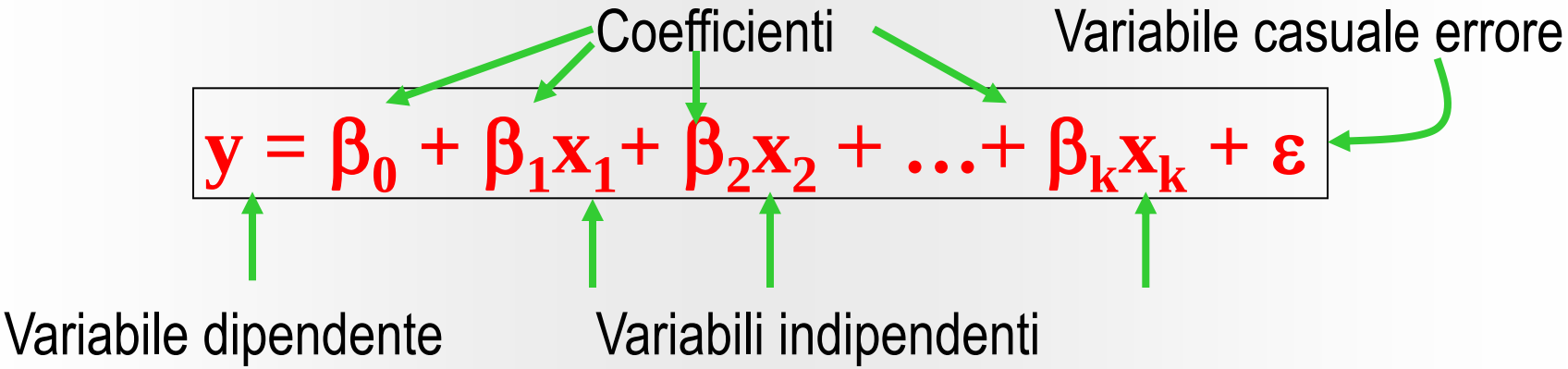
Regressione Lineare Multipla

Introduzione

- La Regressione lineare multipla rappresenta una estensione del modello di regressione semplice
- Questa tecnica è utilizzata per studiare le variazioni di una variabile dipendente, in funzione di più variabili indipendenti
- L'obiettivo è costruire un modello che approssimi i dati meglio del modello di regressione lineare semplice.

8.2 Il Modello e le ipotesi di base

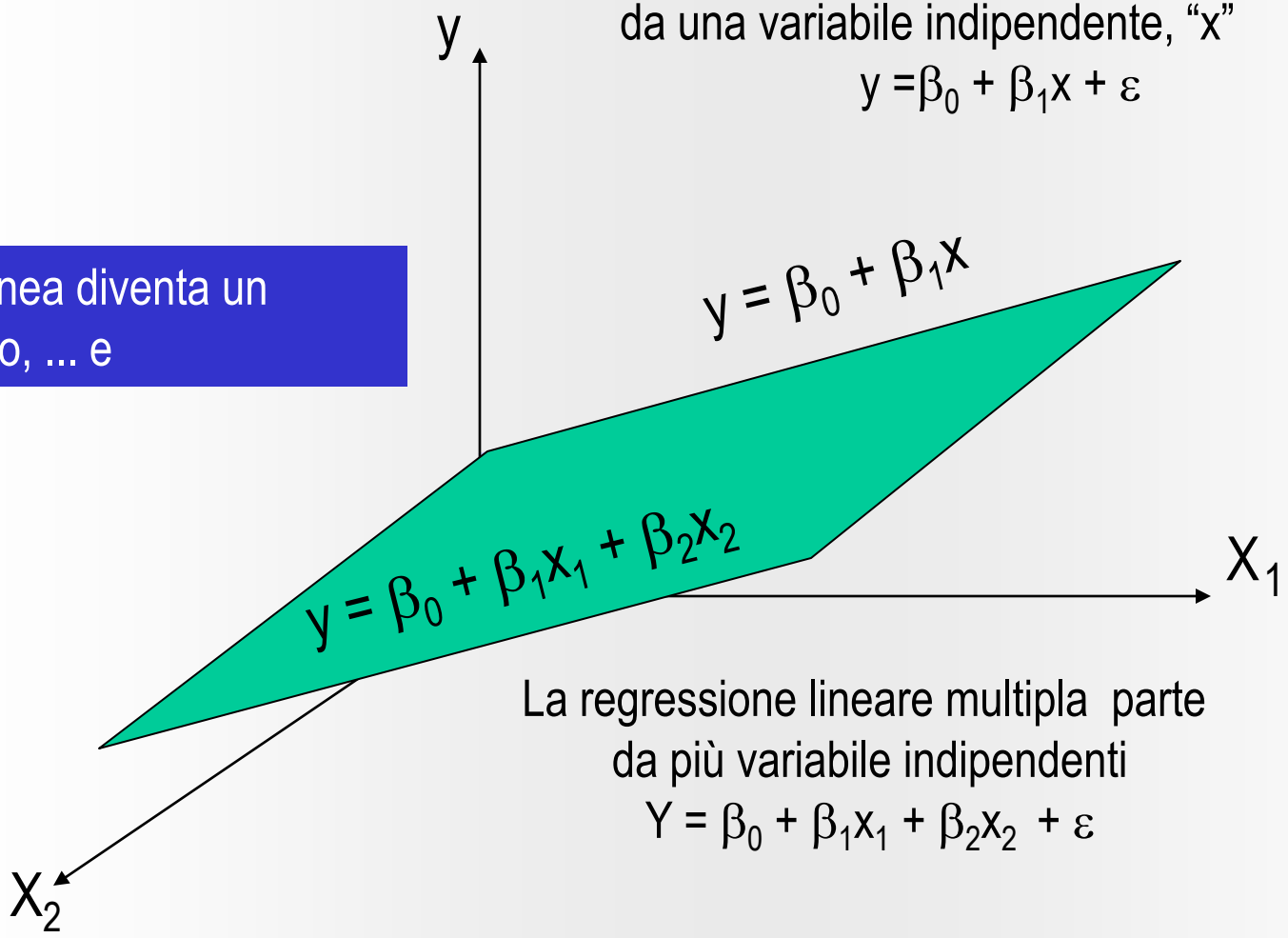
- Partendo da k variabili indipendenti, si stima la variabile dipendente Y



La regressione lineare semplice parte da una variabile indipendente, “x”

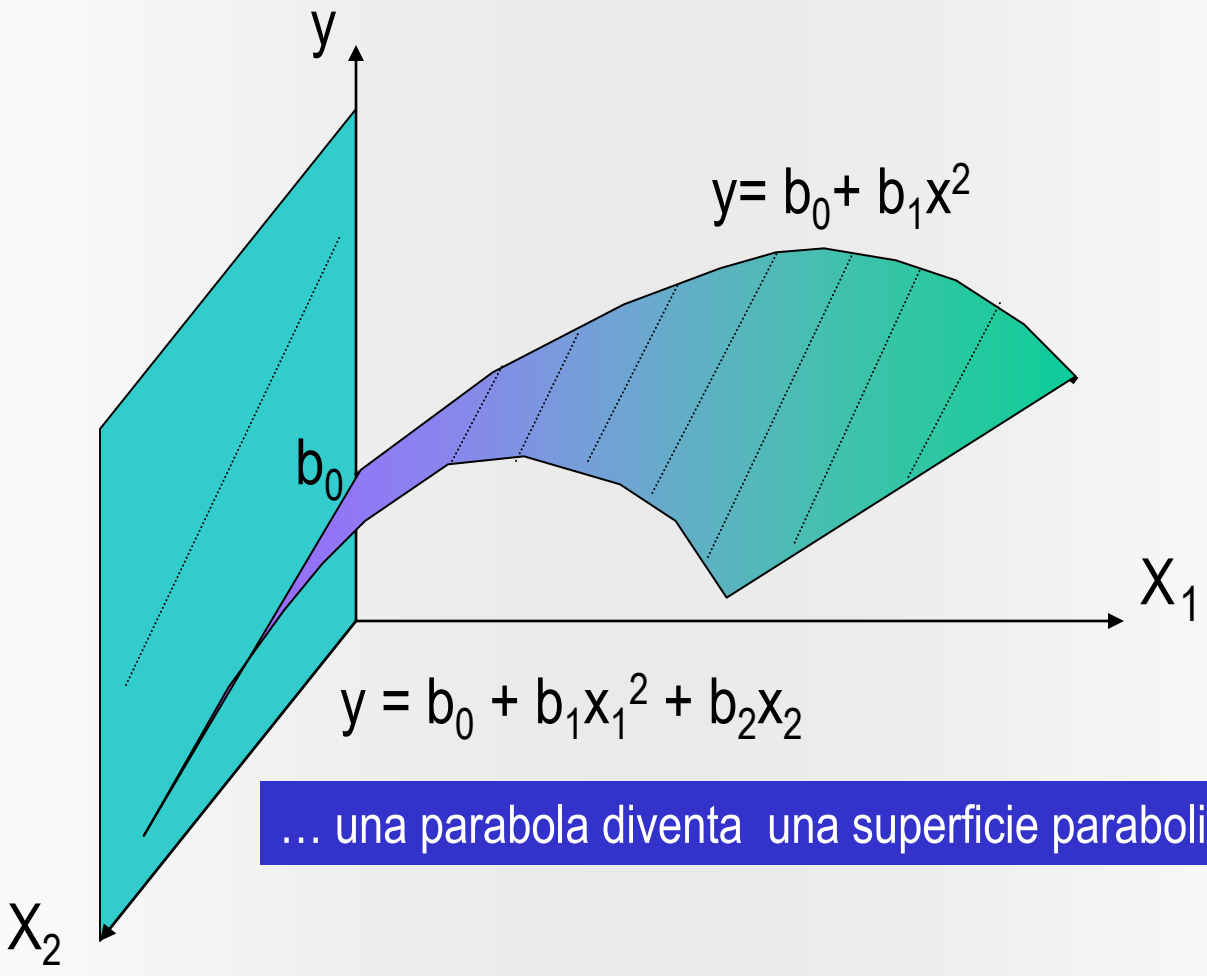
$y = \beta_0 + \beta_1 x + \varepsilon$

La linea diventa un piano, ... e



La regressione lineare multipla parte da più variabile indipendenti

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$



… una parabola diventa una superficie parabolica

- Utilizzando l'algebra lineare, si ha:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ y_i \\ \cdot \\ y_n \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \varepsilon_i \\ \cdot \\ \varepsilon_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdot & \cdot & X_{1k} \\ 1 & X_{21} & X_{22} & & & X_{2k} \\ & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & X_{i1} & X_{i2} & \cdot & \cdot & X_{ik} \\ & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & X_{n1} & X_{n2} & \cdot & \cdot & X_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \beta_i \\ \cdot \\ \beta_k \end{bmatrix}$$

Vettore (nx1) Vettore (nx1) Matrice (nxk) Vettore (kx1)

Il modello può essere scritto nella forma compatta $\mathbf{y = X\beta + \varepsilon}$

e la stima dei m.q. $\boldsymbol{\beta} = (\mathbf{X'X})^{-1}\mathbf{x'y}$

- Le ipotesi su cui si basa il modello di regressione multipla costituiscono una generalizzazione delle ipotesi introdotte nel caso del modello di regressione semplice

Ipotesi 1: La grandezza $X\beta$ definisce la parte sistematica del modello mentre la variabile ε definisce la componente di errore

Ipotesi 2: La matrice delle v.c. X viene fissata alla particolare realizzazione $X=x$

Ipotesi 3: La matrice X ha rango uguale a k $(k<n)$ e conseguentemente la matrice $X'X$ non è singolare

Ipotesi 4: Il vettore casuale ε ha valore atteso 0 $(E(\varepsilon/X))$ e quindi conseguentemente $E(y/X)=X\beta$

Ipotesi 5: La matrice delle varianze e covarianze del vettore casuale ε è data da $E(\varepsilon\varepsilon'/X)=\sigma^2 I_k$ dove I_k è la matrice di Identità di ordine k

- Le ipotesi 4 e 5 riguardano la v.c. errore ε ed implicano che :
  - La v.c. ε si distribuisca come una normale con media zero e varianza costante *(omoschedasticità):*

$$\text{Var}\,(\varepsilon_i \backslash X) = \sigma^2$$

  - Gli errori sono indipendenti ovvero:

$$\text{Cov}\,(\varepsilon_i\, \varepsilon_j \backslash X) = 0$$

- Queste ipotesi sono necessarie per :
  - Stimare i coefficienti del modello
  - Valutare i risultati del modello

Stima dei coefficienti e valutazione del modello

- La procedura:
  - Ottenere i coefficienti e le statistiche del modello utilizzando un software statistico;
  - Diagnosticare le violazioni alle ipotesi del modello, cercando di risolvere i problemi se presenti;
  - Valutare la bontà di adattamento del modello utilizzato;
  - Se il modello supera i test diagnostici, utilizzare i coefficienti per prevedere i valori della y.

Esempio : Dove localizzare un nuovo albergo?

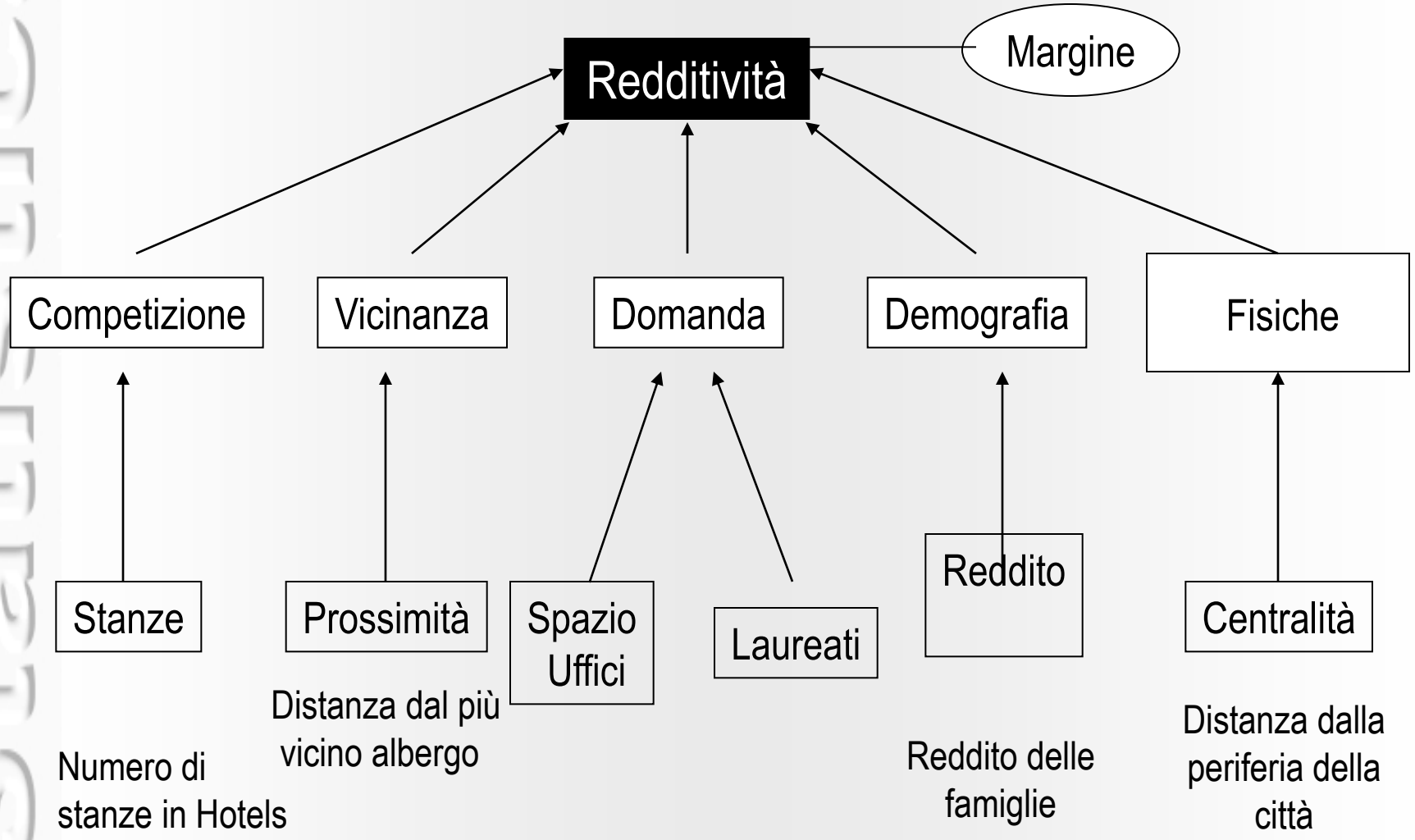
– La Società Internazionale Holiday (SIH)

sta pianificando un'espansione.

– Il management vuole predire quale area geografica sia più redditizia

– I predittori di reddito possono essere inclusi nelle seguenti classi:

  - Competizione
  - Consapevolezza del mercato
  - Generatori di domanda
  - Statistiche demografiche
  - Qualità fisiche


Redditività
Margine
Competizione
Vicinanza
Domanda
Demografia
Fisiche
Stanze
Prossimità
Spazio Uffici
Laureati
Reddito
Centralità
Numero di stanze in Hotels nel raggio di tre Km2
Distanza dal più vicino albergo
Reddito delle famiglie
Distanza dalla periferia della città

– I dati sono stati raccolti estraendo casualmente 100 alberghi e costruendo il seguente modello

$$\textbf{Redditività} = \beta_0 + \beta_1 \textbf{Stanze} + \beta_2 \textbf{Prossimità} + \beta_3 \textbf{Uffici} + \beta_4 + \beta_5 \textbf{Reddito} + \beta_6 \textbf{Centralità} + \textbf{Errore} \qquad \varepsilon$$

alberghi	Redditività	Stanze	Vicinanza	uffoici	Laureati	Reddito	Centralità
1	55,5	3203	0,1	549	8	37	12,1
2	33,8	2810	1,5	496	17,5	39	0,4
3	49	2890	1,9	254	20	39	12,2
4	31,9	3422	1	434	15,5	36	2,7
5	57,4	2687	3,4	678	15,5	32	7,9
6	49	3759	1,4	635	19	41	4

Questa è l'equazione della regressione (equazione predittiva)

Redditività = 72.455 **-** 0.008**Stanze -**1.646**Prossimità**
+ 0.02**Uffici** +0.212**Laureati**
- 0.413**Reddito** + 0.225**Centralità**

R quadro corretto	0,
Errore Standard	5,
N. osservazioni	

ANOVA

	df	*SS*	*MS*	*F*	*Significatività F*
Regressione	6	3123,832006	52	6	
Residua	93	2825,625894	30	8	
Totale	99	5949,4579			

Valutiamo questa equazione

	Coefficienti	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercetta		7,893103745	9,179483	1,11388E-14	56,78048735	88,12874
Stanze		0,00125527	-6,068708	2,7662E-08	-0,010110582	-0,005125
Prossimità		0,632836913	-2,601361	0,010803327	-2,902924523	-0,38955
Uffici		0,003410442	5,795594	9,24331E-08	0,012993085	0,026538
Laureati		0,133427935	1,587246	0,115851278	-0,053178229	0,476744
Reddito		0,139552395	-2,960337	0,003898802	-0,690245235	-0,135999
Centralità		0,178708888	1,260475	0,210651417	-0,12962198	0,580138

- L'errore standard delle stime
  - Stimiamo l'errore standard delle stime

$$s_\varepsilon = \sqrt{\frac{SSE}{n-k-1}}$$

  - Confrontiamo s_ε con il valore medio di y
    - Dall'output l'Errore Standard è = 5.5121
    - Calcolando il valore medio di y si ha
  - s_ε non è particolarmente piccolo $\bar{y} = 45.739$
  - Si può dire che il modello non approssimi bene i dati?

- Il Coefficiente di determinazione
  - La definizione è

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

  - Dall'output, $R^2 = 0.5251$
  - Il 52.51% della variazione della misura di redditività è spiegato dal modello di regressione lineare descritto sopra
  - Quando consideriamo i gradi di libertà
    R^2 corretto = 1-[SSE/(n-k-1)] / [SS(Totale)/(n-1)] =
    = 49.44%

- Il test di validità del modello

  – C'è almeno una variabili indipendente legata linearmente alla variabile dipendente?

  – Per rispondere a questa domanda testiamo l'ipotesi:

$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$

H_1: Almeno un β_i non è uguale a zero

  – Se almeno un β_i non è uguale a zero, il modello è valido

- Per testare questa ipotesi si utilizza la procedura dell'analisi della varianza
- Il test F
  - Si costruisce la statistica F

[Varianza di y] = SSR + SSE.
Un valore grande di **F** deriva da un valore grande di SSR.
Allora, molta della variabilità di y è spiegata dal modello di regressione
L'ipotesi nulla può essere rifiutata, dunque il modello è valido.

$$F = \frac{MSR}{MSE}$$

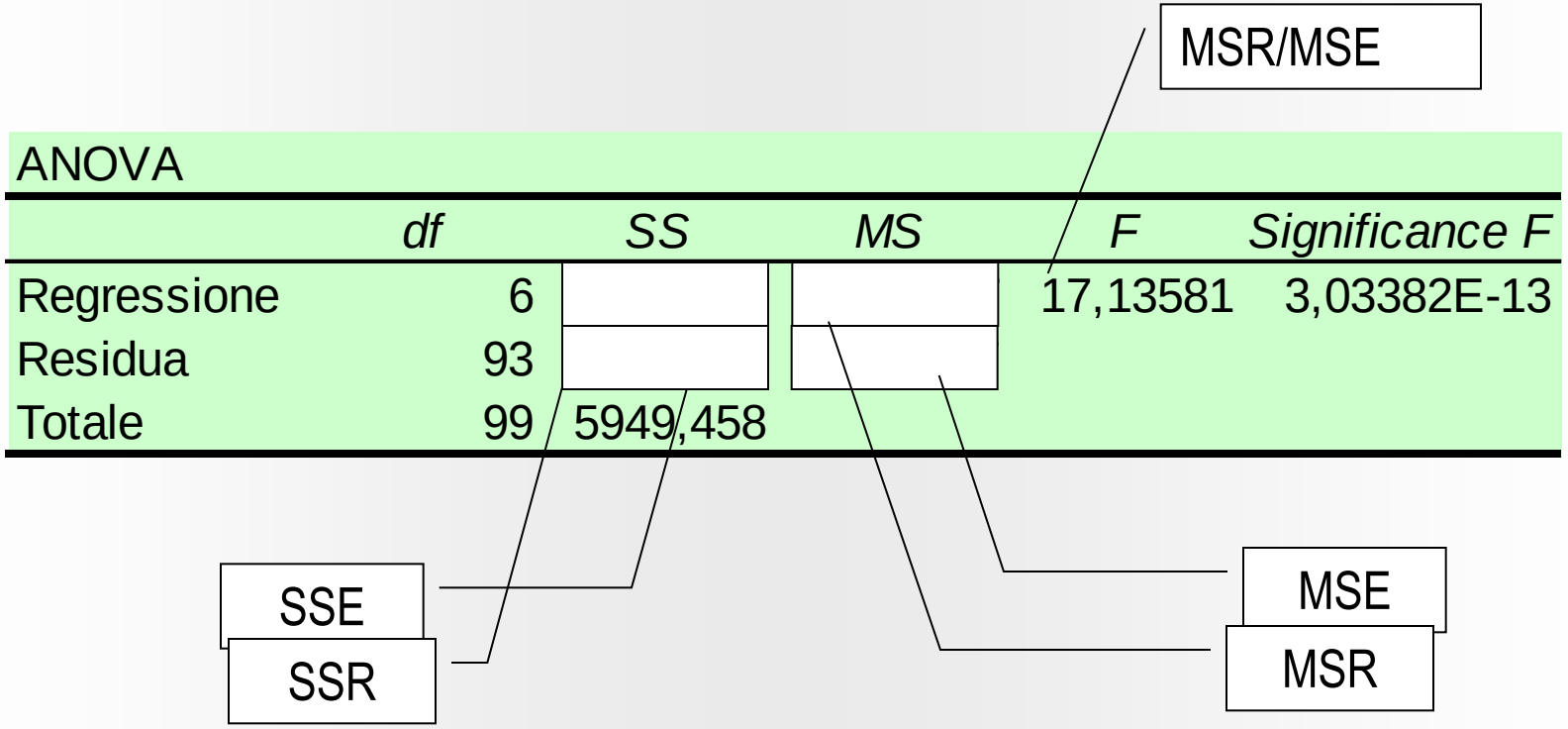
MSR=SSR/k

MSE=SSE/(n-k-1)

$F > F_{\alpha,k,n-k-1}$

La condizione richiesta deve essere soddisfatta

Esempio - continua



Esempio - continua

ANOVA					
					cance F
Regressior					382E-13
Residua					
Totale					

Conclusione: C'è sufficiente evidenza per rifiutare l'ipotesi nulla a favore dell'ipotesi alternativa.
Almeno un coefficiente β_i non è uguale a zero, ovvero almeno una variabile indipendente è legata linearmente a y.

Il modello di regressione lineare è valido

$F_{\alpha,k,n-k-1} = F_{0.05,6,100-6-1}=2.17$
$F = 17.14 > 2.17$

Il p-value = $3.03382(10)^{-13}$
Chiaramente $\alpha = 0.05 > 3.03382(10)^{-13}$ e l'ipotesi nulla è rifiutata

- Interpretiamo i coefficienti

  - $b_0 = 72.5$ E' l'intercetta, ovvero il valore di y quando tutte le variabili indipendenti hanno valore zero. Se però il range di almeno una variabile indipendente non comprende lo zero, l'intercetta non è interpretabile.
  - $b_1 = -.0076$ In questo modello per ogni 1000 stanze in più, all'interno di una area di 3 Km dall'albergo SIH la redditività decresce in media del 7.6% (assumendo costanti le altre variabili).

- $b_2 = -1.65$ Per ogni Km in più dal più vicino concorrente la redditività media decresce del 1.65%

- $b_3 = .02$ Per ogni 1000 m^2 di spazio dedicato agli uffici, la redditività media cresce dello 0.02%.

- $b_4 = .21$ Per 100 Laureati in più, la redditività media cresce dello 0.21%.

- $b_5 = -.41$ Per 1000 Euro in più del reddito medio delle famiglie la redditività media decresce dello 0.41%

- $b_6 = .23$ Per ogni km in più di distanza dalla periferia, la redditività media cresce dello 0.23%

- Il test dei coefficienti
  - Le ipotesi per ogni β_i

$H_0: \beta_i = 0$

$H_1: \beta_i \neq 0$

Test statistico

$$t = \frac{b_i - \beta_i}{s_{b_i}}$$

d.f. = n - k -1

  - Excel output

	Coefficients	*Standard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercetta	72,45461	7,893104	9,179483	1,11E-14	56,78048735	88,12874
Stanze	-0,00762	0,001255	-6,06871	2,77E-08	-0,010110582	-0,00513
Prossimità	-1,64624	0,632837	-2,60136	0,010803	-2,902924523	-0,38955
Uffici	0,019766	0,00341	5,795594	9,24E-08	0,012993085	0,026538
Laureati	0,211783	0,133428	1,587246	0,115851	-0,053178229	0,476744
Reddito	-0,41312	0,139552	-2,96034	0,003899	-0,690245235	-0,136
Centralità	0,225258	0,178709	1,260475	0,210651	-0,12962198	0,580138

- L’utilizzo dell’equazione di regressione lineare
  - Il modello può essere utilizzato per:
    - Predire un intervallo per un particolare valore di y, a partire da un set di valori dati di x_i.
    - Produrre una stima ad intervallo per un valore atteso di y, a partire da un set di valori dati di x_i.
  - Il modello può essere utilizzato per capire le relazioni tra le variabili indipendenti x_i, e la variabile dipendente y, attraverso l’interpretazione dei coefficienti β_i

- Esempio 8.1: Previsione
  - Prevedere la redditività di un albergo con le seguenti caratteristiche:
    - 3815 stanze nel raggio di 3 km^2,
    - Gradi concorrenti a 3,4 Km,
    - 476.000 m^2 metri dedicati ad uffici,
    - 24.500 studenti iscritti all'università,
    - 39.000 Euro il reddito medio delle famiglie,
    - 3,6 Km la distanza dalla periferia della città.

Redditività = 72.455 **-** 0.008**(3815) -**1.646**(3,4)** + 0.02(**476,000)**
+0.212**(24.500)** - 0.413(**39.000)** + 0.225**(3,6) = 37.1%**

Diagnostica di Regressione

- Bisogna verificare che sia rispettate le ipotesi alla base del modello
  - La variabile errore si distribuisce in modo normale? — Istogramma dei residui
  - La varianza dell'errore è costante? — Plot dei residui verso $\hat{y}$
  - Sono gli errori indipendenti? — Plot dei residui
  - Vi sono outliers?
  - C'è il problema della multicollinearità?

- Esempio : Prezzi degli appartamenti e multicollinearità
  - Un venditore di appartamenti sostiene che il prezzo possa essere determinato usando la dimensione della casa, il numero di stanze e la dimensione dello stabile in cui è inserita
  - Un campione di 100 case è estratto in modo casuale

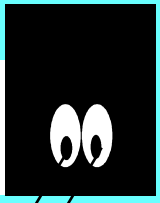
Prezzo	Stanze	C_ Dim	S_Dim
124100	3	1290	3900
218300	4	2080	6600
117800	3	1250	3750
.	.	.	.
.	.	.	.

  - Si analizzano le relazione tra le variabili

- Il modello proposto è:
  PREZZO = $\beta_0 + \beta_1$**Stanze** + β_2**C-Dim** +β_3**S-dim** + ε
  - Excel soluzione

Regressione	
R multiplo	0,74833
R quadro	0,559998
R quadro corretto	0,546248
Errore Standard	25022,71
N. osservazioni	100

Il modello è valido ma non variabile è legata al prezzo !!



ANOVA

	df	*SS*	*MS*	*F*	*Significatività F*
Regressione	3	76501718347	25500572782	40,7269	4,56894E-17
Residua	96	60109046053	626135896,4		
Totale	99	1,36611E+11			

	Coefficienti	*Errore Standard*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercetta	37717,59	14176,74195	2,660526279	0,009145	9576,962637	65858,23
Stanze	2306,081	6994,19244	0,329713665	0,742335	-11577,2921	16189,45
C-Dim	74,29681	52,97857934	1,402393325	0,164023	-30,8649232	179,4585
S-Dim	-4,363783	17,0240013	-0,256331212	0,798244	-38,1561842	29,42862

- Inoltre
  - Quando, regrediamo il prezzo per ogni variabile singolarmente, si trova che ogni variabile è strettamente legata al prezzo
  - La multicolinearità è la causa di questo problema

	Prezzo	*Stanze*	*C-Dim*	*S-Dim*
Prezzo	1			
Stanze	0,645411	1		
C-Dim	0,747762		1	
S-Dim	0,740874			1

- La multicollinearità causa due tipi di difficoltà:
  - La statistica **t** appare molto piccola
  - Il coefficiente β non può essere interpretato come "pendenza"

- Soluzione alle violazioni delle ipotesi alla base del modello

  - I problemi di **Non normalità** o di **eteroscedasticità** possono essere risolti usando una trasformazione della variabile di y.
  - La trasformazione può migliorare la relazione lineare tra la variabile dipendente e le variabili indipendenti.
  - La maggior parte dei programmi software consentono di trasformare facilmente la variabile y

- Una breve lista di trasformazioni
  - » $y' = \log y$ (con $y > 0$)
    - Si usa quando s_ε cresce con y, o
    - quando la distribuzione dell'errore ha una asimmetria positiva
  - » $y' = y^2$
    - Si usa quando s^2_ε è proporzionale a E(y), o
    - quando la distribuzione dell'errore ha una asimmetria negativa
  - » $y' = y^{1/2}$ (con $y > 0$)
    - Si usa quando s^2_ε è proporzionale a E(y)
  - » $y' = 1/y$
    - Si usa quando s^2_ε cresce significativamente al crescere di y

- Esempio : Analisi, diagnostica, trasformazioni

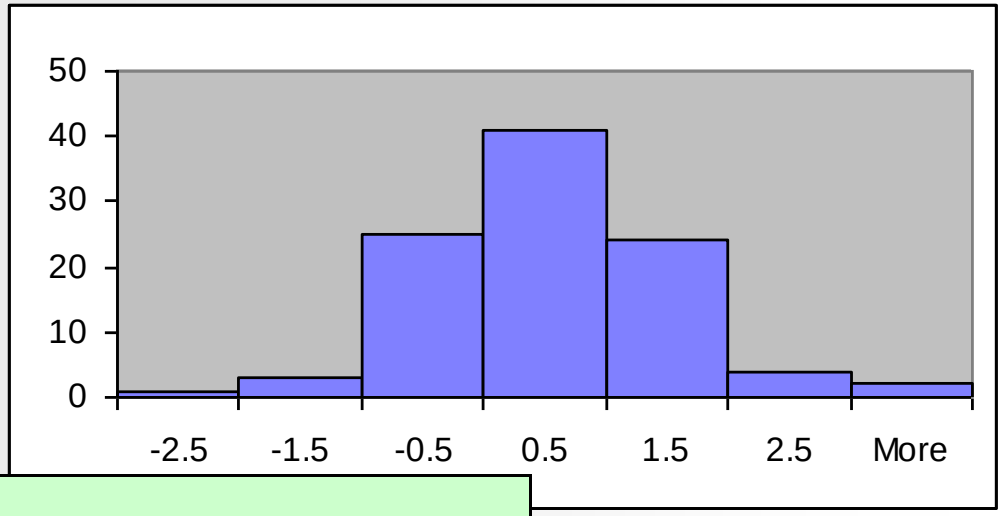
  – Un professore vuole sapere se il tempo limite per la compilazione di un test incide sui risultati
  – Un campione casuale di 100 studenti viene diviso in 5 gruppi
  – Ogni gruppo ha un differente tempo limite

Tempo	40	45	50	55	60
Punteggi	*20*	*24*	*26*	*30*	*32*
	23	*26*	*25*	*32*	*31*

Si analizzano i risultati includendo la diagnostica

Il modello:

Punteggio = $\beta_0 + \beta_1$**tempo** + ε



OUTPUT

Regressione	
R Multiplo	0,862539519
R quadro	0,743974422
R quadro corretto	0,741361916
Errore Standard	2,304609401
N. osservazioni	100

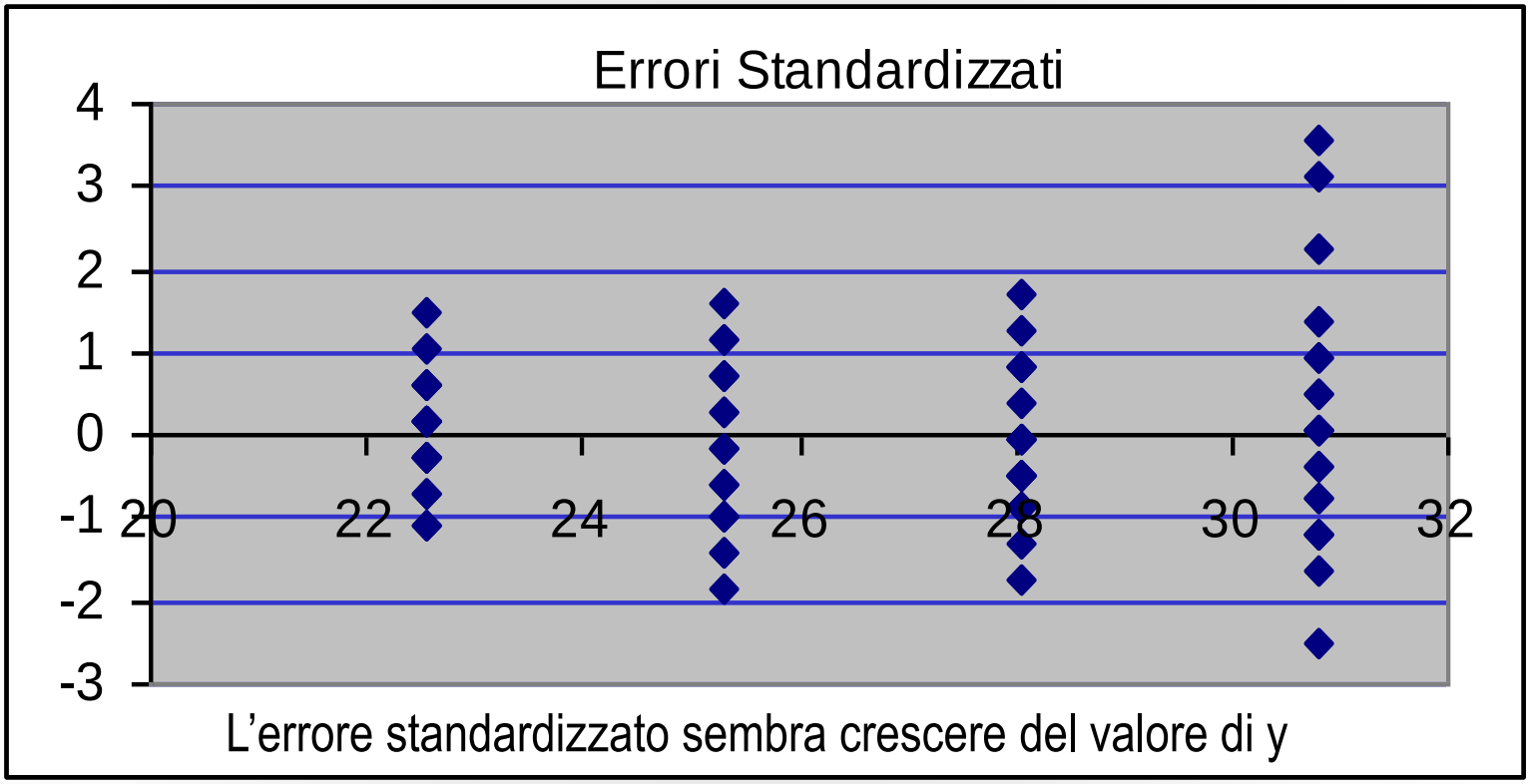
Il modello è buono e fornisce un buon "fit"

L'errore sembra si distribuisca in modo normale

ANOVA

	df	*SS*	*MS*	*F*	*Significatività F*
Regressione	1	1512,5	1512,5	284,7742555	9,41548E-31
Residua	98	520,5	5,311224		
Totale	99	2033			

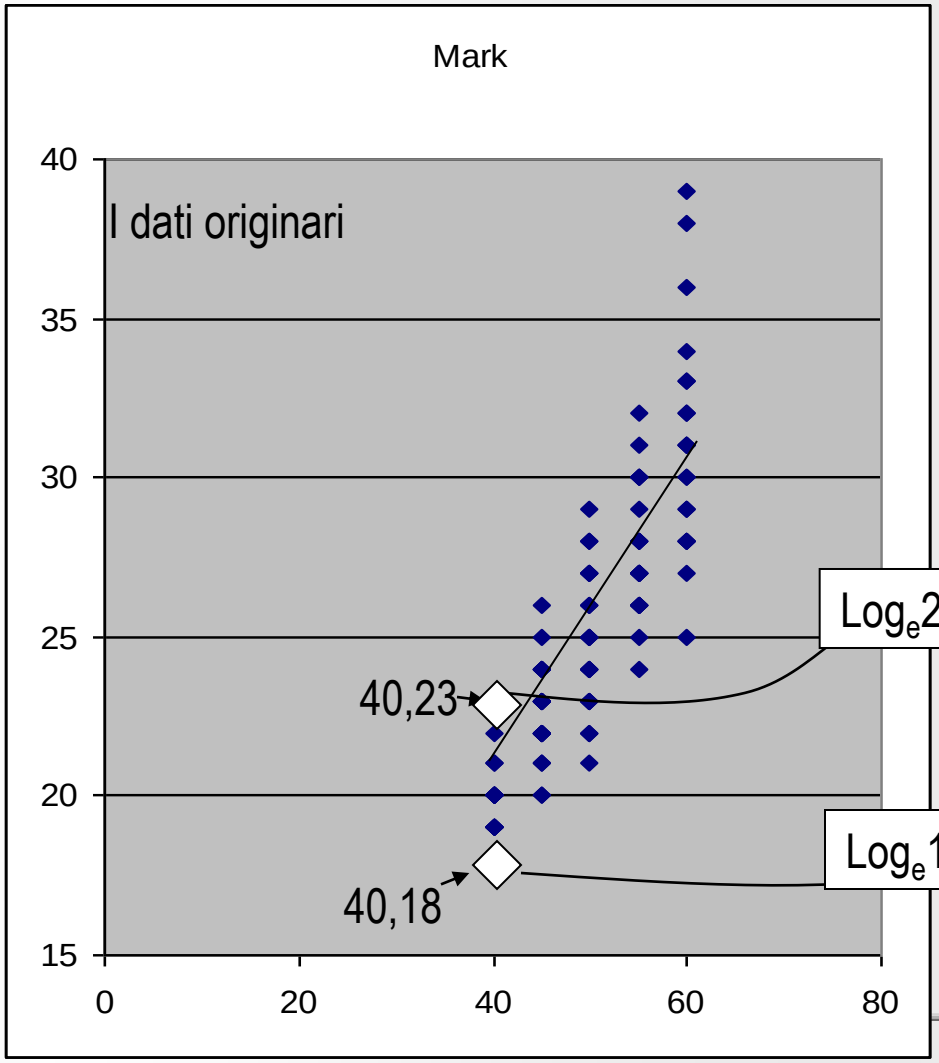
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercetta	-2,2	1,645820309	-1,336719	0,18440922	-5,466076778	1,066077
Time	0,55	0,032592099	16,87526	9,41548E-31	0,485322042	0,614678

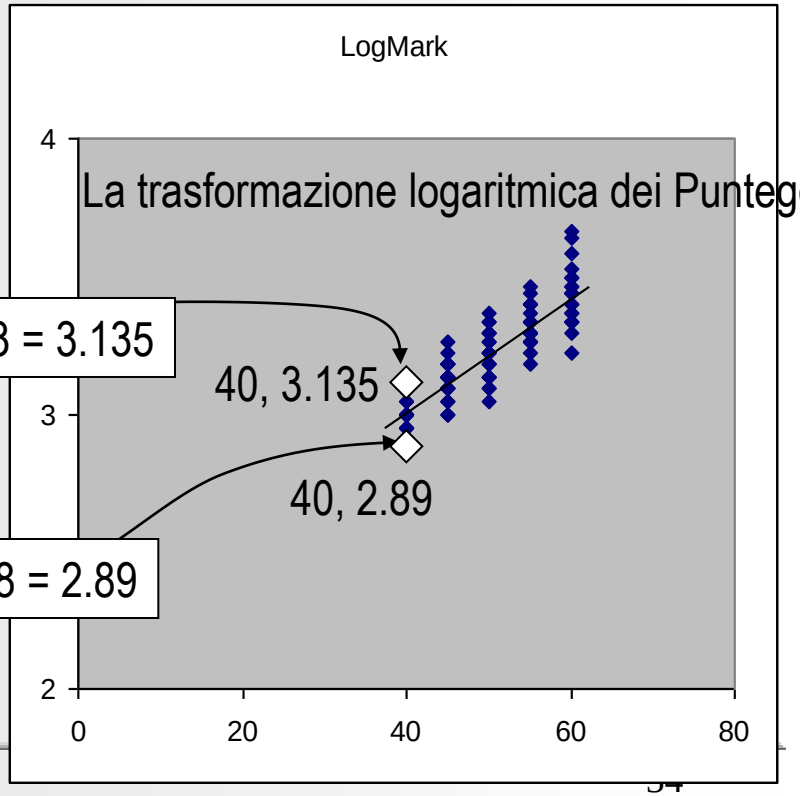


Due trasformazioni possono essere usate per risolvere questo problema:

1. $y' = \log_e y$
2. $y' = 1/y$

Vediamo cosa succede applicando la trasformazione





$Log_e 23 = 3.135$

$Log_e 18 = 2.89$

La nuova regressione è:

Log Punteggio= 2.1295 + .0217Tempo

Il modello:

LOG Punteggio = β'_0 + β'_1**Tempo** + ε'

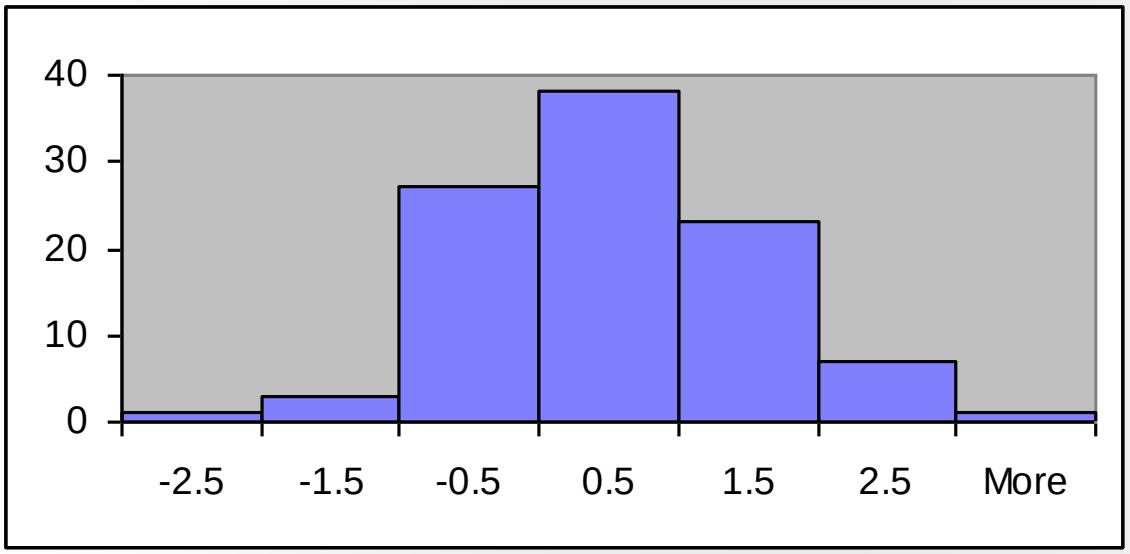
Regressione	
R Multiplo	0,878300373
R quadro	0,771411546
R quadro corretto	0,769079011
Errore Standard	0,0844372
N. osservazioni	100

Il modello è buono e fornisce un buon "fit"

ANOVA

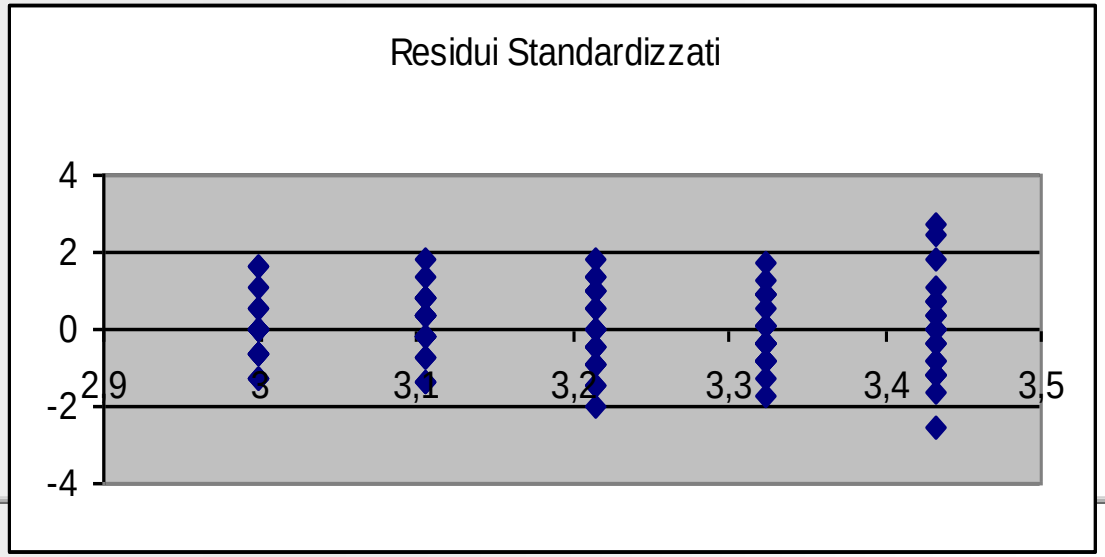
	df	*SS*	*MS*	*F*	*Significance F*
Regressione	1	2,357901025	2,357901	330,7180661	3,58062E-33
Residua	98	0,698704801	0,0071296		
Totale	99	3,056605826			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercetta	2,129582054	0,060300222	35,316322	1,51409E-57	2,009918227	2,2492459
Tempo	0,021715898	0,001194122	18,185656	3,58062E-33	0,019346201	0,0240856



Gli errori si distribuiscono in modo normale

Gli errori standardizzati variano al variare di y ma in modo inferiore a prima



Consideriamo il **Tempo** = 55 minuti

Log Punteggio = 2.1295 + .0217Tempo = 2.1295 + .0217(55) = 3.323

Per trovare il valore di y , bisogna calcolare l'antilogaritmo:

$\text{antilog}_e 3.323 = e^{3.323} = 27.743$

La Diagnostica di Regressione per le serie storiche

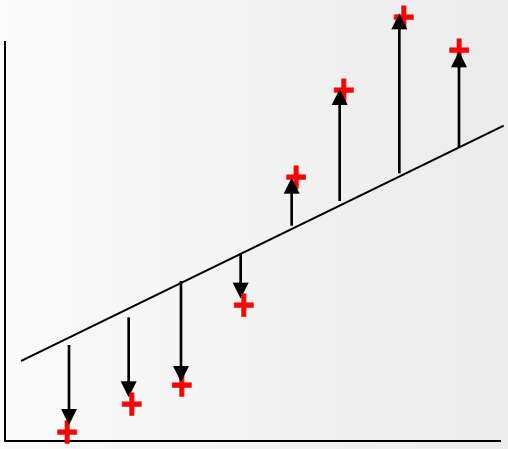
- Il test di Durbin - Watson
  - Questo test serve a dimostrare la violazione dell'ipotesi di indipendenza seriale o semplicemente di autocorrelazione tra gli errori
  - Se c'è autocorrelazione, le variabili errore non sono indipendenti

Residui al tempo i

$$d = \frac{\sum_{i=2}^{n} (r_i - r_{i-1})^2}{\sum_{i=1}^{n} r_i^2}$$

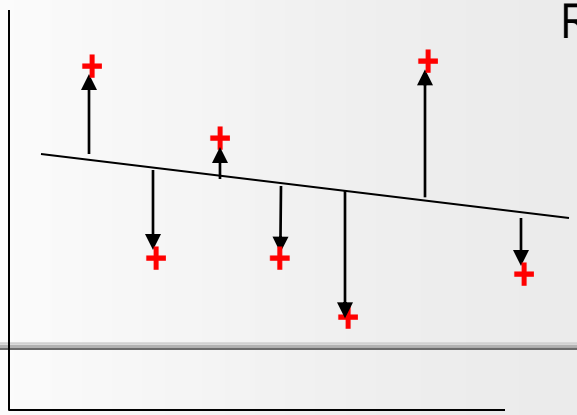
$$Il\ range\ di\ \ d\ \ è\ 0 \le d \le 4$$

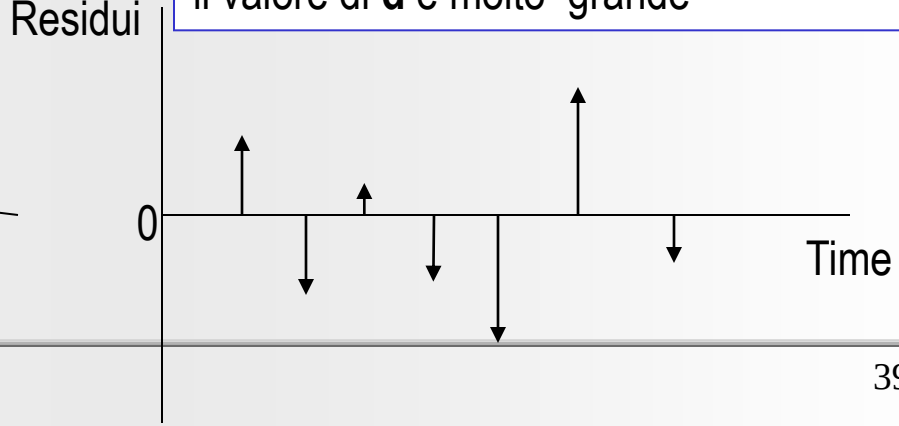
Autocorrelazione positiva di I° grado



L'autocorrelazione positiva si ha quando residui consecutivi tendono ad essere simili. Così il valore di **d** molto piccolo.

Autocorrelazione negativa di I° grado



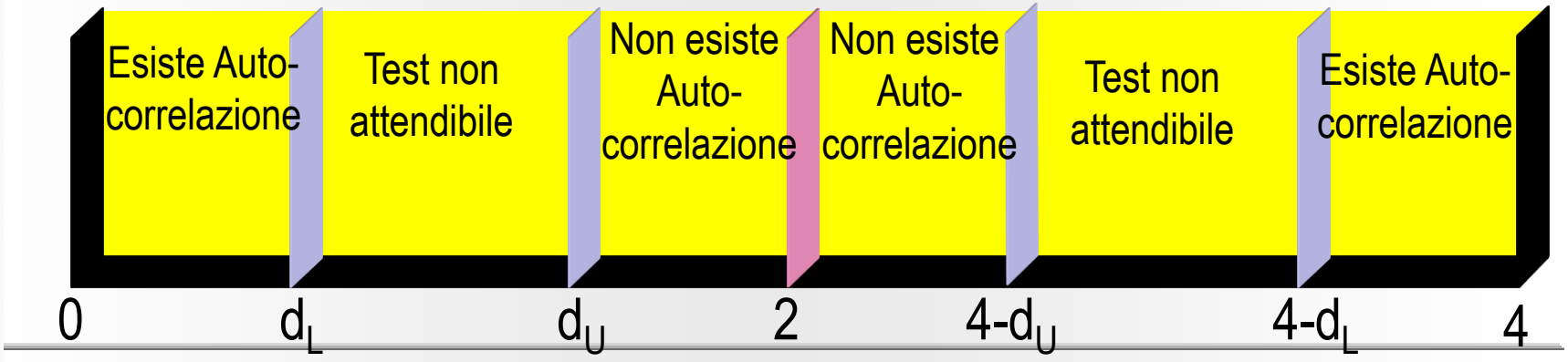


L'autocorrelazione negativa si ha quando residui consecutivi tendono differire sensibilmente. Così il valore di **d** è molto grande

- *Test ad una coda per l'autocorrelazione* ***positiva***
  - Se $d<d_L$ c'è sufficiente evidenza per dimostrare che c'è autocorrelazione
  - Se $d>d_U$ non c'è sufficiente evidenza per dimostrare che c'è autocorrelazione
  - Se d è tra d_L e d_U il test non è attendibile.

- *Test ad una coda per l'autocorrelazione* ***negativa***
  - Se $d>4-d_L$, c'è correlazione negativa
  - If $d<4-d_U$, non c'è correlazione negativa
  - Se d cade tra $4-d_U$ e $4-d_L$ il test non è attendibile.

- Test a due code per l'autocorrelazione
  - Se $d<d_L$ o $d>4-d_L$ esiste l'autocorrelazione
  - Se d cade tra d_L e d_U o tra $4-d_U$ e $4-d_L$ il test non è attendibile
  - Se d cade tra d_U e $4-d_U$ non c'è sufficiente evidenza per affermare che c'è autocorrelazione



- Esempio
  - Come il clima influisce sulla vendita dei biglietti di un impianto di risalita di una stazione sciistica?
  - Sono stati raccolti i dati di 20 anni di vendita di biglietti, di altezza della neve e di temperatura media durante la settimana natalizia
  - Il modello ipotizzato è:

    $$\textbf{Biglietti} = \beta_0 + \beta_1 \textbf{Neve} + \beta_2 \textbf{Temperatura} + \varepsilon$$

  - La regressione ha dato i seguenti risultati

OUTPUT

Regressione	
R Multiplo	0,34645292
R quadro	0,12002962
R quadro corretto	0,0165037
Errore Standard	1711,6764
N. osservazioni	20

Il modello non sembra buono:

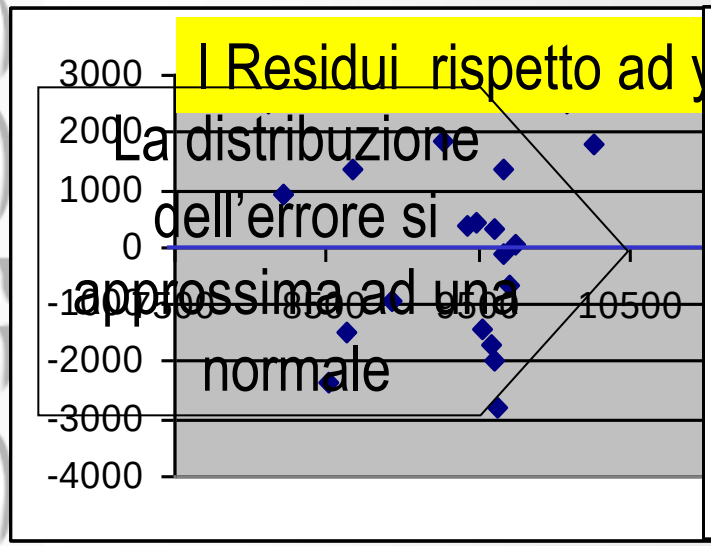
- L'adattamento è basso (R-square=0.12),
- Non è significativo (Signif. F =0.33)
- Nessuna variabile è legata linearmente alla vendita

La diagnostica rileva le seguenti violazioni delle ipotesi di base

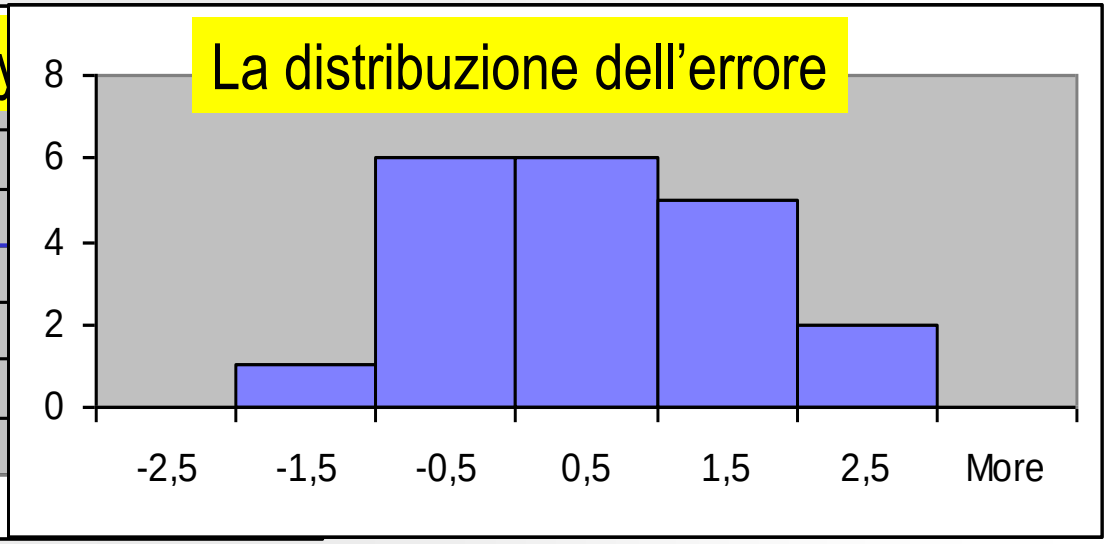
ANOVA

	df	*SS*	*MS*	*F*	*Signif. F*
Regressione	2	6793798,248	3396899,1	1,1594	0,3372706
Residui	17	49807213,95	2929836,1		
Totale	19	56601012,2			

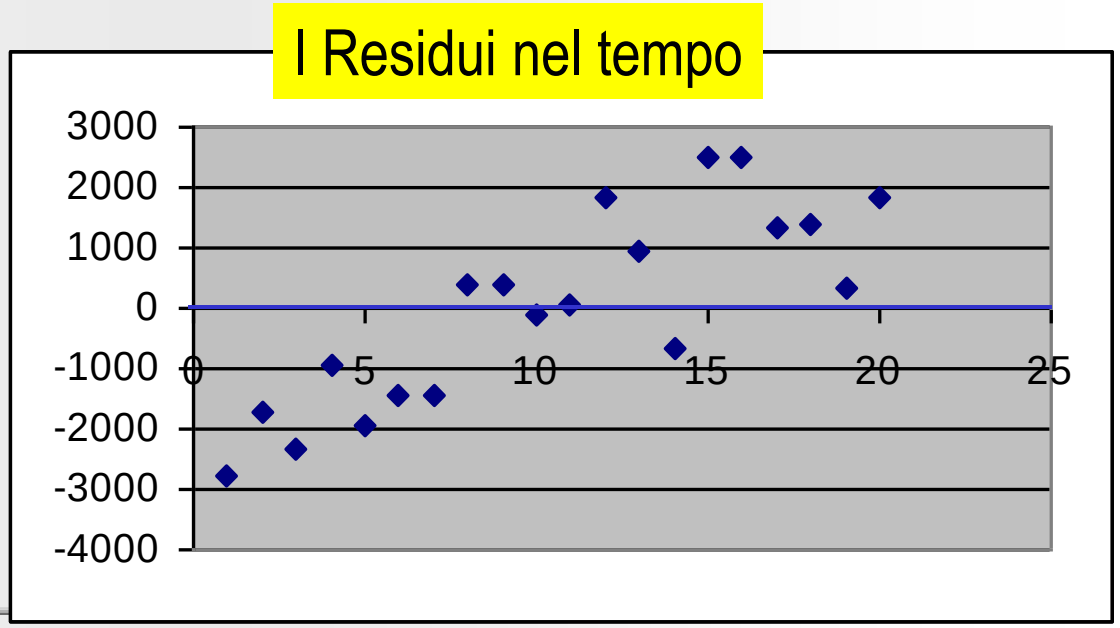
	Coefficienti	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercetta	8308,01142	903,7284951	9,1930391	5E-08	6401,3083	10214,715
Neve	74,5932493	51,57482923	1,4463111	0,1663	-34,22028	183,40678
Temperatura	-8,75373781	19,70435896	-0,444254	0,6625	-50,32636	32,818884



La distribuzione dell'errore si approssima ad una normale



Gli errori non sono indipendenti



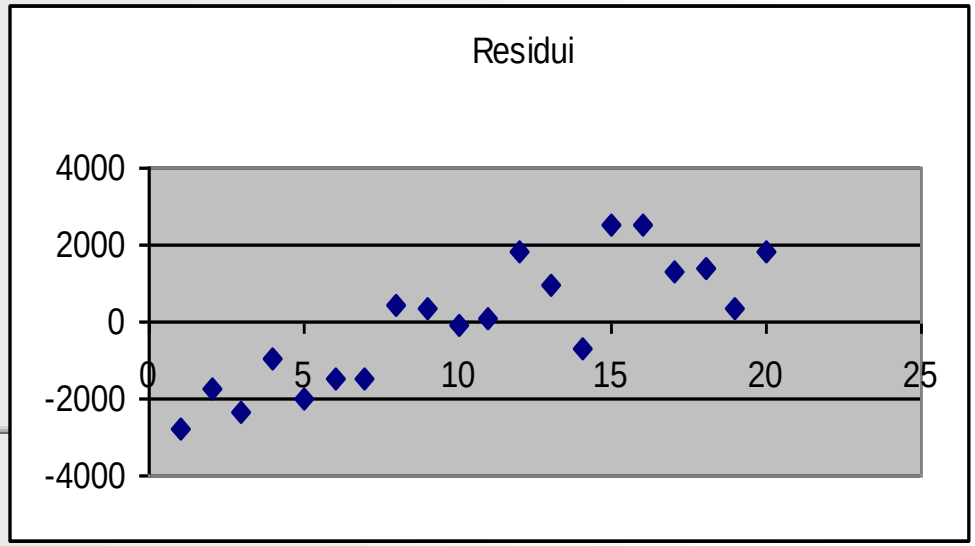
Test per l' auto-correlazione positiva:
n=20, k=2. Dalla tavola Durbin-Watson abbiamo:
d_L=1.10, d_U=1.54. La statistica d=0.59

Conclusione: Poiché $d<d_L$, c'è sufficiente evidenza statistica per dire che c'è autocorrelazione

Durbin-Watson Statistic

-2793.99
-1723.23 *d = 0.5931*
-2342.03
-956.955
-1963.73
.
.

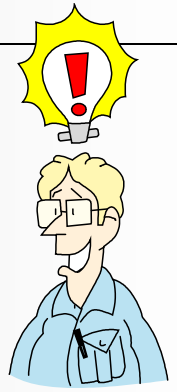
I residui



Il mo

La autocorrelazione è causata dal tempo.
Per correggere il problema, nel modello bisogna aggiungere la variabile dipendente "tempo"

tura+ β_3**Anni**+ε



- Tutte le ipotesi di base sono rispettate
- L'adattamento del modello ai dati è alto $R^2 = 0.74$.
- Il modello è buono. La significatività è: F = 5.93 E-5.
- **L a neve** e **gli anni** sono legate linearmente alla vendita di biglietti.
- **La Temperatura** non è linearmente legata alla vendita di biglietti