Università di Pavia

Test diagnostici

Eduardo Rossi

Fase di controllo diagnostico: controllo della coerenza tra quanto direttamente osservato e le ipotesi statistiche adottate.

Ipotesi MRLM gaussiano:

1. $E[\epsilon|\mathbf{X}] = \mathbf{0}$ non ci sono errori sistematici.
2. I parametri $\boldsymbol{\beta}$ e σ^2 sono costanti nel campione osservato.
3. $E[\epsilon\epsilon'|\mathbf{X}] = \sigma^2\mathbf{I}_N$ omoschedasticità.
4. $\epsilon|\mathbf{X} \sim N(\mathbf{0}, \sigma_2\mathbf{I}_N)$ gaussianità dei disturbi.

Nei test diagnostici l'ipotesi nulla e la sua alternativa vengono usate solo nel test. Il rifiuto della nulla non porta a ritenere valida l'alternativa (test non informativi). Deve essere interpretato semplicemente come un segnale di non coerenza tra quanto ipotizzato e le osservazioni disponibili. Le cause del rifiuto possono essere diverse e non sempre identificabili con la specifica alternativa considerata.
L'esito è la rispecificazione del modello.

Procedure di test che si basano sull'analisi dei residui delle stime OLS.

Test F basati sul confronto tra il confronto tra la somma dei quadrati dei residui di un modello vincolato (*regressione ausiliaria*) e quella di un modello non vincolato.

Test diagnostici presentati nella forma di variabili omesse, si verifica se nel modello si sta commettendo un errore di specificazione dovuto alla non introduzione di regressori rilevanti.

Modello allargato

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\eta}$$

Modello scorrettamente specificato

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$\mathbf{W}$ $(N \times h)$ regressori potenzialmente omessi, $\boldsymbol{\gamma}$ $(h \times 1)$.

Lo stimatore $\widehat{\boldsymbol{\beta}}$ calcolato con il modello scorrettamente specificato è distorto rispetto a $\boldsymbol{\beta}^*$:

$$\begin{aligned}
E\left[\widehat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{W}\right] &= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X},\mathbf{W}\right] \\
&= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}^* + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\eta})|\mathbf{X},\mathbf{W}\right] \\
&= \boldsymbol{\beta}^* + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{\gamma} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\boldsymbol{\eta}|\mathbf{X},\mathbf{W}] \\
&= \boldsymbol{\beta}^* + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{\gamma}
\end{aligned}$$

Per verificare la presenza di $\mathbf{W}$ applichiamo il test F per la verifica dell'ipotesi nulla

$$H_0 : \boldsymbol{\gamma} = \mathbf{0}$$

Per il modello sotto l'ipotesi nulla:

$$\mathbf{y} = \mathbf{P}_X\mathbf{y} + \mathbf{M}_X\mathbf{y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{M}_X\mathbf{y}$$

$$\mathbf{M}_X\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\eta}$$

il modello ausiliario risulta essere:

$$\widehat{\boldsymbol{\epsilon}} = \mathbf{X}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}) + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\eta}$$

da cui

$$\widehat{\boldsymbol{\eta}} = \mathbf{M}_{X,W}\widehat{\boldsymbol{\epsilon}} = \mathbf{M}_{X,W}\mathbf{M}_X\mathbf{y} = \mathbf{M}_{X,W}\left(\mathbf{I}_N - \mathbf{P}_X\right)\mathbf{y} = \mathbf{M}_{X,W}\mathbf{y}$$

Le somme dei quadrati dei residui del modello ristretto e non ristretto sono:

$$\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}} = \mathbf{y}'\mathbf{M}_X\mathbf{y}$$

$$\widehat{\boldsymbol{\eta}}'\widehat{\boldsymbol{\eta}} = \mathbf{y}'\mathbf{M}_{X,W}\mathbf{y}$$

La statistica test $\widehat{F}$

$$\widehat{F} = \frac{\mathbf{y}'\mathbf{M}_X\mathbf{y} - \mathbf{y}'\mathbf{M}_{X,W}\mathbf{y}}{\mathbf{y}'\mathbf{M}_{X,W}\mathbf{y}} \frac{T-K-h}{h} \sim F_{(h,N-K-h)}$$

- Se il numero delle variabili in $\mathbf{W}$, h, è troppo elevato il test non risulta potente.
- Le variabili che entrano in $\mathbf{W}$ non sono note a priori. Sono variabili *proxy*, variabili che sono utili a cogliere possibili errori commessi nella specificazione. In caso di rifiuto non indicano con certezza quale sia la direzione in cui muoversi per una riformulazione del modello.

Una delle ipotesi del MRLM è che i parametri siano costanti su tutto il campione, condizione che deve essere vera per calcolare delle previsioni utili.

Chow (1960) ha proposto due test per verificare:

1. Stabilità dei parametri
2. Capacità previsiva del modello

Funzione del consumo

$$C_t = \boldsymbol{\beta}_1 + \beta_2 Y_t + \epsilon_t \quad t = 1, 2, \ldots, T$$

se osserviamo che il consumo in un certo intervallo di tempo è più basso (o più alto) allora possiamo pensare che il processo generatore dei dati (il modello vero) sia:

$$C_t = \beta_1 + \beta_2 Y_t + \alpha D_t + \epsilon_t \quad t = 1, 2, \ldots, T$$

dove D_t è una variabile *dummy* cioè una grandezza che assume valore 0 per un certo numero di periodi ed 1 negli altri.

Per esempio, il consumo è più basso durante il periodo bellico rispetto a quello del tempo di pace

$$D_t = \begin{cases} 0 & t \in T_p \\ 1 & t \in T_g \end{cases}$$

$T_p + T_g = T$, in questo modo modifichiamo l'intercetta nel tempo di pace rispetto al tempo di guerra:

$$C_t = \beta_1 + \alpha + \beta_2 Y_t + \epsilon_t \quad t \in T_g$$

$$C_t = \beta_1 + \beta_2 Y_t + \epsilon_t \quad t \in T_p$$

Cambiamento di pendenza:

$$C_t = \beta_1 + \beta_2 Y_t + \gamma(D_t Y_t) + \epsilon_t \quad t = 1, 2, \ldots, T$$

vi è una modificazione della propensione marginale al consumo:

$$C_t = \beta_1 + (\beta_2 + \gamma) Y_t + \epsilon_t \quad t \in T_g$$

$$C_t = \beta_1 + \beta_2 Y_t + \epsilon_t \quad t \in T_p$$

Modello con cambiamento di pendenza e di intercetta:

$$C_t = \beta_1 + \beta_2 Y_t + \alpha D_t + \gamma(D_t Y_t) + \epsilon_t \quad t = 1, 2, \ldots, T$$

$$E[C_t|Y_t] = \begin{cases} \beta_1 + \beta_2 Y_t & t \in T_p \\ (\beta_1 + \alpha) + (\beta_2 + \gamma) Y_t & t \in T_g \end{cases}$$

Vediamo come si costruisce la matrice dei regressori $\mathbf{X}$ quando vi sono delle dummy. Cambiamento di intercetta:

$$C_t = \beta_1 + \beta_2 Y_t + \alpha D_t + \epsilon_t \quad t = 1, 2, \ldots, T$$

$$\mathbf{X} = \begin{bmatrix} 1 & Y_1 & 0 \\ 1 & Y_2 & 0 \\ \vdots & \vdots & \vdots \\ 1 & Y_s & 1 \\ 1 & Y_{s+1} & 1 \\ 1 & Y_{s+2} & 1 \\ 1 & Y_{s+3} & 0 \\ \vdots & \vdots & \vdots \\ 1 & Y_T & 0 \end{bmatrix}$$

Cambiamento di pendenza:

$$C_t = \beta_1 + \beta_2 Y_t + \gamma(D_t Y_t) + \epsilon_t \quad t = 1, 2, \ldots, T$$

$$\mathbf{X} = \begin{bmatrix} 1 & Y_1 & 0 \\ 1 & Y_2 & 0 \\ \vdots & \vdots & \vdots \\ 1 & Y_s & Y_s \\ 1 & Y_{s+1} & Y_{s+1} \\ 1 & Y_{s+2} & Y_{s+2} \\ 1 & Y_{s+3} & 0 \\ \vdots & \vdots & \vdots \\ 1 & Y_T & 0 \end{bmatrix}$$

Cambiamento di intercetta e di pendenza:

$$C_t = \beta_1 + \beta_2 Y_t + \alpha D_t + \gamma(D_t Y_t) + \epsilon_t \quad t = 1, 2, \ldots, T$$

$$\mathbf{X} = \begin{bmatrix} 1 & Y_1 & 0 & 0 \\ 1 & Y_2 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & Y_s & 1 & Y_s \\ 1 & Y_{s+1} & 1 & Y_{s+1} \\ 1 & Y_{s+2} & 1 & Y_{s+2} \\ 1 & Y_{s+3} & 0 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & Y_T & 0 & 0 \end{bmatrix}$$

- Se i valori dei test t di $\widehat{\alpha}$ e $\widehat{\gamma}$ sono maggiori di 2, allora i due parametri sono statisticamente diversi da zero, si conclude che esiste un cambiamento di intercetta e di pendenza.
- Il modello più usato per la spiegazione del consumo è quella con cambiamento di intercetta: il t corrispondente a $\widehat{\gamma}$ è minore di 2 $(\alpha \neq 0, \gamma = 0)$.
- Se entrambi i test t sono minori di 2 dobbiamo testare in modo congiunto

$$H_0 : \alpha = \gamma = 0$$

Con dati trimestrali il consumo presenta un picco nel secondo trimestre e nel quarto. Supponiamo che esistano 4 relazioni diverse del consumo nell'arco dell'anno:

$$
\begin{aligned}
C_t &= \beta_1 + \beta_2 Y_t + \alpha_1 D_{1t} + \epsilon_t \\
C_t &= \beta_1 + \beta_2 Y_t + \alpha_2 D_{2t} + \epsilon_t \\
C_t &= \beta_1 + \beta_2 Y_t + \alpha_3 D_{3t} + \epsilon_t \\
C_t &= \beta_1 + \beta_2 Y_t + \epsilon_t
\end{aligned}
$$

$$D_{1t} = \begin{cases} 0 & 2,3,4 \quad \text{trimestre} \\ 1 & 1 \quad\quad\;\; \text{trimestre} \end{cases}$$

$$D_{2t} = \begin{cases} 0 & 1,3,4 \quad \text{trimestre} \\ 1 & 2 \quad\quad\;\; \text{trimestre} \end{cases}$$

$$D_{3t} = \begin{cases} 0 & 1,2,4 \quad \text{trimestre} \\ 1 & 3 \quad\quad\;\; \text{trimestre} \end{cases}$$

La quarta equazione non ha dummy. E' l'equazione di riferimento cioè la base di partenza rispetto alla quale c'è la correzione di intercetta.

$T = 8$, 8 trimestri

$$\mathbf{X} = \begin{bmatrix} 1 & Y_1 & 1 & 0 & 0 \\ 1 & Y_2 & 0 & 1 & 0 \\ 1 & Y_3 & 0 & 0 & 1 \\ 1 & Y_4 & 0 & 0 & 0 \\ 1 & Y_5 & 1 & 0 & 0 \\ 1 & Y_6 & 0 & 1 & 0 \\ 1 & Y_7 & 0 & 0 & 1 \\ 1 & Y_8 & 0 & 0 & 0 \end{bmatrix}$$

Le ultime tre colonne rappresentano le 3 dummy.

Attenzione: Se inseriamo la quarta dummy

$$\mathbf{X} = \begin{bmatrix} 1 & Y_1 & 1 & 0 & 0 & 0 \\ 1 & Y_2 & 0 & 1 & 0 & 0 \\ 1 & Y_3 & 0 & 0 & 1 & 0 \\ 1 & Y_4 & 0 & 0 & 0 & 1 \\ 1 & Y_5 & 1 & 0 & 0 & 0 \\ 1 & Y_6 & 0 & 1 & 0 & 0 \\ 1 & Y_7 & 0 & 0 & 1 & 0 \\ 1 & Y_8 & 0 & 0 & 0 & 1 \end{bmatrix}$$

la prima colonna

$$\mathbf{X}_{\cdot 1} = \mathbf{X}_{\cdot 3} + \mathbf{X}_{\cdot 4} + \mathbf{X}_{\cdot 5} + \mathbf{X}_{\cdot 6}$$

abbiamo una matrice di rango ridotto (collinearità perfetta).

Con dati trimestrali si usano 3 dummy, con dati mensili si usano 11 dummy.

La presenza di outlier la si può accertare, in prima istanza, attraverso l'analisi dei residui. Quando vi sono residui molto grandi è probabile che siamo in presenza di un outlier. In questo caso possiamo usare una dummy per quella osservazione, ad esempio

$$C_t = \beta_1 + \beta_2 Y_t + \alpha_1 D_{1974} + \epsilon_t$$

$$D_{1974} = \begin{cases} 0 & t \neq 1974 \\ 1 & t = 1974 \end{cases}$$

Specificazioni alternative in cui tutti i coefficienti sono differenti in due più sottocampioni. Per esempio i coefficienti possono essere diversi dopo eventi di politica economica o duranti periodi caratterizzati da condizioni particolari (esempio eventi bellici). In questi casi si parla di *break strutturale*.

$$H_0: \quad y_t = \mathbf{x}_t'\boldsymbol{\beta} + \epsilon_t \quad t = 1, \ldots, T \tag{1}$$

$$H_1: \begin{cases} y_t = \mathbf{x}_t'\boldsymbol{\beta} + \epsilon_t & t = 1, \ldots, T_1 \\ y_t = \mathbf{x}_t'\boldsymbol{\beta} + \mathbf{x}_t'\boldsymbol{\gamma} + \epsilon_t & t = T_1 + 1, \ldots, T \end{cases} \tag{2}$$

Break strutturale in T_1, da questa data in avanti i parametri cambiano.

E' cruciale identificare il momento della rottura T_1

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_{T_1} \\ \mathbf{y}_h \end{bmatrix} \begin{matrix} (T_1 \times 1) \\ (h \times 1) \end{matrix} \tag{3}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{T_1} \\ \mathbf{X}_h \end{bmatrix} \begin{matrix} (T_1 \times K) \\ (h \times K) \end{matrix} \tag{4}$$

$$h \equiv T - T_1$$

L'ipotesi nulla può essere riscritta come:

$$H_0 : \begin{cases} \mathbf{y}_{T_1} & = & \mathbf{X}_{T_1}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{T_1} \\ \mathbf{y}_h & = & \mathbf{X}_h\boldsymbol{\beta} + \boldsymbol{\epsilon}_h \end{cases} \tag{5}$$

$$\left[\begin{array}{c} \mathbf{y}_{T_1} \\ \mathbf{y}_h \end{array} \right] = \left[\begin{array}{c} \mathbf{X}_{T_1} \\ \mathbf{X}_h \end{array} \right] \boldsymbol{\beta} + \left[\begin{array}{c} \epsilon_{T_1} \\ \epsilon_h \end{array} \right] \tag{6}$$

$$H_1 : \begin{cases} \mathbf{y}_{T_1} & = & \mathbf{X}_{T_1}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{T_1} \\ \mathbf{y}_h & = & \mathbf{X}_h\boldsymbol{\beta} + \mathbf{X}_h\boldsymbol{\gamma} + \boldsymbol{\epsilon}_h \end{cases} \tag{7}$$

$$\boldsymbol{\gamma} : \quad (K \times 1)$$

$$\begin{bmatrix} \mathbf{y}_{T_1} \\ \mathbf{y}_h \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{T_1} & \mathbf{0} \\ \mathbf{X}_h & \mathbf{X}_h \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_{T_1} \\ \boldsymbol{\epsilon}_h \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{y}_{T_1} \\ \mathbf{y}_h \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{T_1} \\ \mathbf{X}_h \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{0} \\ \mathbf{X}_h \end{bmatrix} \boldsymbol{\gamma} + \begin{bmatrix} \epsilon_{T_1} \\ \epsilon_h \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{8}$$

L'ipotesi nulla diventa:

$$H_0 : \boldsymbol{\gamma} = \mathbf{0} \tag{9}$$

La statistica test è:

$$\frac{(\widetilde{\boldsymbol{\epsilon}}'\widetilde{\boldsymbol{\epsilon}} - \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}})}{\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}} \; \frac{T-2K}{K} \sim F(K, T-2K) \tag{10}$$

Le fonti degli errori di previsione sono

- Incertezza sul modello (la varianza del disturbo e l'incertezza delle stime)
- Errori di previsione delle variabili esplicative (errori di scenario)

Il campione è diviso in due parti. Stimiamo i parametri con le prime T_1 osservazioni e quindi costruiamo false previsionicon le rimanenti osservazioni da T_1 a T. In questo modo eliminiamo l'incertezza dello scenario. In questo modo testiamo la presenza di un errore di previsione sistematico.

$$H_0: \quad y_t = \mathbf{x}_t'\boldsymbol{\beta} + \epsilon_t \quad t = 1, \ldots, T \tag{11}$$

$$H_1: \begin{cases} y_t = \mathbf{x}_t'\boldsymbol{\beta} + \epsilon_t & t = 1, \ldots, T_1 \\ y_t = \mathbf{x}_t'\boldsymbol{\beta} + \gamma_t + \epsilon_t & t = T_1 + 1, \ldots, T \end{cases} \tag{12}$$

γ_t è l'errore sistematico di previsione

$$H_0 : \begin{cases} \mathbf{y}_{T_1} & = & \mathbf{X}_{T_1}\boldsymbol{\beta} + \epsilon_{T_1} \\ \mathbf{y}_h & = & \mathbf{X}_h\boldsymbol{\beta} + \epsilon_h \end{cases} \tag{13}$$

$$H_1 : \begin{cases} \mathbf{y}_{T_1} & = & \mathbf{X}_{T_1}\boldsymbol{\beta} + \epsilon_{T_1} \\ \mathbf{y}_h & = & \mathbf{X}_h\boldsymbol{\beta} + \mathbf{I}_h\gamma + \epsilon_h \end{cases} \tag{14}$$

$$\boldsymbol{\gamma} \quad (h \times 1)$$

$$\begin{bmatrix} \mathbf{y}_{T_1} \\ \mathbf{y}_h \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{T_1} & \mathbf{0} \\ \mathbf{X}_h & \mathbf{I}_h \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} + \begin{bmatrix} \epsilon_{T_1} \\ \epsilon_h \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{y}_{T_1} \\ \mathbf{y}_h \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{T_1} \\ \mathbf{X}_h \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_h \end{bmatrix} \boldsymbol{\gamma} + \begin{bmatrix} \epsilon_{T_1} \\ \epsilon_h \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{15}$$

L'ipotesi nulla è

$$H_0 : \boldsymbol{\gamma} = \mathbf{0} \tag{16}$$

il modello sotto l'ipotesi nulla è:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

La statistica test è

$$\frac{\widetilde{\boldsymbol{\epsilon}}'\widetilde{\boldsymbol{\epsilon}} - \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}}{\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}} \; \frac{T-K-h}{h} \sim F(h, T-K-h) \tag{17}$$

REgression **S**pecification **E**rror **T**est (Ramsey (1969)).
Test sia sulla forma funzionale del modello sia sulla possibile omissione di variabili rilevanti.

Non avendo una precisa conoscenza dell'errore di specificazione, Ramsey suggerisce di approssimarlo con qualche trasformazione della media condizionale.

Nella pratica econometrica si utilizzano un certo numero di potenze dei valori calcolati della y.

Modello di partenza

$$y_t = \mathbf{x}_t' \boldsymbol{\beta} + \epsilon_t$$

$$\widehat{y}_t = \mathbf{x}_t' \widehat{\boldsymbol{\beta}}$$

Modello non vincolato

$$y_t = \mathbf{x}_t' \boldsymbol{\beta} + \gamma_1 \widehat{y}_t^2 + \gamma_2 \widehat{y}_t^3 + \ldots + \gamma_{p-1} \widehat{y}_t^p + \eta_t$$

Il test RESET non è utile per trovare alternative specifiche. Segnala una scorretta specificazione.

L'ipotesi di assenza di correlazione negli errori

$$E[\epsilon_t \epsilon_\tau | \mathbf{X}] = E[\epsilon_t \epsilon_\tau] = 0 \qquad \forall \ t \neq \tau$$

Il test DW serve ad accertare la presenza di autocorrelazione nel termine d'errore. Gli errori non sono direttamente osservabili. Per la costruzione del test si fa ricorso alle stime, ovvero i residui

$$\widehat{\epsilon}_t = y_t - \mathbf{x}_t \widehat{\boldsymbol{\beta}}$$

si parla anche di test di autocorrelazione dei residui.

Autocorrelazione del primo ordine:

$$\epsilon_t = \rho \epsilon_{t-1} + u_t$$

$$H_0 : \rho = 0$$

$$E[(\epsilon_t - \epsilon_{t-1})^2] = E(\epsilon_t^2) + E(\epsilon_{t-1}^2) - 2E(\epsilon_t \epsilon_{t-1})$$

con l'ipotesi di omoschedastictà

$$E(\epsilon_t^2) = E(\epsilon_{t-1}^2)$$

si ottiene

$$\frac{E[(\epsilon_t - \epsilon_{t-1})^2]}{E(\epsilon_t^2)} = 2 - 2\frac{E(\epsilon_t \epsilon_{t-1})}{E(\epsilon_t^2)} = 2(1-\rho)$$

Se $\rho = 0$ allora

$$\frac{E[(\epsilon_t - \epsilon_{t-1})^2]}{E(\epsilon_t^2)} = 2$$

La statistica test

$$DW = \frac{\sum_{t=2}^N (\widehat{\epsilon}_t - \widehat{\epsilon}_{t-1})^2}{\sum_{t=1}^N \widehat{\epsilon}_t^2} = \frac{\sum_{t=2}^N \widehat{\epsilon}_t^2 + \sum_{t=2}^N \widehat{\epsilon}_{t-1}^2 - 2\sum_{t=2}^N \widehat{\epsilon}_t \widehat{\epsilon}_{t-1}}{\sum_{t=1}^N \widehat{\epsilon}_t^2}$$

per N grande

$$\sum_{t=2}^{N} \widehat{\epsilon}_t^2 \approx \sum_{t=2}^{N} \widehat{\epsilon}_{t-1}^2$$

$$DW = \frac{2\sum_{t=2}^{N} \widehat{\epsilon}_t^2 - 2\sum_{t=2}^{N} \widehat{\epsilon}_t \widehat{\epsilon}_{t-1}}{\sum_{t=1}^{N} \widehat{\epsilon}_t^2} \approx 2(1-\widehat{\rho})$$

$$\widehat{\rho} = \frac{\sum_{t=2}^{N} \widehat{\epsilon}_t \widehat{\epsilon}_{t-1}/N}{\sum_{t=1}^{N} \widehat{\epsilon}_t^2/N}$$

$$\widehat{\rho} \approx 1 \quad DW \approx 0$$

$$\widehat{\rho} \approx -1 \quad DW \approx 4$$

$$\widehat{\rho} \approx 0 \quad DW \approx 2$$

Distribuzione campionaria centrata su 2. Sotto H_0 la distribuzione campionaria non è nota poichè dipende dai valori assunti dai regressori.

La distribuzione della statistica si trova compresa tra due distribuzioni che non dipendono da valori ignoti. d_L livello inferiore, d_U livello superiore. La distribuzione di d_U e d_L dipende solo da N e K, i valori al 5% e al 10% sono stati tabulati. Tabulati valori critici in corrispondenza di determinati livelli di siginificatività (5% e 10%).

Regola di decisione del test DW prevede due regioni di indecisione che non è possibile eliminare.

H_0	Decisione	Se
$H_0 : \rho = 0, H_1 : \rho > 0$	Rifiuto	$0 < d < d_L$
$H_0 : \rho = 0, H_1 : \rho > 0$	Indecisione	$d_L \leq d \leq d_U$
$H_0 : \rho = 0, H_1 : \rho < 0$	Rifiuto	$4 - d_L < d < 4$
$H_0 : \rho = 0, H_1 : \rho < 0$	Indecisione	$4 - d_U \leq d \leq 4 - d_L$
$H_0 : \rho = 0, H_1 : \rho \neq 0$	Non rifiuto	$d_U < d < 4 - d_U$

Per $N = 40$ ed una variabile esplicativa $d_L = 1,44$ e $d_U = 1,54$ al 5%, se $DW = 0,123$ allora è minore di d_L, non possiamo rifiutare l'ipotesi che ci sia correlazione seriale positiva nei residui.

Il DW è inappropriato quando tra le variabili esplicative vi è un'endogena ritardata.

Test per verificare la presenza di schemi di autocorrelazione superiore al primo ordine.

$$y_t = \mathbf{x}_t' \boldsymbol{\beta} + \epsilon_t$$

con

$$\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + \ldots + \rho_p \epsilon_{t-p} + u_t$$

con

$$E[u_t u_\tau | \mathbf{X}] = 0 \qquad \forall t \neq \tau$$

se $\epsilon_t \sim AR(2)$,

$$\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + u_t$$

$$y_t = \mathbf{x}_t' \boldsymbol{\beta} + \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + u_t$$

$$H_0 : \rho_1 = \rho_2 = 0$$

sotto l'ipotesi nulla $\epsilon_t = u_t$

$$y_t = \mathbf{x}_t' \boldsymbol{\beta} + \epsilon_t$$

$$\mathbf{W} = \begin{bmatrix} 0 & 0 \\ \widehat{\epsilon}_1 & 0 \\ \widehat{\epsilon}_2 & \widehat{\epsilon}_1 \\ \vdots & \vdots \\ \widehat{\epsilon}_{N-1} & \widehat{\epsilon}_{N-2} \end{bmatrix}$$

$$\widehat{\boldsymbol{\epsilon}} = \mathbf{X}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}) + \mathbf{W}\boldsymbol{\rho} + \mathbf{u}$$

Non sappiamo come si distribuisce la statistica test $\widehat{F}$ per $H_0 : \boldsymbol{\rho} = \mathbf{0}$, quando N è grande

$$\widehat{F} \sim F_{(2,N-K-2)}.$$

Per $\epsilon_t \sim AR(p)$

$$\mathbf{W} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ \widehat{\epsilon}_1 & 0 & \dots & 0 \\ \widehat{\epsilon}_2 & \widehat{\epsilon}_1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ \widehat{\epsilon}_{N-1} & \widehat{\epsilon}_{N-2} & \dots & \widehat{\epsilon}_{N-p} \end{bmatrix}$$

Il test si basa sulla sostituzione degli errori con i residui. Mentre l'autocorrelazione degli errori implica l'autocorrelazione dei residui non è necessariamente vero il contrario. Se abbiamo un basso valore della statistica test DW questo significa che vi è autocorrelazione del primo ordine dei residui ma non significa che ci sia autocorrelazione del primo ordine negli errori.

In generale, l'autocorrelazione dei residui è prodotta dalla mancata (o dalla scorretta) specificazione dinamica del modello.

Eteroschedasticità: $E(\epsilon_t^2|\mathbf{X}) = \sigma_t^2$

Analisi visiva dei residui. Grafico dei residui in funzione del regressore a cui si sospetta collegata la varianza. Sein questo grafico il valore assoluto dei residui rimane costanteper tutto il campione allora si può concludere che non vi è eteroschedasticità.

Quando si è in presenza di eteroschedasticità la varianza condizionale di ϵ_t

$$Var(\epsilon_t|\mathbf{X}, \mathbf{W}) = \sigma_t^2 = f(\mathbf{X}, \mathbf{W}) > 0$$

Le variabili in $\mathbf{W}$ possono coincidere con quelle in $\mathbf{X}$. POssiamo scrivere

$$\epsilon_t = E(\epsilon_t^2|\mathbf{X}, \mathbf{W}) = f(\mathbf{X}, \mathbf{W}) + u_t$$

con $E(u_t|\mathbf{X}, \mathbf{W}) = 0, \forall t$.
Il test di White (1980) si basa sulla scelta delle variabili in $\mathbf{W}$ e su una particolare formulazione della $f(\mathbf{X}, \mathbf{W})$.

White propone di adottare come $f(\mathbf{X}, \mathbf{W})$ una formulazione lineare e come $\mathbf{W}$:

- la costante
- le x_{tk}, $k = 1, \ldots, K$
- i quadrati delle variabili esplicative x_{tk}^2, $k = 2, \ldots, K$
- i prodotti incrociati $x_{tk}x_{tj}$ per $k, j = 2, \ldots, K$

perchè se $E(\epsilon_t^2|\mathbf{X}) = \sigma^2, \forall t$ allora la variabile casuale ϵ_t^2 è incorrelata sia con gli elementi di $\mathbf{X}$ sia con loro trasformazioni.

Con un modello con due regressori e la costante:

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \epsilon_t$$

la regressione ausiliaria risulta:

$$\widehat{\epsilon}_t^2 = \delta_1 + \delta_2 x_{t2} + \delta_3 x_{t3} + \delta_4 x_{t2}^2 + \delta_5 x_{t3}^2 + \delta_6 x_{t2} x_{t3} + u_t$$

testiamo l'ipotesi nulla che $\delta_2 = \delta_3 = \ldots = \delta_6 = 0$ mediante un test F che per grandi campioni ha una distribuzione approssimata da una distribuzione F.

In generale in un modello con K regressori l'equazione ausiliaria da stimare è:

$$\widehat{\epsilon}_t^2 = \delta_1 + \sum_{k=2}^{K} \delta_k x_{tk} + \sum_{k=2}^{K} \delta_{K-1+j} x_{tj}^2 + \sum_{k=2}^{K-1} \sum_{j=k+1}^{K} \delta_{kj} x_{tk} x_{tj} + u_t$$

ipotesi nulla $\delta_2 = \ldots = 0$. La distribuzione approssimata della statistica F è una $F_{(2(K-1)+\sum_{j=2}^{K-1}(K-j)),N-(2(K-1)+\sum_{j=2}^{K-1}(K-j))}$.

- Regressori ridondanti devono essere omessi (dummy)
- Grande numero di regressori. Potenziale perdita di potenza del test. Per ovviare a questo problema si considerano solo i quadrati delle variabili esplicative, oppure $\widehat{y}_t^2$:

  $$\widehat{\epsilon}_t^2 = \delta_1 + \delta_2 \widehat{y}_t + \delta_3 \widehat{y}_t^2 + \eta_t$$

  ipotesi nulla $\delta_2 = \delta_3 = 0$.

Il test di White non richiede alcuna ipotesi circa la natura e la forma dell'eteroschedasticità.