UNIVERSITÁ DEGLI STUDI DI TERAMO

ECONOMIA

METODI STATISTICI PER L'ANALISI ECONOMICA E AZIENDALE

LA REGRESSIONE LINEARE SEMPLICE

FABRIZIO ANTOLINI
fantolini@unite.it

MODELLO DI REGRESSIONE LINEARE SEMPLICE

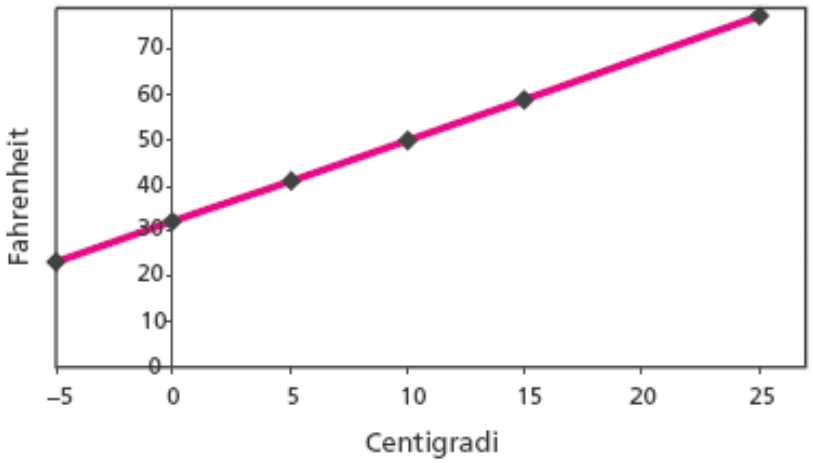
Date due variabili, X e Y, si è interessati a comprendere come la variabile Y (dipendente o risposta) sia influenzata dalla X (esplicativa o indipendente).

- Y è funzione di X se ad ogni valore di X corrisponde un solo valore di Y
- La relazione funzionale è lineare, se possiamo scrivere:

$$Y = \beta_0 + \beta_1 X$$

β_0 è l'intercetta

β_1 è il coefficiente angolare.



Una relazione statistica tra la X e Y può essere descritta da:

$$Y = f(X) + \varepsilon$$

- $f(X)$ definisce il contributo della X
- rappresenta il contributo di tutti i fattori non osservati. E' una componente stocastica ed è una variabile casuale.

IPOTESI DI BASE

Introducendo opportune assunzioni si ottiene il modello di regressione lineare semplice.

Assunzione 1:

$$Y_i = \beta_0 + \beta_i X_i + \varepsilon_i$$

per ogni osservazione $i=1,\dots n$

- implica che la funzione $f(X)$è lineare.

Assunzione 2:

Le ε_i sono variabili casuali indipendenti con $E(\varepsilon_i) = 0$ valore atteso, e varianza costante $V(\varepsilon_i) = \sigma^2$ per ogni osservazione $i=1,\dots n$

- implica che per ogni valore fissato di X, la Y possiede sempre lo stesso grado di variabilità (ipotesi di omoschedasticità)
- Inoltre, poiché la ε_i è una variabile casuale, anche Y è una variabile casuale; pertanto, le osservazioni Y_i sono realizzazioni di variabili casuali indipendenti, con valore atteso $E(Y_i|X = x_i) = \beta_0 + \beta_i X_i$ e con varianza $V(Y_i|X = x_i) = \sigma^2$

Assunzione 3:

I valori X_i della variabile esplicativa X sono noti e senza errore.

IPOTESI DI BASE

Assunzione 3:

I valori X_i della variabile esplicativa X sono noti e senza errore.

Assunzione 4:

Le variabili casuali ε_i hanno distribuzione Normale.

- Tenendo conto di questa ulteriore assunzione il modello di regressione lineare può essere definito nel seguente modo:

$$Y_i \sim N(\beta_0 + \beta_i X_{i,}, \sigma^2)$$

L'introduzione di tale assunzione ha alcune importanti implicazioni:

- Gli stimatori β_0 e β_1 si distribuiscono come una Normale bivariata
- Si ha che:

$$\frac{B_1 - \beta_1}{S(B_1)} \sim t_{n-2} \; e \; \frac{B_0 - \beta_0}{S(B_0)} \sim t_{n-2}$$

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

IPOTESI DI BASE

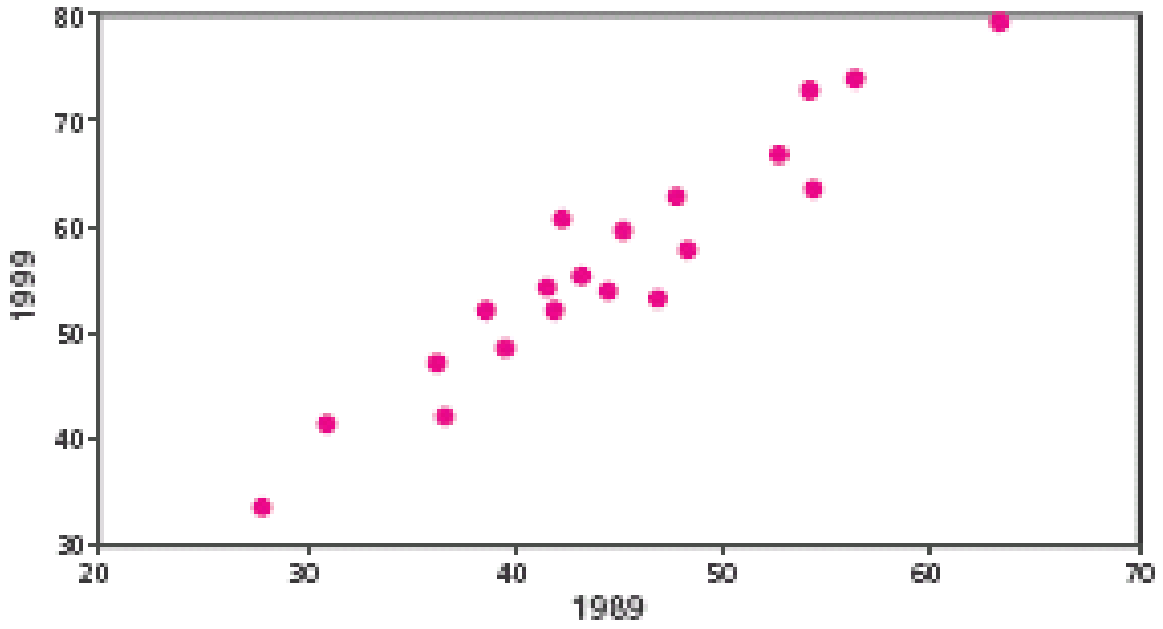
Esempio

Su un campione di 20 aree amministrative si osserva il reddito pro-capite nel 1989 X e 1999 Y.

Area	(X)	(Y)
1	47,8	63
2	27,9	33,4
3	36,6	42
4	54,2	72,8
5	41,9	52
6	44,4	54
7	54,3	63,4
8	42,3	60,7
9	48,2	58
10	41,5	54,4
11	43,2	55,5
12	56,3	74
13	63,3	79,2
14	46,8	53,1
15	45,2	59,6
16	38,7	52
17	36,3	47,2
18	39,5	48,7
19	30,9	41,4
20	52,6	66,9

Si ipotizza il seguente modello:

$$Y_i = \beta_0 + \beta_i X_i + \varepsilon_i$$



IPOTESI DI BASE

Si ottengono le seguente stime dei coefficienti del modello:

$$\widehat{\boldsymbol{\beta_0}} = 0{,}595 \text{ e } \widehat{\boldsymbol{\beta_1}} = 1{,}255$$

ossia la retta di regressione:

$$\boldsymbol{\hat{y}_i = 0{,}595 + 1{,}255 * x_i}$$

Il coefficiente di correlazione è:

$$\boldsymbol{\rho_{XY}} = 0{,}956$$

SQT = 24697,6 da cui:

$$\boldsymbol{R^2_{XY}} = (0{,}956)^2 = 0{,}914$$

ossia circa il 91% della variabilità totale di $\boldsymbol{Y}$ è spiegata dal modello di regressione.

Vedremo successivamente il significato associato ai valori calcolati.

.

STIMA PUNTUALE DEI COEFFICIENTI DI REGRESSIONE E INTERVALLI DI CONFIDENZA PER I PARAMETRI

Indicheremo con: $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} * x_i$

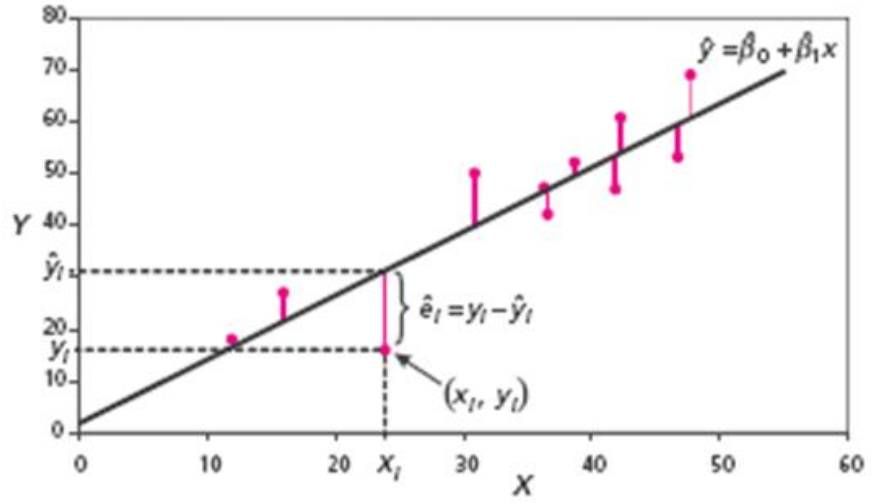
il valore di Y fornito dalla retta stimata dove $\widehat{\beta_0}$ e $\widehat{\beta_1}$ sono le stime dei coefficienti di regressione.

METODO DI STIMA DEI MINIMI QUADRATI

Consiste nel ricercare le stime di β_0 e β_1 , che rendono minima la funzione di perdita:

$$G(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 + \beta_i x_i)^2$$

Chiameremo *residuo i-esimo* la differenza tra il valore osservato y_i e quello fornito dalla retta stimata, $\widehat{y_i}$



STIMA PUNTUALE DEI COEFFICIENTI DI REGRESSIONE E INTERVALLI DI CONFIDENZA PER I PARAMETRI

PROCEDIMENTO:

- Porre uguali a zero le derivate prime rispetto ai parametri:

$$\begin{cases} \dfrac{\partial G(\beta_0, \beta_1)}{\partial \beta_0} = 0 \\ \dfrac{\partial G(\beta_0, \beta_1)}{\partial \beta_1} = 0 \end{cases}$$

- Risolvendo il sistema si ottengono le <u>stime dei minimi quadrati dei coefficienti di regressione</u>:

$$\widehat{\boldsymbol{\beta_1}} = \frac{\boldsymbol{\sigma_{XY}}}{\boldsymbol{\sigma_X^2}}$$

$$\widehat{\boldsymbol{\beta_0}} = \boldsymbol{\bar{y}} - \widehat{\boldsymbol{\beta_1}}\boldsymbol{\bar{x}}$$

STIMA PUNTUALE DEI COEFFICIENTI DI REGRESSIONE E INTERVALLI DI CONFIDENZA PER I PARAMETRI

PROPRIETÀ DEGLI STIMATORI DEI COEFFICIENTI

- $\boldsymbol{\beta_0}$ e $\boldsymbol{\beta_1}$ sono stimatori corretti
- Nella classe degli stimatori corretti di $\boldsymbol{\beta_0}$ e $\boldsymbol{\beta_1}$ che sono funzioni lineari delle $\boldsymbol{Y_i}$, gli stimatori dei minimi quadrati sono i più efficienti (Gauss-Markov).
- La varianza e covarianza degli stimatori dei minimi quadrati sono:

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

$$V(B_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

$$COV((B_0, B_1) = -\sigma^2 \frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Una misura della variabilità degli stimatori dei coefficienti di regressione e della risposta media è data dagli errori standard, ossia le radici quadrate delle varianze:

$$\sigma(B_0) = \sqrt{V(B_0)} \qquad \sigma(B_1) = \sqrt{V(B_1)}$$

STIMA PUNTUALE DEI COEFFICIENTI DI REGRESSIONE E INTERVALLI DI CONFIDENZA PER I PARAMETRI

PROPRIETÀ DEGLI STIMATORI DEI COEFFICIENTI

Gli errori standard dipendono dalla quantità ignota:

$$\boldsymbol{\sigma^2 = V(B_0) = V(\varepsilon_i)}$$

pertanto, la si sostituisce con una sua stima $\boldsymbol{s^2}$ ottenendo gli stimatori:

$$\boldsymbol{s(B_0)}$$

$$\boldsymbol{s(B_1)}$$

$$\boldsymbol{s(\widehat{Y_i})}$$

Lo stimatore che si utilizza per ottenere la stima della varianza è dato da:

$$\boldsymbol{s^2 = \frac{\sum_{i=1}^{n} \widehat{e_i}^2}{n-2}}$$

N.B: La radice quadrata di $\boldsymbol{s^2}$ è una misura della variabilità degli scostamenti dei valori osservati da quelli previsti dal modello e per tale ragione viene usualmente chiamato errore standard di regressione.

INTERVALLI DI CONFIDENZA PER I PARAMETRI

Gli intervalli di confidenza per i parametri $\boldsymbol{\beta_0}$ e $\boldsymbol{\beta_1}$ a un livello di confidenza $\boldsymbol{1\text{-}\alpha}$ sono dati da:

$$\boldsymbol{\beta_0} \mp \boldsymbol{t_{\alpha/2} s(B_0)}$$

$$\boldsymbol{\beta_1} \mp \boldsymbol{t_{\alpha/2} s(B_1)}$$

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

$$\boldsymbol{t = \frac{B_1 - \beta_1}{S(B_1)} \sim t_{n-2}}$$

In corrispondenza del sistema d'ipotesi:

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

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

INTERVALLI DI CONFIDENZA PER I PARAMETRI

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.

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

$$t = \frac{B_0 - \beta_0}{S(B_0)} \sim t_{n-2}$$

INTERVALLI DI CONFIDENZA PER I PARAMETRI

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_{n-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.

INTERVALLI DI CONFIDENZA PER I PARAMETRI

Esempio

Si vuole verificare il sistema d'ipotesi:

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

La statistica test, sotto l'ipotesi nulla, è:

$$t = \frac{1,255}{0,091} = 13,79$$

A un livello di significatività α = 0,999 corrisponde un valore della t-Student con 18 gradi di libertà pari a:

$$t_{0,005} = 2,8784$$

Pertanto, t= 13,79 > 2,8784 quindi si rifiuta la H_0

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

INTERVALLI DI CONFIDENZA PER I PARAMETRI

Le stime dei minimi quadrati possiedono un'importante proprietà, nota come decomposizione della varianza totale:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{e}_i)^2$$

Somma totale dei quadrati (SQT):

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

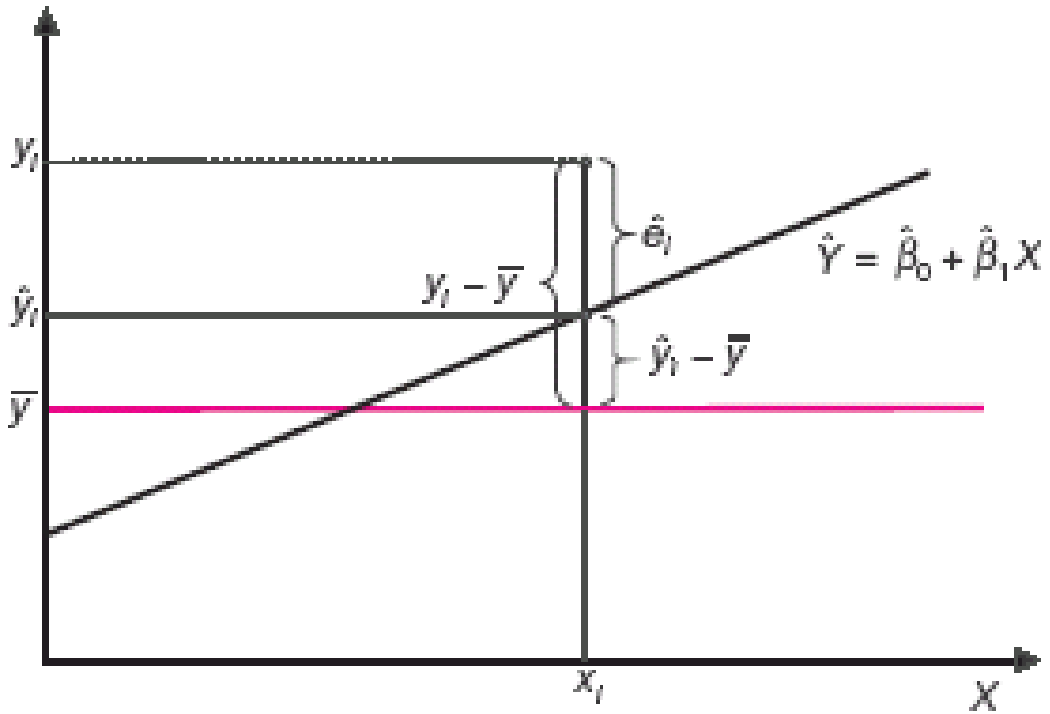
Somma dei quadrati della regressione (SQR):

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

Somma dei quadrati degli errori (SQE):

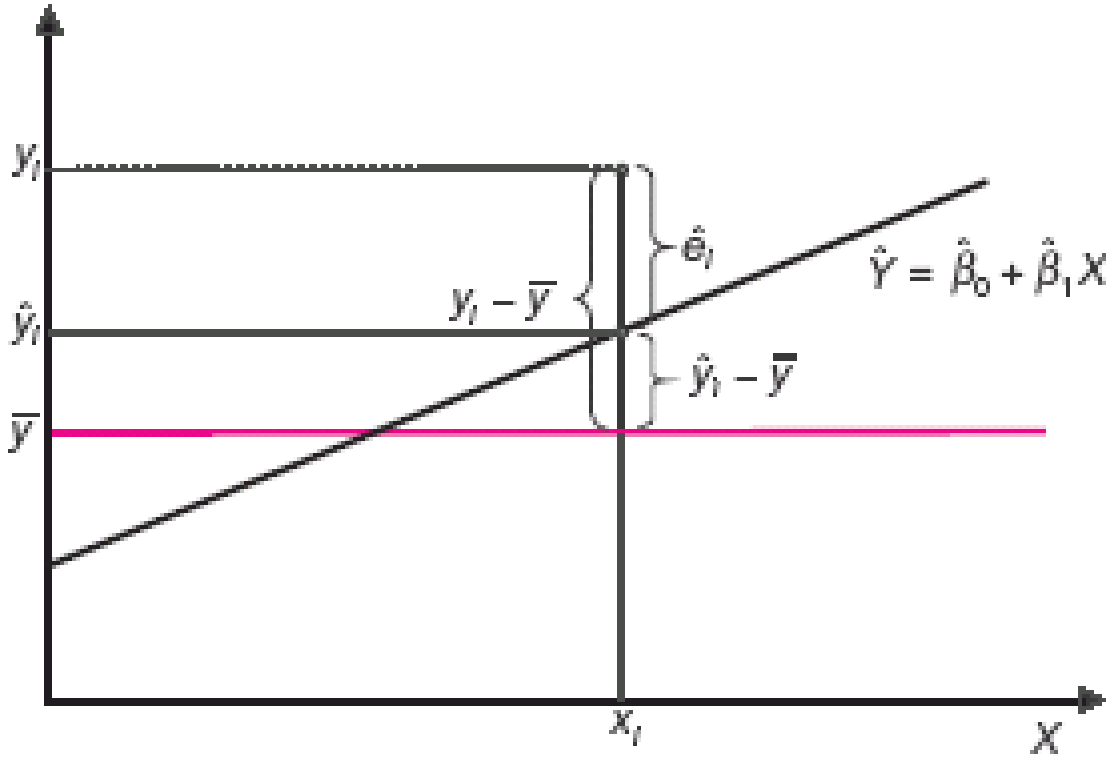
$$SQE = \sum_{i=1}^{n}(\hat{e}_i)^2$$

INTERVALLI DI CONFIDENZA PER I PARAMETRI



- **SQR = 0;** SQE = SQT e i valori stimati sono tutti uguali alla media campionaria
- **SQR = SQT;** SQE = 0 e tutti i valori stimati sono uguali a quelli osservati.

INTERVALLI DI CONFIDENZA PER I PARAMETRI



- **SQR = 0;** SQE = SQT e i valori stimati sono tutti uguali alla media campionaria
- **SQR = SQT;** SQE = 0 e tutti i valori stimati sono uguali a quelli osservati.

COEFFICIENTE DI DETERMINAZIONE

Dalla relazione $\boldsymbol{SQT = SQR + SQE}$ si può definire un indice che misura la bontà di adattamento della retta di regressione:

$$\boldsymbol{R^2_{XY} = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}}$$

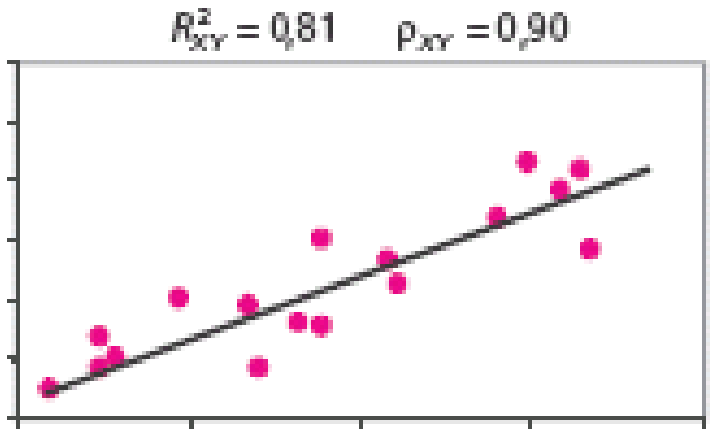
Il rapporto $\boldsymbol{R^2_{XY}}$ è detto coefficiente di determinazione e indica la proporzione di variabilità di $\boldsymbol{Y}$ spiegata dalla variabile esplicativa $\boldsymbol{X}$, attraverso il modello di regressione.

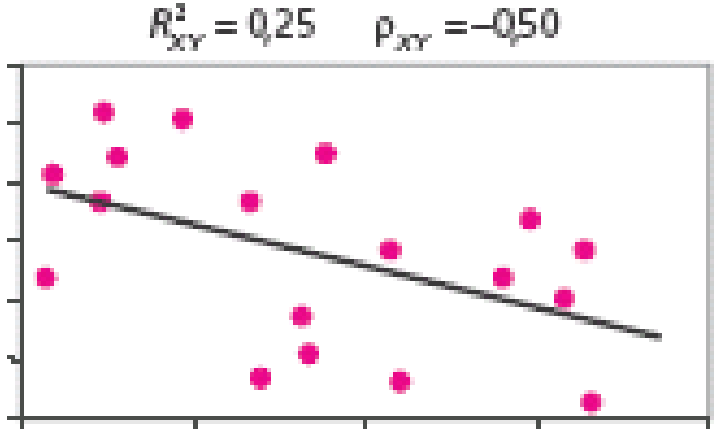
N.B: si può dimostrare che il coefficiente di determinazione corrisponde al quadrato del coefficiente di correlazione lineare:

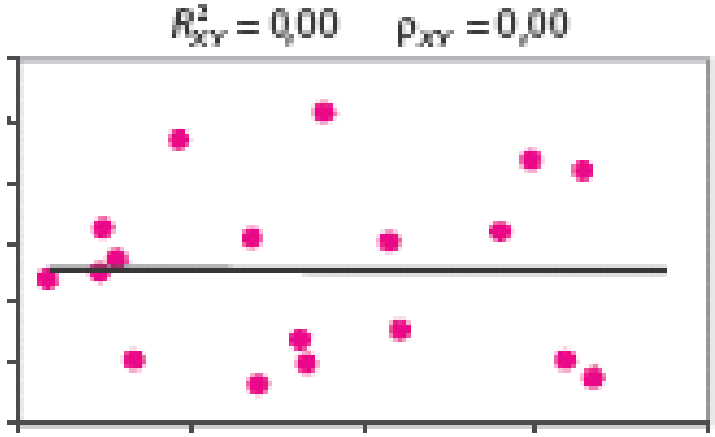
$$\boldsymbol{R^2_{XY} = (\rho_{XY})^2 = \left(\frac{\sigma_{xy}}{\sigma_X \sigma_Y}\right)^2}$$

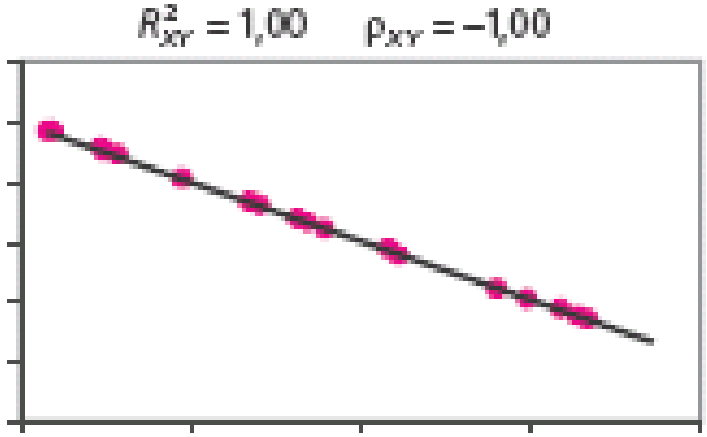
COEFFICIENTE DI DETERMINAZIONE

Esempio









ANALISI DEI RESIDUI

Quando si realizza un modello di regressione lineare, una delle prime cose da fare è l'analisi dei residui. La retta di regressione è infatti una semplificazione della realtà e non coglie tutta la variabilità presente in un insieme di dati.

La parte di variabilità che non è spiegata dal modello costituisce proprio il residuo della regressione.

Affinché il modello di regressione riesca ad avere un buon potere predittivo, questo errore deve essere una variazione imprevedibile nella variabile risposta.

I valori residui in un'analisi di regressione rappresentato proprio la parte di errore di previsione del modello di regressione. I residui, detti anche scarti, rappresentano infatti le differenze tra i valori osservati nel dataset e i valori stimati calcolati con l'equazione di regressione.

In altre parole, i residui indicano la variabilità dei dati attorno alla retta di regressione.

Come abbiamo visto, le proprietà degli stimatori dei parametri del modello richiedono alcune assunzioni, ed è quindi verificare la validità di tali assunzioni.

ANALISI DEI RESIDUI

Un tecnica di verifica si basa sull'analisi dei residui.

- I residui hanno una distribuzione normale?
- Le variabili indipendenti sono incorrelate con l'errore?
- La varianza dei residui è omogenea?
- La distribuzione dei residui è lineare?
- Ci sono degli *outliers* che influenzano la pendenza della retta?
- I residui sono tra loro correlati?

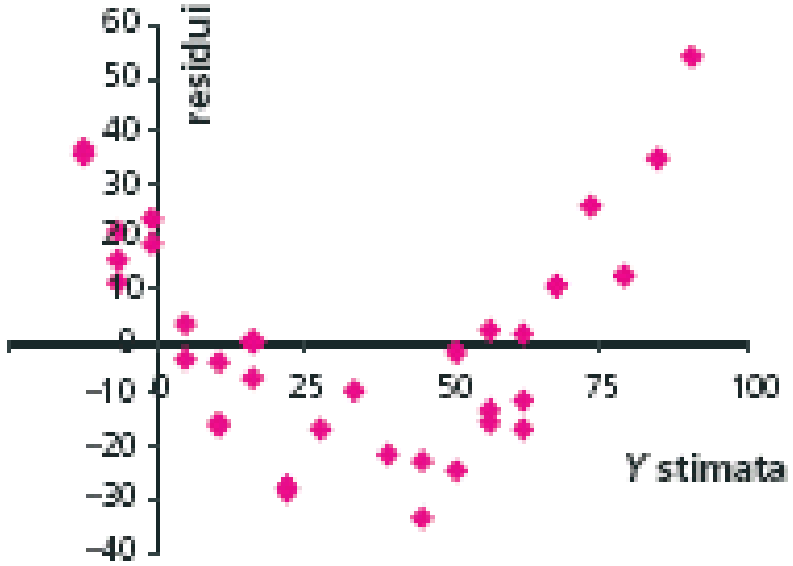
L'idea di base è che se le assunzioni sono vere, ossia se il modello è ben specificato, allora i residui $\widehat{e_i}$ rifletteranno le proprietà attribuite ai termini di errore ε_i.

ANALISI DEI RESIDUI

METODO GRAFICO: GRAFICO DEI RESIDUI

Si assume che la funzione di regressione sia di tipo lineare; ciò può essere verificato analizzando il grafico dei residui.

Esempio



Secondo l'ipotesi di linearità, i dati devono infatti distribuirsi in modo casuale intorno allo 0. Nel seguente grafico è evidente che la relazione non è di tipo lineare. In altre parole, in questo caso sapendo quale è il valore stimato si può predire quale sarà il valore del residuo.

ANALISI DEI RESIDUI

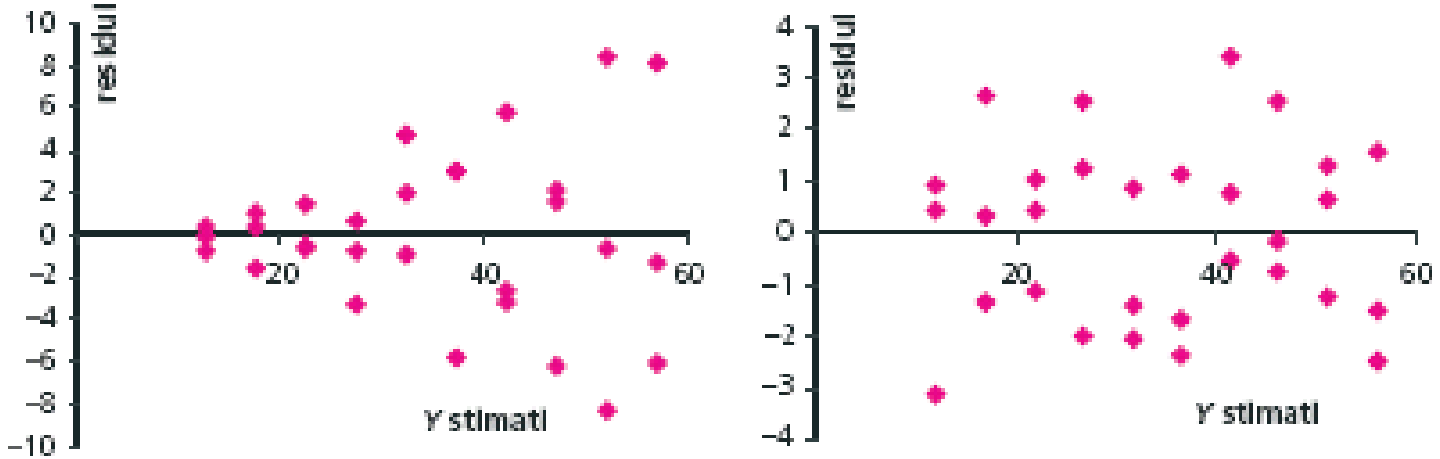
ASSUNZIONE DI OMOSCHEDACITA'

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

Per verificare l'ipotesi di omogeneità delle varianze dei residui, è necessario creare un grafico a dispersione. I valori stimati della y si riportano sull'asse orizzontale delle x. Sull'asse verticale, invece, si indicano i valori dei residui.

N.B: 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.

Esempio



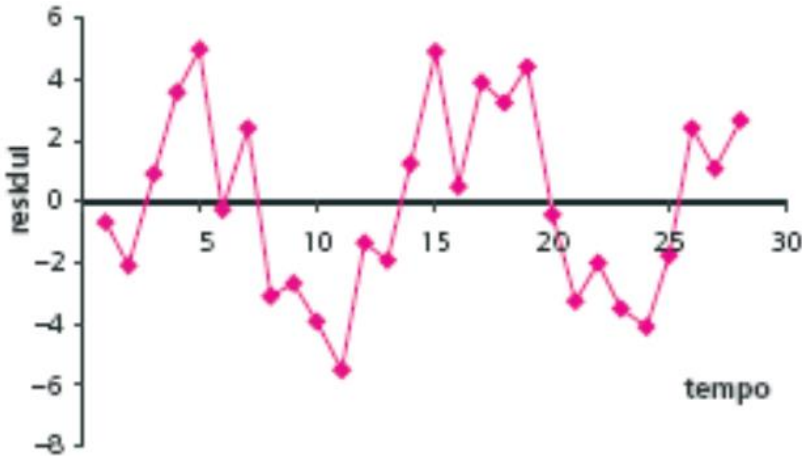
ANALISI DEI RESIDUI

ASSUNZIONE DI INDIPENDENZA

Se la variabile esplicativa è correlata con il termine dell'errore, tale variabile esplicativa non può essere utilizzata per predire quale sarà l'errore del modello di regressione poiché la componente di errore di un modello di previsione deve essere imprevedibile.

Per verificare questa assunzione è possibile costruire un grafico di dispersione nel quale sull'asse orizzontale vanno posti i valori della variabile esplicativa, mentre sull'asse verticale i valori dei residui.

Esempio



Osservando, ad esempio, il grafico che studia la relazione tra il tempo (x) ed i residui di un modello di regressione che aveva come obiettivo quello di predire la distanza percorsa da un auto, l'ipotesi di indipendenza non è confermata poiché è individuabile una relazione tra le due variabili.

ANALISI DEI RESIDUI

ASSUNZIONE DI NORMALITA'

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:

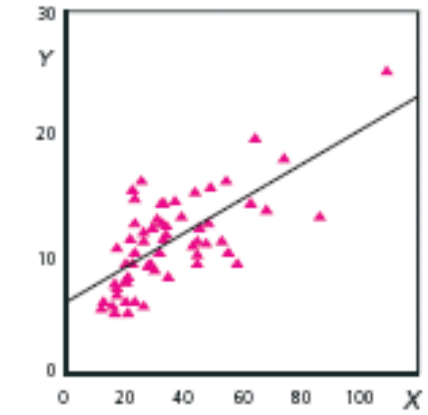
$$\hat{e}_i^* = \frac{\hat{e}_i}{s} \sim N(0,1)$$

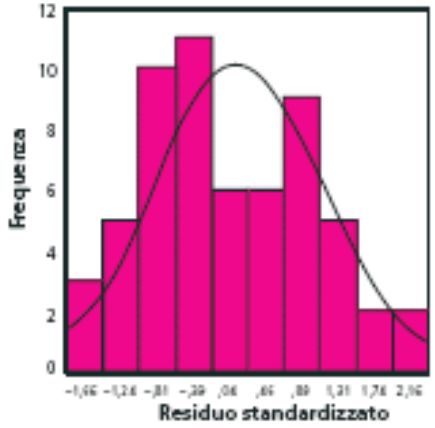
Esistono più grafici per controllare la Normalità dei residui standardizzati:

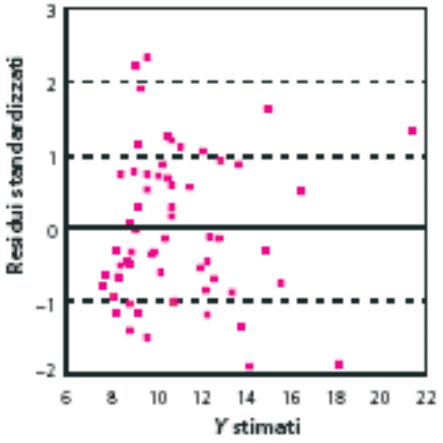
- La retta di regressione
- L'istogramma delle frequenze
- Il grafico dei residui standardizzati (il 98% dei residui standardizzati deve oscillare tra -2 e +2)
- Il grafico di normalità q-q plot in cui quanto più i punti si allineano lungo la bisettrice, tanto più è verificata l'ipotesi di normalità.

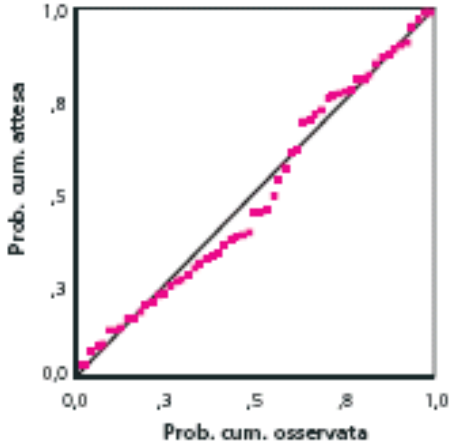
ANALISI DEI RESIDUI

ASSUNZIONE DI NORMALITA'









In particolare, il q-q plot (o grafico dei quantili) riporta sull'asse delle ascisse i quantili teorici di una distribuzione Normale mentre i quantili dei residui standardizzati sono invece riportati sull'asse verticale.

- L'idea è che se i residui avessero una distribuzione normale, i loro quantili dovrebbero coincidere con quelli della distribuzione Normale. A livello visivo, questo significa che i punti dovrebbero disporsi lungo la bisettrice.

N.B: Nella pratica, non capita quasi mai che i punti si dispongano esattamente lungo la bisettrice. Per poter dire che gli errori hanno una distribuzione normale ci si accontenta quindi che i punti siano vicino alla linea presente nel grafico.

ASSUNZIONE DI INCORRELAZIONE DEI RESIDUI

L'ultima ipotesi sui residui richiede di verificare che i residui non siano tra loro autocorrelati.

Intuitivamente l'autocorrelazione dei residui si verifica, ad esempio, quando si hanno delle misure ripetute nel tempo sugli stessi individui (oppure su identiche aree). In queste situazioni, i modelli di regressione lineare non sono adatti a descrivere i dati ed è preferibile utilizzare dei modelli basati sulle analisi longitudinali dei dati.

Senza approfondire la costruzione del test, per poter capire se i residui siano tra loro autocorrelati si ricorre principalmente al coefficiente *d di Durbin-Watson*.

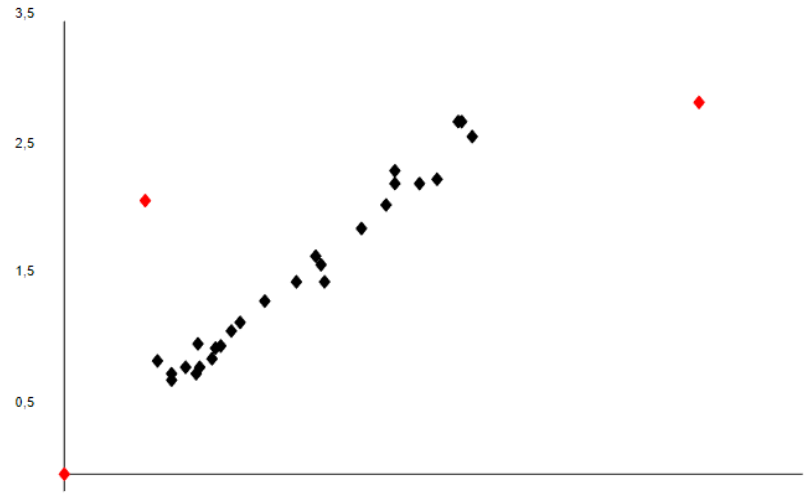
- Questo indice è compreso tra 0 e 4
- Valori vicini a 2 indicano che non c'è autocorrelazione
- Valori piccoli indicano, invece, che residui successivi tra loro sono, in media, vicini in valore l'uno all'altro, o correlati positivamente.
- Valori grandi indicano che residui successivi tra loro sono, in media, molto differenti in valore l'uno dall'altro, o correlati negativamente.

ANALISI DEI RESIDUI

VALORI ANOMALI

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

Esempio



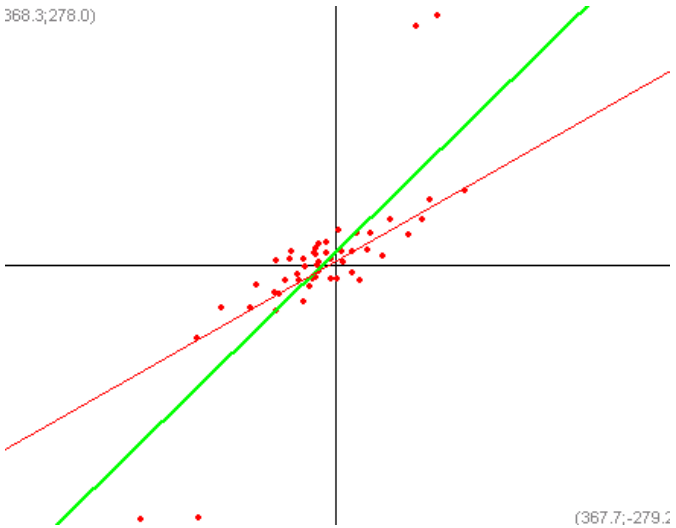
Nel seguente grafico sono riportate le misurazioni di un indice di inquinamento effettuate nel 1994 e nel 1995 sulle acque di 30 laghi. Si può notare che la maggior parte dei punti segue un trend lineare. Tuttavia, i punti in rosso si discostano dal resto dei dati e potrebbero perciò essere considerati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

- 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.

ANALISI DEI RESIDUI

VALORI ANOMALI

Esempio



Sono mostrate due rette di regressione:

- la prima, in verde, è calcolata su tutti i punti del piano
- la seconda, in rosso, è calcolata escludendo i due punti più in alto e i due più in basso.

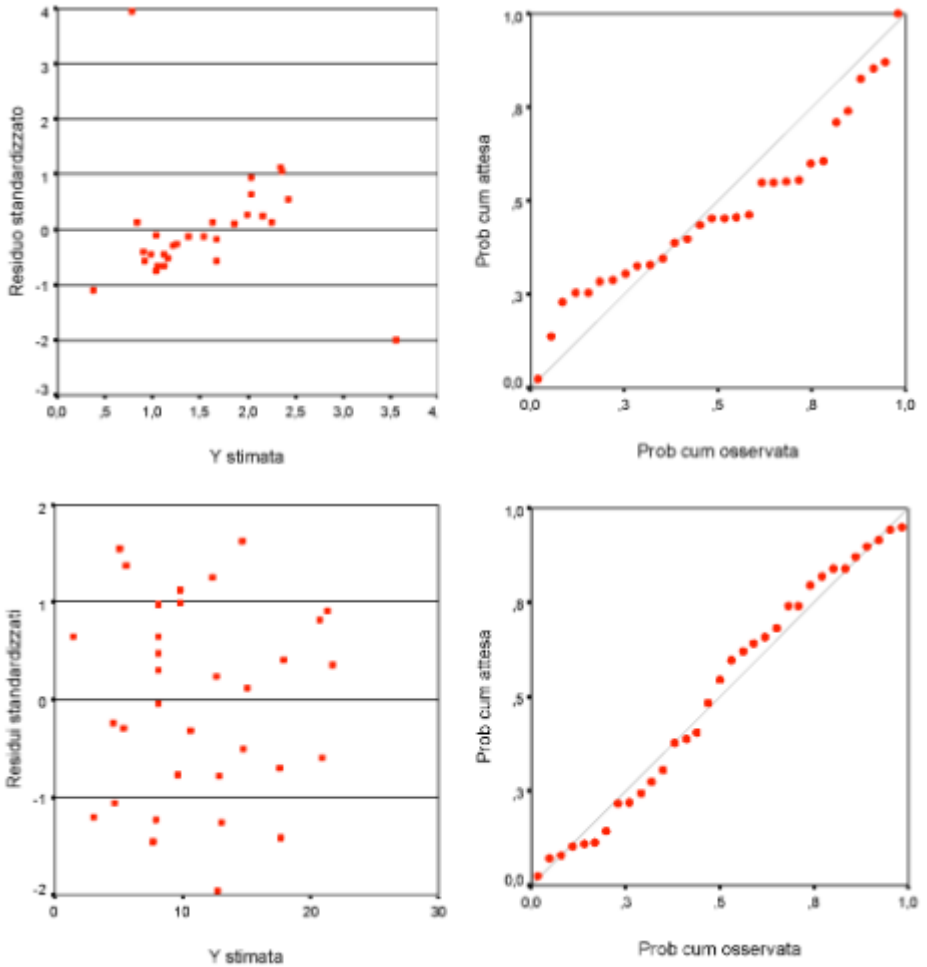
<u>Le due rette differiscono sia per il coefficiente angolare che per l'intercetta.</u>

N.B: Attraverso il grafico dei residui standardizzati e il grafico di normalità q-q plot è possibile identificare deviazioni dovute alla presenza di valori anomali.

ANALISI DEI RESIDUI

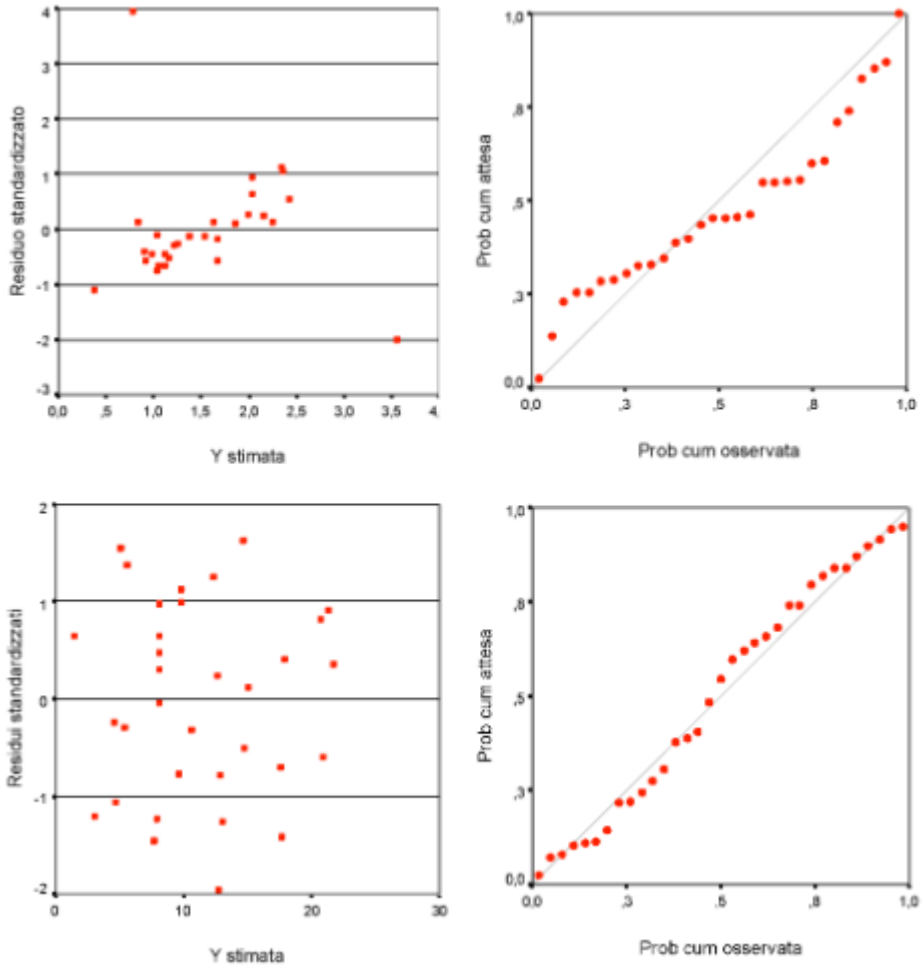
VALORI ANOMALI

Esempio



VALORI ANOMALI

Esempio



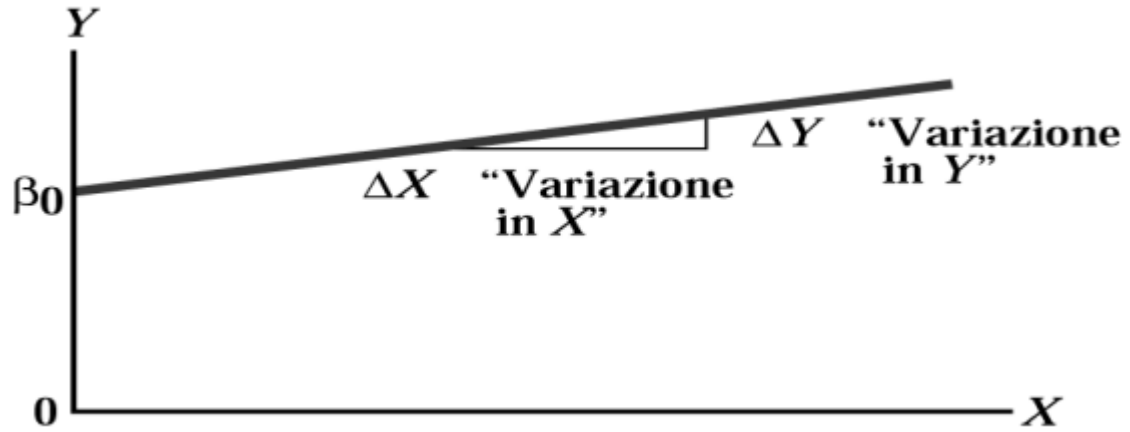
Nelle due figure in alto sono mostrati il grafico dei residui standardizzati e il grafico di normalità q-q plot:

- 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.

INTERPRETARE I RISULTATI DI UNA REGRESSIONE SEMPLICE

Come abbiamo visto, la regressione semplice genera un'equazione per descrivere la relazione statistica tra una variabile predittiva (***X***) e la variabile risposta (***Y***).

- Il coefficiente di regressione rappresenta la variazione media nella variabile di risposta per un'unità di variazione nella variabile predittiva
- l'obiettivo è quello di costruire un modello attraverso cui prevedere i valori di una variabile **dipendente o risposta** a partire dai valori di una variabile **indipendente o esplicativa**



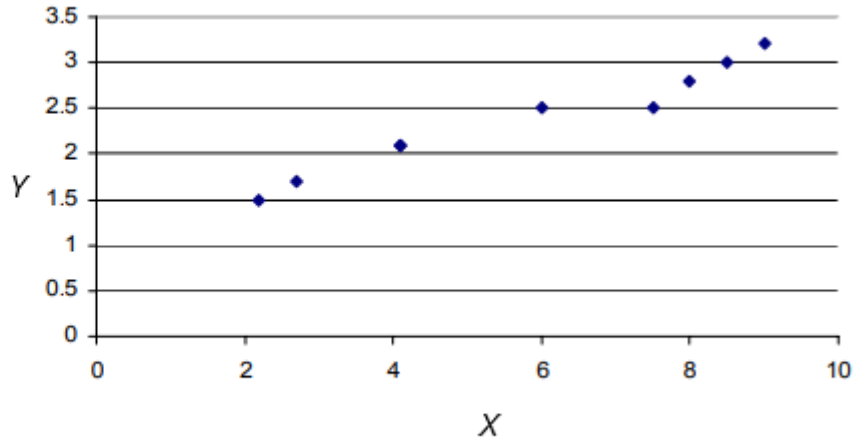
- L'inclinazione $\boldsymbol{\beta_1}$ indica come varia ***Y*** in corrispondenza di una variazione unitaria di ***X***.
- L'intercetta $\boldsymbol{\beta_0}$ corrisponde al valore medio di ***Y*** quando ***X*** è uguale a 0.
- Il segno di $\boldsymbol{\beta_1}$ indica se la relazione lineare è positiva o negativa.

INTERPRETARE I RISULTATI DI UNA REGRESSIONE SEMPLICE

Esempio

Un produttore desidera ottenere una misura della qualità di un prodotto ma la procedura è troppo costosa. Decide allora di stimare questa misura (***Y***) a partire dall'osservazione di un'altra misura (***X***) più semplice meno costosa da ottenere.

unità prodotto	*Y*	*X*
1	4.1	2.1
2	2.2	1.5
3	2.7	1.7
4	6	2.5
5	8.5	3
6	4.1	2.1
7	9	3.2
8	8	2.8
9	7.5	2.5



Sotto le ipotesi viste in precedenza, sappiamo che i parametri del modello possono essere stimati ricorrendo ai dati del campione. Ricordiamo che la regressione ha come obiettivo quello di individuare la retta che meglio si adatti ai dati.

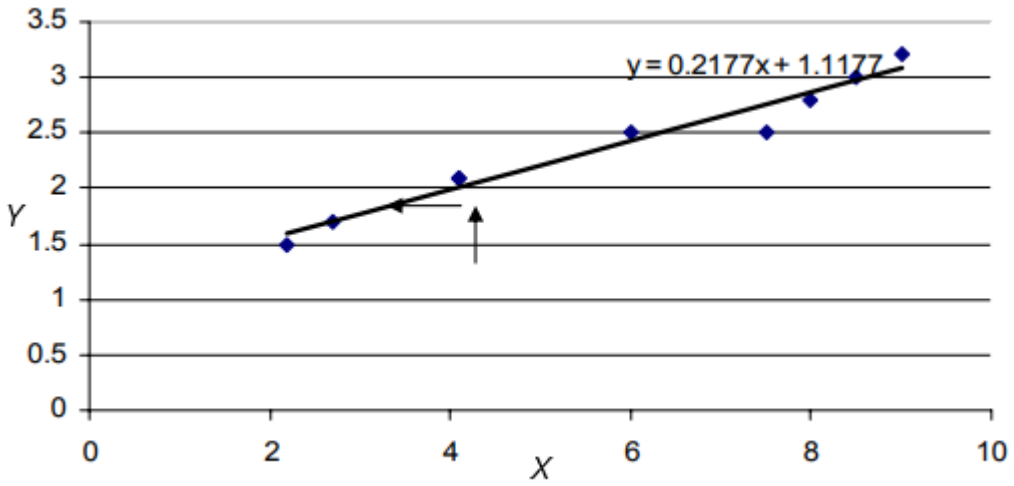
Il criterio più semplice è quello di valutare le differenze tra i valori osservati (Y_i) e i valori previsti ($\hat{Y}_i$)

INTERPRETARE I RISULTATI DI UNA REGRESSIONE SEMPLICE

Esempio

Utilizzando il **Metodo dei Minimi Quadrati** riusciamo a trovare i valori $\mathbf{b_0}$ e $\mathbf{b_1}$ che rendono minima la somma dei quadrati delle differenze tra i valori osservati $\boldsymbol{Y_i}$ e i valori stimati $\hat{\boldsymbol{Y}}_i$. Chiaramente i valori $\mathbf{b_0}$ e $\mathbf{b_1}$ sono chiamati coefficienti di regressione.

Nell'esempio precedente, applicando il metodo dei minimi quadrati si ottiene la seguente retta di regressione:



Risulta:
$\mathbf{b_1}$ = 0,2177
$\mathbf{b_0}$ = 1,1177

Perciò se aumenta di un'unità il valore di ***X***, il valore previsto di ***Y*** subisce un incremento di 0,2177.
Se ***X*** assume valore 0, il valore previsto per ***Y*** è pari a 1,1177.

Tramite l'equazione $\boldsymbol{Y} = 1{,}1177 + 0{,}2177\boldsymbol{X}$ è possibile prevedere i valori di ***Y*** in funzione di quelli osservati di ***X***.

Se ad esempio osservassimo un valore di ***X*** pari a 4,5 il valore stimato di ***Y*** sarebbe 2,1.