

Regressione Lineare Semplice e Correlazione

Introduzione

- ✓ La Regressione è una tecnica di analisi della relazione tra due variabili quantitative
- ✓ Questa tecnica è utilizzata per calcolare il valore (y) di una variabile dipendente, in funzione del valore di un'altra variabile indipendente (x_1, x_2, \dots, x_k .)
- ✓ La funzione di regressione che viene individuata esprime la relazione di dipendenza in media della variabile Y dalla variabile X

Il modello

- Il modello lineare

$$y = \beta_0 + \beta_1 x + \varepsilon$$

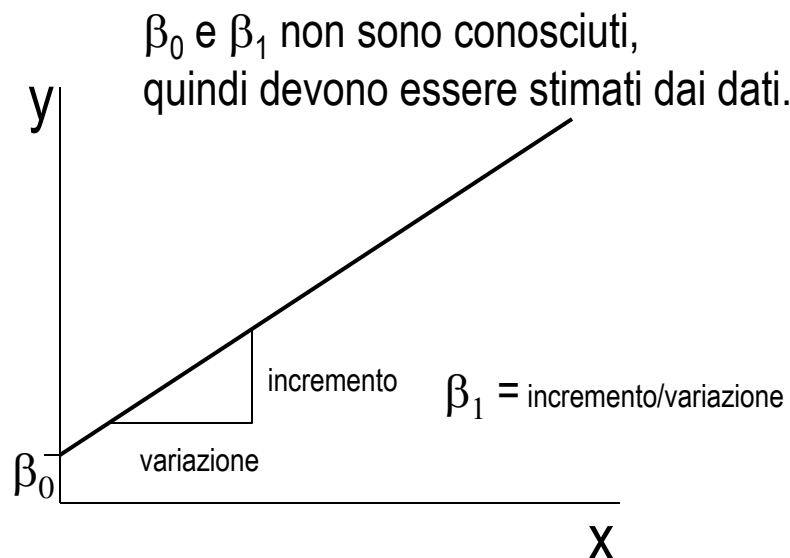
y = variabile dipendente

x = variabile indipendente

β_0 = y-intercetta

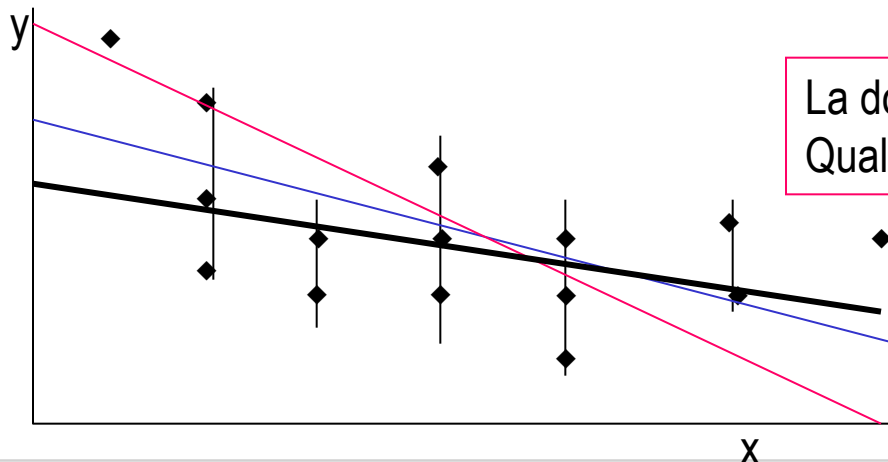
β_1 = coefficiente angolare

ε = variabile errore



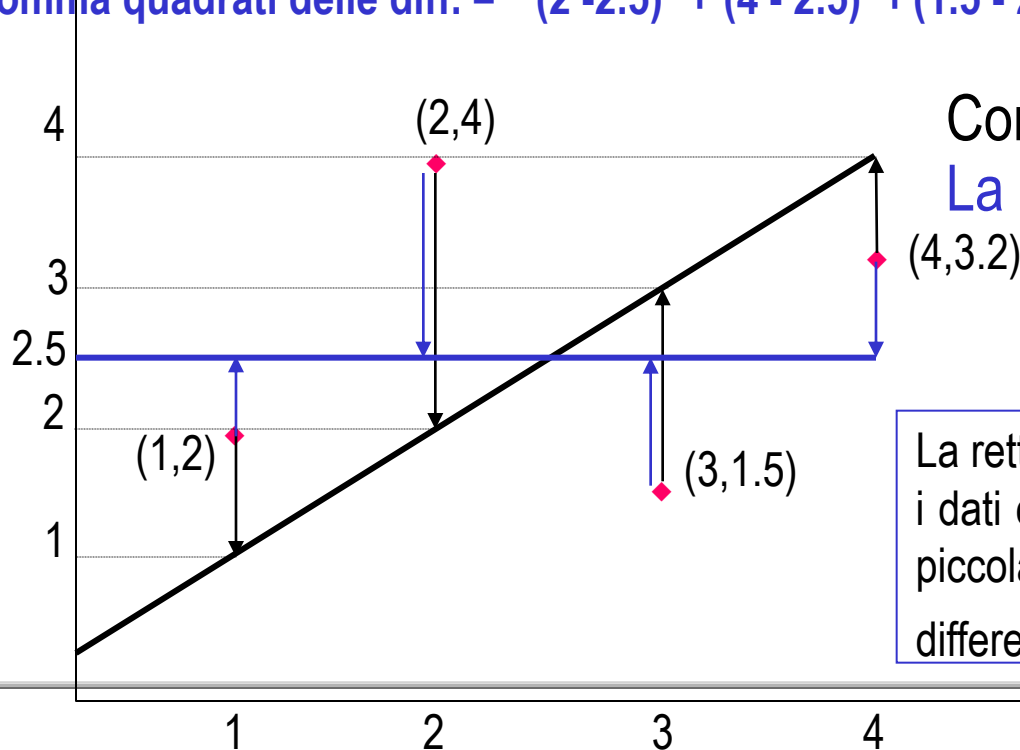
Stima dei Coefficienti

- Le stime sono determinate da:
 - Estrazione del campione dalla popolazione di riferimento
 - Calcolo delle statistiche semplici
 - Ricerca della migliore retta di interpolazione dei dati



La retta di regressione è quella che minimizza
la somma dei quadrati delle differenze tra le osservazioni e la retta

$$\begin{aligned}\text{Somma quadrati delle diff.} &= (2 - 1)^2 + (4 - 2)^2 + (1.5 - 3)^2 + (3.2 - 4)^2 = 6.89 \\ \text{Somma quadrati delle diff.} &= (2 - 2.5)^2 + (4 - 2.5)^2 + (1.5 - 2.5)^2 + (3.2 - 2.5)^2 = 3.99\end{aligned}$$



Confrontiamo due rette
La seconda è orizzontale

La retta che interpola meglio
i dati è quella a cui corrisponde la più
piccola somma dei quadrati delle
differenze



$$\sum_h \sum_i [y_{ih} - (b_0 + b_1 x_h)]^2 = \min$$

Derivando rispetto a b_0 e b_1 e ponendo le derivate parziali uguali a zero, otteniamo la formula

$$b_1 = \frac{\text{cov}(X, Y)}{s_x^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

L'equazione di regressione che stima l'equazione del modello lineare è

$$\hat{y} = b_0 + b_1 x$$

• Esempio

Relazione tra i Km effettuati e il prezzo di un'auto usata

- Un venditore di auto usate vuole capire la relazione tra i Km effettuati e il prezzo della macchina usata
- Un campione casuale di 100 auto è selezionato e i dati Trovare la retta di regressione.

Auto	Km.	Prezzo
1	37388	5318
2	44758	5061
3	45833	5008
4	30862	5795
5	31705	5784
6	34010	5359
.	.	.
.	.	.
.	.	.

Variabile indipendente x

Variabile dipendente y



Esempio 7.1

• Soluzione

- Per calcolare b_0 and b_1 abbiamo bisogno di calcolare:

$$\bar{x} = 36,009.45; \quad s_x^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = 43,528,688$$

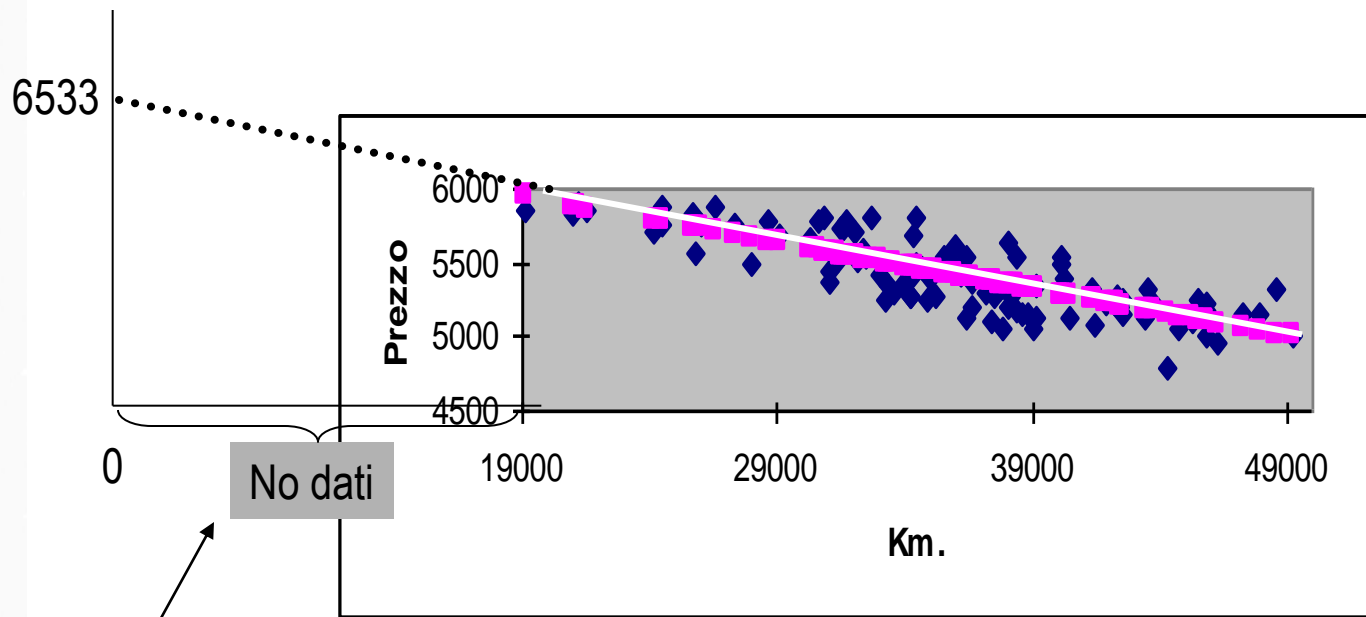
$$\bar{y} = 5,411.41; \quad \text{cov}(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = -1,356,256$$

dove $n = 100$.

$$b_1 = \frac{\text{cov}(X, Y)}{s_x^2} = \frac{-1,356,256}{43,528,688} = -.0312$$

$$b_0 = \bar{y} - b_1 \bar{x} = 5411.41 - (-.0312)(36,009.45) = 6,533$$

$$\hat{y} = b_0 + b_1 x = 6,533 - .0312x$$



$$\hat{y} = 6,533 - .0312x$$



L'intercetta è $b_0 = 6533$.

Questo è il coefficiente angolare.
Per ogni chilometro addizionale, il prezzo decresce
in media di € 0.0312

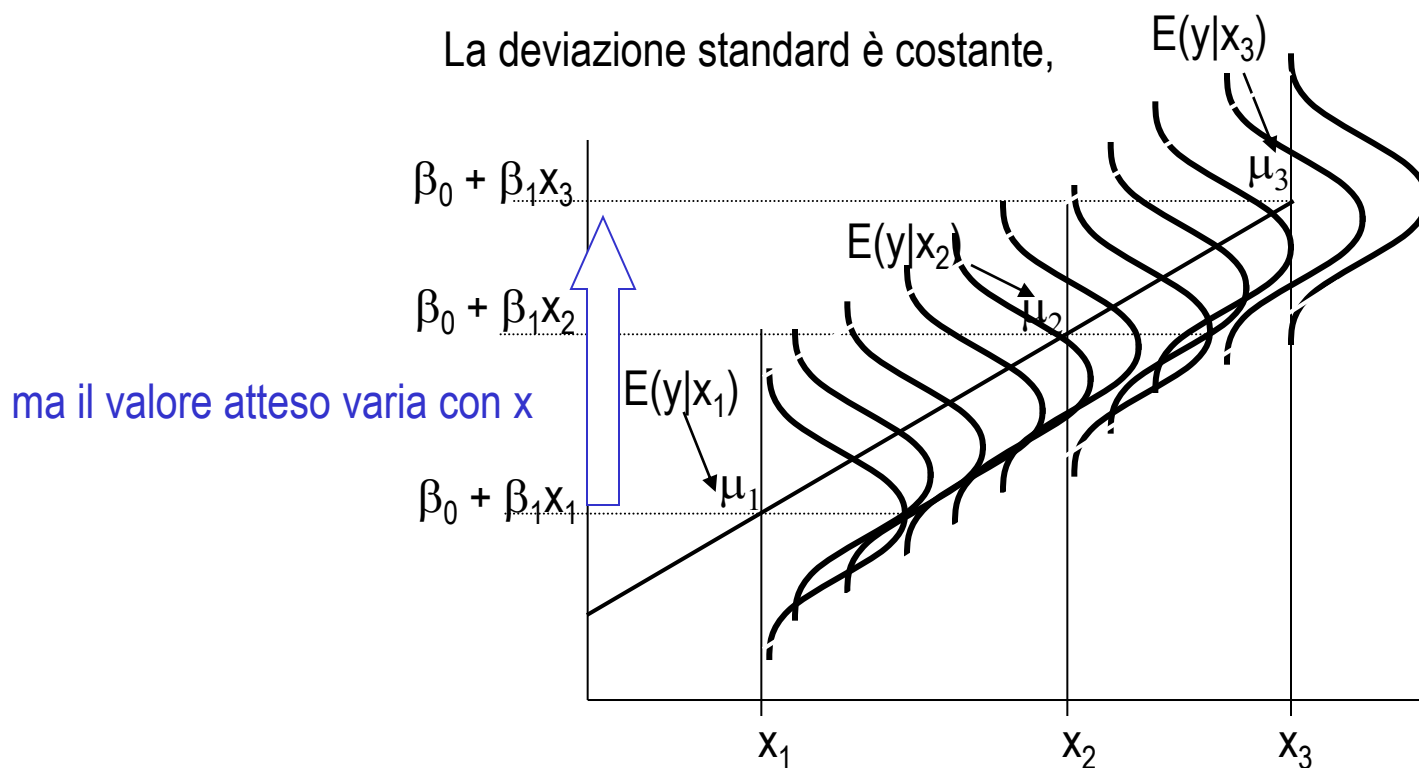
L'intercetta può essere interpretata come:
"Il prezzo delle auto che non sono mai state guidate"

La variabile Errore

Le ipotesi alla base del modello

- L'errore ε è una parte critica del modello di regressione
- Devono essere soddisfatte quattro ipotesi *forti* sulla variabile casuale ε :
 - ε si distribuisce in modo normale
 - Il valore atteso di ε è zero ovvero $E(\varepsilon_i) = 0$
 - La deviazione standard di ε è σ_ε per tutti i valori di x ovvero $E(\varepsilon_i^2) = \sigma_\varepsilon^2$
 - I set di errori associati a differenti valori di y sono tutti tra loro indipendenti ovvero $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$

Per le prime tre ipotesi alla base del modello:
y si distribuisce in modo normale con valore atteso $E(y) = \beta_0 + \beta_1 x$, e deviazione standard σ_ε



Valutazione del modello

- ✓ Il metodo dei minimi quadrati produce una regressione lineare anche quando non ci sia una relazione lineare tra x ed y .
- ✓ E' importante, perciò, valutare la bontà di adattamento del modello lineare
- ✓ Numerosi metodi sono utilizzati per fare ciò:
 - Test dei coefficienti
 - Indici sintetici

- **Somma dei quadrati degli errori**

- La somma dei quadrati degli scarti tra i punti e la retta di regressione è una misura di come la retta approssimi bene la nube dei punti.

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



$$SSE = (n-1)s_Y^2 - \frac{\text{cov}(X, Y)^2}{s_X^2}$$

• L'errore standard delle stime

- Il valore atteso di ε è uguale a 0
- Se σ_ε è piccolo, gli errori tendono a concentrarsi attorno alla media (=0). Dunque il modello approssima bene i dati
- Così, possiamo usare σ_ε come una misura di adattabilità del modello lineare
- Uno stimatore non distorto di σ_ε^2 è dato da s_ε^2

Errore Standard delle Stime

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$



- Esempio .
 - Calcolare l'errore standard delle stime

Soluzione

$$s_Y^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-1} = \frac{6,434,890}{99} = 64,999$$

$$SSE = (n-1)s_Y^2 - \frac{\text{cov}(X, Y)^2}{s_x^2} = 99(64,999) - \frac{(-1,356,256)^2}{43,528,688} = 2,252,363$$

Calcolati prima

Dunque,

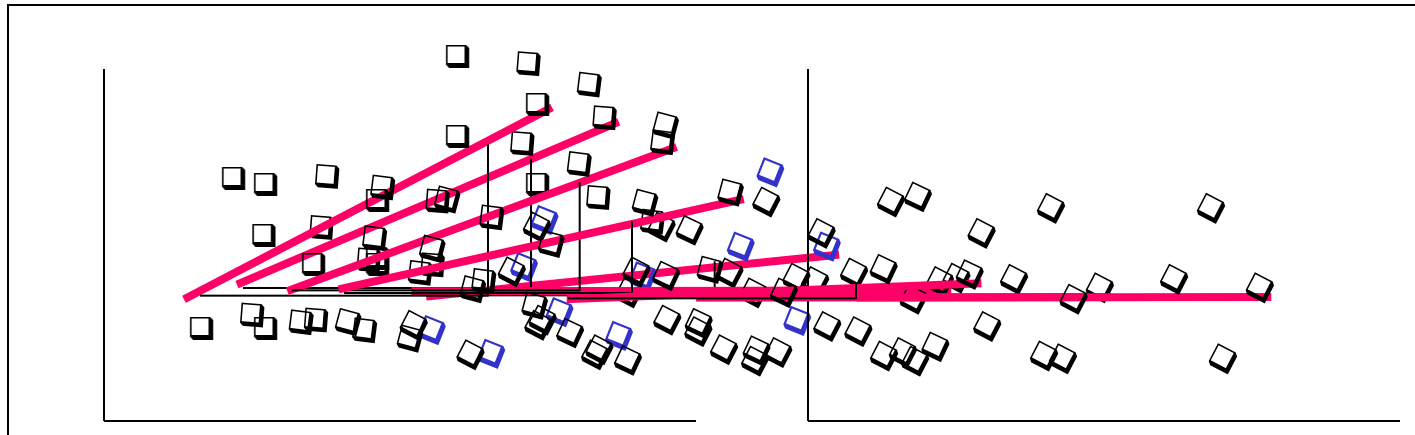
$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{2,251,363}{98}} = 151.6$$

Il modello approssima bene i dati, soprattutto se confrontiamo s_ε con il valore medio di y .

$$s_\varepsilon = 151.6, \bar{y} = 5,411.4$$

• Test della pendenza della retta

- Quando non esiste una relazione lineare tra le due variabili la retta di regressione è orizzontale



Relazione lineare

La pendenza non è uguale a zero

Relazione non lineare

La pendenza è uguale a zero

- Possiamo fare inferenza su β_1 partendo da b_1 , facendo il seguente test di ipotesi:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0 (< 0; > 0)$$

– La *statistica test* è

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

dove

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}}$$

↑
Errore standard di b_1 .

- Se la variabile casuale *errore* si distribuisce in modo normale la statistica è una *t* di Student con $n-2$ g.d.



Esempio 7.1

- Soluzione dell'esempio

- Per calcolare “t” abbiamo bisogno dei valori di b_1 e di s_{b_1}

$$b_1 = -.312$$

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}} = \frac{151.6}{\sqrt{(99)(43,528,688)}} = .00231$$

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{-.312 - 0}{.00231} = -13.49$$

P-value= 4.4 4E-24

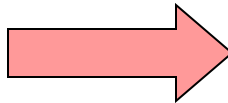
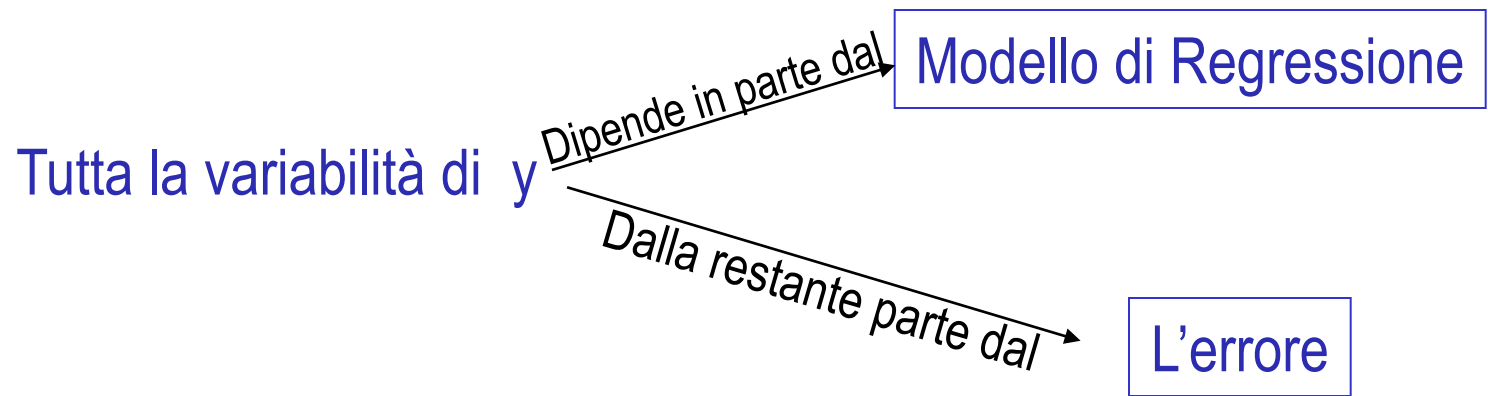
C'è una schiacciante evidenza della dipendenza lineare del prezzo dell'auto usata, dal numero di Km effettuati

- **Coefficiente di determinazione**

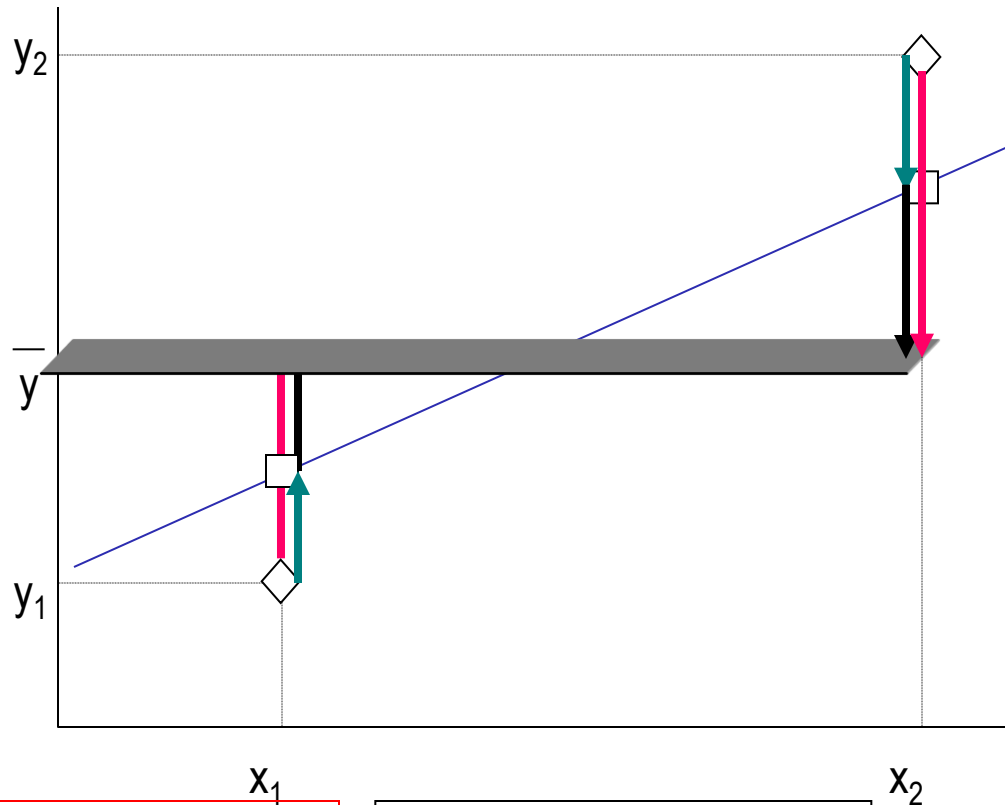
- Quando vogliamo misurare la *forza* della relazione lineare, usiamo l'indice di Determinazione lineare R^2

$$R^2 = \frac{[\text{cov}(X, Y)]^2}{s_x^2 s_y^2} \quad o \quad R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2}$$

- Per capire tale coefficiente bisogna ricordare che :



Consideriamo due punti (x_1, y_1) e (x_2, y_2) di un campione



Variazione Totale in y =


Variazione espressa dalla
retta di regressione

+ Variazione dell'errore

$$(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 = (\hat{y}_1 - \bar{y})^2 + (\hat{y}_2 - \bar{y})^2 + (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2$$

Variazione in $y = SSR + SSE$

- R^2 misura la proporzione di variabilità di y espressa dalla variabilità di x

$$R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2} = \frac{\sum (y_i - \bar{y})^2 - SSE}{\sum (y_i - \bar{y})^2} = \frac{SSR}{\sum (y_i - \bar{y})^2}$$


- R^2 varia tra 0 e 1
 - Quando è uguale ad 1 ($R^2 = 1$), i punti giacciono sulla retta di regressione
 - Quando è uguale ad 0 ($R^2 = 0$), non c'è relazione lineare tra x e y



Esempio 7.1

- Esempio .
 - Trovare il coefficiente di determinazione

•Soluzione

$$R^2 = \frac{[\text{cov}(X, Y)]^2}{s_x^2 s_y^2} = \frac{[-1,356,256]^2}{(43,528,688)(64,999)} = .6501$$

Il 65% della varianza del prezzo è spiegata dalla variazione dei Km segnati dal tachimetro. Il restante 35% non viene spiegato dal modello



Esempio 7.1

Uso del modello di Regressione lineare

- Se siamo soddisfatti della bontà di adattamento della retta di regressione, possiamo utilizzare l'equazione stimata per *predire* valori di y
 - Esempio
 - Prevedere il prezzo una una macchina con 40,000 Km

$$\hat{y} = 6533 - .0312x = 6533 - .0312(40,000) = 5,285$$

Esempio 7.6

• Intervallo di confidenza

– Due sono gli intervalli importanti per le previsioni di y .

- Intervallo di previsione – per un valore particolare di y
- Intervallo di confidenza – per il valore atteso di y

– Intervallo di previsione

$$\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

– Intervallo di confidenza

$$\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{\frac{1}{n} + \frac{(x_g - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

L'intervallo di previsione è più ampio dell'intervallo di confidenza

• Esempio

- Calcolare un intervallo di previsione per una macchina con 40,000 Km
- Soluzione

- L'intervallo di previsione al 95% =

$$\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$[6533 - .0312(40000)] \pm 1.984(151.6) \sqrt{1 + \frac{1}{100} + \frac{(40,000 - 36,009)^2}{\sum 4,309,340,160}} = 5,285 \pm 303$$

Diagram illustrating the components of the prediction interval formula:

- \hat{y} is the predicted mean response.
- $t_{\alpha/2}$ is the critical value from the t-distribution.
- s_{ε} is the standard error of the estimate.
- The term under the square root represents the variance of the prediction.

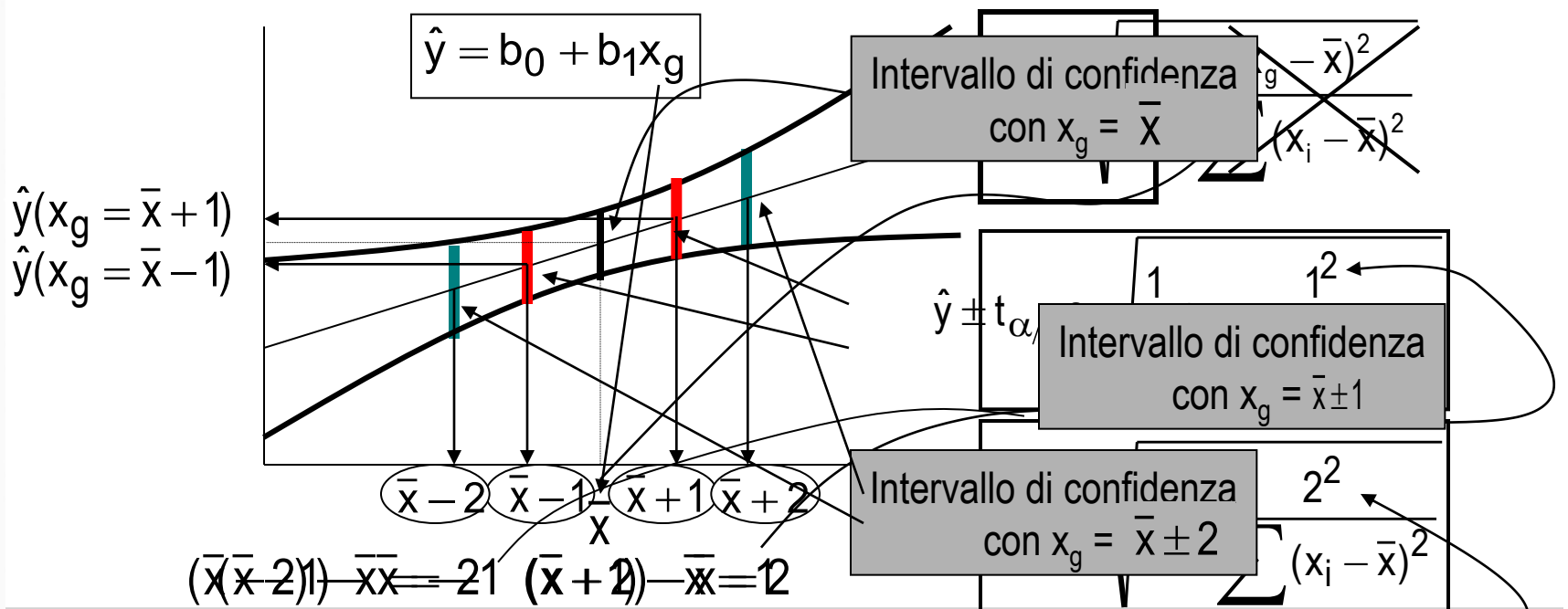
- Il venditore di auto vuole prendere un lotto di 40,000 KM. Calcolare l'intervallo di confidenza per y al 95%
- Soluzione

$$\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{\frac{1}{n} + \frac{(x_g - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$[6533 - .0312(40000)] \pm 1.984(151.6) \sqrt{\frac{1}{100} + \frac{(40,000 - 36,009)^2}{4,309,340,160}} = 5,285 \pm 35$$

- L'effetto di un valore dato di x nell'intervallo

- Appena x_g si allontana da \bar{x} l'intervallo diventa più grande. Il più piccolo intervallo è trovato per \bar{x} .



Coefficiente di correlazione

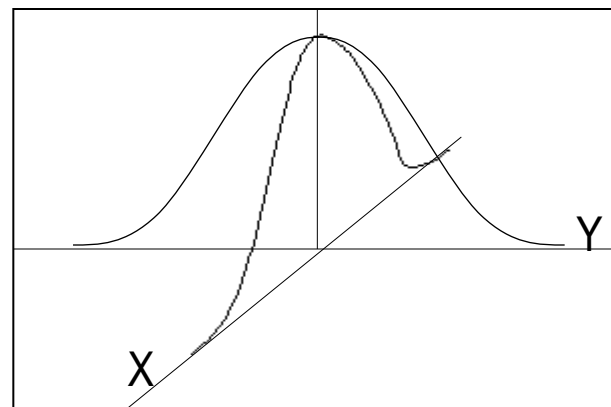
- Il coefficiente di correlazione è utilizzato per misurare il legame tra due variabili.
- Assume un valore tra -1 e 1
 - Se $r = -1$ (associazione negativa) o $r = +1$ (associazione positiva) ogni punto giace sulla retta di regressione.
 - Se $r = 0$ non c'è legame lineare.
- Il coefficiente di correlazione può essere utilizzato per testare una relazione lineare tra due variabili.

- Test del coefficiente di correlazione
 - Quando non c'è relazione lineare $\rho = 0$.
 - Le ipotesi sono:
 - $H_0: \rho = 0$
 - $H_1: \rho \neq 0$
 - La statistica test è:

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

dove r è il coefficiente di correlazione nel campione

calcolato da $r = \frac{\text{cov}(X, Y)}{s_x s_y}$



La statistica è una t di Student con $n - 2$ g.d.l.



Esempio 7.1

• Esempio Test di relazione lineare

- Effettuare un test sul coefficiente di correlazione dell'esempio 7.1 per vedere se c'è relazione lineare

• Soluzione

- $H_0: \rho = 0$
 $H_1: \rho \neq 0$
- La zona di rifiuto è
 $|t| > t_{\alpha/2, n-2} = t_{.025, 98} = 1.984$
- Nel campione il coefficiente di correlazione è
 $r = \text{cov}(X, Y) / s_x s_y = -.806$

Il valore della statistica t è

$$t = r \sqrt{\frac{n-2}{1-r^2}} = -13.49$$

Conclusione:

C'è sufficiente evidenza ad un livello di significatività $\alpha = 5\%$ per dire che c'è un legame lineare tra le due variabili.

La Diagnostica di Regressione

- Prima di utilizzare un modello di regressione per fare inferenza, bisogna verificare
 - che le ipotesi alla base del modello siano rispettate
 - che non ci siano dati anomali che possano inficiare i risultati
 - Come vedere se le ipotesi forti:
 - ε si distribuisce in modo normale
 - La varianza di ε è costante per tutti i valori di x :
 $E(\varepsilon_i^2) = \sigma_\varepsilon^2$
 - Gli errori sono tra loro indipendenti:
 $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$
- sono rispettate?



Esempio 7.1

- Analisi dei residui

Analizzando i residui (o i residui standardizzati), si può vedere se ci sono violazioni alle ipotesi poste alla base del modello

- Non normalità

Esempio

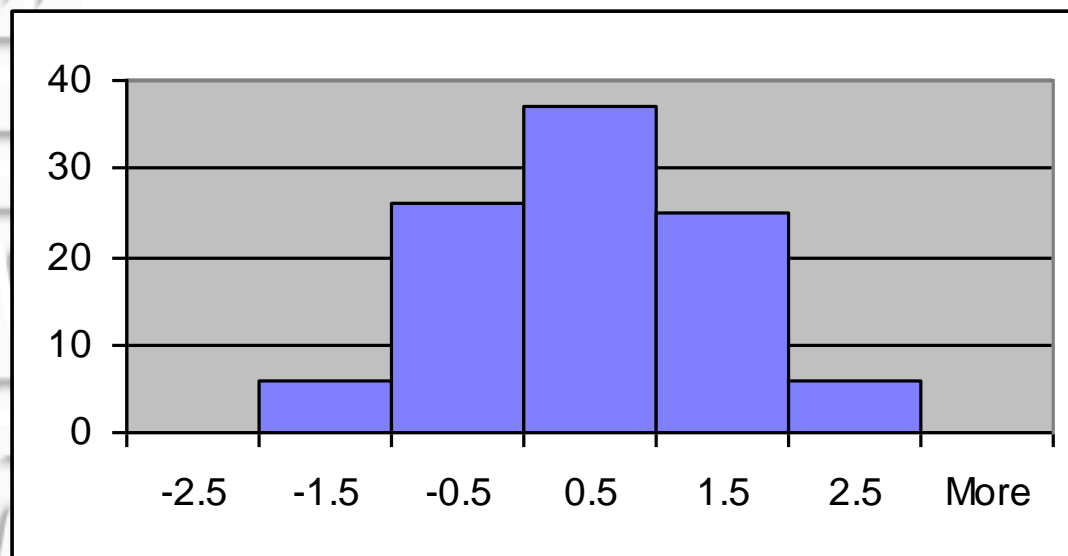
- Sui dati dell'Esempio costruiamo gli istogrammi dei residui standardizzati
- Esaminiamo gli istogrammi e guardiamo alla forma della distribuzione centrata attorno allo zero

RESIDUI OUTPUT			Lista parziale
Osservazioni	Residui	Residui Standardizzati	
1	-50,45749927	-0,334595895	
2	-77,82496482	-0,516076186	
3	-97,33039568	-0,645421421	
4	223,2070978	1,480140312	
5	238,4730715	1,58137268	

Per ogni residuo calcoliamo:

$$s_{r_i} = s_{\varepsilon} \sqrt{1 - h_i} \quad \text{dove}$$

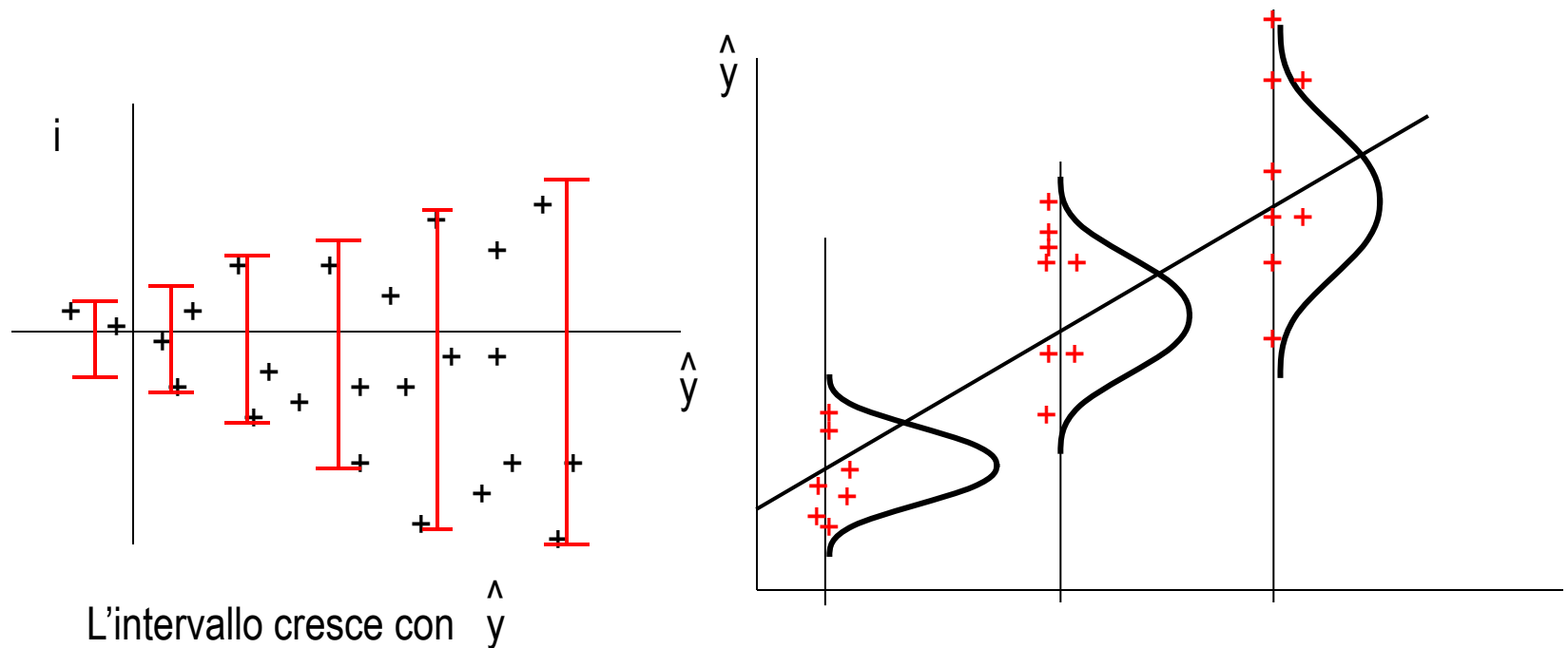
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}$$



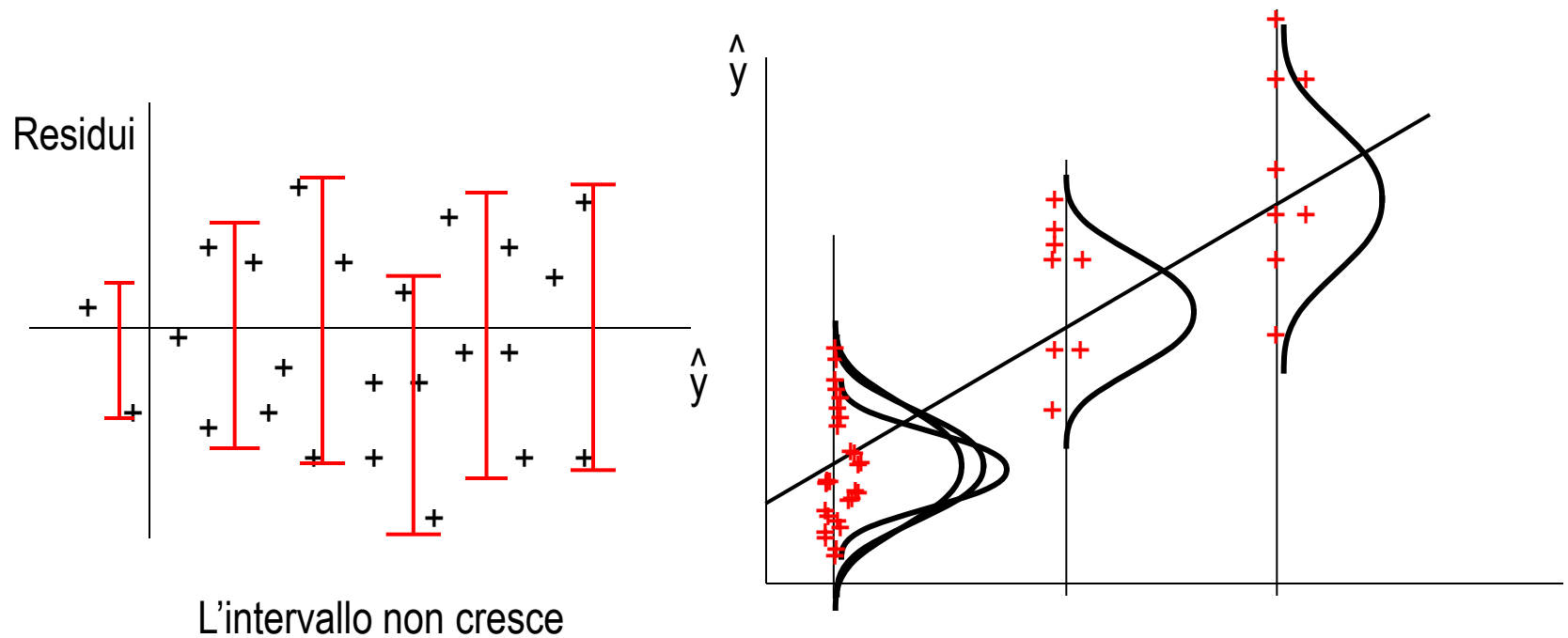
Possiamo inoltre fare il test χ^2 di normalità

• Eteroschedasticità

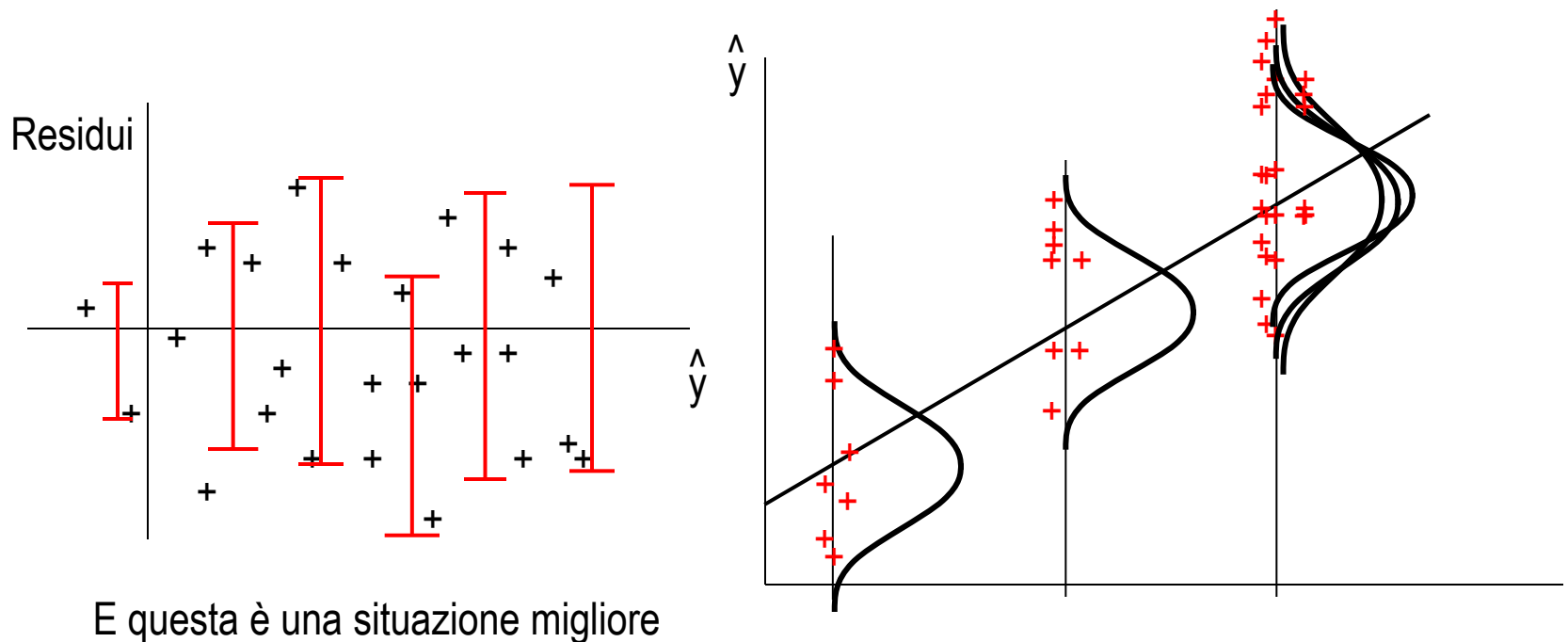
- Quando la varianza di ε non è costante per tutti i valori di x , allora si ha **eteroschedasticità**



- Quando la varianza di ε è costante per tutti i valori di x , allora c'è **omoschedasticità**

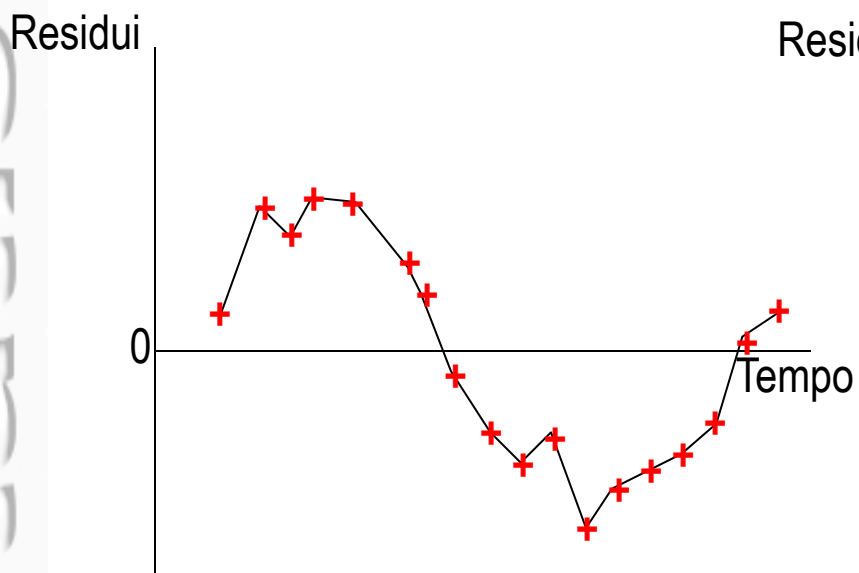


- Quando la varianza di ε è costante per tutti i valori di x , allora c'è **omoschedasticità**

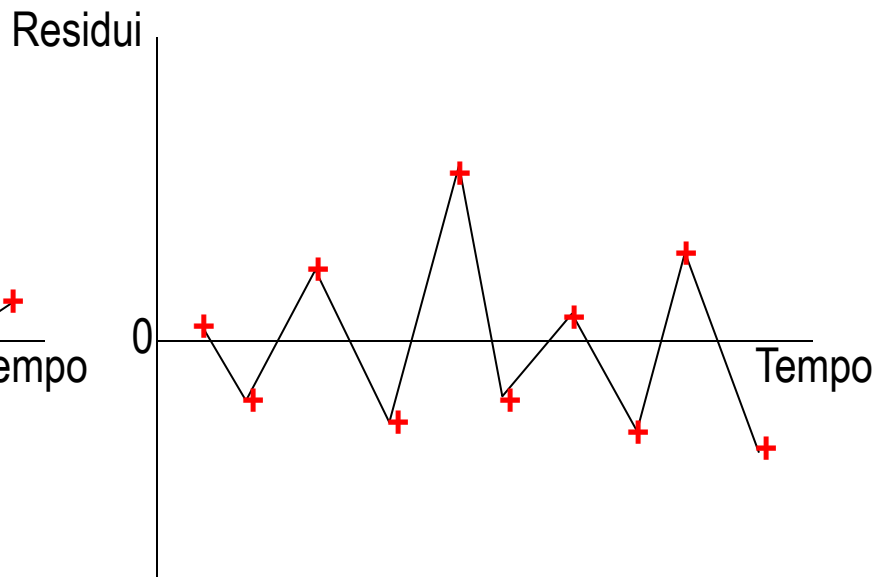


- Non indipendenza delle variabili errore
 - Quando le variabili errore non sono indipendenti si parla di autocorrelazione dei residui (soprattutto per le Serie Storiche)

Esempi di autocorrelazione dei residui



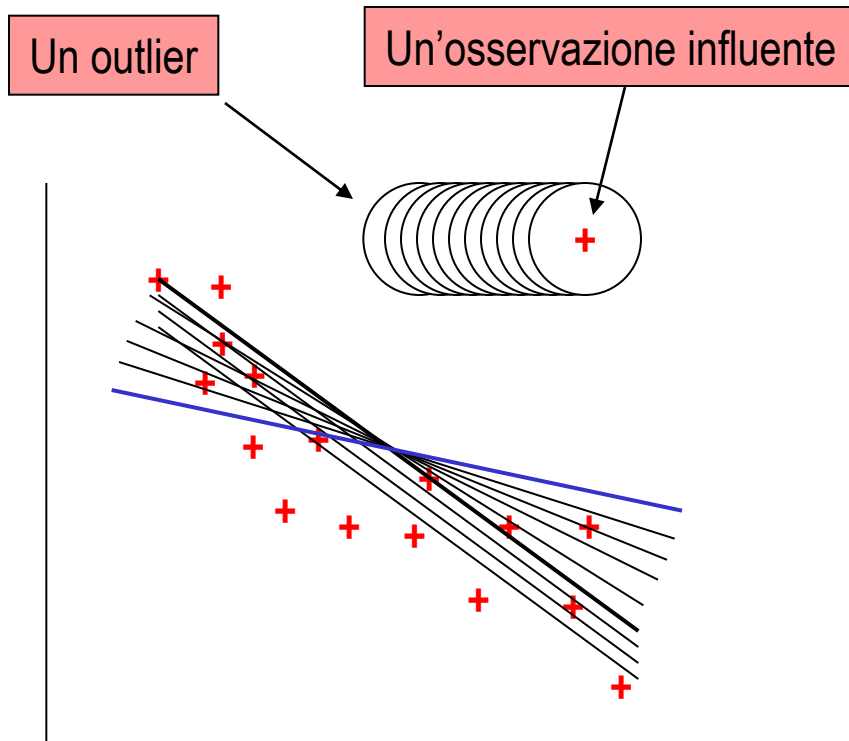
Andamento positivo dei residui
alternato con un andamento negativo



Oscillazione dei residui attorno
allo zero

•Outliers

Un *outlier* è un valore o troppo piccolo o troppo grande, che può influenzare la retta di regressione e per questo deve essere identificato con un scatter-plot



... ma, può influenzare ancora di più l'analisi!!

Gli outliers portano uno spostamento della retta di regressione