CAPITOLO 6 – Misura delle relazioni tra variabili per le decisioni aziendali
Paragrafo 6.2 – Analisi e impiego della correlazione semplice e del modello di regressione lineare semplice

6

Paragrafo 6.2
Analisi e impiego della correlazione semplice e del modello di regressione lineare semplice

Argomenti

- Analisi e misura della correlazione semplice

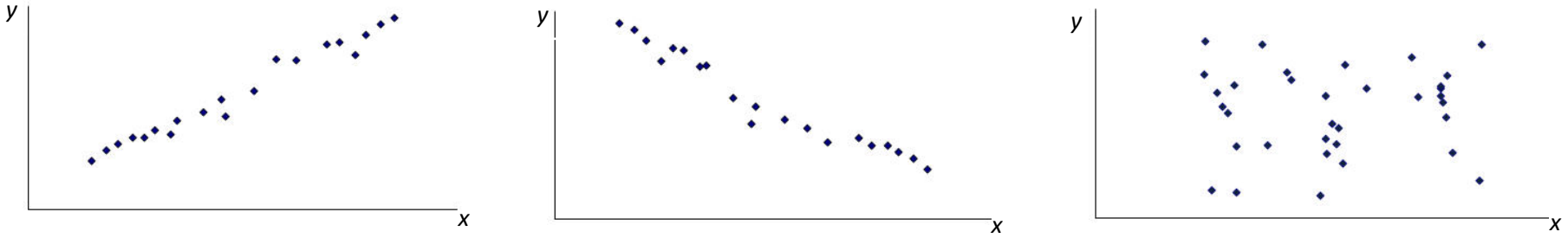
- Modello di regressione lineare semplice

- Analisi dei residui

- Casi di relazioni non lineari e presenza di dati anomali

L’analisi e la misura della correlazione

Come è noto, l’analisi della correlazione tra due variabili quantitative si conduce inizialmente attraverso la predisposizione del diagramma di dispersione (*scatter plot* o *scattergram*): la disposizione dei punti corrispondenti alle n coppie di valori di Y e di X fornisce una prima indicazione sull’eventuale legame



Le tre rappresentazioni riportate in Figura mettono in evidenza, rispettivamente, una correlazione positiva, una correlazione negativa e una mancanza di correlazione

Per avere una misura dell'intensità della relazione si calcola l'**indice** parametrico di **correlazione di Pearson** ρ_{xy}

L'indice viene utilizzato quando le variabili sono continue, cioè misurate su una scala a intervalli o razionale, sono distribuite normalmente e la loro relazione deve essere di tipo lineare:

$$corr(x,y) = \rho_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sigma_x^2 \cdot \sigma_y^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i y_i) - \bar{x}\bar{y}}{\sigma_x \cdot \sigma_y} = \frac{\mathrm{cov}(x,y)}{\sigma_x \cdot \sigma_y}$$

I valori che il coefficiente può assumere sono compresi tra -1 e +1, i valori estremi esprimono una relazione perfetta tra le variabili, mentre in assenza di una relazione il valore del coefficiente è 0. Il segno positivo o negativo corrisponde a quello della covarianza

Quando l'indice di correlazione viene calcolato su dati provenienti da un campione, piuttosto che da un'intera popolazione obiettivo, occorre fare inferenza sull'indice procedendo con un **test di ipotesi** sulla **significatività** del suo valore

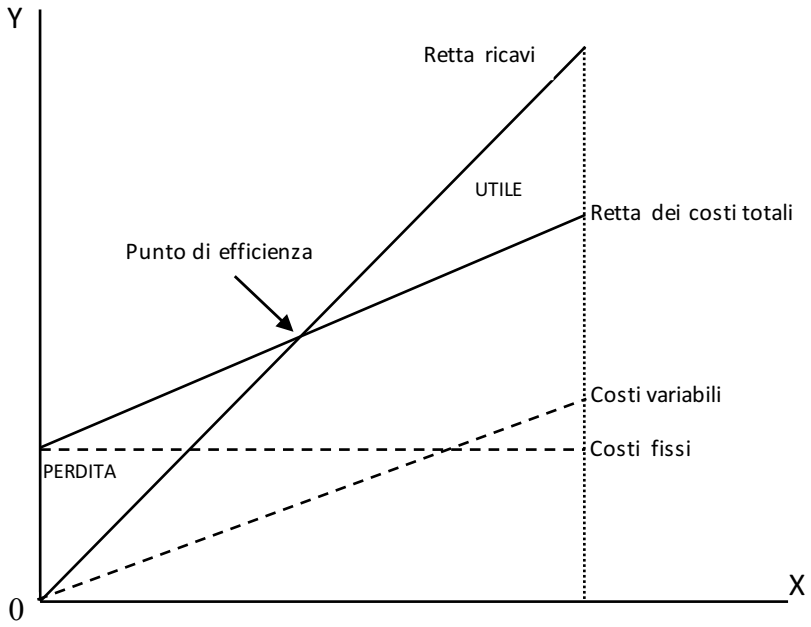
La statistica test per la verifica dell'ipotesi nulla $H_0 : \rho = 0$ (contro l'ipotesi $H_1 : \rho \neq 0$) è:

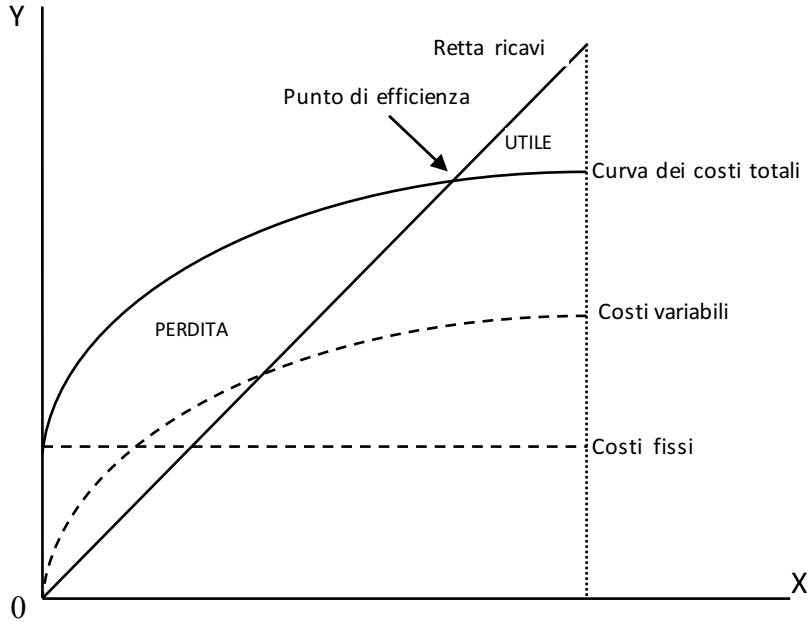
$$t = \frac{r - 0}{\sqrt{(1 - r^2)/(n-2)}} = r\sqrt{\frac{n-2}{1-r^2}}$$

dove *r* è il valore dell'indice calcolato sui dati campionari e $\sqrt{(1-r^2)/(n-2)}$ è l'errore standard di *r*. Se la variabili X e Y sono casuali e distribuite normalmente, allora sotto l'ipotesi che sia vera, la statistica test *t* è distribuita come una variabile *t* di Student con n-2 gradi di libertà. Occorre prestare attenzione alla presenza di valori anomali nei dati che potrebbero inficiare la natura delle distribuzioni e conseguentemente la validità del test

Esempio: studio della relazione tra costi e volume di produzione

Affrontiamo l'analisi dei costi di produzione in relazione al volume delle vendite, che in termini aziendali è detta *break-even point analysis*. Ricordiamo che in genere essa si basa sulla costruzione di un grafico come quello riportato nella Figura qui sotto





Il grafico mette in evidenza la relazione esistente tra i costi (distinti in costanti e variabili) e l'ammontare della produzione, nonché l'ammontare dei ricavi in relazione alle unità prodotte e vendute, dati i prezzi di vendita

Il punto di efficienza si ha per un volume di vendite di unità, dove i ricavi sono uguali ai costi; sotto tale punto le vendite di una quantità inferiore dei prodotti da luogo a una perdita e viceversa sopra il punto l'azienda ricava un utile

Nella costruzione del diagramma si presuppone che i costi siano sotto controllo e che non si verifichino sensibili fluttuazioni dei prezzi dei materiali, della manodopera e dei prodotti finiti, e che i costi variabili siano direttamente proporzionali, in modo deterministico, in relazione al volume di produzione

Le ipotesi che si possono fare a priori sulle relazioni tra i vari costi e il volume della produzione devono essere verificate empiricamente attraverso una misura della correlazione esistente tra le variabili, sia, soprattutto, attraverso la misura della dipendenza dei costi di produzione dal volume delle vendite (con la stima dei parametri di un modello di regressione)

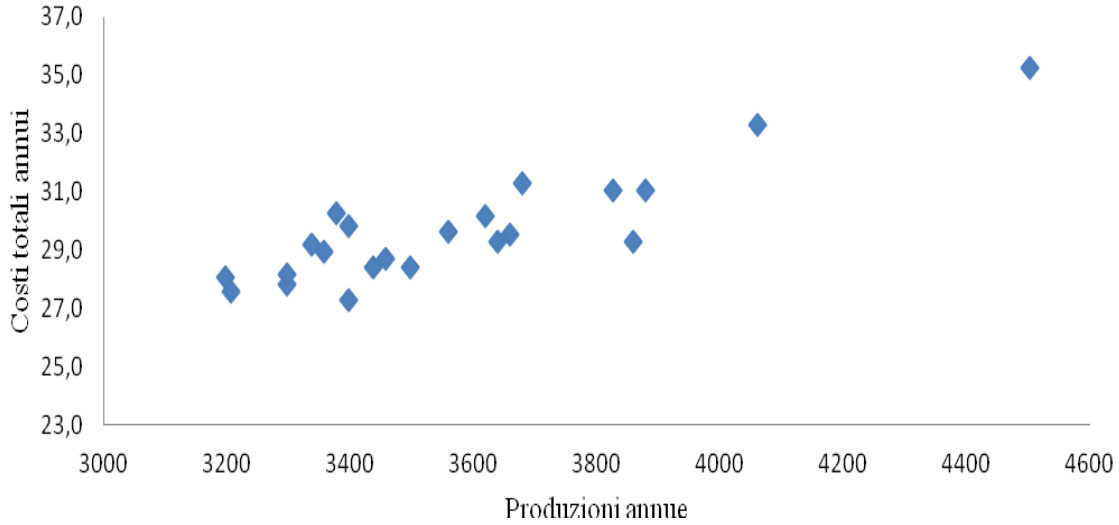
Supponiamo che un'azienda del settore alimentare abbia diversi stabilimenti (oppure molti reparti all'interno di uno stabilimento) di dimensioni diverse secondo il numero di addetti, che producono lo stesso tipo di prodotto con volumi di produzione e di vendite differenti, e che sono localizzati in tutto il territorio nazionale, ad esempio nei capoluoghi di regione Italiani. L'azienda ha rilevato, attraverso un campione casuale, i dati sui volumi di produzione e costi totali riportati nella Tabella 1

Tabella 1

Capoluoghi in cui sono localizzati gli stabilimenti	*Produzioni annue in tonnellate*	*Costi totali annui in milioni di Euro*
Ancona	3557,7	29,61
Aosta	3296,4	27,83
Bari	3437,1	28,39
Bologna	3336,6	29,17
Cagliari	3618,0	30,17
Campobasso	3376,8	30,28
Catanzaro	3195,9	28,06
Firenze	4060,2	33,28
Genova	3859,2	29,28
L'Aquila	3658,2	29,51
Milano Nord	3678,3	31,28
Milano Sud	3825,0	31,06
Napoli	3396,9	29,83
Palermo	3497,4	28,39
Perugia	3296,4	28,17
Potenza	3638,1	29,28
Roma Nord	3879,3	31,06
Roma Sud	4502,4	35,27
Torino	3396,9	27,28
Trento	3457,2	28,72
Trieste	3206,0	27,56
Venezia	3356,7	28,94

L'obiettivo dell'analisi è individuare e misurare la relazione tra i costi e volume della produzione, che sono le due variabili in gioco. La verifica della eventuale relazione tra le due variabili e la successiva determinazione della forma analitica di tale relazione serviranno ovviamente per costruire il diagramma di efficienza

La prima cosa da fare per la verifica dell'esistenza della relazione è la costruzione del diagramma dispersione



Come si vede, all'aumentare del livello di produzione, aumenta il valore del costo totale: esiste pertanto una relazione positiva tra le due variabili

Tuttavia al manager interessa conoscere anche l'intensità di questo legame, che si ottiene calcolando l'indice di correlazione di Pearson

Calcoli

$$\bar{x}\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{78526,68}{22}=3569,39 \qquad \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{652,39}{22}=29,65$$

$$\sigma_{xy}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=500,52 \qquad \sigma_{xy}=\frac{1}{n}\sum_{i=1}^{n}(x_iy_i)-\bar{x}\bar{y}=\frac{1}{22}*2339644,8\text{-}(3569,39*29,65)=500,52$$

$$corr(x,y)=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_iy_i)-\bar{x}\bar{y}}{\sigma_x\cdot\sigma_y}=\frac{500,52}{\sqrt{92531,01*3,42}}=\frac{500,52}{304,19*1,85}=\frac{500,52}{562,14}=0,89$$

Per l'azienda di prodotti alimentari il valore dell'indice di correlazione tra le due variabili è molto alto

Si può concludere, pertanto, che la relazione tra il volume dei costi indicato con *y*, e il volume della produzione (*x*) è quasi del 90%

Essendo i dati ottenuti attraverso un campione si esegue un test a due code dell'ipotesi nulla di assenza di correlazione ad un livello di significatività pari, ad esempio, a α=0,05. Il valore della statistica t è:

$$t = 0,89\sqrt{\frac{22-2}{1-(0,89)^2}} = 8,74$$

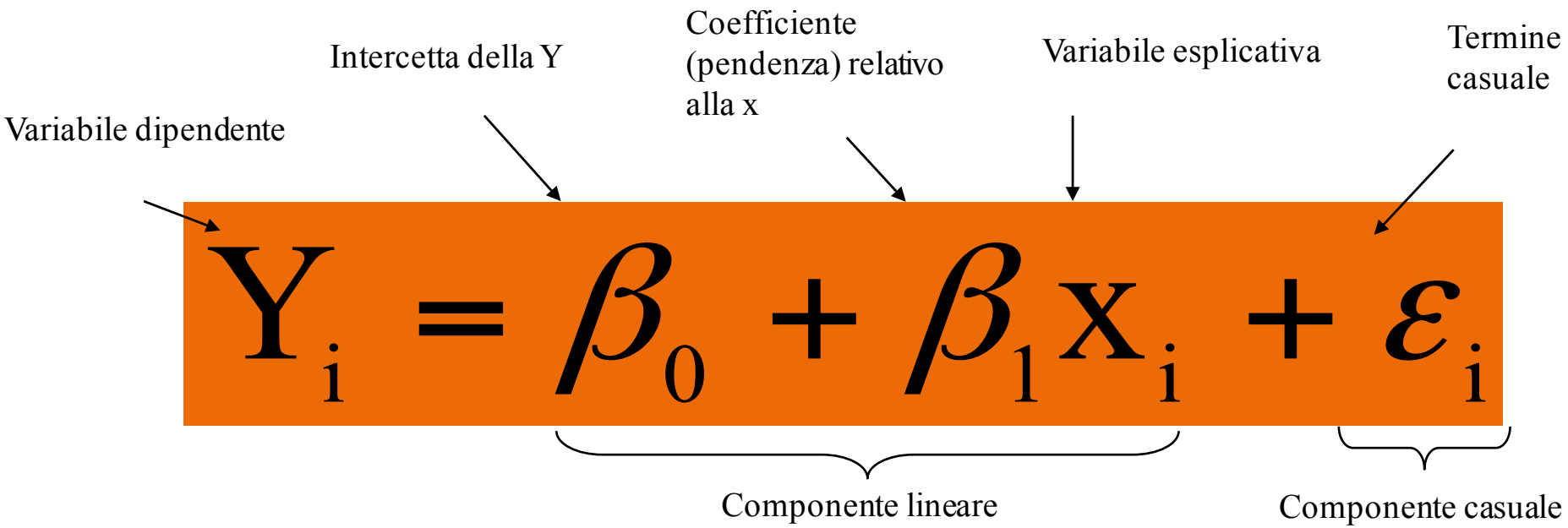
Per una distribuzione *t* di Student con 20 gradi di libertà, la probabilità di osservare un valore maggiore di |8,74| è p=0,000000028. Si rifiuta quindi l'ipotesi nulla: sulla base dei dati campionari osservati, c'è evidenza che la correlazione reale nella popolazione sia diversa da 0

Il modello di regressione lineare semplice

L'indice di correlazione misura una relazione lineare biunivoca tra due variabili. Ma in azienda, nel caso dell'analisi dei costi descritto nell'esempio, nasce quasi sempre l'esigenza di studiare e misurare la correlazione intesa come una relazione di dipendenza di una variabile (o più variabili) rispetto a un'altra variabile (o più variabili)

Focalizzando l'attenzione sulla relazione tra due variabili, la relazione di dipendenza lineare di Y da X può essere espressa attraverso l'impiego di un **modello di regressione lineare semplice**

Modello di regressione lineare semplice



La "valutazione" dei valori di e aiuterà a quantificare l'entità della dipendenza di Y da X, a visualizzare la relazione tra le due variabili, e a "prevedere" Y in funzione di X (interpolazione/estrapolazione).

$\mu = \beta_0 + \beta_1 x$ componente deterministica o retta di regressione

β_0 è il valore di μ quando *x* è uguale a zero; β_1 indica di quanto aumenta μ se *x* aumenta di una unità β_1 è la derivata prima della componente deterministica rispetto a *x*. Si ricorda il legame tra il coefficiente di correlazione lineare (misura del legame biunivoco) tra X e Y e il coefficiente della retta (che misura la dipendenza lineare) di Y da X: $\rho_{xy} = b_1 \frac{\sigma_y}{\sigma_x}$

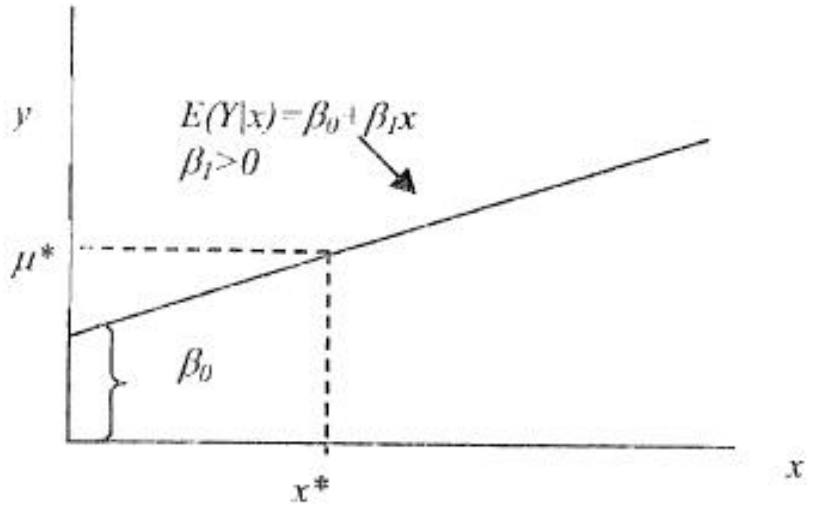
β_0 è espresso nella stessa unità di misura della variabile risposta;

β_1 ha unità di misura che dipende da *y* e da *x*

L'elasticità di μ rispetto a x^3 è pari a $\beta_1 / (\mu / x)$. Se tale grandezza è, ad esempio, *0,3* significa che al variare di *x* del *1%*, *y* varia dello *0,3%*

Nel modello probabilistico la componente deterministica rappresenta la linea delle medie $E(Y/x) = \beta_0 + \beta_1 x + \varepsilon$. μ è detto anche valore atteso condizionato della v.c. *Y* rispetto ad un determinato valore di *x;* ad esempio, nell'espressione sotto, μ^* è il valore atteso della variabile risposta nel punto *x** e cioè *condizionatamente ad x=x**

$$E(Y/x = x^*) = \mu^* = \beta_0 + \beta_1 x^*$$



Componente deterministica del modello

Specificazione della componente deterministica

Definiamo il modello con riferimento a un campione di *n* coppie di osservazioni sulle due variabili. Nelle formule verrà aggiunto il pedice *i* per indicare la generica unità statistica *i*

Modello probabilistico:

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ per *i=1, ... ,n*

da cui deriviamo la formula della retta di regressione (retta delle medie)

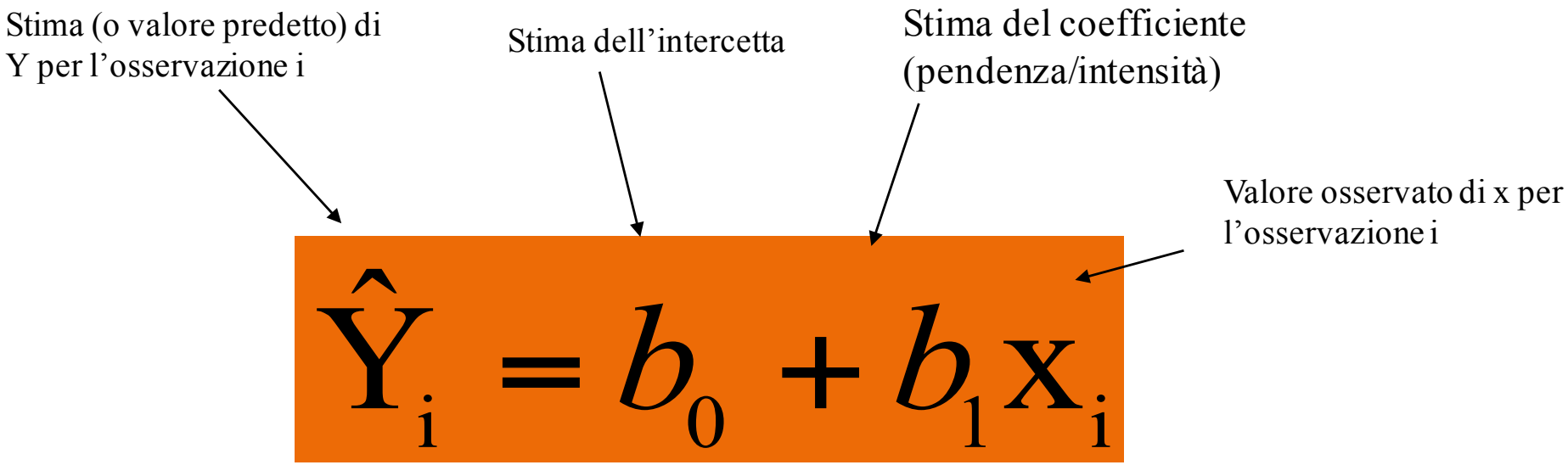
$$E(Y_i / x_i) = \mu_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Dalla espressione del valore atteso condizionato sopra definito si deduce che per ogni *i*

Da ciò si deduce che $E(\varepsilon_i) = 0$

Modello di regressione lineare semplice

La retta dei Minimi Quadrati Ordinari è quella che, fra tutte le rette possibili, minimizza la somma dei quadrati degli scarti fra valori osservati e i valori interpolati della variabile risposta, da cui ricava che soluzione del sistema di equazioni che ammette una unica soluzione



La retta dei Minimi Quadrati Ordinari (MQO) e la stima dei parametri

$$b_1 = \frac{SS_{xy}}{SS_{xx}} \qquad b_0 = \bar{y} - b_1\bar{x}$$

dove

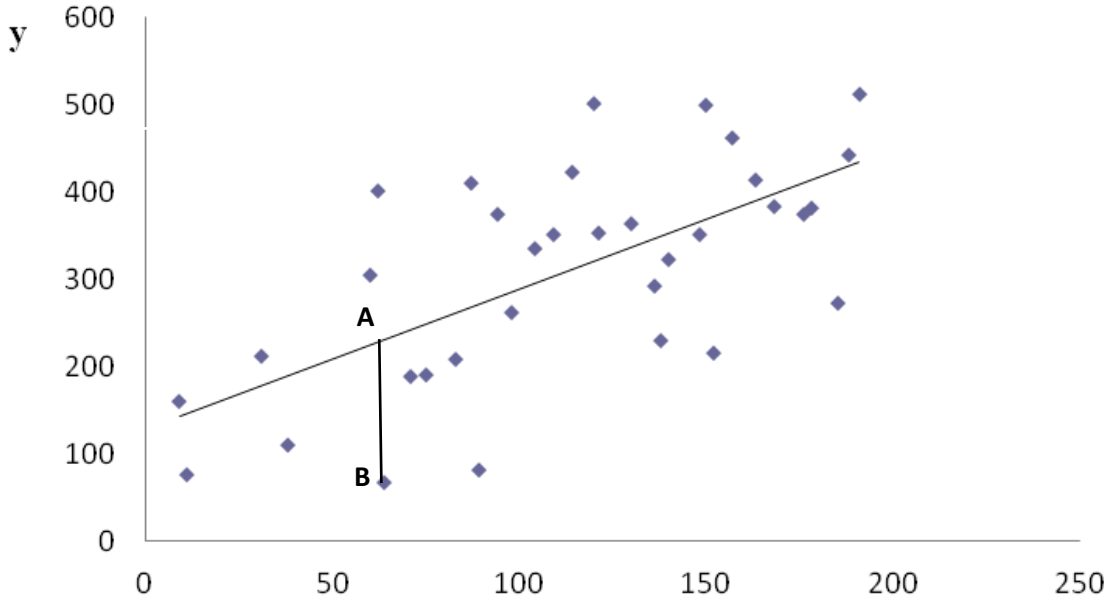
$$ss_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \textit{(devianza di x)}$$

$$ss_{xy} = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) \qquad \textit{(codevianza di xy)}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} y_i h_i \qquad h_i = \frac{(x - \bar{x})}{ss_{xx}}$$

con *n = dimensione campionaria*

Il metodo dei Minimi Quadrati Ordinari



B

Le distanze fra valori osservati e valori interpolati sono rappresentate dai segmenti verticali di tipo AB. Tali distanze, denominate anche scarti e indicate con d_i, vengono calcolate come: $d_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$

La somma dei quadrati degli scarti che è minima in corrispondenza dei valori stimati dai MQO è:

$$sse = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_i) \right]^2$$

sse (sum of squared errors) indica il minimo valore che può assumere la somma dei quadrati dagli scarti

La bontà di adattamento del modello

Si può dare un giudizio descrittivo sull'adattamento del modello, mediante l'indice di determinazione R^2 che è uguale al rapporto fra devianza di regressione e devianza totale della variabile dipendente

$$\frac{Devianza\ \ di\ \ regressione}{Devianza\ \ totale} = 1 - \frac{Devianza\ \ residua}{Devianza\ \ totale} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Indicatore dell'adattamento globale e la capacità esplicativa complessiva del modello in rapporto ai dati campionari

$0 \leq R^2 \leq 1$

$R^2=0$, il modello non spiega per niente la variabile risposta (la devianza di regressione è nulla e la variabilità di Y non dipende dalla relazione con X). La retta di regressione è parallela all'asse X (caso di indipendenza interpolativa)

$R^2=1$, il modello spiega perfettamente la variabile risposta. I punti sono allineati sulla retta di regressione

Specificazione del modello di regressione lineare semplice di I° ordine

a. $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ per *i=1, … ,n*

b. $E(\varepsilon_i) = 0 \longrightarrow E(Y_i/x_i) = \mu_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ per $\forall$ *i=1,…,n*

c. $V(\varepsilon_i) = \sigma^2 \longrightarrow V(Y_i) = \sigma^2$ per $\forall$ *i=1,…,n*

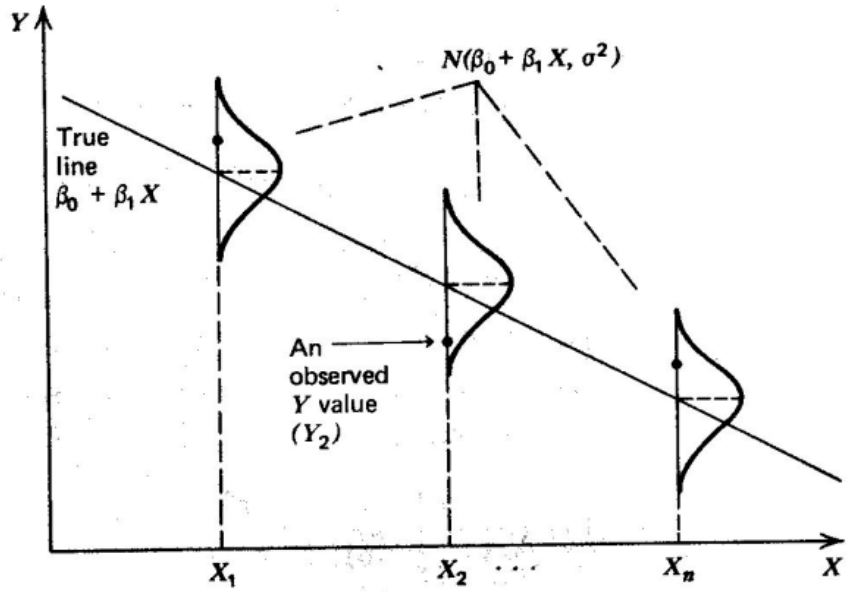
d. la distribuzione di ε_i è normale con varianza omoschedastica:

$$\varepsilon_i \sim N(0,\sigma^2) \longrightarrow Y_i \sim N((\beta_0 + \beta_1 x_i),\sigma^2))$$

e. gli errori sono indipendenti, vale a dire che il disturbo associato all'osservazione i-esima non ha nessun effetto su quello dell'osservazione j-esima

$Cov(\varepsilon_i, \varepsilon_j) = 0$ per $\forall$ $i \neq j$ con *i=1,…,n*

Le implicazioni delle prime quattro ipotesi possono essere facilmente apprezzate dalla Figura. Le distribuzioni degli errori sono centrate sulla media ovvero su un punto che si trova sulla retta.Le distribuzioni sono tutte uguali, al variare della *x,* perché $\varepsilon_i \sim N(0,\sigma^2)$ per ogni *i.* L'ipotesi di normalità della componente di disturbo è utile per poter effettuare analisi inferenziale come: stima di intervallo e test delle ipotesi sui parametri della componente deterministica



Distribuzione del disturbo e retta di regressione

Stimatori del MQO del modello di regressione lineare semplice I° ordine

$$B_1 = \frac{SS_{xy}}{ss_{xx}} \qquad B_0 = \bar{Y} - B_1\bar{x}$$

dove

$$ss_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \textit{(devianza di x)}$$

$$SS_{xy} = \sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x}) \quad \textit{(codevianza di xy)}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}Y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n}Y_i h_i \qquad h_i = \frac{(x - \bar{x})}{ss_{xx}}$$

con *n = dimensione campionaria*

Definiamo la variabile casuale: $\hat{Y}_i = B_0 + B_1 x_i$
Questa variabile casuale può essere considerata:

1. uno stimatore della media $\mu = \beta_0 + \beta_1 x$

2. un predittore del valore della variabile casuale $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ si parla in questo caso di predittore e non di stimatore perché esso ha lo scopo di "stimare" la determinazione di una variabile casuale e non il valore di un parametro (costante)

Nel caso 1. abbiamo che $\hat{Y}_i$ è stimatore corretto di μ_i. L'errore di stima pari a

$$W_i = \hat{Y}_i - \mu_i = \hat{Y}_i - (\beta_0 + \beta_1 x_i) = (B_0 - \beta_0) + (B_1 - \beta_1)x_i \,.$$

La varianza di e_i (che è uguale alla varianza di $\hat{Y}_i$) è

$$V(W_i) = E(\hat{Y}_i - \mu_i)^2 = V(\hat{Y}_i) = \sigma^2 \left(\frac{1}{n} + h_i^2 \right).$$

Pertanto, quanto più è grande $V(\hat{Y}_i)$ tanto meno è precisa la statistica $\hat{Y}_i$ nella stima di μ_i.

Nel caso 2., viene introdotto l'errore di previsione *E*$_i$ che è

$$E_i = \hat{Y}_i - Y_i = B_0 + B_1 x_i - (\beta_0 + \beta_1 x_i + \varepsilon_i) = (B_0 - \beta_0) + (B_1 - \beta_1)x_i + \varepsilon_i$$

La varianza di *E*$_i$ è:

$$V(E_i) = E\left[(\hat{Y}_i - \mu_i) - (Y_i - \mu) \right]^2 = \sigma^2 + V(W_i) = \sigma^2 \left(1 + \frac{1}{n} + h_i^2 \right).$$

Quanto più grande è questa varianza tanto peggiore sarà $\hat{Y}_i$ quale predittore di Y_i. Si noti che la varianza dell'errore di previsione è più grande di quella dell'errore di stima. In quest'ultimo è presente solo la variabilità campionaria nella stima di μ_i; nell'errore di previsione è presente anche la componente stocastica del modello.

La variabile casuale Y_i e il suo predittore $\hat{Y}_i$ hanno lo stesso valore atteso e cioè $E(Y_i) = E(\hat{Y}_i) = \mu_i$.

La stima corretta della varianza σ^2 viene ricavata attraverso i residui dei MQO e precisamente

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

La somma dei quadrati dei residui viene divisa per la grandezza [numero di osservazioni- numero dei parametri-1)]. La grandezza

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$$

è lo stimatore corretto di σ^2.

Sfruttando l'ipotesi di normalità della componente stocastica è possibile determinare intervalli di confidenza per i parametri e per μ_i e intervallo di previsione per Y_i.

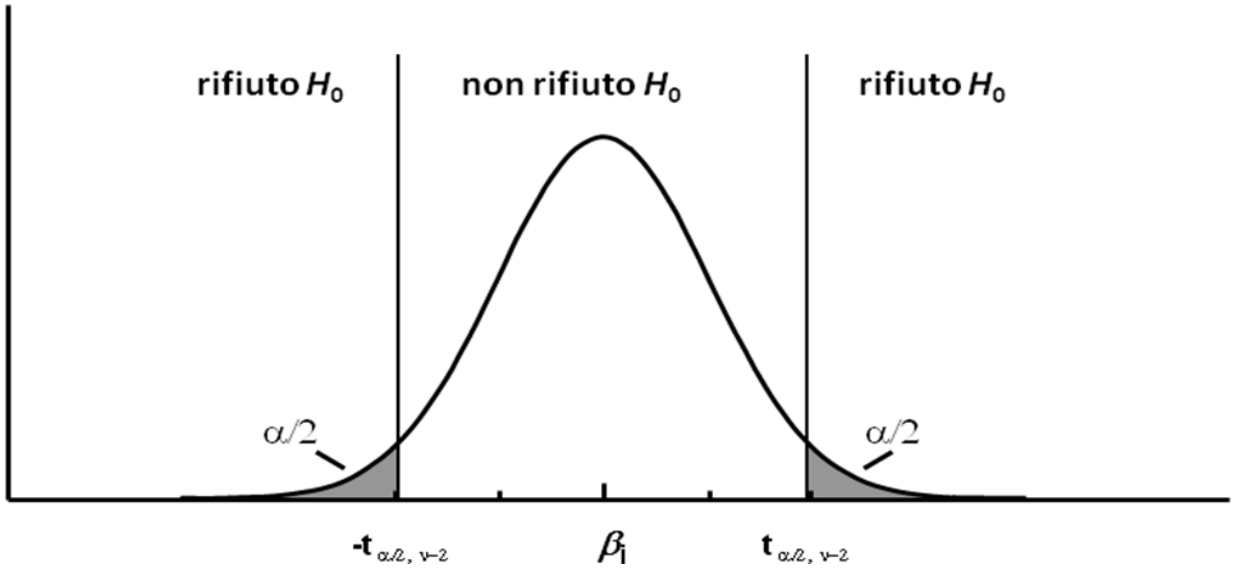
La significatività delle stime

E' significativo o meno l'effetto della variabile esplicativa sulla variabile dipendente?

Si vuole verificare l'ipotesi $H_0 : \beta_j = 0$ contro l'ipotesi alternativa $H_1 : \beta_j \neq 0$ (test bidirezionale) con un livello di significatività α ,

La statistica test è $\frac{B_1 - 0}{\sqrt{\hat{V}(B_1)}} \sim t_{n-2}$

Occorrerà individuare la regione di rifiuto data dai valori della statistica *t* che sono maggiori del valore teorico $t_{\alpha/2,n-2}$, dedotto dalla tavola delle probabilità della distribuzione t di Student



$P(|t_{n-2}|>|t_{calc}|)=\alpha$= Pr(rifiutare $H_0|H_0$), quindi gli intervalli $[-\infty, -t_{\alpha/2,n-2})$ e $(\infty, t_{\alpha/2,n-2}]$ sono *la regione di rifiuto di* H_0 e di conseguenza *la regione di non rifiuto di* H_0 è $[-t_{\alpha/2,n-2}, t_{\alpha/2,n-2}]$

L’analisi dei residui

L’ultima fase dell’analisi consiste nello studio della componente stocastica del modello come ulteriore verifica della bontà del modello adattato e per controllare se valgono alcune assunzioni fondamentali alla base del modello lineare:

le ipotesi di normalità e di omoschedasticità

Solo se questi controlli convalidano le assunzioni fatte si può considerare valida l’operazione di stima mediante i MQO e quindi anche il risultato del test delle ipotesi

Ma come si fa se la variabile non è osservabile ?
Si analizzano pertanto i residui di regressione $r_i = y_i - \hat{y}_i$

Il residuo i-esimo è la determinazione di una variabile casuale con media nulla ma con varianza che dipende da *i* (e cioè dai valori delle variabili esplicative)
Il procedimento corretto sarebbe quello di studiare i residui standardizzati

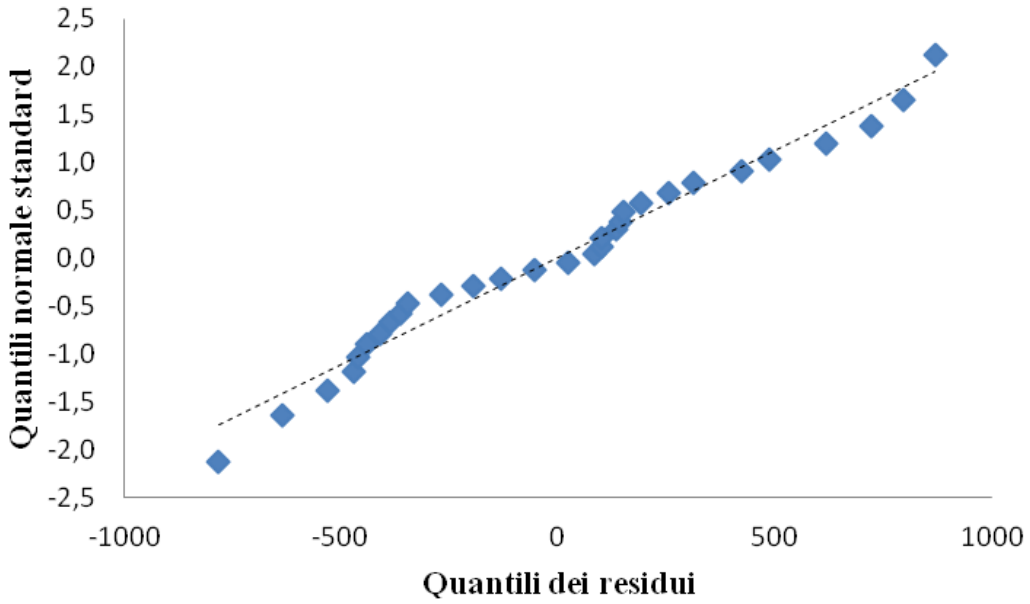
$$r_{i,std} = \frac{r_i}{s(r_i)}$$

dove $s(r_i)$ è la devazione standard stimata del residuo *i*

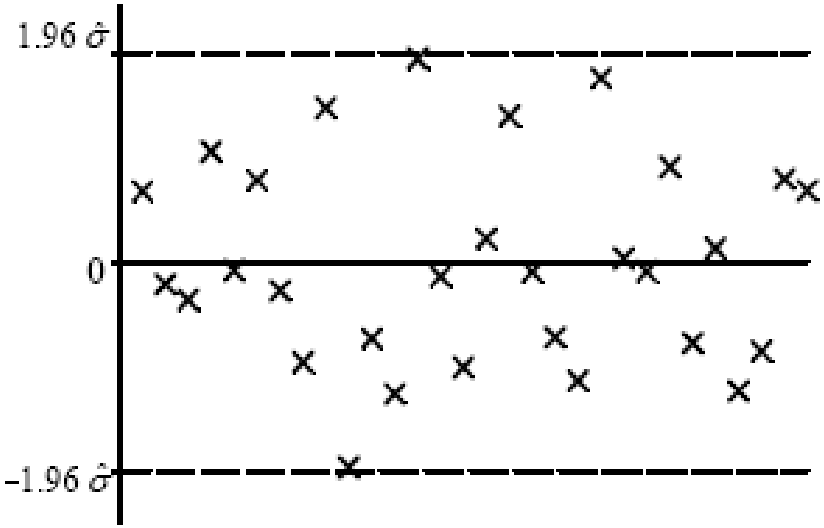
Essi sarebbero in vero delle variabili casuali t-Student tuttavia, per una numerosità campionaria n abbastanza grande, si può applicare la distribuzione normale

Verifica empirica dell'ipotesi di normalità

I residui dovrebbero presentare una distribuzione empirica piuttosto simile alla normale. Per confrontare la distribuzione dei residui con la distribuzione normale può essere utilizzato un grafico detto q-qnormal plot



Un'altra rappresentazione utile per verificare l'ipotesi di normalità



I residui dovrebbero disporsi in maniera casuale intorno all'asse delle ascisse. Sotto l'ipotesi di normalità, il 95% di essi dovrebbe essere compreso tra $\pm 1{,}96 \cdot \hat{\sigma}$

Verifica empirica dell'ipotesi di omoschedasticità

La presenza di strutture particolari nel grafico dei residui può indicare errori di specificazione nel modello

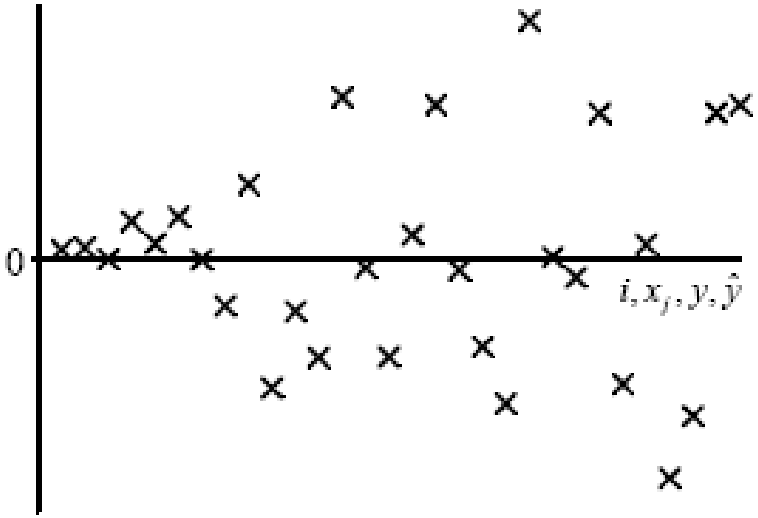
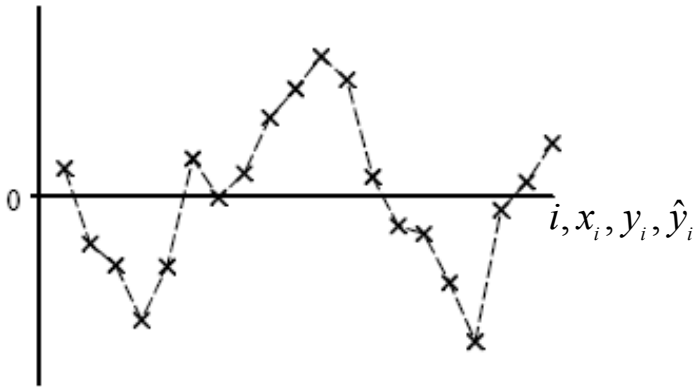
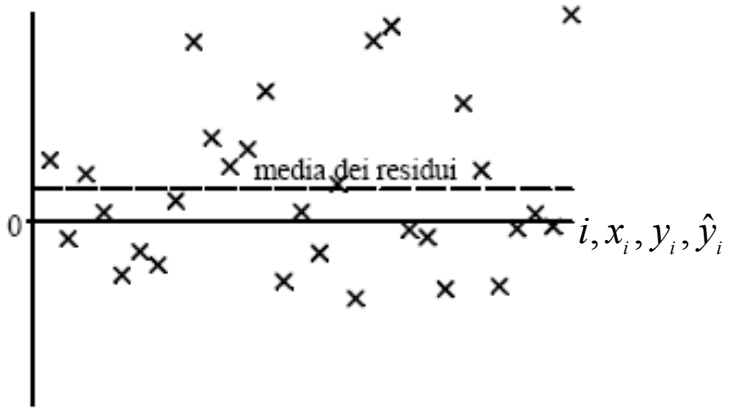


Grafico dei residui in presenza di eteroschedasticità

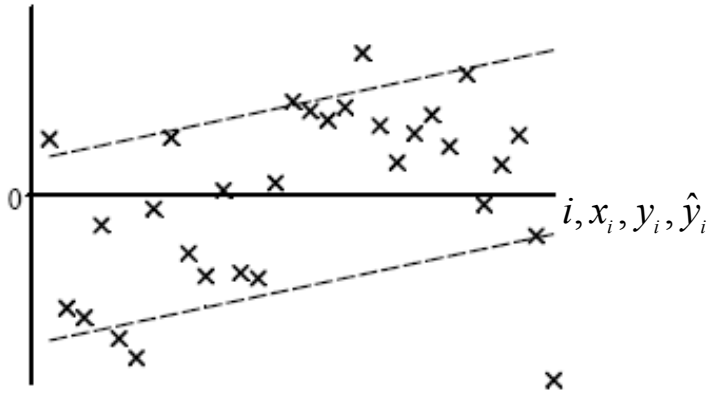
Altri casi particolari



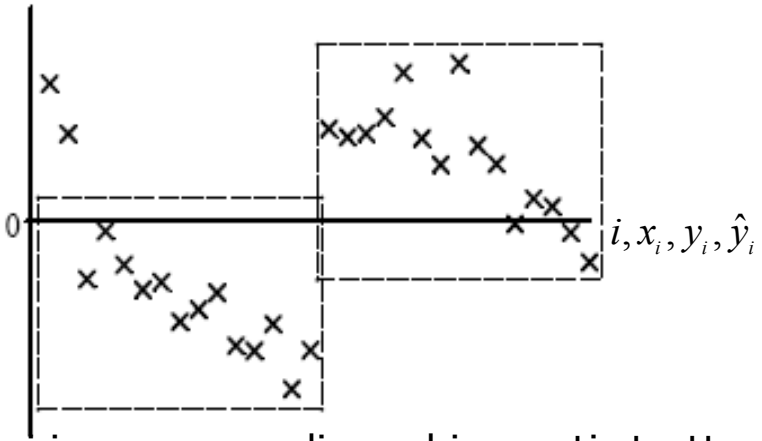
Residui in presenza di autocorrelazione



Residui nel caso di una stima del modello senza intercetta



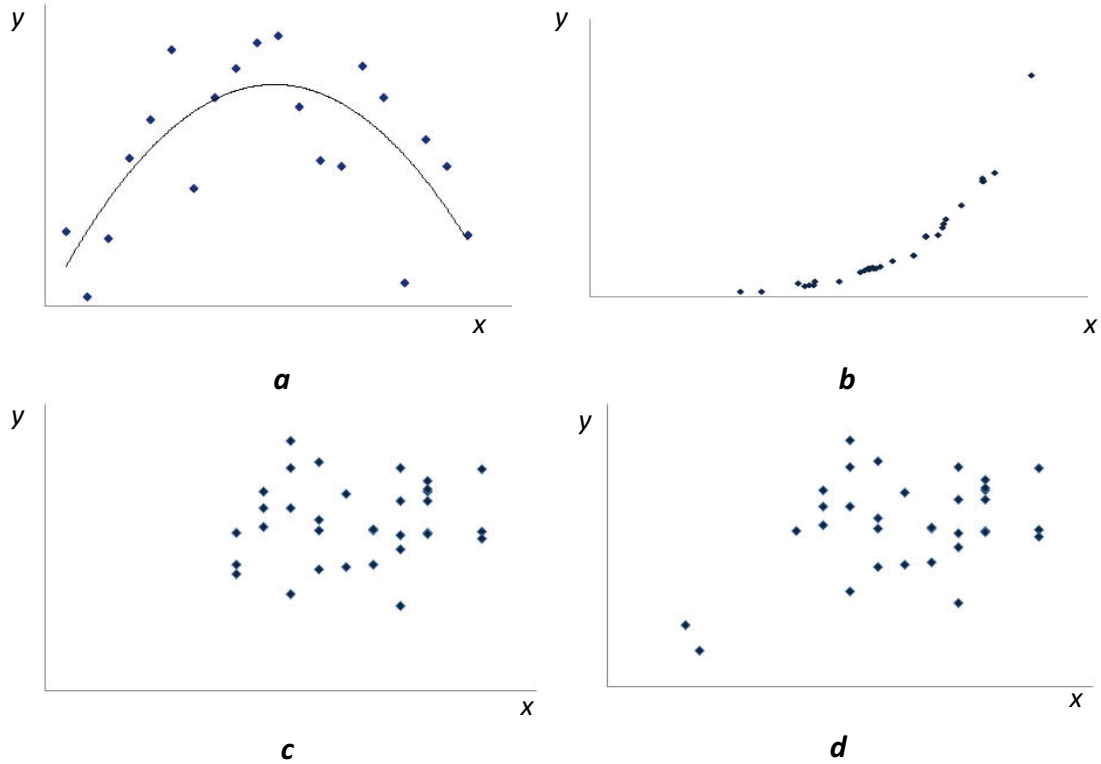
Residui nel caso di una stima del modello con una variabile esplicativa omessa



Residui in presenza di cambiamenti strutturali

Casi di relazioni non lineari e presenza di dati anomali

Il coefficiente di correlazione non risulta più una misura di intensità e segno del legame tra variabili sia quando la relazione fra le variabili è di tipo non lineare, sia in presenza di valori anomali (che sono molto diversi dal resto dei valori osservati della variabile)



Legenda:

(a) relazione quadratica

(b) relazione esponenziale

(c) relazione quasi nulla

(d) relazione quasi nulla con la presenza anche di valori anomali

Relazioni non lineari tra variabili e presenza di valori anomali

Quando le relazioni non sono lineari occorre utilizzare una forma funzionale appropriata che esprima la vera relazione esistente tra le variabili

La funzione di una relazione quadratica può essere di utilità pratica, anche per la costruzione del diagramma del punto di efficienza

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \qquad \text{per } i=1, \ldots, n$$

dove

β_0 intercetta del modello

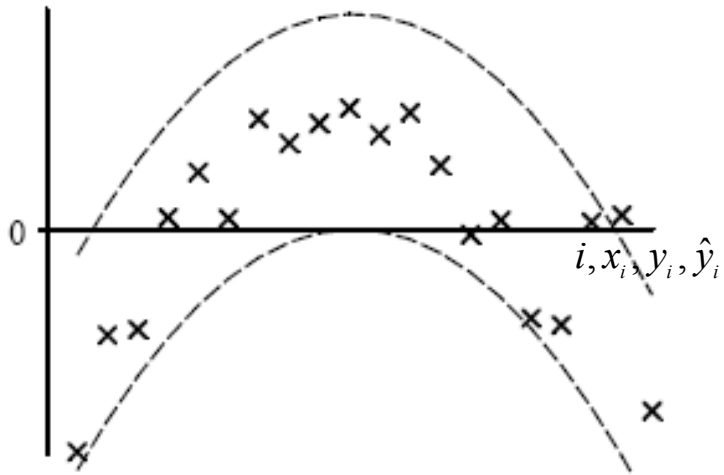
β_1 coefficiente dell'effetto lineare di x su y

β_2 coefficiente dell'effetto quadrato di x su y.

Se β_1 è positivo, a fronte di un incremento unitario di x la funzione cresce di β_1, ma questo incremento varia di $2 \cdot \beta_2$ al variare unitario di x; se β_1 è minore di $2 \cdot \beta_2$ x, il modello declina. Se β_2 è negativo, allora l'incremento di y decresce di $2 \cdot \beta_2$ per un incremento unitario di x il decremento di y aumenta di $2 \cdot \beta_2$

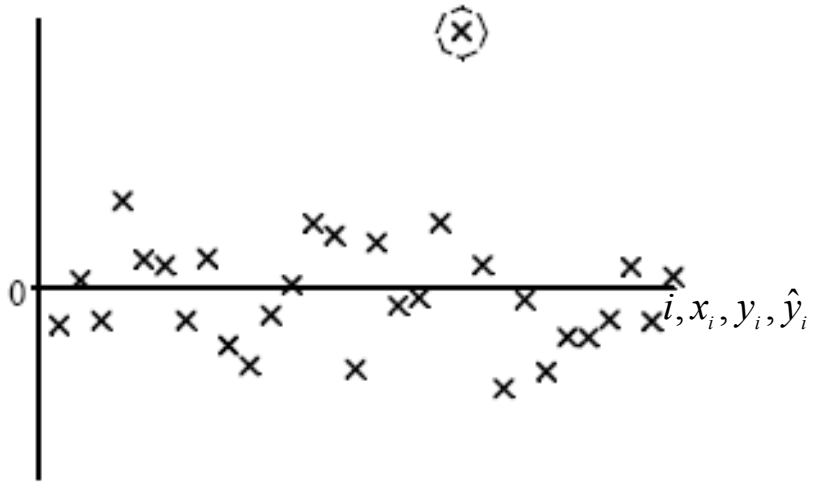
Il modello viene stimato con il metodo dei Minimi Quadrati Ordinari.

Andamento non lineare dei residui che sottintende un modello di regressione non lineare



Valori anomali (outliers)

La presenza di valori anomali nei dati genera la presenza di residui corrispondenti ai valori anomali, che sono molto distanti dagli altri residui



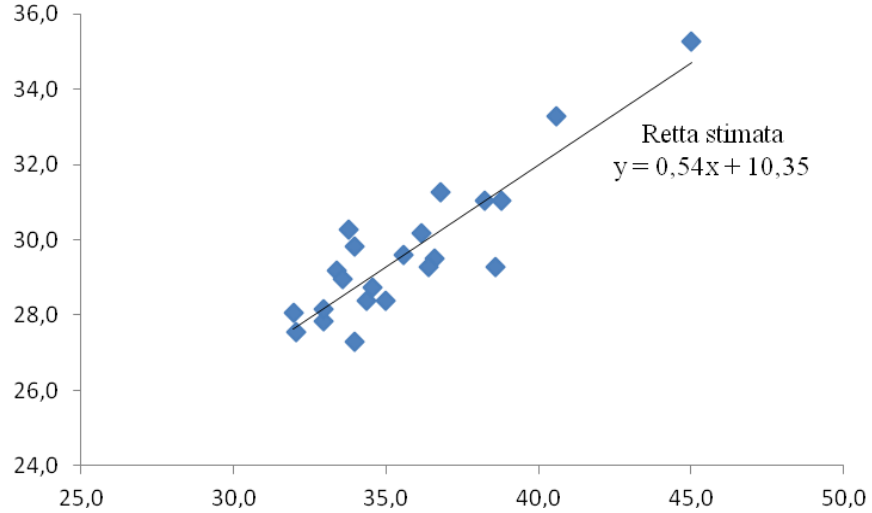
ESEMPIO sul modello di regressione semplice

Dopo avere effettuato un'analisi di correlazione, il cui risultato ha prodotto un coefficiente molto alto (0,89), l'azienda del settore alimentare si pone l'obiettivo di individuare una relazione funzionale tra le due variabili costi totali e volume di produzione, stimando un modello di regressione lineare semplice, ovviamente utilizzando i dati della Tabella 1

Risultati della stima del modello lineare semplice con il metodo MQO

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*	*Intervallo di confidenza al 95%*	
Intercetta	10,35	2,22	4,67	$0{,}15 \cdot 10^{-3}$	5,73	14,97
*Produzioni annue in centinaia di tonnellate**	0,54	0,06	8,75	$0{,}03 \cdot 10^{-6}$	0,41	0,67

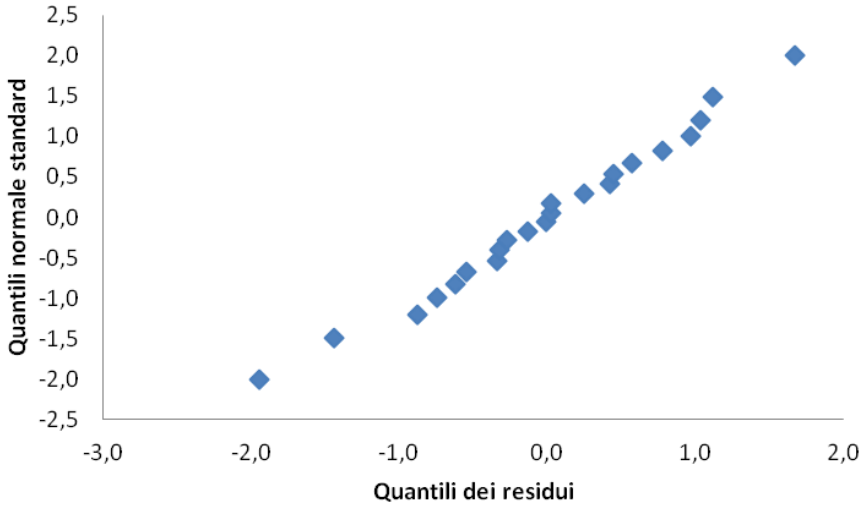
** I dati originari presentati in Tabella sono stati riproporzionati rispetto all'ordine di grandezza della variabile costi, pertanto le produzioni annue sono espresse in centinaia di tonnellate*



Unità campionarie e retta di regressione stimata

L'azienda osserva il valore dell'indice di determinazione *che* R^2= 0,79 e quindi la capacità del modello di rappresentare le relazione tra le due variabili è abbastanza buono, ma non conferma l'ipotesi che spesso viene adottata di diretta proporzionalità dei costi variabili al variare del volume di produzione

Con l'analisi empirica sui residui, mediante l'osservazione dei grafici del q-q normal plot e dei residui standardizzati rispetto alla variabile risposta stimata, il manager verifica la conformità del modello adottato e delle analisi svolte alle ipotesi inferenziali sottostanti il modello teorico



Q-qnormal plot dei residui

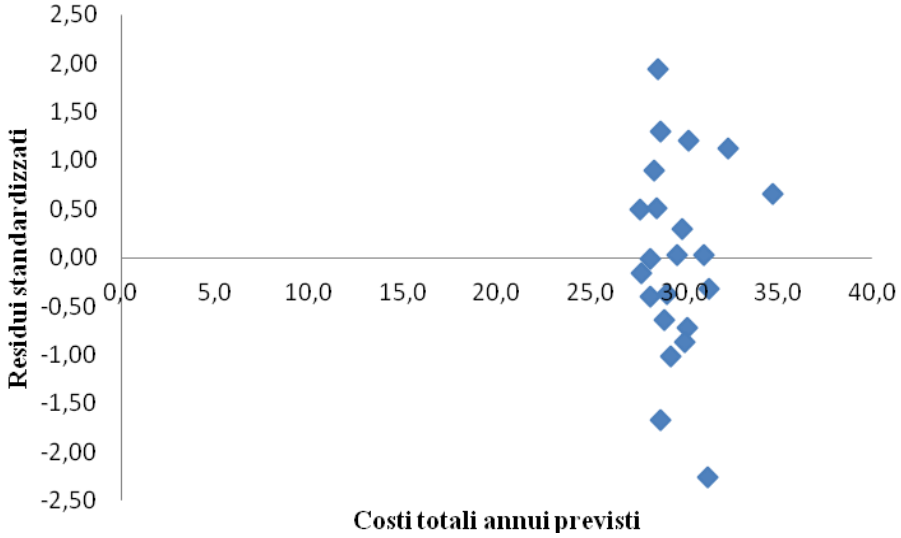


Grafico dei residui standardizzati

Competenze acquisite

- applicazione dell'indice di correlazione di Pearson;
- interpretazione dell'indice di correlazione di Pearson;
- applicazione del modello di regressione lineare semplice;
- interpretazione dei risultati del modello di regressione semplice stimato;
- impiego di alcuni particolari grafici nell'analisi di regressione;