

1

Le Inferenze sul modello di regressione

PREVEDONO:

- Assunzione di normalità degli errori e inferenza sui parametri
- Analisi della Varianza
- Inferenza per la risposta media e la previsione
- Analisi dei residui
- Valori anomali

Assunzione di normalità degli errori

Assunzione:

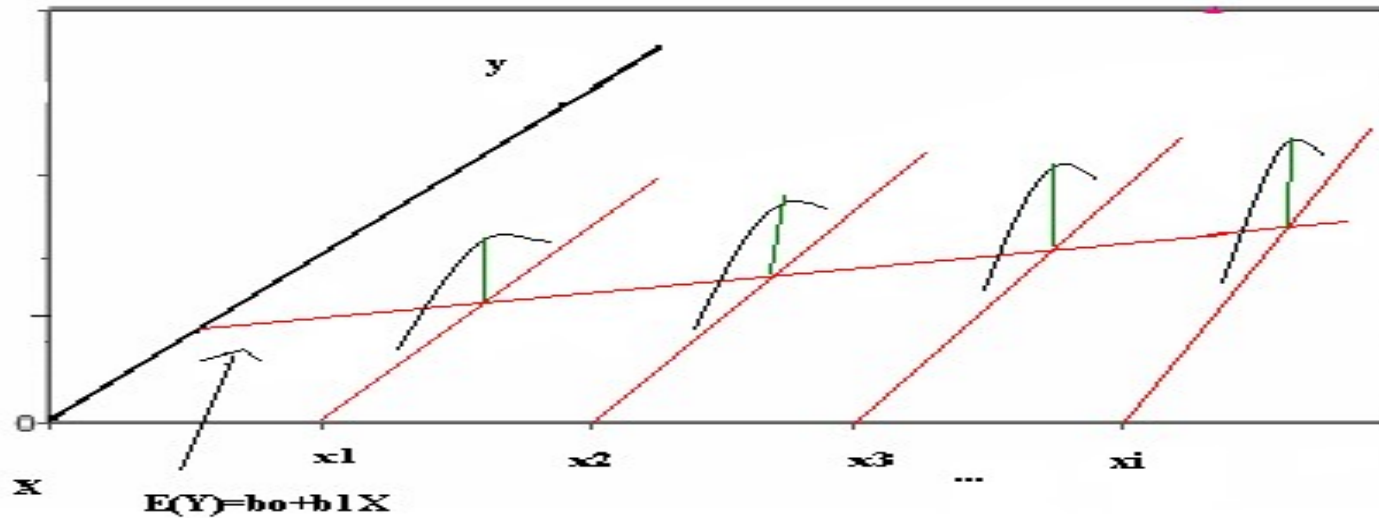
Le variabili casuali ε_i hanno distribuzione Normale.



Tenendo conto di questa ulteriore assunzione (insieme di ipotesi forti) il modello di regressione lineare può essere definito nel seguente modo:

Le osservazioni y_i sono realizzazioni di variabili casuali Normali con valore atteso $\beta_0 + \beta_1 x_i$ e varianza σ^2 ossia:

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$



1. **Normalità:** gli errori per ogni valore di x hanno distribuzione normale; il modello di regressione è robusto rispetto a scostamenti da tale ipotesi: le inferenze su retta e coefficienti non risultano seriamente compromesse da una distribuzione degli errori solo approssimativamente normale.
2. **Omoschedasticità:** variabilità costante per ciascun valore x (sia per i valori di x piccoli che elevati gli errori devono variare di un medesimo ammontare); tale ipotesi è cruciale: se non soddisfatta si debbono o trasformare opportunamente i dati o ricorrere a metodi di stima diversa (M.Q. ponderati)
3. **Indipendenza:** per ciascun valore di x ; importante quando i dati sono frutto di osservazioni nel corso del tempo: gli errori prodottisi in un certo periodo potrebbero essere correlati con gli errori del periodo precedente

Assunzione di normalità degli errori

L'introduzione di tale assunzione ha alcune importanti implicazioni:

- **Gli stimatori dei m.q. B_0 e B_1 , (combinazioni lineari della v.c. Y_j) si distribuiscono come una Normale bivariata; con valori attesi, varianze e covarianza date dalle precedenti espressioni.**

- **Si ha che:**

$$\frac{B_1 - \beta_1}{s(B_1)} \sim t_{n-2} \qquad \frac{B_0 - \beta_0}{s(B_0)} \sim t_{n-2}$$

dove t_{n-2} indica una v.c. t-Student con n-2 gradi di libertà e $s(B_0)$ e $s(B_1)$ indicano gli errori standard di B_0 e B_1 .

5

Vediamo che B_0 e B_1 sono stimatori lineari:

$$\begin{aligned}
 \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) &= \sum_{i=1}^n (x_i - \bar{x})Y_i - \sum_{i=1}^n (x_i - \bar{x})\bar{Y} = \\
 \sum_{i=1}^n (x_i - \bar{x})Y_i - \bar{Y} \sum_{i=1}^n x_i + n\bar{Y}\bar{x} &= \sum_{i=1}^n (x_i - \bar{x})Y_i - \bar{Y}n\bar{x} + n\bar{Y}\bar{x} = \sum_{i=1}^n (x_i - \bar{x})Y_i = \sum_{i=1}^n k_i Y_i \\
 \Rightarrow B_1 &= \frac{\sum_{i=1}^n k_i Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n \frac{k_i}{\sum_{i=1}^n (x_i - \bar{x})^2} Y_i = \sum_{i=1}^n k_i^* Y_i
 \end{aligned}$$

idem per B_0

6

Quindi poiché le y_i sono distribuite normalmente e B_1 è una funzione lineare di variabili normali indipendenti, ciò implica che anche esso si distribuisce normalmente.

“Diciamo”:

$$B_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

ma

$$Var(B_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ e se la varianza di } P \text{ } \sigma^2 \text{ non è nota}$$

ed è stimata con S^2 uno stimatore non distorto per la varianza di B_1 è dato da:

$$S^2(B_1) = \frac{S^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Le inferenze

Il risultato precedente:

$$\frac{B_1 - \beta_1}{S(B_1)} \sim t(n - 2) \quad \text{e} \quad \frac{B_0 - \beta_0}{S(B_0)} \sim t(n - 2)$$

Ci forniscono le statistiche campionarie cui fare riferimento per determinare gli intervalli di confidenza e condurre i test d'ipotesi sui parametri del modello



Intervalli di confidenza per i parametri

Gli intervalli di confidenza per i parametri β_0 e β_1 a un livello di confidenza $1 - \alpha$ sono dati da:

$$B_0 \pm t_{\alpha/2} s(B_0) \quad \text{e} \quad B_1 \pm t_{\alpha/2} s(B_1)$$

Esempio:

Le stime dei parametri del modello di regressione lineare semplice che pone (Y) il reddito medio pro-capite del 1999 (in 20 regioni) in funzione di quello del 1989 sono:

$$\hat{\beta}_1 = 1,255 \quad \hat{\beta}_0 = 0,595$$

con errori standard $s(B_1) = 0,091$ $s(B_0) = 4,111$

Fissato un livello di confidenza $1 - \alpha = 0,95$ il valore della t-Student con 10 g.d.l. è $t_{0,025} = 2,101$. Pertanto, si trovano gli intervalli di confidenza:

$$\text{int.conf.}(\beta_0) : 0,595 \pm 8,637$$

$$\text{int.conf.}(\beta_1) : 1,255 \pm 0,191$$

ossia

$$- 8,042 \leq \beta_0 \leq 9,232$$

$$1,064 \leq \beta_1 \leq 1,446$$

Verifica d'ipotesi per i parametri

Sotto l'ipotesi nulla $\beta_1 = b_1$ la statistica test è:

$$t = \frac{B_1 - b_1}{s(B_1)} \sim t_{n-2}$$

In corrispondenza del sistema d'ipotesi:

$$H_0 : \beta_1 = b_1 \text{ contro } H_1 : \beta_1 \neq b_1$$

A un livello di significatività α la regione di rifiuto è data dai valori della statistica test superiori in valore assoluto a $t_{\alpha/2}$.

La verifica d'ipotesi più frequente è:

$$H_0 : \beta_1 = 0 \text{ contro } H_1 : \beta_1 \neq 0$$

con

$$t = \frac{B_1}{s(B_1)} \sim t_{n-2}$$



ossia che la Y sia indipendente in media dalla X.

Verifica d'ipotesi per i parametri

Sotto l'ipotesi nulla che il parametro è $\beta_0 = b_0$ la statistica test è:

$$t = \frac{B_0 - b_0}{s(B_0)} \sim t_{n-2}$$

In corrispondenza del sistema d'ipotesi:

$$H_0 : \beta_0 = b_0 \text{ contro } H_1 : \beta_0 \neq b_0$$

A un livello di significatività α la regione di rifiuto è data dai valori della statistica test superiori in valore assoluto a $t_{\alpha/2}$.

La verifica d'ipotesi più frequente è:

$$H_0 : \beta_0 = 0 \text{ contro } H_1 : \beta_0 \neq 0$$

con

$$t = \frac{B_0}{s(B_0)} \sim t_{n-2}$$

 ossia che per $X=0$ il valore medio di Y sia nullo.

Esempio

Si vuole verificare il sistema d'ipotesi:

$$H_0 : \beta_1 = 0 \quad \text{contro} \quad H_1 : \beta_1 \neq 0$$

La statistica test, sotto l'ipotesi nulla, è: $t = \frac{1,255}{0,091} = 13,79$

A un livello di significatività $\alpha = 0,01$ corrisponde un valore della t-Student con 18 g.d.l. pari a: $t_{0,005} = 2,8784$

pertanto $t = 13,79 > 2,8784 = t_{0,005}$  **rifiuto** H_0

Osservato il valore del p-value=0,000, possiamo rifiutare l'ipotesi nulla per qualsiasi ragionevole livello di α .

Esiste pertanto una relazione lineare tra il reddito pro-capite del 1999 e quello del 1989.

Analisi della Varianza

Dalla **decomposizione della varianza totale** si può ricavare una procedura alternativa alla verifica d'ipotesi (è molto utile con più variabili esplicative).

Gli elementi dell'ANOVA (analisi della varianza) sono: SQT, SQE, SQR e i corrispondenti **gradi di libertà** (numero di termini sommati meno il numero di vincoli che tali quantità devono soddisfare).

$$SQT = \sum_{i=1}^n (y_i - \bar{y})^2 \quad \longrightarrow \quad \mathbf{n-1} \text{ g.d.l.}$$

$$SQE = \sum_{i=1}^n \hat{e}_i^2 \quad \longrightarrow \quad \mathbf{n-2} \text{ g.d.l.}$$

$$SQR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad \longrightarrow \quad \mathbf{1} \text{ g.d.l.}$$

Il test F seguente è basato sull'assunzione che sotto l'ipotesi nulla ($\beta_1=0$) sia SQE che SQR possono essere utilizzati per ottenere stimatori indipendenti della varianza degli errori σ^2 .

Per effettuare tale test si ottengono due stime campionarie della varianza σ^2 chiamate medie dei quadrati; precisamente, si dividono SQR ed SQE per i rispettivi g.d.l.; SQR ha un grado di libertà poiché si riferisce al solo coefficiente angolare; mentre SQE ne ha n meno due:



La somma dei quadrati suddivisa per il corrispondente numero di g.d.l. è detta **media dei quadrati** (MQ).

$$\begin{aligned}
 SQE &= \sum_{i=1}^n \hat{e}_i^2 & \longrightarrow & MQE = \sum_{i=1}^n \hat{e}_i^2 / n - 2 \\
 SQR &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 & \longrightarrow & MQR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / 1
 \end{aligned}$$

Poiché si è visto che la distribuzione F può essere vista come la distribuzione del rapporto tra due stime indipendenti della varianza in presenza di varianze uguali delle popolazioni: se MQE ed MQR sono indipendenti e se H_0 è vera entrambe sono stime della varianza della popolazione σ^2 . Sotto l'ipotesi nulla si dimostra che:

$$F = \frac{MQR}{MQE} \sim F(1, n - 2)$$

Poiché si è visto che la distribuzione F può essere vista come la distribuzione del rapporto tra due stime indipendenti della varianza in presenza di varianze uguali delle popolazioni: se MQE ed MQR sono indipendenti e se H_0 è vera entrambe sono stime (corrette) della varianza della popolazione σ^2 . Sotto l'ipotesi nulla si dimostra che:

Tavola ANOVA

e	Sorgente di variazione	Somma dei quadrati	Gradi di libertà	Media dei quadrati	F
	Regressione	SQR	1	$MQR = SQR / 1$	$F = MQR / MQE$
	Residuo	SQE	$n - 2$	$MQE = SQE / (n - 2)$	
	Totale	SQT	$n - 1$		

Analisi della Varianza

Nell'ultima colonna viene riportata la **statistica test F** dell'ANOVA che è definita come:

$$F = MQR/MQE$$

Per $F \approx 1$, MQR e MQE sono simili ossia l'ipotesi $\beta_1 = 0$

Per $F \gg 1$, si tenderà verso l'ipotesi $\beta_1 \neq 0$

Sotto l'ipotesi nulla $\beta_1 = 0$, la statistica test F ha distribuzione **F -Fisher** con parametri dati dai g.d.l. associati a MQR e MQE, ossia **1 e $n-2$** .

Fissato un livello di significatività α e calcolato il valore di F_α , se

$F > F_\alpha$  si rifiuta l'ipotesi nulla

Analisi della Varianza

ESEMPIO

Per l'esempio precedente si osserva la seguente tavola ANOVA:

Tavola ANOVA

Sorgente di variazione	Somma dei quadrati	Gradi di libertà	Media dei quadrati	<i>F</i>
Regressione	2283,6	1	2283,60	<i>F</i> = 192,06
Residuo	214,0	18	11,89	
Totale	2497,6	19		

Il p-value corrispondente a $F=192,06$ sulla distribuzione *F-Fisher* con 1 e 18 g.d.l. è praticamente uguale a zero, pertanto possiamo rifiutare l'ipotesi nulla $\beta_1 = 0$

Caso Multiplo: cenni

Nel caso di più regressori:

$$X_1, X_2, X_3, \dots, X_j, \dots, X_p$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \dots + \beta_j x_{ji} + \dots + \beta_p x_{pi} + \varepsilon_i$$

$$i=1, 2, \dots, n$$

ed in generale

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_j X_j + \dots + \beta_p X_p + \varepsilon$$

Sotto le analoghe ipotesi per il modello, il test F seguente è solitamente basato sull'assunzione di l'ipotesi nulla

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_j = \dots \beta_p = 0$$

E che anche qui sia SQE che SQR possono essere utilizzati per ottenere stimatori indipendenti della varianza degli errori σ^2 .

Per effettuare tale test si ottengono quindi le due stime campionarie della varianza σ^2 chiamate medie dei quadrati; ma questa quando si dividono SQR ed SQE per i rispettivi g.d.l. occorre tener conto che SQR ha **p** gradi di libertà poiché si riferisce ai **p** coefficiente delle X ; mentre SQE ne ha due n meno **$p+1$** (vengono stimati tale numero di coefficienti con i dati)



La tavola ANOVA:

Tavola ANOVA

Sorgente di variazione	Somma dei quadrati	Gradi di liberta	Media dei quadrati	F
Regressione	SQR	p	$MQR = SQR / p$	$F = \frac{MQR}{MQE}$
Residuo	SQE	n-p-1	$MQE = SQE / (n-p-1)$	MQE
Totale	SQT	n-1		

Ed il test F viene condotto secondo l'analogo meccanismo

Inferenza per la risposta media

La funzione di regressione per un dato valore $X=x_i$, restituisce la stima del valor medio $E(Y_i|x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i = \hat{Y}_i$

Sotto l'assunzione di Normalità, lo stimatore $\hat{Y}_i$ è distribuito come una Normale e inoltre:

$$\frac{\hat{Y}_i - E(\hat{Y}_i|x_i)}{s(\hat{Y}_i)} \sim t - Student(n - 2 \text{ g.d.l.})$$

segue che l'intervallo di confidenza per il **valore medio della** Y_i a un livello di confidenza $1 - \alpha$ è

$$\hat{Y}_i \pm t_{\alpha/2} s(\hat{Y}_i)$$

dove

$$s(\hat{Y}_i) = \sqrt{s^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{h=1}^n (x_h - \bar{x})^2} \right]}$$

Inferenza per la risposta media

$$\hat{Y}_i \pm t_{\alpha/2} s(\hat{Y}_i) \qquad s(\hat{Y}_i) = \sqrt{s^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{h=1}^n (x_h - \bar{x})^2} \right)}$$

Dall'espressione dell'intervallo di confidenza si osserva che:

- l'intervallo aumenta al crescere dell'errore standard $\sqrt{s^2}$
- l'intervallo diminuisce al crescere della dimensione campionaria n
- l'intervallo aumenta al crescere della distanza di x_i da $\bar{x}$

Inferenza per la previsione

Quando il modello di regressione lineare è utilizzato a **fini previsivi**, bisogna distinguere tra:

- previsione del **valor medio** di Y_i , ossia $E(Y_i|x_i)$
- previsione del **valore singolo** della Y_i

La stima puntuale delle due quantità è la stessa e pari a $\hat{\beta}_0 + \hat{\beta}_1 x_i = \hat{y}_i$ ma lo stimatore del valore singolo presenta un errore standard maggiore dello stimatore per il valore medio.

L'errore standard dello stimatore del singolo valore è:

$$s(Y_i - \hat{Y}_i) = \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{h=1}^n (x_h - \bar{x})^2}}$$

la cui stima si ottiene sostituendo a σ la stima $\sqrt{s^2}$

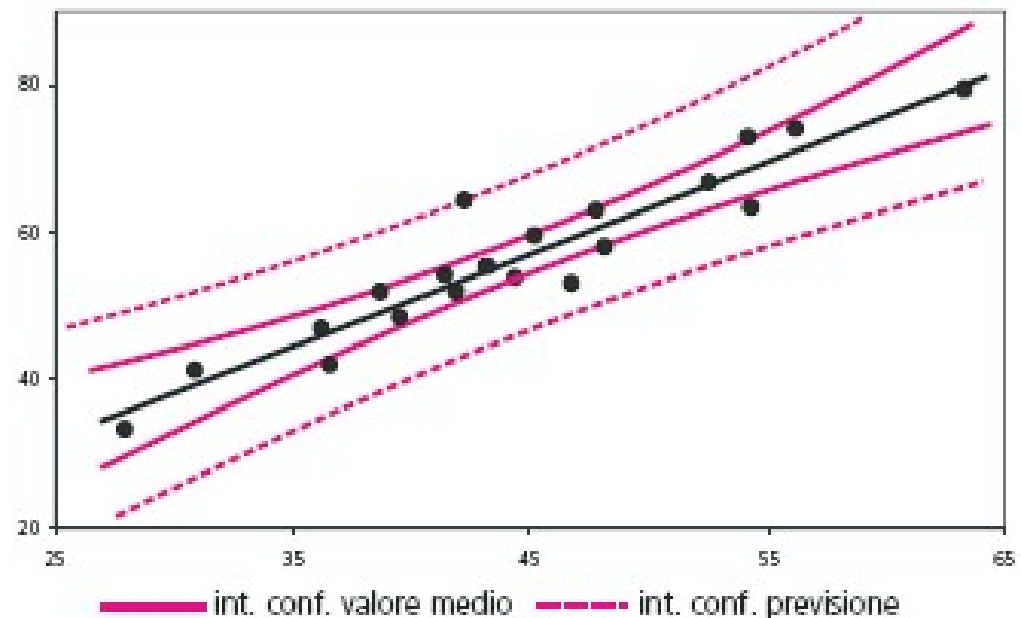
Inferenza per la previsione

Si dimostra che: $\frac{Y_i - \hat{Y}_i}{s(Y_i - \hat{Y}_i)} \sim t_{n-2}$

e ne consegue che l'intervallo di confidenza per la previsione del singolo valore Y_i a livello $1 - \alpha$ è dato da:

$$\hat{Y}_i \pm t_{\alpha/2} s(Y_i - \hat{Y}_i)$$

Intervalli di confidenza al 95% per il valore medio e il valore previsto.

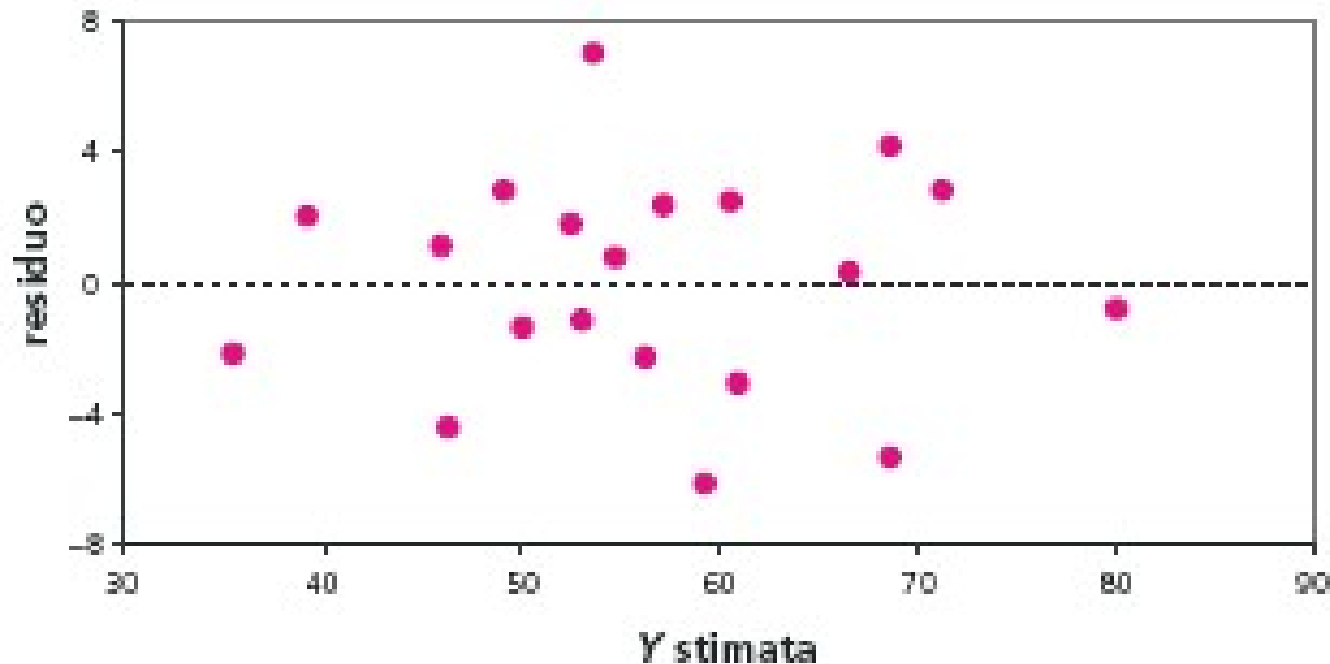


Analisi dei residui

Le proprietà degli stimatori dei parametri del modello richiedono alcune assunzioni. E' utile verificare la validità di tali assunzioni. Una tecnica di verifica si basa sull'**analisi dei residui**.

Se le assunzioni sono vere, ossia se il modello è **ben specificato**, i residui $\hat{\varepsilon}_j$ rifletteranno le proprietà attribuite ai termini di errore ε_j .

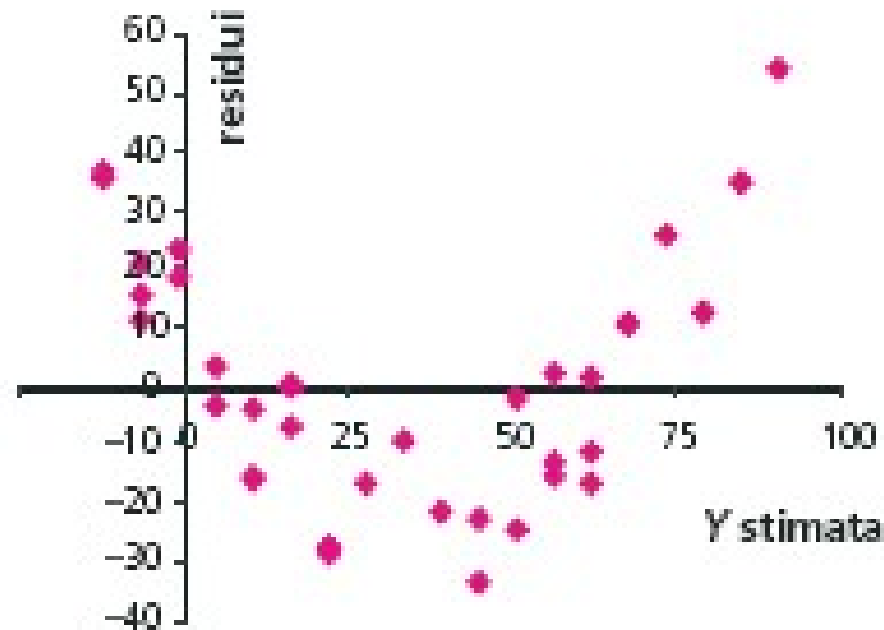
Un metodo grafico: **grafico dei residui**.



Analisi dei residui

Assunzione di linearità

Si assume che la funzione di regressione sia di tipo lineare; ciò può essere verificato analizzando il grafico dei residui. Nel seguente grafico è evidente che la relazione non è di tipo lineare.

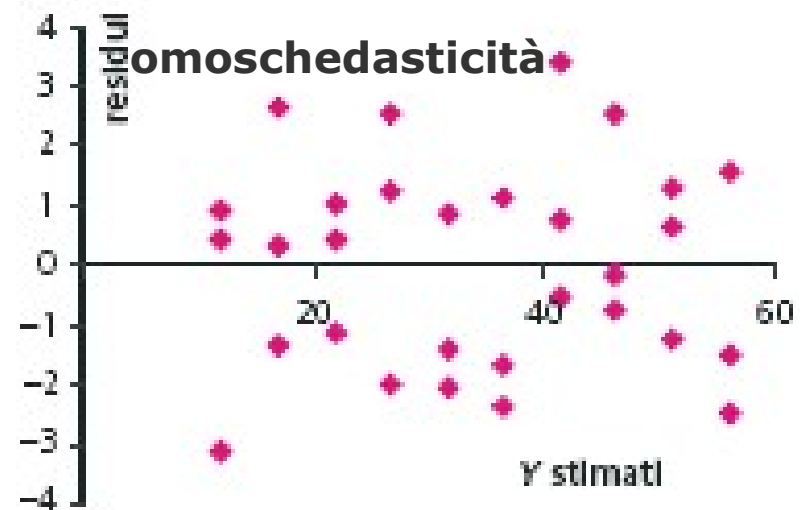
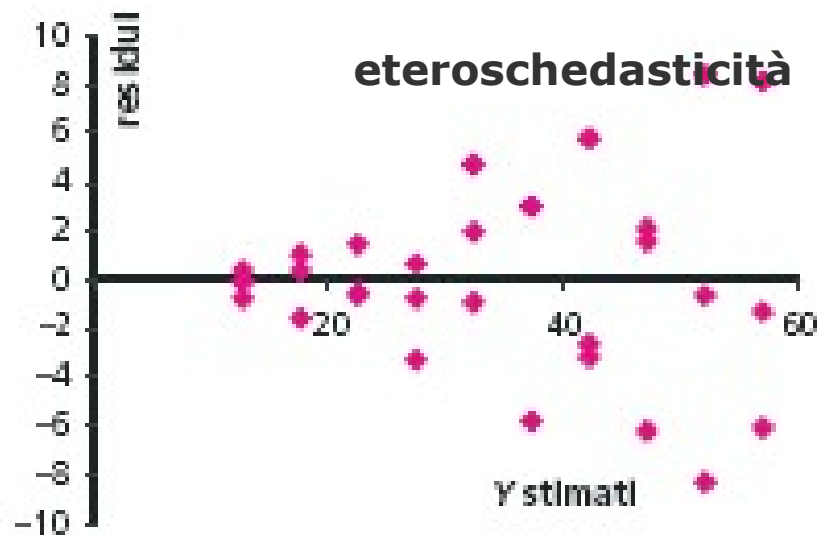


Analisi dei residui

Assunzione di omoschedasticità

Si assume che la varianza della y_i sia costante per ogni valore della variabile esplicativa. Altrimenti si parla di **eteroschedasticità**.

In presenza di omoschedasticità il grafico dei residui dovrebbe presentarsi approssimativamente come una nuvola di punti che si dispone in modo casuale all'interno di una fascia orizzontale.



Analisi dei residui

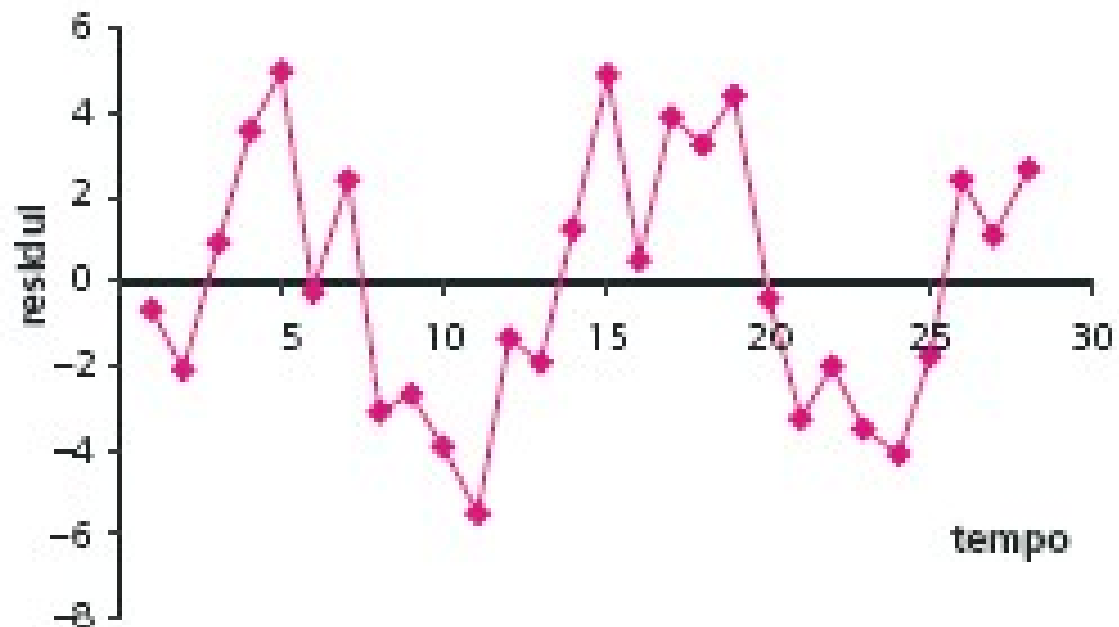
Assunzione di indipendenza

Se le osservazioni sono in una sequenza temporale, in genere gli errori non sono indipendenti. Disponendo in un grafico i residui secondo l'ordine temporale di osservazione, possiamo avere che:

I residui contigui tendono ad assumere stesso segno: **autocorrelazione positiva**. Sul grafico i residui mostrano comportamenti ciclici intorno allo zero.

I residui contigui tendono ad assumere segno opposto: **autocorrelazione negativa**. Sul grafico i residui tenderanno sistematicamente a cambiare segno.

**Esempio di
residui con
autocorrelazione
positiva**



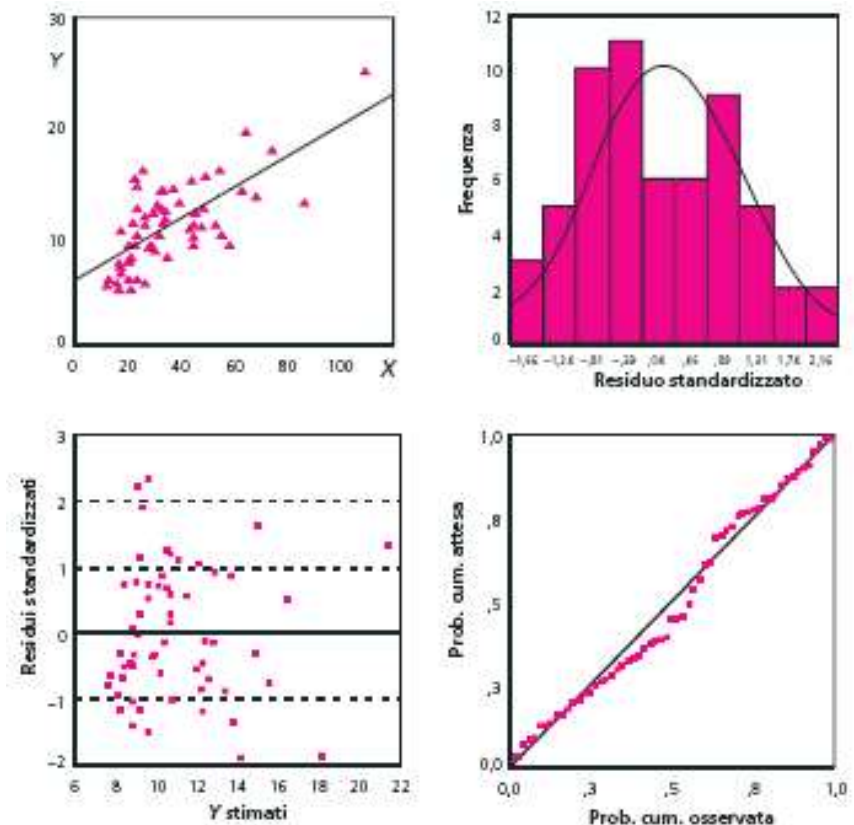
Analisi dei residui

Assunzione di Normalità

Un modo semplice di verificare tale assunzione è di considerare i **residui standardizzati**, che devono distribuirsi, al crescere di n , secondo una Normale standardizzata, cioè:

$$\hat{e}_i^* = \frac{\hat{e}_i}{s} \sim N(0,1)$$

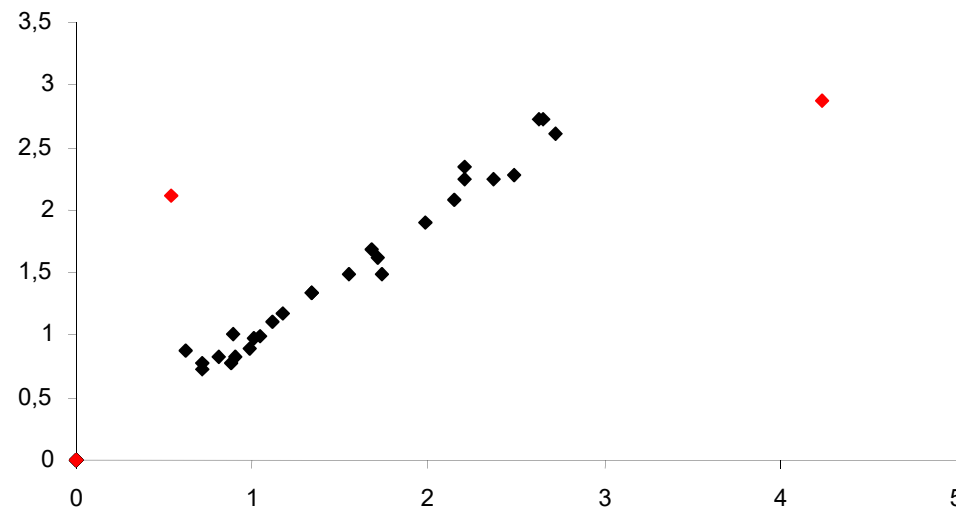
Nelle figure accanto sono presentati alcuni grafici per controllare la Normalità dei residui standardizzati: La retta di regressione; l'istogramma di frequenze; il **grafico dei residui standardizzati** (il 98% dei residui standardizzati devono oscillare tra -2 e +2); il **grafico di normalità P-P** in cui quanto più i punti si allineano lungo la bisettrice, tanto più è verificata l'ipotesi di normalità.



Valori anomali

Con il termine **valore anomalo** si indicano quelle osservazioni che per qualche ragione vengono considerate “diverse” dal resto dei dati.

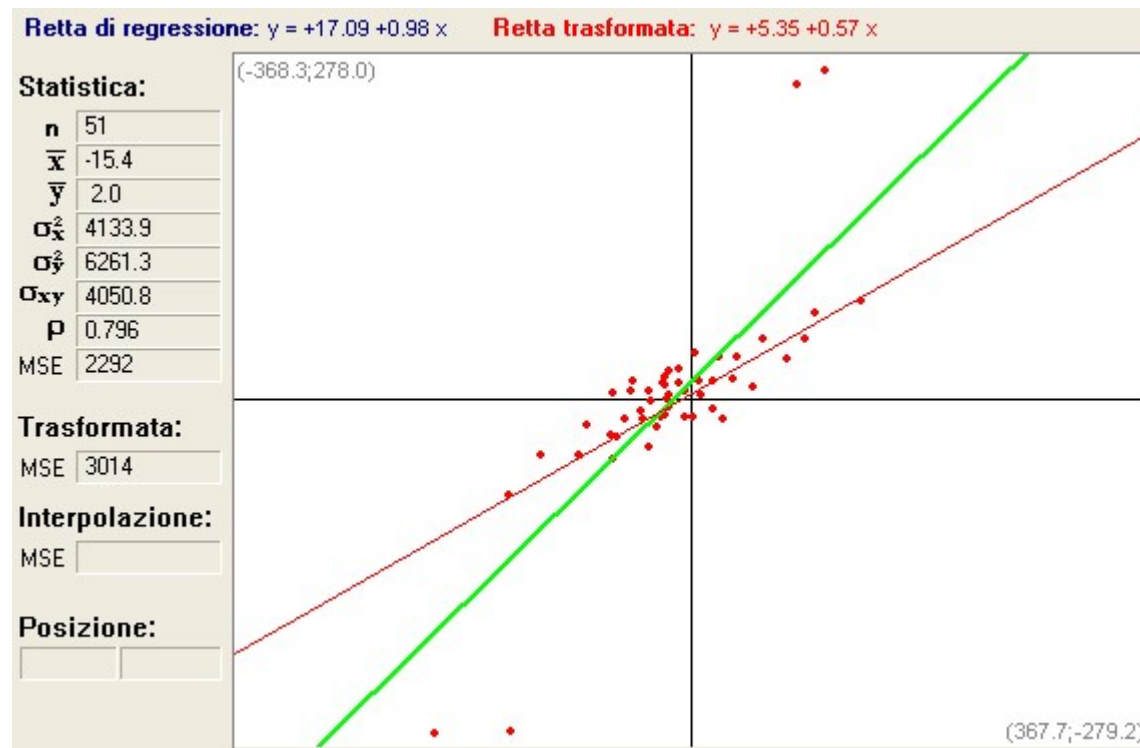
Esempio. Nel seguente grafico sono riportate le misurazioni di un indice di inquinamento effettuate nel 1994 e nel 1995 sulle acque di 30 laghi. Si può notare che la maggior parte dei punti segue un trend lineare. Tuttavia, i punti in rosso si discostano dal resto dei dati e potrebbero perciò essere considerati dei potenziali valori anomali.



Valori anomali

Un valore posizionato lontano dalla retta di regressione potrebbe essere considerato un valore anomalo e dovrebbe presentare in valore assoluto un residuo standardizzato molto elevato.

➡ La presenza di valori anomali può avere degli effetti rilevanti sulla regressione.

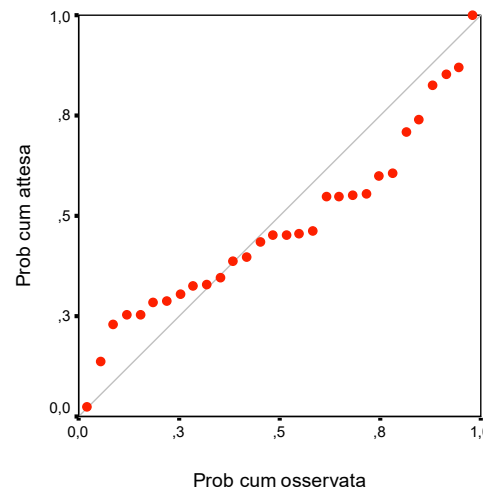
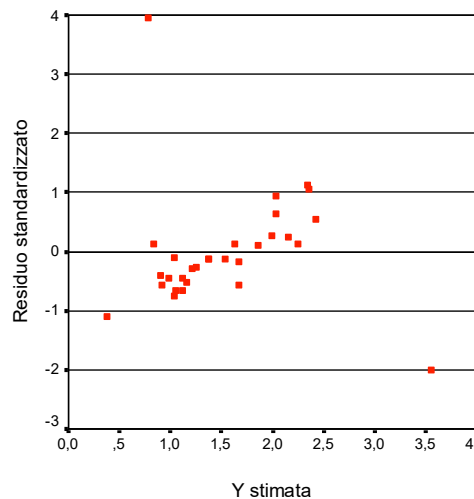


Sono mostrate due rette di regressione:
la prima, in **verde**, è calcolata su tutti i punti del piano (equazione in blu);
la seconda, in **rosso**, è calcolata escludendo i due punti più in alto e i due più in basso (equaz. in rosso).

Le due rette differiscono sia per il coefficiente angolare che per l'intercetta.

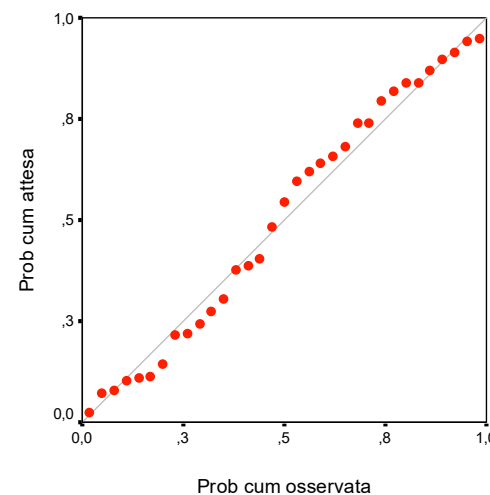
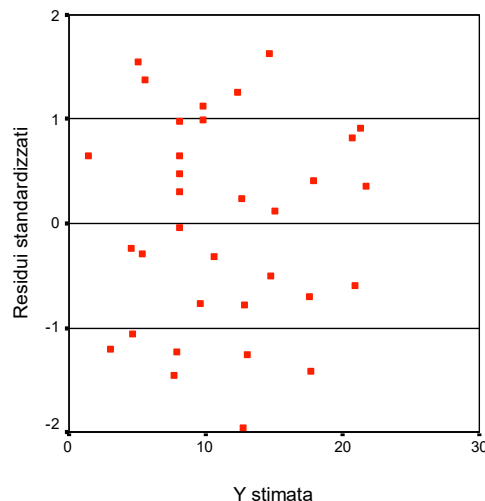
Valori anomali

Attraverso il grafico dei residui standardizzati e il grafico di normalità P-P è possibile identificare deviazioni dovute alla presenza di valori anomali.



Nelle due figure in alto sono mostrati il grafico dei residui standardizzati e il grafico di normalità P-P.

Due dei tre punti "sospetti" presentano nel grafico degli scostamenti anomali.



Eliminando questi due punti, si ottengono i due grafici successivi, che mostrano una maggiore aderenza dei dati alle assunzioni del modello.