Stagionalità con le dummies

I coefficienti di stagionalità (nel caso questa sia costante) si possono stimare con il modello di regressione cono variabili dicotome.

Partendo dal modello moltiplicativo $y_t = (T_t C_t) * S_t * u_t$

Si ipotizza che →

$$\begin{cases} T_t = \exp\left\{\sum_{i=0}^{m} \beta_i t^i\right\} \text{ polinomio in t} \\ C_t = \text{ciclicità inglobata nel ciclo-trend} \\ S_t = \sum_{j=1}^{k} \alpha_j S_{t,j} \text{ Stagionalità.} \end{cases}$$

k=periodo di stagionalità

$S_{t,j}$ sono i coefficienti fissi di stagonlaità

Il modello, nei logaritmi, è $Ln(y_t) = \sum_{i=0}^{m} \beta_i t^i + \sum_{j=1}^{k} \alpha_j S_{t,j} + e_t$

Stagionalità con le dummies/2

La stima dei parametri può effettuarsi con la regressione multipla ed anzi, le sue diagnostiche danno suggerimenti sull' impatto della stagionalità.

La stima dei parametri "α" deve rispettare il vincolo

$$\sum_{j=1}^{k} \alpha_j = -\beta_0$$

che esprime l'esaurirsi degli effetti stagionali ALL'INTERNO DELL'ANNO

Nell'ambito della formulazione ADDITIVA NEI LOGARITMI il significato degli "α" è quello già visto nel modello di regressione

α_j **misura come si modifica il trend, per gli effetti stagionali, nella stagione j-esima dell'anno.**

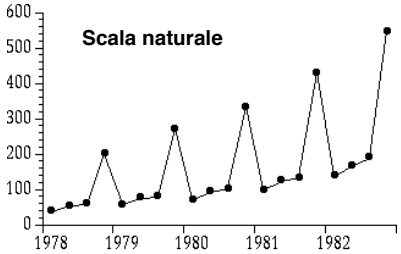
$$Ln(\hat{y}_{t,j}) = \sum_{i=0}^{m} \beta_i t^i + \alpha_j; \quad t = 1,2,\cdots,n; \; j = 1,2,\cdots,k$$

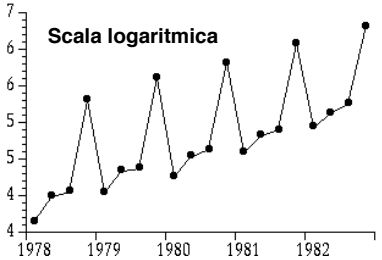
Esempio

La TOYS "R" è forse il maggiore rivenditore di giocattoli nel mondo (solo negli USA controllava 150 supermercati specifici del settore).

Ecco l'andamento delle sue vendite

Anni	Trim.	Vendite
1978	Feb.1-Apr.30	38,0
	Mag.1-Lug.31	53,6
	Ago.1-Ott.31	57,5
	Nov.1-Gen.31	200,0
1979	Feb.1-Apr.30	56,5
	Mag.1-Lug.31	75,8
	Ago.1-Ott.31	78,3
	Nov.1-Gen.31	269,7
1980	Feb.1-Apr.30	70,2
	Mag.1-Lug.31	92,7
	Ago.1-Ott.31	101,8
	Nov.1-Gen.31	332,6
1981	Feb.1-Apr.30	97,3
	Mag.1-Lug.31	123,7
	Ago.1-Ott.31	132,9
	Nov.1-Gen.31	429,4
1982	Feb.1-Apr.30	138,3
	Mag.1-Lug.31	167,6
	Ago.1-Ott.31	189,9
	Nov.1-Gen.31	545,9

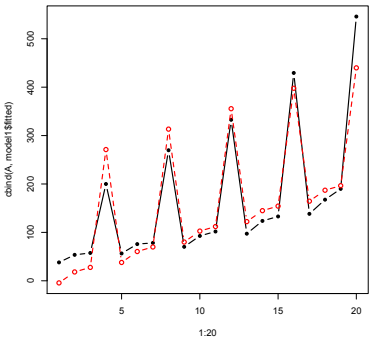




Esempio/2

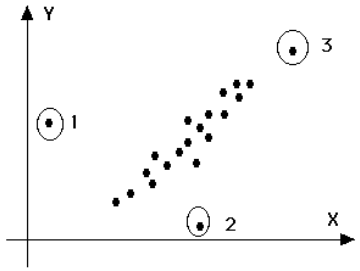
```
lm(formula = A ~ x1 + Q - 1)

Residuals:
   Min     1Q Median     3Q    Max
-71.11 -21.77  -9.92  21.44 105.97

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
x1   110.788     17.688   6.264 1.52e-05 ***
Q1Q  -14.901     24.353  -0.612    0.550
Q2Q   -2.833     25.436  -0.111    0.913
Q3Q   -3.984     26.582  -0.150    0.883
Q4Q  228.905     27.782   8.239 5.98e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 42.62 on 15 degrees of freedom
Multiple R-squared:  0.969, Adjusted R-squared:  0.9586
F-statistic: 93.69 on 5 and 15 DF,  p-value: 9.095e-11
```

Red= fitted



Valori remoti

Alcune osservazioni possono risultare così "remote" dalle altre da avere una influenza eccessiva sul modello e determinare un pessimo FITTING



I punti "1" e "2" sembrano in netto contrasto con la configurazione complessiva dei dati. Il "3", pur essendo anomalo, sembra collocarsi nel trend del fenomeno

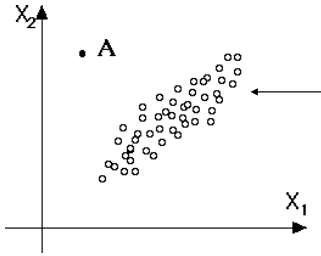
Da notare che il punto "1" è anomalo rispetto alla "X", il "2" rispetto alla "Y" ed il "3" rispetto ad entrambe

L'effetto della "1" e della "2" potrebbe non essere eccessivo: il valore della Y è coerente con l'andamento generale. La "2" invece infuisce molto (in modo negativo) dato che il suo "Y" è molto scentrato.

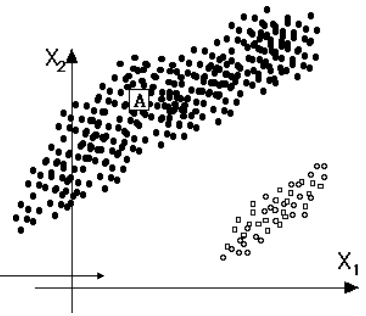
Valori remoti/2

Nelle applicazioni a dati reali qualche rilevazione scaturisce da circostanze inusuali: catastrofi naturali, problemi internazionali, cambiamenti politici, scioperi o serrate, etc.

C'è poi il rischio che certi dati siano sbagliati per mero errore materiale



In questi casi è necessario accertarsi che i valori remoti o anomali ci siano, ma su questa strada c'è incertezza



Non c'è alcuna garanzia che il punto "A" sia ANOMALO e gli altri NORMALI.

Un ampliamento delle rilevazioni potrebbe dar luogo ad uno scatter diverso

Diagnostiche per i valori remoti

Più importante è valutare l'influenza dei valori remoti sul modello.

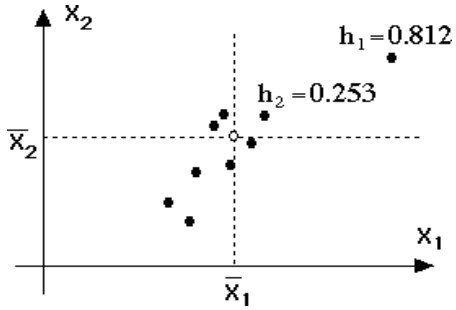
Nel caso della regressione lineare semplice è sufficiente lo studio dello scatterplot. Se i dati sono multidimensionali è necessario ricorrere a speciali formule.

Nel prosieguo studieremo tre diagnostiche:

- Un indice che esprima la posizione della i-esima osservazione rispetto alle altre
- Un indice che esprima l'effetto di eliminare l' i-esima osservazione sui valori stimati
- Un indice che esprima l'effetto di eliminare l i-esima osservazione sulla stima dei parametri

Esistono anche misure basate sull'effetto di cancellazione di più di una osservazione, ma non saranno considerate nel nostro corso

Uso della matrice Hat



Una leva prossima ad uno indica che l'osservazione è molto discosta dal "nucleo" dei dati ed un valore prossimo a zero significa che si colloca in prossimità del punto medio

Se la leva dell'i-esimo dato è grande essa contribuisce fortemente a determinare il valore stimato della risposta.

Come si è detto, stimate sono una combinazione lineare delle osservate

$$\hat{y} = Hy$$

Maggiore è h_i maggiore sarà il peso di X_i sul valore stimato. Al limite, se fosse $h_i=1$ allora

$$\hat{y}_i = y_i$$

Quindi il modello sarebbe VINCOLATO a stimare esattamente y_i col rischio di viziare l'adattamento delle altre osservazioni

Un valore di soglia per la leva

Quant'è che il valore della leva è tanto grande da preoccupare per il fitting del modello?

In media, il valore di h_i è pari a $\bar{h} = \frac{\sum_{i=1}^{n} h_i}{n} = \frac{m+1}{n}$

Congettura

sarà considerato "eccessiva" una leva SUPERIORE al doppio della media

h_i è da considerarsi eccessivo se $h_i \geq 2\left(\frac{m+1}{n}\right)$

Purché n>2(m+1)

Le indicazioni ottenute con la leva prescindono dai valori osservati sulla dipendente, ma quantificano la "forza" che eserciterà la osservata y_i sulla stimata $\hat{y}_i$

Maggiore è la leva h_i, maggiore sarà l'influenza del punto i-esimo sulla regressione

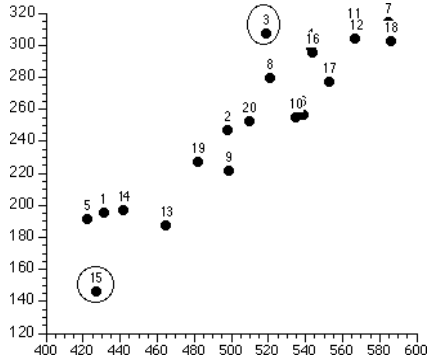
Esempio

Ecco alcuni dati regionali: due regressori e la leva. Il valore di soglia è

Regione	X_{i1}	X_{i2}	h_i
1	19.5	43.1	0.201
2	24.7	49.8	0.059
3	30.7	51.9	0.372
4	29.8	54.3	0.111
5	19.1	42.2	0.248
6	25.6	53.9	0.129
7	31.4	58.5	0.156
8	27.9	52.1	0.096
9	22.1	49.9	0.115
10	25.5	53.5	0.110
11	31.1	56.6	0.120
12	30.4	56.7	0.109
13	18.7	46.5	0.178
14	19.7	44.2	0.148
15	14.6	42.7	0.333
16	29.5	54.4	0.059
17	27.7	55.3	0.106
18	30.2	58.6	0.197
19	22.7	48.2	0.067
20	25.2	51.0	0.050

$$2\left(\frac{m+1}{n}\right) \Rightarrow 2\left(\frac{2+1}{20}\right) = 0.3$$

che evidenzia come anomale le rilevazioni "3" e "15".



Entrambe le osservazioni hanno leva molto alta rispetto alla terza leva

Residui SD

Per la diagnostica dei valori anomali sono molto informativi gli errori ottenuti dopo aver cancellato l'i-esima osservazione (deleted).

In questo modo il valore stimato non può essere influenzato da forzature verso il valore osservato y_i in quanto la i-esima osservazione è esclusa.

Il calcolo dei residui SD (*Studentized Deleted*) può essere effettuato con le quantità già ottenute dal classico modello di regressione

$$d_i^* = \hat{e}_i \sqrt{\frac{n-m-1}{SSE(1-h_{ii}) - \hat{e}_i^2}}$$

Maggiore è d_i più influente è l'osservazione per determinare y_i

Tali valori andrebbero confrontati con i quantili della t-Student con n-(m+1) gradi di libertà

In linea di massima se $|d_i^*| > 1.6$

si può ritenere che l'effetto della i-esima osservazione sia eccessivo ovvero che sia un valore anomalo

Esempio

Regione	h_i	e_i	d^*_i
1	0.201	-1.68	-0.75
2	0.059	3.64	1.58
3	0.372	-3.17	-1.70
4	0.111	-3.16	-1.39
5	0.248	0.00	0.00
6	0.129	-0.36	-0.15
7	0.156	0.72	0.31
8	0.096	4.02	1.82
9	0.115	2.66	1.15
10	0.110	-2.48	-1.07
11	0.120	0.34	0.14
12	0.109	2.23	0.95
13	0.178	-3.95	-1.88
14	0.148	3.45	1.57
15	0.333	0.57	0.27
16	0.059	0.64	0.26
17	0.106	-0.85	-0.35
18	0.197	-0.78	-0.34
19	0.067	-2.86	-1.21
20	0.050	1.04	0.42

Se il residuo SD è grande, il dato corrispondente potrebbe essere anomalo.

Lo studio dei residui SD evidenzia le osservazioni 3, 8, 13 come "anomale".

In realtà anche i residui semplici avrebbero dato la stessa indicazione, ma segnalando anche come remote altre osservazioni che invece risultano normali.

Da notare che solo per l'osservazione 3 coincidono le indicazioni della leva e degli SD

N.B. più grande è l'ampiezza del campione, maggiore sarà il numero di osservazioni che potrebbe apparire anomalo (senza esserlo).

Diagnostica sui parametri

E' possibile misurare l'effetto sui β stimati della potenziale esclusione della osservazione i-esima senza ripetere i calcoli,

$$\hat{\beta}_{new} = \hat{\beta}_{old} - \frac{\hat{e}_i W_{old}^{-1} \mathbf{x}_i^t}{1 - h_{ii}}$$

Una sintesi di queste variazioni è la DISTANZA DI COOK

$$c_i = \frac{\left(\hat{\beta}_{old} - \hat{\beta}_{new}\right)^t W_{old}^{-1} \left(\hat{\beta}_{old} - \hat{\beta}_{new}\right)}{m\hat{\sigma}}$$

che mostra l'equivalenza tra variazione nei parametri e variazione nei valori stimati dovuta alla cancellazione della i-esima osservazione

Il calcolo della formula è basato su quantità già usate per le altre diagnostiche

$$c_i = \frac{\hat{e}_i^2}{m\hat{\sigma}} \left[\frac{h_{ii}}{(1 - h_{ii})} \right]^2$$

Maggiore è il residuo $\hat{e}_i$ oppure più grande è la leva h_i più grande sarà la distanza.

E' per questo che la distanza di Cook si affianca bene alle altre misure

Ancora sulla distanza di Cook

E' difficile stabilire un valore di soglia per le c_i. Ci si può però basare sul valore di equilibrio della leva

$$h_i = \left(\frac{m+1}{2} \right)$$

nonché su di un valore standard per

$$\left| \frac{e_i}{m \hat{\sigma}} \right| \cong 2$$

Molto empiricamente, consideriamo elevata la distanza di Cook se

$$c_i > \frac{4}{n - (m+1)}$$

Esempio

Regione	h_i	e_i	c_i
1	0.201	-1.68	0.046
2	0.059	3.64	0.045
3	0.372	-3.17	0.488
4	0.111	-3.16	0.072
5	0.248	0.00	0.000
6	0.129	-0.36	0.001
7	0.156	0.72	0.006
8	0.096	4.02	0.098
9	0.115	2.66	0.054
10	0.110	-2.48	0.044
11	0.120	0.34	0.001
12	0.109	2.23	0.035
13	0.178	-3.95	0.212
14	0.148	3.45	0.125
15	0.333	0.57	0.013
16	0.059	0.64	0.001
17	0.106	-0.85	0.005
18	0.197	-0.78	0.010
19	0.067	-2.86	0.032
20	0.050	1.04	0.003

La distanza di Cook conferma come dati anomali quelli relativi alla 3^osservazione.

La 13^ è sospetta perché la sua c_i è vicina al valore di soglia: 0.24

Ma si tratta di reali anomalie?

La differenza nella stima dell'i-esimo valore della dipendente è

$$\hat{y}_{io} - \hat{y}_{in} = \frac{h_i \hat{e}_i}{1 - h_i}$$

$$\text{in particolare } \hat{y}_{3o} - \hat{y}_{3n} = \frac{h_3 e_3}{1 - h_3} = \frac{0.372 * (-3.17)}{1 - 0.372} = 1.88$$

che rispetto al valore osservato: 30.7 costituisce appena il 6.3%. Nonostante le indicazioni, la 3^non è un valore anomalo

Che fare in caso di anomalia?

L'osservazione A=(y_i, x_{i1},x_{i2},…,X_{im}) è giudicata anomala se sembra NON seguire la struttura del modello laddove la stragrande maggioranza degli altri dati vi si adatta bene

Se A è considerato anomalo si può…

- Escluderlo dal data set con guadagno sul fitting del modello. Attenzione! Per alcuni fenomeni non è serio eliminare dei dati (pensate ad esempio alle osservazioni sulle massime dei fiumi, delle piogge, delle eruzioni vulcaniche, etc.
- Farlo intervenire con un peso ridotto in modo da attenuarne l'impatto. Attenzione! Si aggiunge un problema: la scelta dei pesi.
- Utilizzare un criterio alternativo ai minimi quadrati che sia meno sensibile ai valori remoti. Ad esempio i minimi assoluti.

Dieto le diagnostiche

In precedenza abbiamo visto due tipi di diagnostiche sulla matrice dei dati

DI RIGA:

Segnalano la presenza di anomalie che, magari in pochi, possono rendere incoerente il modello di regressione con l'intero data set.

DI COLONNA:

Il segno ed il valor p dei parametri indicano se i dati sul regressore sono sufficientemente significativi per spiegare i valori della risposta

Le misure studiante dipendono tutte dalla matrice dei prodotti incrociati

$$X^t X$$

La collinearità

Ricordiamo che le colonne di X sono LID (linearmente dipendenti) se

$$\sum_{j=1}^{m} \lambda_j x_j = 0 \qquad \text{per almeno un } \lambda_j \neq 0$$

Se la relazione è esatta la matrice inversa di W=(X'X) non esiste e W è detta SINGOLARE

In generale ci sono sempre relazioni tra i regressori che comportano un certo grado di dipendenza lineare.

Questo fenomeno è detto COLLINEARITA' (o MULTICOLLINEARITA').

La collinearità è un fenomeno di gruppo, che riguarda almeno due regressori e che può colpire, in vario grado, gruppi diversi di regressori

Le colline non c'entrano

Co-lavoratori=Collaboratori
Co-lineari=collineari

Considerazioni sui ranghi

Il rango della matrice di varianze-covarianze non può essere più grande di quello delle matrici componenti:

$$\text{ran}(W) \leq \text{Min}\{\text{ran}(CX), \text{ran}(X^tC)\} = \text{ran}(CX)$$

ma anche CX è un prodotto per cui

$$\text{ran}(W) \leq \text{Min}\{\text{ran}(C), \text{ran}(X)\}$$

La matrice di centramento C è idempotente e per tali matrici si ha

$$\text{ran}(C) = \text{Tr}(C)$$

Poiché

$$c_{ii} = 1 - \frac{1}{n} \Rightarrow \text{Tr}(C) = \sum_{i=1}^{n} c_{ii} = \sum_{i=1}^{n}\left(1 - \frac{1}{n}\right) = n * \left(1 - \frac{1}{n}\right) = n - 1$$

Quindi, dato che n è molto più grande di m+1, il rango di W sarà sempre determinato dal rango della matrice dei dati X cioé dal numero di regressori

Correlazione e collinearità

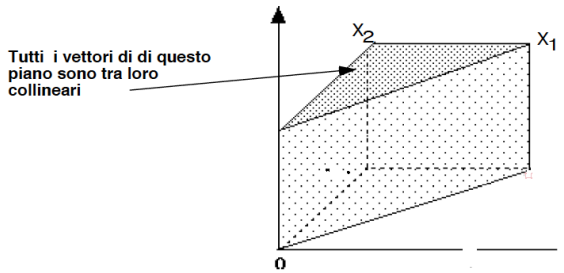
La perfetta correlazione tra due regressori genera singolarità. Poiché

$$|r_{ij}| = 1 \Rightarrow x_i = a + bx_j \quad \text{ovvero} \quad x_j = c + dx_i$$

abbiamo

$$\lambda_1 x_i + \lambda_2 x_j = 0 \quad \text{con} \quad \lambda_1 \text{ o } \lambda_2 \neq 0$$

per cui una colonna è LID



Esempio

L'analisi della matrice di correlazione può segnalare le singolarità o le quasi singolarità (correlazioni superiori a 0.9).

Correlation Analysis

Pearson Corr Coeff /Prob>|r| under $H_0: \rho = 0$

	Y	X1	X2
Y	1.00000	0.90932	0.93117
	0.0	0.0120	0.0069
X1	0.90932	1.00000	0.74118
	0.0120	0.0	0.0918
X2	0.93117	0.74118	1.00000
	0.0069	0.0918	0.0

Diagonale unitaria

r_{Y1} r_{Y2} r_{12}

Se una delle entrate è superiore a 0.98 le stime dei parametri negli OLS sono da considerarsui poco affidabili

Correlazione e collinearità/2

L'assenza di correlazioni elevate non esclude la collinearità. Essa nasce da

$$\sum_{k \in K} \lambda_k x_k \cong 0 \quad \text{dove } K = \{\text{insieme di interi tra 1 e m}\}$$

cioè una relazioe di uguaglianza approssimata

può quindi succedere che r_{12}, r_{13}, r_{23} **siano piccoli, ma la regressione**

$$x_1 = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3$$

possa dare un R^2 **pari ad uno**

Correlazione e collinearità/3

La relazione di dipendenza lineare si può scrivere come

$$X_i \cong \sum_{k \neq i} \lambda_k X_k$$

dove, il simbolo ≅ indica "approssimativamente uguale", dato che la matrice dei dati "X" ha rango pieno e l'uguaglianza è impossibile

La collinearità è un fenomeno di gradualità e non di esistenza. Più stretta è la relazione ≅ maggiore sarà la collinearità.

Poiché i regressori risentono quasi sempre di effetti esterni al modello, la stima OLS sconta comunque un certo grado di multicollinearità.

La speranza è che non sia a livelli dannosi. Occorre perciò accertare tale livello per poi intervenire in modo da contenerne gli effetti negativi

Natura del problema

Consideriamo i seguenti dati

Dato	y	x_1	x_2
1	23	2	6
2	83	8	9
3	63	6	8
4	103	10	10

Due persone, separatamente, stimano modello

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

Stima di A: $y_i = -87 + x_{i1} + 18x_{i2}$

Stima di B: $y_i = -7 + 9x_{i1} + 2x_{i2}$

Entrambi i modelli si adattano perfettamente alle osservate.

Come è possibile?

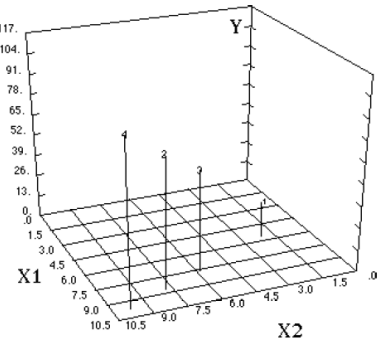
Dato	y_i	Stima-A	Stima-B
1	23	23	23
2	83	83	83
3	63	63	63
4	103	103	103

Natura del problema/2

In realtà ci sono infiniti modelli, tutti diversi e tutti perfetti.

La causa è che tra i regressori c'è una relazione lineare esatta

$x_{i2} = 5 + 0.5x_{i1}$ (è ovvio che $r_{12}=1$)



i punti si allineano ed una delle dimensioni è superflua.

D'altra parte gli scarti tra teoriche ed osservate sono tutti nulli.

Conclusioni provvisorie:

La collinearità tra i regressori è compatibile con elevati ;

A causa di essa esistono modelli diversi che danno un buon adattamento;

I parametri non possono essere interpretati come effetto di un regressore TENUTI COSTANTI GLI ALTRI perché variando un regressore, varia anche l'altro.

Cause della collinearità

La collinearità è una sorta di quasi-singolarità della matrice dei regressori.

Quali ne sono le cause?

- **Le tecniche di rilevazione dei dati: ad esempio l'intervallo limitato di alcuni regressori oppure presenza di errori di misurazione simili su regressori diversi.**
- **Correlazione spuria o latente: i regressori pur essendo, in principio, poco legati, risentono di fenomeni esterni che agiscono su entrambi.**
- **Non coerenza dei dati di un regressore con la specificazione del modello (ad esempio quando si usa una polinomiale di grado più elevato del necessario).**
- **Il modello è applicato ad un numero ridotto di casi.**

Effetti della collinearità

La collinearità danneggia la stima dei parametri e la loro precisione.

Ad esempio, nel modello con due regressori si ha

$$\beta_1^* = \frac{r_{1y} - r_{12}r_{2y}}{1 - r_{12}^2}; \quad s(\beta_1^*) = \frac{\hat{\sigma}}{\sqrt{1 - r_{12}^2}}$$

$$\beta_2^* = \frac{r_{2y} - r_{12}r_{1y}}{1 - r_{12}^2}; \quad s(\beta_2^*) = \frac{\hat{\sigma}}{\sqrt{1 - r_{12}^2}}$$

al tendere di $|r_{12}| \to 1$ i valori dei parametri e la loro Dev.Std. diventano infinitamente grandi per cui la stima è poco attendibile.

Da notare che, essendo

$$t_1 = \frac{r_{1y} - r_{12}r_{2y}}{\hat{\sigma}\sqrt{1 - r_{12}^2}}; \quad t_2 = \frac{r_{2y} - r_{12}r_{1y}}{\hat{\sigma}\sqrt{1 - r_{12}^2}}$$

la significatività delle stime tenderebbe a zero, anche se i regressori fossero in realtà utili per spiegare la dipendente

Effetti della Collinearità/2

La collinearità si riferisce alla matrice dei regressori X ovvero alle variabili indipendenti ed a come queste compaiono nel modello.

Non è un problema STATISTICO, ma NUMERICO che ha però molti risvolti statistici.

Per illustrare tale situazione studieremo due casi si regressione multipla

- **Un modello con regressori ortogonali**
- **Un modello con regressori fortemente correlati**

Regressori non correlati

La produttività di un gruppo di operai è collegata all'ampiezza del gruppo ed all'entità del bonus di produzione

Ciclo i	Numero X_{i1}	Bonus X_{i2}	Produt y_i
1	4	2	42
2	4	2	39
3	4	3	48
4	4	3	51
5	6	2	49
6	6	2	53
7	6	3	61
8	6	3	60

$$R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad r_{1y} = 0.7419 \quad r_{2y} = 0.6384$$

I due regressori sono ortogonali e la matice di correlazione coincide col la matrice identità

$$\text{due variabili esplicative}: \quad y_i = 0.375 + 5.375x_{i1} + 9.250x_{i2}$$
$$\text{variabile esplicativa } 1: \quad y_i = 23.5 + 5.375x_{i1}$$
$$\text{variabile esplicativa } 2: \quad y_i = 27.5 + 9.250x_{i2}$$

Se i regressori sono incorrelati il loro effetto è lo stesso sia che siano tutti presenti che se presenti separatamente

Esempio

Studiamo le seguenti variabili

$y = Ln(\text{indice prezzi al consumo})$
$x_1 = Ln(\text{indice a prezzi al correnti})$
$x_2 = Ln(\text{PIL a prezzi costanti})$
$x_3 = x_1 - x_2 = Ln(\text{deflattore PIL})$

coinvolgiamole nei modelli

Anno	IPC	Pil-Cor	Pil-Cos	Defl.
1950	4.37	6.28	5.66	-0.62
1960	4.52	6.60	6.23	-0.37
1970	4.73	6.98	6.89	-0.09
1980	5.54	7.30	7.87	0.57
1982	5.58	7.30	8.03	0.73
1983	5.60	7.33	8.10	0.77
1984	5.54	7.43	8.24	0.81

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2}; \quad y_i = \alpha_1 x_{i1} + \alpha_2 x_{i3}$$

nessuno dei regressori ha influenza sulla dipendente (livello di inflazione). L'unica sarebbe la terza, che però nel primo è assente.

$$y = \underset{(2.997)}{0.273}x_1 + \underset{(4.972)}{0.442}x_2; \quad R^2 = 0.9994; \hat{\sigma} = 0.127$$

$$y = \underset{(92.356)}{0.715}x_1 + \underset{(4.972)}{0.442}x_3; \quad \overline{R}^2 = 0.9994; \hat{\sigma} = 0.127$$

t-Student

a) perchè aumenta il t-student della X_1?

b) Perchè cambia solo la precisione della X_1?

c) Cosa c'è di diverso nei due modelli?

Misura del condizionamento

Un'altra misura è il *condition number*

$$C(X) = \frac{\lambda_{max}\left[(X'X)\right]}{\lambda_{min}\left[(X'X)\right]}$$

Un valore prossimo ad uno indica una matrice di regressori con equivariabilità verso ogni direzione. Valori elevati indicano collinearità

L'ordine di grandezza di C(X) esprime il numero di cifre che si degradano nel calcolo della matrice inversa.

Se X è misurata con 7 cifre significative e C(X)=10'000 allora la stima dei parametri sarà accurata fino alla terza cifra decimale (7-4=3).

Valori di C(X)<1'000 sono piuttosto tranquilli. Preoccupano valori C(X)>100'000

Relazioni tra i regressori

Per individuare le quasi-singolarità dovremmo esplorare tutti i submodelli definibili tra i regressori. Il loro numero è

$$m * \frac{(m-1)^2}{2}$$

per cui se m=2 si studia solo un submodello; ma se m=6 occorre studiarne 75 e questo non sempre è possibile.

In genere ci si limita a quelli che pongono ciascun regressore in relazione con tutti gli altri badando in particolare a valore di R^2 in

$$X_i = \sum_{j \neq i} \beta_j X_j$$

Non è però necessario effettuare materialmente i calcoli. Gli R^2 parziali sono infatti ottenibili dai calcoli già effettuati per la stima ordinaria

Variance inflation factors

La Dev.Std delle stime nel modello con regressori standardardizzati è

$$s\left(\beta_i^*\right) = \hat{\sigma}\sqrt{v_{ii}^*}$$

dove v_{ii}^* è l'elemento sulla diagonale dell'inversa della matrice di correlazione

Questi elementi sono noti come VIF *(Variance Inflation Factors)* ed indicano quanto la variabilità del parametro dipenda dai legami tra tutti i regressori

Si dimostra infatti che $v_{ii}^* = \dfrac{1}{\left(1 - R_i^2\right)}$

dove R_i^2 è il coefficiente di determinazione multipla del modello in cui x_i è come dipnendente rispetto agli altri regressori.

Dato che R_i^2 varia tra zero ed uno, i VIF hanno un campo di variazione che parte da uno e va all'infinito

Significato dei VIF

Se il regressore i-esimo non è legato linearmente agli altri si avrà $R_i^2 = 0$ ed il suo VIF sarà pari ad uno.

Man mano che R_i^2 aumenta anche il VIF cresce e tende ad infinito al tendere di R_i^2 ad uno

Ricordiamo che $$Var\left(\hat{\beta}_j\right) = VIF_j \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}$$

Il secondo fattore è la varianza della stima del parametro nella regressione semplice usando come regressore solo x_i

Il VIF dell'i-esimo regressore esprime l'incremento di variabilità generatosi nel modello aggiungendo dei regressori diversi dalla x_i

Altre misure di collinearità

Il maggiore tra gli "m" VIF (uno per ogni regressore) è un indicatore di collinearità.

Empiricamente si ritiene che

$$\max_{1 \le i \le m}\left\{\mathrm{VIF}_i\right\} \ge 10$$

sia una soglia che indichi livelli pericolosi di collinearità nella X.

Un'altro indicatore, ma che aggiunge poco al precedente, è la media dei VIF

$$\overline{\mathrm{VIF}_i} = \frac{\sum_{i=1}^{m}\mathrm{VIF}_i}{m}$$

Esempio

Nella tabella ci sono dei risultati relativi alla stima dei dati regionali già visti

Regressore	Parametri β_i^*	Determin. R_i^2	Tolleranza $1 - R_i^2$	VIF $\dfrac{1}{1 - R_i^2}$
x1	4.264	0.9986	0.0014	708.84
x2	-1.562	0.9904	0.0096	104.61
x3	-2.983	0.9982	0.0012	564.34

Tutti i VIF sono in questo caso molto elevati. I dati sono poco coerenti con il modello, ovvero la specificazione di quest'ultimo non è corretta.

La tolleranza è talvolta adoperata per indicare se includere o meno il regressore (dovrebbe essere superiore a 0.01).

Da notare che $$\mathrm{Tolleranza}_i = \frac{1}{\mathrm{VIF}_i}$$

Rimedi per la collinearità

Per la regressione polinomiale

Comportano gradi elevati dei polinomi e quindi collinearità nei regressori che aumenta con il crescere del grado del polinomio.

Come rimedio si possono usare i polinomi ortogonali.

Centrare le variabili può anche essere d'aiuto.

Con dati sotto controllo

Si possono aggiungere nuove rilevazioni che possano spezzare le relazioni collineari tra regressori.

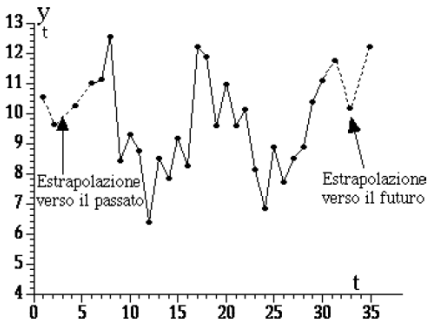
Si possono usare gli scores delle componenti principali come regressori.

Regressione e serie storiche

La capacità estrapolativa del modello è fondamentale quando i dati sono delle time series o SERIE STORICHE

Molte applicazioni considerano situazioni in cui variabile dipendente e regressori sono rilevati sequenzialmente nel tempo.

$$y_t = \sum_{i=0}^{m} \beta_j x_{ti} + u_t \qquad (x_{t0} \equiv 1)$$



L'estrapolazione consiste ora nel proiettare (all'indietro o in avanti) nel tempo il modello di regressione

Regressione e serie storiche/2

La variabile tempo può entrare nel modello di regressione in vari modi

- Regressione con componente polinomiale

$$y_t = \sum_{i=0}^{m} \beta_j x_{ti} + \sum_{i=1}^{p} \beta_{m+i} t^i + u_t$$

Funzioni di t come regressori

- Regressione con componente autoregressiva

$$y_t = \sum_{i=0}^{m} \beta_j x_{ti} + \sum_{i=1}^{p} \beta_{m+i} y_{t-i} + u_t$$

Le ritardate (lagged) della risposta come regressori

- Regressione con ritardi distribuiti

$$y_t = \sum_{i=0}^{m} \beta_j x_{ti} + \sum_{i=1}^{p} \sum_{j=1}^{k_i} \lambda_{j,i} x_{t-j,i} + u_t$$

Oltre ai regressori a destra compaiono pure le loro ritardate

Regressione e serie storiche/3

Le tre formulazioni non sono una novità: ai soliti regressori se ne aggiungono altri non derivati da variabili indipendenti

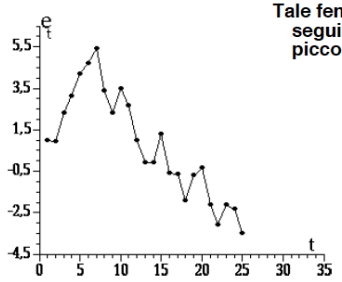
L'aspetto temporale introduce però nuovi problemi di stima che debbono essere approfonditi.

Da un lato i regressori, essendo tutti soggetti alla medesime forze evolutive, potrebbero risultare collineari.

D'altra parte, la natura dinamica di queste relazioni fa si che l'effetto degli errori e/o dei regressori non si esplichi solo nei valori correnti, ma prosegua per altre rilevazioni (autocorrelazione)

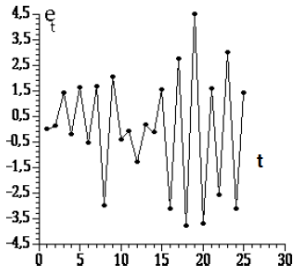
Problemi di autocorrelazione

L'ordinamento temporale causa spesso il problema della AUTOCORRELAZIONE o CORRELAZIONE SERIALE tra gli errori



Tale fenomeno si manifesta se uno scarto elevato è seguito da uno scarto più elevato ovvero un errore piccolo è seguito da un errore più piccolo.

$$e_t = y_t - \hat{y}_t$$



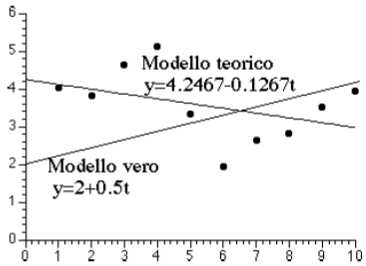
Può anche aversi un effetto di compensazione: un errore per difetto è seguito da uno per eccesso o viceversa

Esempio con simulazione

La simulazione consiste nella attivazione di un certo modello a partire da valori completamente noti e da errori casuali ε_t generati con il computer

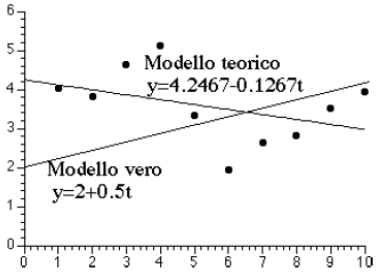
$$\begin{cases} y_t = 2 + 0.5t + u_t; t = 1,2,\ldots,10 \\ u_t = u_{t-1} + \varepsilon_t; u_0 = 1. \end{cases}$$

t	$\hat{y}_t$	ε_t	u_t	y_t
1	2.5	0.5	1.5	4.0
2	3.0	-0.7	0.8	3.8
3	3.5	0.3	1.1	4.6
4	4.0	0.0	1.1	5.1
5	4.5	-2.3	-1.2	3.3
6	5.0	-1.9	-3.1	1.9
7	5.5	0.2	-2.9	2.6
8	6.0	-0.3	-3.2	2.8
9	6.5	0.2	-3.0	3.5
10	7.0	-0.1	-3.1	3.9



L'autocorrelazione rende molto imprecisa la stima dei parametri

Effetti dell'autocorrelazione



La variabilità intorno al modello teorico è molto più bassa che intorno al modello vero.

Ciò implica un SQM più basso del dovuto con conseguente riduzione della varianza dei parametri e aumento dei t-Student

N.B.

L'autocorrelazione può portare a giudizi avventati sulla significatività dei parametri e sulla validità del modello di regressione

Molti studiosi sostengono che le previsioni basate su modelli con errori autocorrelati non sono affidabili

Accertare l'autocorrelazione

E' ovvio che l'autocorrelazione riguardi le serie storiche perché nei dati privi di ordinamento basterebbe riordinare le righe della matrice dei dati per eliminarla.

Gli effetti dell'autocorrelazione sono molto negativi ed è quindi essenziale accertarne la presenza.

In che modo affiora l'autocorrelazione?

Una verifica molto semplice, ma anche molto soggettiva è quella grafica.

Il grafico più efficace per l'analisi temporale dei residui è il

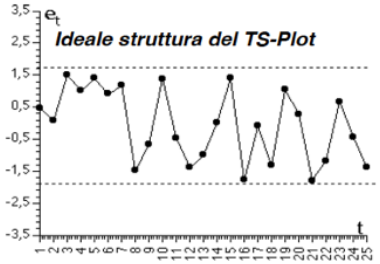
- **Time sequence plot. Ordinate: errori (ovvero errori standardizzati oppure studentized); ascisse:tempo.**

Altri grafici utili sono:

- **Ordinate: errori; ascisse: valori teorici**
- **Ordinate: errori; ascisse: regressori (uno o due)**

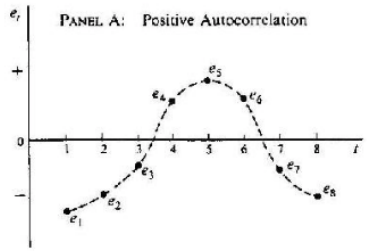
Accertamento grafico

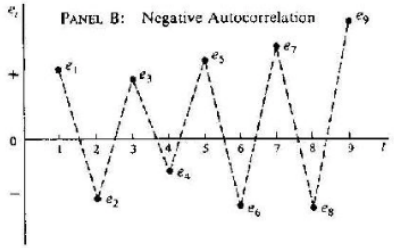
L'analisi grafica è sempre utilizzabile purché non debba essere ripetuta per una miriade di modelli.



Noi ci aspettiamo un andamento privo di struttura.

In presenza di autocorrelazione gli errori tendono a presentarsi secondo uno schema NON erratico.





Esempio

Nella tabella sono raccolti i dati trimestrali sui consumi e dello stock di moneta

Anno	Tr.	Cons.	Moneta	Anno	Tr.	Cons.	Moneta
1952	1	214.6	159.3	1954	3	238.7	173.9
	2	217.7	161.2		4	243.2	176.1
	3	219.6	162.8	1955	1	249.4	178.0
	4	227.2	164.6		2	254.3	179.1
1953	1	230.9	165.9		3	260.9	180.2
	2	233.3	167.9		4	263.3	181.2
	3	234.1	168.3	1956	1	265.6	181.6
	4	232.3	169.7		2	268.2	182.5
1954	1	233.7	170.5		3	270.4	183.3
	2	236.5	171.6		4	275.6	184.3

La versione ristretta della teoria quantitativa della moneta sostiene che

$$C_t = \beta_0 + \beta_1 M_t + u_t$$

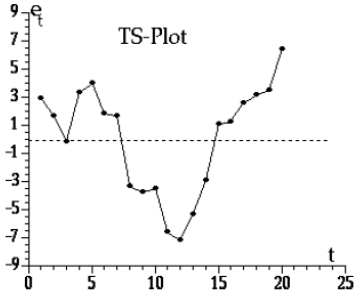
Il coefficiente β_1 è detto moltiplicatore ed è molto importante in questa teoria

Risultati della stima

$$C_t = \underset{(7.794)}{-154.719} + \underset{(20.080)}{2.3} M_t \qquad R^2 = 0.955;\ \hat{\sigma} = 3.983$$

Poichè la dimensione temporale è presente non siamo sorpresi di trovare degli errori correlatl.

I residui correlati sono compatibili con valori elevati dell'R-quadro.



Rimedio dell'autocorrelazione

Se si riscontra autocorrelazione, il primo passo è di ristrutturare il modello includendovi le variabili che assorbano gli effetti non già inseriti (se possibile).

Se il modello risponde ai nostri canoni, l'unica conclusione possibile è che ci sia autocorrelazione PURA.

Una prassi diffusa (non necessariamente da condividere) è di costruire il modello a partire da una lista di regressori possibili e di decidere quali e quanti inserirne in base ad indici numerico-statistici *(stepwise regression)*

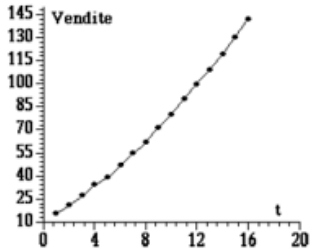
Nessuna prassi empirica può garantire la definizione dell'insieme corretto di regressori e si rischia di considerare ad errori autocorrelati un modello solo specificato male

Ristrutturazione del modello

Un compromesso è la scelta del regressore con cui riportare una certa variabile indipendente nel modello.

Un caso emblematico sono le regressioni polinomiali in cui il grado del polinomio non è vincolato ad un certo grado.

t	Vendite		
1	15.24	9	71.31
2	20.82	10	80.18
3	27.44	11	89.79
4	34.07	12	99.58
5	39.29	13	109.27
6	46.52	14	118.99
7	54.96	15	129.89
8	62.12	16	142.01

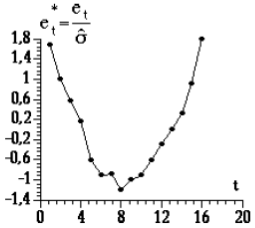


Dal grafico non è chiaro se il tempo interviene linearmente oppure con accelerazione (trend quadratico)

Ristrutturazione del modello/2

Gli OLS danno

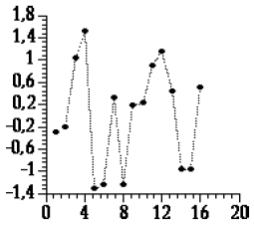
$$V_t = \underset{(0.915)}{-0.241} + \underset{(36.959)}{8.422}t;\quad R^2 = 0.989;\quad \hat{\sigma} = 4.202$$



L'autocorrelazione è evidente, ma è anche evidente l'omissione, tra i regressori, del termine quadratico

Per il modello quadratico si ha

$$V_t = \underset{(22.183)}{10.285} + \underset{(39.035)}{4.913}t + \underset{(28.674)}{0.206}t^2;\quad R^2 = 0.999;\quad \hat{\sigma} = 0.544$$



L'autocorrelazione è stata rimossa. Non del tutto come è ovvio per ogni applicazione realistica

Errori autoregressivi

L'autocorrelazione può esplicarsi in molti modi. Per il momento esaminiamo un caso semplice che però approssima situazioni più complesse

REGRESSIONE LINEARE CON RESIDUI AUTOREGRESSIVI DEL 1^ ORDINE

$$\begin{cases} y_t = \sum_{i=0}^{m} \beta_j x_{ti} + u_t \\ u_t = \rho u_{t-1} + \varepsilon_t \quad con \quad |\rho < 1| \end{cases}$$

1^ordine significa che l'errore al tempo t è legato solo a quello precedente: tempo t-1

dove gli ε_t sono incorrelati.

Questa formulazione aggiunge un parametro incognito, ρ, a quella classica

Da sottolineare che l'applicazione di questo modello richiede la disponibilità di un numero di rilevazioni non piccolo. Almeno n>14

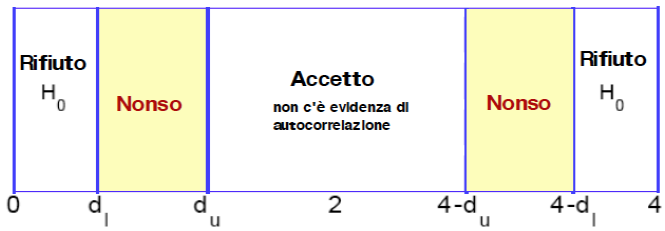
L'indice Durbin-Watson

Supponendo che viga il modello di prima, si può calcolare un indice che quantifica il grado di autocorrelazione degli errori.

$$\text{Indice Durbin – Watson:}\quad DW = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Come per tutti gli indici già studiati è difficile stabilirne la esatta validità, però esistono dei valori di soglia che permettono alcune decisioni conclusive.

I valori di soglia dipendono dal numero di rilevazioni e dal grado di confidenza, ritenuto tollerabile nelle decisioni.



Esempio

Una compagnia distributrice di bibite intende stimare la vendita annuale di aranciata in funzione della spesa annuale di pubblicità. I dati raccolti per n=20 anni sono in tabella insieme alla popolazione target

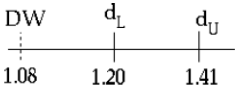
Anno	t	Vendite y_t	Spesa x_{t1}	Target x_{t2}
1960	1	3083	75	825000
1961	2	3149	78	830445
1962	3	3218	80	838750
1963	4	3239	82	842940
1964	5	3295	84	846315
1965	6	3374	88	852240
1966	7	3475	93	860760
1967	8	3569	97	865925
1968	9	3597	99	871640
1969	10	3725	104	877745
1970	11	3794	109	886520
1971	12	3959	115	894500
1972	13	4043	120	900400
1973	14	4194	127	904005
1974	15	4318	135	908525
1975	16	4493	144	912160
1976	17	4683	153	917630
1977	18	4850	161	922220
1978	19	5005	170	925910
1979	20	5236	182	929610

Modello semplice

$$\hat{y}_t = 1608.508 + 20.091 x_{t1}$$

$$DW = \frac{\sum_{t=2}^{20} (e_t - e_{t-1})^2}{\sum_{t=1}^{20} e_t^2} = \frac{8195.2065}{7587.9154} = 1.08$$

Per α=5%, n= 20 troviamo:

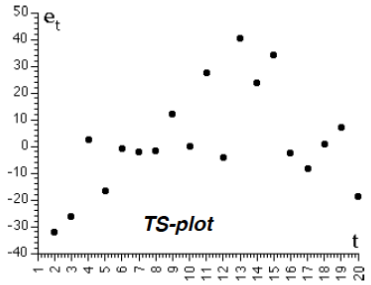


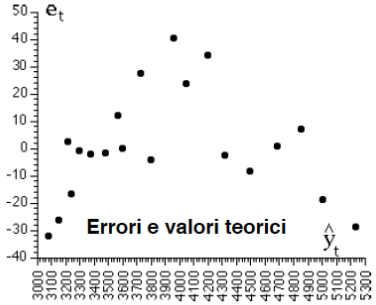
Poichè DW<DL dobbiamo ritenere che ci sia autocorrelazione

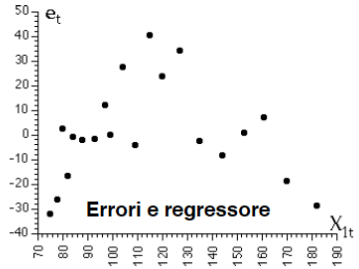
Analisi grafica

Se i residui tendono a collocarsi tutti in modo erratico all'interno di una banda ristretta NON c'è autocorrelazione.

Nei grafici c'è invece una struttura che conferma la presenza di effetti cumulativi negli errori







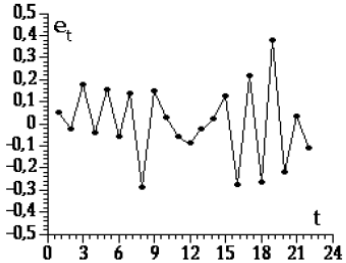
Esempio

Si ritiene che ci sia forte parallelismo tra il tasso di disoccupazione totale D e quello nei settori manifatturieri M

Il modello di regressione produce i risultati:

$$D_t = \underset{(7.181)}{-3.809} + \underset{(36.547)}{1.98M_t} \qquad DW=3.45;\ DW^{-}=0.55$$

$$\bar{R}^2 = 0.985; \quad \hat{\sigma} = 0.176; \quad \hat{\rho} = -0.753$$



Anno	t	M_t	D_t
1967	1	9.75	15.55
1968	2	9.85	15.67
1969	3	9.60	15.37
1970	4	9.25	14.46
1971	5	8.90	13.97
1972	6	8.85	13.66
1973	7	8.80	13.75
1974	8	8.75	13.23
1975	9	9.45	15.05
1976	10	9.95	15.92
1977	11	9.80	15.54
1978	12	9.45	14.81
1979	13	9.80	15.57
1980	14	9.45	14.92
1981	15	9.85	15.82
1982	16	9.45	14.63
1983	17	9.90	16.01
1984	18	11.25	18.20
1985	19	11.20	18.75
1986	20	10.85	17.45
1987	21	10.15	16.32
1988	22	10.55	16.97

L'andamento degli errori evidenzia una autocorrelazione negativa.
Infatti DW-<DL=1.24

Autocorrelazione negativa

Nelle serie economiche, per varie ragioni (non tutte chiare) è più frequente una autocorrelazione positiva.

Può però accadere che si riscontri autocorrelazione negativa:

Errori adiacenti tendono ad avere segno opposto ed il time sequence plot assume la tipica forma di oscillogramma

Se si sospetta che ci sia correlazione seriale negativa, l'indice appropriato è

$$DW^{-} = 4 - DW$$