

## **CAP. 6 – REGRESSIONE LINEARE SEMPLICE**

### **6.1 Introduzione**

Nel capitolo 2 è stato introdotto, quale *momento misto centrale di ordine 1,1*, uno specifico indice per la misura della relazione tra due caratteri quantitativi: il **coefficiente di correlazione lineare di Bravais-Pearson**

$$\rho = \rho_{xy} = \rho_{yx} = \frac{\overline{\mu}_{11}}{\overline{\mu}_{xy}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sigma_{yx}}{\sigma_y \sigma_x} = \frac{\text{Codev}(y,x)}{\sqrt{\text{Dev}(y)} \cdot \sqrt{\text{Dev}(x)}}$$

In quella sede è stato chiarito che il coefficiente stesso deve essere interpretato esclusivamente come misura di interdipendenza lineare che assume valore  $\pm 1$  se e solo se i due caratteri sono legati da una relazione lineare del tipo

$$y = a + b x$$

cioè, se e solo se, noto il valore assunto da uno dei due caratteri, il valore assunto dall'altro carattere risulta univocamente determinato dal legame lineare.

Ovviamente, nelle situazioni reali, una tale condizione si realizza molto raramente, molto più frequenti sono, invece, le situazioni in cui è ipotizzabile un qualche legame tra i due caratteri e nelle quali la relazione lineare, come si avrà modo di chiarire nelle pagine successive, viene assunta come misura di prima approssimazione del legame stesso.

### **6.2 Regressione lineare semplice**

Si supponga che le manifestazioni di uno specifico fenomeno, ad esempio la domanda di un certo bene di consumo da parte delle famiglie, siano rappresentate dal simbolo algebrico  $y$  e che sia possibile osservare  $n$  manifestazioni del fenomeno stesso  $y_1, y_2, \dots, y_i, \dots, y_n$ . Si ipotizzi, inoltre, che altri caratteri, ad esempio reddito disponibile, prezzo del bene, prezzo di beni sostitutivi, ecc., influiscano sulle

determinazioni  $y_i$ . Se con i simboli algebrici  $x_1, x_2, \dots, x_j, \dots, x_m$  si rappresentano tali caratteri, è ipotizzabile tra la variabile  $y$  e le variabili  $x_j$  una relazione del tipo

$$y = f(x_1, x_2, \dots, x_j, \dots, x_m)$$

che, per ciascuna unità statistica di osservazione (la famiglia), diventa

$$y_i = f(x_{1i}, x_{2i}, \dots, x_{ji}, \dots, x_{mi}) \text{ per } i = 1, 2, \dots, n.$$

Si supponga ora che le  $m$  variabili  $x_j$  possano essere distinte in tre gruppi: le prime  $k$  variabili  $(x_1, x_2, \dots, x_k)$  rappresentano fenomeni osservabili e sono molto influenti sul fenomeno  $y$ , le successive  $h$  variabili  $(x_{k+1}, x_{k+2}, \dots, x_{k+h})$ , sempre molto influenti su  $y$ , non sono osservabili, mentre le residue variabili  $(x_{k+h+1}, x_{k+h+2}, \dots, x_m)$  sono poco influenti su  $y$  e/o non sono osservabili.

Per quanto sopra detto e introducendo l'ipotesi di additività degli effetti, si può riscrivere la relazione precedente nel modo seguente

$$y = f(x_1, x_2, \dots, x_k, x_{k+1}, x_{k+2}, \dots, x_{k+h}) + v$$

dove  $v$  riassume in un'unica variabile l'effetto combinato dei fattori poco influenti.

Essendo, comunque, non osservabili le variabili  $(x_{k+1}, x_{k+2}, \dots, x_{k+h})$  si è costretti ad introdurre un'ulteriore approssimazione e, sempre nell'ipotesi di additività degli effetti, la relazione iniziale diventa

$$y = f(x_1, x_2, \dots, x_k) + w + v$$

dove la variabile  $w$  rappresenta l'effetto di fattori influenti ma non osservabili. Ovviamente, l'approssimazione ora introdotta potrebbe risultare non del tutto accettabile e compromettere, quindi, la capacità rappresentativa del modello.

Se si introduce un'ulteriore approssimazione: la linearità (dove la linearità va intesa nel senso sotto precisato) degli effetti dei fattori influenti ed osservabili si ha

$$y = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_k \cdot x_k + z + w + v$$

dove  $z$  rappresenta l'effetto non lineare su  $y$  delle variabili  $x_1, x_2, \dots, x_k$ .

La relazione (modello analitico rappresentativo del legame tra il carattere  $y$ , variabile dipendente, ed i caratteri  $x_1, x_2, \dots, x_k$ , variabili indipendenti o variabili esplicative) può essere riscritta nella forma

$$y = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_k \cdot x_k + u$$

dove  $u = z + w + v$  rappresenta la cosiddetta **componente accidentale** e  $y^* = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_k \cdot x_k$  rappresenta la **componente sistematica del modello**, usualmente detto, **di regressione lineare multipla**. Da sottolineare in proposito che la linearità è riferita ai coefficienti  $\beta_0, \beta_1, \dots, \beta_k$  e non alle variabili  $x_1, x_2, \dots, x_k$ , cioè, la generica variabile  $x_i$  può rappresentare, sia la manifestazione osservata  $x$  di un fenomeno che si ritiene influente su  $y$ , sia qualunque trasformazione nota di tale manifestazione ad esempio  $x^2, x^3, 1/x, \log x$ , ecc. .

Per  $k = 1$ , e ponendo  $x_1 = x$ , si ottiene l'espressione del **modello di regressione lineare semplice** (una sola variabile esplicativa)

$$y = \beta_0 + \beta_1 \cdot x + u = y^* + u$$

che, per le  $n$  osservazioni disponibili, diventa

$$y_i = y_i^* + u_i = \beta_0 + \beta_1 \cdot x_i + u_i \quad \text{per } i = 1, 2, \dots, n$$

dove, tenendo conto di quanto sottolineato a proposito della linearità,  $x_i$  può rappresentare, sia la manifestazione diretta (osservazione) del fenomeno rappresentato con il simbolo algebrico  $x$ , sia una qualunque trasformazione nota di tale manifestazione.

Usualmente si dispone di  $n$  coppie di osservazioni  $(y_i, x_i)$  sulle due variabili di interesse che a seconda della situazione in esame, possono essere rappresentate graficamente nelle Fig. 1 e Fig. 2 e che evidenziano, rispettivamente, il caso di una sola osservazione  $y_i$  ( $i = 1, 2, \dots, n$ ) in corrispondenza di ciascuna modalità  $x_i$  (cfr. Fig. 1), il caso di più osservazioni  $y_{ij}$  ( $i = 1, 2, \dots, s; j = 1, 2, \dots, n_i$ ) in corrispondenza di ciascuna modalità  $x_i$  (cfr. Fig. 2)..

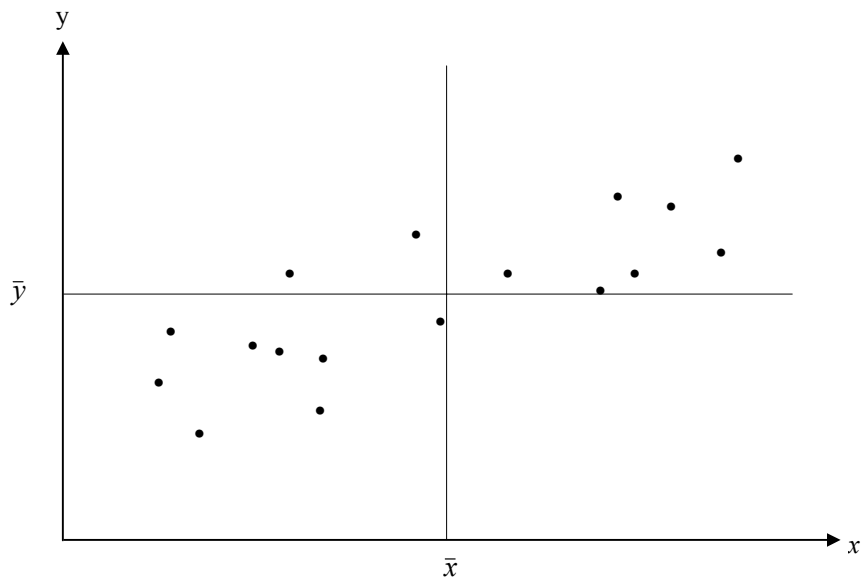


Fig. 1 – Distribuzione ipotetica di coppie di osservazioni (una sola osservazione  $y$  in corrispondenza di ciascuna modalità osservata della  $x$ ).

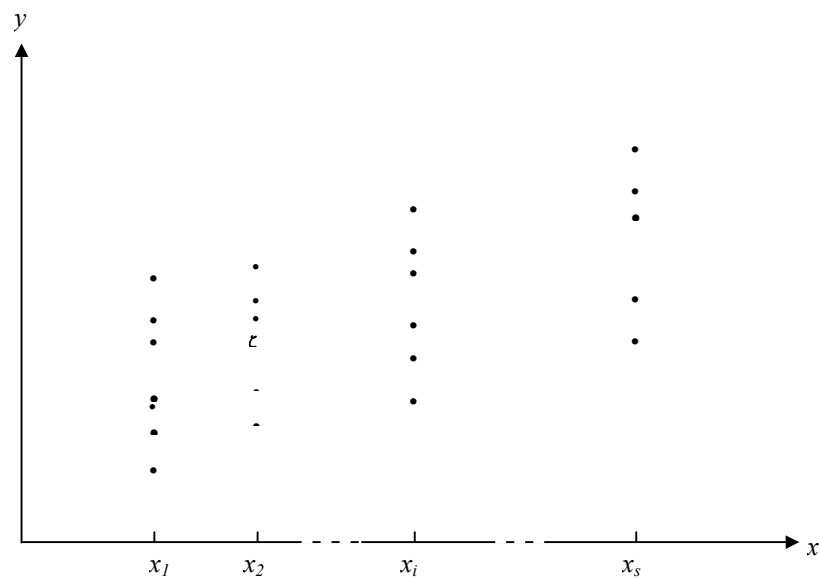


Fig. 2 - Distribuzione ipotetica di coppie di osservazioni (più osservazioni di  $y$  in corrispondenza di ciascuna modalità osservata della  $x$ ).

In entrambe le situazioni prospettate è ipotizzabile una relazione di tipo lineare tra le due variabili; infatti, si può osservare come le due rette sovrapposte alle nuvole di punti (cfr. Figg: 3 e 4) rappresentino in modo abbastanza soddisfacente l'andamento dei punti stessi.

$$y_i^* = \beta_0 + \beta_1 \cdot x_i \quad \text{per } i = 1, 2, \dots, n.$$

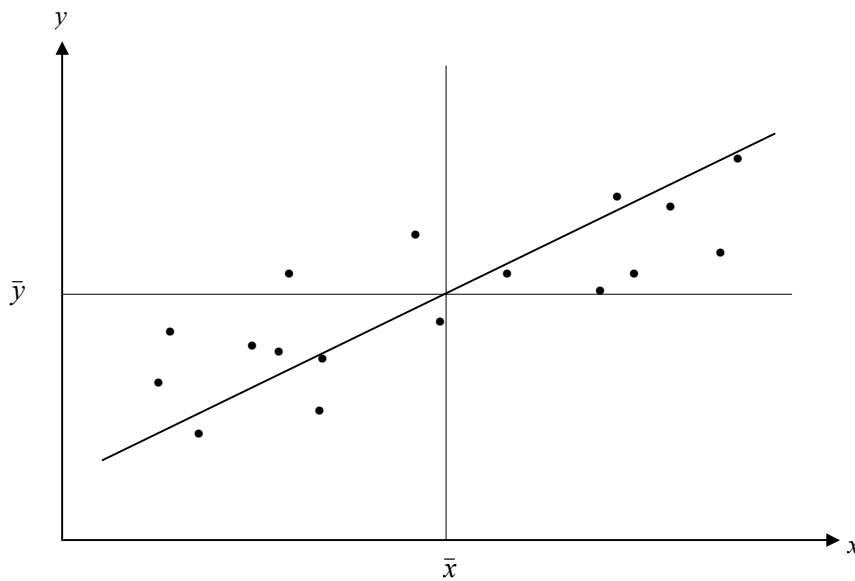


Fig. 3 – Distribuzione ipotetica di coppie di osservazioni e retta interpolante (una sola osservazione  $y$  in corrispondenza di ciascuna modalità osservata della  $x$  ).

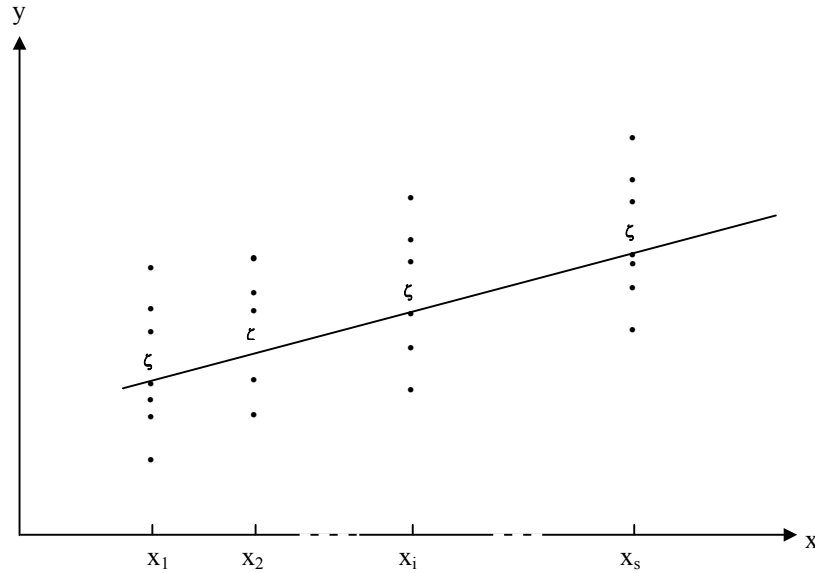


Fig. 4 - Distribuzione ipotetica di coppie di osservazioni e retta interpolante (più osservazioni di  $y$  in corrispondenza di ciascuna modalità osservata della  $x$  ).

Di rette sovrapponibili ai punti ne esistono un numero infinito, si tratta, allora, di individuare quella ritenuta migliore sulla scorta di un prefissato criterio di ottimalità, Il

problema dal punto di vista statistico è, dunque, quello di procedere alla stima ottimale dei due parametri incogniti (coefficienti che definiscono la retta)  $\beta_0$  (**intercetta**) e  $\beta_1$  (**coefficiente di regressione**) o, più in generale, utilizzare le  $n$  coppie di informazioni campionarie  $(y_i, x_i)$  per “fare” inferenza sul modello lineare che si ritiene possa rappresentare in maniera soddisfacente il legame che sussiste tra le due variabili di interesse e che in una sorta di popolazione teorica (super-popolazione) dovrebbe, prescindendo dalla componente accidentale, risultare di tipo deterministico

Se la relazione fosse perfetta in corrispondenza di ciascun valore  $x_i$  si dovrebbe osservare un unico valore  $y_i$  uguale ad  $y_i^*$ , in realtà, come già sottolineato, una tale eventualità la si riscontra molto raramente nella pratica operativa: la componente sistematica del modello spiega soltanto una parte della variabile dipendente; ad esempio, se si pensa che la domanda di un certo bene dipenda dal reddito disponibile è abbastanza ovvio ipotizzare che non tutti i soggetti in possesso di uno stesso ammontare di reddito domandino la stessa quantità del bene, la relazione tra reddito (variabile  $x$ ) e quantità del bene domandato (variabile  $y$ ) è, pertanto del tipo  $y_i = y_i^* + u_i$  e non  $y_i = y_i^*$ .

Nel modello introdotto le quantità note sono dunque  $y_i$  ed  $x_i$  mentre le quantità non note sono  $\beta_0$ ,  $\beta_1$  e, quindi,  $y_i^*$  e  $u_i$ . Si tratterà allora di utilizzare le informazioni disponibili per procedere ad una stima (puntuale o di intervallo) delle entità incognite o, eventualmente, alla verifica di ipotesi statistiche sulle entità stesse. In realtà, le entità incognite sono i due coefficienti  $\beta_0$  e  $\beta_1$  che una volta noti consentono di trarre conclusioni sia su  $y_i^*$  che su  $u_i$ .

### 6.3 Ipotesi di specificazione (caso A)

Sul modello di regressione lineare semplice vengono usualmente introdotte delle ipotesi che specificano le condizioni di base che si ritiene, quantomeno in via di prima approssimazione, debbano essere soddisfatte e che per la loro natura caratterizzano in modo particolare il modello stesso che viene detto **modello classico di regressione lineare semplice**.

Le ipotesi di specificazione riguardano la variabile (esplicativa o indipendente)  $x_i$  e, soprattutto la componente accidentale  $u_i$ :

1. le  $x_i$  ( $i = 1, 2, \dots, n$ ) sono quantità costanti in ripetuti campioni, sono, cioè, o variabili matematiche o determinazioni di variabili casuali, in quest'ultimo caso l'analisi viene effettuata condizionatamente ai valori  $x_1, x_2, \dots, x_n$ ;

2. le variabili casuali  $u_i$  hanno valore atteso (media) nullo

$$E(u_i) = 0 \quad \text{per } i = 1, 2, \dots, n;$$

3. le variabili casuali  $u_i$  hanno varianza costante (**omoschedasticità**)

$$Var(u_i) = E(u_i) = \sigma^2 \quad \text{per } i = 1, 2, \dots, n;$$

4. le variabili casuali  $u_i$  sono incorrelate (**incorrelazione**)

$$Cov(u_i, u_j) = E(u_i, u_j) = 0 \quad \text{per } i \neq j = 1, 2, \dots, n.$$

Le conseguenze sulle variabili  $y_i$  delle ipotesi introdotte sono:

- a.  $E(y_i) = E(y_i/x_i) = \beta_0 + \beta_1 \cdot x_i = y_i^* \quad \text{per } i = 1, 2, \dots, n;$

- b.  $Var(y_i) = Var(y_i/x_i) = \sigma^2 \quad \text{per } i = 1, 2, \dots, n;$

- c.  $Cov(y_i, y_j) = 0 \quad \text{per } i \neq j = 1, 2, \dots, n.$

Sulla scorta delle ipotesi di specificazione introdotte, si può procedere alla stima puntuale dei due coefficienti incogniti  $\beta_0$  e  $\beta_1$ .

Se con  $\hat{\beta}_0$  e con  $\hat{\beta}_1$  si indicano le due stime ottenute, ne risulta di conseguenza che la stima di  $y_i^*$  è data da

$$\hat{y}_i^* = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i \quad \text{per } i = 1, 2, \dots, n$$

mentre la stima di  $u_i$  è data da  $\hat{u}_i = y_i - \hat{y}_i^*$  che viene detto **residuo di regressione** o **errore di regressione**. Inoltre, a ragione dell'ipotesi  $E(u_i) = 0$ , si ha

$$\hat{y}_i^* = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i = \hat{y}_i.$$

### 6. 3. 1 Stima dei minimi quadrati

Si è già avuto modo di accennare in precedenza al **metodo di stima dei minimi quadrati** sottolineando, in particolare, il largo impiego del metodo stesso nell'ambito

dei modelli statistici lineari, il modello classico di regressione lineare costituisce la specificazione più semplice di tale classe di modelli.

Se si pone

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - y_i^*)^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 \cdot x_i)^2$$

il **metodo di stima dei minimi quadrati** si sostanzia nella ricerca dei valori  $\beta_0$  e  $\beta_1$  che minimizzano la somma dei quadrati degli scarti sopra definita. Per individuare tale minimo basterà determinare il punto di stazionarietà (che è sicuramente un punto di minimo avendo a che fare con una funzione quadratica il cui punto di massimo è infinito) della funzione  $Q(\beta_0, \beta_1)$  che si ottiene risolvendo il sistema:

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = 0$$

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

che diventa

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \left[ \sum_{i=1}^n (y_i - \beta_0 - \beta_1 \cdot x_i)^2 \right] = -2 \left( \sum_{i=1}^n y_i - n \beta_0 - \beta_1 \cdot \sum_{i=1}^n x_i \right) = 0$$

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = \frac{\partial}{\partial \beta_1} \left[ \sum_{i=1}^n (y_i - \beta_0 - \beta_1 \cdot x_i)^2 \right] = -2 \left( \sum_{i=1}^n y_i x_i - \beta_0 \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2 \right) = 0$$

cioè

$$\sum_{i=1}^n y_i = n \beta_0 + \beta_1 \sum_{i=1}^n x_i$$

$$\sum_{i=1}^n y_i x_i = \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2$$

Risolvendo il sistema delle due equazioni nelle due incognite  $\beta_0$  e  $\beta_1$  si ottiene

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{Codev}(y, x)}{\text{Dev}(x)} = \frac{\sigma_{xy}}{\sigma_x^2} = b_{y/x}$$

dove  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  e  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ .

E' possibile a questo punto riproporre le Figg. 3 e 4 dove le rette interpolanti non sono più rette generiche ma quelle (cfr. Figg. 5 e 6) che derivano dall'applicazione del metodo dei minimi quadrati (**rette dei minimi quadrati**). Nella Fig. 6 è stata inserita anche l'ipotesi di normalità dei valori assunti dalla variabile  $y$  in corrispondenza di ciascun valore assunto dalla variabile  $x$ ; ma su quest'ultimo aspetto si avrà modo di tornare successivamente.

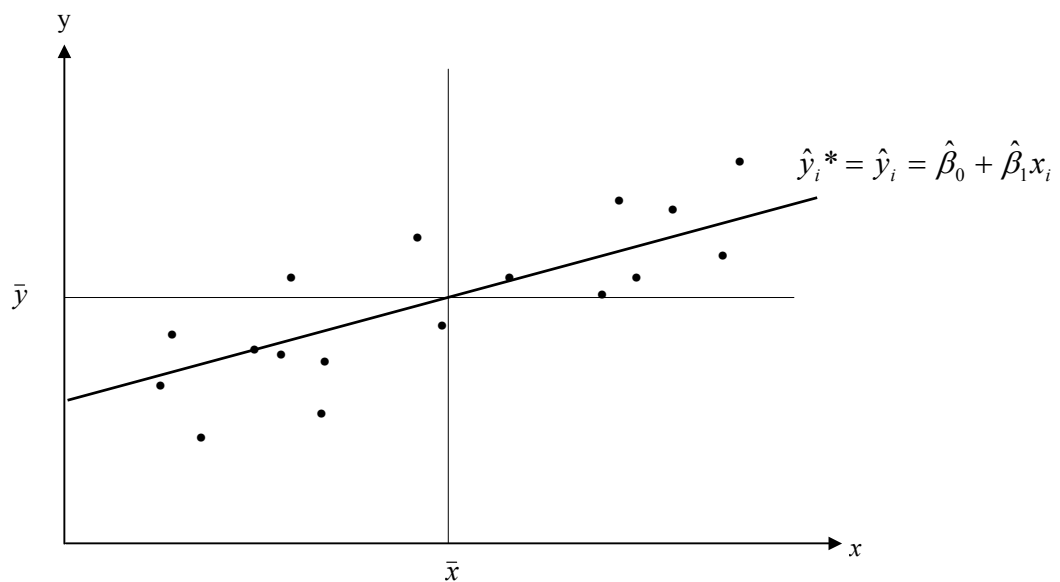


Fig. 5 – Distribuzione ipotetica di coppie di osservazioni e retta dei minimi quadrati (una sola osservazione  $y$  in corrispondenza di ciascuna modalità osservata della  $x$ ).

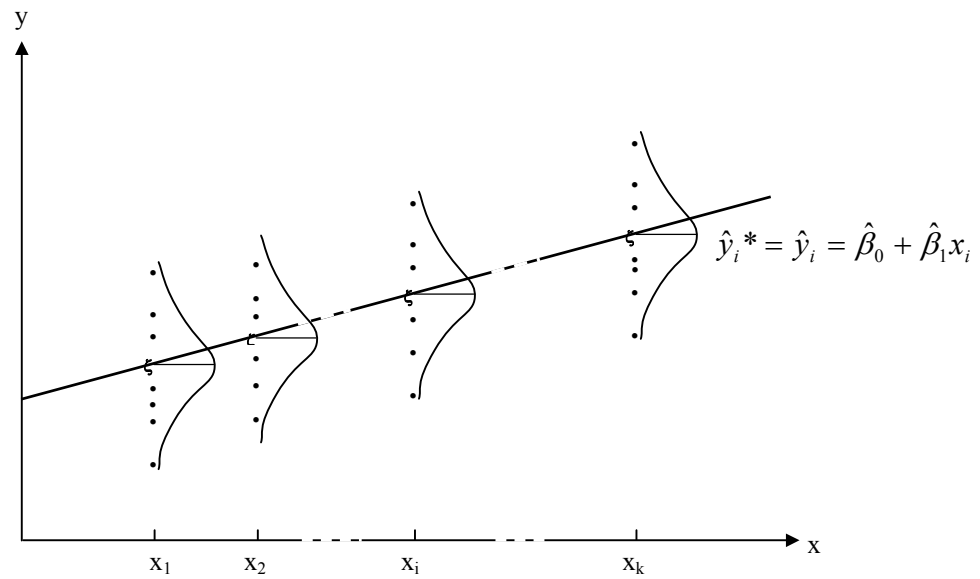


Fig. 6 - Distribuzione ipotetica di coppie di osservazioni e retta dei minimi quadrati (più osservazioni di  $y$  in corrispondenza di ciascuna modalità osservata della  $x$ ).

Le varianze degli stimatori sono:

$$Var(\hat{\beta}_0) = \sigma_{\hat{\beta}_0}^2 = \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \cdot \sigma^2$$

$$Var(\hat{\beta}_1) = \sigma_{\hat{\beta}_1}^2 = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sigma^2$$

$$\hat{V}ar(\hat{y}_i^*) = \hat{\sigma}_{\hat{y}_i^*}^2 = \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2} \right) \cdot \hat{\sigma}^2$$

infatti, valendo le relazioni di uguaglianza:

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \cdot \bar{x} = \frac{1}{n} \sum_{j=1}^n y_j - \frac{\sum_{j=1}^n (x_j - \bar{x}) \cdot y_j}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \bar{x} = \\ &= \sum_{j=1}^n \left( \frac{1}{n} - \frac{(x_j - \bar{x}) \cdot \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \cdot y_j = \sum_{j=1}^n a_j \cdot y_j \\ \text{dove } a_j &= \frac{1}{n} - \frac{(x_j - \bar{x}) \cdot \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_1 &= \frac{\sum_{j=1}^n (x_j - \bar{x}) \cdot y_j}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{j=1}^n \left( \frac{x_j - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \cdot y_j = \sum_{j=1}^n b_j \cdot y_j \\ \text{dove } b_j &= \frac{x_j - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{y}_i^* &= \sum_{j=1}^n \left( \frac{1}{n} + \frac{(x_i - \bar{x})}{\sum_{r=1}^n (x_r - \bar{x})^2} (x_j - \bar{x}) \right) \cdot y_j = \sum_{j=1}^n c_j \cdot y_j \\ \text{dove } c_j &= \frac{1}{n} + \frac{(x_i - \bar{x})}{\sum_{r=1}^n (x_r - \bar{x})^2} (x_j - \bar{x}) \end{aligned}$$

e ricordando che la varianza di una combinazione lineare di variabili casuali indipendenti è pari alla combinazione delle varianze delle singole variabili casuali con coefficienti elevati al quadrato si ha:

$$Var \left( \sum_{i=1}^n a_i \cdot y_i \right) = \sum_{i=1}^n a_i^2 \cdot Var(y_i) = \sigma^2 \sum_{i=1}^n a_i^2, \text{ da cui:}$$

$$\begin{aligned}
 \text{Var}(\hat{\beta}_0) &= \text{Var}\left(\sum_{j=1}^n a_j \cdot y_j\right) = \sum_{j=1}^n a_j^2 \cdot \text{var}(y_j) = \sum_{j=1}^n \left(\frac{1}{n} - \frac{(x_j - \bar{x}) \cdot \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^2 \cdot \sigma^2 = \\
 &= \left(\sum_{j=1}^n \frac{1}{n^2} + \sum_{j=1}^n \frac{(x_j - \bar{x})^2 \cdot \bar{x}^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2}\right) \cdot \sigma^2 = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \cdot \sigma^2 \\
 \text{Var}(\hat{\beta}_1) &= \text{Var}\left(\sum_{j=1}^n b_j \cdot y_j\right) = \sum_{j=1}^n b_j^2 \cdot \text{var}(y_j) = \sum_{j=1}^n \left(\frac{x_j - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \cdot \sigma^2 = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sigma^2 \\
 \text{Var}(\hat{y}_i^*) &= \text{Var}\left(\sum_{j=1}^n c_j \cdot y_j\right) = \sum_{j=1}^n c_j^2 \cdot \text{var}(y_j) = \sum_{j=1}^n \left(\frac{1}{n} + \frac{(x_i - \bar{x})}{\sum_{r=1}^n (x_r - \bar{x})^2} (x_j - \bar{x})\right)^2 \cdot \sigma^2 = \\
 &= \left[\sum_{i=1}^n \frac{1}{n^2} + \sum_{i=1}^n \left(\frac{(x_i - \bar{x})}{\sum_{r=1}^n (x_r - \bar{x})^2} (x_j - \bar{x})\right)^2\right] \cdot \sigma^2 = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{r=1}^n (x_j - \bar{x})^2}\right) \cdot \sigma^2
 \end{aligned}$$

Seguendo la stessa procedura risulta facile anche la derivazione della covarianza tra le due variabili casuali stima  $\hat{\beta}_0$  e  $\hat{\beta}_1$ . Si ha

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{j=1}^n \left(\frac{1}{n} - \frac{(x_j - \bar{x}) \cdot \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \cdot \left(\frac{x_j - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \cdot \sigma^2 = -\frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sigma^2$$

Per quanto sopra detto si ottiene

$$\hat{y}_i^* = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

Se si procede al calcolo della varianza dello stimatore  $\hat{y}_i^*$  basandosi su questa espressione si ha:

$$\begin{aligned} \text{Var}(\hat{y}_i^*) &= \sigma_{\hat{y}_i^*}^2 = \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i) = \text{Var}(\hat{\beta}_0) + x_i^2 \text{Var}(\hat{\beta}_1) + 2x_i \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \\ &= \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \cdot \sigma^2 + x_i^2 \cdot \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sigma^2 - 2x_i \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sigma^2 \\ &= \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2} \right) \cdot \sigma^2 \end{aligned}$$

che coincide con l'espressione già ottenuta.

Si sottolinea che la quantità  $\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$  è il migliore previsore (stima) di  $y_i$ , cioè, quello che sotto le ipotesi specificate minimizza l'errore quadratico medio (= alla varianza trattandosi di stimatore non distorto) e che le stime dei minimi quadrati godono delle proprietà specificate dal teorema che segue.

**Teorema 1 (Gauss-Markov):** Le stime dei minimi quadrati di  $\beta_0$  e  $\beta_1$  sono di minima varianza nell'ambito delle stime lineari e corrette (**BLUE** dall'inglese **Best Linear Unbiased Estimator**).

*Dimostrazione*

Si procederà alla dimostrazione per  $\hat{\beta}_1$ , considerazioni analoghe possono essere svolte nei confronti di  $\hat{\beta}_0$ .

Lo stimatore  $\hat{\beta}_1$  è lineare e corretto; infatti:

$$\hat{\beta}_1 = \frac{\sum_{j=1}^n (x_j - \bar{x}) \cdot y_j}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{j=1}^n \left( \frac{x_j - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \cdot y_j = \sum_{j=1}^n b_j \cdot y_j \quad \text{(linearità)}$$

inoltre

$$\begin{aligned}
 E(\hat{\beta}_1) &= E\left[ \sum_{j=1}^n \left( \frac{x_j - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \cdot y_j \right] = \sum_{j=1}^n \left( \frac{x_j - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \cdot E(y_j) = \\
 &= \sum_{j=1}^n \left( \frac{x_j - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \cdot (\beta_0 + \beta_1 \cdot x_j) = \frac{\sum_{j=1}^n x_j (x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \beta_1 = \beta_1
 \end{aligned}$$

(correttezza)

Si consideri ora un generico stimatore lineare e corretto di  $\beta_1$ , ad esempio

$$\hat{\beta}_1 = \sum_{j=1}^n \alpha_j y_j, \text{ dove, per il vincolo di correttezza deve risultare}$$

$$E(\hat{\beta}_1) = E\left( \sum_{j=1}^n \alpha_j y_j \right) = \sum_{j=1}^n \alpha_j E(y_j) = \sum_{j=1}^n \alpha_j (\beta_0 + \beta_1 \cdot x_j) = \beta_1$$

cioè

$$\sum_{j=1}^n \alpha_j = 0 \quad \text{e} \quad \sum_{j=1}^n \alpha_j \cdot x_j = 1$$

Tenendo conto di quanto sopra scritto, si vuol dimostrare che  $Var\hat{\beta}_1 \geq Var\hat{\beta}_1$ .

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \sum_{j=1}^n \alpha_j^2 \cdot \text{Var}(y_j) = \sigma^2 \cdot \sum_{j=1}^n \alpha_j^2 = \sigma^2 \cdot \sum_{j=1}^n (\alpha_j - b_j + b_j)^2 = \\ \text{dove } b_j &= \frac{x_j - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \sigma^2 \cdot \left[ \sum_{j=1}^n (\alpha_j - b_j)^2 + \sum_{j=1}^n b_j^2 + 2 \cdot \sum_{j=1}^n (\alpha_j - b_j) \cdot b_j \right] = \\ &= \sigma^2 \cdot \sum_{j=1}^n (\alpha_j - b_j)^2 + \sigma^2 \cdot \sum_{j=1}^n b_j^2 + 2 \left[ \sum_{j=1}^n \alpha_j \cdot b_j - \sum_{j=1}^n b_j^2 \right] = \\ &= \sigma^2 \cdot \sum_{j=1}^n (\alpha_j - b_j)^2 + \text{Var}(\hat{\beta}_0) + 2 \cdot \left[ \sum_{j=1}^n \frac{\alpha_j \cdot x_j}{\sum_{i=1}^n (x_i - \bar{x})^2} - \sum_{j=1}^n \frac{\alpha_j \cdot \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} - \sum_{j=1}^n \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] = \end{aligned}$$

ma, per il vincolo di correttezza

$$\sum_{j=1}^n \alpha_j = 0 \quad \text{e} \quad \sum_{j=1}^n \alpha_j \cdot x_j = 1, \text{ quindi}$$

$$\text{Var}(\hat{\beta}_1) = \sigma^2 \cdot \sum_{j=1}^n (\alpha_j - b_j)^2 + \text{Var}(\hat{\beta}_0) \geq \text{Var}\hat{\beta}_0$$

dove, il segno di uguaglianza vale solo quando  $\alpha_j = b_j$ .

Come si può osservare le varianze degli stimatori  $\hat{\beta}_0$ ,  $\hat{\beta}_1$  e  $\hat{y}_i^*$  dipendono dalla varianza  $\sigma^2$  (parametro di disturbo), usualmente incognita, della componente accidentale. Una stima corretta di tale parametro è data da

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

dove, come già sottolineato,  $\hat{u}_i = (y_i - \hat{y}_i^*)$  rappresentano i residui di regressione (cfr. Fig. 7), mentre la stima della varianza della componente accidentale viene usualmente detta **varianza residua** e misura la parte (stimata) della **variabilità** della  $y_i$  (variabile dipendente) **non spiegata dalla variabile esplicativa**  $x_i$  (variabile indipendente).

Da quanto detto risultano le seguenti **stime delle varianze degli stimatori**

$$\hat{Var}(\hat{\beta}_0) = \hat{\sigma}_{\hat{\beta}_0}^2 = \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \cdot \hat{\sigma}^2$$

$$\hat{Var}(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1}^2 = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \hat{\sigma}^2$$

$$\hat{Var}(\hat{y}_i^*) = \hat{\sigma}_{\hat{y}_i^*}^2 = \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2} \right) \cdot \hat{\sigma}^2$$

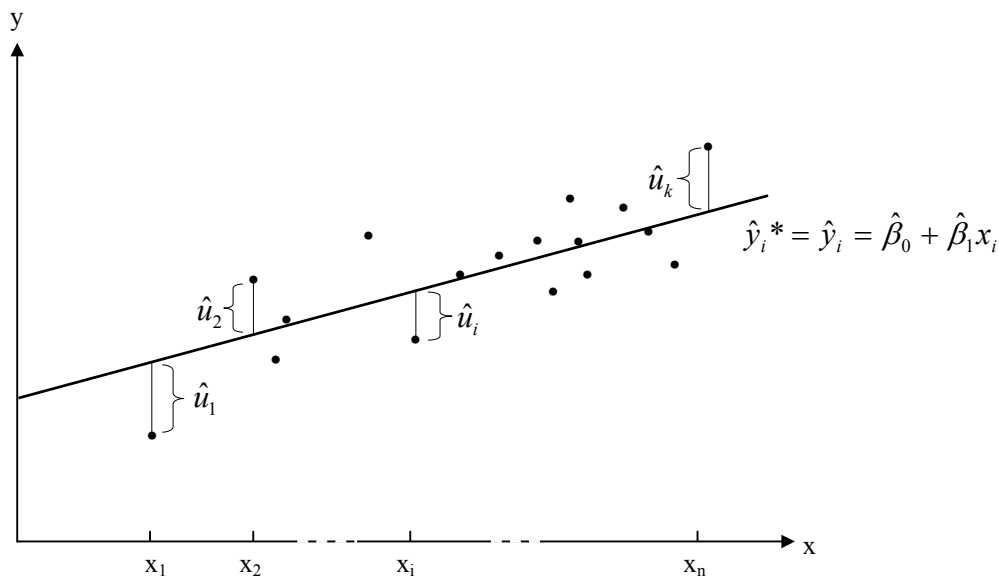


Fig. 7 - Distribuzione ipotetica di coppie di osservazioni e retta dei minimi quadrati e residui di regressione (una sola osservazione  $y$  in corrispondenza di ciascuna modalità osservata della  $x$ ).

#### 6. 4 Ipotesi di specificazione (caso B: normalità della componente accidentale)

Se alle quattro ipotesi di specificazione introdotte in precedenza si aggiunge l'ulteriore ipotesi di **normalità della distribuzione della componente accidentale**

$$u_i \sim N(0, \sigma^2) \quad \text{per } i = 1, 2, \dots, n$$

ne deriva, come conseguenza diretta, la normalità della distribuzione delle  $y_i$

$$\text{i. } y_i \sim N(\beta_0 + \beta_1 \cdot x_i, \sigma^2)$$

inoltre, ricordando che nel caso di variabili casuali normali la correlazione nulla implica l'indipendenza, le variabili casuali  $y_i$  risultano statisticamente indipendenti, da cui:

$$\text{ii. } \hat{\beta}_0 \sim N(\beta_0, \sigma_{\hat{\beta}_0}^2)$$

$$\text{iii. } \hat{\beta}_1 \sim N(\beta_1, \sigma_{\hat{\beta}_1}^2)$$

$$\text{iv. } \hat{y}_i^* \sim N(y_i^*, \sigma_{\hat{y}_i^*}^2)$$

$$\text{v. } \frac{(n-2) \cdot \hat{\sigma}^2}{\sigma^2} \sim \frac{\sum_{i=1}^n \hat{u}_i^2}{\sigma^2} \sim \chi_{n-2}^2$$

L'ipotesi di normalità già introdotta nella Fig. 6, trova una più esplicita rappresentazione nella Fig. 8.

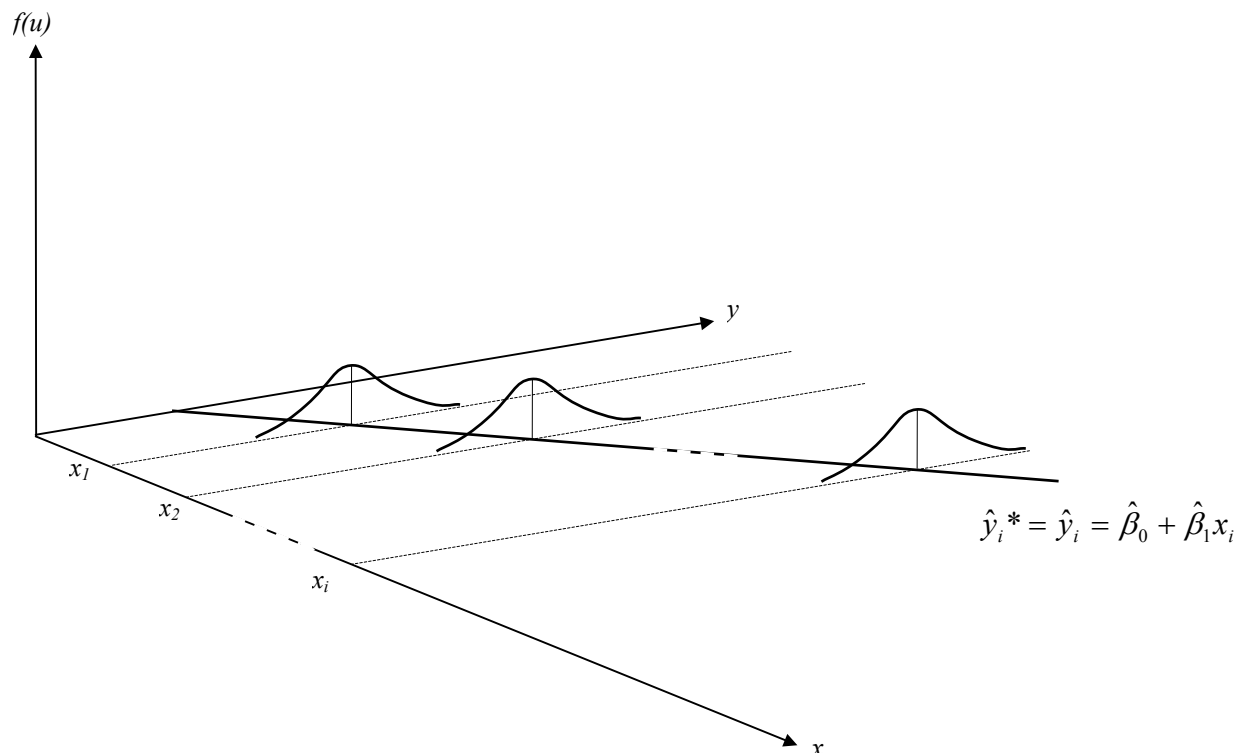


Fig. 8 – Ipotesi di distribuzione normale della componente accidentale nel modello di regressione lineare semplice

Le conseguenze espresse ai punti i., ii., iii., iv. e v. sono di immediata verifica; infatti:

- i) le variabili  $y_i = \beta_0 + \beta_1 \cdot x_i + u_i$  sono distribuite normalmente in quanto trasformazioni di variabili casuali normali;
- ii) le variabili  $\hat{\beta}_0$ ,  $\hat{\beta}_1$  e  $\hat{y}_i^*$  sono distribuite normalmente in quanto espresse da combinazioni lineari di variabili casuali normali indipendenti.

Meno immediata è la conseguenza espressa al punto v. . In proposito si deve sottolineare che gli  $(n-2)$  gradi di libertà derivano dal fatto che rispetto agli  $n$  gradi di libertà originari (le  $n$  osservazioni campionarie), due gradi di libertà si perdono nella operazione di stima; infatti, vengono imposti due vincoli per ottenere le stime di  $\beta_0$  e  $\beta_1$ . Pertanto, mentre le  $y_i$  costituiscono  $n$  variabili casuali indipendenti, le  $n$  variabili casuali  $\hat{y}_i^*$ , devono soddisfare i due vincoli introdotti per ottenere le stime  $\beta_0$  e  $\beta_1$ . Inoltre, nell'universo dei campioni, le due variabili casuali stima  $\beta_0$  e  $\beta_1$  hanno

distribuzione indipendente dalla variabile casuale  $W = \frac{\sum_{i=1}^n \hat{u}_i^2}{\sigma^2}$  che ha, come già sottolineato, una distribuzione di tipo  $\chi^2$  con  $n-2$  gradi di libertà.

#### 6.4.1 Stime di massima verosimiglianza

L'introduzione dell'ipotesi di normalità consente il calcolo della verosimiglianza del campione e di procedere, pertanto, all'uso del metodo della massima verosimiglianza per ottenere la stima dei parametri incogniti  $\beta_0, \beta_1$  e  $\sigma^2$ .

La verosimiglianza del campione è data da

$$L(\beta_0, \beta_1, \sigma^2 / y_1, y_2, \dots, y_n; x_1, x_2, \dots, x_n) = L(\beta_0, \beta_1, \sigma^2 / \underline{y}, \underline{x}) = L(\beta_0, \beta_1, \sigma^2) =$$

$$= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 \cdot x_i)^2} = (2\pi\sigma^2)^{-n/2} \cdot e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 \cdot x_i)^2}$$

Le **stime di massima verosimiglianza** dei parametri incogniti si ottengono facilmente derivando ed uguagliando a zero le derivate del logaritmo della verosimiglianza.

Risulta facile verificare che le stime di massima verosimiglianza  $\tilde{\beta}_0$  e  $\tilde{\beta}_1$  coincidono con le stime dei minimi quadrati  $\hat{\beta}_0$  e  $\hat{\beta}_1$ , mentre la stima di massima verosimiglianza della varianza  $\sigma^2$  è data da  $\tilde{\sigma}^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n}$ ; ovviamente,  $\tilde{u}_i = \hat{u}_i$  e  $\tilde{y}_i = \hat{y}_i = \tilde{y}_i^* = \hat{y}_i^*$ .

Si segnala che per derivare le stime di massima verosimiglianza  $\tilde{\beta}_0$  e  $\tilde{\beta}_1$  si può anche evitare il ricorso alla derivazione della verosimiglianza (o della log-verosimiglianza); infatti, al riguardo basta osservare che il massimo della verosimiglianza rispetto a  $\beta_0$  e  $\beta_1$  si ottiene quando è minima la quantità

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 \cdot x_i)^2$$

che è l'espressione di base del metodo dei minimi quadrati.

Relativamente alle stime di massima verosimiglianza ottenute si deve sottolineare che gli stimatori  $\tilde{\beta}_0$  e  $\tilde{\beta}_1$  pur coincidendo numericamente con gli stimatori  $\hat{\beta}_0$  e  $\hat{\beta}_1$  da questi si diversificano in quanto (**Teorema di Rao**) sono di minima varianza nell'ambito degli stimatori corretti (**BUE** dall'inglese **Best Unbiased Estimator**), inoltre, la stima  $\tilde{\sigma}^2$  della varianza  $\sigma^2$  non è corretta, cioè,  $E(\tilde{\sigma}^2) \neq \sigma^2$ .

#### 6.4.2 Stime di intervallo

Per quanto sopra richiamato, si può procedere facilmente alla derivazione delle stime di intervallo (intervalli di confidenza) per i parametri incogniti  $\beta_0$ ,  $\beta_1$  e  $\sigma^2$  e per le quantità  $y_i^*$  e  $y_i$ . Infatti, facendo riferimento alla situazione più usuale, che è quella della non conoscenza del valore assunto dal parametro di disturbo  $\sigma^2$  (varianza della componente accidentale), per  $\alpha$  prefissato si ottengono gli intervalli sotto riportati

$$P\left(\hat{\beta}_0 - t_{\alpha/2} \cdot \hat{\sigma}_{\hat{\beta}_0} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2} \cdot \hat{\sigma}_{\hat{\beta}_0}\right) = 1 - \alpha$$

$$P\left(\hat{\beta}_1 - t_{\alpha/2} \cdot \hat{\sigma}_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2} \cdot \hat{\sigma}_{\hat{\beta}_1}\right) = 1 - \alpha$$

$$P \left[ \frac{(n-2) \cdot \hat{\sigma}^2}{\chi_{\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-2) \cdot \hat{\sigma}^2}{\chi_{1-\alpha/2}^2} \right] = 1 - \alpha$$

Si segnala che l'ultimo intervallo è stato derivato distribuendo simmetricamente il valore di  $\alpha$  nelle due code della distribuzione e che l'intervallo per  $\beta_0$  si ottiene attraverso i passaggi sotto riportati (ragionamento analogo vale per l'intervallo relativo a  $\beta_1$ ).

Poiché

$$\hat{\beta}_0 \sim N(\beta_0, \sigma_{\hat{\beta}_0}^2)$$

si avrà

$$Z_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} \sim N(0, 1)$$

che non è elemento pivotale essendo incognita la varianza  $\sigma_{\hat{\beta}_0}^2$  dove è presente la varianza della componente accidentale; infatti

$$\sigma_{\hat{\beta}_0}^2 = \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \cdot \sigma^2$$

ma, se si tiene presente che

$$W = \frac{(n-2) \cdot \hat{\sigma}^2}{\sigma^2} \sim \chi_{n-2}^2$$

e che  $Z_{\hat{\beta}_0}$  e  $W$  sono variabili casuali indipendenti, si ha

$$T_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} / \sqrt{\frac{W}{(n-2)}} = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t_{n-2}$$

che rappresenta la variabile casuale  $t$  di Student con  $(n-2)$  gradi di libertà (elemento pivotale) che consente la derivazione dell'intervallo sopra riportato applicando il procedimento di derivazione degli intervalli di confidenza illustrato nel Cap. 4.

L'intervallo di stima relativo alle variabili  $y$  assume particolare rilevanza; infatti, un tale intervallo può interessare sia valori corrispondenti a valori osservati di  $x$ , cioè  $(x_1, x_2, \dots, x_n)$ , sia valori non osservati di tale variabile. Ad esempio, si potrebbe aver interesse a determinare un intervallo di stima per  $y_p^*$  che corrisponde ad un valore

non osservato  $x_0$  ma assumibile dalla variabile  $x$ ; in quest'ultimo caso, l'intervallo assume la particolare connotazione di **intervallo di previsione** e la quantità  $y_p - \hat{\beta}_0 - \hat{\beta}_1 x_p = \hat{u}_p$  viene detto **errore di previsione**.

Ipotizzando la non conoscenza della varianza  $\sigma^2$  della componente accidentale, l'intervallo per un generico valore  $y_i^*$  può essere determinato facendo riferimento alla variabile casuale  $t$  di Student (elemento pivotale)

$$T_{y_i^*} = \frac{\hat{y}_i^* - y_i^*}{\hat{\sigma}_{\hat{y}_i^*}} = \frac{\hat{y}_i^* - y_i^*}{\hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}}} =$$

ed anche, se interessa l'intervallo di previsione per  $y_p^*$

$$T_{y_p^*} = \frac{\hat{y}_p^* - y_p^*}{\hat{\sigma}_{\hat{y}_p^*}} = \frac{\hat{y}_p^* - y_p^*}{\hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}}}$$

Gli intervalli, per un prefissato livello di confidenza  $1-\alpha$ , sono rispettivamente

$$P\left(\hat{y}_i^* - t_{\alpha/2} \cdot \hat{\sigma}_{\hat{y}_i^*} \leq y_i^* \leq \hat{y}_i^* + t_{\alpha/2} \cdot \hat{\sigma}_{\hat{y}_i^*}\right) = 1-\alpha$$

ed anche

$$P\left[\hat{y}_i^* - t_{\alpha/2} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}} \leq y_i^* \leq \hat{y}_i^* + t_{\alpha/2} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}}\right] = 1-\alpha$$

$$P\left(\hat{y}_p^* - t_{\alpha/2} \cdot \hat{\sigma}_{\hat{y}_p^*} \leq y_p^* \leq \hat{y}_p^* + t_{\alpha/2} \cdot \hat{\sigma}_{\hat{y}_p^*}\right) = 1-\alpha$$

ed anche

$$P\left[\hat{y}_p^* - t_{\alpha/2} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}} \leq y_p^* \leq \hat{y}_p^* + t_{\alpha/2} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}}\right] = 1-\alpha$$

Capita spesso, e ciò avviene soprattutto quando si vogliono effettuare previsioni, di essere interessati alla determinazione di intervalli di stima non per il valore teorico  $y^*$  (cioè il valore che dovrebbe assumere la variabile dipendente in assenza di effetti accidentali e che è uguale, per le ipotesi di specificazione introdotte, al valore medio di  $y$ ) ma per il valore effettivo  $y$  (valore osservato od osservabile che include, quindi, anche l'effetto della componente accidentale).

Per perseguire un tale obiettivo si deve osservare che, come già sottolineato, le stime puntuali di un generico valore  $\hat{y}_h^*$  e  $\hat{y}_h$ , corrispondente ad una determinazione  $x_h$  ( $h = i$  o qualunque altro indice), coincidono, cioè  $\hat{y}_h^* = \hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_h$ , le loro varianze sono però diverse; infatti, se si considera l'errore di previsione  $\hat{u}_h = y_h - \hat{y}_h$  si ha:

$$\begin{aligned} E(\hat{u}_h) &= E(\beta_0 + \beta_1 x_h + u_h - \hat{\beta}_0 + \hat{\beta}_1 x_h) = 0 \\ \text{Var}(\hat{u}_h) &= E(\hat{u}_h^2) = E\left[\left(\beta_0 + \beta_1 x_h + u_h - \hat{\beta}_0 - \hat{\beta}_1 x_h\right)^2\right] = \\ &= E\left\{\left[\left(\beta_0 - \hat{\beta}_0\right) + \left(\beta_1 - \hat{\beta}_1\right) x_h + u_h\right]^2\right\} = \\ &= \text{Var}(\hat{\beta}_0) + \text{Var}(\hat{\beta}_1) \cdot x_h^2 + \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \cdot x_h + \text{Var}(u_h) = \\ &= \sigma^2 \cdot \left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}\right) \end{aligned}$$

pertanto l'intervallo di confidenza per  $y_h$  è dato da

$$P\left(\hat{y}_h - t_{\alpha/2} \cdot \hat{\sigma}_{\hat{y}_h} \leq y_h \leq \hat{y}_h + t_{\alpha/2} \cdot \hat{\sigma}_{\hat{y}_h}\right) = 1 - \alpha$$

ed anche

$$P\left[\hat{y}_h - t_{\alpha/2} \cdot \hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}} \leq y_h \leq \hat{y}_h + t_{\alpha/2} \cdot \hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}}\right] = 1 - \alpha$$

L'intervallo di stima per  $y_h$  risulta più ampio di quello relativo ad  $y_h^*$ ; infatti: alla variabilità dovuta alla stima di  $\beta_0$  e  $\beta_1$  si aggiunge la variabilità indotta dalla componente accidentale  $u_h$ ; inoltre, l'ampiezza degli intervalli così determinati

dipendono fortemente dallo scarto  $(x_h - \bar{x})$  e risultano tanto più ampi quanto più il valore di riferimento della  $x$  si allontana dal suo valore medio  $\bar{x}$ . L'evidenziazione grafica di tale situazione è riportata nella Fig. 9.

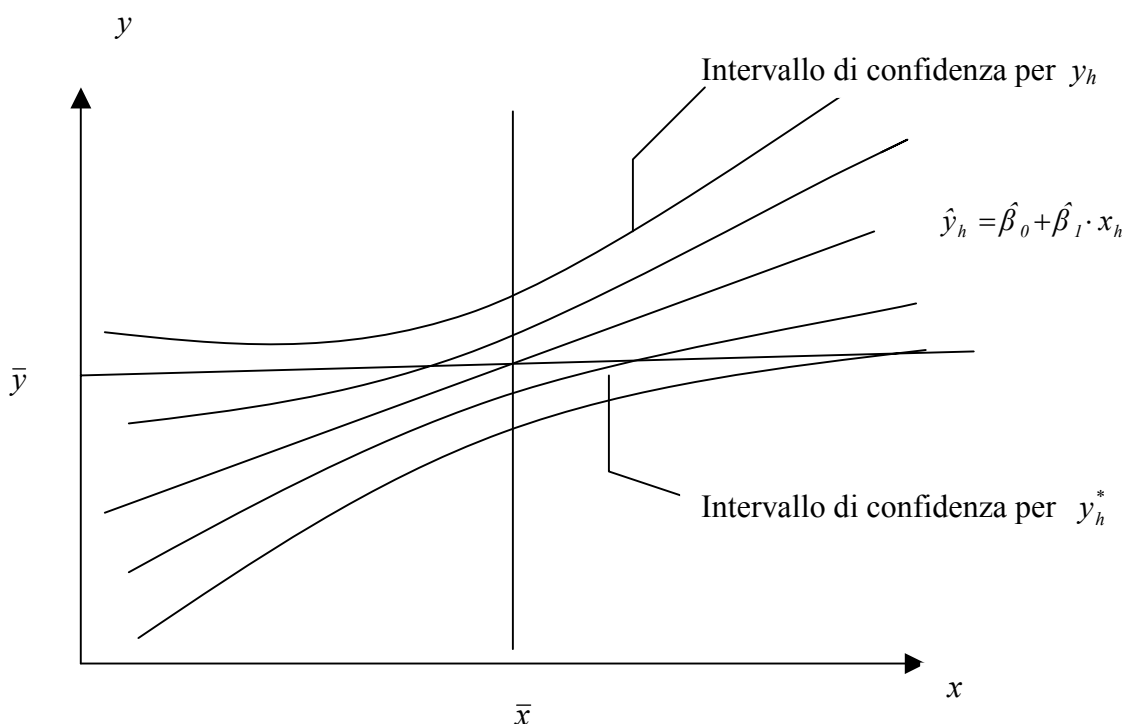


Fig. 9 – Intervalli di confidenza per i valori medi  $y_h^*$  e per i valori individuali  $y_h$ .

### 6. 4. 3 Test delle ipotesi

Per quanto detto nelle pagine precedenti e nel Cap. 5, è ora possibile risolvere facilmente qualunque problema di test delle ipotesi riguardo alle entità incognite presenti nel modello di regressione lineare semplice. Infatti, sotto la condizione di normalità della distribuzione della componente accidentale, basterà fare riferimento alle variabili casuali (variabili casuali test)  $T_{\hat{\beta}_0}, T_{\hat{\beta}_1}, T_{\hat{y}_i^*}, T_{\hat{y}_i}$  e  $W$  sopra definite.

Se, ad esempio, si volesse risolvere il problema di test delle ipotesi

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

la regione di rifiuto dell'ipotesi nulla (nessun effetto della supposta variabile esplicativa  $x$  sulla variabile dipendente  $y$ ) risulterebbe definita dai semintervalli

$-\infty \text{ --- } | -t_{\alpha/2}, t_{\alpha/2} \text{ --- } +\infty :$

Se il problema di test fosse

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 < 0$$

cioè, di effetto nullo contro effetto negativo (e questo potrebbe essere un caso di interesse quando, ad esempio,  $x$  rappresenta il prezzo di un certo bene ed  $y$  la domanda del bene stesso: al crescere del prezzo la domanda del bene dovrebbe diminuire). La regione critica del test (rifiuto dell'ipotesi nulla) è costituita dal semintervallo  $-\infty \text{ --- } | -t_{\alpha}$ .

Le procedure di test sopra richiamate derivano dall'applicazione del test del rapporto di verosimiglianza che, come già sottolineato, fornisce (quando esiste, ed i casi considerati rientrano in questa categoria) il test uniformemente più potente, nel caso di ipotesi alternativa unidirezionale, il test uniformemente più potente nella classe dei test non distorti, nel caso di ipotesi alternativa bidirezionale.

## 6.5 Trasformazioni di modelli non lineari nei coefficienti

È stato precisato che la linearità del modello di regressione semplice è riferita ai coefficienti e non alla variabile; infatti, ad esempio, il modello  $y = \beta_0 + \beta_1 \cdot x^3$  è perfettamente equivalente al modello  $y = \beta_0 + \beta_1 \cdot x$  sopra considerato. L'equivalenza è del tutto ovvia, infatti, se si pone  $z = x^3$ , si ottiene il modello di regressione lineare semplice  $y = \beta_0 + \beta_1 \cdot z$ .

Le considerazioni svolte valgono quindi per tutti i modelli lineari nei parametri incogniti che li caratterizzano. E', tuttavia, possibile in molti casi di interesse applicare le stesse procedure a modelli non lineari nei parametri, è ciò accade tutte le volte in cui risulta possibile ricondursi alla situazione di linearità operando opportune trasformazioni del modello non lineare. Ovviamente, quando si operano delle trasformazioni sia le ipotesi di specificazioni sia le conclusioni cui si perviene vanno riferite al modello trasformato e non al modello originario. Alcuni esempi significativi sono quelli sotto riportati.

i)  $y = \beta_0 \cdot x^{\beta_1} \cdot e^u$  ,

la cui trasformazione logaritmica fornisce (**modello doppio logaritmico**)

$$\log \cdot y = \log \cdot \beta_0 + \beta_1 \cdot \log \cdot x + u$$

ii)  $y = e^{\beta_0 + \beta_1 \cdot x} \cdot e^u$

$$y = \beta_0 \cdot e^{\beta_1 \cdot x} \cdot e^u$$

$$e^y = \beta_0 \cdot x^{\beta_1} \cdot e^u$$

le cui trasformate logaritmiche danno (**modello semilogaritmico**) rispettivamente

$$\log \cdot y = \beta_0 + \beta_1 \cdot x + u$$

$$\log \cdot y = \log \cdot \beta_0 + \beta_1 \cdot x + u$$

$$y = \log \cdot \beta_0 + \beta_1 \cdot \log \cdot x + u.$$

## 6. 6 Coefficiente di correlazione lineare e analisi della varianza

Il coefficiente di correlazione lineare  $\rho_{yx} = \rho_{xy} = \rho$  è stato introdotto come indice relativo di concordanza (rapporto tra l'indice assoluto di concordanza covarianza  $\sigma_{yx} = \sigma_{xy}$  ed il valore massimo che  $|\sigma_{yx}|$  può assumere e che è dato dal prodotto tra gli scostamenti quadratici medi  $\sigma_y \cdot \sigma_x$ ), cioè

$$\rho = \frac{\sigma_{yx}}{\sigma_y \sigma_x} = \frac{\text{Codev}(y, x)}{\sqrt{\text{Dev}(y) \cdot \text{Dev}(x)}}$$

Tale coefficiente può essere visto anche come media geometrica dei due coefficienti di regressione  $b_{y/x} = \frac{\sigma_{yx}}{\sigma_x^2} = \frac{\text{Codev}(y, x)}{\text{Dev}(x)}$  e  $b_{x/y} = \frac{\sigma_{yx}}{\sigma_y^2} = \frac{\text{Codev}(y, x)}{\text{Dev}(y)}$ .

Infatti, in riferimento al modello  $y_i = \beta_0 + \beta_1 \cdot x_i + u_i$  la stima dei minimi quadrati (e

della massima verosimiglianza) di  $\beta_1$  sia pari a  $\hat{\beta}_1 = \frac{\text{Codev}(y, x)}{\text{Dev}(x)} = \frac{\sigma_{yx}}{\sigma_x^2} = b_{y/x}$ , se si

ipotizza un modello lineare del tipo  $x_i = \alpha_0 + \alpha_1 \cdot y_i + v_i$  e si introducono le usuali ipotesi di specificazione, la stima dei minimi quadrati (e della massima verosimiglianza)

di  $\alpha_1$  è pari a  $\hat{\alpha}_1 = \frac{\text{Codev}(y, x)}{\text{Dev}(y)} = \frac{\sigma_{yx}}{\sigma_y^2} = b_{x/y}$  dal che risulta quanto affermato:

$$\rho = \sqrt{\hat{\beta}_1 \cdot \hat{\alpha}_1} = \sqrt{b_{y/x} \cdot b_{x/y}} = \frac{\sigma_{yx}}{\sigma_y \sigma_x} = \frac{\text{Codev}(y, x)}{\sqrt{\text{Dev}(y) \cdot \text{Dev}(x)}}$$

Una terza, forse la più interessante, interpretazione del coefficiente di correlazione lineare di Bravais-Pearson deriva dalle osservazioni che seguono.

Dato il modello

$$y_i = \beta_0 + \beta_1 \cdot x_i + u_i \quad \text{per } i = 1, 2, \dots, n$$

che soddisfa alle ipotesi di specificazione introdotte, la **devianza totale** della variabile  $y$  è data da

$$\begin{aligned} \text{Dev}(T) = \text{Dev}(y) &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i^* + \hat{y}_i^* - \bar{y})^2 = \\ &= \sum_{i=1}^n (y_i - \hat{y}_i^*)^2 + \sum_{i=1}^n (\hat{y}_i^* - \bar{y})^2 = \text{Dev}(r) + \text{Dev}(R) \end{aligned}$$

dove  $\text{Dev}(r) = \sum_{i=1}^n (y_i - \hat{y}_i^*)^2$  viene detta **devianza residua** e misura la parte della devianza totale della variabile  $y$  che non risulta spiegata dalla supposta relazione con la variabile  $x$ ;  $\text{Dev}(R) = \sum_{i=1}^n (\hat{y}_i^* - \bar{y})^2$  viene detta **devianza di regressione** e misura quanta parte della devianza di  $y$  è spiegata dalla relazione lineare con la variabile  $x$ .

Se in corrispondenza di ciascuna modalità  $x_i$  ( $i = 1, 2, \dots, s$ ) della variabile  $x$ , si disponesse di più osservazioni  $y_{ij}$  ( $j = 1, 2, \dots, n_i$ ), si potrebbe procedere alla seguente scomposizione della devianza totale della variabile  $y$

$$\begin{aligned} \text{Dev}(T) = \text{Dev}(y) &= \sum_{i=1}^s \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^s \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i^* + \hat{y}_i^* - \bar{y}_i + \bar{y}_i - \bar{y})^2 = \\ &= \sum_{i=1}^s \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i^*)^2 + \sum_{i=1}^s \sum_{j=1}^{n_i} (\hat{y}_i^* - \bar{y}_i)^2 + \sum_{i=1}^s \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 = \\ &= \sum_{i=1}^s \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^s \sum_{j=1}^{n_i} (\bar{y}_i - \hat{y}_i^*)^2 + \sum_{i=1}^s \sum_{j=1}^{n_i} (\hat{y}_i^* - \bar{y})^2 \end{aligned}$$

dove:  $\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$  e le tre diverse devianze ottenute dalla scomposizione (si ricorda che i doppi prodotti sono tutti nulli) sono di facile interpretazione: in un caso come primo elemento di riferimento si considerano i valori che si trovano sulla **retta di**

**regressione** (cfr. Fig. 10), nel secondo caso il primo elemento di riferimento sono i valori (medie di gruppo) che si trovano sulla **spezzata di regressione**.

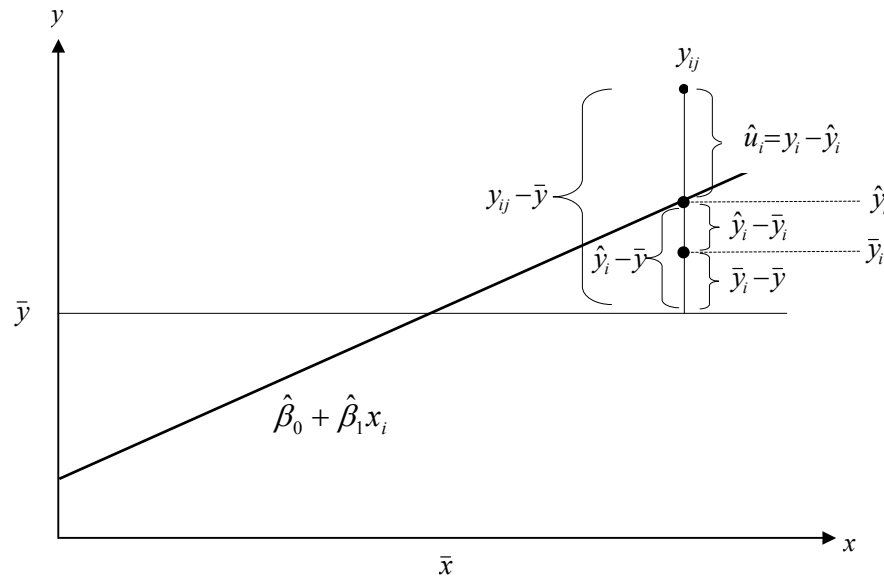


Fig. 10 – Scomposizione della devianza totale della variabile  $y$

Tornando alla scissione della devianza totale della variabile  $y$  nelle due componenti: devianza di regressione e devianza residua, si può introdurre l'indice, usualmente detto **di determinazione**

$$R^2 = \frac{Dev(R)}{Dev(T)} = 1 - \frac{Dev(r)}{Dev(T)}$$

che, ovviamente, assume valori compresi nell'intervallo  $0 \text{ --- } 1$ : assume valore  $0$  quando tutti i valori  $\hat{y}_i = \hat{y}_i^*$  che si trovano sulla retta di regressione sono uguali tra loro e, quindi, uguali a  $\bar{y}$  (media della variabile  $y$ ); assume valore  $1$  quando tutti gli scarti  $(y_i - \hat{y}_i)$  sono uguali a zero, cioè, quando tutti i punti osservati si trovano sulla retta di regressione (adattamento totale del modello).

Tenendo presente che

$$\begin{aligned} Dev(R) &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (\beta_0 + \beta_1 \cdot x_i - \bar{y})^2 = \sum_{i=1}^n (\bar{y} - \beta_1 \cdot \bar{x} + \beta_1 \cdot x_i - \bar{y})^2 = \\ &= \hat{\beta}_1^2 \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{\sigma_{yx}^2}{\sigma_x^4} \cdot n \cdot \sigma_x^2 = n \frac{\sigma_{yx}^2}{\sigma_x^2} \end{aligned}$$

si avrà

$$R^2 = \frac{Dev(R)}{Dev(T)} = \frac{n \cdot \sigma_{yx}^2 / \sigma_x^2}{n \cdot \sigma_y^2} = \frac{\sigma_{yx}^2}{\sigma_x^2 \cdot \sigma_y^2} = \rho^2$$

cioè: l'indice di determinazione è uguale al quadrato del coefficiente di correlazione lineare, il che consente d'interpretare tale quadrato come misura della proporzione della variabilità totale della variabile  $y$  che risulta spiegata dalla supposta relazione lineare con la variabile  $x$ .

Se si vuole sottoporre a test l'ipotesi di un effetto "significativo" della variabile  $x$  sulla variabile  $y$ , si può procedere come sopra indicato, cioè formulando l'ipotesi:

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_1 : \beta_1 &\neq 0 \end{aligned}$$

od anche facendo ricorso ad un test di bontà di adattamento del modello.

Si è già osservato che

$$W = \frac{(n-2) \cdot \hat{\sigma}^2}{\sigma^2} \sim \chi_{n-2}^2$$

se si considera ora la variabile

$$V = \frac{Dev(R)}{\sigma^2} = \frac{\sum_{i=1}^n (\hat{y}_i^* - \bar{y})^2}{\sigma^2}$$

che ha legge di distribuzione  $\chi^2$  con un grado di libertà ed è indipendente dalla variabile  $W$ ; che si ricorda ha legge di distribuzione  $\chi^2$  con  $(n-2)$  gradi di libertà, la

variabile (rapporto tra due variabili  $\chi^2$  indipendenti divise per i rispettivi gradi di libertà)

$$F = \frac{W}{V/(n-2)} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sigma^2} \cdot \frac{(n-2) \cdot \hat{\sigma}^2}{\sigma^2} / (n-2)$$

ha, sotto l'ipotesi  $H_0 : \beta_1 = 0$  (quando l'ipotesi è vera) legge di distribuzione del tipo  $F$  di Fisher-Snedecor con 1 e  $(n-2)$  gradi di libertà.

Da rilevare che sotto l'ipotesi  $H_0 : \beta_1 = 0$  contro l'ipotesi  $H_1 : \beta_1 \neq 0$  vale l'uguaglianza  $T_{n-2}^2 = F_{1,n-2}$ , il che porta a concludere che nel caso di regressione lineare semplice la procedura per sottoporre a test l'ipotesi di adattamento del modello e l'ipotesi (bidirezionale) sul coefficiente angolare della retta di regressione sono del tutto equivalenti. In proposito vale la pena, infine, segnalare che tale procedura equivale anche a quella relativa al test diretto sul coefficiente di correlazione  $\rho$ ; infatti, sotto l'ipotesi  $H_0 : \rho = 0$  contro l'ipotesi alternativa  $H_1 : \rho \neq 0$ , la variabile casuale test di riferimento è

$$T_{\rho} = \frac{\hat{\rho} \cdot \sqrt{(n-2)}}{\sqrt{1-\hat{\rho}^2}} = \hat{\beta}_1 \cdot \sqrt{\frac{Dev(x)}{Dev(y)}} \cdot \sqrt{1 - \frac{Dev(R)}{Dev(T)} / (n-2)} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\beta_1}} = T_{\hat{\beta}_1}$$