Argomenti

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

Regressione e correlazione

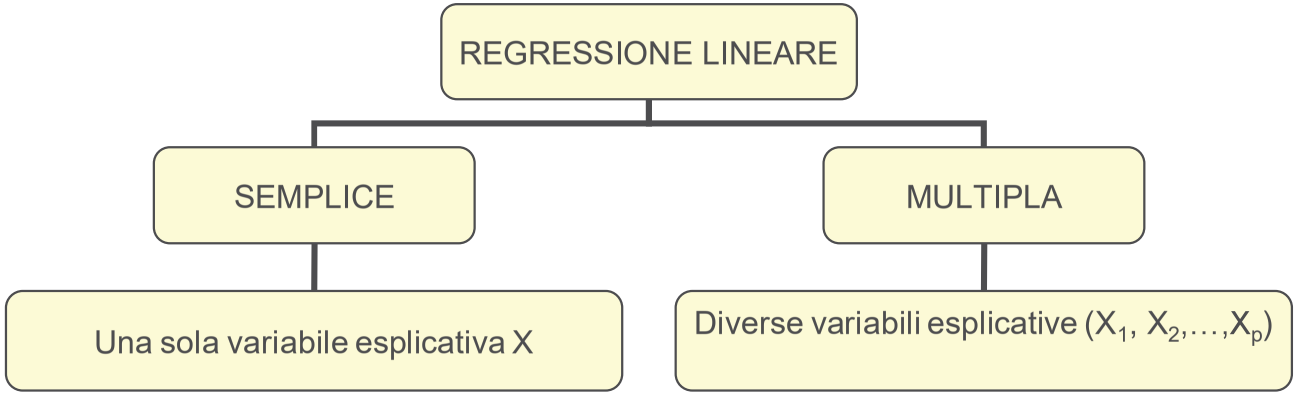
- Esistono molti metodi di inferenza statistica che si riferiscono ad una sola variabile statistica.
- Obiettivo della lezione: studio della relazione tra due variabili.
- Tecniche oggetto di studio:

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 e con 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

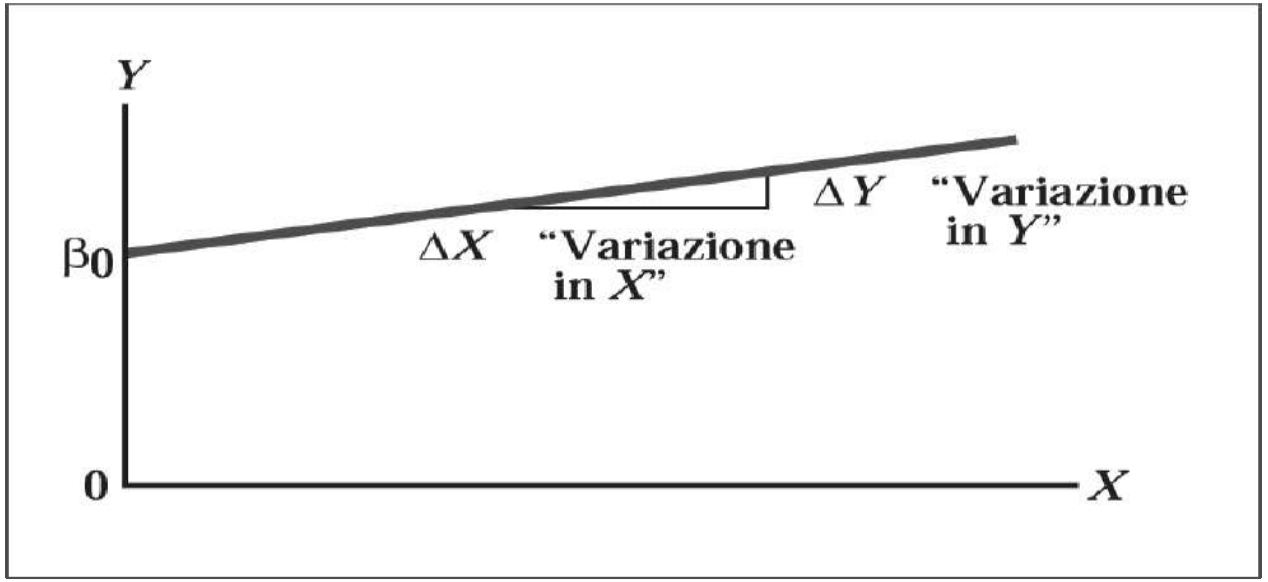
Il modello di regressione

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.

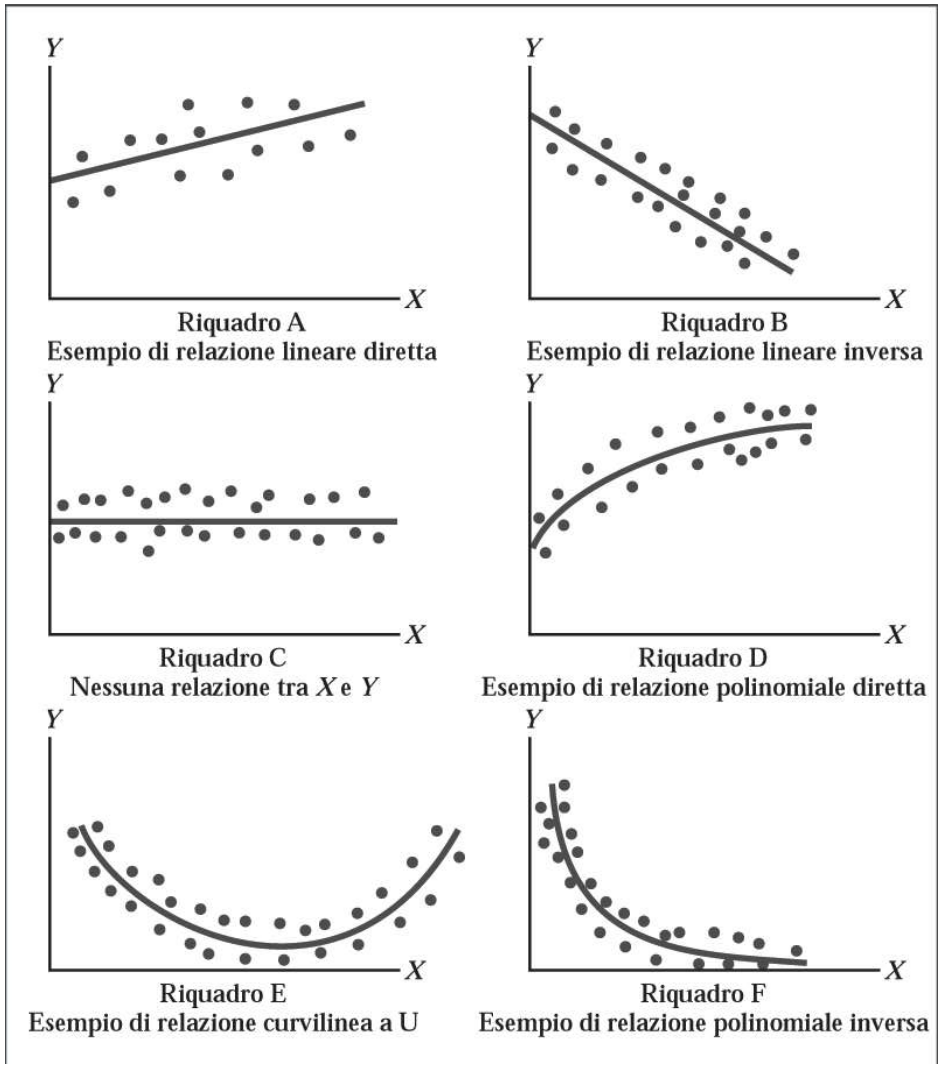
Il segno di β_1 indica se la relazione lineare è positiva o negativa.

Esempio di relazione lineare positiva



Il modello di regressione

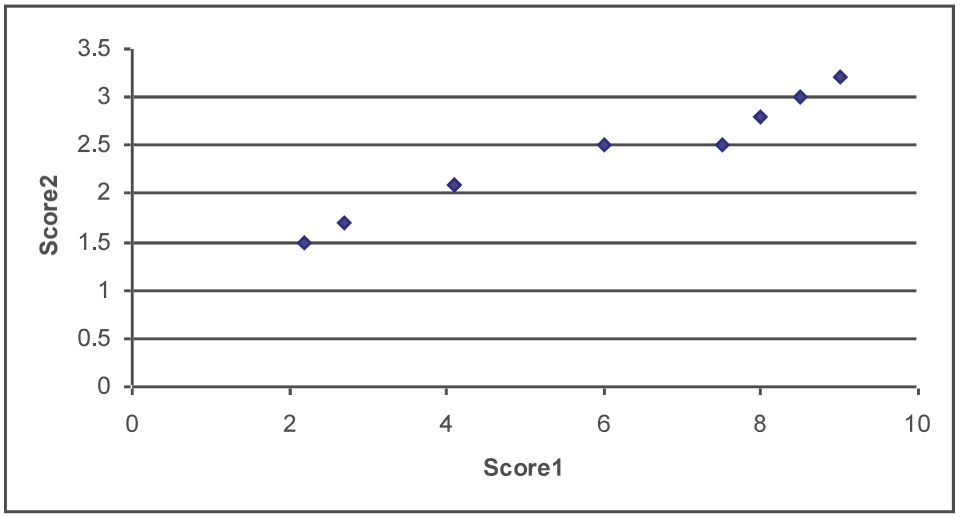
La scelta del modello matematico appropriato è suggerita dal modo in cui si distribuiscono i valori delle due variabili nel diagramma di dispersione



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.

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*)

$$SQT = \text{variabilità totale}$$

$$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \qquad (9.19)$$

$$= \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

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

$$SQR = \text{variabilità spiegata dalla regressione}$$

$$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = b_0 \sum_{i=1}^{n} Y_i + b_1 \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \qquad (9.20)$$

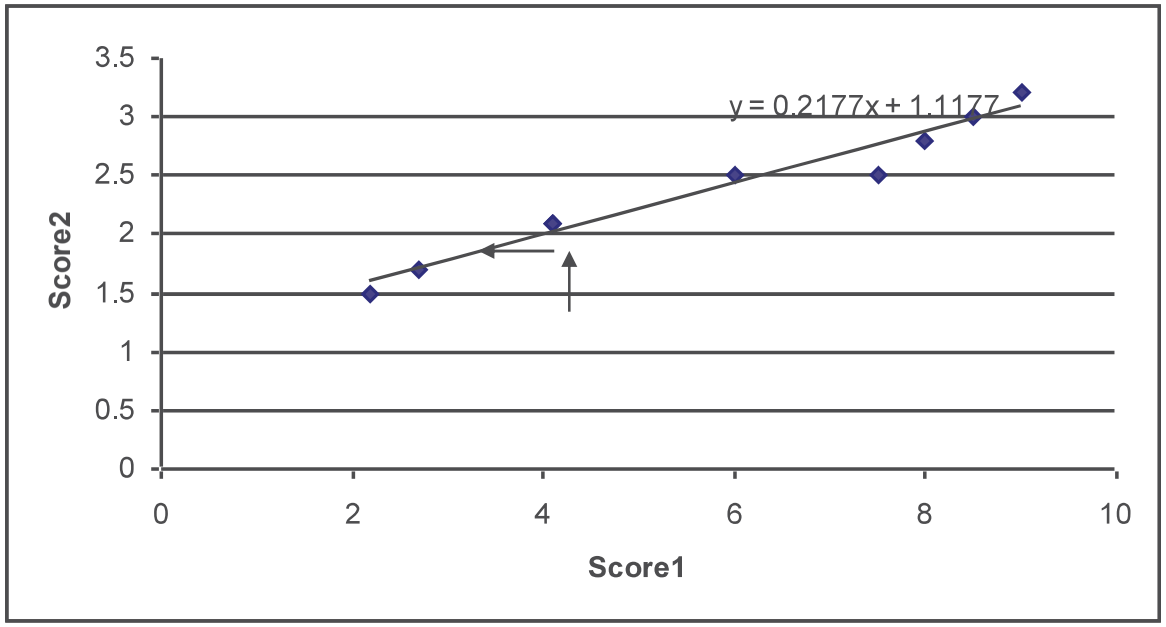
Formula per il calcolo della somma dei quadrati degli errori (SQE)

$$SQE = \text{variabilità residua}$$

$$= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} Y_i^2 - b_0 \sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i Y_i \qquad (9.21)$$

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; 8,5]. 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 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

$$SQT = SQR + SQE \tag{9.3}$$

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.

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

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*.

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

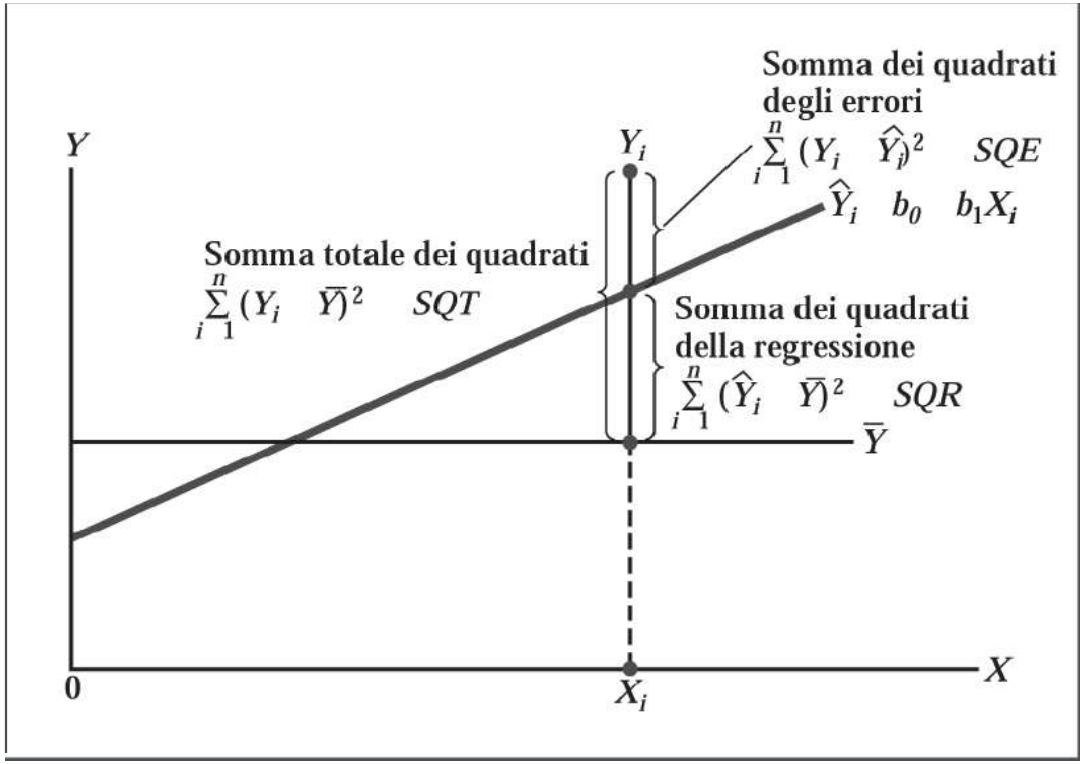
$$= SQT - SQE$$

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

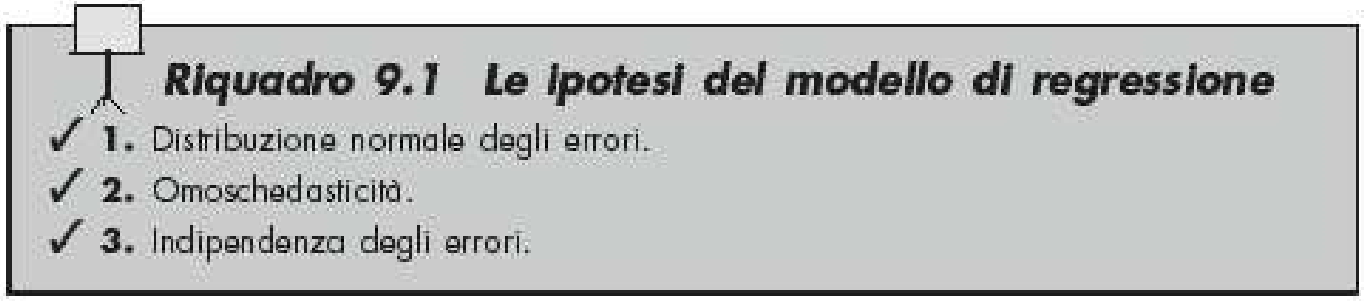
Y_i = il valore di Y in corrispondenza X_i

$\hat{Y}_i$ = 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)