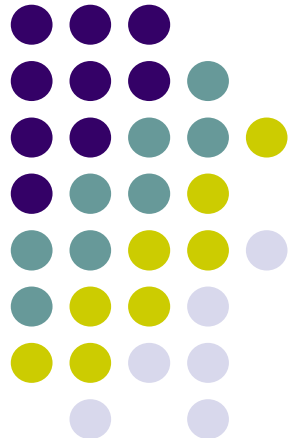
Levine, Krehbiel, Berenson

Statistica

Casa editrice: Pearson

Capitolo 12

La regressione lineare semplice



Insegnamento: Statistica
Corso di Laurea Triennale in Economia

Dipartimento di Economia e Management, Università di Ferrara
Docenti: Prof. Stefano Bonnini, Dott.ssa Angela Grassi

Argomenti

- Regressione e correlazione
- Regressione lineare semplice
  - Il modello di regressione
  - Equazione della retta di regressione
  - Misure di variabilità
  - Assunzioni del modello
  - Analisi dei residui
  - Inferenza sull'inclinazione della retta
  - Le trappole della regressione
  - I calcoli della regressione lineare semplice

Regressione e correlazione

Esistono molti metodi di inferenza statistica che si riferiscono ad una sola variabile statistica.

Obiettivo della lezione: studio della relazione tra due variabili.

Tecniche oggetto di studio:

regressione → Costruire un modello attraverso cui **prevedere** i valori di una **variabile dipendente** o **risposta** (quantitativa) a partire dai valori di una o più **variabili indipendenti** o **esplicative**

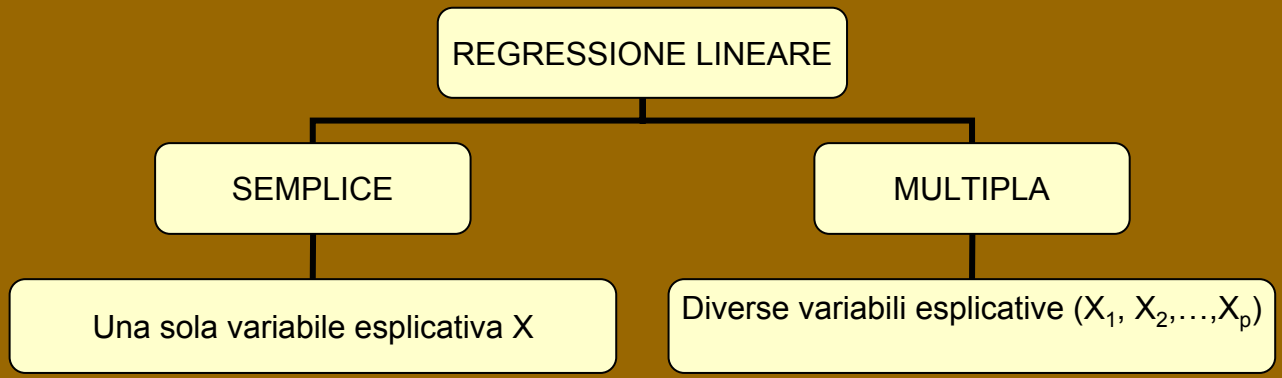
correlazione → Studio della associazione tra variabili quantitative

Regressione lineare

Solitamente nel modello di regressione si indica con

Y la variabile dipendente

X la variabile esplicativa



Il modello di regressione

Per studiare la relazione tra due variabili è utile il diagramma di dispersione in cui si riportano i valori della variabile esplicativa X sull'asse delle ascisse e i valori della variabile dipendente Y sull'asse delle ordinate.

La relazione tra due variabili può essere espressa mediante funzioni matematiche più o meno complesse tramite un modello di regressione.

Il modello di regressione lineare semplice è adatto quando i valori delle variabili X e Y si distribuiscono lungo una retta nel diagramma di dispersione.

Il modello di regressione lineare semplice

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (9.1)$$

dove

β_0 = l'intercetta per la popolazione

β_1 = l'inclinazione per la popolazione

ϵ_i = l'errore casuale in Y corrispondente all'i–esima osservazione

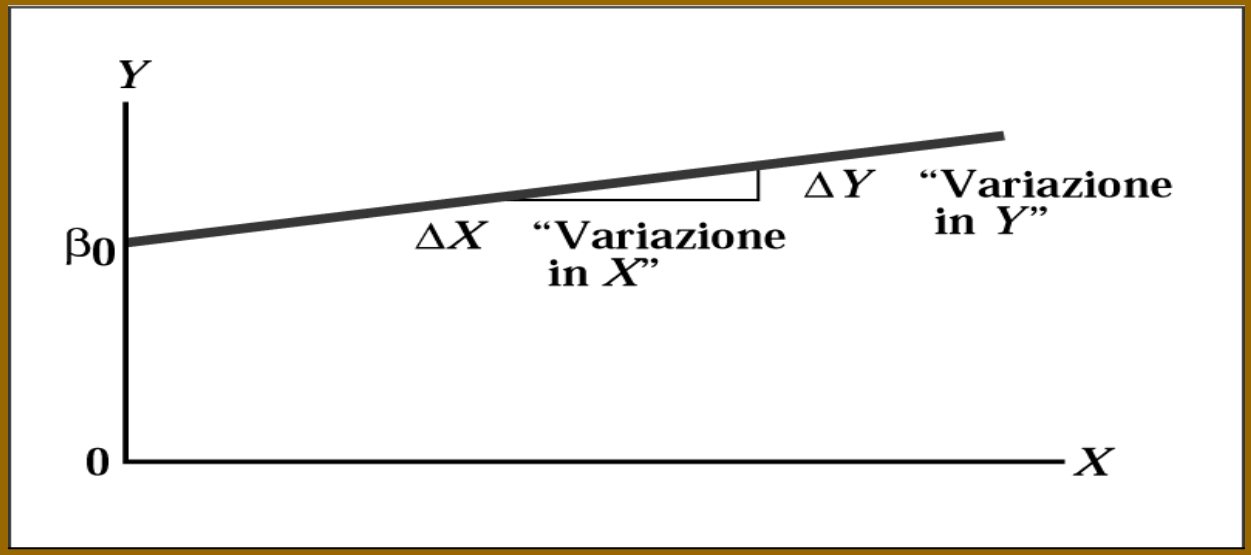
Il modello di regressione

L'inclinazione β_1 indica come varia Y in corrispondenza di una variazione unitaria di X.

L'intercetta β_0 corrisponde al valore medio di Y quando X è uguale a 0.

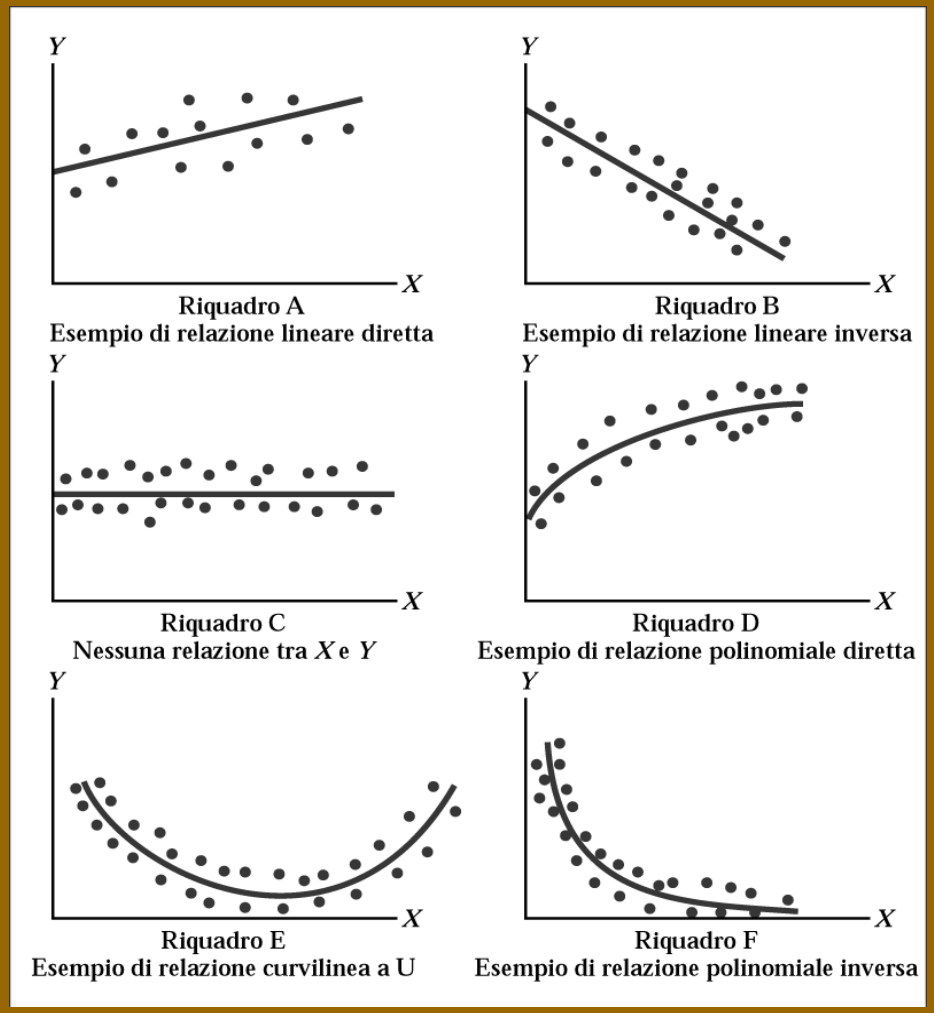
Il segno di β_1 indica se la relazione lineare è positiva o negativa.

Esempio di relazione lineare positiva



Il modello di regressione

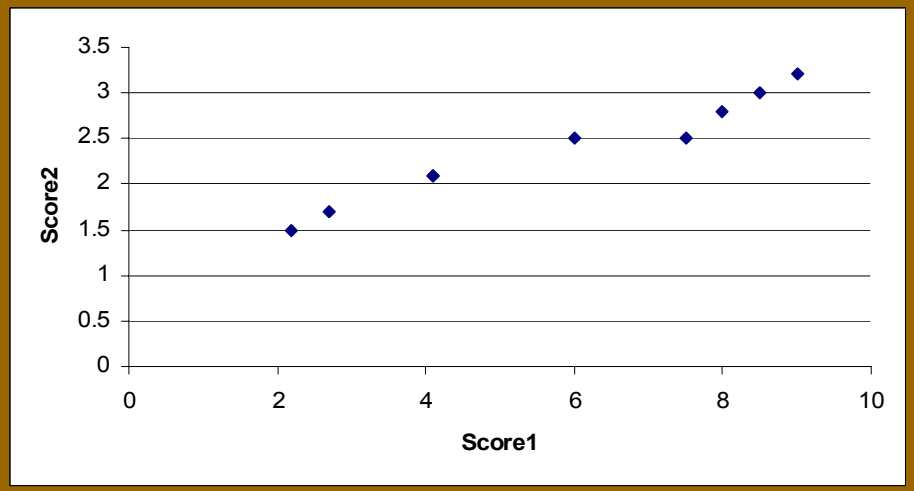
La scelta del modello matematico appropriato è suggerita dal modo in cui si distribuiscono i valori delle due variabili nel diagramma di dispersione



Il modello di regressione

Esempio: un produttore desidera ottenere una misura della qualità di un prodotto ma la procedura è troppo costosa. Decide allora di stimare questa misura (score 2) a partire dall'osservazione di un'altra misura (score 1) più semplice meno costosa da ottenere.

Unità di prodotto	Score1	Score2
1	4.1	2.1
2	2.2	1.5
3	2.7	1.7
4	6	2.5
5	8.5	3
6	4.1	2.1
7	9	3.2
8	8	2.8
9	7.5	2.5



Equazione della retta di regressione

Si dimostra che sotto certe ipotesi i parametri del modello β_0 e β_1 possono essere stimati ricorrendo ai dati del campione. Indichiamo con b_0 e b_1 le stime ottenute.

L'equazione campionaria del modello di regressione lineare

La previsione di Y in base al modello di regressione lineare è data dalla somma tra l'intercetta campionaria e il prodotto tra il valore di X e l'inclinazione campionaria

$$\hat{Y}_i = b_0 + b_1 X_i \qquad (9.2)$$

dove

$\hat{Y}_i$ = previsione di Y per l'osservazione i

X_i = valore di X per l'osservazione i

La regressione ha come obiettivo quello di individuare la retta che meglio si adatta ai dati.

Esistono vari modi per valutare la capacità di adattamento

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

Equazione della retta di regressione

Il metodo dei minimi quadrati consiste nel determinare b_0 e b_1 rendendo minima la somma dei quadrati delle differenze tra i valori osservati Y_i e i valori stimati $\hat{Y}_i$.

si tratterà di *minimizzare* la somma dei loro quadrati:

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

dove

Y_i = il vero valore di Y per l'osservazione di i

$\hat{Y}_i$ = il valore previsto di Y per l'osservazione di i

Dal momento che in base al modello proposto $\hat{Y}_i = b_0 + b_1 X_i$, si tratta di minimizzare la seguente espressione:

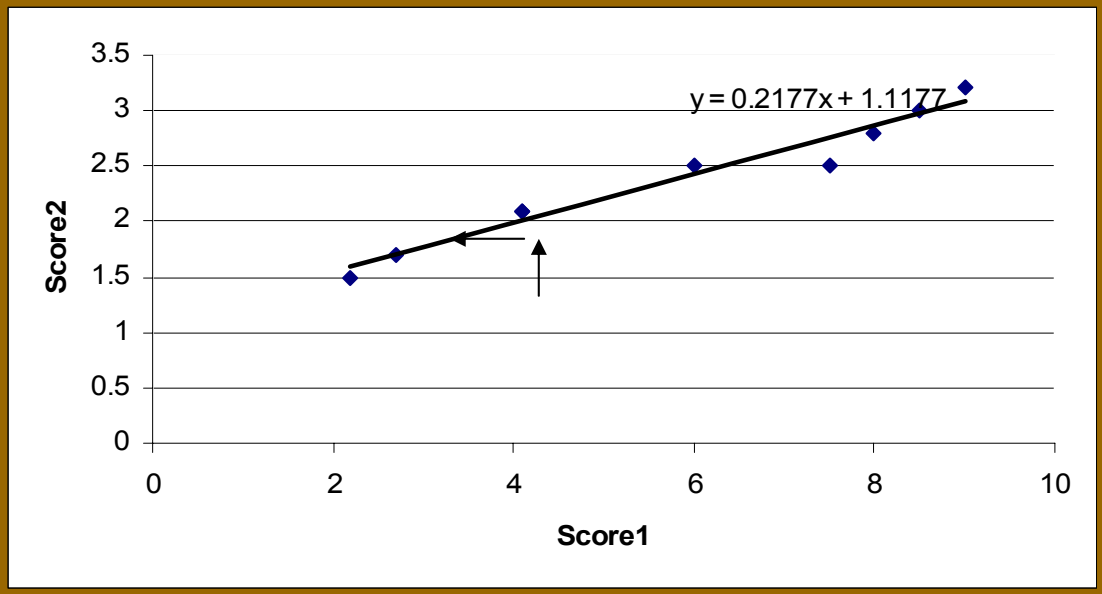
$$\sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

rispetto alle due incognite b_0 e b_1.

I valori b_0 e b_1 sono chiamati coefficienti di regressione.

Equazione della retta di regressione

Nell'esempio precedente in cui si intendeva prevedere il valore di una misura di qualità score2 in funzione di un'altra misura score1, applicando il metodo dei minimi quadrati si ottiene la seguente retta di regressione:



Tramite l'equazione score2 = 1,1177 + 0,2177 score1 è possibile prevedere i valori di score2 in funzione di quelli osservati di score1. Se ad esempio osservassimo un valore di score1 pari a 4,5 il valore stimato di score2 sarebbe 2,1.

Risulta:

b_1 = 0,2177

b_0 = 1,1177

Perciò se aumenta di un'unità il valore di score1, il valore previsto di score2 subisce un incremento di 0,2177.

Se score1 assume valore 0, il valore previsto per score2 è pari a 1,1177.

Equazione della retta di regressione

La previsione di un valore di Y in corrispondenza di un certo valore di X può essere definita in due modi, in relazione all'intervallo di valori di X usati per stimare il modello:

- interpolazione: se la previsione di Y corrisponde ad un valore di X interno all'intervallo
- estrapolazione: se la previsione di Y corrisponde ad un valore di X che non cade nell'intervallo

Nell'esempio precedente l'intervallo per la variabile indipendente (score1) è [2,2; 9,0]. Calcolando la previsione di score2 per un valore di score1 pari a 4,5 abbiamo effettuato un interpolazione. Se volessimo calcolare la previsione di score2 in corrispondenza del valore 9,5 per score1, faremmo un estrapolazione.

Misure di variabilità

Le seguenti misure di variabilità consentono di valutare le capacità previsive del modello statistico proposto.

Variabilità totale (somma totale dei quadrati) → variabilità di Y

Variabilità spiegata (somma dei quadr. della regress.) → variabilità di Ŷ

Variabilità non spiegata (somma dei quadr. degli errori) → variabilità dell'errore

Le misure di variabilità nella regressione

Somma totale dei quadrati = somma dei quadrati della regressione + somma dei quadrati degli errori

$$SQT = SQR + SQE \quad (9.3)$$

La somma totale dei quadrati (*SQT*)

La somma totale dei quadrati (*SQT*) è data dalla somma dei quadrati delle differenze tra i valori osservati di *Y* e la loro media.

$$SQT = \text{variabilità totale} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \quad (9.4)$$

La somma dei quadrati della regressione (*SQR*)

La somma dei quadrati della regressione (*SQR*) è data dalla somma dei quadrati delle differenze tra i valori previsti di *Y* e la media di *Y*.

$$SQR = \text{variabilità spiegata} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \quad (9.5)$$

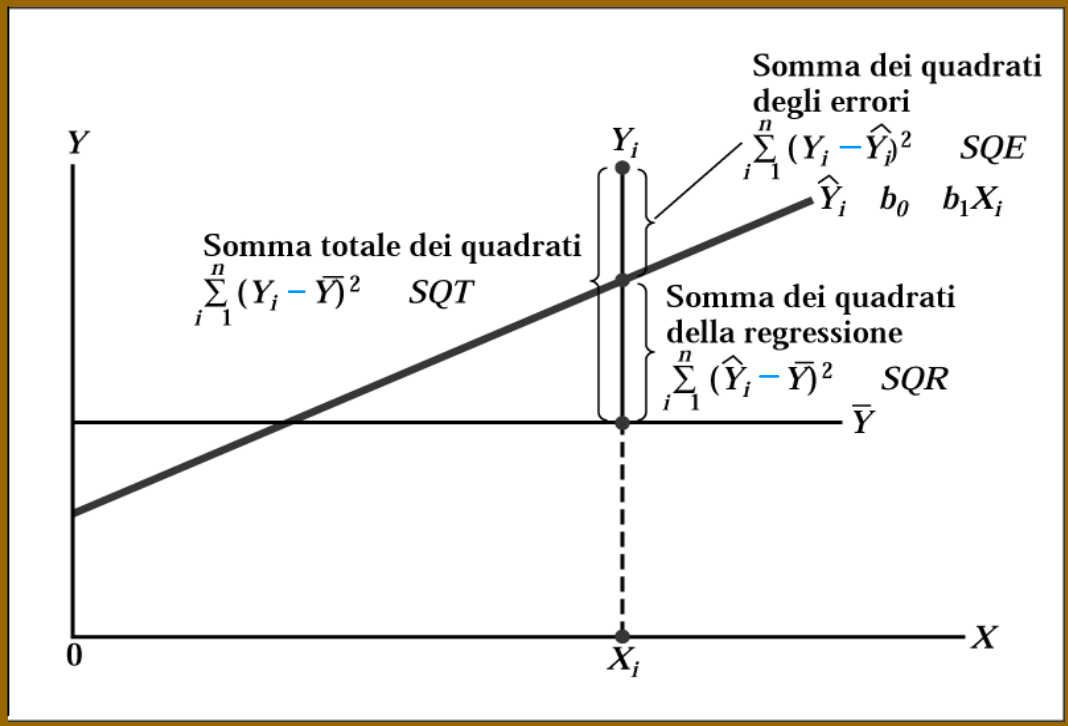
$$= SQT - SQE$$

Misure di variabilità

La somma dei quadrati degli errori (*SQE*)

La somma dei quadrati degli errori (*SQE*) è data dalla somma dei quadrati delle differenze tra i valori osservati e i valori previsti di Y

$$SQE = \text{variabilità non spiegata} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad (9.6)$$



Misure di variabilità

Il coefficiente di determinazione è una misura utile per valutare il modello di regressione

Il coefficiente di determinazione

Il coefficiente di determinazione è dato dal rapporto tra la somma dei quadrati della regressione e la somma totale dei quadrati.

$$r^2 = \frac{SQR}{SQT} \qquad (9.7)$$

Esso misura la parte di variabilità di Y spiegata dalla variabile X nel modello di regressione.

L'errore standard della stima è una misura della variabilità degli scostamenti dei valori osservati da quelli previsti.

L'errore standard della stima

$$S_{YX} = \sqrt{\frac{SQE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} \qquad (9.8)$$

dove

Y_i = il valore di Y in corrispondenza X_i

$\hat{Y}_i$ = il valore previsto di Y in corrispondenza di X_i

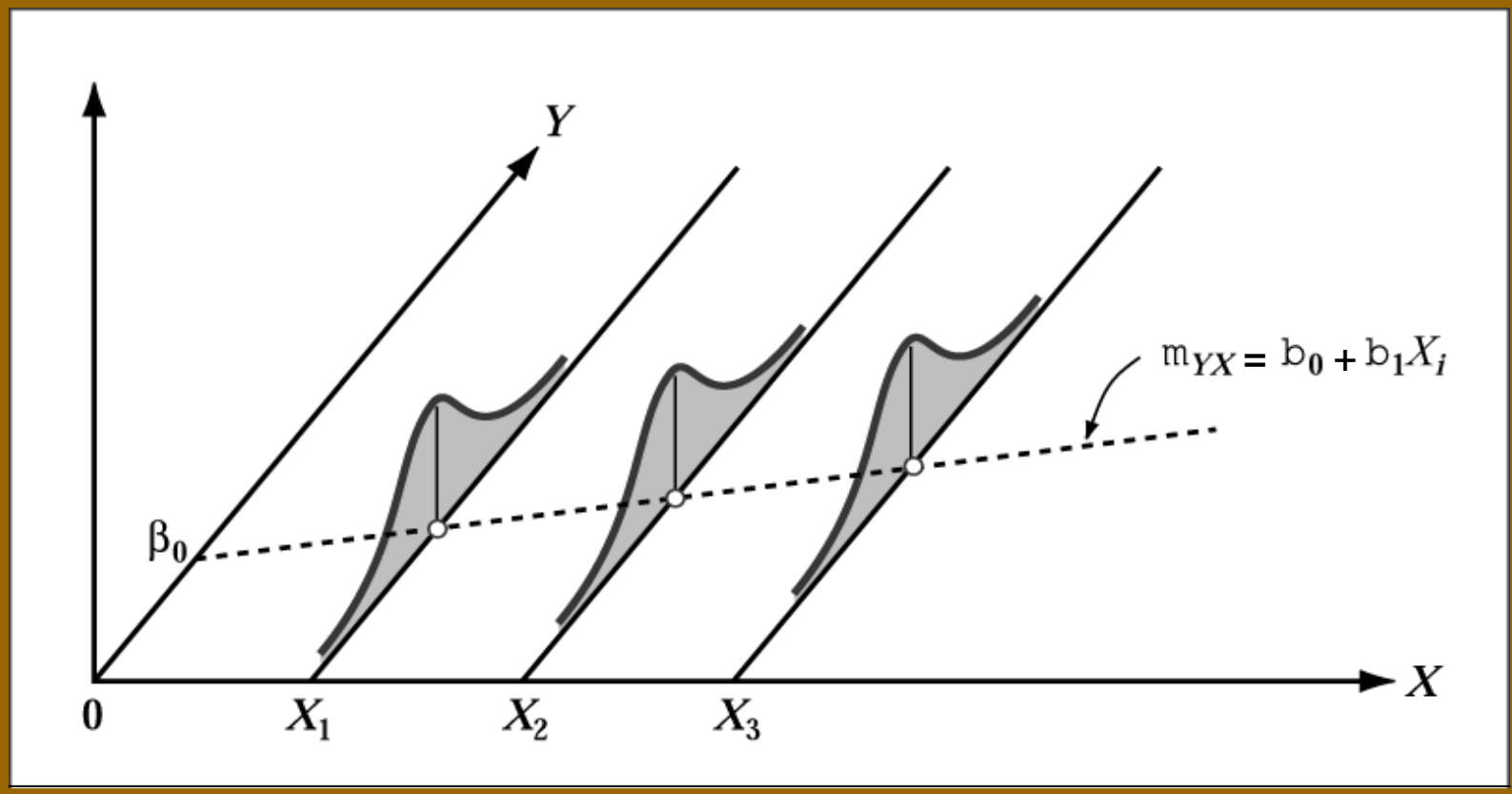
SQE = somma dei quadrati degli errori

Nell'esempio precedente risulta $r^2 = 0{,}96$ e $S_{YX} = 0{,}13$.

Le assunzioni del modello

- Distribuzione normale degli errori: gli errori devono avere, per ogni valore di X, una distribuzione normale. Il modello di regressione è comunque robusto rispetto a scostamenti dall'ipotesi di normalità
- Omoschedasticità: la variabilità degli errori è costante per ciascun valore di X.
- Indipendenza degli errori: gli errori devono essere indipendenti per ciascun valore di X (importante soprattutto per osservazioni nel corso del tempo)

Le assunzioni del modello



Analisi dei residui

Il residuo e_i è una stima dell'errore che commetto nel prevedere Y_i tramite $\hat{Y}_i$.

Il residuo

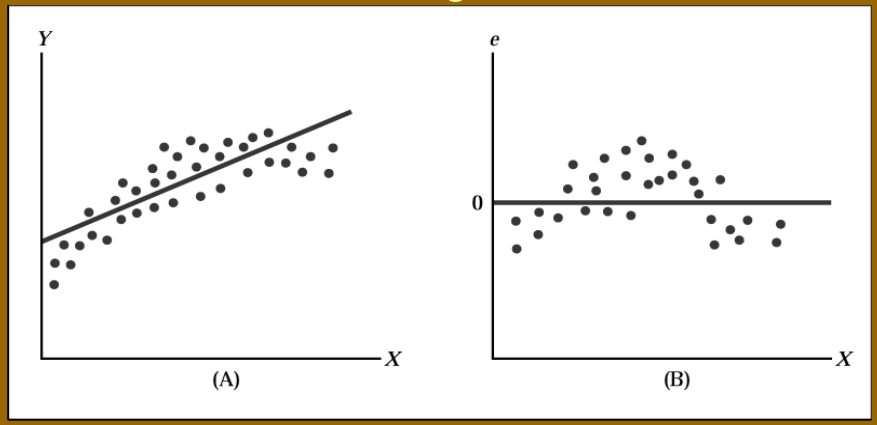
Il residuo è uguale alla differenza tra valore osservato e il valore previsto di Y:

$$e_i = Y_i - \hat{Y}_i \qquad (9.9)$$

Per stimare la capacità di adattamento ai dati della retta di regressione è opportuna una analisi grafica → grafico di dispersione dei residui (ordinate) e dei valori di X (ascisse).

Se si evidenzia una relazione particolare il modello non è adeguato.

Nell'esempio a lato il modello di regressione lineare non sembra appropriato. Il grafico a destra evidenzia lo scarso adattamento ai dati del modello (lack of fit). Quindi il modello polinomiale è più appropriato.

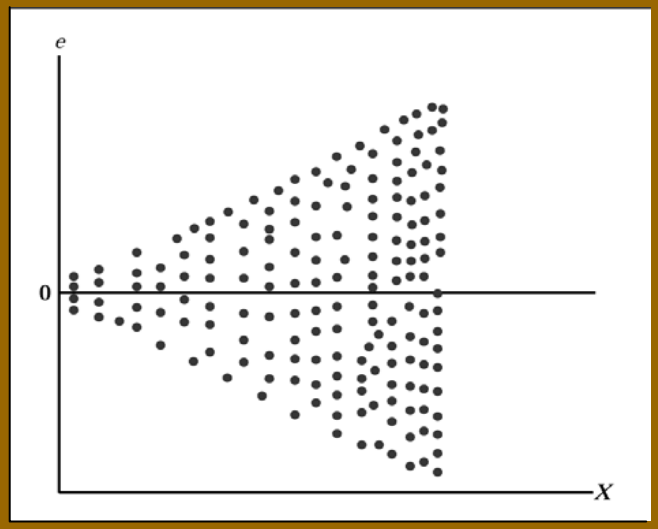


Analisi dei residui

Valutazione delle ipotesi:

- Omoschedasticità: il grafico dei residui rispetto a X consente di stabilire anche se la variabilità degli errori varia a seconda dei valori di X

Il grafico a lato evidenzia ad esempio che la variabilità dei residui aumenta all'aumentare dei valori di X.

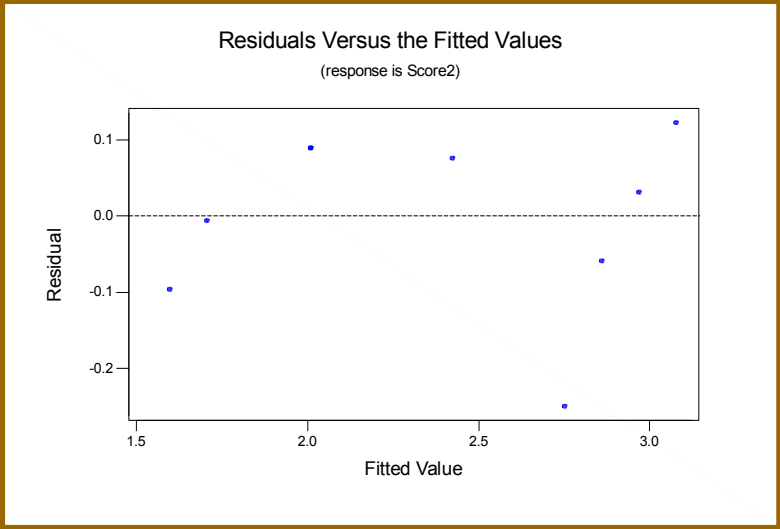


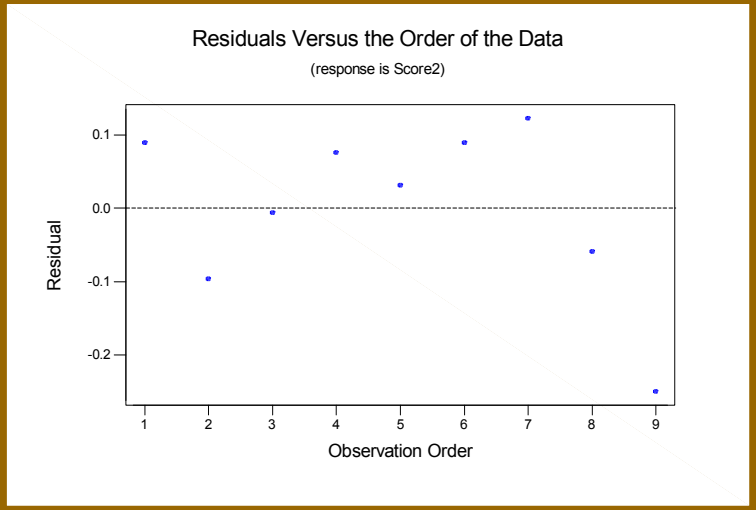
- Normalità: rappresentazione della distribuzione di frequenze dei residui (es. istogramma)
- Indipendenza: rappresentando i residui nell'ordine con cui sono stati raccolti i dati emerge un'eventuale autocorrelazione tra osservazioni successive.

Analisi dei residui

Dall'esempio precedente risulta che i residui non si distribuiscono in modo regolare al variare delle stime della variabile dipendente (e quindi anche al variare della X). Il modello quindi non è ben specificato.

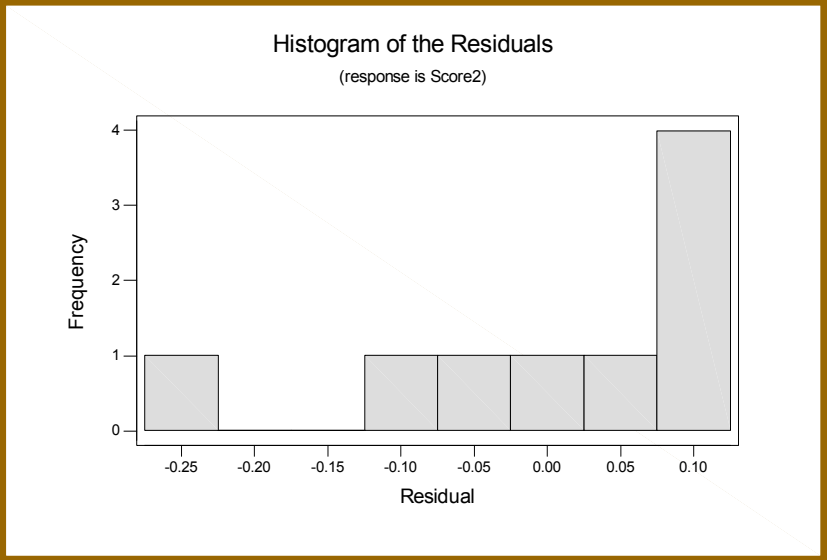
Il grafico dei residui rispetto al tempo non sembra evidenziare l'esistenza di autocorrelazione dei primi.

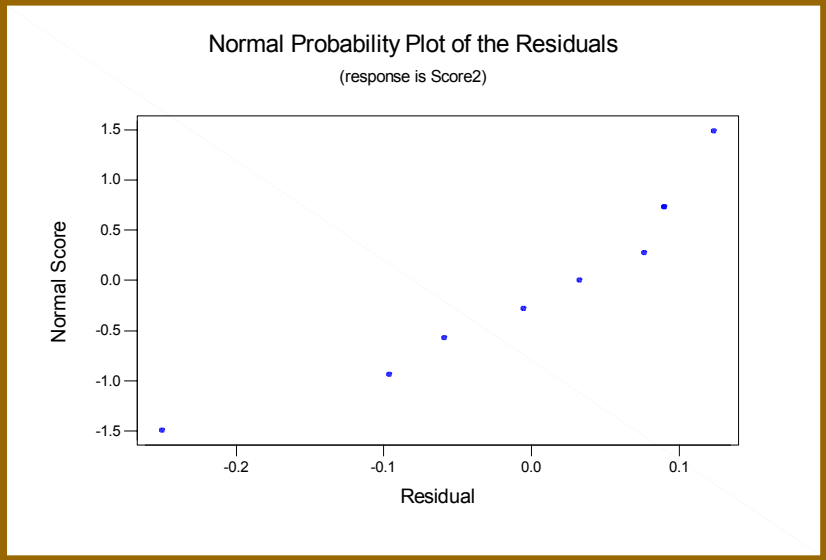




Analisi dei residui

Per quanto riguarda la normalità dei residui, l'istogramma delle frequenze e il normal probability plot ci portano ad escludere che la condizione sia verificata.





Inferenza sull'inclinazione della retta di regressione

Possiamo stabilire se tra le variabili X e Y sussiste una relazione lineare significativa sottoponendo a verifica l'ipotesi che β_1 (inclinazione della popolazione) sia uguale a zero.

H_0: $\beta_1 = 0$ (non vi è una relazione lineare)

H_1: $\beta_1 \neq 0$ (vi è una relazione lineare)

Il test *t* per la verifica di ipotesi sull'inclinazione β_1

La statistica *t* è data dalla differenza tra l'inclinazione campionaria e l'inclinazione ipotizzata della popolazione, il tutto diviso per l'errore standard dell'inclinazione.

$$t = \frac{b_1 - \beta_1}{S_{b_1}}$$

dove

(9.11)

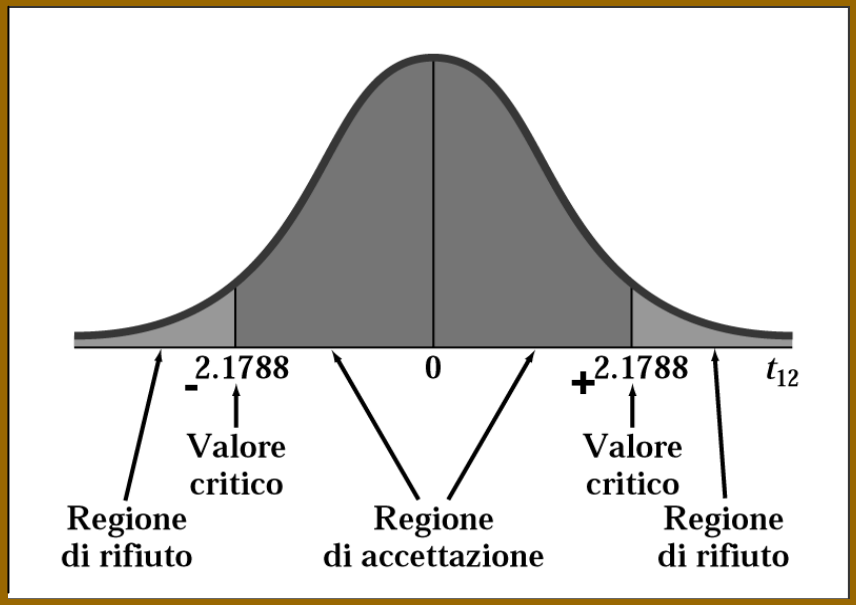
$$S_{b_1} = \frac{S_{YX}}{\sqrt{SQX}}$$

$$SQX = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

La statistica *t* ha una distribuzione *t* di Student con $n - 2$ gradi di libertà.

Inferenza sull'inclinazione della retta di regressione

Se ad esempio α=0,05 e n=14, allora le regioni di accettazione e di rifiuto sono definite come segue:



Nell'esempio del modello di regressione in cui score1 è variabile esplicativa e score2 variabile dipendente abbiamo che b_1=0,2177 n=8

$t=b_1/S_{b1}=12,51>t_6 = 2,45$

perciò rigetto l'ipotesi che l'inclinazione sia nulla a favore dell'ipotesi che esista inclinazione significativa.

Inferenza sull'inclinazione della retta di regressione

La significatività dell'inclinazione della retta può essere sottoposta a verifica anche ricorrendo al test F:

Il test *F* per la verifica di ipotesi sull'inclinazione β_1

La statistica F è data dal rapporto tra la media dei quadrati della regressione (MQR) e la media dei quadrati dell'errore (MQE):

$$F = \frac{MQR}{MQE} \tag{9.12}$$

dove

$$MQR = \frac{SQR}{p}$$

$$MQE = \frac{SQE}{n - p - 1}$$

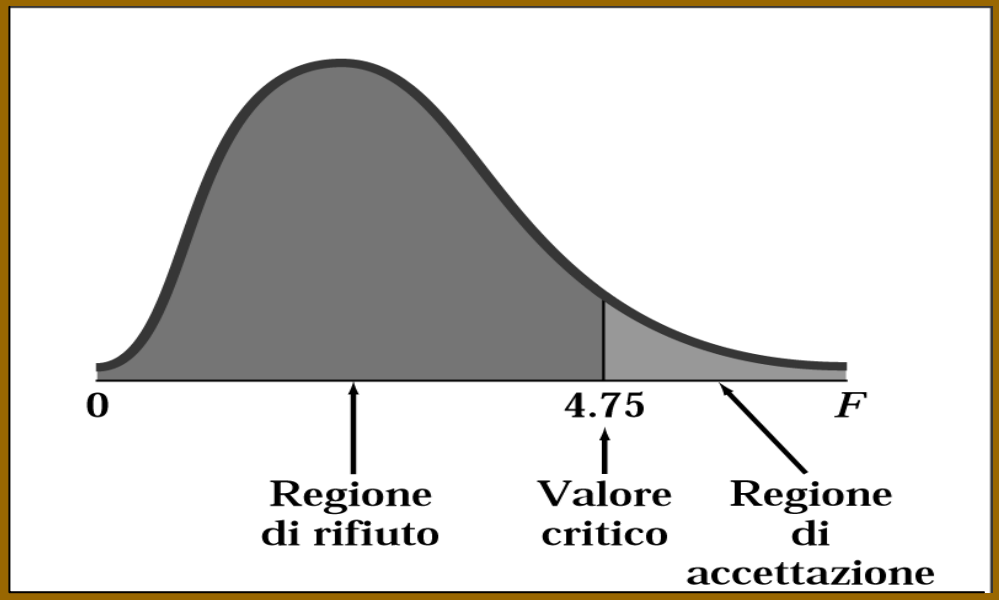
Tabella 9.5 *Tabella dell'ANOVA per la verifica della significatività del coefficiente di regressione*

Fonte	*Gdl*	Somma dei quadrati	Media dei quadrati (Varianza)	*F*
Regressione	p	SQR	$MQR = \frac{SQR}{p}$	$F = \frac{MQR}{MQE}$
Residuo	$n - p - 1$	SQE	$MQE = \frac{SQE}{n - p - 1}$	
Totale	$n - 1$	STQ		

Inferenza sull'inclinazione della retta di regressione

La regola decisionale è la seguente:

Rifiuto H_0 se $F > F_U$ con F_U valore critico che lascia a destra probabilità pari ad α.



Nell'esempio del modello di regressione in cui score1 è variabile esplicativa e score2 variabile dipendente abbiamo che F=156,56 > $F_{1,6}$ = 5,99 quindi rigetto l'ipotesi di inclinazione non significativa.

Inferenza sull'inclinazione della retta di regressione

Un altro modo per verificare la significatività di β_1 è quello di costruire un intervallo di confidenza per il parametro.

Se il valore ipotizzato $\beta_1 = 0$ è incluso nell'intervallo accetto l'ipotesi di inclinazione non significativa.

L'intervallo di confidenza per l'inclinazione

L'intervallo di confidenza per β_1 si ottiene addizionando e sottraendo all'inclinazione campionaria b_1 il prodotto tra il valore critico della statistica t e l'errore standard dell'inclinazione.

$$b_1 \pm t_{n-2} S_{b_1} \qquad (9.13)$$

Nel nostro esempio abbiamo $\beta_1 = 0{,}21767$ $t_6 = 2{,}45$ $S_{b1} = 0{,}01740$ perciò al livello di confidenza del 95% il vero valore di β_1 è compreso nell'intervallo [0,17504;0,2603]. Lo zero non cade nell'intervallo, perciò rigetto l'ipotesi nulla.

Stima della previsione

Oltre ad ottenere previsioni per i valori di Y (stime puntuali della media di Y) si possono ottenere intervalli di confidenza per la media della variabile risposta:

L'intervallo di confidenza per $\mu_{Y|X_i}$ la media di Y

$$\hat{Y}_i \pm t_{n-2} S_{YX} \sqrt{h_i} \qquad (9.14)$$

dove

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

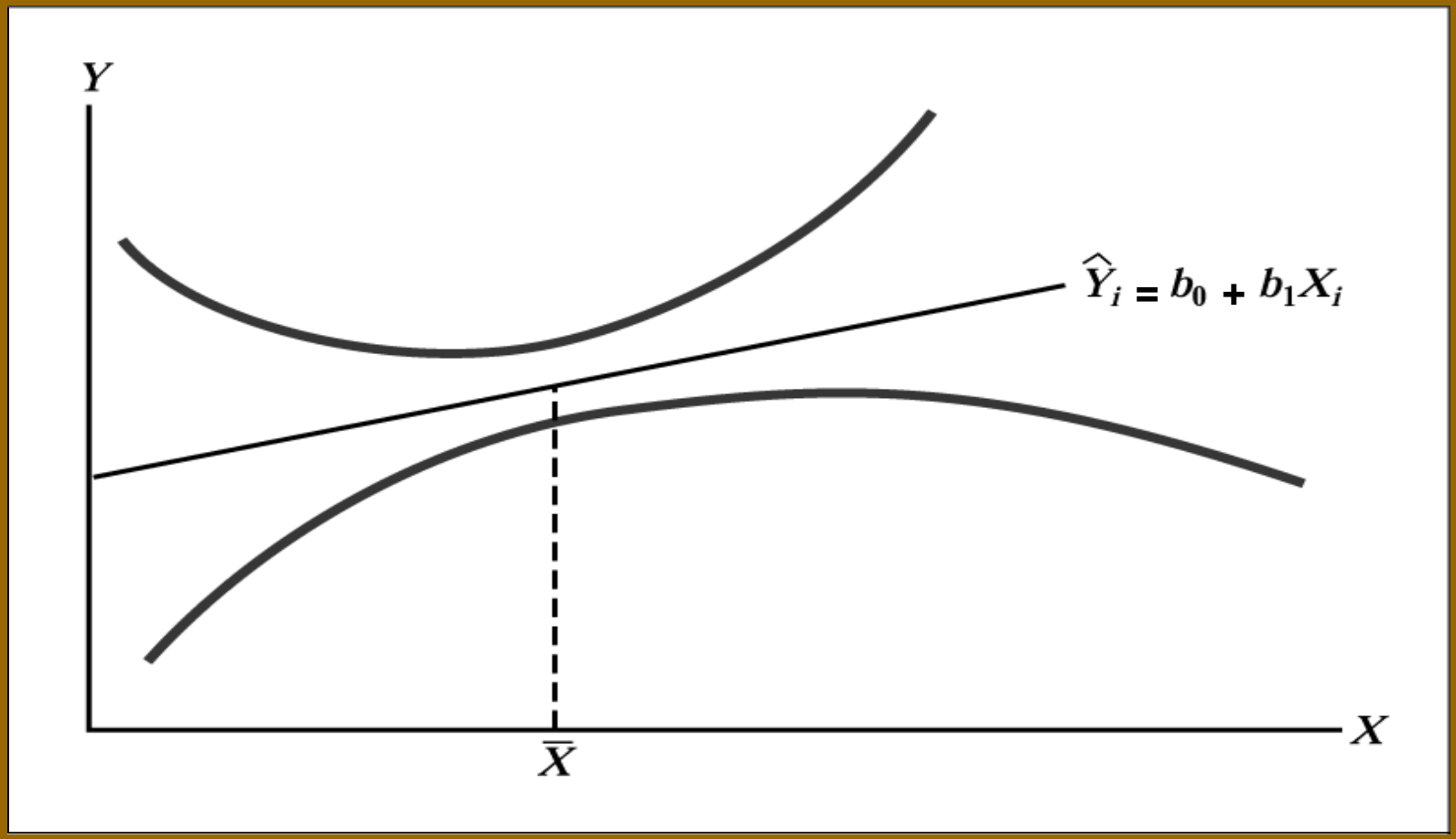
$\hat{Y}_i$ = previsione del valore medio Y: $\hat{Y}_i = b_0 + b_1 X_i$

S_{YX} = errore standard della stima

n = ampiezza del campione

X_i = i–esimo del valore di X

Stima della previsione



Stima della previsione

E' possibile ottenere un intervallo di confidenza per la previsione di un singolo valore di Y. La formula è molto simile a quella dell'intervallo di confidenza per la media anche se in questo caso si stima un valore e non un parametro:

L'intervallo di confidenza per la previsione di una singola risposta Y_t

$$\hat{Y}_t \pm t_{n-2} S_{YX} \sqrt{1 + h_t} \qquad (9.15)$$

dove

h_t, $\hat{Y}_t$, S_{YX}, n e X_t sono definiti come nell'equazione 9.15

Le trappole dell'analisi di regressione

Il modello di regressione è una tecnica statistica molto utilizzata.

Spesso però viene impiegata in modo non corretto.

Riquadro 9.2 Le difficoltà del modello di regressione

- ✓ **1.** Scarsa conoscenza delle assunzioni alla base del modello.
- ✓ **2.** Scarsa conoscenza del modo in cui valutare tali assunzioni.
- ✓ **3.** Scarsa conoscenza dei modelli alternativi a quello di regressione lineare semplice.
- ✓ **4.** Uso del modello di regressione senza una conoscenza adeguata della teoria sottostante.

L'analisi grafica molto spesso consente di rilevare eventuali informazioni che le analisi numeriche non evidenziano.

Le trappole dell'analisi di regressione

Ad esempio, a partire da quattro dataset diversi, è possibile ottenere gli stessi risultati in termini di statistiche di regressione pur trattandosi di situazioni molto diverse tra loro.

Tabella 9.6 *Quattro insiemi di dati artificiali*

Dataset A		Dataset B		Dataset C		Dataset D	
X_i	Y_i	X_i	Y_i	X_i	Y_i	X_i	Y_i
10	8.04	10	9.14	10	7.46	8	6.58
14	9.96	14	8.10	14	8.84	8	5.76
5	5.68	5	4.74	5	5.73	8	7.71
8	6.95	8	8.14	8	6.77	8	8.84
9	8.81	9	8.77	9	7.11	8	8.47
12	10.84	12	9.13	12	8.15	8	7.04
4	4.26	4	3.10	4	5.39	8	5.25
7	4.82	7	7.26	7	6.42	19	12.50
11	8.33	11	9.26	11	7.81	8	5.56
13	7.58	13	8.74	13	12.74	8	7.91
6	7.24	6	6.13	6	6.08	8	6.89

$$SQR = \text{variabilità spiegata} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = 27.51$$

$$SQE = \text{variabilità non spiegata} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 13.763$$

$$SQT = \text{variabilità totale} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 41.273$$

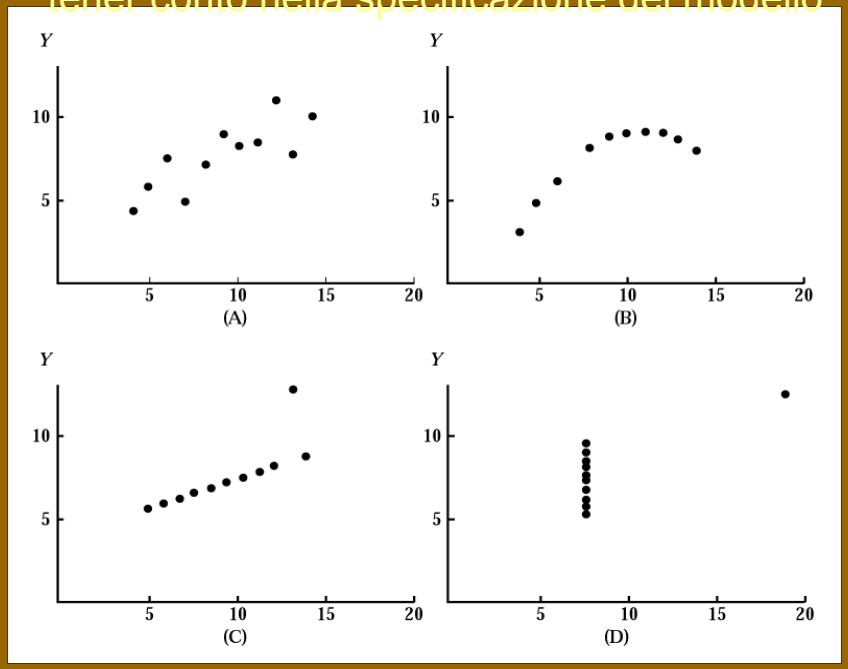
Le trappole dell'analisi di regressione

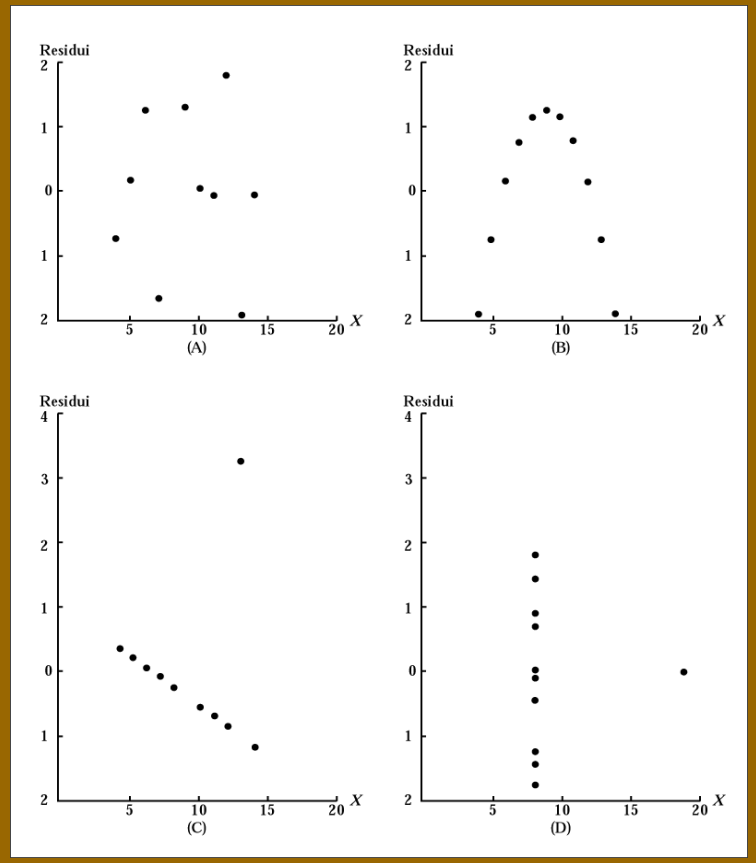
Caso A: il modello di regressione lineare semplice sembra appropriato

Caso B: sembra più appropriato un modello polinomiale (di secondo grado)

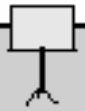
Caso C: presenza di un outlier che deve essere eliminato prima di procedere alle stime

Caso D: valore anomalo di X di cui si dovrebbe tener conto nella specificazione del modello





Le trappole dell'analisi di regressione



Riquadro 9.3 Una strategia per evitare le trappole della regressione

- ✓ **1.** Cominciate l'analisi sempre con un'attenta osservazione del diagramma di dispersione, per cogliere l'eventuale relazione tra *X* e *Y*.
- ✓ **2.** Verificate se le ipotesi alla base del modello di regressione sono soddisfatte dopo la stima del modello e prima di passare a impiegarne i risultati.
- ✓ **3.** Rappresentate graficamente i residui rispetto alla variabile dipendente per stabilire se il modello si adatta ai dati e se l'ipotesi di omoschedasticità è rispettata.
- ✓ **4.** Usate l'istogramma, il diagramma ramo-foglia o il diagramma scatola e baffi dei residui per verificare in quale misura l'ipotesi di normalità degli errori è rispettata.
- ✓ **5.** Se i dati sono raccolti in ordine sequenziale, rappresentate graficamente i residui nell'ordine con cui i dati sono stati raccolti e calcolate la statistica di Dubin-Watson.
- ✓ **6.** Se alla luce dei punti 3-5 ritenete che le ipotesi alla base del modello di regressione lineare siano violate, ricorrete ad altri metodi di stima del modello o ad altri modelli.
- ✓ **7.** Se alla luce dei punti 3-5 ritenete che le ipotesi alla base del modello di regressione lineare non siano violate, potete procedere ad alcune inferenze sul modello. Sottoponete a verifica la significatività dei coefficienti e costruite gli intervalli di confidenza per la risposta media e per la previsione.

I calcoli della regressione lineare semplice

Applicando il metodo dei minimi quadrati per la stima dei coefficienti della retta di regressione si ha:

Equazioni da risolvere per applicare il metodo dei minimi quadrati

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i \qquad (9.16a)$$

$$\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 \qquad (9.16b)$$

Formula per il calcolo dell'inclinazione b_1

$$b_1 = \frac{SQXY}{SQX} \qquad (9.17)$$

$\bar{X})$dove

$$SQXY = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}$$

$$SQX = \sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}$$

Formula per il calcolo dell'intercetta b_0

$$b_0 = \bar{Y} - b_1 \bar{X} \qquad (9.18)$$

dove

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \quad \text{e} \quad \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

I calcoli della regressione lineare semplice

Calcolo delle misure di variabilità:

Formula per il calcolo della somma totale dei quadrati (*SQT*)

$$SQT = \text{variabilità totale} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \quad (9.19)$$

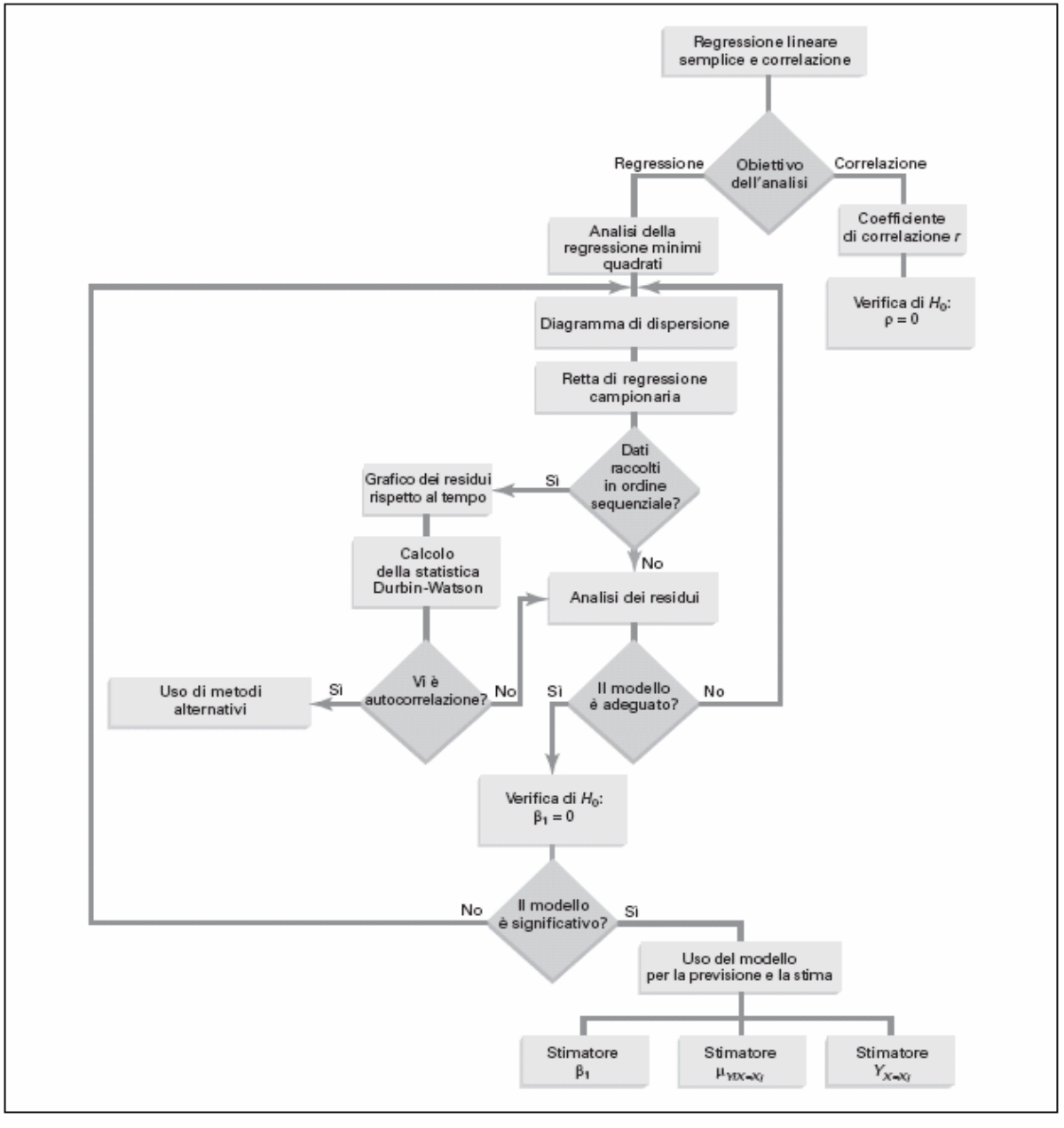
Formula per il calcolo della somma dei quadrati della regressione (SQR)

$$SQR = \text{variabilità spiegata dalla regressione} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = b_0\sum_{i=1}^{n} Y_i + b_1\sum_{i=1}^{n} X_iY_i - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \quad (9.20)$$

Formula per il calcolo della somma dei quadrati degli errori (SQE)

$$SQE = \text{variabilità residua} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} Y_i^2 - b_0\sum_{i=1}^{n} Y_i - b_1\sum_{i=1}^{n} X_iY_i \quad (9.21)$$

Riepilogo


Regressione lineare semplice e correlazione
Obiettivo dell'analisi
Regressione
Correlazione
Analisi della regressione minimi quadrati
Coefficiente di correlazione r
Verifica di H_0: $\rho = 0$
Diagramma di dispersione
Retta di regressione campionaria
Dati raccolti in ordine sequenziale?
Sì
Grafico dei residui rispetto al tempo
Calcolo della statistica Durbin-Watson
No
Analisi dei residui
Vi è autocorrelazione?
Sì
No
Uso di metodi alternativi
Il modello è adeguato?
Sì
No
Verifica di H_0: $\beta_1 = 0$
Il modello è significativo?
No
Sì
Uso del modello per la previsione e la stima
Stimatore β_1
Stimatore $\mu_{Y|X=X_i}$
Stimatore $Y_{X=X_i}$