Chẩn đoán phần dư

Như đã nêu trong một bà i giảng trướ c đây, các phần dư bình phươ ng tối thiểu phản ánh các tính chất của nhữ ng thà nh phần nhiễu ngẫu nhiên. Do điều nà y đúng, nên chúng ta có thể sử dụ ng nhữ ng phần dư nà y để kiểm tra một số trong các giả định đã đưa ra về nhữ ng nhiễu ngẫu nhiên. Là m việc nà y là quan trọ ng vì nhữ ng tính chất thống kê tốt đẹ p của các hà m ướ c lượ ng bình phươ ng tối thiểu phụ thuộc và o các giả định cổ điển.

Trong bà i giảng nà y chúng ta sẽ khảo sát sự tự tươ ng quan giữ a nhữ ng phần dư nà y.

Tự tươ ng quan (Tươ ng quan chuỗi)

Định nghiã

Giả định cổ điển về các sai số ngẫu nhiên không tự tươ ng quan là thế nà y:

$$E[\varepsilon_t \varepsilon_s] = 0 \qquad \text{vớ i mọ i } t \neq s$$

Nếu giả định nà y không đúng, thì các sai số ngẫu nhiên là tự tươ ng quan (hay tươ ng quan chuỗi).

Dễ dà ng xác định nhữ ng hậu quả đối vớ i hà m ướ c lượ ng OLS . Dạ ng tự tươ ng quan đơ n giản nhất là tự tươ ng quan *bậc nhất* :

$$\varepsilon_t = \rho\varepsilon_{t-1} + \xi_t$$

Hệ số ρ đượ c gọ i là hệ số tự tươ ng quan bậc nhất và thà nh phần sai số ngẫu nhiên ξ_t đượ c giả định là thỏa mã n các giả định cổ điển.

Các bậc tự tươ ng quan cao hơ n đượ c trình bà y một cách tươ ng tự :

$$\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \xi_t \qquad \text{(tự tươ ng quan bậc hai)}$$

Tự tươ ng quan nói chung đượ c coi như một hiện tượ ng của chuỗi thờ i gian do thự c tế là nhữ ng tác động của các sai số ngẫu nhiên có thể đượ c phân phối theo thờ i gian. Ngoà i ra , nếu sai số ngẫu nhiên nà y có chứa nhữ ng tác động của các biến thay đổi theo thờ i gian, thì các tác động của chúng có thể đượ c phản ánh trong tự tươ ng quan..

Cung cấp các minh họ a bằng đồ thị .

Chúng ta không kỳ vọ ng tự tương quan sẽ xuất hiện trong dữ liệu chéo, nhưng có các trườ ng hợ p mà trong đó các sai số ngẫu nhiên có thể tỏ ra là "tự tương quan ". Trườ ng hợ p thứ nhất đượ c gọ i là tương quan không gian cù ng thờ i (contemporaneous spatial correlation). Điều nà y có thể xảy ra khi các đơn vị chéo kế bên nhau chia xẻ nhữ ng tác động ngẫu nhiên tương tự . Ví dụ , tất cả các tỉnh bờ biển đều có thể bị tác động tương tự bởi một trận bã o. Ví dụ thứ hai đượ c gọ i là tự tương quan "hiển hiện (apparent)" và có thể là do xác định mô hình sai cho phép hồi qui (một biến hồi qui quan trọ ng có thể đã bị bỏ hay dạ ng hà m số nà y có thể là không phù hợ p).

Các hà m ước lượ ng OLS khi hiện diện tự tương quan

Xét mô hình hồi qui đơn biến vơi tự tương quan bậc nhất:

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t \quad \text{trong đó} \quad \varepsilon_t = \rho\varepsilon_{t-1} + \xi_t$$

Chúng ta yêu cầu $-1 < \rho < 1$ để cho mô hình ổn định.

Trong nhữ ng trườ ng hợ p nà y, dễ dà ng chứng minh rằng các hà m ước lượ ng OLS là không chệch. Cũ ng dễ dà ng tính các phương sai của chúng, và chúng ta tìm ra rằng nhũ ng biểu thức dà nh cho nhữ ng phương sai nà y khác với nhữ ng gì đượ c tự động tính bởi các chương trình OLS.

Như trước đây, vấn đề nà y xuất hiện là liệu có phải các hà m ước lượ ng OLS giữ lạ i đượ c tính chất phương sai tối thiểu của chúng. Dễ dà ng chứng minh rằng chúng không giữ đượ c .

Lấy phương trình đầu trừ phương trình thứ hai sau đây:

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t$$

$$\rho Y_{t-1} = \rho\beta_1 + \beta_2 \rho X_{t-1} + \rho\varepsilon_{t-1}$$

$$\Rightarrow \quad Y_t - \rho Y_{t-1} = (1-\rho)\beta_1 + \beta_2(X_t - \rho X_{t-1}) + \xi_t$$

Do sai số ngẫu nhiên của phương trình hiệu số nà y có các tính chất cổ điển, nên hà m ước lượ ng OLS của hệ số góc nà y sẽ là BLUE. Hà m ước lượ ng nà y trông giống thế nà y:

$$\tilde{\beta}_2 = \frac{\sum(X_t - \rho X_{t-1})(Y_t - \rho X_{t-1})}{\sum(X_t - \rho X_{t-1})^2}$$

Hà m ước lượ ng nà y khác với hà m ước lượ ng OLS, và do hà m nà y là TỐT NHẤT, nên chúng ta biết rằng hà m OLS không thể là tốt nhất.

Do đó, OLS là không chệch nhưng không hiệu quả.

Khó khăn khác trong việc sử dụ ng hà m ước lượ ng OLS là các công thức dà nh cho nhữ ng phương sai của hệ số là không chính xác và hà m ước lượ ng chuẩn của các phương sai sai số là chệch.

Lờ i giải tạ o ra bởi hà m ước lượ ng $\tilde{\beta}_2$ là cơ sở cho phương pháp ước lượ ng Cochrane-Orcutt (một trườ ng hợ p đặc biệt của "các bình phương tối thiểu tổng quát ").

Hà m ước lượ ng $\tilde{\beta}_2$ không thể đượ c tính toán trự c tiếp, vì chúng ta không biết giá trị thự c của ρ. Như chúng ta sẽ thấy , kiểm định cho tự tương quan dự a trên một hà m ước lượ ng cho ρ , nên nếu chúng ta tìm ra tự tương quan, chúng ta cũ ng tìm ra thông tin mà chúng ta cần để tính hà m ước lượ ng Cochrane-Orcutt .

Kiểm định truyền thống : Kiểm định Durbin-Watson

Kiểm định Durbin-Watson cho tự tương quan bậc nhất dự a trên các phần dư bình phương tối thiểu với giả định rằng mô hình hồi qui nà y có một thà nh phần tung độ gốc và rằng nó không có bất cứ giá trị trễ nà o của biến phụ thuộc như là các biến hồi qui (Y_{t-s} không đượ c phép như là một biến giải thích). Ngoà i ra, nhữ ng biến hồi qui nà y đượ c coi như phi ngẫu nhiên, và trong dữ liệu không bị thiếu quan sát nà o.

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

Để phát triển một trự c giác cho việc trị thống kê kiểm định nà y hoạ t động ra sao, hã y nhân đẻ bỏ ngoặc và nhóm lạ i tổng ở tử số:

$$d = \frac{\sum_{t=2}^{T} e_t^2 + \sum_{t=2}^{T} e_{t-1}^2 - 2\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_t^2}$$

Nếu kích cơ mẫu là lơ n, thì các thà nh phần sau gần bằng nhau :

$$\sum_{t=2}^{T} e_t^2 \quad \text{và} \quad \sum_{t=2}^{T} e_{t-1}^2 \quad \text{và} \quad \sum_{t=1}^{T} e_t^2$$

Vì thế chúng ta có thể viết rằng trị thống kê Durbin-Watson xấp xỉ bằng :

$$d \cong 2 \times \left(1 - \frac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_t^2} \right)$$

Hã y nhơ lạ i rằng quá trình tự tương quan bậc nhất trông giống như thế nà y : $\varepsilon_t = \rho\varepsilon_{t-1} + \xi_t$ và thay các thà nh phần sai số ngẫu nhiên không thể quan sát bằng các phần dư OLS :

$$e_t = \rho e_{t-1} + v_t$$

Đây là một mô hình hồi qui đơn biến . Chúng ta có thể ươ c lượ ng hệ số tự tương quan bậc nhất như sau :

$$\hat{\rho} = \frac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=2}^{T} e_{t-1}^2}$$

Nó hoà n toà n gần vơ i thà nh phần trong phép xấp xỉ vơ i trị thống kê Durbin-Watson, nên chúng ta có thể viết:

$$d \cong 2 \times (1 - \hat{\rho})$$

Cách khác

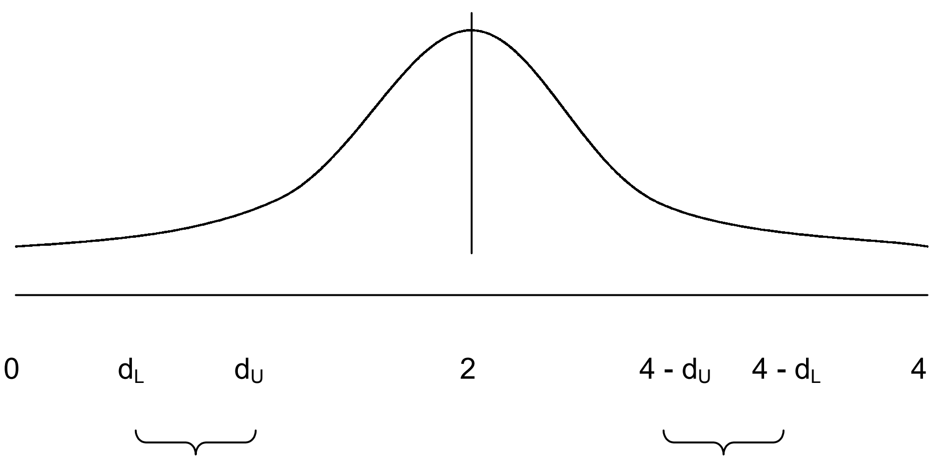
$$\hat{\rho} = 1 - \frac{d}{2}$$

Để giải thích trị thống kê DW , hã y ghi nhận như sau:

$$\hat{\rho} = -1 \Rightarrow d = 4$$
$$\hat{\rho} = 0 \Rightarrow d = 2$$
$$\hat{\rho} = 1 \Rightarrow d = 0$$

Dướ i giả thuyết "không" rằng không có tự tương quan, trị thống kê DW cần phải gần bằng 2.

Các giá trị tớ i hạ n đối vớ i trị thống kê DW đượ c trình bà y thà nh bảng trong các sách kinh tế lượ ng. Tuy nhiên, có một khó khăn ở đó. Nhữ ng giá trị tớ i hạ n nà y phụ thuộc và o bản chất của nhữ ng biến hồi qui, vì thế các trình bà y dướ i dạ ng bảng là không hoà n thiện : chúng chỉ cung cấp các biên trên và dướ i cho nhữ ng giá trị tớ i hạ n nà y. Vù ng giữ a các biên trên và dướ i của các giá trị tớ i hạ n đượ c gọ i là *vù ng vô định.* Nếu d tính toán rơ i và o vù ng vô định nà y, thì chúng ta không thể thể bác bỏ mà cũ ng không thể không bác bỏ giả thuyết "không" nà y.



Các kỹ thuật cù ng thờ i : Tự tương quan và các hà m tự tương quan riêng phần

Xét một cấu trúc tự tương quan bậc P :

$$\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \rho_3\varepsilon_{t-3} + \cdots + \rho_P\varepsilon_{t-P} + \xi_t$$

Hà m tự tương quan nà y là tập hợ p của các ước lượ ng

$$r_p = \frac{\sum_{(t=p+1)}^{R} e_t\, e_{t-p}}{\sum_{(t=1)}^{R} e_t^2}$$

Hà m tự tương quan nà y là đồ thị của r_p với $p = 1 \ldots P$. Như chúng ta sẽ thấy ngay sau đây, EViews rất dễ dà ng tạ o ra hà m nà y và vẽ đồ thị của nó.

Giả thuyết rằng không tồn tạ i tự tương quan cho tới bậc p có thể đượ c kiểm định bằng cách sử dụ ng trị thống kê Q của Ljung-Box :

$$Q_{LB} = T\left(T+2\right)\sum_{j=1}^{p}\frac{r_j^2}{\left(T-j\right)}$$

Dưới giả thuyết "không" về không tồn tạ i tự tương quan , trị thống kê Q có phân phối xác suất tiệm cận với $\chi^2_{(p)}$.

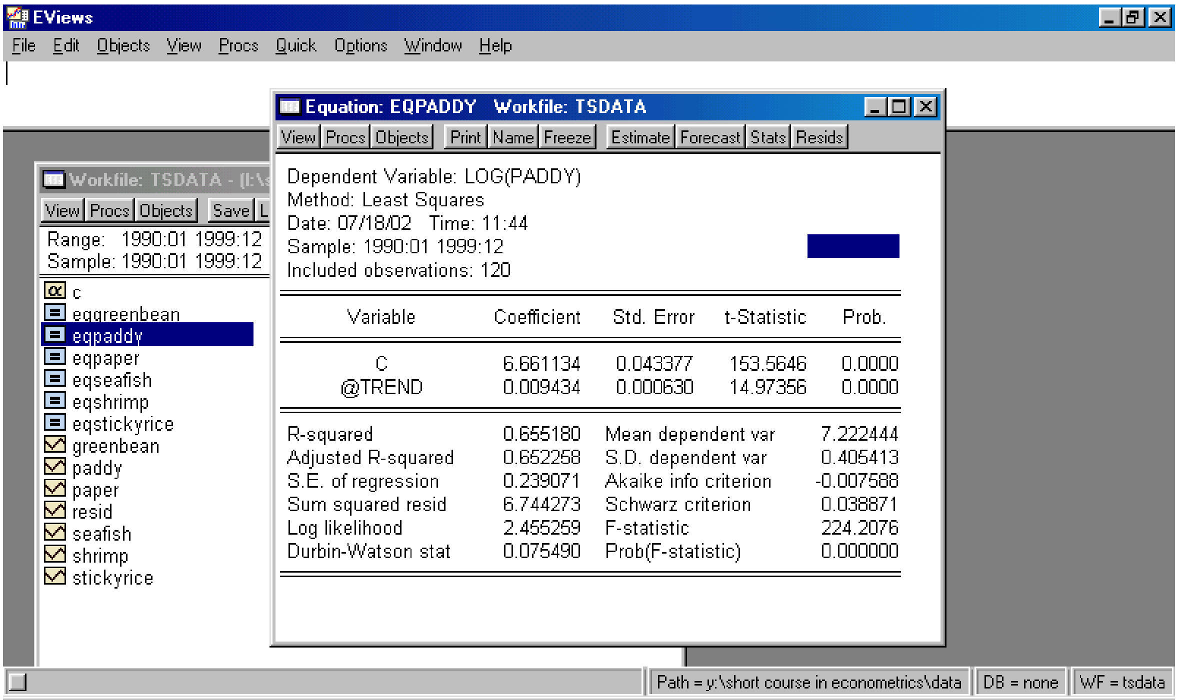
Bây giờ hã y xét xác định mô hình tự hồi qui sau :

$$e_t = \varphi_1 e_{t-1} + \varphi_2 e_{t-2} + \cdots + \varphi_P e_{t-P} + v_t$$

Nhữ ng hệ số tự tương quan riêng phần φ_j $\left(j = 1, \cdots, P \right)$ đo các tương quan giữ a các phần dư cách xa nhau j thờ i đoạ n , trong khi kiểm soát nhữ ng tương quan tạ i các độ trễ khác. Nếu tự tương quan có thể đượ c thâu tóm bởi một phép tự hồi qui bậc thấp hơn P thì phép tự tương quan riêng phần thứ P sẽ gần với zero.

Minh hoạ trên EViews

Đây là phép hồi qui logarit-tuyến tính (tăng trưởng) của giá lúa theo thời gian.



Dependent Variable: LOG(PADDY)
Method: Least Squares
Date: 07/18/02 Time: 11:44
Sample: 1990:01 1999:12
Included observations: 120

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.661134	0.043377	153.5646	0.0000
@TREND	0.009434	0.000630	14.97356	0.0000

R-squared	0.655180	Mean dependent var	7.222444
Adjusted R-squared	0.652258	S.D. dependent var	0.405413
S.E. of regression	0.239071	Akaike info criterion	-0.007588
Sum squared resid	6.744273	Schwarz criterion	0.038871
Log likelihood	2.455259	F-statistic	224.2076
Durbin-Watson stat	0.075490	Prob(F-statistic)	0.000000

Việc nhìn qua trị thống kê Durbin-Watson gợi ý rằng các phần dư hồi qui của chúng ta thể hiện tính tự tương quan đồng biến rất mạnh. Để thu được chi tiết hơn, hãy nhấp theo thứ tự sau :

View / Residual Tests / Corellogram - Q-Statistics

Hãy chọn đưa vào 12 độ trễ (một qui tắc nằm lòng là đưa vào $\sqrt{T}$ độ trễ, trong đó T là kích cỡ mẫu).

Trong trường hợp này, biểu đồ tương quan cung cấp một minh hoạ xuất sắc cho biểu đồ tương quan mô tả quá trình tự tương quan bậc nhất.

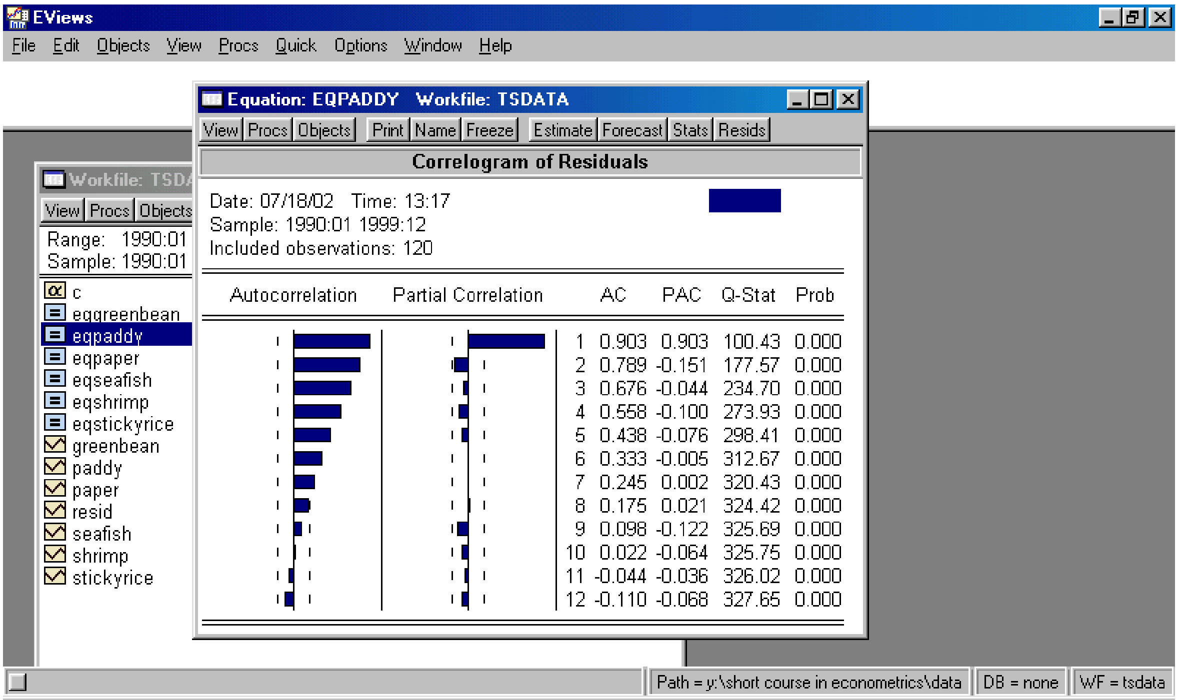
Hệ số tương quan riêng phần thứ nhất có ý nghĩa mạnh, nhưng những hệ số tương quan riêng phần còn lại không có ý nghĩa. Những hệ số tự tương quan giảm theo cấp số nhân như được kỳ vọng. Để thấy vì sao điều này được kỳ vọng, hãy viết

$$\varepsilon_t = \rho\varepsilon_{t-1} + \xi_t$$

và $\varepsilon_{t-1} = \rho\varepsilon_{t-2} + \xi_{t-1}$ và thay biểu thức thứ hai và o biểu thức thứ nhất :

$$\varepsilon_t = \rho^2\varepsilon_{t-2} + \xi_t + \rho\varepsilon_{t-1}$$

Chúng ta thấy rằng hệ số tương quan giữ a ε_t và ε_{t-2} là ρ^2.



Chúng ta cũ ng cần chú ý tới nhữ ng giá trị của trị thống kê Q và nhữ ng giá trị p tương ứng. Do hệ số tương quan thứ nhất nà y có ý nghĩa, nên giả thuyết rằng bất cứ tương quan nà o có thấp hơn thứ p đều là zero bị bác bỏ.

Bây giờ , khi chúng ta đã xác định rằng các nhiễu hồi qui của chúng ta đượ c đặc trưng bởi tự tương quan bậc nhất, chúng ta là m gì đây? Phương pháp truyền thống là ước lượ ng Cochrane-Orcutt; EViews cung cấp một hà m ước lượ ng tương tự nhưng tốt hơn.

Mô hình tự hồi qui bậc nhất

Trước đây, khi chúng ta viết

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t$$

$$\rho Y_{t-1} = \rho\beta_1 + \beta_2 \rho X_{t-1} + \rho\varepsilon_{t-1}$$

$$\Rightarrow \quad Y_t - \rho Y_{t-1} = (1-\rho)\beta_1 + \beta_2 (X_t - \rho X_{t-1}) + \xi_t$$

chúng ta đã khẳng định rằng ướ c lượ ng OLS của phươ ng trình thứ ba sẽ là BLUE. Không may, chúng ta đã không thể là m đượ c việc nà y vì giá trị p là chưa biết.

Một cách tiến hà nh là biến đổi lạ i phương trình nà y như sau :

$$Y_t = \beta_1' - \rho Y_{t-1} + \beta_2 X_t + \beta_3 X_{t-1} + \xi_t$$

Việc thự c hiện OLS trên phươ ng trình nà y cho ta các hà m ướ c lượ ng nhất quán của β_1', β_2, β_3 và ρ .

Hã y ghi nhận rằng chúng ta ướ c lượ ng 4 tham số ở đây, nhưng trong mô hình ban đầu chỉ có ba tham số: β_1, β_2 và ρ .

Chúng ta đã bị mất thông tin đượ c chưá trong hai giớ i hạ n :

$$\beta_1' = (1-\rho)\beta_1 \quad \text{và} \quad \beta_3 = \rho\beta_2$$

Trong trườ ng hợ p nà y, chúng ta có thể yêu cầu EViews thự c hiện một thuật toán bình phươ ng tối thiểu phi tuyến (NLS) áp đặt hai giớ i hạ n nà y :

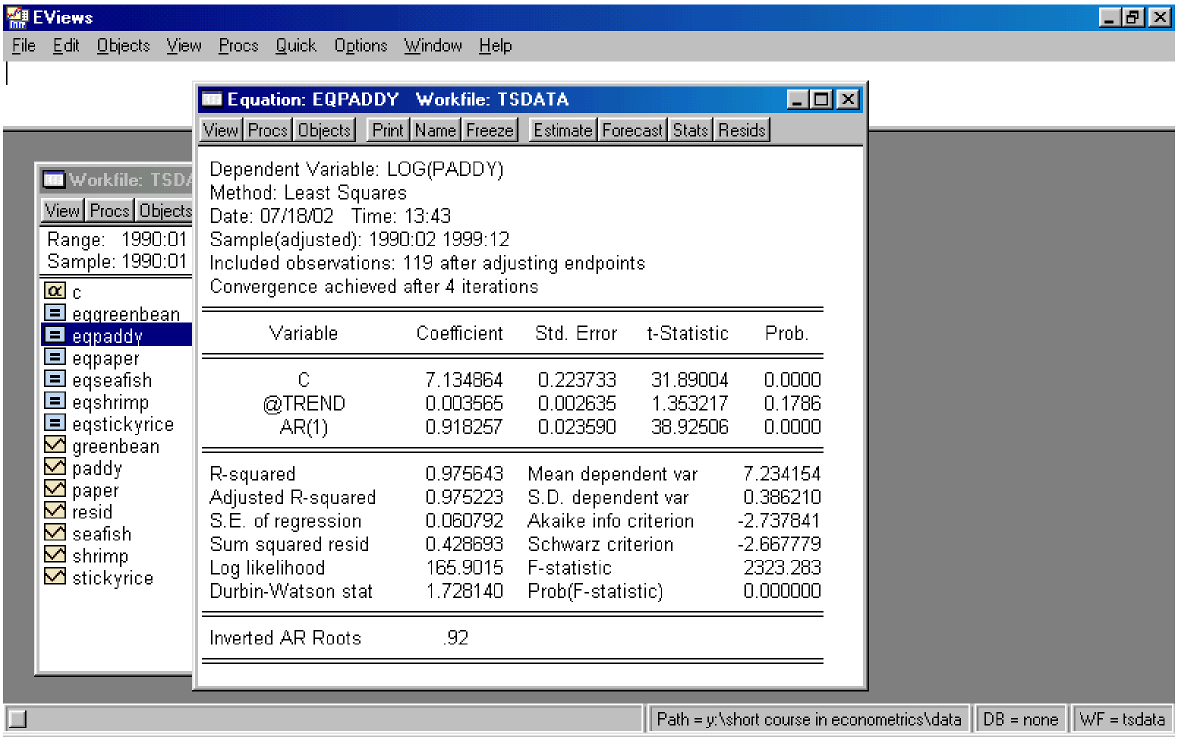
Trong cửa sổ xác định phươ ng trình, hã y đánh máy phươ ng trình nà y như sau đây:

```
log(paddy)  c  @trend  ar(1)
```

Biến ar(1) nói vớ i EViews hã y ướ c lượ ng đặc trưng mô hình tự hồi qui

$$Y_t - \rho Y_{t-1} = (1-\rho)\beta_1 + \beta_2 (X_t - \rho X_{t-1}) + \xi_t$$

tuỳ thuộc và o các giới hạ n đã nêu trên đây. Kết quả cho phương trình log(paddy) là đây :



Dependent Variable: LOG(PADDY)
Method: Least Squares
Date: 07/18/02 Time: 13:43
Sample(adjusted): 1990:02 1999:12
Included observations: 119 after adjusting endpoints
Convergence achieved after 4 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7.134864	0.223733	31.89004	0.0000
@TREND	0.003565	0.002635	1.353217	0.1786
AR(1)	0.918257	0.023590	38.92506	0.0000

R-squared	0.975643	Mean dependent var	7.234154
Adjusted R-squared	0.975223	S.D. dependent var	0.386210
S.E. of regression	0.060792	Akaike info criterion	-2.737841
Sum squared resid	0.428693	Schwarz criterion	-2.667779
Log likelihood	165.9015	F-statistic	2323.283
Durbin-Watson stat	1.728140	Prob(F-statistic)	0.000000

Inverted AR Roots .92

Các trị thống kê t trong kết quả nà y là tiệm cận với chuẩn chuẩn hoá. Chúng ta thấy rằng ước lượ ng điểm nà y của hệ số tự tương quan bậc nhất là 0,94 và răng nó có ý nghĩa thống kê . Trị thống kê F trong kết quả nà y là tiệm cận với Chi bình phương. Chúng ta thấy rằng mức ý nghĩa chung của phép hồi qui nà y rất cao. Cuối cù ng, chúng ta ghi nhận rằng trị thống kê Durbin-Watson đã cải thiện một cách đáng kể.

Để kiểm tra xem có phải là các phần dư của mô hình tự hồi qui bậc nhất của chúng ta bây giờ không cò n tự tương quan, hã y nhấp

View / Residual Tests / Corellogram - Q-Stats

Chúng ta thấy không có các hệ số tương quan hay các hệ số tương quan riêng phần có ý nghĩa, vì thế bây giờ chúng ta tin tưởng rằng các kết quả của chúng ta là không bị xói mò n bởi sự hiện diện của tự tương quan.

Tính chuẩn của các phần dư

Chúng ta có thể sử dụ ng kiểm định Jarque-Bera để kiểmtra tính chuẩn của nhữ ng phần dư nà y . Hã y nhấp

View / Residual Tests / Histogram - Normality Test

Giả thuyết rằng độ nghiêng (skewness) là zero và kurtosis là ba bị bác bỏ mạnh mẽ. Tuy nhiên, biểu đồ tần suất không cho thấy độ nghiêng mạnh, và kích cỡ mẫu là khá lớn, vì thế chúng ta có thể tin tưởng dựa vào định lý giới hạn trung tâm(CLT). CLT khẳng định rằng các hàm ước lượng NLS của chúng ta có phân phối xác suất xấp xỉ chuẩn, rằng các trị thống kê t của chúng ta có phân phối xác suất xấp xỉ chuẩn chuẩn hoá, và rằng các trị thống kê F của Wald có phân phối xác suất xấp xỉ Chi bình phương.

Các biến giả để thâu tóm phá vỡ cấu trúc

Phần này dành cho Bài tập 15

Trong cửa sổ phương trình (Equation) hãy nhấp **Resids**. Ghi nhận rằng dữ liệu này dường như tuân theo một xu hướng dốc hơn nhiều trước năm1993 so với sau đó. Có thể là khó xác định chính xác phá vỡ cấu trúc này, nhưng đối với tôi thì trông như nó xuất hiện trong tháng 6/1991.

Hãy tạo ra một biến giả có giá trị zero trước tháng 6/1991 và có giá trị 1,0 sau tháng 6/1991.

Đầu tiên, hãy tạo ra một biến sẽ tạo điều kiện cho Anh/Chị biết giá trị mà @trend tạo ra.

GENR period = @trend

Bây giờ hãy ghi nhận rằng period = 17 trong tháng 6/1991.

Bây giờ hãy đánh máy lệnh sau vào cửa sổ lệnh:

genr dum = @trend > 17

Hãy xoá giai đoạn của biến để giữ cho workfile của Anh/Chị được gọn gàng.

Kiểm tra biến dum để chắc chắn rằng Anh/Chị đã thu được điều mà Anh/Chị kỳ vọng.

Bây giờ hãy ước lượng một phương trình mới:

Quick / Estimate Equation

log(paddy) c @trend dum @trend*dum

Một phép kiểm tra nhanh những phần dư này xác nhận rằng một lần nữa chúng ta lại có tự tương quan bậc nhất, vì thế chúng ta chạy lại mô hình này như sau :

log(paddy) c @trend dum @trend*dum ar(1)

Qua kiểm tra

View / Residual Tests / Corellogram Q-Statistics

chúng ta tìm ra rằng những phần dư mới không có tự tương quan và qua kiểm tra
View / Residual Tests / White Heteroskedasticity

chúng ta tìm ra (sử dụng dạng chi bình phương của kiểm định này) rằng chúng ta không bác bỏ giả thuyết rằng không có tính không đồng nhất của phương sai .