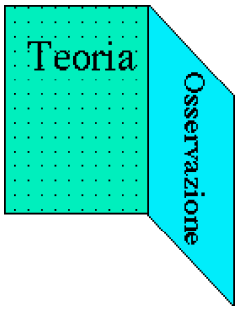
Dei modelli

I problemi statistici incontrati nello studio delle discipline socioeconomiche nascono dal dualismo fra evidenza empirica e analisi teorica.



Tale dualismo è incorporato nella nozione di modello scientifico

Non sono possibili costruzioni teoriche senza ripetizione.

I fenomeni socio-economici non si ripetono se considerati solo in modo superficiale e descrittivo.

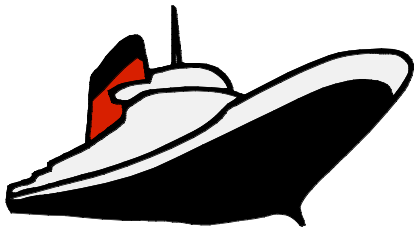
Se però ci si limita ai fattori più rilevanti troviamo delle ricorrenze sulle quali impostare il modello

Semplificazione ed astrazione

Il modello è una rappresentazione semplificata ed astratta di una realtà. Con esso si può lavorare su una realtà più grande, complessa e mutevole.

Esso sta alla realtà come il quadro sta alla fotografia: questa riporta tutto, quello solo ciò che ha colpito l'ispirazione dell'artista.

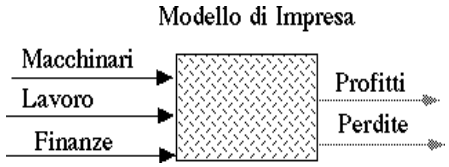
Il modello dà risposte in ragione della sua vicinanza al fenomeno che rappresenta



Per studiare il comportamento della nave non si userà una barchetta di carta, ma una serie di equazioni, disegni e modelli in scala.

Esempio: modello di impresa

Un' impresa può essere rappresentata come una combinazione di inputs per produrre profitti



Il modello descrive e spiega schematicamente una certa situazione imprenditoriale.

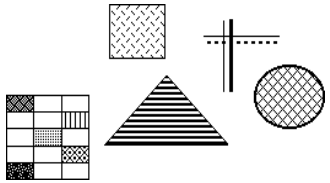
Le conclusioni basate sul modello non sono neutrali: sono legate alle ipotesi in esso inglobate.

Ogni compromesso ragionevole tra semplicità e realtà è un modello

Estensioni del modello

il limite del modello di impresa è che non tiene conto dell'interazione con le altre imprese. Per includerla occorre allargare o cambiare il modello

L'approccio deve essere pluralistico: non c'è un unico, universale modello scientifico.



E' necessario predisporre modelli differenti per affrontare le varie situazioni incontrate nelle scienze economiche e sociali

Gli elementi soggettivi sono tali e tanti che autori diversi, pur lavorando sulla stessa realtà pervengono a modelli diversi e talvolta contrapposti

Tipi di modelli

VERBALI Descrizione a parole di una situazione verificabile:

La diminuzione del Tasso Ufficiale di Sconto favorisce gli investimenti

MATEMATICI Gli aspetti essenziali di una situazione sono espressi con delle equazioni

La 1ª legge del moto di Newton

$$f(t) = \beta_0 + \beta_1 t \quad con \quad \begin{cases} \beta_0 = posizione\ \ iniziale \\ \beta_1 = velocità\ \ uniforme \\ f(t) = distanza\ percorsa \end{cases}$$

I modelli studiati dalla statistica -sotto qualsiasi forma- debbono avere precisi riscontri nella realtà

Tipi di modelli/2

FISICI Si costruisce un apparato che rappresenta IN SCALA la situazione di studio oppure ne rappresenta una parte.

Lo studio del CX per le auto nella galleria del vento

ANALOGICI Delle relazioni non fisiche sono simulate con dei meccanismi fisici

La diffusione di un dialetto attraverso i cerchi concentrici che si producono sull'acqua

Esempio: il sistema economico

La costruzione della teoria economica consiste nella elaborazione di un sistema di equazioni interconnesse (il modello).

Le equazioni devono comporsi di variabili Indipendenti, dipendenti e Parametri.

Conoscendo i parametri siamo in grado di sapere quali saranno le variazioni nelle dipendenti per ogni combinazione di livelli nelle indipendenti

Esempio di modello econometrico multiequazionale →

$$\begin{cases} Y = C + I \\ C = a_1 + a_2 Y + a_3 r \\ I = a_4 + a_5 r \\ r = 11\% \end{cases} \quad con \quad \begin{cases} Y = \text{Re}\,ddito \\ C = Consumi \\ I = Investimenti \\ r = Tasso\ \ di\ \ Sconto \end{cases}$$

$a_1\ a_2\ a_3\ a_4\ a_5$ **= parametri**

Queste relazioni debbono essere coerenti al loro interno (non avere contraddizioni) e rispecchiare la realtà osservata

Il sistema economico/2

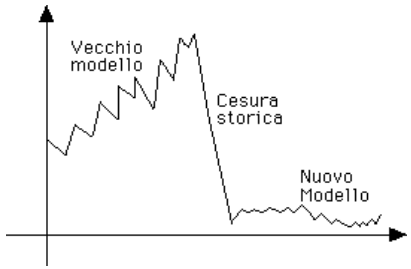
il funzionamento del sistema, PER I SUOI EFFETTI CUMULATIVI, provoca un graduale modificarsi nei parametri

il modello si applica finché i parametri sono invariati o cambiati di poco

Da un certo punto in poi le variazioni nei parametri superano l'elasticità assunta dal modello.

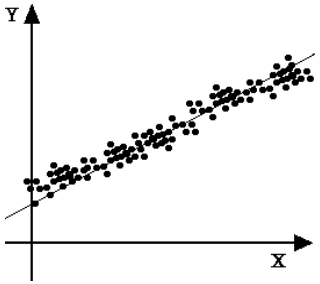
A questo punto il modello crolla perché nel fenomeno in esame c'è un cambiamento strutturale: una vera e propria cesura storica.

Un modello ben costruito dovrebbe spiegare il funzionamento dell'economia, ma contenere anche elementi di autodistruzione.



Relazione tra due variabili

Dopo aver rappresentato graficamente i dati a mezzo dello scatterplot si è interessati a determinare una curva che passi vicino ai punti



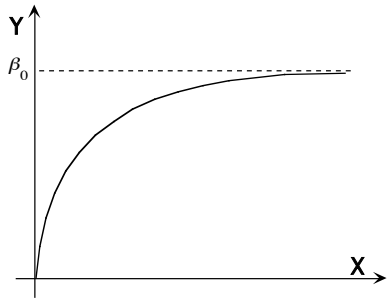
- **Per sostituire uno schema semplice alla nube dei punti**
- **Per sintetizzare le tendenze di fondo**
- **Per ricostruire o determinare il valore della Y noto quello della X o viceversa**

il presupposto è che esiste una variabile (la "X" detta indipendente o esogena) che è causa o comunque agisce sull'altra (la "Y" detta dipendente o endogena).

Esempio di costruzione del modello/2

La funzione "f" è al momento indeterminata: si sa che un certo legame esiste, ma non si riesce a darne una esatta espressione analitica.

E' noto che, a parità di forma, una specie non può superare una dimensione data. Per cui la relazione tra X ed Y è di tipo crescente, ma gli aumenti devono avvenire a ritmo decrescente



il modello potenza è particolarmente adatto per rappresentare tali situazioni.

$$Y = \beta_0\left[1-(\beta_1)^x\right],\ 0<\beta_1<1$$

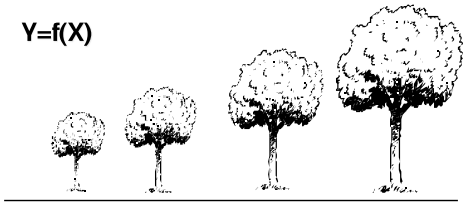
Esempio di costruzione di un modello

La teoria di un fenomeno può spesso essere sintetizzata da un modello espresso da una equazione.

Sia "Y" l'ampiezza in cm del diametro alla base del tronco di una data specie arborea e sia "X" l'età .

L'idea che il diametro sia più grande secondo l'età può essere espressa dalla relazione funzionale:

Y=f(X)



Queste variazioni assicurano all'albero adeguata resistenza e flessibilità.

Esempio di costruzione del modello/3

Nel modello si individuano

Y= variabile ENDOGENA-DIPENDENTE-SPIEGATA-INTERNA-CONSEGUENTE

X= variabile ESOGENA-INDIPENDENTE-ESPLICATIVA-ESTERNA-ANTECEDENTE

Esistono moltissimi fattori che incidono sull'accrescimento: la quota, il tipo di suolo, l'esposizione, l'impianto arboreo, etc.

Tali fattori non solo incidono su "Y" ma si influenzano anche tra di loro determinando una rete complessa di interrelazioni che il modello ignora.

La "X" è un "riassunto" dei fattori determinanti, ovvero si sceglie "X" perché è considerata il risultato del loro comune interagire.

La variabile endogena

Una variabile ha questo ruolo se:

- **Rappresenta il fenomeno che si intende spiegare, prevedere, controllare**
- **E' una risposta ad uno più stimoli in un dato organismo**
- **E' l'output di un sistema che ha uno o più fattori in input**
- **Esprime l'obiettivo raggiungibile per uno o più tipi di interventi.**

La variabile esogena

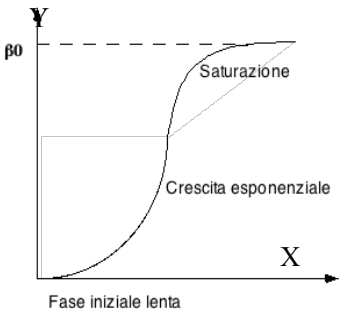
Una variabile ha questo ruolo se:

- **E' un riassunto dei fattori determinanti (età dell'albero)**
- **E' controllabile (spese in pubblicità, aumento delle vendite)**
- **E' precedente la endogena (quotazione di oggi, quotazione di domani)**
- **E' ritenuta una causa determinante della endogena (ore di studio e voto d'esame)**

Che cos'è il modello?

E l'insieme delle ipotesi e delle equazioni che stabiliscono una certa relazione tra due o più variabili.

La curva logistica $Y = \frac{\beta_0}{1 + e^{-\beta_1 X}}$ **ingloba le seguenti ipotesi:**



Avvio difficile, crescita lenta e faticosa;

Accelerazione dello sviluppo che appare quasi incontrollabile.

Rallentamento del fenomeno che finisce con l'attestarsi sull'asintoto β_0.

Tale modello descrive bene l'aumento delle popolazioni di persone, di imprese, di batteri, di prodotti, etc.

Esempio: modello Keynesiano semplice

Intende fornire una spiegazione del funzionamento del sistema economico

La variabile esplicativa è la domanda globale ovvero gli investimenti e segnatamente quelli di natura pubblica "F".

La variabile "F" si configura come "esogena" in quanto soggetta -almeno in parte- a controllo governativo.

La variabile endogena Y è il livello di produzione e del reddito della nazione in condizione di sottoccupazione

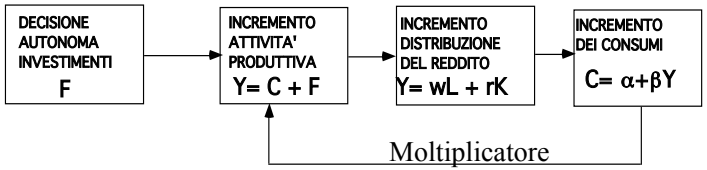
$$Y = f(F)$$

il modello Keynesiano semplice/2

F *variabile autonoma*

$$Y = C + F$$
$$wL + rK = Y$$
$$C = \alpha + \beta Y$$

$$Y = \alpha + \beta Y + F \Rightarrow (1-\beta)Y = \alpha + F$$
$$Y = \left[\frac{1}{1-\beta}\right]F + \frac{\alpha}{1-\beta}$$



La realizzazione degli investimenti attiva la produzione che a sua volta genera un aumento di reddito ai fattori lavoro e capitale.

Tale aumento induce maggiori consumi che si concreta in una maggiore produzione di beni

il modello Keynesiano semplice/3

La quantità

$$\left[\frac{1}{1-\beta}\right]$$

rappresenta il "moltiplicatore" e misura l'impatto sul reddito generato da un incremento unitario di investimenti "F".

"β" è la propensione marginale al consumo ovvero l'incremento indotto nei consumi da un aumento unitario di reddito:

$$0 < \beta < 1$$

Se la propensione marginale al consumo vale 0.60 il moltiplicatore varrà 2.5 per cui un investimento pari a 1'000 produrrà -attraverso il circuito economico- un incremento di reddito/produzione pari a 2'500.

Finalità del modello

DESCRITTIVE

al modello si chiede solo di rappresentare bene la realtà osservata.

La capacità dei Chip di Memoria si quadruplica ogni 3 anni

INTERPRETATIVE

il modello deve mettere in evidenza i legami tra i fenomeni coinvolti in forme e modi riconducibili a precise teorizzazioni.

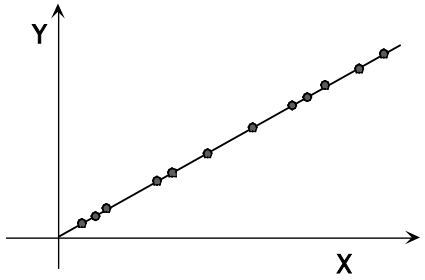
La produzione è una funzione lineare omogenea di capitale e lavoro (equazione Cobb-Douglas)

PREVISIONALI

il modello deve fornire previsioni sull'andamento futuro del fenomeno

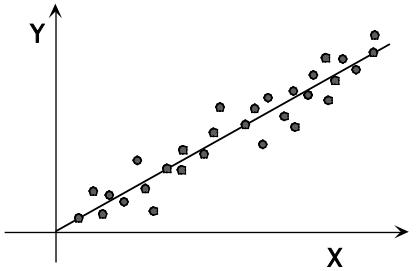
Le esportazioni di beni durevoli aumentano linearmente nel tempo

Relazioni stocastiche e deterministiche



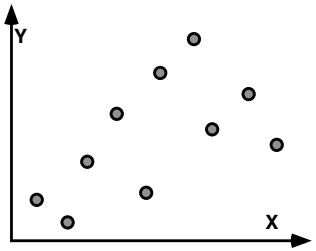
DETERMINISTICA: ad ogni età corrisponde un determinato diametro del tronco

STOCASTICA: il diametro del tronco aumenta con l'età, ma l'incremento non è UNIVOCO: talvolta aumenta di più, altre volte di meno



Relazioni esatte ed approssimate

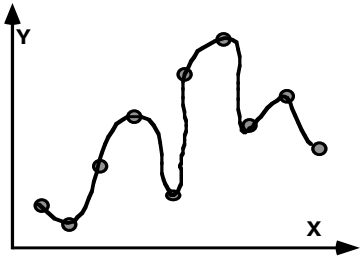
Consideriamo il seguente scatterplot:



Si conoscono i valori Y_i corrispondenti ai valori X_i;

Siamo alla ricerca di una funzione f(X) i cui valori $f(X_i)$ siano "vicini" alle Y_i .

E' possibile determinare un adattamento perfetto: se (X_i,Y_i) i=1,...,n sono "n" coppie di valori distinte allora un polinomio di grado "n-1" o inferiore si adatta perfettamente ai punti.



Relazioni esatte ed approssimate/2

Ad esempio, il polinomio di Lagrange

passa esattamente per gli "n" punti

$$f(X)=\sum_{i=1}^{n} y_i \left[\frac{\prod_{\substack{j=1\\ j\neq i}}^{n}\left(X-X_j\right)}{\prod_{\substack{j=1\\ j\neq i}}^{n}\left(X_i-X_j\right)}\right]$$

il calcolo dei valori può essere facilmente programmato al computer

Di solito i polinomi di Lagrange hanno un grado troppo elevato (comunque non superiore a n-1) per essere usabili in modo rapido e semplice.

Se si aggiunge un nuovo punto il calcolo deve essere ripetuto

In statistica si rinuncia alla perfetta interpolazione matematica per un adattamento approssimato, ma più essenziale e stabile

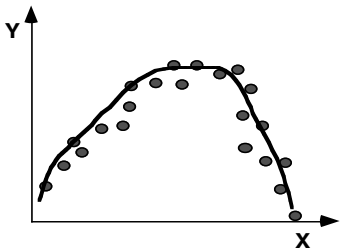
Relazioni esatte ed approssimate/3

il compromesso tra bontà di adattamento e semplicità del modello deriva da forti dose di convenzionalismo

SEMPLICITA' **Si definisce un sistema di curve "semplici" e flessibile**

$$f\left(X;\theta_1,\theta_2,\ldots,\theta_m\right)$$

ADATTABILITA'' **Si cerca di riconoscere la struttura del modello "f" nello scatterplot per poi adattarvi quella più idonea.**



L'andamento degli n=22 punti ricorda in modo indiscutibile la parabola.

Parabole ne esistono infinite, quale scegliere?

N.B. l'identificazione del modello è ostacolata dalla presenza di errori

Modello di regressione lineare semplice

Supponiamo di disporre di "n" coppie di osservazioni

y	x
y_1	x_1
y_2	x_2
M	M
y_i	x_i
M	M
y_n	x_n

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

COMPONENTE STOCASTICA (Non osservabile)

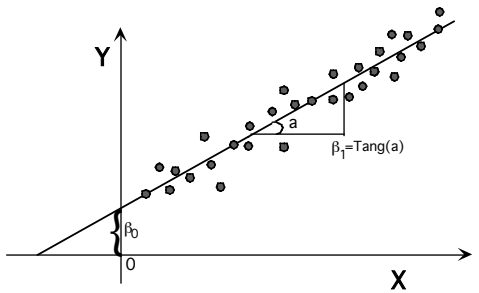
COMPONENTE DETERMINISTICA

β_0= intercetta.

valore a cui tende la Y quando la X tende a zero.

β_1 = Coefficiente angolare.

variazione in Y per un aumento unitario in X



La terminologia

Modello. **Perché è un insieme di ipotesi rispetto al legame esistente tra la variabile esogena ed endogena. Le ipotesi in genere danno luogo ad una equazione, lineare nel nostro caso**

Regressione **. E' una etichetta storica dovuta agli studi di Francis Galton (1889) sull'effetto di regressione: la tendenza a prevalere dei valori medi.**

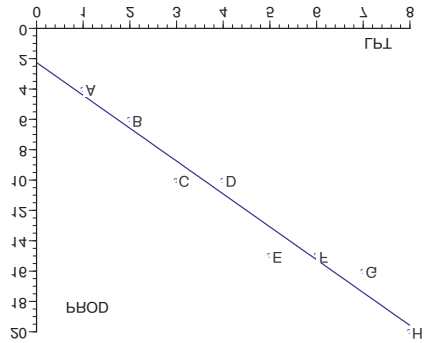
Lineare. **Perché i parametri incogniti vi compaiono con potenza 1**

Semplice. **Perchè c'è una sola variabile esplicativa (REGRESSORE) in contrapposizione a *multipla* termine usato quando vi sono più variabili esplicative.**

Esempio di modello di regressione

L'ing. Consolata Mirabelli è responsabile della produzione di semilavorati. Nel breve periodo controlla solo il lavoro part-time. L'ing. intende conoscere che relazione (se c'è) tra questo fattore e la produzione

Prova	L.P.T.	PROD
A	1	4
B	2	6
C	3	10
D	4	10
E	5	15
F	6	15
G	7	16
H	8	20

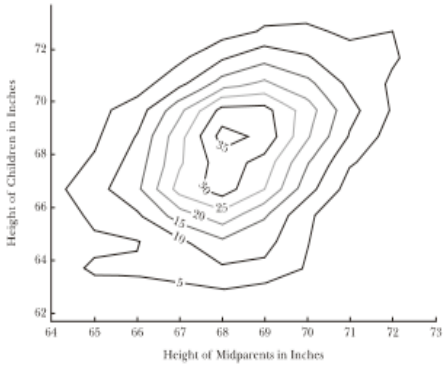


Cosa indica lo scatterplot?

1) Una sicura tendenza all'aumento della produzione dovuta all' aumento di LPT
2) Un'altrettando sicura dispersione intorno alla tendenza (espressa dalla retta)

Alle origine del concetto di regressione

Figure 4
Galton's Original Smoother‡Contour Plots: Contours after Galton Smoother



Sir Francis Galton notò che i figli di padri alti erano più alti della media, ma meno di quanto non eccedessero dalla media i loro padri. I figli di padri bassi erano in media bassi, ma meno bassi della media generale di quanto non lo fossero i padri.

Ipotizzò quindi una generale tendenza al livellamento delle altezze.

L'effetto di regressione

Il principio del ritorno alla media lo si ritrova in varie occasioni

Un docente che loda gli studenti per il buon risultato raggiunto in una prova vedrà un esito peggiore nella prova successiva (Metodo Fata turchina)

il docente che sgrida gli studenti per la pessima riuscita di un test otterrà risultati molto migliori nella seguente prova (Metodo sergente Harman nel film full metal jacket)

Un buon governo sarà seguito da una amministrazione inefficace e ad un premier inadeguato succederà un brillante primo ministro.

Nelle competizioni articolate su due fasi è frequente notare il ribaltamento degli esiti tra la prima e seconda prova: i migliori che peggiorano ed i peggiori che migliorano.

L'effetto di regressione/2

Per ottenere un buon risultato in un'impresa difficile concorrono due fattori:

- Talento/Genio
- Sorte

il successo in una prova ardua implica che entrambi i fattori hanno agito a favore.

Nella seconda prova il talento/genio magari migliorano o agiscono con la stessa intensità

La Sorte è capricciosa e imprevedibile e non si ripete.

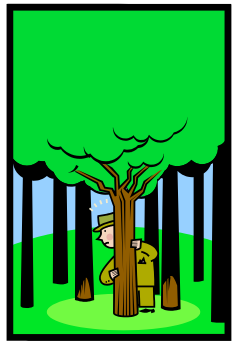
Ed ecco l'effetto di regressione alla media in cui gli scarti si annullano tutti.

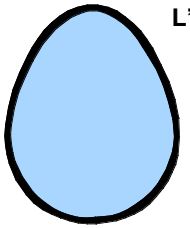
Proposta del modello lineare:

il rasoio di Occam

Se è necessario dare una soluzione ad un problema di cui si sa poco, la risposta più semplice comporta meno rischi in caso di errore ed è spesso quella giusta.

Smarriti in una foresta se ne esce spesso procedendo in linea retta.





L'uovo di Colombo

Principio di semplicità di Galilei

La natura procede per vie semplici ed offre così la sicura scelta tra le varie spiegazioni possibili dei suoi fenomeni

Proposta del modello lineare/2

il legame più semplice tra due variabili è quello lineare

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

Ipotizziamo che l'ordinata "Y" sia dovuta alla combinazione ADDITIVA di due valori: la parte deterministica (lineare) ed un errore

Si cammina nel piano fino a che non si guarda l'orizzonte per ricordarsi che la terra è una sfera.

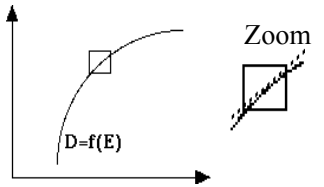
il termine "u" è il risultato di:

- Errori e carenze nella misurazione e nella rilevazione di "Y" e di "X"
- Insufficienza del solo fattore X a "spiegare" da solo la Y
- Inadeguatezza della "semplice" relazione lineare

Una ragione di più

Teorema di Taylor.

Se la funzione "f" che lega "D" ad "E"' ha derivate prime e seconde continue in un intorno del punto E0, in tale intorno la "f" è ben approssimata dalla retta



In generale si può dire che la scelta del modello lineare è motivata da

- Ragioni di semplicità
- Esigenze di sintesi
- Approssimazione funzionale

Una ragione formale

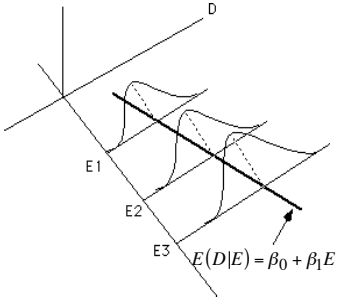
Supponiamo di scegliere un campione casuale di unità e di rilevare su ogni unità due variabili continue: "X" e "Y"

La casualità di tale esperimento è descritta da una densità bivariata f(X,Y).
Ipotizziamo che sia valido il modello gaussiano.

Ne consegue:

$$E(D|E) = \beta_0 + \beta_1 E;\ dove:\ \beta_0 = \mu_D - \rho\frac{\sigma_D}{\sigma_E}\mu_E;\ \beta_1 = \rho\frac{\sigma_D}{\sigma_E}$$

In questo modello il valore atteso di una variabile condizionato al valore dell'altra, è -necessariamente- una funzione lineare della condizionante.



Che succede se il coefficiente di correlazione "ρ" è nullo?

Limiti del modello lineare

I sistemi dinamici hanno la proprietà di non poter essere compresi se non in modo globale.

Questa regola ammette una sola eccezione: i sistemi integrabili, fra i quali si collocano in prima fila i sistemi lineari".

Si perde però di vista l'instabilità potenziale.

Se un fenomeno ha effetti cumulativi tali che:

$$Y_{n+1} = 100 * Y_n \Rightarrow Y_n = Y_0 100^n$$

una causa infinitesima può avere effetti catastrofici:

il battito d'ali di una farfalla nei Caraibi imprime al vento una forza pari a 0.000'000'000'000'000'1 nodi, ma dopo solo 9 passaggi il vento ha una forza di 100 nodi che travolgerà New York

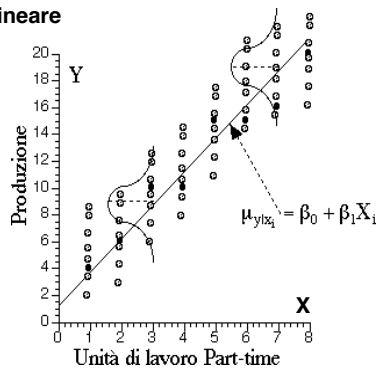
La specificazione del modello

il modello di regressione incorpora le indicazioni assumendo

- **Per ogni valore della "X" esiste una distribuzione di probabilità della "Y" (non è necessario che "X" sia una variabile casuale).**
- **Il valore atteso di questa distribuzione di "Y" varia sistematicamente al variare di "X"**
- **Le variazioni seguono una funzione lineare**

Per ogni livello di lavoro part time si ha una distribuzione di probabilità della produzione

Ad esempio, per X=7, la produzione sarà un valore originato dalla estrazione casuale dalla distribuzione YIX=7



Le ipotesi : linearità

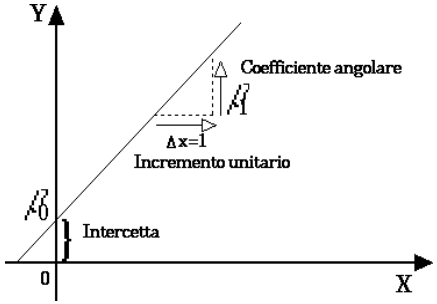
La relazione tra le due variabili è espressa da una retta

$$Y_i = \beta_0 + \beta_1 X_i + e_i;\quad i = 1,2,\ldots,n$$

"n" è il numero di coppie di valori considerati

C'è una componente di struttura espressa dalla retta intorno alla quale la casualità ed altre forze fanno oscillare i valori.

Ogni "e_i" è una variabile casuale e così lo è anche ogni "Y_i"



Ipotesi: relazione tra esogena ed errori

E' un punto controverso

il modello di regressione si regge sia nel caso che la "X" sia una variabile casuale o che assuma un numero limitato e fisso di valori.

In genere si ipotizza che i valori della "X" costituiscano un blocco fisso e noto di costanti numeriche per cui, l'errore -variabile casuale- ne è indipendente.

In alternativa, si ipotizza che la "X" sia una variabile casuale, ma in questo caso occorre l'ipotesi di incorrelazione:

$$Cov(X_i, e_j) = 0 \quad per \ \ ogni \ "i" \ \ e \ "j"$$

La scelta sarà dettata dal particolare contesto applicativo

Ipotesi: valore atteso degli errori nullo

$$E(e_i) = 0; \quad i = 1,2,\ldots,n$$

errori per difetto e per eccesso debbono compensarsi nell'ambito delle "n" coppie di valori

il valore atteso della endogena non è influenzato dal valore atteso degli errori:

$$E(Y_i) = E(\beta_0 + \beta_1 X_i + e_i) = E(\beta_0 + \beta_1 X_i) + E(e_i) = \beta_0 + \beta_1 X_i + 0$$
$$= \beta_0 + \beta_1 X_i; \quad i = 1,2,\ldots,n$$

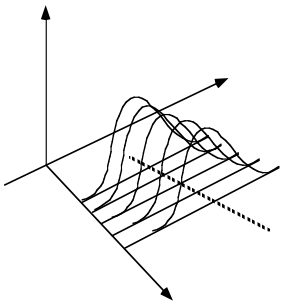
Non è necessario che la la media degli errori sia nulla, ma che sia costante rispetto all'indice "i".

$$Se \ \ E(e_i) = 3 \Rightarrow E(Y_i) = \beta_0 + \beta_1 X_i + 3 = \beta_0' + \beta_1 X_i; \ i = 1,2,\ldots,n$$

incognito è "β0" e incognito è β'0

Ipotesi: omoschedasticità

$$\sigma^2(e_i) = \sigma^2; \quad i = 1,2,\ldots,n$$



la variabilità dell'errore è la stessa per ogni osservazione.

Ciò implica che la variabilità delle distribuzione condizionate YIXi è la stessa per ogni "i".

$$Se \ \ \sigma^2(Y_i) = \sigma^2(\beta_0 + \beta_1 X_i + e_i)$$
$$= \sigma^2(\beta_0 + \beta_1 X_i) + \sigma^2(+e_i) + 2Cov(\beta_0 + \beta_1 X_i, e_i)$$
$$= 0 + \sigma^2(e_i) + 2 * 0 = \sigma^2$$

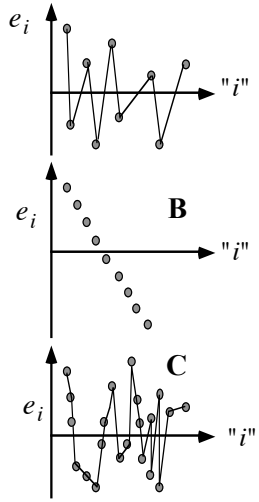
Perché il valore è fisso oppure fissato come condizionante per ogni "i"

ipotesi sulla incorrelazione tra esogena ed errori

Ipotesi: incorrelazione tra errori diversi

$$Cov(e_i, e_j) = 0 \quad per \ \ ogni \ \ i \neq j$$

non ci debbono essere strutture evidenti nell'andamento degli errori



ESEMPI DI CORRELAZIONE

A) Correlazione negativa tra errori successivi: un errore per eccesso è sempre seguit da un errore per difetto.

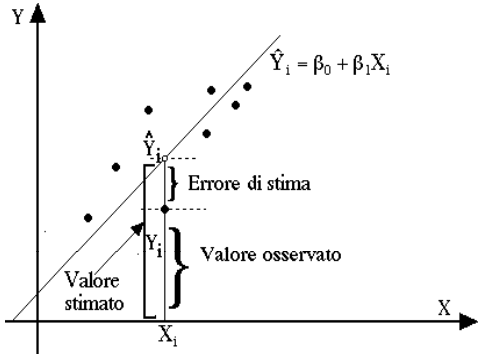
B) Correlazione positiva tra errori successivi: un errore per difetto o per eccesso è seguito da un errore con lo stesso segno.

C) In generale la serie degli errori deve essere "casuale" cioè se si escludono alcuni dei valori, comunque scelti, la struttura non diviene riconoscibile (ciò vale per "n" grande).

Stima dei parametri

Se per due punti passa una sola retta fra più di due punti non allineati ne passano infinite.

Ogni scelta determina degli errori dovuti alla sostituzione di un valore presunto o teorico ad un valore osservato



$\hat{y}_i$ = **valore stimato (ottenuto in base al modello)**

Occorre stabilire un criterio che ci permetta di scegliere quella che passa più vicino ai punti.

Criteri di calcolo

La vicinanza della retta è espressa con una sintesi degli scarti relativi tra valori osservati e valori stimati.

Poiché la somma dei valori è fissa il denominatore è ignorato

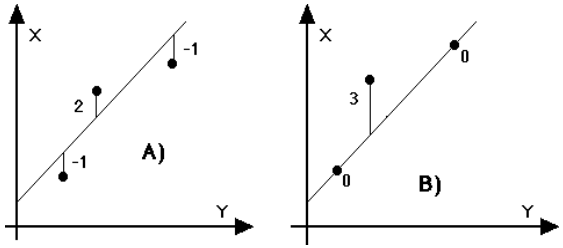
$$Q_r(\beta_0,\beta_1) = \frac{\sum_{i=1}^{n}|Y_i - \hat{Y}_i|^r}{\frac{\sum_{i=1}^{n}|Y_i|^r}{n}} = \frac{\sum_{i=1}^{n}|Y_i - \hat{Y}_i|^r}{\sum_{i=1}^{n}|Y_i|^r}$$

(di solito r=1 oppure r=2)

La scelta del criterio determina la scelta della retta.

Secondo Q_1 l'adattamento è migliore con la retta B.

E' il contrario secondo Q_2.



$Q_1(A) = |-1| + |+2| + |-1| = 4$

$Q_2(A) = 1^2 + 2^2 + 1^2 = 6$

$Q_1(B) = |0| + |+3| + |0| = 3$

$Q_2(B) = 0^2 + 3^2 + 0^2 = 9$

Scelta tra scarti assoluti e al quadrato

il criterio dei minimi assoluti -proposto dall'astronomo Boscovich e ripreso da Laplace- risale almeno al 1755.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_i) = 0$$ **Somma degli scarti negativi e somma degli scarti positivi uguali in valore assoluto**

$$\sum_{i=1}^{n}|y_i - \beta_0 - \beta_1 X_i|$$ **Minima rispetto ai parametri incogniti β_0, β_1**

Pro — **Gli scarti hanno lo stesso ordine di grandezza dei valori per cui il criterio risulta semplice e naturale**

Contro — **La soluzione è ottenuta con algoritmi di calcolo numerico.**

La soluzione non è necessariamente univoca (si pensi alla mediana per "n" pari se la retta migliore è parallela all'asse della "X")

Le trattazione delle proprietà statistiche è difficile e poco generale.

Scelta tra scarti assoluti e al quadrati/2

il criterio dei minimi quadrati risale a Legendre e Gauss.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_i) = 0$$ **Somma degli scarti negativi e somma degli scarti positivi uguali in valore assoluto**

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_i)^2$$ **Minima rispetto ai parametri incogniti β_0, β_1**

Pro — **Espressione univoca e semplice della soluzione. Trattazione chiara e rigorosa delle proprietà statistiche**

Contro — **Peso eccessivo agli scarti più grandi. Dati i due valori Y_1=12 e Y_2=8 ed ipotizziamo uno scarto del 20% in entrambi, si ottiene:**

$$\frac{(y_1 - \hat{y}_1)^2}{y_1} = \frac{(12-9)^2}{12} = 0.75; \quad \frac{|12-9|}{12} = 0.25$$

$$\frac{(y_2 - \hat{y}_2)^2}{y_2} = \frac{(8-6)^2}{8} = 0.50; \quad \frac{|8-6|}{8} = 0.25$$

Soluzione dei minimi quadrati

Partiamo dall'errore i-esimo:

$$(y_i - \hat{y}_i) = y_i - \beta_0 - \beta_1 X_i = y_i - \beta_0 - \beta_1 X_i \pm \bar{y} \pm \beta_1 \bar{x}$$
$$= (y_i - \bar{y}) + (\bar{y} - \beta_0 - \beta_1 \bar{x}) - \beta_1 (X_i - \bar{x})$$

che evidenzia il ruolo del punto $(\bar{x}, \bar{y})$ **come baricentro dei dati osservati**

Elevando al quadrato e sviluppando si ottiene:

$$(y_i - \hat{y}_i)^2 = \left[(y_i - \bar{y}) + (\bar{y} - \beta_0 - \beta_1 \bar{x}) - \beta_1 (x_i - \bar{x})\right]^2 = (y_i - \bar{y})^2 + (\bar{y} - \beta_0 - \beta_1 \bar{x})^2 + \beta_1^2 (x_i - \bar{x})^2 +$$
$$+ 2(y_i - \bar{y})(\bar{y} - \beta_0 - \beta_1 \bar{x}) - 2\beta_1 (y_i - \bar{y})(x_i - \bar{x}) - 2\beta_1 (\bar{y} - \beta_0 - \beta_1 \bar{x})(x_i - \bar{x})$$

Considerando la somma di tutti gli "n" termini e ricordando che la somma degli scarti dalla media aritmetica e nulla si arriva a:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} (\bar{y} - \beta_0 - \beta_1 \bar{x})^2 + \beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2\beta_1 \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

Soluzione dei minimi quadrati/2

Definiamo: $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2;\ S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2;\ S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y});$

Tali quantità sono note come devianze e codevianze

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = S_{yy} + n(\bar{y} - \beta_0 - \beta_1 \bar{x})^2 + \beta_1^2 S_{xx} - 2\beta_1 S_{xy}$$

Ne consegue:

$$= S_{yy} + n(\bar{y} - \beta_0 - \beta_1 \bar{x})^2 + \beta_1^2 S_{xx} - 2\beta_1 S_{xy} + \frac{S_{xy}^2}{S_{xx}} - \frac{S_{xy}^2}{S_{xx}}$$
$$= S_{yy} + n(\bar{y} - \beta_0 - \beta_1 \bar{x})^2 + S_{xx}\left[\beta_1^2 - 2\beta_1 \frac{S_{xy}}{S_{xx}} + \frac{S_{xy}^2}{S_{xx}}\right] - \frac{S_{xy}^2}{S_{xx}}$$
$$= S_{yy} + n(\bar{y} - \beta_0 - \beta_1 \bar{x})^2 + S_{xx}\left[\beta_1 - \frac{S_{xy}}{S_{xx}}\right]^2 - \frac{S_{xy}^2}{S_{xx}}$$

La somma degli errori dipende dalle incognite solo attraverso dei termini al quadrato per cui il minimo si ottiene azzerando quei termini e cioè

$$\hat{\beta}_1 = \frac{S_{yx}}{S_{xx}};\quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Stima della varianza degli errori

Oltre ai due parametri della retta esiste un'altro parametro incognito: σ^2

Le assunzioni del modello indicano gli errori "e" come v.c. non osservabili, possiamo però stimarne i valori con gli errori osservati

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad \text{per} \quad i = 1, 2, \ldots, n$$

Uno stimatore "naturale" di σ^2 **sarebbe allora la varianza degli errori stimati. In realtà si usa uno stimatore basato su questa, ma con correzione sul numero di osservazioni**

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

che è uno stimatore non distorto e consistente della varianza degli errori.

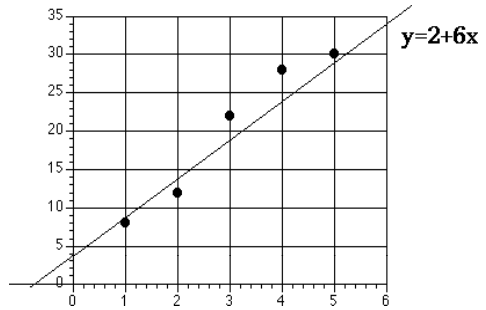
Ai fini del calcolo è facile mostrare che

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\mu}_y)^2 - \hat{\beta}_1 * \sum_{i=1}^{n} (y_i - \hat{\mu}_y)(x_i - \hat{\mu}_x)$$

Per cui il calcolo di s_e^2 **usa quantità già pronte**

Esempio

```
x<-1:5;y<-c(8,12,22,28,30)
Ols<-lm(y~x)
summary(Ols)
```




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0000     2.4221   0.826  0.46950
x             6.0000     0.7303   8.216  0.00377 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.309 on 3 degrees of freedom
Multiple R-squared: 0.9574, Adjusted R-squared: 0.9433
F-statistic:  67.5 on 1 and 3 DF,  p-value: 0.003774
```

Esercizio

Il prezzo e l'epoca dell'usato per una particolare auto è stato rilevato presso alcuni rivenditori

X	Y
2	10.5
5	3.2
1	11.7
6	4.8
4	7.3
5	3.6
3	8.8
2	9.9

1) Calcolare le stime dei parametri
2) Disegnare lo scatterplot e la retta teorica

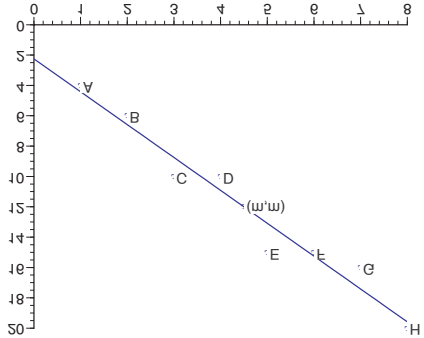
Proprietà della retta di regressione

La retta di regressione passa per il punto di coordinate $(\bar{x},\bar{y})$

La retta stimata può essere scritta come:

$$y = \bar{y} + \hat{\beta}_1(x - \bar{x}) \;\Rightarrow\; \bar{y} + \hat{\beta}_1(\bar{x} - \bar{x}) = \bar{y}$$

Non si tratta di un vincolo aggiuntivo, ma una caratteristica intrinseca al metodo dei minimi quadrati



Proprietà della retta di regressione/2

La somma degli scarti tra osservate e teoriche è nulla:

$$\sum_{i=1}^{n} y_i - \hat{y}_i = \sum_{i=1}^{n} y_i - \bar{y} - \hat{\beta}_1(x - \bar{x}) \;\Rightarrow\; \sum_{i=1}^{n}(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n}(x - \bar{x}) = 0 - \hat{\beta}_1 * 0 = 0$$

Ciò implica che Media osservate = Media teoriche

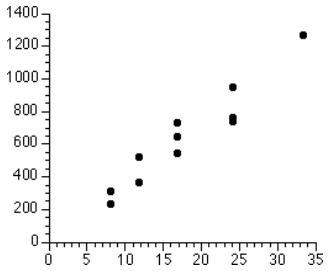
$$\frac{\sum_{i=1}^{n} \hat{y}_i}{n} = \frac{\sum_{i=1}^{n} \bar{y} + \hat{\beta}_1(x - \bar{x})}{n} = \frac{n\bar{y} + \hat{\beta}_1 \sum_{i=1}^{n}(\bar{x} - \bar{x})}{n} = \frac{n\bar{y} + 0}{n} = \bar{y}$$

Esempio

L'urbanista Palmira Morrone investiga la relazione tra flusso di traffico X (mgl di auto per 24 ore) ed il contenuto di piombo Y nella cortegga degli alberi che fiancheggiano una superstrada (peso a secco in *μ*g/g)

i	1	2	3	4	5	6	7	8	9	10	11
x_i	8.3	8.3	12.1	12.1	17.0	17.0	17.0	24.3	24.3	24.3	33.6
y_i	227	312	362	521	640	539	728	945	738	759	1263

a) Disegnare lo scatterplot;
b) Stimare i parametri;
c) Calcolare i valori teorici
d) Verificare che le proprietà indicate



$\hat{\beta}_1 = 36.18385$
$\hat{\beta}_0 = -12.84159$

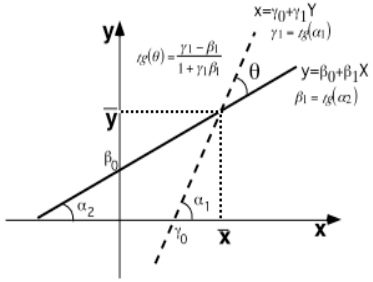
	y_i	$\hat{y}_i$	$\hat{e}_i$
	227	287.48	-60.48
	312	287.48	24.52
	362	424.98	-62.98
	521	424.98	96.02
	640	602.28	37.72
	539	602.28	-63.28
	728	602.28	125.72
	945	866.43	78.57
	738	866.43	-128.43
	759	866.43	-107.43
	1263	1202.94	60.06
MEDIE	639.4545	639.4545	0.0000

Proprietà della retta di regressione/3

il ruolo di esogena ed endogena può essere scambiato:

$$y_i = \beta_0 + \beta_1 x_i + e_i \Rightarrow \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad dove \begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} \end{cases}$$

$$x_i = \gamma_0 + \gamma_1 y_i + e_i' \Rightarrow \hat{x}_i = \hat{\gamma}_0 + \hat{\gamma}_1 y_i \quad dove \begin{cases} \hat{\gamma}_0 = \bar{x} - \hat{\gamma}_1 \bar{y} \\ \hat{\gamma}_1 = \dfrac{S_{xy}}{S_{yy}} \end{cases}$$



Le due rette interpolanti sono legate:

$$\hat{\gamma}_1 = \frac{S_{xy}}{S_{yy}} = \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} * \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} * r$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} * \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} * r$$

$$\hat{\gamma}_1 * \hat{\beta}_1 = r^2$$

$$tg(\theta) = \frac{r^2 - 1}{r}$$

i coefficienti angolari hanno sempre lo stesso segno per cui le due rette non sono mai perpendicolari

Le due rette sono parallele (e coincidenti) se e solo se Y=X dato che ora tg(θ)=0

Misura dell'adattamento

I minimi quadrati ci garantiscono il miglior adattamento possibile, ma questo potrebbe non essere abbastanza.

Dobbiamo trovare misure in grado di quantificare il grado di scostamento tra valori stimati e valori osservati.

Come protagonisti principali avremo

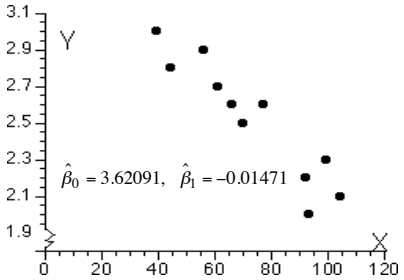
I valori osservati, I valori teorici e il numero delle osservazioni

I valori dei coefficienti

SQM degli errori

L'aziendalista Costantina Tenuta fa parte di una commissione chiamata a valutare una serie di progetti per l'idoneità al finanziamento. Per controllarne la congruità pone in relazione il numero X dei progetti per area e il tempo medio di completamento Y.

Settori destinatari	Progetti	Tempi medi di compl.
Edilizia demaniale	105	2.1
Opera stradali extraurbane	94	2.0
Disinquinamento	93	2.2
Ferrovie	78	2.6
Edilizia Sanitaria	71	2.5
Edilizia scolastica	67	2.6
Porti commerciali	62	2.7
Infrastrutture urbane	57	2.9
Energia	45	2.8
Smaltimento RSU	40	3.0
Ferrovie Metgropolitane	36	3.4
Archivi, Biblioteche	30	3.2
Ferrovie in concessione	12	3.3
Altri	100	2.3



$$s_e = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}} = 0.1479$$

E' nullo solo in caso di perfetta relazione lineare

Non varia però entro limiti predefiniti. Possiamo solo dire che un adattamento è peggiore o migliore di un altro, ma non se un dato adattamento è buono o no perché dipende dalla scala di misurazione delle variabili

Correlazione teoriche-osservate

Una possibilità di valutare l'adattamento potrebbe basarsi su:

$$\frac{Cov(y_i, \hat{y}_i)}{\sigma(y_i)\sigma(\hat{y}_i)} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})\left(\bar{y} + \hat{\beta}_1 (x_i - \bar{x}) - \bar{y}\right)}{\sqrt{S_{yy} \sum_{i=1}^{n} \left(\bar{y} + \hat{\beta}_1 (x_i - \bar{x}) - \bar{y}\right)^2}} = \frac{\hat{\beta}_1 S_{xy}}{\sqrt{S_{yy} \hat{\beta}_1^2 S_{xx}}} =$$

$$= \frac{\hat{\beta}_1}{|\hat{\beta}_1|} r$$

Ne consegue che l'adattamento è anche misurabile da:

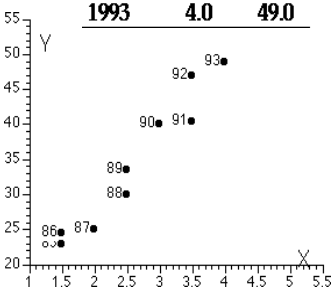
$$\left| \frac{Cov(y_i, \hat{y}_i)}{\sigma(y_i)\sigma(\hat{y}_i)} \right| = \left| \frac{\hat{\beta}_1}{|\hat{\beta}_1|} r \right| = |r|$$

cioè dal valore assoluto del coefficiente di correlazione tra osservate e stimate che coincide con il valore assoluto del coefficiente di correlazione "r" tra X ed Y.

Esempio

La dott.ssa Sarina Bonofiglio, analista finanziario, sta studiando la relazione tra X= Tasso medio sui prestiti nel sistema interbancario e Y=importo della cedola semestrale di un titolo obbligazionario.

Anni	TMPSI	CeSe.
1985	1.5	23.0
1986	1.5	24.5
1987	2.0	25.0
1988	2.5	30.0
1989	2.5	33.5
1990	3.0	40.0
1991	3.5	40.5
1992	3.5	47.0
1993	4.0	49.0



a) Disegnare lo scatterplot
b) Calcolare i parametri
c) Misurare l'adattamento con r(x,y)
d) Supponendo che il dato del 1993 sia non affidabile perché affetto dalla crisi nello SME calcolare il valore interpolato.
e) Quale sarà la cedola semestrale se nel 1994 il TMPSI arriva a 5.5?

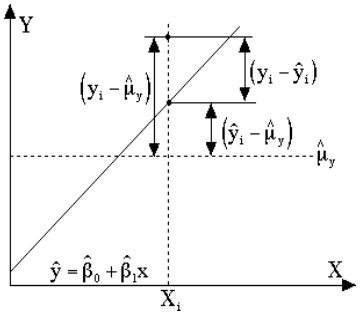
$$\hat{\beta}_0 = 6.448718;\ \hat{\beta}_1 = 10.602564;\ s_e^2 = 6.4803$$

$$r(x,y) = 0.9703$$

$$\hat{y}_{93} = 6.448718 + 10.602564 * 4.0 = 48.89$$

$$\hat{y}_{94} = 6.448718 + 10.602564 * 5.5 = 64.76$$

Coefficiente di determinazione (R^2)



La variabilità di "Y" può essere scomposta in due parti distinte. Infatti, l'identità

$$\sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n}\left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right]$$

rimane anche quando si considerano i quadrati (se la retta è quella dei minimi quadrati)

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}\left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right]^2 = \sum_{i=1}^{n}\left[(y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})\right] \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{y}_i - 2\bar{y}\sum_{i=1}^{n}(y_i - \hat{y}_i) \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\hat{\beta}_0\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) + 2\hat{\beta}_1\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)x_i \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\end{aligned}$$

Ancora sull'R^2

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

Dividendo per "n" si ha la seguente relazione:

Varianza totale=Varianza errori+Varianza stime

La varianza delle stime è la parte di variabilità (attitudine a presentare modalità diverse) che il nostro modello riesce a spiegare, quella degli errori è la parte che rimane ignota.

Varianza totale=Varianza NON spiegata+Varianza spiegata

Formula dell' R^2

Dividendo i membri per la devianza totale si ha

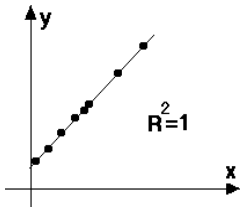
$$1 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} + \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

il 1° addendo è il rapporto tra varianza non spiegata e varianza totale, il 2° è il rapporto tra varianza spiegata e varianza totale.

Questo rapporto è usato come indice della bontà di adattamento ed è noto come il COEFFICIENTE DI DETERMINAZIONE

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

Casi estremi

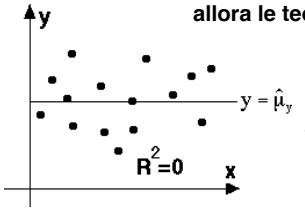


Se tutte le osservate sono allineate su di una retta, teoriche ed osservate coincidono e quindi

$$Se\ \ y_i = \hat{y}_i\ \ per\ \ ogni\ "i" \Rightarrow R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1;$$

$$\hat{y}_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x})$$

Se la retta di regressione è piatta (coefficiente angolare nullo) allora le teoriche sono tutte pari alla media e quindi



$$Se\ \ \hat{y}_i = \bar{y}\ \ per\ \ ogni\ "i" \Rightarrow R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - 1 = 0$$

Esempio

Studio della relazione tra il massimo del battito cardiaco sotto stress ed età

X	Y	$(x-M_x)$	$(y-M_y)$	$(x-M_x)(y-M_y)$	$(x-M_x)^2$	$(y-M_y)^2$	y_i	$(y-y_i)^2$
10	210	-25	25	-625	625	625	212.2058	4.8656
20	200	-15	15	-225	225	225	201.3234	1.7514
25	195	-10	10	-100	100	100	195.8822	0.7783
35	190	0	5	0	0	25	184.9998	25.0020
40	185	5	0	0	25	0	179.5586	29.6088
45	175	10	-10	-100	100	100	174.1174	0.7790
50	165	15	-20	-300	225	400	168.6762	13.5144
55	160	20	-25	-500	400	625	163.2350	10.4652
35	185			-1850	1700	2100		86.7647

$$s_e = \sqrt{\frac{86.7647}{6}} = 3.8027$$

$$\beta_0 = 223.0812;\quad \beta_1 = 1.0882;$$

$$R^2 = 1 - \frac{86.7647}{2100} = 0.9572$$

$$r(y,\hat{y}) = \left|\frac{-1850}{\sqrt{1700 * 2100}}\right| = 0.9791$$

$$R^2 = (-0.9791)^2 = 0.9587$$

Esempio

Reddito	Superficie
22	106
26	117
45	128
37	132
28	80
50	95
56	168
34	65
60	150
40	120
45	110
36	45
80	205
120	230

SUMMARY OUTPUT

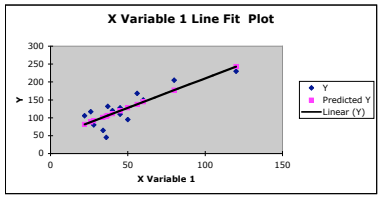
Regression Statistics	
Multiple R	0.832286982
R Square	0.692701621
Adjusted R Sq	0.667093423
Standard Erro	29.31437326
Observations	14

	Beta	*SE*	*t Stat*	*P-value*
Intercept	44.9592	17.2813	2.6016	0.0232
X Variable 1	1.6518	0.3176	5.2010	0.0002

Si supponga che la proprietaria di un'agenzia immobiliare voglia stabilire la relazione tra reddito familiare e superficie di un appartamento.

RESIDUAL OUTPUT

Observation	*Predicted Y*	*Residuals*
1	81.2988	24.7012
2	87.9060	29.0940
3	119.2901	8.7099
4	106.0757	25.9243
5	91.2096	-11.2096
6	127.5491	-32.5491
7	137.4599	30.5401
8	101.1203	-36.1203
9	144.0671	5.9329
10	111.0311	8.9689
11	119.2901	-9.2901
12	104.4239	-59.4239
13	177.1031	27.8969
14	243.1750438	-13.17504381

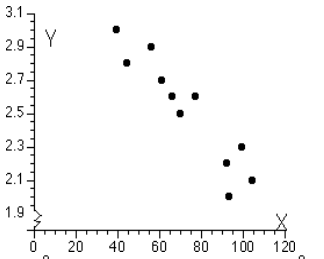


Esercizio

L'aziendalista Costantina Tenuta fa parte di una commissione chiamata a valutare una serie di progetti per l'idoneità al finanziamento. Per controllarne la congruità pone in relazione il numero X dei progetti per area e il tempo medio di completamento Y.

Settori di destinazione	Proget.	TMC
Edilizia demaniale	105	2.1
Altri	100	2.3
Opere straddali extraurb	94	2.0
Disinquinamento	93	2.2
Ferrovie	78	2.6
Edilizia sanitaria	71	2.5
Edilizia scolastica	67	2.6
Porti commerciali	62	2.7
Infrastrutture urbane	57	2.9
Energia	45	2.8
Smaltimento R.S.U.	40	3.0
Ferrovie metropol.	36	3.4
Archivi, Biblioteche	30	32.2
Ferrocie in concessione	12	3.3

a) Disegnare lo scatterplot; b) Stimare i parametri; c) stimare la varianza degli errori



$$\sum y_i = 37.6;\quad \sum y_i^2 = 103.54;\quad \sum x_i = 890;\quad \sum x_i^2 = 67.182;$$
$$\sum x_i y_i = 2234.30$$

$$\hat{\beta}_1 = -0.0147109;\quad \hat{\beta}_0 = 3.6209072$$

$$s_e^2 = \frac{103.54 - 3.6209072 * 37.6 - (-0.0147109) * 2234.30}{14 - 2} = \frac{0.2624532}{12} = 0.0219$$

Proprietà degli stimatori ai M.Q.

Le formule delle stime ottenute con il metodo dei minimi quadrati

$$\begin{cases}\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \\ \hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}\end{cases}$$

evidenziano la dipendenza delle espressioni dalle osservazioni campionarie "y".

Per accertare le proprietà degli stimatori dei parametri conviene ricordare che

$$S_{xy} = \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n} y_i(x_i-\bar{x}) - \bar{y}\sum_{i=1}^{n}(y_i-\bar{y}) = \sum_{i=1}^{n} y_i(x_i-\bar{x})$$

$$S_{xx} = \sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x}) = \sum_{i=1}^{n} x_i(x_i-\bar{x}) - \bar{x}\sum_{i=1}^{n}(x_i-\bar{x}) = \sum_{i=1}^{n} x_i(x_i-\bar{x})$$

che discendono entrambe dal fatto che la media aritmetica rende nulla la somma degli scarti.

Centramento

Gli stimatori ottenuti con i minimi quadrati sono centrati

Coefficiente angolare

$$E(\hat{\beta}_1) = E\left[\frac{\sum_{i=1}^{n} y_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] = \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}(x_i-\bar{x})E(y_i) = \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}(x_i-\bar{x})E(\beta_0+\beta_1 x_i+e_i) =$$

$$= \frac{\beta_0\sum_{i=1}^{n}(x_i-\bar{x}) + \beta_1\sum_{i=1}^{n}(x_i-\bar{x})x_i + \sum_{i=1}^{n}(x_i-\bar{x})0}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{0+\beta_1\sum_{i=1}^{n}(x_i-\bar{x})^2+0}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \beta_1$$

Intercetta

$$E(\hat{\beta}_0) = E\left[\bar{y}-\hat{\beta}_1\bar{x}\right] = \beta_0+\beta_1\bar{x}-\bar{x}E(\hat{\beta}) = \beta_0+\beta_1\bar{x}-\bar{x}\beta_1 = \beta_0$$

Consistenza

All'aumentare di n la varianza della y e della x raggiungono un <u>limite definito</u> (sorte benigna) per cui la varianza degli stimatore tende a zero.

Coefficiente angolare

$$\sigma^2(\hat{\beta}_1) = \sigma^2\left[\frac{\sum_{i=1}^{n} y_i(x_i-\bar{x})}{S_{xx}}\right] = \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\sigma^2(y_i) = \frac{\sigma^2}{S_{xx}^2}S_{xx} = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{n}*\frac{1}{\sigma^2(x)}$$

Intercetta

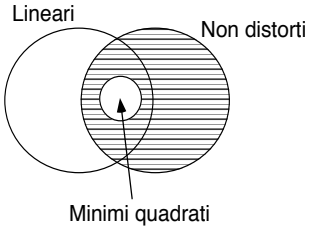
$$\sigma^2(\hat{\beta}_0) = \sigma^2\left[\bar{y}-\hat{\beta}_1\bar{x}\right] = \frac{\sigma^2}{n}+\bar{x}^2\sigma^2(\hat{\beta}_1) - 2Cov(\bar{y},\hat{\beta}) = \frac{\sigma^2}{n}+\bar{x}^2\frac{\sigma^2}{S_{xx}} - 2*0 = \frac{\sigma^2}{n}\left[1+\frac{\bar{x}^2}{\sigma^2(x)}\right]$$

La covarianza è nulla dato che $E\left[(\bar{y}-\beta_0-\beta_1\bar{x})(\hat{\beta}_1-\beta_1)\right] = E\left[(0)*(\hat{\beta}_1-\beta_1)\right] = 0$

Teorema di Gauss-Markov

E' la principale giustificazione teorica del criterio dei minimi quadrati

Nell'ambito degli stimatori lineari (cioè combinazioni lineari delle variabili casuali campionarie) non distorti, quelli ottenuti con i minimi quadrati (inclusa la stima della varianza degli errori) hanno la varianza inferiore.



La proprietà della varianza minima non richiede la normalità.

Possono però esistere stimatori non lineari o non distorti con varianza inferiore a quella dei minimi quadrati

Si ricorda che la varianza più piccola forniscono -a parità di ampiezza campionaria- intervalli di confidenza più corti cioè più accurati

Test nel modello di regressione

Per espletare l'inferenza nel modello di regressione lineare sono necesarie due ipotesi alternative

- **"n" è abbastanza grande da attivare il teorema limite centrale**
- $e_i \sim N(0;\sigma^2)$

  gli errori sono gaussiani (e quindi indipendenti oltreché incorrelati)

Queste condizioni permettono di stabilire quale sia la distribuzione degli stimatori in quanto funzioni di variabili casuali.

In particolare, si dimostra che $\hat{\beta}_0,\ \hat{\beta}_1$ sono stimatori di massima verosimiglianza con annesse proprietà e difetti.

Gaussianità

Poiché usiamo variabili casuali indipendenti ed identicamente distribuite l'aggiunta della condizione

$$y_i \sim N(\beta_0 + \beta_1 x_i; \sigma^2); \quad i = 1,2,\ldots,n$$

porterà -grazie alla riproducibilità- alla distribuzione gaussiana

$$\hat{\beta}_0 \sim N\left[\beta_0, \frac{\sigma^2}{n}\left(1 + \frac{n\bar{x}^2}{S_{xx}}\right)\right]; \qquad \hat{\beta}_1 \sim N\left[\beta_1, \frac{\sigma^2}{S_{xx}}\right]$$

Inoltre, $$\frac{(n-2)s_e^2}{\sigma^2} \sim \chi^2(n-2)$$

L'ipotesi di gaussianità (diretta delle y oppure ottenuta grazie ad "n" grande ed al conseguente limite centrale) è indispensabile per la validità delle procedure inferenziali relativamente ai parametri del modello di regressione.

Test sul coefficiente angolare

La prima è più importante verifica riguarda l'esistenza o meno di una relazione tra la Y e la X

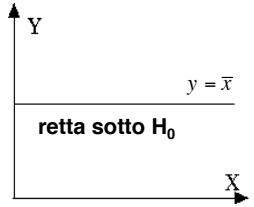
ESISTE O NON ESISTE UNA RELAZIONE TRA ENDOGENA ED ESOGENA?

A questa domanda rispondiamo solo PARZIALMENTE: la "Y" varia o non varia LINEARMENTE con la "X"?.

Questo si traduce nella verifica dell'ipotesi $\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$

infatti, β_1 esprime la variazione media nella endogena a partire da una variazione unitaria nella esogena

Se H0 non potesse essere rifiutata la retta di regressione si presenterebbe come parallela all'asse delle X e non in grado di spiegare nulla della "Y"



Continua test su β_1

La statistica test necessaria per la verifica si ottiene come per i test sulla media. In particolare si coinvolge la media e la varianza dello stimatore

$$E(\hat{\beta}_1) = \beta_1; \quad \sigma^2(\hat{\beta}_1) = \frac{\hat{s}_e^2}{SS_{xx}} \qquad \text{dove } SS_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

in un rapporto t di Student $$t_c = \frac{\hat{\beta}_1 - (\beta_1 \text{ sotto } H_0)}{\frac{s_e}{\sqrt{SS_{xx}}}} \Rightarrow \frac{\hat{\beta}_1}{\frac{s_e}{\sqrt{SS_{xx}}}}$$

Le regioni di rifiuto, per le varie tipologie di ipotesi alternativa, sono

$$H_1 = \beta_1 < 0 \Rightarrow t_c \leq -t_{\alpha,n-2}; \quad H_1 = \beta_1 > 0 \Rightarrow t_c \geq t_{\alpha,n-2}; \quad H_1 = \beta_1 \neq 0 \Rightarrow |t_c| \geq t_{\frac{\alpha}{2},n-2};$$

in cui i gradi di libertà corrispondono al divisore di s_e^2

Se si applica il T.L.C. si cambieranno le soglie critiche con i livelli della distribuzione gaussiana

Esempio

Si ritiene che il consumo di energia elettrica sia determinato da un modello lineare nella temperatura media del giorno

giorno	T.M.	C.E.L.
1	95	214
2	82	152
3	90	156
4	81	129
5	99	254
6	100	266
7	93	210
8	95	204
9	93	213
10	87	150

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = 180709 - \frac{915*1948}{10} = 2556$$

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \hat{\mu}_x)^2 = \sum_{i=1}^{n} x_i - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 84103 - \frac{(915)^2}{10} = 380.5$$

$$\hat{\mu}_y = 194.8;\quad \hat{\mu}_x = 91.5;\quad \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{2556}{380.5} = 6.7175;$$

$$R^2 = 0.99$$

$$\hat{\beta}_0 = \hat{\mu}_y - \hat{\beta}_1\hat{\mu}_x = 194.8 - (6.7175)*91.5 = -419.85$$

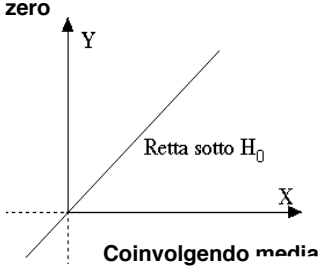
$$s_e^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{2093.43}{8} = 261.68 \Rightarrow s_e = \sqrt{261.68} = 16.18 \qquad t_c = \frac{6.7175}{16.18/\sqrt{380.5}} = 8.10$$

Posto α=5% si ha $\text{Bidirezionale}: t_{0.05,8} = 2.306;\quad \text{Unidirezionale}: t_{0.025,8} = 1.86$

al 5% si rifiuta l'ipotesi nulla. La si rifiuta anche all'1‰

Test sull'intercetta

La verifica dell'intercetta è poco interessante dato che non ha incidenza sulla bontà di adattamento. In genere si sottopone a verifica l'ipotesi che sia uguale a zero



$$\begin{cases} H_0: \beta_0 = 0 \\ H_1: \beta_0 \neq 0 \end{cases}$$

se NON si rifiuta si intende che la retta ha equazione

$$y_i = \beta_1 x_i + e_i$$

cioè passa per l'origine stabilendo una proporzionalità diretta tra la "Y" e la "X"

Coinvolgendo media e varianza dello stimatore ai minimi quadrati in un rapporto t di Student:

$$\sigma\left(\hat{\beta}_0\right) = \frac{\sigma}{\sqrt{n}}\sqrt{\left(1 + \frac{n\bar{x}^2}{S_{xx}}\right)} \qquad t_c = \frac{\hat{\beta}_0}{\sigma\left(\hat{\beta}_0\right)}$$

si applicano le stesse regioni di rifiuto usata per il coefficiente angolare

Nel caso di temperature e consumo di energia si ha: $t_c = \frac{-419.85}{16.18*\sqrt{\frac{84103}{380.5}}} = \frac{-419.85}{240.551} = -1.745$

L'ipotesi si rifiuta al 10%, ma si accetta al 20%

Esercizio

Si supponga che la proprietaria di un'agenzia immobiliare voglia stabilire la relazione tra reddito familiare e superficie dell'appartamento. A questo fine considera un campione casuale di n=10 clienti

Re ddito	22	26	45	37	28	50	56	34	60	40
Superficie	16	17	26	24	22	21	32	18	30	20

$$SS_{xx} = 1489.6$$

$$SS_{yy} = 262.4$$

$$SS_{xy} = 527.2$$

$$\hat{\mu}_x = 39.8$$

$$\hat{\mu}_y = 22.6$$

a) $\hat{\beta}_1 = \frac{527.2}{1489.6} = 0.354;\quad \hat{\beta}_0 = 22.6 - 0.354*39.8 = 8.51$

b) $s_e^2 = \frac{262.4 - \frac{(527.2)^2}{1489.6}}{8} = 9.477$

c) $R^2 = 1 - \frac{75.81}{262.4} = 0.71$

d) $t\left(\hat{\beta}_1\right) = \frac{0.354}{3.078/\sqrt{1489.6}} = 4.44;\quad t\left(\hat{\beta}_0\right) = \frac{8.51}{3.078*\sqrt{\frac{1489.6 + 10*39.8}{1489.6}}} = 2.678$

per g=8 il coefficiente angolare è significativo almeno all'1%. L'intercetta solo al 5%

Analisi dei residui

Uno dei modi più efficaci per valutare l'adattamento del modello di regressione è l'analisi grafica dei residui stimati

$$\hat{e}_i = y_i - \hat{y}_i = (y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x})$$

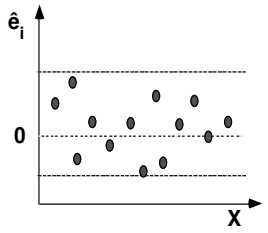
Se le ipotesi del modello sono corrette i residui hanno

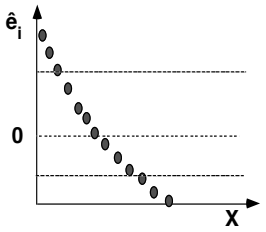
- **Media nulla (questo è vero con le stime ai minimi quadrati)**
- **Avere varianza finita (che non aumenta con "y" o con "e").**
- **Non avere una struttura riconoscibile.**

L'atteggiamento è prudenziale: l'analisi dei residui -basata anche su valutzioni grafiche- deve segnalare manifeste incongruenze o condraddizioni conglamate senza che il vaglio di queste analisi sia una prova definitiva della validità del modello

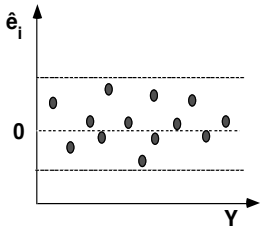
Analisi dei residui/2

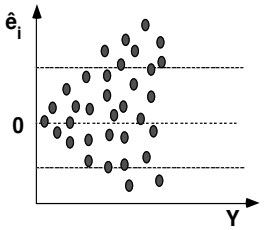
In particolare si usano due scatterplot: (e,x) e (e,y)





E' violata l'ipotesi di incorrelazione tra esogena e residui
Ciò però è impossibile nel modello lineare semplice per le proprietà della retta ai minimi quadrati

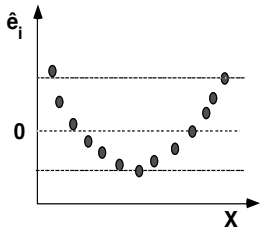




E' violata l'ipotesi di omoschedasticità e.o di varianza finita delgi errori

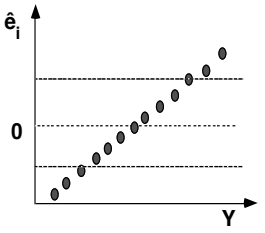
Analisi dei residui/3

L'analisi dei residui evidenzia inoltre alcune possibili manchevolezze



I dati sembrano sono in realtà disposti secondo un arco di parabola che la retta non riesce a cogliere.

I rsidui hanno una precisa struttura che indica anche il tipo di correzione

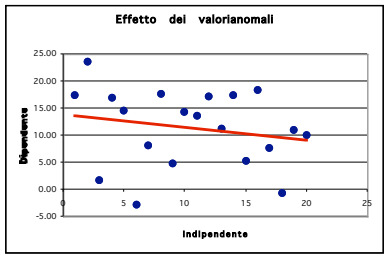


L'andamento dei residui evidenza una forte correlazioen con la endogena.

Ciò implica che l'esogena introdotto non ha spiegato quasi nulla

Effetto dei valori anomali

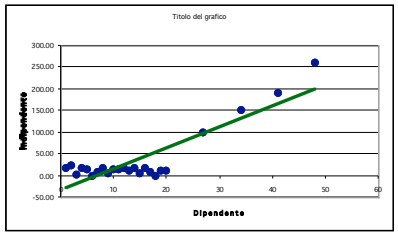
X	Y
1	17.27
2	23.49
3	1.72
4	16.87
5	14.58
6	-2.76
7	8.04
8	17.65
9	4.75
10	14.39
11	13.61
12	17.22
13	11.22
14	17.39
15	5.31
16	18.39
17	7.64
18	-0.73
19	10.86
20	9.89
27	**100.00**
34	**150.00**
41	**190.00**
48	**260.00**



L'impatto della "X" può essere fuorviato dalla presenza di alcuni valori remoti.

R multiplo	0.2064
R al quadrato	0.0426
R al quadrato	-0.0106
Errore standa	7.0803
Osservazioni	20

	Beta	*Stat t*	*p-value*
Intercetta	13.9197	4.2322	0.0005
Variabile X 1	-0.2457	-0.8950	0.3826



I minimi quadrati trascurano il blocco di 20 dati tra i quali non c'è relazione significativa (o è negativa). Invece, pone attenzione ai quattro punti tra i quali c'è una relazione positiva

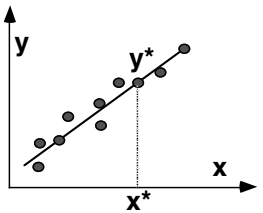
R multiplo	0.8795
R al quadrato	0.7735
R al quadrato	0.7632
Errore standa	32.7075
Osservazioni	24

	beta	*Stat t*	*p-value*
Intercetta	-34.9737	-3.2383	0.0038
Variabile X 1	4.9060	8.6687	0.0000

Le previsioni

Sia X* un valore qualsiasi della variabile indipendente X. Ad esso corrisponde il valore stimato:

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$



che è uno stimatore non distorto

Infatti:

$$E\left(\hat{y}^* - y^*\right) = E\left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right) - E(y^*) = \beta_0 + \beta_1 x^* - \beta_0 - \beta_1 x^* = 0$$

L'errore di previsione è quindi nullo -in media- Ignoriamo però cosa succede di volta in volta nella singola occasione.

Le previsioni/2

il valore previsto y* può essere considerato:

- **Una stima puntuale del valore atteso della y dato che la x è pari a x* (qui è vista come un parametro incognito)** $y^* = E(y|x = x^*)$

- **La previsione del valore della y che corrisponderà ad x=x* (qui è il valore possibile di una variabile)**

In entrambi i casi occorre superare la stima puntuale ed arrivare a quella intervallare

Poiché y* dipende dagli stimatori $\hat{\beta}_0$ e $\hat{\beta}_1$ **è essa stessa una variabile casuale e quindi dotata di variabilità campionaria che deve essere stimata.**

Lo stimatore dipende dalla particolare angolare adottata.

La variabilità dei valori -y* parametro

è la tendenza a variare della stima qualora il calcolo sia ripetuto più volte aggiungendo nuovi dati

$$s(y^*) = s_e * \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \qquad \textbf{dove:} \qquad s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}; \qquad S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Poiché y* è una funzione lineare di v.c. normali è essa stessa normale :

$$(\hat{\beta}_0 + \hat{\beta}_1 x^*) - t_{\frac{\alpha}{2};n-2}\, s(y^*) \le (\beta_0 + \beta_1 x^*) \le (\hat{\beta}_0 + \hat{\beta}_1 x^*) + t_{\frac{\alpha}{2};n-2}\, s(y^*)$$

è un intervallo di confidenza al livello di significatività dell' (1-α)% per il parametro: valore atteso della variabile condizionata quando x=x*

Esempio

Si abbiano i seguenti dati

x	1000	1100	1200	1250	1300	1400	1450
y	220	280	350	375	450	470	500

$\sum x_i y_i = 3384750; \quad \hat{\beta}_0 = -417.70;$

$\hat{\beta}_1 = 0.640; \quad s_e = 17.62; \quad \bar{x} = 1242.86$

Calcoliamo l'errore standard della previsione per x*=1200

$$s(\hat{\beta}_0 + \hat{\beta}_1 1200) = 17.62\sqrt{0.1429 + 0.0121} = 6.937$$

La soglia della "t" di Student per α=0.005 e g=5 è $t_{\frac{0.005}{2};5} = 4.032$

Ne consegue che l'intervallo di interpolazione è

$$-417.70 + 0.640 * 1200 \pm 4.032 * 6.937$$

$$322.33 \le y^* \le 378.27$$

Con un livello di fiducia estremamente alto asseriamo che intervalli di questo tipo
tenderanno a racchiudere il valore incognito y*

La varianza dei valori -y* variabile

Se si vuole ottenere la stima del valore della "y" quando la x raggiunge il livello x=x* si usa una procedura diversa.

il valore è una variabile casuale per cui non parleremo di intervallo di confidenza, ma di intervallo di previsione

L'errore di previsione è dato da $(\beta_0 + \beta_1 x^*) - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$ **che ora è visto come differenza tra due variabili casuali.**

C'è maggiore incertezza e l'intervallo sarà più ampio

$$\mathrm{Var}\left[(\beta_0 + \beta_1 x^*) - (\hat{\beta}_0 + \hat{\beta}_1 x^*)\right] = \mathrm{var}\left[(\beta_0 + \beta_1 x^*)\right] + \mathrm{var}\left[(\hat{\beta}_0 + \hat{\beta}_1 x^*)\right] = s_e^2 + s_e^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$$

$$= s_e^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$$

La covarianza è nulla dato che y* è indipendente dalle altre y

La precisione dei valori -y* variabile/2

il valore atteso dell'errore di previsione è zero, per cui la T: ha distribuzione "t" di student con (n-2) gradi di libertà

$$T = \frac{(\beta_0 + \beta_1 x^*) - (\hat{\beta}_0 + \hat{\beta}_1 x^*)}{s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}$$

il termine seguente: $(x^* - \bar{x})^2$

evidenzia la dipendenza dell'errore di previsione dallo scentramento del valore della x* per cui si vuole la previsione:

Maggiore è la lontananza del valore fissato di x*, più grande sarà l'errore di previsione.

Esercizio

x	398	292	352	575	568	450	550	408	484	350	503	600	600
y	0.15	0.05	0.23	0.43	0.23	0.40	0.44	0.44	0.45	0.09	0.59	0.63	0.60

$\sum x_i = 6130;\ \sum x_i^2 = 3022050;\ \sum y_i = 4.73;\ \sum y_i^2 = 2.1785;\ \sum x_i y_i = 2418.74$

$\hat{\beta}_1 = 0.00143;\ \hat{\beta}_0 = -0.311;\ s_e = 0.131$

Se il valore futuro della X dovesse attestarsi al valore X=500 allora l'intervallo di previsione al 5% per la Y sarà dato da

$$-0.311 + 0.00143 * 500 \pm 2.201 * 0.131 * \sqrt{1 + \frac{1}{13} * \frac{13 * (500 - 471.54)^2}{1709750}} \qquad t_{\frac{0.05}{2},11} = 2.201$$

$0.10 \le y \le 0.70$

Un intervallo molto ampio dovuto alla forte variabilità dei residui (cioè scarso adattamento)

$$\frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n(x^* - \bar{x})^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Funzioni lineari

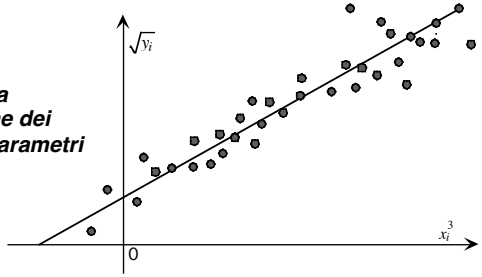
La linearità del modello di regressione è legata solo al modo in cui compaiono i parametri e non alle variabili.

In questo senso i modelli

$$\sqrt{y_i} = \beta_0 + \beta_1 x_i^3 + u_i \qquad e^{y_i} = \beta_0 + \beta_1 Ln(|x_i - 7|) + e_i$$

sono lineari dato che i parametri vi compaiono direttamente e con potenza uno.

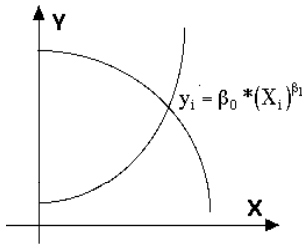
Quello che si riporta sugli assi ha importanza per l'interpretazione dei risultati, non per la stima dei parametri

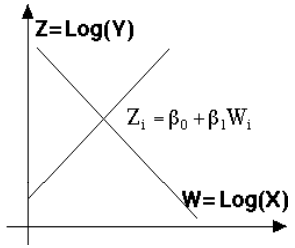


Funzioni linearizzabili

Il modello di regressione si estende alle forme analitiche LINEARIZZABILI

Si tratta di espressioni che diventano lineari con opportune trasformazioni.





Il modello è linearizzabile se modificando opportunamente gli assi, la relazione appare lineare

Modelli intrinsecamente lineari

Il modello è è tale se modificando opportunamente gli assi, la relazione appare lineare, ma non nei parametri originali

ERRORI ADDITIVI

$$y_i = e^a + b^2 x_i + u_i \Rightarrow y_i = \beta_0 + \beta_1 x_i + u_i \quad \beta_0 = e^a, \ \beta_1 = b^2$$

ERRORI MOLTIPLICATIVI

$$a(z_i)^b = c(w_i)^b e_i \Rightarrow Ln(a) + bLn(z_i) = Ln(c) + dLn(w_i) + Ln(e_i)$$

$$\Rightarrow y_i = \beta_0 + \beta_1 x_i + u_i \quad \beta_0 = \frac{Ln(c/a)}{b}, \ \beta_1 = d/b$$

Da notare che per errori moltiplicativi si deve in genere anche ipotizzare che

$$e_i > 0; \ E(e_i) = 1$$

Altrimenti sarebbe impossibile la linearizzazione

Esempio

La relazione tra percentuali cumulate di redditi Q_i e percentuali cumulate di redditieri P_i può essere rappresentata dalla curva di Lorenz

$$Q_i = P_i^a (2 - P_i)^b e_i$$

Determinare la stima dei parametri

La forma analitica è linearizzabile con la trasformazione seguente:

$$\frac{Ln(Q_i)}{Ln(P_i)} = a + b\frac{Ln(2-P_i)}{Ln(P_i)} + U_i \qquad Y_i = \frac{Ln(Q_i)}{Ln(P_i)}; \quad X_i = \frac{Ln(2-P_i)}{Ln(P_i)}; \quad U_i = \frac{Ln(E_i)}{Ln(P_i)}$$

P_i	Q_i	Y_i	X_i	$(Y_i - M_y)$	$(X_i - M_x)$	$C(x,y)$	$(X_i - M_x)^2$
0.1	0.0270	1.5686	-0.2788	-0.3170	0.3099	-0.0982	0.0777
0.2	0.0762	1.5996	-0.3652	-0.2860	0.2235	-0.0639	0.1334
0.3	0.1354	1.6608	-0.4407	-0.2248	0.1480	-0.0333	0.1942
0.4	0.2059	1.7247	-0.5129	-0.1609	0.0758	-0.0122	0.2631
0.5	0.2874	1.7989	-0.5850	-0.0867	0.0037	-0.0003	0.3422
0.6	0.3809	1.8895	-0.6587	0.0039	-0.0700	-0.0003	0.4339
0.7	0.4891	2.0052	-0.7356	0.1196	-0.1469	-0.0176	0.5411
0.8	0.6147	2.1808	-0.8171	0.2952	-0.2284	-0.0674	0.6676
0.9	0.7650	2.5425	-0.9046	0.6569	-0.3159	-0.2075	0.8183
		16.9705	-5.2985	0.0000	0.0000	-0.5007	3.4715

$M_x = 1.8856$

$M_y = -0.5887$

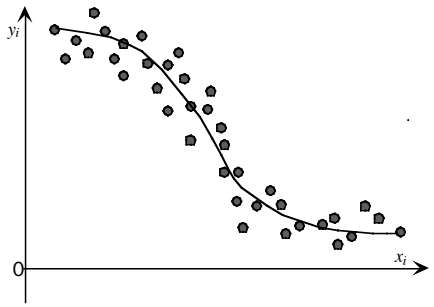
$a' = 1.8001$

$b' = -0.1442$

Modelli non lineari

I modelli si dicono NON LINEARI se in nessun modo è possibile ricondurli ad una forma lineare diretta o intrinseca nei parametri

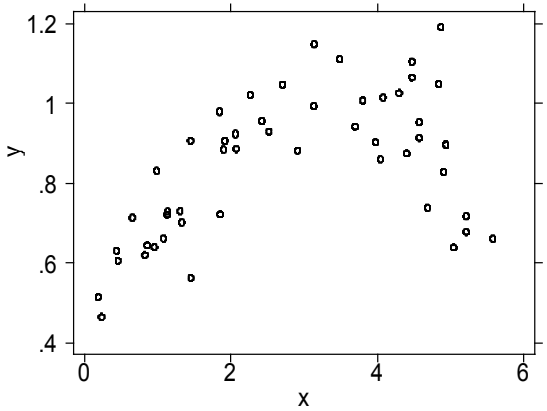
$$y_i = \beta_0 (\beta_1)^{x_i} + e_i$$



In questo caso la stima dei parametri avviene con procedure di ottimizzazione

Non sono semplici da utilizzazre, ma diventa sempre più facile utilizzarle

Relazione non lineare



La legge Yerkes-Dodson descrive il legame ad "U rovesciata" tra l'intensità dello stimolo e la qualità attesa della performance.

Regressione per serie evolutive

La situazione è quella di un fenomeno che segua un ordinamento unidimensionale, il cui valore attuale dipende essenzialmente da quelli accaduti in precedenza.

- **L'indice MIB**
- **Prospezione verticale di un terreno**
- **Spese alimentari**

Se t=1,2,...,n è l'indice che individua i vari punti nei quali il fenomeno viene rilevato, la regressione per serie evolutive avrà espressione:

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t$$

dove Y_{t-1} è la cosiddetta variabile "ritardata di lag 1".

Regressione per serie evolutive/2

Si distinguono due situazioni:

STATICA COMPARATA

Le osservazioni sulla endogena e sulla esogena, sono relative allo stesso fenomeno ma rilevato su diverse unità in epoche diverse

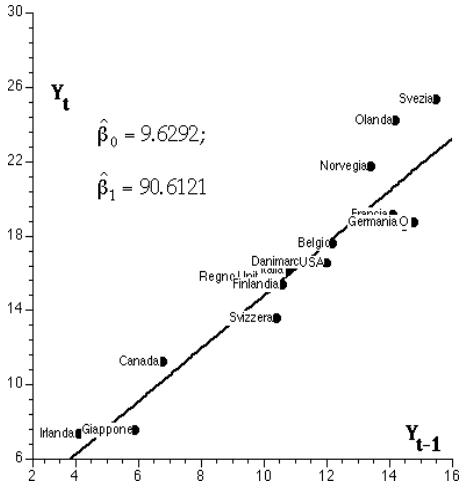
AUTOREGRESSIONE

La variabile esogena é data, per ogni osservazione (cioé in relazione dinamica), dalla endogena ritardata

Esempio: statica comparata

A partire dai dati sulla pressione fiscale in due epoche diverse e per vari paesi "occidentali" determinate i parametri della retta di regressione

Paesi	Indip. Y_{t-1}	Dip. Y_t
Irlanda	4.10	7.30
Olanda	14.20	24.20
Canada	6.80	11.20
Svezia	15.50	25.30
Norvegia	13.40	21.70
Regno Unito	10.10	15.70
Italia	10.80	16.10
Danimarca	11.50	16.60
Finlandia	10.60	15.30
Belgio	12.20	17.60
USA	12.00	16.50
Francia	14.10	19.10
Svizzera	10.40	13.50
Austria	14.50	18.50
Giappone	5.90	7.50
Germania O	14.80	18.70



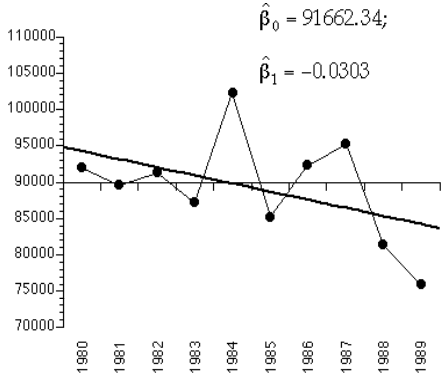
Esempio: autoregressione

Nell'esempio precedente una stessa variabile è osservata in due tempi diversi per le medesime unità. Lo stesso modello può essere applicato in situazioni in cui il valore della dipendente al tempo "t" è legato linearmente al valore della stessa dipendente al tempo "t-1"

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + e_t$$

Produzione di frumento

Anni	Indip. Y_{t-1}	Dip. Y_t
1980	91952	89590
1981	89590	91241
1982	91241	87174
1983	87174	102269
1984	102269	85171
1985	85171	92287
1986	92287	95205
1987	95205	81412
1988	81412	75882



Analisi del trend

il TREND è il sentiero predefinito che si immagina il fenomeno tenda a seguire a meno di piccoli ed incontrollabili errori. Inoltre, se spostato dal trend, tende a ritornarci

Un dato fenomeno è osservato periodicamente e si ipotizza che l'intensità rilevata dipenda proprio dal momento di osservazione

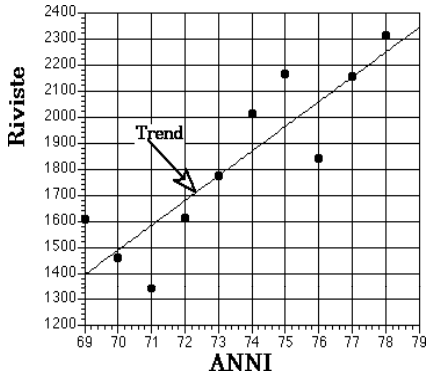
$$Y_t = \beta_0 + \beta_1 t + e_t$$

Pubblicazioni di riviste

Anni t	Riviste Y_t
69	1603
70	1455
71	1338
72	1605
73	1768
74	2005
75	2160
76	1836
77	2152
78	2309

$\hat{\beta}_0 = -5137.2$

$\hat{\beta}_1 = 94.697$



Analisi del trend/2

In questa formulazione la variabile esogena "t" può variare in un qualsiasi insieme di valori equispaziati:

$t \in \{1,2,\ldots,n\};$ $\quad t \in \{10,20,\ldots,10*n\};$

$t \in \{1990,1991,\ldots,1989+n\};$ $\quad t \in \{1.5,3.0,\ldots,1.5*n\};$

$t \in \{1,3,5,7\ldots,2n+1\};$

Ne consegue che la variabile esogena debba essere interpretata come un insieme fisso di costanti.

Uso della retta di regressione

INTERPOLAZIONE

Lo scopo è trovare i valori della dipendente o di sostituirne i valori particolarmente anomali, per valori noti della indipendente.

ESTRAPOLAZIONE

Determinazione del valore della dipendente che corrisponde ad un valore della indipendente non necessariamente osservato.

CONTROLLO

Determinazione del valore della indipendente idoneo a determinare un fissato livello della dipendente

In ogni caso si ottengono dei VALORI TEORICI costituenti la stima dei VALORI VERI che rimangono comunque sconosciuti

Esempio

Unità di lavoro part-time e aumento di produzione

Prova	L.P.T.	PROD
A	1	4
B	2	6
C	3	10
D	4	10
E	5	15
F	6	15
G	7	16
H	8	20

$$PROD = 2.25 * 2.1667 * LPT$$

Ogni unità di lavoro part-time addizionale è responsabile di 2.1667 tonn. di produzione.

Se il lavoro part-time non fosse impiegato la produzione media sarebbe a 2.25 tonn.

Supponiamo che si decida di impiegare 10 LPT quale sarà l'incremento di produzione?

$$\hat{y}_{10} = 2.25 + 2.1667 * 10 = 2.25 + 21.667 = 23.917$$

Se invece si volesse stabilire quante LPT impiegare per ottenere 16 semilavorati allora

$$16 = 2.25 + 2.1667 * \hat{X} \Rightarrow \hat{X} = \frac{(16 - 2.25)}{2.1667} = 6.346$$

Perché fallisce un modello

ERRATA TEORIZZAZIONE

Le relazioni ipotizzate non reggono alla prova dei fatti per cui il modello non si conforma alla realtà osservata (ad esempio manca una variabile o è inserita un'altra non pertinente ovvero si è forzata la linearità)

E' difficile accertare l'influenza di tale eventualità.

Il modello è una visione semplificata della realtà, ma potrebbe esserne una visione semplicistica

$$Voto\ d'esame = f(Simpatia\ ispirata\ agli\ esaminatori)$$

$$Produzione\ agraria = f(Entità\ delle\ piogge)$$

$$Scorte\ magazzino = f(Vendite)$$

Si tatta di una limitazione intrinseca alla modellistica che si controlla solo presupponendo la validità del modello.

Perché fallisce un modello/2

ERRATA FORMULAZIONE

Le variabili sono state correttamente individuate, ma usate in modo sbagliato.

Ad esempio, la curva di Gompertz, spesso usata dagli attuari, per la costruzione delle tavole di mortalità ha equazione:

$$y = \beta_0 * e^{-\beta_1 e^{-\beta_2 X}}$$

Se però allo scatterplot viene adattato il modello

$$y = \beta_0 + \beta_1 X + \beta_2 X^2$$

le sue capacità esplicative saranno limitate ed occorre riformulare il modello.

Perché fallisce un modello/3

SCARSA QUALITA' DEI DATI

Se i dati acquisiti sui fenomeni coinvolti nel modello sono inattendibili sarà scadente anche il modello

GARBAGE IN -------> GARBAGE OUT

I risultati di una elaborazione statistica non possono essere più attendibili dei dati da essa utilizzati

Invece di utilizzare un numero indice sintetico dei prezzi per l'intera collettività nazionale si utilizza un indice per la scala mobile dei salari.

Gli strumenti di misurazione contengono errori sistematici o sono stati volontariamente alterati