Khoảng tin cậy và kiểm định giả thiết về các hệ số hồi quy

Bởi:

Phạm Trí Cao

Khoảng tin cậy và kiểm định giả thiết về các hệ số hồi quy

Khoảng tin cậy cho các hệ số hồi quy

Thực sự chúng ta không biết σ^2 nên ta dùng ước lượng không chệch của nó là

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Sai số chuẩn của hệ số hồi quy cho độ dốc

$$se(\beta_2) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

Từ

$$\hat{\beta}_2 \sim N(\beta_2, \sigma_{\hat{\beta}_2}^2)$$

với

$$\sigma_{\hat{\beta}_2} = \frac{\sqrt{\sigma^2}}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

ta có

$$Z = \frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}} \sim N(0,1)$$

(3.14)

Từ tính chất của phương sai mẫu ta có

$$(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

(3.15)

Từ (3.14) và (3.15) Ta xây dựng trị thống kê

$$\frac{\frac{\hat{\beta}_2-\beta_2}{\sigma_{\beta_2}}}{\sqrt{\frac{(n-2)\frac{\hat{\sigma}^2}{\sigma^2}}{n-2}}} \sim \frac{Z}{\sqrt{\frac{\chi^2_{n-2}}{n-2}}} \sim t_{(n-2)}$$

(3.16)

Biến đổi vế trái chúng ta được

$$\frac{\frac{\hat{\beta}_2-\beta_2}{\sigma_{\beta_2}}}{\sqrt{\frac{(n-2)\frac{\hat{\sigma}^2}{\sigma^2}}{n-2}}} = \frac{\hat{\beta}_2-\beta_2}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}\sigma^2_{\beta_2}}} = \frac{\hat{\beta}_2-\beta_2}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2} * \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}}} = \frac{\hat{\beta}_2-\beta_2}{se(\hat{\beta}_2)}$$

Thay vào (3.16) ta được

$$\frac{\hat{\beta}_2-\beta_2}{se(\hat{\beta}_2)} \sim t_{(n-2)}$$

(3.17)

Chứng minh tương tự ta có

$$\frac{\hat{\beta}_1-\beta_1}{se(\hat{\beta}_1)} \sim t_{(n-2)}$$

(3.18)

Ước lượng khoảng cho hệ số hồi quy với mức ý nghĩa α như sau

$$\hat{\beta}_1 - t_{(n-2,1-\alpha/2)} se(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{(n-2,1-\alpha/2)} se(\hat{\beta}_1)$$

(3.19)

$$\hat{\beta}_2 - t_{(n-2,1-\alpha/2)} se(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{(n-2,1-\alpha/2)} se(\hat{\beta}_2)$$

(3.20)

Kiểm định giả thiết về hệ số hồi quy

Chúng ta quan tâm nhiều đến ý nghĩa thống kê độ dốc (β_2) của phương trình hồi quy hơn là tung độ gốc (β_1). Cho nên từ đây đến cuối chương chủ yếu chúng ta kiểm định giả thiết thống kê về độ dốc.

Giả thiết

$$H_0 : \beta_2 = \beta_2^*$$
$$H_1 : \beta_2 \neq \beta_2^*$$

Phát biểu mệnh đề xác suất

$$P\left(t_{(n-2,\alpha/2)} \le \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} \le t_{(n-2,1-\alpha/2)} \right) = 1 - \alpha$$

Quy tắc quyết định

Nếu

$$\frac{\hat{\beta}_2 - \beta_2^*}{se(\hat{\beta}_2)} < t_{(n-2,\alpha/2)}$$

hoặc

$$\frac{\hat{\beta}_2 - \beta_2^*}{se(\hat{\beta}_2)} > t_{(n-2,1-\alpha/2)}$$

thì bác bỏ H_0.

Nếu

$$t_{(n-2,\alpha/2)} \le \frac{\hat{\beta}_2 - \beta_2^*}{se(\hat{\beta}_2)} \le t_{(n-2,1-\alpha/2)}$$

thì ta không thể bác bỏ H_0.

Quy tắc thực hành-Trị thống kê t trong các phần mềm kinh tế lượng

Trong thực tế chúng ta thường xét xem biến độc lập X có tác động lên biến phụ thuộc Y hay không. Vậy khi thực hiện hồi quy chúng ta kỳ vọng $\beta_2 \neq 0$. Mức ý nghĩa hay được dùng trong phân tích hồi quy là β=5%.

Giả thiết

$$\mathbf{H_0 : \beta_2 = 0}$$
$$\mathbf{H_1 : \beta_2 \neq 0}$$

Trị thống kê trở thành

$$\text{t-stat} = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)}$$

Quy tắc quyết định

Nếu

$$/\text{t-stat}/ > t_{(n-2,97,5\%)}$$

thì bác bỏ H_0.

Nếu

$$/\text{t-stat}/ > t_{(n-2,97,5\%)}$$

thì không thể bác bỏ H_0.

Tra bảng phân phối Student chúng ta thấy khi bậc tự do n trên 20 thì trị thống kê $t_{97,5\%}$ thì xấp xỉ 2.

Quy tắc thực hành

Nếu /t-stat/ > 2 thì bác bỏ giả thiết $\beta_2 = 0$.

Nếu /t-stat/≤ 2 thì ta không thể bác bỏ giả thiết $\beta_2=0$.

Trong các phần mềm bảng tính có tính toán hồi quy, người ta mặc định mức ý nghĩa α=5% và giả thiết H_0: β_i=0. Thủ tục tính toán hồi quy của Excel cung cấp cho ta các hệ số hồi quy, trị thống kê t, ước lượng khoảng của hệ số hồi quy và giá trị p

Ở chương 2 chúng ta đã biết ước kiểm định trên ước lượng khoảng, trị thống kê và giá trị p là tương đương nhau.
.Sau đây là kết quả hồi quy được tính toán bằng thủ tục hồi quy của một vài phần mềm thông dụng.

Excel

Kết quả Regresstion cho dữ liệu của ví dụ 3.1. (Chỉ trích phần hệ số hồi quy)

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	92,24091128	33,61088673	2,744376012	0,010462	23,39205354	161,089769
X	0,611539034	0,067713437	9,031280327	8,68E-10	0,472834189	0,750243878

Intercept: Tung độ gốc

Coefficients : Hệ số hồi quy

Standard Error : Sai số chuẩn của ước lượng hệ số

t Stat : Trị thống kê $t_{(n-2)}$

P-value : Giá trị p

Lower95%: Giá trị tới hạn dưới của khoảng ước lượng với độ tin cậy 95%.

Upper95% : Giá trị tới hạn trên của khoảng ước lượng với độ tin cậy 95%.

Bác bỏ H_0 khi /t-stat/ > 2 hoặc p-value < 0,05 hoặc khoảng (Lower;Upper) không chứa 0.

Như đã trình bày ở chương 2, đây thực ra là 3 cách diễn đạt từ một mệnh đề xác suất nên kết luận từ 3 trị thống kê t, p và ước lượng khoảng là tương đương nhau.

Eviews

Thủ tục Make Equation cho kết quả như sau(chỉ trích phần hệ số hồi quy):

Khoảng tin cậy và kiểm định giả thiết về các hệ số hồi quy

Dependent Variable: Y

Method: Least Squares

Included observations: 30 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	92.24091	33.61089	2.744376	0.0105
X	0.611539	0.067713	9.031280	0.0000

C : Tung độ gốc

Coefficient : Hệ số hồi quy

Std. Error : Sai số chuẩn của ước lượng hệ số

t – Statistic : Trị thống kê $t_{(n-2)}$

Prob: Giá trị p.Bác bỏ H_0 khi /t-Statistic/ > 2 hoặc Prob < 0,05.

SPSS

Thủ tục Regression->Linear. (Chỉ trích phần hệ số hồi quy).

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	92,241	33,611		2,744	,010
	X	,612	,068	,863	9,031	,000

Constant: Tung độ gốc

Unstandardized Coefficients: Các hệ số hồi quy

Standardized Coefficients: Các hệ số hồi quy chuẩn hoá

Khái niệm này nằm ngoài khuôn khổ của giáo trình.

.

t: t-StatSig: Giá trị p.

Bác bỏ H_0 khi /t/ >2 hoặc Sig < 0,05

Định lý Gauss-Markov

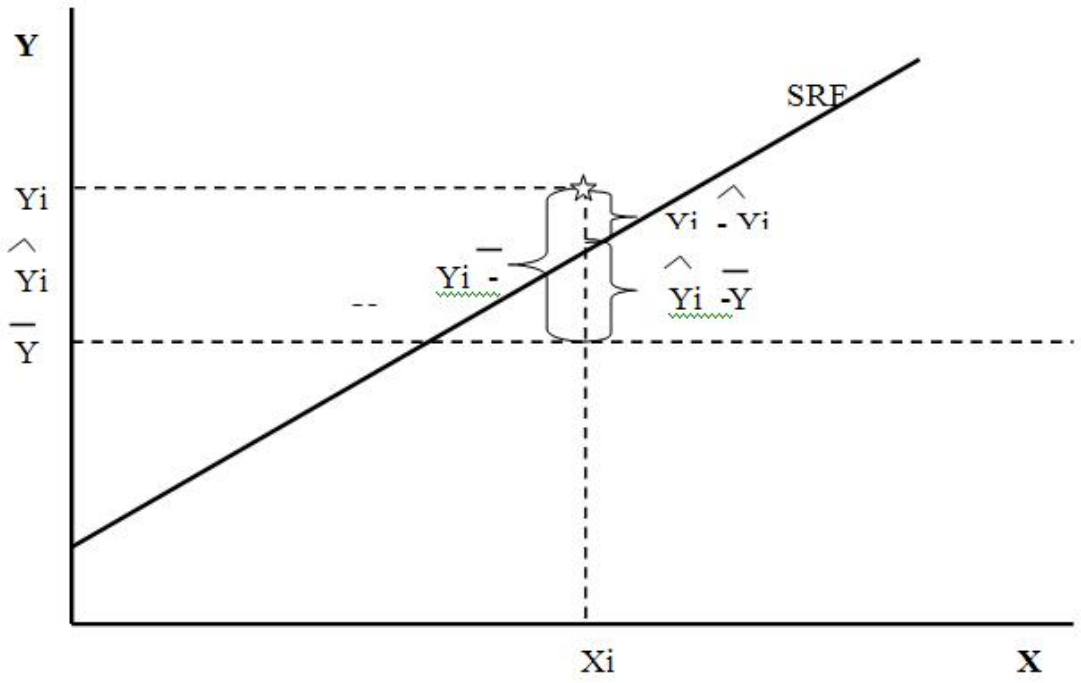
Với các giả định của mô hình hồi quy tuyến tính cổ điển, hàm hồi quy tuyến tính theo phương pháp bình phương tối thiểu là ước lượng tuyến tính không thiên lệch tốt nhất.

Chúng ta sẽ không chứng minh đinh lý này.

Phần chứng minh các tính chất ở phần này có ở Gujarati, Basic Econometrics-3rd Edition, trang 97-98.

Độ thích hợp của hàm hồi quy – R 2

Làm thế nào chúng ta đo lường mức độ phù hợp của hàm hồi quy tìm được cho dữ liệu mẫu. Thước đo độ phù hợp của mô hình đối với dữ liệu là R^2. Để có cái nhìn trực quan về R^2, chúng ta xem xét đồ thị sau



Hình 3.5. Phân tích độ thích hợp của hồi quy

$Y_i - \bar{Y}$: biến thiên của biến phụ thuộc Y, đo lường độ lệch của giá trị Yi so với giá trị trung bình $\bar{Y}$.

$\hat{Y}_i - \bar{Y}$: biến thiên của Y được giải thích bởi hàm hồi quy

$e_i = Y_i - \hat{Y}_i$: biến thiên của Y không giải thích được bởi hàm hồi quy hay sai số hồi quy.

Trên mỗi Xi chúng ta kỳ vọng e_i nhỏ nhất, hay phần lớn biến thiên của biến phụ thuộc được giải thích bởi biến độc lập. Nhưng một hàm hồi quy tốt phải có tính chất mang tính tổng quát hơn. Trong hồi quy tuyến tính cổ điển, người ta chọn tính chất tổng bình phương biến thiên không giải thích được là nhỏ nhất.

Ta có

$$\mathbf{Y_i = \hat{Y} + e_i}$$
$$\mathbf{Y_i - \overline{Y} = \hat{Y} - \overline{Y} + e_i}$$
$$\mathbf{y_i = \hat{y}_i + e_i}$$

Với $y_i = Y_i - \bar{Y}$ và $\hat{y}_i = \hat{Y} - \bar{Y}$

Vậy

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2 + 2\sum_{i=1}^{n} \hat{y}_i e_i$$

(3.21)

Số hạng cuối cùng của (3.21) bằng 0.

Vậy

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2$$

Đặt

$$TSS = \sum_{i=1}^{n} y_i^2 \text{ , } ESS = \sum_{i=1}^{n} \hat{y}_i^2 \text{ và } RSS = \sum_{i=1}^{n} e_i^2$$

TSS(Total Sum of Squares): Tổng bình phương biến thiên của Y.

ESS(Explained Sum of Squares): Tổng bình phương phần biến thiên giải thích được bằng hàm hồi quy của Y.

RSS(Residual Sum of Squares) : Tổng bình phương phần biến thiên không giải thích được bằng hàm hồi quy của Y hay tổng bình phương phần dư.Ta có:

TSS = ESS + RSS

Đặt

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$R^2 = \frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{\hat{\beta}_2^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \hat{\beta}_2^2 \frac{\left(\sum_{i=1}^{n} x_i^2 \Big/ n-1 \right)}{\left(\sum_{i=1}^{n} y_i^2 \Big/ n-1 \right)} = \hat{\beta}_2^2 \frac{S_x^2}{S_y^2}$$

Mặt khác ta có

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

Vậy

$$R^2 = \frac{\left(\sum_{i=1}^{n} x_i y_i \right)^2}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2} = r_{X,Y}^2$$

(3.22)

Vậy đối với hồi quy hai biến R^2 là bình phương của hệ số tương quan.

Tính chất của R^2

$0 \leq R^2 \leq 1$. Với $R^2=0$ thể hiện X và Y độc lập thống kê. $R^2 = 1$ thể hiện X và Y phụ thuộc tuyến tính hoàn hảo.

R^2 không xét đến quan hệ nhân quả.

Dự báo bằng mô hình hồi quy hai biến

Dựa trên X_0 xác định chúng ta dự báo Y_0.

Ước lượng điểm cho Y_0 là :

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$$

.

Để ước lượng khoảng chúng ta phải tìm phân phối xác suất của $\hat{Y}_i$.

Dự báo giá trị trung bình

$$E(Y_0 | X = X_0)$$

Từ

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$$

Suy ra

$$var(\hat{Y}_0) = var(\hat{\beta}_1 + \hat{\beta}_2 X_0) = var(\hat{\beta}_1) + X_0^2 var(\hat{\beta}_2) + 2X_0 cov(\hat{\beta}_1, \hat{\beta}_2)$$

(3.23)

Thay biểu thức của

$var(\hat{\beta}_1)$, $var(\hat{\beta}_2)$ và $cov(\hat{\beta}_1, \hat{\beta}_2)$

ở mục 3.3.4 vào (3.23) và rút gọn

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

Dự báo giá trị cụ thể của Y_0

Từ

$$Y_0 - \hat{Y}_0 = (\beta_1 - \hat{\beta}_1) + (\beta - \hat{\beta}_2)X_0 + e_0$$

Ta có

$$E(Y_0 - \hat{Y}_0) = E(\beta_1 - \hat{\beta}_1) + X_0 E(\beta - \hat{\beta}_2) + E(e_0) = 0$$

và

$$var(Y_0 - \hat{Y}_0) = var(\hat{\beta}_1) + X_0^2 var(\hat{\beta}_2) + 2X_0 cov(\hat{\beta}_1, \hat{\beta}_2) + var(e_0)$$

(3.25)

Số hạng cuối cùng

$$var(e_0) = \sigma^2$$

. Vậy

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

(3.26)

Sai số chuẩn của dự báo

Cho giá trị của Y_0

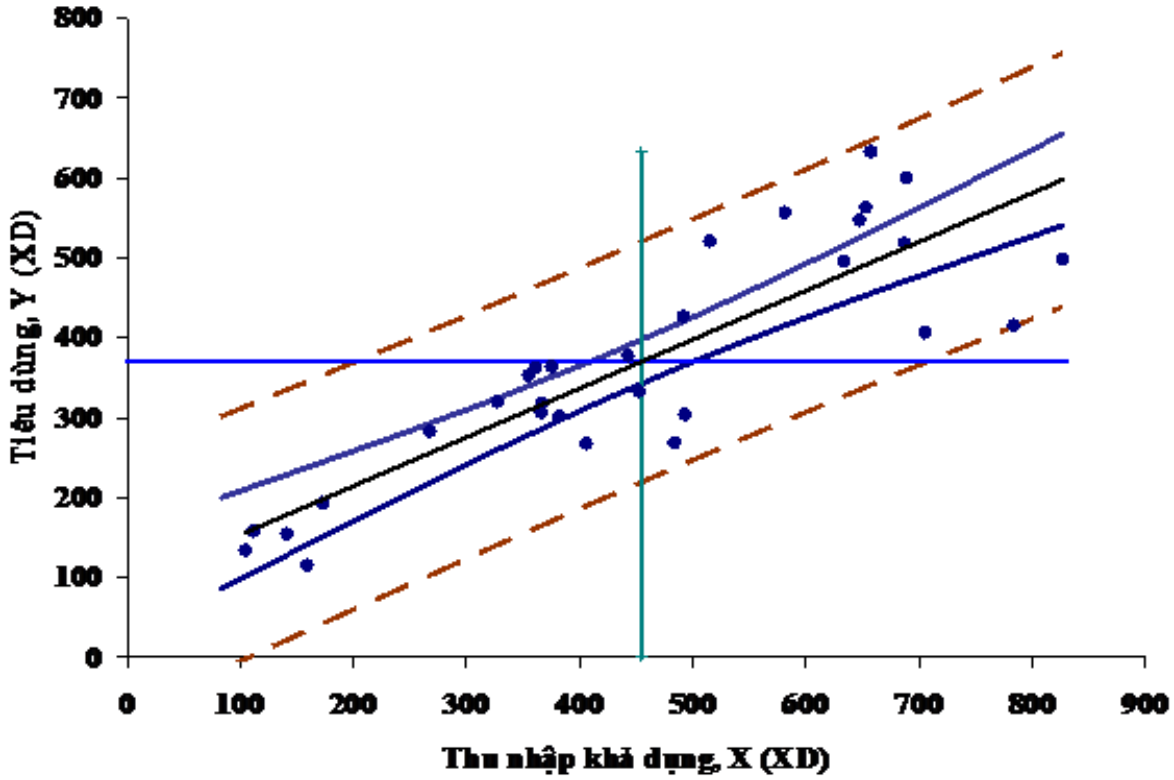
$$se(\hat{Y}_0) = \sigma \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} x_i^2} \right)^{1/2}$$

Khoảng tin cậy cho dự báo

$\hat{Y}_o \pm t_{(n-2,1-\alpha/2)} se(\hat{Y}_o)$

Nhận xét: X 0 càng lệch ra khỏi giá trị trung bình thì dự sai số của dự báo càng lớn. Chúng ta sẽ thấy rõ điều này qua đồ thị sau.

Ước lượng khoảng cho Y_0 trung bìnhY trung bìnhƯớc lượng khoảng cho Y_0X trung bình



Hình 3.6. Ước lượng khoảng cho Y_0.