

Minimi Quadrati Generalizzati

Nel definire lo stimatore BLUE dei parametri del modello di regressione si è risolto un problema più ampio

Si è ipotizzato che

$$E(uu') = V = \sigma^2 \Omega$$

In cui solo σ^2 fosse incognito.

Le stime dei minimi quadrati erano

$$\tilde{\beta} = [X'V^{-1}X]^{-1}X'V^{-1}y; \quad \text{Var}(\tilde{\beta}) = [X'V^{-1}X]^{-1}$$

Se $\Omega=I$ otteniamo gli OLS.

Sono note altre forme di Ω che ricadono in questa formulazione che è detta dei minimi quadrati generalizzati se $\Omega \neq I$

Errori eteroschedastici

L'ipotesi che la varianza degli errori rimanga la stessa in ogni osservazione della dipendente e per ogni combinazione di regressori non è sempre realistica.

ESEMPIO: nella regressione

$$\text{Spesa vacanze} = \beta_0 + \beta_1 \text{Reddito}_{\text{familiare}} + u$$

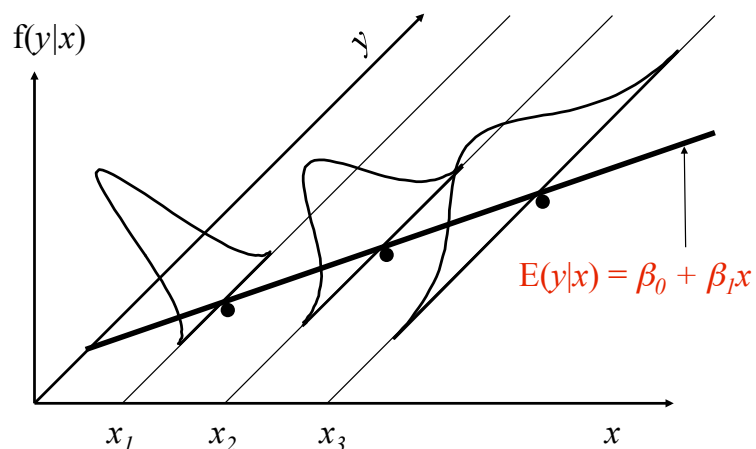
Nuclei familiari a basso reddito hanno poco da scegliere.

All'aumentare del reddito aumentano le opportunità e quindi la variabilità

Non fare vacanza per i primi è una costrizione, per i secondi è una scelta



Rappresentazione della eteroschedasticità

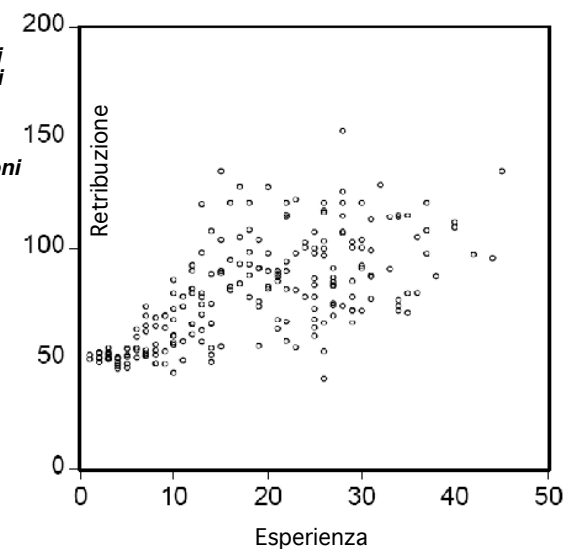


La eteroschedasticità ovvero la diversificazione della varianza degli errori ha conseguenze importanti sulla stima del modello di regressione

Un esempio di eteroschedasticità

E' vero che le retribuzioni medie aumentano con gli anni di esperienza.

E' poco ragionevole attendersi che le variazioni tra i soggetti siano le stesse ad ogni livello di esperienza



Cause della eteroschedasticità



Le unità campionarie provengono da sottopopolazioni diverse ovvero da aree diverse.

Ad esempio, nelle imprese di minori dimensioni la variabilità delle vendite è più contenuta che in quelle maggiori che tendono ad essere più volatili.



Nel campione sono più presenti unità con maggiore o minore variabilità rispetto alle altre

Ad esempio, le unità rilevate al centro-città potrebbero essere poche se quelle in periferia sono più numerose per includere nel campione tutte le periferie



Uno o più regressori importanti sono stati omessi ovvero non compaiono nella forma più adatta.

Talvolta inserire $\text{Log}(x)$ oppure $\sqrt{x}$ invece di x può eliminare la eteroschedasticità.

La mancanza di un regressore importante non è un problema di eteroschedasticità, ma di corretta specificazione del modello

Errori eteroschedastici/2

Per migliorare l'aderenza ai fenomeni reali ipotizziamo errori incorrelati, ma eteroschedastici

$$E(uu^t) = \sigma^2 \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_n \end{bmatrix} = \sigma^2 \Omega$$

In cui la variabilità non è costante, ma cambia anche se non necessariamente le varianze sono tutte diverse.

E' possibile infatti immaginare gruppi di osservazioni con varianza omogenea, che è però diversa nei vari gruppi

In questo caso la matrice Ω è diagonale formata da elementi positivi, non necessariamente tutti distinti.

Dove si incontra



E' comune nei data set in cui si riscontra una forte escursione nei valori: sono presenti valori molto piccoli e molto grandi



E' anche presente nei dati cross-section in cui si aggregano dati con diversa natura e quindi potenzialmente diversa variabilità



Nelle serie storiche con un forte trend monotono si riscontra spesso una diversa struttura di variabilità tra la fase iniziale e quella finale del fenomeno



Non è infrequente trovarla nelle indagini campionarie in cui la fase di acquisizione si modifica in aspetti rilevanti.

Stima dei parametri

Se V è una matrice simmetrica e definita positiva possiamo scrivere

$$V = (Q\Lambda^{0.5})(Q\Lambda^{0.5})^t$$

Dove

Q è la matrice formata con gli autovettori di norma unitaria associati agli autovalori di V ;
 $\Lambda^{0.5}$ è la matrice diagonale i cui elementi non nulli sono le radici degli autovalori di V .

Se la matrice V è diagonale gli autovalori coincidono con i suoi elementi non nulli e gli autovettori sono i vettori identità (cioè $Q=I$).

Quindi

$$\tilde{\beta} = [X^t \Lambda^{-0.5} \Lambda^{-0.5} X]^{-1} X^t \Lambda^{-0.5} \Lambda^{-0.5} y \Rightarrow \tilde{\beta} = [W^t W]^{-1} W^t z$$

Dove ora $W=X \Lambda^{-0.5}$ e $z=y \Lambda^{-0.5}$.

Stima dei parametri/2

Lo stimatore ai minimi quadrati in presenza di errori eteroschedastici è lo stimatore OLS, ma applicato a variabili pesate

$$z = \begin{bmatrix} \frac{y_1}{\sqrt{\lambda_1}} \\ \frac{y_2}{\sqrt{\lambda_2}} \\ \dots \\ \frac{y_n}{\sqrt{\lambda_n}} \end{bmatrix}; \quad W = \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} & \frac{x_{11}}{\sqrt{\lambda_1}} & \frac{x_{12}}{\sqrt{\lambda_1}} & \dots & \frac{x_{1m}}{\sqrt{\lambda_1}} \\ \frac{1}{\sqrt{\lambda_2}} & \frac{x_{21}}{\sqrt{\lambda_2}} & \frac{x_{22}}{\sqrt{\lambda_2}} & \dots & \frac{x_{2m}}{\sqrt{\lambda_2}} \\ \dots & \dots & \dots & \dots & \dots \\ \frac{1}{\sqrt{\lambda_n}} & \frac{x_{n1}}{\sqrt{\lambda_n}} & \frac{x_{n2}}{\sqrt{\lambda_n}} & \dots & \frac{x_{nm}}{\sqrt{\lambda_n}} \end{bmatrix}$$

Ovvero un modello di regressione lineare multipla così formulato

$$\frac{y_i}{\sqrt{\lambda_i}} = \beta_0 \left(\frac{1}{\sqrt{\lambda_i}} \right) + \beta_1 \left(\frac{x_{i1}}{\sqrt{\lambda_i}} \right) + \beta_2 \left(\frac{x_{i2}}{\sqrt{\lambda_i}} \right) + \dots + \beta_m \left(\frac{x_{im}}{\sqrt{\lambda_i}} \right) + \frac{u_i}{\sqrt{\lambda_i}}$$

L'i-esimo dato ($y_i, x_{i1}, x_{i2}, \dots, x_{im}$) è pesato con il reciproco della radice dell'autovalore λ_i : i casi con maggiore variabilità pesano meno laddove sono più rilevanti le osservazioni con minore dispersione

Conseguenze sugli stimatori/2

Che succede se si stimano i parametri come se $\Omega = I$, ma in realtà $\Omega \neq I$?

Gli stimatori rimangono non distorti.

Ovviamente non sono più gli stimatori BLUE del modello, ma anzi hanno una variabilità che può essere molto superiore o molto inferiore.

Le diagnostiche basate sui t-Student, sull'F-Fisher e sull' R^2 sono più alte (o più basse) di quanto non dovrebbero

C'è il rischio di giudicare buono o cattivo un modello che lo è solo in virtù di una falsa premessa sulla omoschedasticità

Conseguenze sugli stimatori

Nulla cambia rispetto alla caratteristica di stimatore BLUE (se Ω è nota)

La bontà di adattamento in termini di R^2 non risente del cambiamento

$$R^2 = \frac{Dev\ Spieg.}{Dev\ Tot.} = \frac{y^T C H y}{y^T C y} = \frac{SSR}{SST}$$

L'impatto della divisione per i pesi λ_i si annulla.

La varianza degli stimatori dipende dai pesi λ_i

$$Var(\tilde{\beta}) = [X^T \Lambda^{-1} X]^{-1}$$

Operatività della tecnica

La eteroschedasticità è un fenomeno che riguarda la potenzialità di errore per le singole osservazioni che si modifica al modificarsi delle osservazioni.

Non avrebbe alcuna conseguenza se la matrice di varianze-covarianze fosse nota ovvero incognita nel solo parametro σ^2 .

In pratica si opera con due livelli di ignoranza



Non sappiamo se gli errori sono omo- oppure etero-schedastici

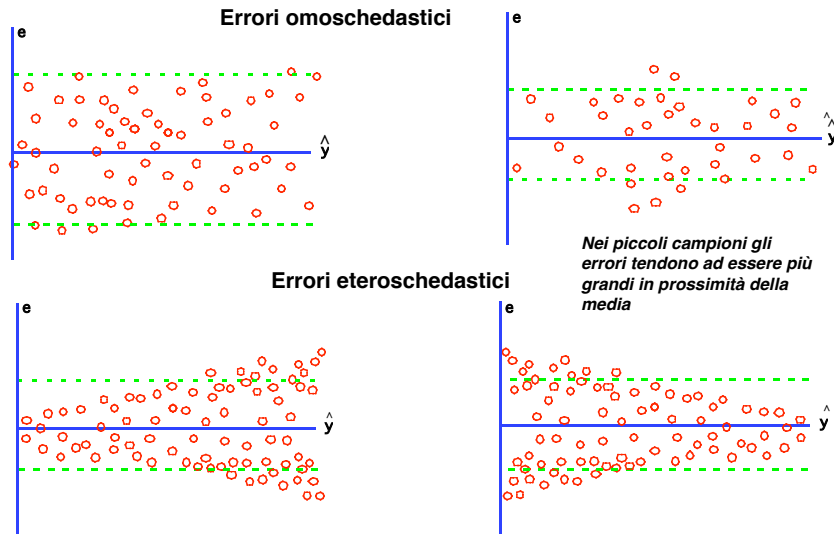


Nel secondo caso non sappiamo quanti e quali siano i pesi λ_i che debbono quindi essere stimati.

I parametri da stimare sono $m+n$ (m in β ed n in Ω) quindi più parametri che dati. Come se ne esce?

Accertamento grafico

Un grafico molto utile è formato ponendo in ascissa i valori stimati della dipendente ed in ordinata i residui stimati (oppure uno dei regressori).



Test di Breusch-Pagan

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2 \\ H_1 : \sigma_i^2 = \sigma^2 \exp\{x_i' \theta\} \end{cases}$$

In questo caso le w_i sono gli stessi regressori e la funzione $h()$ è la esponenziale.

ALGORITMO

1) Calcolare gli OLS

$$\hat{\beta} = (X'X)^{-1} X'y \Rightarrow e_i^2 = (y_i - x_i' \hat{\beta})^2, i = 1, 2, \dots, n$$

2) Calcolare l' R^2 della regressione

$$e_i^2 = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \dots + \theta_m x_{im} + v_i$$

3) Calcolare la statistica $F(m, n-m-1)$ $F_h = \frac{R^2}{1-R^2} \left(\frac{n-m-1}{m} \right)$

4) Determinare il p-value e decidere sulla presenza di eteroschedasticità

Test di eteroschedasticità

I grafici possono dare un'idea di massima. E' opportuno integrarli con delle indicazioni più formali.

Non dobbiamo trascurare che c'è un possibile problema di identificazione: più parametri che dati.

Dobbiamo superare questo problema

In varie occasioni si ipotizza una ETEROSCHEDASTICITA' MOLTIPLICATIVA

$$u_i = c_i \varepsilon_i \quad \text{con} \quad E(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2$$

dove

$$c_i^2 = h(w_i' \theta) > 0 \quad \text{con} \quad h(0) = 1, \quad h'(0) \neq 0$$

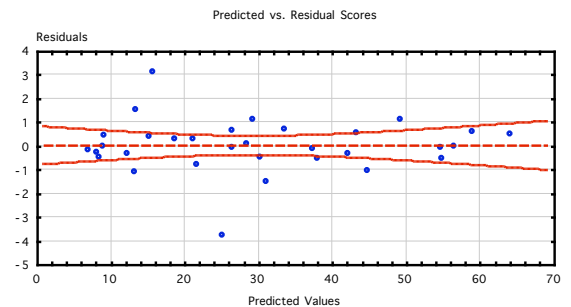
Le w_i sono delle funzioni dei regressori.

$h()$ non è specificata.

C_i è un fattore di proporzionalità. Maggiore è c_i maggiore è la variabilità del dato i -esimo

y	X1	X2	X3
21.37	4.8	3.8	7.3
26.33	6.6	4.1	9.7
27.03	6.6	4.2	9.6
6.80	0.8	0.3	2.5
37.48	9.3	6.8	14.3
54.54	14.8	10.3	19.8
15.56	3.1	1.6	5.9
56.47	15.3	10.7	20.5
7.94	1.4	0.6	2.9
9.49	1.1	0.4	3.8
20.86	5.3	3.1	7.9
41.71	11.4	7.4	15.2
64.53	16.8	12.7	23.5
14.89	3.4	1.8	4.1
59.61	16	11.3	21.4
29.71	7.7	4.9	11.1
28.58	7.2	4.5	10.5
37.25	9.5	7.2	13.3
30.25	7.4	4.7	10.7
7.73	1.9	0.3	2.4
18.84	3.2	1.7	6.2
43.78	11.5	7.7	15.8
11.87	2.2	1.4	4.5
34.21	8.4	5.4	12.7
29.50	7.9	5.1	11.4
43.65	11.2	8.3	16.7
21.36	6.3	3.8	9.2
50.30	13.3	9.2	17.7
12.17	2.9	1.5	4.7
54.30	14.1	10.6	20.3
18.94	4.4	2.5	6.8
8.86	0.1	1.4	3.7

Esempio_25



$$R^2 = 0.9958, \quad \hat{\beta}_0 = 2.5742, \quad \hat{\beta}_1 = 0.0850, \quad \hat{\beta}_2 = 1.0371, \quad \hat{\beta}_3 = 1.2781$$

$$R_h^2 = 0.08 \Rightarrow F_h = \left(\frac{0.08}{0.92} \right) \left(\frac{28}{4} \right) = 0.61 \quad (p\text{-value} = 0.68)$$

Se si afferma che c'è eteroschedasticità (si rifiuta H_0) si sbaglia circa sette volte su dieci. Non è consigliabile rifiutare.

Test di White

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2 \\ H_1 : \sigma_i^2 \neq \sigma_j^2 \text{ per } i \neq j \end{cases}$$

ALGORITMO

1) Calcolare gli OLS

$$\hat{\beta} = (X^T X)^{-1} X^T y \Rightarrow e_i^2 = (y_i - x_i^T \hat{\beta})^2, i = 1, 2, \dots, n$$

2) Calcolare l'R² della regressione

$$e_i^2 = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \dots + \theta_m x_{im} + \theta_{m+1} x_{i1}^2 + \theta_{m+2} x_{i2}^2 + \dots + \theta_{2m} x_{im}^2 + \sum_{r=1}^{m-1} \sum_{s=r+1}^m \theta_{rs} x_{ir} x_{is} + v_i$$

Tutte le coppie di regressori

3) Calcolare la statistica F(m,n-m-1)

$$F_h = \frac{R^2}{1-R^2} \left(\frac{n-m-1}{m} \right)$$

4) Determinare il p-value e decidere sulla presenza di eteroschedasticità

Test di White/2

Se i regressori sono molti e le osservazioni poche il test di White non è indicato.

Se m=4 occorre regredire gli errori contro ben 14 variabili indipendenti

Esiste una versione “ridotta” del test che lega la variabilità dei residui ai valori stimati ed al loro quadrato (che in fondo sono funzioni lineari di tutti i regressori).

$$e_i^2 = \theta_0 + \theta_1 \hat{y}_i + \theta_2 \hat{y}_i^2 + v_i$$

A questo punto si calcola la statistica F(m,n-m-1)

$$F_h = \frac{R^2}{1-R^2} \left(\frac{n-3}{2} \right)$$

Si determina quindi il p-value del test e si decide sulla presenza di eteroschedasticità

Test di Park

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2 \\ H_1 : \sigma_i^2 \neq \sigma_j^2 \text{ per } i \neq j \end{cases} \quad \begin{aligned} &\text{Qui si ritiene che } \text{var}(u_i) = \sigma^2(|z_i| + 1)^\beta \\ &\text{Dove } z_i \text{ può essere uno dei regressori} \end{aligned}$$

ALGORITMO

1) Calcolare gli OLS

$$\hat{\beta} = (X^T X)^{-1} X^T y \Rightarrow e_i^2 = (y_i - x_i^T \hat{\beta})^2, i = 1, 2, \dots, n$$

2) Stimare i parametri della regressione

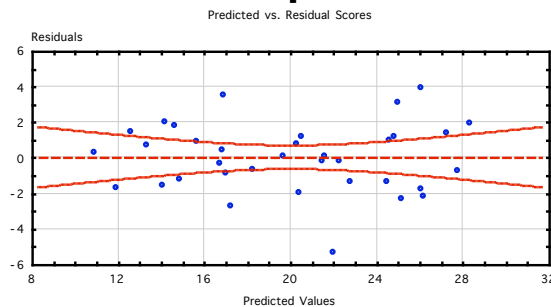
$$\text{Ln}(e_i^2) = \theta_0 + \theta_1 \text{Ln}(|z_i| + 1)$$

Se si è sicuri che $z_i \neq 0$ si può omettere il +1 dall'argomento del logaritmo

3) Valutare il p-value della t-studenti di θ_1 e decidere sulla presenza di eteroschedasticità

y	X1	X2	X3	x4
21.69	4	6	-8	6
22.82	9	12	-8	0
16.41	9	9	-12	-1
28.63	9	13	-7	0
27.06	10	8	-5	2
21.31	1	9	-6	5
22.09	4	10	-8	4
24.36	4	6	-3	6
21.43	9	6	-4	-1
25.60	7	10	-4	-1
10.25	2	9	-8	-3
21.68	6	6	-7	4
16.55	2	12	-11	1
24.01	9	11	-10	4
14.03	1	14	-12	-1
13.61	3	8	-12	4
17.61	2	10	-6	0
16.41	1	5	-4	0
14.03	2	11	-11	-1
16.14	8	6	-12	3
20.44	5	9	-11	2
30.31	5	14	-6	3
26.03	8	14	-11	2
23.15	4	12	-4	0
14.54	6	8	-7	-2
21.07	3	12	-4	-3
17.30	8	9	-11	-1
18.43	3	11	-4	-2
12.51	5	5	-11	3
28.16	10	11	-11	3
29.98	10	5	-3	1
11.20	8	5	-11	-3
16.68	6	11	-10	3
19.78	6	7	-3	-3
16.23	1	7	-8	2

Esempio



$$F_h = \frac{0.04}{0.96} \left(\frac{32}{2} \right) = 0.67 \Rightarrow p\text{-value} = 0.41$$

Se si afferma che c'è eteroschedasticità (si rifiuta H_0) si sbaglia quattro volte su dieci.

Non è consigliabile rifiutare.

Test di Park/2

La difficoltà maggiore con questo test è la scelta dei valori della z_i .

Su questo possiamo dare solo indicazioni

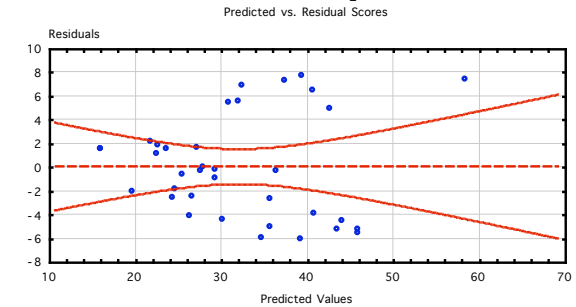
Dovrebbe essere un **FATTORE DIMENSIONALE** ovvero i quadrati degli errori dovrebbero variare con i valori di questa variabile.

La relazione tra errori e fattore dimensionale dovrebbe essere almeno **monotona**.

La disponibilità del fattore dimensionale potrebbe tornare utile per eliminare o ridurre il problema della eteroschedasticità.

y	X1	X2
24.0	2	7.8
27.4	4	9.8
21.8	8	1.2
17.6	6	0.2
24.9	5	6.5
29.1	7	6.8
25.2	3	7.9
27.9	9	2.7
24.2	6	6.0
17.5	1	4.3
24.5	3	7.0
17.5	1	4.3
36.1	9	9.8
22.8	7	2.9
28.5	8	5.4
23.7	6	2.6
33.1	9	9.3
28.9	4	9.4
39.6	2	26.5
65.9	9	28.3
30.6	6	13.6
22.3	1	13.1
37.0	3	22.3
28.7	5	14.2
25.7	3	13.3
38.2	1	27.4
40.4	8	19.3
44.7	6	15.1
33.2	5	18.1
37.6	3	14.9
39.3	1	18.2
47.6	9	15.1
47.1	4	19.6
40.6	2	28.0
36.4	1	16.9
47.2	6	17.8

Esempio



$$p\text{-value per } \theta_1 \text{ in } \ln(e_i^2) = \theta_0 + \theta_1 \ln(|x_{i1}|): 0.1041$$

$$p\text{-value per } \theta_1 \text{ in } \ln(e_i^2) = \theta_0 + \theta_1 \ln(|x_{i2}|): 0.0035$$

C'è una fortissima evidenza di eteroschedasticità, almeno in relazione al 2° regressore

Esercizio

GDP	LFG	EQP	NEQ	GAP	0.0149	0.0242	0.0384	0.0516	0.7885	-0.0047	0.0283	0.0358	0.0842	0.8579
0.0089	0.0118	0.0214	0.2286	0.6079	0.0148	0.0303	0.0446	0.0954	0.8850	0.0260	0.0150	0.0701	0.2199	0.3755
0.0332	0.0014	0.0991	0.1349	0.5809	0.0484	0.0359	0.0767	0.1233	0.7471	0.0295	0.0258	0.0263	0.0880	0.9180
0.0256	0.0061	0.0684	0.1653	0.4109	0.0115	0.0170	0.0278	0.1448	0.9356	0.0295	0.0279	0.0388	0.2212	0.8015
0.0124	0.0209	0.0167	0.1133	0.8634	0.0345	0.0213	0.0221	0.1179	0.9243	0.0261	0.0299	0.0189	0.1011	0.8458
0.0676	0.0239	0.1310	0.1490	0.9474	0.0288	0.0081	0.0814	0.1879	0.6457	0.0107	0.0271	0.0267	0.0933	0.7406
0.0437	0.0306	0.0646	0.1588	0.8498	0.0452	0.0305	0.1112	0.1788	0.6816	0.0179	0.0253	0.0445	0.0974	0.8747
0.0458	0.0169	0.0415	0.0885	0.9333	0.0362	0.0038	0.0683	0.1790	0.5441	0.0318	0.0118	0.0729	0.1571	0.8033
0.0169	0.0261	0.0771	0.1529	0.1783	0.0278	0.0274	0.0243	0.0957	0.9207	-0.0011	0.0274	0.0193	0.0807	0.8884
0.0021	0.0216	0.0154	0.2846	0.5402	0.0055	0.0201	0.0609	0.1455	0.8229	0.0373	0.0069	0.0397	0.1305	0.6613
0.0239	0.0266	0.0229	0.1553	0.7695	0.0535	0.0117	0.1223	0.2464	0.7484	0.0137	0.0207	0.0138	0.1352	0.8555
0.0121	0.0354	0.0433	0.1067	0.7043	0.0146	0.0346	0.0462	0.1268	0.9415	0.0184	0.0276	0.0860	0.0940	0.9762
0.0187	0.0115	0.0688	0.1834	0.4079	0.0479	0.0282	0.0557	0.1842	0.8807	0.0341	0.0278	0.0395	0.1412	0.9174
0.0199	0.0280	0.0321	0.1379	0.8293	0.0236	0.0064	0.0711	0.1944	0.2863	0.0279	0.0256	0.0428	0.0972	0.7838
0.0283	0.0274	0.0303	0.2097	0.8205	-0.0102	0.0203	0.0219	0.0481	0.9217	0.0189	0.0048	0.0694	0.1132	0.4307
0.0046	0.0316	0.0223	0.0577	0.8414	0.0153	0.0226	0.0361	0.0935	0.9628	0.0133	0.0189	0.0762	0.1356	0.0000
0.0094	0.0206	0.0212	0.0288	0.9805	0.0332	0.0316	0.0446	0.1878	0.7853	0.0041	0.0052	0.0155	0.1154	0.5782
0.0301	0.0083	0.1206	0.2494	0.5589	0.0044	0.0184	0.0433	0.0267	0.9478	0.0120	0.0378	0.0340	0.0760	0.4974
0.0292	0.0089	0.0879	0.1767	0.4708	0.0198	0.0349	0.0273	0.1687	0.5921	-0.0110	0.0275	0.0702	0.2012	0.8695
0.0259	0.0047	0.0890	0.1885	0.4585	0.0243	0.0281	0.0260	0.0540	0.8405	0.0110	0.0309	0.0843	0.1257	0.8875
0.0446	0.0044	0.0655	0.2245	0.7924	0.0231	0.0146	0.0778	0.1781	0.3605					

De Long and Summers (1991) studied the national growth of 61 countries from 1960 to 1985 using OLS. The regression equation they used is
 $GDP = \beta_0 + \beta_1 LFG + \beta_2 GAP + \beta_3 EQP + \beta_4 NEQ + u$
 where
 GDP growth per worker LFG labor force growth
 GAP relative GDP gap EQP equipment investment
 NEP non- equipment investment .

Stimare il modello di regressione e verificare la presenza di eteroschedasticità

Rimedi: minimi quadrati generalizzati

Se i grafici ed i test concordano nel segnalare la presenza di una variabilità non omogenea nei dati occorre ridurre il problema.

Se si è usato il test di Park e si è accertato che c'è eteroschedasticità dovuta al regressore x_k cioè si pensa $Var(u_i) = \sigma^2(X_{ik})^2$, si può agire sul modello

$$\frac{y_i}{|x_{ik}|} = \alpha_0 + \beta_0 \left(\frac{1}{|x_{ik}|} \right) + \beta_1 \left(\frac{x_{i1}}{|x_{ik}|} \right) + \beta_2 \left(\frac{x_{i2}}{|x_{ik}|} \right) + \dots + \beta_k + \dots + \beta_m \left(\frac{x_{im}}{|x_{ik}|} \right) + \frac{u_i}{|x_{ik}|}$$

$$z_i = \beta_k + \beta_0 w_{i0} + \beta_1 w_{i1} + \beta_2 w_{i2} + \dots + \beta_m w_{im} + v_i$$

il coefficiente del primo regressore del nuovo modello è l'intercetta del modello originario.

L'intercetta del nuovo modello è il coefficiente angolare del fattore di proporzionalità nel vecchio modello.

Continua esempio

y	X0	X1	X2	Z	W1	W2	W0
24.0	1.0	2	7.8	3.07	0.13	0.72	0.36
27.4	1.0	4	9.8	2.79	0.10	0.41	1.00
21.8	1.0	8	1.2	18.20	0.83	6.67	1.00
17.6	1.0	6	0.2	88.20	5.00	30.00	1.00
24.9	1.0	5	6.5	3.83	0.15	0.77	1.00
29.1	1.0	7	6.8	4.27	0.15	1.03	1.00
25.2	1.0	3	7.9	3.19	0.13	0.38	1.00
27.9	1.0	9	2.7	10.35	0.37	3.33	1.00
24.2	1.0	6	6.0	4.03	0.17	1.00	1.00
17.5	1.0	1	4.3	4.06	0.23	0.23	1.00
24.5	1.0	3	7.0	3.50	0.14	0.43	1.00
17.5	1.0	1	4.3	4.06	0.23	0.23	1.00
36.1	1.0	9	9.8	3.68	0.10	0.92	1.00
22.8	1.0	7	2.9	7.86	0.34	2.41	1.00
28.5	1.0	8	5.4	5.27	0.19	1.48	1.00
23.7	1.0	6	2.6	9.12	0.38	2.31	1.00
33.1	1.0	9	9.3	3.55	0.11	0.97	1.00
28.9	1.0	4	9.4	3.07	0.11	0.43	1.00
39.6	1.0	2	26.5	1.49	0.04	0.08	1.00
65.9	1.0	9	28.3	2.33	0.04	0.32	1.00
30.6	1.0	6	13.6	2.25	0.07	0.44	1.00
22.3	1.0	1	13.1	1.70	0.08	0.08	1.00
37.0	1.0	3	22.3	1.66	0.04	0.13	1.00
28.7	1.0	5	14.2	2.02	0.07	0.35	1.00
25.7	1.0	3	13.3	1.93	0.08	0.23	1.00
38.2	1.0	1	27.4	1.39	0.04	0.04	1.00
40.4	1.0	8	19.3	2.09	0.05	0.41	1.00
44.7	1.0	6	15.1	2.96	0.07	0.40	1.00
33.2	1.0	5	18.1	1.84	0.06	0.28	1.00
37.6	1.0	3	14.9	2.52	0.07	0.20	1.00
39.3	1.0	1	18.2	2.16	0.05	0.05	1.00
47.6	1.0	9	15.1	3.15	0.07	0.60	1.00
47.1	1.0	4	19.6	2.40	0.05	0.20	1.00
40.6	1.0	2	28.0	1.45	0.04	0.07	1.00
36.4	1.0	1	16.9	2.15	0.06	0.06	1.00
47.2	1.0	6	17.8	2.65	0.06	0.34	1.00

Test di White per la regressione di z
su w_1 e w_2

$$F_h = \frac{0.021}{1-0.021} \left(\frac{33}{2} \right) = 0.35 \Rightarrow p\text{-value} = 0.71$$

L'eteroschedasticità è stata rimossa

$$\begin{aligned} \text{Var}(v_i) &= \text{Var}\left(\frac{u_i}{|x_{ik}|}\right) = \frac{1}{|x_{ik}|^2} \text{Var}(u_i) \\ &= \frac{1}{|x_{ik}|^2} \sigma^2 |x_{ik}|^2 = \sigma^2 \text{ per ogni } i \end{aligned}$$

*Questa procedura crea delle perplessità.
Che ruolo ha il regressore che opera come
fattore di proporzionalità?*

*Agisce in proprio o subisce l'effetto di un
fattore esterno comune alla dipendente?*

Rimedi: uso dei valori stimati

Se nessuno dei regressori appare collegato alla eteroschedasticità

Ovvero non riusciamo a determinare il tipo di legame che unisce uno o più regressori agli errori

possiamo ipotizzare che

$$\text{Var}(u_i) = \sigma^2 \hat{y}_i$$

E quindi agire sul modello

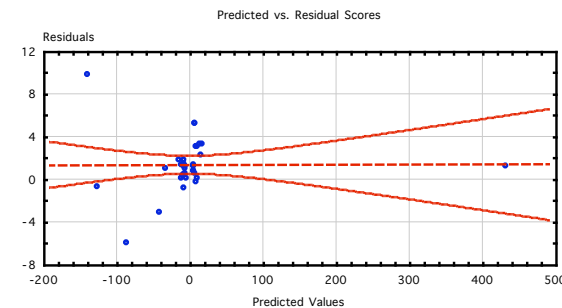
$$\frac{y_i}{\hat{y}_i} = \beta_0 \left(\frac{1}{\hat{y}_i} \right) + \beta_1 \left(\frac{x_{i1}}{\hat{y}_i} \right) + \beta_2 \left(\frac{x_{i2}}{\hat{y}_i} \right) + \dots + \beta_m \left(\frac{x_{im}}{\hat{y}_i} \right) + \frac{u_i}{\hat{y}_i}$$

$$z_i = \beta_0 w_{i0} + \beta_1 w_{i1} + \beta_2 w_{i2} + \dots + \beta_m w_{im} + v_i$$

In questo caso il modello è da stimare senza intercetta.

Prosegue esempio

y	X0	X1	X2	Z	W1	W2
24.0	1.0	2	7.8	11.09	0.93	3.61
27.4	1.0	4	9.8	-92.65	-13.55	-33.19
21.8	1.0	8	1.2	-8.69	-3.18	-0.48
17.6	1.0	6	0.2	-8.61	-2.93	-0.10
24.9	1.0	5	6.5	-44.54	-8.94	-11.63
29.1	1.0	7	6.8	-126.99	-30.59	-29.72
25.2	1.0	3	7.9	16.51	1.97	5.18
27.9	1.0	9	2.7	433.04	139.49	41.85
24.2	1.0	6	6.0	-10.09	-2.50	-2.50
17.5	1.0	1	4.3	11.14	0.64	2.74
24.5	1.0	3	7.0	12.78	1.56	3.65
17.5	1.0	1	4.3	11.14	0.64	2.74
36.1	1.0	9	9.8	-130.68	-32.62	-35.52
22.8	1.0	7	2.9	-12.24	-3.76	-1.56
28.5	1.0	8	5.4	-32.49	-9.13	-6.16
23.7	1.0	6	2.6	20.24	5.12	2.22
33.1	1.0	9	9.3	-12.34	-3.36	-3.47
28.9	1.0	4	9.4	16.97	2.35	5.53
39.6	1.0	2	26.5	-8.83	-0.45	-5.91
65.9	1.0	9	28.3	8.81	1.20	3.78
30.6	1.0	6	13.6	-6.08	-1.19	-2.70
22.3	1.0	1	13.1	-5.50	-0.25	-3.23
37.0	1.0	3	22.3	-9.59	-0.78	-5.79
28.7	1.0	5	14.2	-4.88	-0.85	-2.41
25.7	1.0	3	13.3	-5.79	-0.68	-3.00
38.2	1.0	1	27.4	-7.29	-0.19	-5.23
40.4	1.0	8	19.3	-7.26	-1.44	-3.47
44.7	1.0	6	15.1	6.15	0.82	2.08
33.2	1.0	5	18.1	-5.48	-0.82	-2.99
37.6	1.0	3	14.9	6.73	0.54	2.67
39.3	1.0	1	18.2	5.72	0.15	2.64
47.6	1.0	9	15.1	9.58	1.81	3.04
47.1	1.0	4	19.6	6.05	0.51	2.52
40.6	1.0	2	28.0	-7.70	-0.38	-5.31
36.4	1.0	1	16.9	6.65	0.18	3.09
47.2	1.0	6	17.8	7.26	0.92	2.74



Questa procedura non sembra avere successo

$$F_h = \frac{0.41}{0.59} \left(\frac{33}{2} \right) = 11.47 \Rightarrow p\text{-value} = 0.0002$$

Nonostante la trasformazione la
eteroschedasticità è rimasta tutta.

Rimedi: minimi quadrati ponderati

I pesi che misurano la variabilità delle osservazioni potrebbero essere noti o resunti tali da ricerche precedenti o analoghe.

Sia w_i $i=1,2,\dots,n$ la variabile che contiene tali indicazioni.

Invece di usare gli OLS possiamo usare i WLS che mirano a minimizzare...

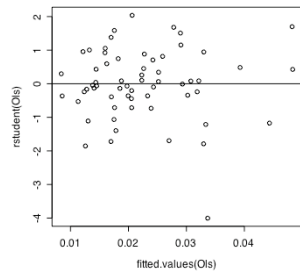
$$S(\beta) = \sum_{i=1}^n w_i \left(y_i - \sum_{j=0}^m \beta_j x_{ij} \right)^2$$

In cui w_i è decrescente per errori crescenti.

Se ad esempio il dato y_i è la media di n_i osservazioni allora $w_i=1/n_i$ è una scelta ragionevole.

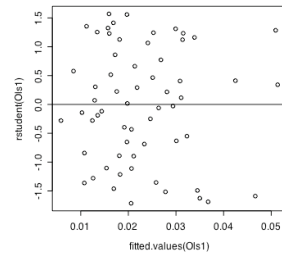
Un tentativo potrebbe essere $w_i = \frac{1}{(\hat{\epsilon}_i)^2 + 0.0001}$, $\hat{\epsilon}_i = \text{residui OLS}$

Applicazione Esercizio



studentized Breusch-Pagan test

data: Ols
BP = 8.4738, df = 4, p-value = 0.07569



studentized Breusch-Pagan test

data: GDP ~ .
BP = 0.8329, df = 1, p-value = 0.3614

L'eteroschedasticità è stata rimossa. Ma non possiamo essere sicuri che il modello abbia conservato la logica iniziale

Rimedi: uso dei logaritmi

I logaritmi attenuano le differenze di scala e quindi potrebbero risultare utili qualora a queste sia da attribuire la eteroschedasticità

Se il fattore di proporzionalità è poco diversificato, il rimedio potrebbe essere efficace.

Ad esempio sugli errori si ha

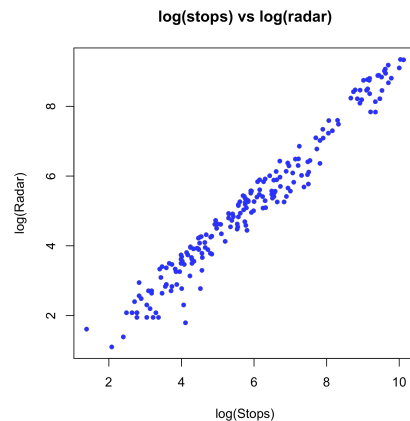
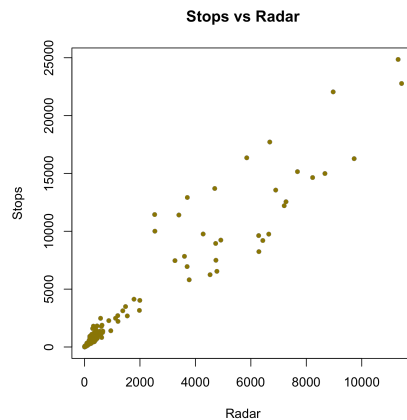
$$\text{Var}[Ln(au_i)] = \text{Var}[Ln(a) + Ln(u_i)] = \text{Var}[Ln(u_i)]$$

Se sono presenti valori negativi si può sottrarre il valore minimo.

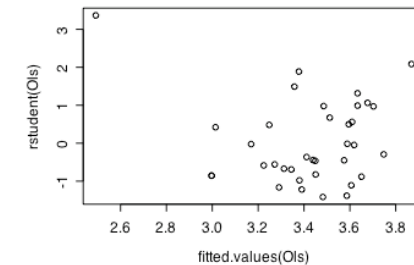
$$Ln(y_i - y_{\min} + 1) = \beta_0 + \beta_1 Ln(x_{i1} - x_{\min,1} + 1) + \dots + \beta_m Ln(x_{im} - x_{\min,m} + 1) + v_i$$

uso dei logaritmi/1

```
traffic <- read.csv("WSPTrafficStops.csv")
```



Ancora esempio



La regressione con i logaritmi appare efficace

studentized Breusch-Pagan test

data: Ols
BP = 1.7683, df = 2, p-value = 0.4131

y	X0	X1	X2	Z	W1	W2
24.0	1.0	2	7.8	3.18	0.69	2.05
27.4	1.0	4	9.8	3.31	1.39	2.28
21.8	1.0	8	1.2	3.08	2.08	0.18
17.6	1.0	6	0.2	2.87	1.79	-1.61
24.9	1.0	5	6.5	3.21	1.61	1.87
29.1	1.0	7	6.8	3.37	1.95	1.92
25.2	1.0	3	7.9	3.23	1.10	2.07
27.9	1.0	9	2.7	3.33	2.20	0.99
24.2	1.0	6	6.0	3.19	1.79	1.79
17.5	1.0	1	4.3	2.86	0.00	1.46
24.5	1.0	3	7.0	3.20	1.10	1.95
17.5	1.0	1	4.3	2.86	0.00	1.46
36.1	1.0	9	9.8	3.59	2.20	2.28
22.8	1.0	7	2.9	3.13	1.95	1.06
28.5	1.0	8	5.4	3.35	2.08	1.69
23.7	1.0	6	2.6	3.17	1.79	0.96
33.1	1.0	9	9.3	3.50	2.20	2.23
28.9	1.0	4	9.4	3.36	1.39	2.24
39.6	1.0	2	26.5	3.68	0.69	3.28
65.9	1.0	9	28.3	4.19	2.20	3.34
30.6	1.0	6	13.6	3.42	1.79	2.61
22.3	1.0	1	13.1	3.11	0.00	2.57
37.0	1.0	3	22.3	3.61	1.10	3.10
28.7	1.0	5	14.2	3.36	1.61	2.65
25.7	1.0	3	13.3	3.24	1.10	2.59
38.2	1.0	1	27.4	3.64	0.00	3.31
40.4	1.0	8	19.3	3.70	2.08	2.96
44.7	1.0	6	15.1	3.80	1.79	2.71
33.2	1.0	5	18.1	3.50	1.61	2.90
37.6	1.0	3	14.9	3.63	1.10	2.70
39.3	1.0	1	18.2	3.67	0.00	2.90
47.6	1.0	9	15.1	3.86	2.20	2.71
47.1	1.0	4	19.6	3.85	1.39	2.98
40.6	1.0	2	28.0	3.70	0.69	3.33
36.4	1.0	1	16.9	3.59	0.00	2.83
47.2	1.0	6	17.8	3.85	1.79	2.88

Deviazioni standard Eicker-White

Si sospetta eteroschedasticità, ma non si può o non si ritiene necessario procedere con i minimi quadrati generalizzati o altre forme di correzione.

Si possono usare degli errori standard robusti rispetto alla violazione della omoschedasticità

$$Var(\tilde{\beta}) = (X'X)^{-1} (X'\Delta X) (X'X)^{-1}$$

$$\Delta = diag\left(\frac{ne_i^2}{n-m}\right), \quad i = 1, 2, \dots, n$$

I parametri rimangono quelli ottenuti con la procedura standard dei minimi quadrati con residui e_i

In questo modo si ottengono gli scarti quadratici medi noti come heteroschedasticity-consistent.

I t-Student hanno ora una maggiore attendibilità

Altre deviazioni standard robuste

MacKinnon and White (1985) coinvolgono la leva delle osservazioni

$$MW : Var(\tilde{\beta}) = (X'X)^{-1} \left[X' diag\left(\frac{e_i^2}{1-h_{ii}}\right) X \right] (X'X)^{-1}$$

Long and Ervin (2000) coinvolgono i residui deleted

$$LI : Var(\tilde{\beta}) = (X'X)^{-1} \left[X' diag\left(\frac{e_i}{1-h_{ii}}\right)^2 X \right] (X'X)^{-1}$$

Secondo certe simulazioni questa forma risulta la più efficiente per assicurare p-value vicini a quelli veri in caso di eteroschedasticità ignorata

Esercizio

x1	x2	x3	Cycles
-1	-1	-1	674
-1	-1	0	370
-1	-1	1	292
-1	0	-1	338
-1	0	0	266
-1	0	1	210
-1	1	-1	170
-1	1	0	118
-1	1	1	90
0	-1	-1	1414
0	-1	0	1198
0	-1	1	634
0	0	-1	1022
0	0	0	620
0	0	1	438
0	1	-1	442
0	1	0	332
0	1	1	220
1	-1	-1	3636
1	-1	0	3184
1	-1	1	2000
1	0	-1	1568
1	0	0	1070
1	0	1	566
1	1	-1	1140
1	1	0	884
1	1	1	360

Wool data: number of cycles to failure of samples of worsted yarn in a 3rd experiment

Stimare il modello di regressione e verificare la presenza di eteroschedasticità

In caso affermativo procedere

- 1) **Al calcolo degli errori standard robusti rispetto alla eteroschedasticità**
- 2) **Ridurre la eteroschedasticità con i minimi quadrati ponderati**