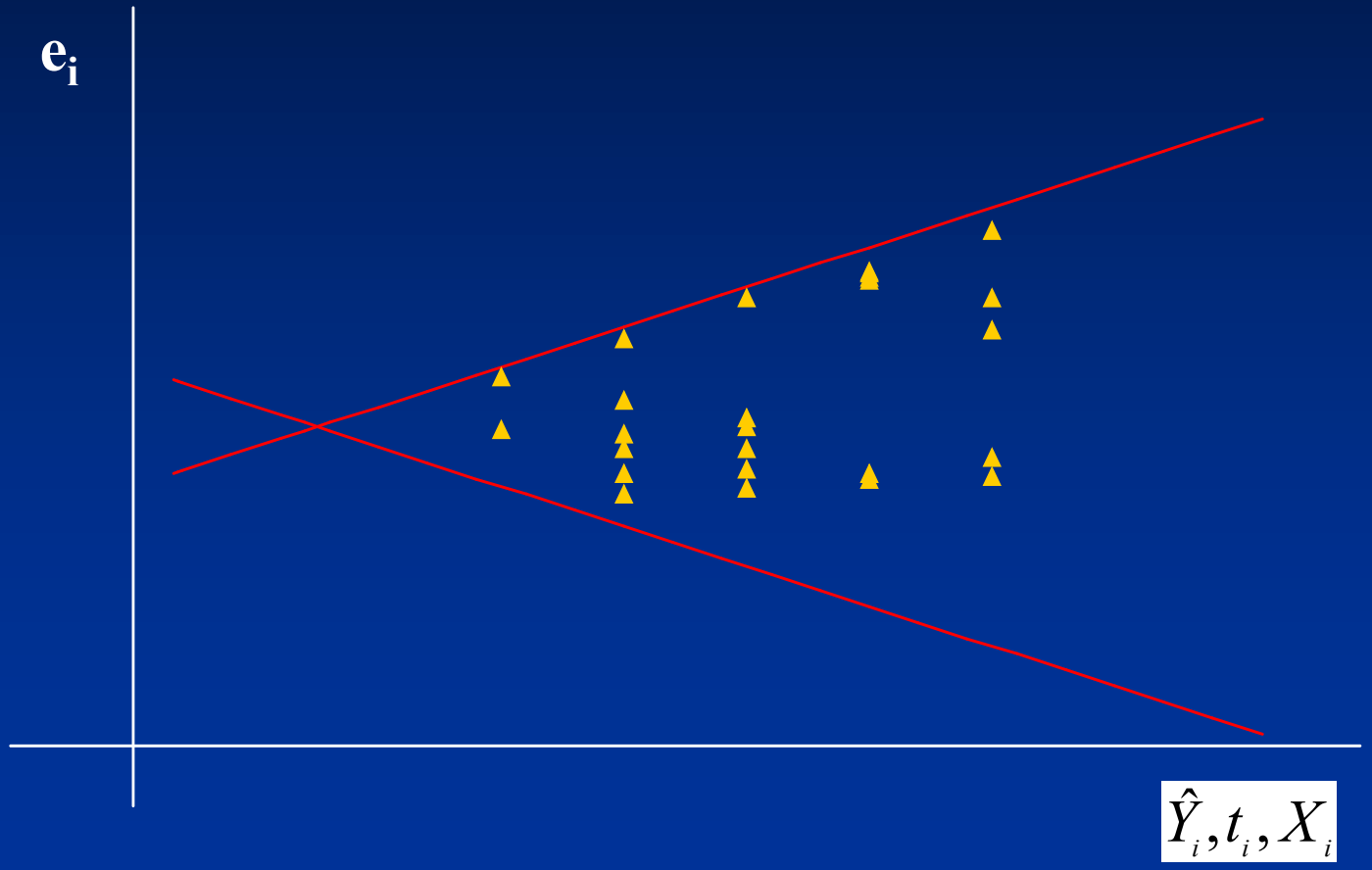
$$ES(\hat{y}_i) = \sqrt{\left(\frac{\sum(y_i - \hat{y}_i)^2}{n-2}\right)\left[\frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]}$$

I gradi di libertà di t tabulato dipendono dai gradi di libertà della varianza residua

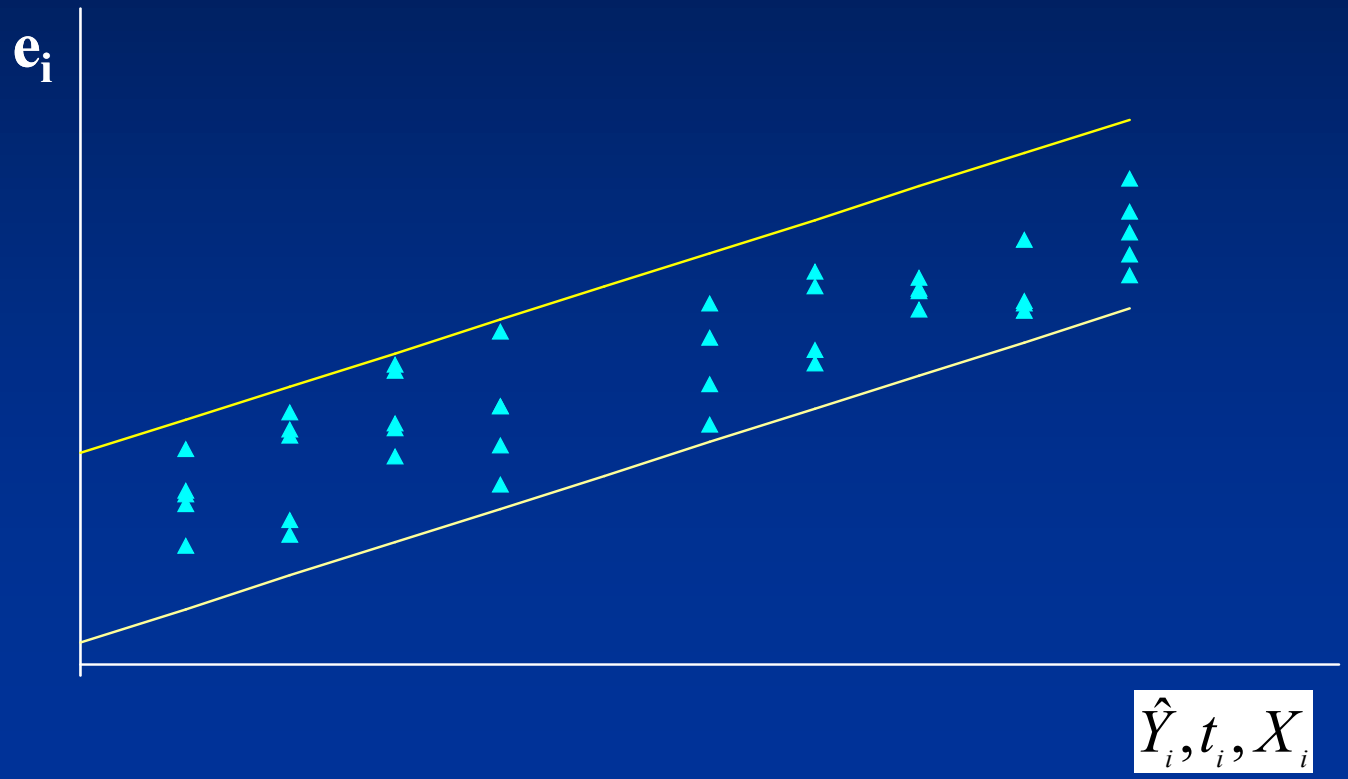
Il grafico evidenzia nessuna anomalia o violazione del modello e nessun effetto del tempo



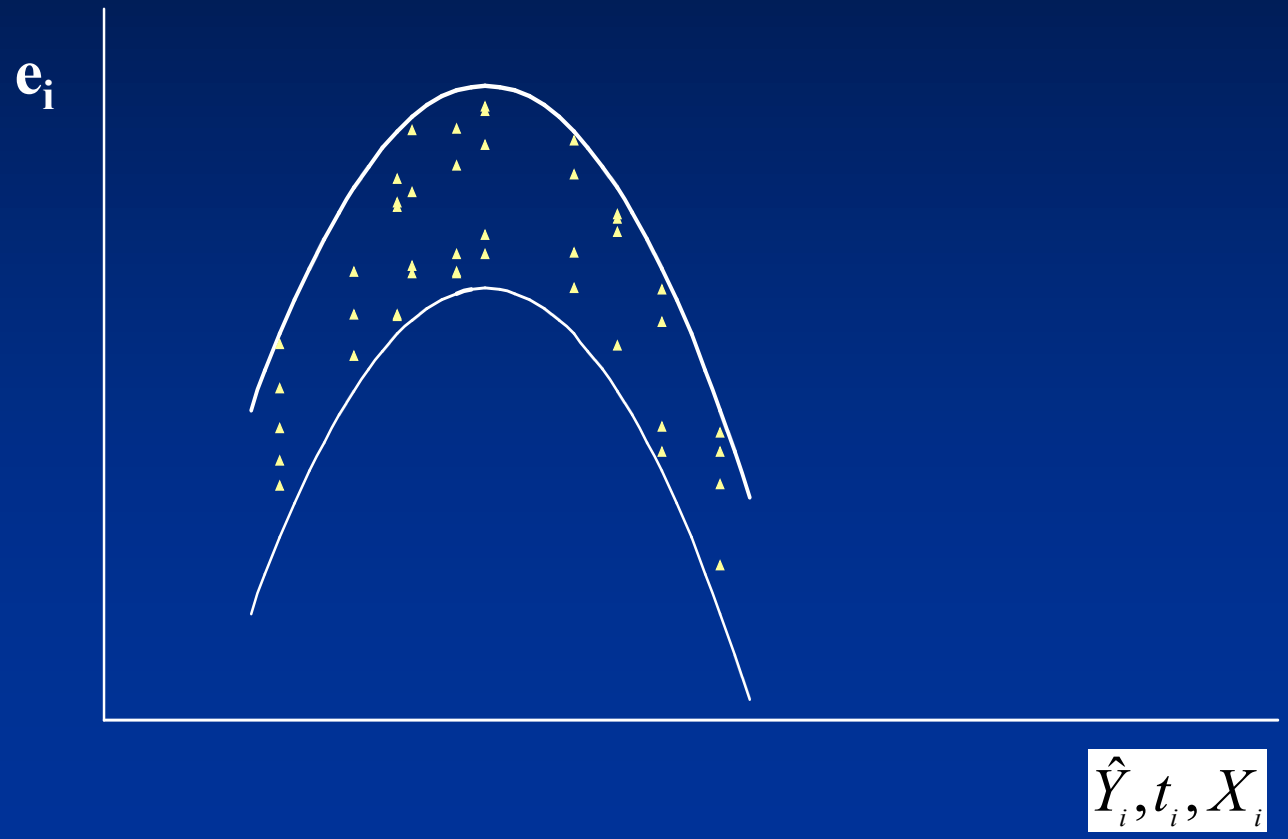
Il grafico evidenzia che la Varianza non è costante, se in funzione del tempo cresce con esso, potrebbe pertanto essere richiesta una trasformazione della variabile dipendente Y o un diverso metodo di stima dei parametri



Il grafico evidenzia errori nell'analisi. E' possibile una sottostima o sovrastima di Y, potrebbe essere stata omessa l'intercetta, gli effetti delle variabili indipendenti X potrebbero essere stati non adeguatamente considerati, potrebbe essere necessario introdurre un termine lineare in funzione del tempo



Il grafico evidenzia che il modello è inadeguato. Potrebbe essere necessario introdurre nuovi termini nel modello, trasformare le Y, considerare interazioni, introdurre termini lineari o quadratici in funzione del tempo



CONCLUSIONI SULLA ANALISI DI REGRESSIONE

- La regressione lineare semplice è una tecnica utile per lo studio della relazione fra una variabile indipendente (x) ed una variabile dipendente (y).

- L'analisi della regressione può sintetizzare la relazione fra queste due variabili e generare una formula che predica i valori di y in base ai valori di x.

Per impiegare le tecniche di regressione nella maniera migliore, i ricercatori dovrebbero tenere bene a mente i tre seguenti suggerimenti:

1. Tracciare il diagramma a punti ed esaminare i dati in anticipo per la ricerca dei valori estremi o di configurazioni particolari di tipo non lineare.

2. Esaminare la disposizione spaziale dei valori residui rispetto ai valori di $\hat{y}$ interpolati. Ricercando i valori estremi e quegli andamenti che possono suggerire l'opportunità di trasformare i dati

3. Nel presentare lo studio fornire al lettore informazioni sufficienti per comprendere a pieno l'analisi e poter riprodurre i risultati.
Le varianze dei coefficienti e la varianza residua dovrebbero essere riportate accanto alla equazione della retta di regressione

Un modello di regressione collega una misurazione Y ad una o più variabili esplicative:

$$Y = a+b_1X_1+b_2X_2+\ldots\ldots\ldots b_kX_k+ e$$

In questo caso la variabile Y è continua e si suppone distribuita normalmente.

Se la variabile Y è dicotomica Y=1 (successo) Y=0 (insuccesso) si utilizza la regressione logistica:

$$P(Y=1/X) = \exp(a+ \Sigma\ b_iX_i)/(1+\exp(a + \Sigma\ b_iX_i))$$

$$\text{Log}(p/1-p) = a + b_1X_1 + b_2X_2 + \ldots\ldots\ldots\ldots + b_kX_k$$

$0 \leq p \leq 1$ $0 \leq p/1-p \leq \infty$ $-\infty \leq \log(p/1-p) \leq +\infty$

Test di verifica dell'ipotesi $b_i = 0$

$Z = |\ b_i\ |$ /stima dell'errore standard di (b_i)

Intervallo di confidenza per b_i

$b_i \pm 1.96$ x (stima errore standard di (b_i))

Stima del rischio

OR = exp(b)