CHƯƠNG 10: ỨNG DỤNG MÔ HÌNH PHÂN TÍCH HỒI QUY TRONG DỰ BÁO

Xét lượng cam bán (Y: tạ/ngày) theo giá cam (X: ngàn đ/kg) ta có bảng số liệu sau:

Y	X
14	2
10	2
12	3
7	4
8	5
9	5
6	5
6	6

Y01
Y02
...
Y0s

$X_0 = 4,5$

Ta thấy trong dữ liệu không có giá cam là 4,5 ngàn đ/kg. Một câu hỏi tự nhiên là: nếu giá cam là $X = X_0 = 4,5$ ngàn đ/kg thì lượng cam bán là bao nhiêu?

Ta không thể trả lời: chờ tôi một tý, để tôi chạy ra chợ hỏi bà bán, xem nếu bà bán giá 4,5 ngàn đ/kg thì bà bán được bao nhiêu tạ/ngày.

Từ mẫu đã có, ta phải trả lời câu hỏi này.

Ta phải ước lượng (dự đoán) được lượng cam bán sẽ là bao nhiêu nếu giá cam là $X = X_0 = 4,5$ ngàn đ/kg.

Ta thấy cùng mức giá thì lượng cam bán sẽ khác nhau. Thí dụ: cùng giá bán, lượng cam bán ngày hôm trước khác ngày hôm sau. Cùng ngày bán, nơi bán khác nhau thì lượng cam bán sẽ khác nhau. Cùng ngày bán, cùng nơi bán, bà già sẽ bán ít/ nhiều hơn thiếu nữ.

Túm lại, ứng với X_0 sẽ có nhiều giá trị của Y, ký hiệu là Y01, Y02, ..., Y0s.

Ta có: $E(Y/X=X0)=E(Y/X0)=\frac{1}{s}\sum_{j=1}^{s} Y_{0j}$

E(Y/X0) =? : *dự báo (ước lượng) giá trị trung bình*

Y0= (Y01,...,Y0s) =? : *dự báo giá trị cá biệt*

Ta có 2 dạng ước lượng là UL điểm và UL khoảng nên dự báo có 2 dạng: *dự báo điểm* và *dự báo khoảng*.

Dự báo với mô hình 2 biến.

Dự báo với mô hình nhiều biến.

10.1 DỰ BÁO VỚI MÔ HÌNH HAI BIẾN

Xét mô hình hồi quy hai biến sau:

$$PRF\begin{cases}E(Y/X)=\alpha+\beta X\\ Y=\alpha+\beta X+U\end{cases};\quad SRF\begin{cases}\hat{Y}_i=\hat{\alpha}+\hat{\beta}X_i\\ Y_i=\hat{\alpha}+\hat{\beta}X_i+\hat{U}_i\end{cases}$$

Một khi mô hình ước lượng *SRF* được xác định là phù hợp tốt, ta có thể dùng để dự báo giá trị trung bình $E(Y/X)$ hay giá trị cá biệt Y.

10.1.1 Dự báo điểm *(Point Prediction)*

Cho $X = X_0$ cho trước.

Dự báo giá trị trung bình chính là dự báo cho $E(Y/X=X_0)$

Dự báo cá biệt ký hiệu là Y_0

Khi thay X_0 vào hàm *SRF*, ta thu được $\hat{Y}_0=\hat{\alpha}+\hat{\beta}X_0$

ta chứng minh được $\hat{Y}_0$ là ước lượng tuyến tính, không chệch tốt nhất của $E(Y/X_0)$ và Y_0, do đó người ta sử dụng $\hat{Y}_0$ là dự báo điểm cho cả giá trị trung bình và giá trị cá biệt của biến phụ thuộc Y.

10.1.2 Dự báo khoảng *(Interval Prediction)*

Để dự báo khoảng, người ta cũng phải căn cứ vào dự báo điểm $\hat{Y}_0$.

Cần lưu ý rằng về bản chất, $\hat{Y}_0$ cũng là đại lượng ngẫu nhiên, vì nó phụ thuộc vào các đại lượng ngẫu nhiên $\hat{\alpha}$, $\hat{\beta}$.

Người ta cũng chứng minh được rằng $\hat{Y}_0$ có phân phối chuẩn dưới giả thiết thành phần nhiễu có phân phối chuẩn, với kỳ vọng là $E(Y/X_0)=\alpha+\beta X_0$ và phương sai xác định bằng biểu thức:

$$var(\hat{Y}_0)=\sigma^2\left[\frac{1}{n}+\frac{(X_0-\overline{X})^2}{\sum x_i^2}\right] \qquad (10.1)$$

$$se(\hat{Y}_0)=\sqrt{var(\hat{Y}_0)} \qquad (10.2)$$

10.1.2.1 Dự báo trung bình (Mean Prediction) $E(Y/X_0)$

Với tính chất $\hat{Y}_0$ có phân phối chuẩn, và sử dụng $\hat{\sigma}^2$ là ước lượng không chệch cho phương sai tổng thể σ^2, thì lúc đó đại lượng ngẫu nhiên t xác định bằng biểu thức

$$t = \frac{\hat{Y}_0 - E(Y/X_0)}{se(\hat{Y}_0)} \qquad (10.3)$$

có phân phối *t-student* với $(n-2)$ bậc tự do, trong đó n là số quan sát của mẫu, 2 chính là số tham số có trong mô hình hồi quy.

Với độ tin cậy $(1-\alpha)$, KTC của giá trị trung bình $E(Y/X_0)$ là:

$$\hat{Y}_0 \pm t_{\alpha/2}^{(n-2)}.se(\hat{Y}_0) \qquad (10.4)$$

Ta có sai số của dự báo $\hat{U}_0 = Y_0 - \hat{Y}_0$ là đại lượng ngẫu nhiên có phương sai xác định như sau:

$$\text{var}(\hat{U}_0) = \text{var}(Y_0 - \hat{Y}_0) = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right]$$

$$= \text{var}(\hat{Y}_0) + \text{var}(U_0) = \text{var}(\hat{Y}_0) + \sigma^2 \qquad (10.5)$$

Khi dùng $\hat{\sigma}^2$ thay thế cho σ^2, ta có:

$$t = \frac{Y_0 - \hat{Y}_0}{se(\hat{U}_0)} \qquad (10.6)$$

có phân phối *t-student* với $(n-2)$ bậc tự do.

Với độ tin cậy $(1-\alpha)$, KTC của giá trị cá biệt cho biến phụ thuộc là:

$$\hat{Y}_0 \pm t_{\alpha/2}^{(n-2)}.se(\hat{U}_0) \qquad (10.7)$$

Thí dụ 10.4.1: Dự báo cho mô hình hồi quy hai biến

Ở tiểu bang *New York* của Mỹ, phần đông dân cư ưa thích trò chơi xổ số có tên là *Lotto*. Giả sử người ta muốn khảo sát mối quan hệ giữa thu nhập và số tiền chơi *Lotto* trong dân cư, bằng cách thiết lập mẫu quan sát, trong đó Y là số tiền mua xổ số trong tuần, X là thu nhập khả dụng (sau thuế) trong tuần, tất cả được tính bằng đô la (bảng 10.1).

Equation: EQ01 Workfile: ...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 05/28/07 Time: 21:26
Sample (adjusted): 1 10
Included observations: 10 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7.618182	3.052340	2.495850	0.0372
X	0.081455	0.011216	7.262425	0.0001

R-squared	0.868297	Mean dependent var	29.00000
Adjusted R-squared	0.851834	S.D. dependent var	6.616478
S.E. of regression	2.546834	Akaike info criterion	4.884436
Sum squared resid	51.89091	Schwarz criterion	4.944953
Log likelihood	-22.42218	F-statistic	52.74282
Durbin-Watson stat	3.039473	Prob(F-statistic)	0.000087

Với tổng số quan sát của mẫu $n = 10$, ta có:

$$\bar{X} = \sum X_i / n = 2625/10 = 262.5;\ \bar{Y} = \sum Y_i / n = 290/10 = 29$$

$$\hat{\beta} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n(\bar{X})^2} = \frac{80325 - 10.(262.5).(29)}{740625 - 10.(262.5)^2} = \frac{4200}{51562.5} = 0.081455$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\,\bar{X} = 29 - (0.081455).(262.5) = 7.618182$$

Vậy hàm hồi quy ước lượng là

$$Y_i = 7.618182 + 0.081455X_i + \hat{U}_i$$

Giả sử ta muốn sử dụng mô hình hồi quy này để dự báo mức chi tiêu mua *Lotto* trung bình của những người có mức thu nhập khả dụng là 340 $/tuần, ký hiệu là $E(Y/X_0 = 340)$, với độ tin cậy là 95%. Trình tự tính toán như sau:

$$TSS = \sum Y_i^2 - n(\bar{Y})^2 = 8804 - 10.(29)^2 = 394$$

$$ESS = \hat{\beta}^2\left(\sum X_i^2 - n\bar{X}^2\right) = \hat{\beta}^2 \sum x_i^2 = (0.081455)^2(740625 - 10.(262.5)^2)$$

$$= (0.081455)^2(51562.5) = 342.1091$$

$$RSS = TSS - ESS = 394 - 342.1091 = 51.89091$$

$$\hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{51.89091}{8} = 6.486364$$

Ta có ước lượng điểm của $E(Y/X_0 = 340)$ chính là $\hat{Y}_0$:

$$\hat{Y}_0 = 7.618182 + 0.081455X_0 = 7.618182 + 0.081455.(340) = 35.31273$$

$$var(\hat{Y}_0) = \sigma^2\left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right]$$

$$\approx 6.486364\left[\frac{1}{10} + \frac{(340 - 262.5)^2}{51562.5}\right] = 1.404199$$

$$se(\hat{Y}_0) = \sqrt{\text{var}(\hat{Y}_0)} = \sqrt{1.404199} = 1.184989$$

Với độ tin cậy 95%, tra bảng *t-student* ta có $t_{\alpha/2}^{(n-2)} = t_{0.025}^{8} = 2.306$

Vậy khoảng tin cậy 95% của $E(Y/X_0 = 340)$ là:

$$\hat{Y}_0 \pm t_{\alpha/2}^{(n-2)}.se(\hat{Y}_0) = (35.31273 \pm (2.306).(1.184989))$$

Hay (32.58014 ; 38.04531)

Giả sử ta muốn dự báo số tiền mua *Lotto* của một người có thu nhập 340 $/tuần (dự báo cá biệt) cùng với độ tin cậy 95%, thì cần tính $\text{var}(\hat{U}_0)$:

$$\text{var}(\hat{U}_0) = \text{var}(Y_0 - \hat{Y}_0) = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right]$$

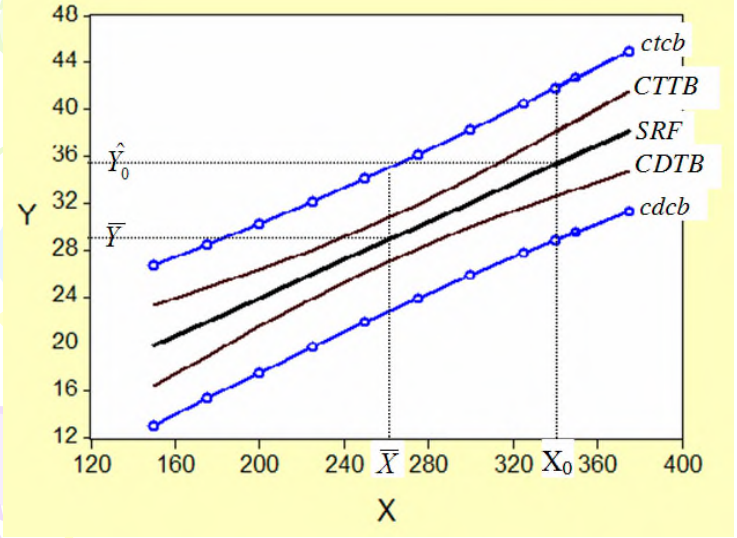
$$= \text{var}(\hat{Y}_0) + \sigma^2 \approx 1.404199 + 6.486364 = 7.890563$$

$$se(\hat{U}_0) = \sqrt{\text{var}(\hat{U}_0)} = \sqrt{7.890563} = 2.809015$$

Vậy, khoảng tin cậy 95% của $Y_0/X_0=340$ là :

$$\hat{Y}_0 \pm t_{\alpha/2}^{(n-2)}.se(\hat{U}_0) = (35.31273 \pm (2.306).(2.809015))$$

Hay (28.83514 ; 41.79031)



Quan sát hình vẽ, nhận thấy:

-Giá trị X_0 nằm trong khoảng biến thiên của số liệu ($150<X_0<375$) thì kết quả dự báo mới đáng tin cậy, X_0 càng gần $\bar{X}$ thì kết quả dự báo càng chính xác.

-Do đó khoảng dự báo hẹp nhất (và do đó có độ chính xác cao nhất) là tại trung bình mẫu $X_0 = \bar{X}$.

10.2 DỰ BÁO VỚI MÔ HÌNH NHIỀU BIẾN

Giả sử mô hình hồi quy nhiều biến dạng ma trận như sau:

PRF ngẫu nhiên : $Y = X.\beta + U$

SRF ngẫu nhiên : $Y = X.\hat{\beta} + \hat{U}$

10.2.2 Dự báo khoảng

10.2.2.1 Dự báo trung bình

Để thực hiện dự báo cho giá trị trung bình, ta cần ước lượng phương sai của giá trị dự báo điểm $\hat{Y}_0$ bằng công thức:

$$\text{var}\left(\hat{Y}_0\right) = X^{0T}.\text{cov}\left(\hat{\beta}\right)X^0 = \sigma^2 X^{0T}.\left(X^T X\right)^{-1}.X^0 \qquad (10.8)$$

$$se\left(\hat{Y}_0\right) = \sqrt{\text{var}\left(\hat{Y}_0\right)} \qquad (10.9)$$

Trong đó $\text{cov}\left(\hat{\beta}\right)$ là ma trận hiệp phương sai của hệ số hồi quy xác định bằng công thức (4.10).

độ tin cậy (1-α), khoảng tin cậy của giá trị trung bình $E(Y/X_0)$ là:

$$\hat{Y}_0 \pm t_{\alpha/2}^{(n-k)}.\hat{se}\left(\hat{Y}_0\right) \qquad (10.10)$$

10.2.2.2 Dự báo cá biệt

Sai số của dự báo trong mô hình nhiều biến được tính bằng công thức:

$$\text{var}\left(\hat{U}_0\right) = \text{var}\left(\hat{Y}_0\right) + \sigma^2 = \sigma^2\left[X^{0T}.\left(X^T X\right)^{-1}.X^0 + 1\right] \qquad (10.11)$$

$$se\left(\hat{U}_0\right) = \sqrt{\text{var}\left(\hat{U}_0\right)} \qquad (10.12)$$

độ tin cậy (1-α), ta có thể ước lượng khoảng giá trị cá biệt cho biến phụ thuộc bằng biểu thức:

$$\hat{Y}_0 \pm t_{\alpha/2}^{(n-k)}.\hat{se}\left(\hat{U}_0\right) \qquad (10.13)$$

Thí dụ 10.4.2: Dự báo cho mô hình hồi quy nhiều biến

Tiếp theo thí dụ 5.1 trong chương 5,

Y : lượng hàng bán được (tấn/tháng)

X : giá hàng (ngàn đ/kg)

D= 0 : thành phố , D=1 : nông thôn

dự báo lượng hàng bán được trung bình của một cửa hàng ở thành phố, khi giá bán là 2.9 ngàn đồng/kg, với độ tin cậy 95%.

$$TSS = Y^T.Y - n\left(\bar{Y}\right)^2 = \sum Y_i^2 - n\left(\bar{Y}\right)^2$$

$$= 3065.04 - 10\left(\frac{174.6}{10}\right)^2 = 16.524$$

$$ESS = \hat{\beta}^T.X^T.Y - n\left(\bar{Y}\right)^2 = \begin{bmatrix}25.6633 & 0.407582 & -3.45971\end{bmatrix}\begin{bmatrix}174.6\\87.8\\420.53\end{bmatrix}$$

$$-10\left(\frac{174.6}{10}\right)^2 = 13.17077$$

$$RSS = TSS - ESS = 3.353227\ ;\ \hat{\sigma}^2 = \frac{RSS}{n-k} = \frac{RSS}{n-3} = 0.479032$$

$$X^{0T} = \begin{pmatrix}1 & D_0 & X_0\end{pmatrix} = \begin{pmatrix}1 & 0 & 2.9\end{pmatrix}$$

$$\hat{Y}_0 = X^{0T}.\hat{\beta} = \hat{\beta}^T.X^0 = \begin{pmatrix}1 & 0 & 2.9\end{pmatrix}\begin{bmatrix}25.6633\\0.407582\\-3.45971\end{bmatrix} = 15.63015$$

$$Var\left(\hat{Y}_0\right) = \sigma^2 X^{0T}\left(X^T X\right)^{-1}X^0 \approx \hat{\sigma}^2 X^{0T}\left(X^T X\right)^{-1}X^0$$

$$= \frac{0.479032}{27.3}\begin{bmatrix}1 & 0 & 2.9\end{bmatrix}\begin{pmatrix}149.46 & -1.86 & -60\\-1.86 & 11.01 & -1.5\\-60 & -1.5 & 25\end{pmatrix}\begin{bmatrix}1\\0\\2.9\end{bmatrix} = 0.205475$$

$$\Rightarrow se\left(\hat{Y}_0\right) = \sqrt{Var\left(\hat{Y}_0\right)} = 0.453294$$

Với độ tin cậy 95% $\Rightarrow \alpha = 0.05 \Rightarrow t_{\alpha/2}^{(n-3)} = t_{0.025}^{7} = =2.365$

Ta có khoảng tin cậy của E(Y/X0) :

$$\left(\hat{Y}_0 - t_{\alpha/2}.se\left(\hat{Y}_0\right) ;\ \hat{Y}_0 + t_{\alpha/2}.se\left(\hat{Y}_0\right)\right)$$

$$\sim \left(15.63015 \pm 2.365 \text{ x } 0.453294\right) \sim \left(14.43572 ;\ 16.70219\right)$$

10.3 ĐÁNH GIÁ ĐỘ CHÍNH XÁC CỦA DỰ BÁO

10.3.1 Phân chia mẫu

Một trong những tiêu chuẩn để đánh giá mô hình tốt theo tiêu chuẩn của *Harvey* là mô hình có khả năng dự báo chính xác.

việc đánh giá mức độ chính xác trong dự báo của mô hình hồi quy đòi hỏi *phải có số liệu thực tế để đối chiếu với giá trị dự báo từ mô hình*. Điều này có thể thực hiện bằng cách thu thập thêm số liệu mới, nhưng thực tế việc thu thập thêm số liệu mới không phải lúc nào cũng dễ dàng thực hiện được.

Do đó người ta giải quyết bằng cách phân chia mẫu, nghĩa là từ mẫu đang có tách thành hai mẫu con.

Mẫu con thứ nhất được sử dụng để ước lượng mô hình hồi quy và gọi là "mẫu khởi động" (*initialization set*).

Mẫu con thứ hai được sử dụng để kiểm tra độ chính xác của các giá trị dự báo từ mô hình hồi quy tìm được từ mẫu khởi động. Mẫu con thứ hai được gọi là "mẫu kiểm tra" (*test set*).

Ta cần cân nhắc việc xác định mẫu kiểm tra sao cho không làm thay đổi nhiều đến kết quả hồi quy dựa trên mẫu khởi động, đồng thời có đủ số quan sát cần thiết trong mẫu kiểm tra để đánh giá được khả năng dự báo của mô hình.

10.3.2 Tiêu chuẩn đo lường thống kê của dự báo

Giả sử mẫu kiểm tra gồm m quan sát,

trong đó ký hiệu Y_i là giá trị thực tế của biến phụ thuộc Y;

$\hat{Y}_i$ là giá trị dự báo điểm của mô hình hồi quy;

$\hat{U}_i = Y_i - \hat{Y}_i$ là sai số của dự báo.

Đánh giá khả năng dự báo của mô hình được dựa trên các sai số dự báo trong mẫu kiểm tra mà không dựa trên mẫu khởi động vì thực tế khi xây dựng mô hình hồi quy, người ta đã tìm cách cực tiểu các phần dư trong mẫu khởi động để xác định các tham số ước lượng.

Vấn đề ở đây là kết quả ước lượng có còn khớp (*fitted*) với các quan sát ngoài mẫu khởi động hay không?

Ta có các tiêu chuẩn đo lường thống kê như sau:

- Sai số trung bình *ME* (*mean error*)

$$ME = \frac{1}{m}\sum_{i=1}^{m}\hat{U}_i \qquad (10.14)$$

- Sai số tuyệt đối trung bình *MAE* (*mean absolute error*)

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|\hat{U}_i\right| \qquad (10.15)$$

- Sai số bình phương trung bình *MSE* (*mean squared error*)

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\hat{U}_i^2 \qquad (10.16)$$

- Căn bậc hai của sai số bình phương trung bình *RMSE*

$$RMSE = \sqrt{MSE} \qquad (10.17)$$

Các chỉ số trên đều phụ thuộc vào đơn vị đo của biến, do đó việc đánh giá các chỉ số trên lớn hay nhỏ không chỉ chú ý thuần túy về mặt giá trị mà còn phải quan tâm đến đơn vị của biến.

Các đo lường thống kê *ME*, *MAE*, *MSE*, *RMSE* chỉ có ý nghĩa khi được đối chiếu hay so sánh giữa các mô hình hồi quy (cùng dạng biến phụ thuộc và cùng cỡ mẫu), hay nói cách khác, việc phân tích độc lập các giá trị của những chỉ số này ít có ý nghĩa.

Các tiêu chuẩn còn lại có thể được phân tích để xem xét khả năng dự báo của một mô hình hồi quy có tốt hay không.

Ngoài ra, người ta còn so sánh các tiêu chuẩn đo lường thống kê giữa các mô hình hồi quy khác nhau nhằm mục đích lựa chọn được mô hình có khả năng dự báo tốt nhất.

các tiêu chuẩn đo lường thống kê không phụ thuộc vào đơn vị đo của biến:

- Sai số phần trăm (tương đối) *PE* (*percentage error*):

$$PE_i = \left(\frac{Y_i - \hat{Y}_i}{Y_i}\right).100 \qquad (10.18)$$

- Sai số phần trăm trung bình *MPE* (*mean percentage error*):

$$MPE = \frac{1}{m}\sum_{i=1}^{m}PE_i \qquad (10.19)$$

- Sai số phần trăm tuyệt đối trung bình *MAPE* (*mean absolute percentage error*):

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left|PE_i\right| \qquad (10.20)$$

Ngoài ra ta còn có hệ số bất đẳng thức *Theil* (*Theil Inequality Coefficient*) như sau:

$$TIC = \frac{\sqrt{\sum_{i=1}^{m}\left(\hat{Y}_i - Y_i\right)^2/m}}{\sqrt{\sum_{i=1}^{m}\hat{Y}_i^2/m} + \sqrt{\sum_{i=1}^{m}Y_i^2/m}} \qquad (10.21)$$

Hệ số *TIC* có giá trị trong [0,1]. Khi *TIC*=0, tức sai lệch giữa giá trị dự báo điểm với giá trị thực tế bằng 0, khi đó hàm hồi quy dự báo chính xác hoàn toàn. Trong thực tế hiếm khi có được giá trị lý tưởng *TIC*=0, mà chỉ kỳ vọng *TIC* càng gần 0 thì càng tốt.

Ta có thể phân tích tử số thành các thành phần sau:

$$\sum\left(\hat{Y}_i - Y_i\right)^2/m = \left(\bar{\hat{Y}} - \bar{Y}\right)^2 + \left(s_{\hat{Y}} - s_Y\right)^2 + 2\left(1 - R_{\hat{Y}Y}\right)s_{\hat{Y}}s_Y$$

Trong đó:

$\bar{\hat{Y}} = \dfrac{\sum_{i=1}^{m}\hat{Y}_i}{m}$: giá trị trung bình của dự báo điểm trong mẫu kiểm tra;

$\bar{Y} = \dfrac{\sum_{i=1}^{m}Y_i}{m}$: giá trị trung bình thực tế trong mẫu kiểm tra;

$s_{\hat{Y}} = \sqrt{\dfrac{\sum\left(\hat{Y}_i - \bar{\hat{Y}}\right)^2}{m}}$: độ lệch chuẩn của giá trị dự báo trong mẫu kiểm tra;

$s_Y = \sqrt{\dfrac{\sum\left(Y_i - \bar{Y}\right)^2}{m}}$: độ lệch chuẩn của giá trị thực tế trong mẫu kiểm tra;

$R_{\hat{Y}Y} = \dfrac{\sum\left(\hat{Y}_i - \bar{\hat{Y}}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum\left(\hat{Y}_i - \bar{\hat{Y}}\right)^2 . \sum\left(Y_i - \bar{Y}\right)^2}}$: hệ số tương quan giữa các giá trị dự báo và giá trị thực tế trong mẫu kiểm tra.

Ta có các tỷ lệ tương ứng như sau:

Tỷ lệ chệch (*Bias Proportion*): cho biết trung bình các giá trị dự báo khác biệt như thế nào so với trung bình các giá tri thực tế.

$$\frac{\left(\bar{\hat{Y}} - \bar{Y}\right)^2}{\sum\left(\hat{Y}_i - Y_i\right)^2/m} \qquad (10.22)$$

Tỷ lệ phương sai (*Variance Proportion*): cho biết mức độ biến thiên của các giá trị dự báo khác biệt như thế nào so với độ biến thiên của các giá trị thực tế.

$$\frac{\left(s_{\hat{Y}} - s_Y\right)^2}{\sum\left(\hat{Y}_i - Y_i\right)^2/m} \qquad (10.23)$$

Tỷ lệ hiệp phương sai (*Covariance Proportion*): cho biết tỉ lệ phần sai số của dự báo không mang tính hệ thống.

$$\frac{2\left(1 - R_{\hat{Y}Y}\right)s_{\hat{Y}}s_Y}{\sum\left(\hat{Y}_i - Y_i\right)^2/m} \qquad (10.24)$$

Tỷ lệ chệch, tỷ lệ phương sai và tỷ lệ hiệp phương sai có tổng luôn bằng 1.

Nếu dự báo là tốt, tỷ lệ chệch và tỷ lệ phương sai sẽ có khuynh hướng nhỏ, và như vậy phần lớn sai số trong dự báo sẽ thuộc về tỷ lệ hiệp phương sai, là phần đo lường thể hiện tính chất không hệ thống (không quy luật).

10.3.3 Dự báo ngoài mẫu

Nếu mục đích chỉ là để kiểm tra khả năng dự báo của mô hình thì giá trị biến độc lập (*X0*) được sử dụng để dự báo được lấy từ trong mẫu kiểm tra.

Tuy nhiên ứng dụng của phân tích hồi quy là sử dụng mô hình hồi quy để dự báo cho biến phụ thuộc – dự báo ngoài phạm vi mẫu phân tích.

Một mô hình hồi quy sau khi tiến hành dự báo trong mẫu nhằm mục đích đánh giá khả năng dự báo chính xác của mô hình có thể được vận dụng để dự báo ngoài mẫu.

Thí dụ 10.4.3: Dự báo bằng cách sử dụng phần mềm Eviews

Sử dụng lại bảng dữ liệu 4.4 trong thí dụ ở chương 4 về :
tỉ lệ tử vong ở trẻ sơ sinh (*CM*),
tỉ lệ phần trăm phụ nữ biết đọc-viết (*FLR*),
thu nhập quốc gia theo đầu người (*PGNP*),
và tỉ lệ sinh đẻ trung bình của một phụ nữ (*TFR*)
của nhóm gồm 64 quốc gia.
sử dụng 54 quan sát đầu tiên làm mẫu khởi động (*initialization set*)
10 quan sát cuối cùng để làm mẫu kiểm tra (*test set*).

Trong thí dụ 4.2, chúng ta đề cập đến hai mô hình sau:

$$CM_i = \beta_0 + \beta_1 FLR_i + \beta_2 PGNP_i + \beta_3 TFR_i + U_i \qquad (4.20)$$

$$CM_i = \beta_0 + \beta_1 FLR_i + \beta_2 TFR_i + U_i \qquad (4.21)$$

Mục tiêu của chúng ta bây giờ là đánh giá năng lực dự báo của hai mô hình trên sử dụng các tiện ích của phần mềm *Eviews*.

Các bước tiến hành như sau:

- Bước 0: xác định phạm vi mẫu khởi động.
- Bước 1: hồi quy các mô hình (4.20), (4.21) trên mẫu khởi động gồm 54 quan sát.
- Bước 2: tính giá trị dự báo điểm của hai mô hình (là $\hat{Y}_i$)
- Bước 3: xác định phạm vi mẫu kiểm tra.
  (bước 2, 3: dùng chức năng Forecast của Eviews)
- Bước 4: tính các sai số (là $\hat{U}_i = Y_i - \hat{Y}_i$) giữa giá trị dự báo điểm (là $\hat{Y}_i$) và giá trị thực tế (là Y_i) trên mẫu kiểm tra, áp dụng các công thức từ (10.14) đến (10.24) đối với từng mô hình hồi quy.
- Bước 5: so sánh độ chính xác của dự báo giữa hai mô hình.

Equation: EQ03_KD Workfile...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: CM
Method: Least Squares
Date: 05/28/07 Time: 23:49
Sample: 1 54
Included observations: 54

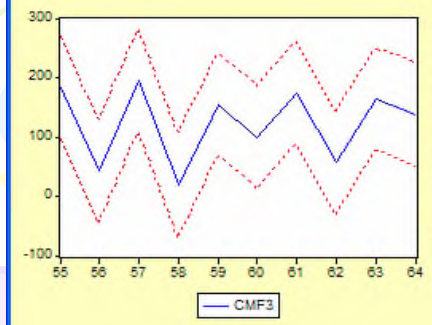
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	162.1708	38.05521	4.261462	0.0001
FLR	-1.930380	0.279037	-6.918000	0.0000
TFR	14.35270	4.916006	2.919585	0.0052

R-squared	0.717037	Mean dependent var	146.8704
Adjusted R-squared	0.705940	S.D. dependent var	78.28155
S.E. of regression	42.44994	Akaike info criterion	10.38848
Sum squared resid	91901.86	Schwarz criterion	10.49898
Log likelihood	-277.4890	F-statistic	64.61781
Durbin-Watson stat	2.161525	Prob(F-statistic)	0.000000

41

Graph: GRAPH3 Workfile: T...

View | Proc | Object | Print | Name | AddText | Line/Shade | Remove | Template | Options | Zoom



Forecast: CMF3
Actual: CM
Forecast sample: 55 64
Included observations: 10

Root Mean Squared Error	36.86195
Mean Absolute Error	24.84072
Mean Abs. Percent Error	28.69349
Theil Inequality Coefficient	0.141319
Bias Proportion	0.068956
Variance Proportion	0.029051
Covariance Proportion	0.901993

42

Equation: EQ04_KD Workfile...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: CM
Method: Least Squares
Date: 05/28/07 Time: 23:51
Sample: 1 54
Included observations: 54

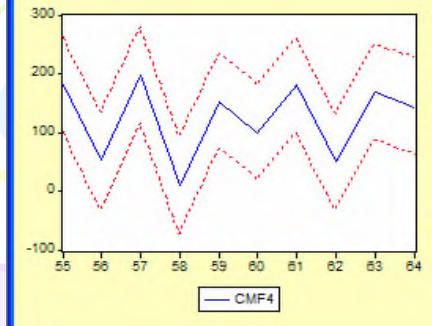
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	162.3474	35.71079	4.546172	0.0000
FLR	-1.782755	0.267052	-6.675694	0.0000
TFR	14.37367	4.613151	3.115802	0.0030
PGNP	-0.005452	0.001938	-2.813594	0.0070

R-squared	0.755714	Mean dependent var	146.8704
Adjusted R-squared	0.741057	S.D. dependent var	78.28155
S.E. of regression	39.83471	Akaike info criterion	10.27854
Sum squared resid	79340.22	Schwarz criterion	10.42587
Log likelihood	-273.5206	F-statistic	51.55936
Durbin-Watson stat	2.154986	Prob(F-statistic)	0.000000

43

Graph: GRAPH4 Workfile: T...

View | Proc | Object | Print | Name | AddText | Line/Shade | Remove | Template | Options | Zoom



Forecast: CMF4
Actual: CM
Forecast sample: 55 64
Included observations: 10

Root Mean Squared Error	36.21988
Mean Absolute Error	22.52055
Mean Abs. Percent Error	27.98974
Theil Inequality Coefficient	0.137640
Bias Proportion	0.092642
Variance Proportion	0.058464
Covariance Proportion	0.848894

44

So sánh các tiêu chuẩn:

RMSE

MAE

MAPE

TIC

Ta thấy của 4 biến tốt hơn. Vậy để dự báo thì ta chọn mô hình 4 biến.

Mời ghé thăm trang web:

- http://kinhteluong.ungdung.googlepages.com
- http://xacsuatthongke.googlepages.com
- http://phamtricao.googlepages.com
- www37.websamba.com/phamtricao
- www.phamtricao.web1000.com