Chương 5

MÔ HÌNH HỒI QUY TUYẾN TÍNH

§ 5.1. MÔ HÌNH HỒI QUY TUYẾN TÍNH ĐƠN

5.1.1. Vấn đề mô hình hồi quy

Nhiều bài toán trong khoa học kỹ thuật đòi hỏi khảo sát quan hệ giữa hai hoặc nhiều biến. Lấy làm ví dụ, chúng ta xét số liệu ở Bảng 5.1, ở đó y chỉ thị độ sạch của oxy sinh ra trong quá trình chưng cất hóa học, còn x là nồng độ phần trăm của hydrocarbon có mặt ở bình ngưng bộ phận chưng cất.

Bảng 5.1. Độ sạch của oxy ứng với tỷ lệ phần trăm hydrocarbon

TT	x(%)	y(%)	TT	x(%)	y(%)	TT	x(%)	y(%)
1	0.99	90.01	8	1.23	91.77	15	1.11	89.85
2	1.02	89.05	9	1.55	99.42	16	1.2	90.39
3	1.15	91.43	10	1.4	93.65	17	1.26	93.25
4	1.29	93.74	11	1.19	93.54	18	1.32	93.41
5	1.46	96.73	12	1.15	92.52	19	1.43	94.98
6	1.36	94.45	13	0.98	90.56	20	0.95	87.33
7	0.87	87.59	14	1.01	89.54	21	1.32	94.01

Khi thể hiện các điểm (x_i, y_i) lên đồ thị, ta nhận được đồ thị rải điểm như ở Hình 5.1. Ta nhận thấy, mặc dầu không có đường cong đơn giản nào đi qua các điểm này, song có thể khẳng định rằng, các điểm ấy dường như nằm phân tán quanh một đường cong với phương trình $y = f(x)$ nào đó. Vậy có thể giả thiết rằng giá trị trung bình của Y – biến chỉ thị độ sạch khi nồng độ phần trăm X của hydrocarbon tại mức x thỏa mãn quan hệ

$$E(Y \mid x) = f(x) \qquad (5.1.1)$$

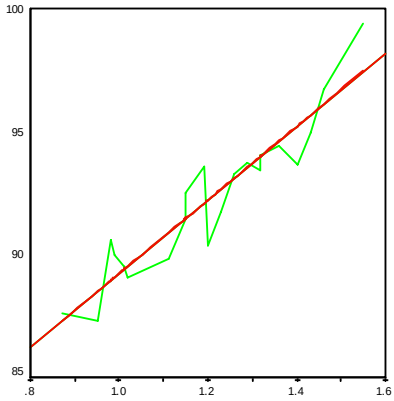
Để tổng quát hóa, chúng ta nên dùng mô hình xác suất bằng cách coi Y là BNN mà ứng với giá trị x của biến X thì

$$Y = f(x) + \varepsilon \qquad (5.1.2)$$

với ε là sai lầm ngẫu nhiên.

Trước hết chúng ta xét trường hợp đơn giản nhất, cũng rất hay xảy ra trong thực tế, khi $f(x) = ax + b$. Khi đó (5.1.2) trở thành

$$Y = ax + b + \varepsilon \qquad (5.1.3)$$



Hình 5.1. Đồ thị rải điểm, đường hồi quy cho số liệu độ sạch của oxy

Mô hình (5.1.3) được gọi là mô hình hồi quy (MHHQ) tuyến tính đơn; x được gọi là biến hồi quy (hay biến độc lập, biến giải thích), Y được gọi là biến phản hồi (hay biến phụ thuộc, biến được giải thích); a, b được gọi là các tham số hồi quy, a: hệ số chặn, b: hệ số góc; đường thẳng $y = ax + b$ được gọi là đường hồi quy (lý thuyết).

Mô hình được gọi là tuyến tính vì nó tuyến tính với các tham số a, b (a, b có lũy thừa 1); được gọi là đơn vì có một biến hồi quy. Ở bài §5.2 chúng ta sẽ xét mô hình hồi quy bội với ít nhất 2 biến hồi quy. Người ta cũng xét mô hình hồi quy phi tuyến, ở đó hàm hồi quy là hàm phi tuyến của các tham số (xem [1], [9]).

Giả sử ở quan sát thứ i biến X nhận giá trị x_i, biến Y nhận giá trị y_i và sai lầm ngẫu nhiên là ε_i. Như vậy, dưới dạng quan sát, mô hình (5.1.3) trở thành

$$\begin{cases} y_1 = a + bx_1 + \varepsilon_1 \\ . \quad . \quad . \quad . \quad . \quad . \quad . \\ y_n = a + bx_n + \varepsilon_n \end{cases} \qquad (5.1.4)$$

Lưu ý rằng y_i là các BNN.

Để khảo sát mô hình chúng ta phải tiến hành các thí nghiệm, các phép đo đạc hay các phép quan sát, gọi chung là quan sát, để có bộ số liệu $\{(x_i, y_i)\}$. Thông qua bộ số liệu này, người ta đưa ra các xấp xỉ (ước lượng) tốt cho các tham số. Mô hình với các hệ số đã ước lượng được gọi là mô hình thực nghiệm (empirical model) hay mô hình lọc (filted model). Dùng mô hình thực nghiệm chúng ta có thể tiến hành một số dự đoán, tính các giá trị cực trị cũng như các khía cạnh của vấn đề điều khiển.

5.1.2. Ước lượng hệ số hồi quy

Bây giờ giả sử các BNN $y_1, ..., y_n$ nhận các giá trị cụ thể nào đó, vẫn ký hiệu là $y_1, ..., y_n$. Khi đó

$$\varepsilon_i = y_i - (ax_i + b) \qquad (5.1.5)$$

thể hiện độ lệch của quan sát thứ i so với đường hồi quy lý thuyết (xem Hình 5.2). Tổng bình phương các độ lệch

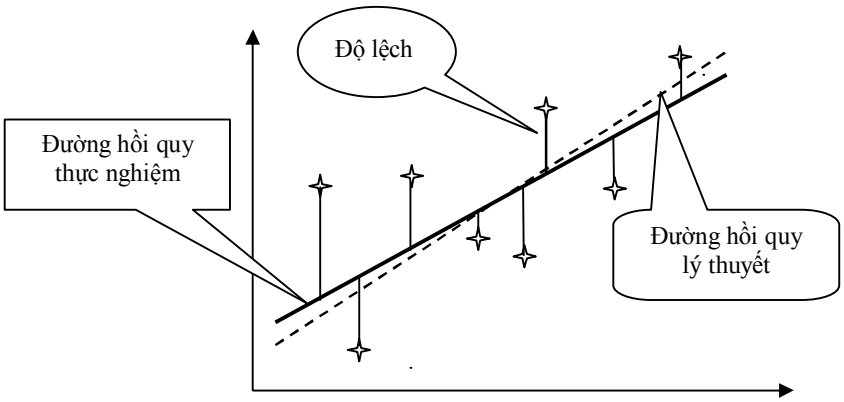
$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

thể hiện "chất lượng" của việc xấp xỉ số liệu bởi đường hồi quy lý thuyết. Ta không thể biết đường hồi quy lý thuyết, việc ta có thể làm là tìm các hệ số a, b để

$$\ell(a,b) = \sum_{i=1}^{n} (y_i - (a + bx_i))^2 \to \min . \qquad (5.1.6)$$

Vì $\ell(a,b)$ là đa thức bậc 2 của 2 ẩn a, b; điều kiện cần để nó đạt cực tiểu là

$$\frac{\partial \ell}{\partial a} = \frac{\partial \ell}{\partial b} = 0 . \qquad (5.1.7)$$



Hình 5.2. Độ lệch và các đường hồi quy lý thuyết, thực nghiệm

Thực ra chứng minh được đây cũng là điều kiện đủ. Đây là hệ 2 phương trình tuyến tính bậc nhất của a, b. không khó khăn gì ta tính được nghiệm của hệ này là:

$$\begin{cases} \hat{b} = \dfrac{\overline{xy} - \overline{x}.\overline{y}}{S_{XX}/n} \\ \hat{a} = \overline{y} - \hat{b}\,\overline{x} \end{cases} \qquad (5.1.8)$$

trong đó

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i;\quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i;\quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i;\quad S_{XX} = \sum_{i=1}^{n} (x_i - \overline{x})^2. \qquad (5.1.9)$$

Với các ƯL này ta được phương trình hồi quy thực nghiệm

$$y = \hat{a}x + \hat{b}. \qquad (5.1.10)$$

Phương pháp tìm các Ư L của hệ số như trên gọi là phương pháp bình phương cực tiểu.

Các phương trình (5.1.5) - (5.1.10) áp dụng với mọi giá trị cụ thể của các BNN $y_1,...,y_n$ nên chúng cũng đúng cho các BNN này. Dưới đây, khi áp dụng các phương trình này và khi không sợ lầm lẫn, ta không phân biệt các BNN $y_1,...,y_n$ với các giá trị cụ thể của chúng.

5.1.3. Tính chất của ước lượng của các hệ số hồi quy

Từ (5.8) ta có ngay $\bar{y} = \hat{a} + \hat{b}\bar{x}$. Như vậy, đường hồi quy đi qua điểm "trung tâm" $(\bar{x}, \bar{y})$ của số liệu.

Lưu ý rằng, ƯL hệ số (5.1.8) hoàn toàn không cần các giả thiết về các thành phần ngẫu nhiên ε_i. Để có các tính chất tốt của ƯL, cần có những giả thiết đặt lên các thành phần ngẫu nhiên này. Giả thiết dễ chấp nhận là chúng có kỳ vọng không, cùng phương sai σ^2, độc lập; giả thiết tiếp sau là chúng có phân bố chuẩn:

$$\varepsilon_1, ..., \varepsilon_n \text{ độc lập, cùng phân bố chuẩn } N(0; \sigma^2). \quad (5.1.11)$$

Khi đó ƯL hệ số có những tính chất thống kê tốt thể hiện ở định lý sau.

Định lý 5.1. Khi điều kiện (5.1.11) thỏa mãn thì:

i) $\hat{a}$ và $\hat{b}$ lần lượt là ƯL không chệch của tham số a và b:

$$E[\hat{a}] = a; \qquad E[\hat{b}] = b \quad (5.1.12)$$

ii) Phương sai của các ƯL $\hat{a}$ và $\hat{b}$ được tính như sau

$$\sigma_a^2 = V[\hat{a}] = \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x})^2}{S_{XX}} \right),$$

$$\sigma_b^2 = V[\hat{b}] = \frac{\sigma^2}{S_{XX}} \quad (5.1.13)$$

iii) ƯL không chệch của phương sai chung σ^2 của mô hình cho bởi

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (5.1.14)$$

với

$\hat{y}_i = \hat{a} + \hat{b}x_i$: dự báo của quan sát thứ i

$e_i = y_i - \hat{y}_i$: phần dư thứ i.

Ý tưởng chứng minh phần i) dựa vào chỗ $\hat{a}$ và $\hat{b}$ là tổ hợp tuyến tính của các BNN chuẩn nên chúng là các BNN chuẩn, rồi thực hiện phép lấy kỳ vọng. Chứng minh phần ii) và iii) dựa vào Định lý 3.20, 3.21 và các phép toán ma trận. Tuy nhiên trình bày chúng rất dài nên không viết ra ở đây; độc giả quan tâm có thể xem ở [1], [9].

Vì σ^2 trong công thức (5.1.13) chưa biết, ta phải dùng xấp xỉ của nó là $\hat{\sigma}^2$. Chúng ta đưa ra định nghĩa.

Định nghĩa. Đối với mô hình HQTT đơn, sai số chuẩn hóa (thực nghiệm) của hệ số góc và hệ số chặn lần lượt được xác định bởi

$$se(\hat{b}) = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}; \qquad se(\hat{a}) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right]} \qquad (5.1.15)$$

trong đó, $\hat{\sigma}^2$ được tính theo (5.1.14).

5.1.4. Kiểm định giả thuyết

Một khâu quan trọng để kiểm tra tính phù hợp của MHHQ là kiểm định giả thuyết. Các hệ số $\hat{a}$, $\hat{b}$, và $\hat{\sigma}^2$ là những BNN nên có thể làm một số kiểm định về chúng. Ta luôn nhớ rằng điều kiện (5.1.11) phải được thỏa mãn. Các đối thuyết đưa ra dưới đây đều là 2 phía. Độc giả có thể đưa ra đối thuyết 1 phía với điều chỉnh thích hợp các ngưỡng phê phán.

a) Sử dụng kiểm định T

Hệ số góc là tham số quan trọng nhất của MHHQ tuyến tính đơn. Xét bài toán kiểm định giả thuyết hai phía:

$$H_0: \ b = b_0 \ / \ H_1: \ b \neq b_0. \qquad (5.1.16)$$

Ở đây, b_0 là giá trị cho trước. Từ giả thiết (5.1.11), y_i là các BNN độc lập và $y_i \sim N(a + bx_i; \sigma^2)$. $\hat{b}$ là tổ hợp tuyến tính của các BNN y_i nên nó cũng có phân bố chuẩn. Theo Định lý 5.1, $\hat{b}$ có phân bố chuẩn $N(b; \sigma^2 / S_{XX})$. Ngoài ra, như trong chứng minh của Định

lý trên, $(n-2)\hat{\sigma}^2/\sigma^2$ có phân bố khi bình phương với n - 2 bậc tự do và độc lập với $\hat{b}$. Theo Định lý 3.21, dưới giả thuyết H_0 thì

$$T_b = \frac{\hat{b} - b_0}{\sqrt{\hat{\sigma}^2 / S_{XX}}} \sim T(n-2). \qquad (5.1.17)$$

Như vậy, chúng ta sẽ bác bỏ H_0 (ở mức ý nghĩa α) nếu

$$|T_b| = \frac{|\hat{b} - b_0|}{se(\hat{b})} = \frac{|\hat{b} - b_0|}{\sqrt{\hat{\sigma}^2 / S_{XX}}} > t_{\frac{\alpha}{2}}(n-2). \qquad (5.1.18)$$

Trường hợp đặc biệt quan trọng là khi $b_0 = 0$:

$$H_0: \ b = 0 / H_1: \ b \neq 0. \qquad (5.1.19)$$

Điều này liên quan đến ý nghĩa (hay tác dụng) của hồi quy (significance of regression): Nếu không bác bỏ H_0 (coi $b = 0$) thì có nghĩa rằng không có một quan hệ tuyến tính nào giữa X và Y (có thể là quan hệ thực sự của X và Y là quan hệ phi tuyến), sự thay đổi của biến X không kéo theo sự thay đổi dự đoán biến Y, X không có (hoặc rất ít) tác dụng để dự đoán Y; dự đoán cho Y tốt nhất nên dùng $\bar{Y}$.

Tương tự, giả thuyết liên quan đến hệ số chặn là

$$H_0: \ a = a_0 / H_1: \ a \neq a_0. \qquad (5.1.20)$$

Bởi vì

$$T_a = \frac{\hat{a} - a_0}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}}{S_{XX}}\right]}} \sim T(n-2) \qquad (5.1.21)$$

nên giả thuyết bị bác bỏ ở mức α nếu

$$|T_a| = \frac{|\hat{a} - a_0|}{se(\hat{a})} = \frac{|\hat{a} - a_0|}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right]}} > t_{\frac{\alpha}{2}}(n-2). \qquad (5.1.22)$$

b) Phân tích phương sai

Phương pháp phân tích phương sai được dùng để kiểm định tính hiệu quả của việc lập mô hình. Trước hết, từ chỗ $y_i - \overline{y} = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i)$, bình phương hai vế rồi lấy tổng ta được:

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (5.1.23)$$

Chúng ta xác định các đại lượng sau đây:

Tổng bình phương đầy đủ: $SS_T = S_{YY} = \sum_{i=1}^{n}(y_i - \overline{y})^2$,

Tổng bình phương hồi quy: $SS_R = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$,

Tổng bình phương các phần dư (các sai số):

$$SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 . \qquad (5.1.24)$$

Biểu thức (5.1.23) được viết lại dưới dạng:

$$SS_T = SS_R + SS_E \qquad (5.1.23')$$

Có thể chứng minh rằng, $SS_R / [\sigma^2 + b^2 S_{XX}]$ và SS_E / σ^2 là những BNN độc lập, có phân bố khi bình phương với 1 và n - 2 bậc tự do tương ứng. Như vậy, nếu giả thuyết $H_0 : b = 0$ là đúng thì

$$F_0 = \frac{SS_R / 1}{SS_E / (n-2)} = \frac{MS_R}{MS_E} \qquad (5.1.25)$$

có phân bố $F(1, n-2)$) (xem Định lý 3.23).

Các đại lượng MS_R, MS_E gọi chung là bình phương trung bình. Nói chung, bình phương trung bình được tính bằng cách lấy tổng bình phương chia cho bậc tự do của nó.

Chúng ta sẽ bác bỏ H_0 nếu $F_0 > f_\alpha(1; n-2)$.

Trong các phần mềm thống kê, thủ tục kiểm định được trình bày ở bảng phân tích phương sai giống như Bảng 5.2.

Bảng 5.2. Phân tích phương sai để kiểm định tính hiệu quả của hồi quy

Nguồn	Tổng các bình phương	Bậc tự do	Bình phương trung bình	F_0	P-giá trị
Hồi quy	SS_R	1	MS_R	$\frac{MS_R}{MS_E}$	P
Sai số	SS_E	n-2	MS_E		
Đầy đủ	SS_T	n-1			

Nếu P-giá trị lớn hơn mức ý nghĩa chọn trước, chúng ta phải chấp nhận giả thuyết $b = 0$, tức là việc xây dựng mô hình không có tác dụng. Cần phải tìm mô hình khác, lấy thêm số liệu…

Lưu ý. Chứng minh được, thủ tục phân tích phương sai và thủ tục kiểm định T cho bài toán kiểm định giả thuyết 2 phía (5.1.16) là tương đương theo nghĩa chấp nhận giả thuyết hay bác bỏ giả thuyết là đồng thời với 2 thủ tục này. Tuy nhiên, kiểm định T linh động hơn, có thể xét kiểm định 1 phía, trong khi phân tích phương sai chỉ có thể xét 1 phía. Mặt khác, phân tích phương sai có thể tổng quát sang trường hợp hồi quy bội xét đến ở bài §5.2.

5.1.5. Khoảng tin cậy

a) Khoảng tin cậy của các tham số

Bởi vì các thống kê T_a, T_b ở (5.1.22), (5.1.17) có phân bố $T(n-2)$ nên dễ dàng xây dựng khoảng tin cậy cho chúng.

Với giả thiết chuẩn (5.1.11), khoảng tin cậy $100(1-\alpha)\%$ cho hệ số chặn a và hệ số góc b lần lượt là

$$\left(\hat{a} \pm t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}\right]}\right),$$

$$\left(\hat{b} \pm t_{\alpha/2}(n-2)\sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}\right). \qquad (5.1.26)$$

b) Khoảng tin cậy cho đáp ứng trung bình

Vì $y_0 = E[Y \mid x_0] = a + bx_0$ nên một ƯL điểm cho giá trị này là $\hat{y}_0 = \hat{a} + \hat{b}x_0$. Đây là ƯL không chệch vì $\hat{a}$ và $\hat{b}$ là ƯL không chệch của a và b. Phương sai của $\hat{a} + \hat{b}x_0$ là $\sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right]$. Tuy nhiên, vì nói chung chúng ta không biết σ^2 mà phải dùng ƯL $\hat{\sigma}^2$ của nó. Dễ thấy rằng $\frac{\hat{y}_0 - y_0}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right]}} \sim T(n-2)$. Từ đó ta có:

Khoảng tin cậy $100(1-\alpha)\%$ cho đáp ứng trung bình khi $x = x_0$ là $(\hat{y}_0 \pm \omega)$, trong đó

$$\begin{cases} \omega = t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right]}, \\ \hat{y}_0 = \hat{a} + \hat{b}x_0. \end{cases} \qquad (5.1.27)$$

c) Dự đoán quan sát tương lai

Một ứng dụng quan trọng của phân tích hồi quy là dự đoán quan sát (cá biệt) của biến Y trong tương lai tại mức x_0 cho trước của biến hồi quy, ký hiệu là $Y \mid x_0$ hay đơn giản là Y_0.

ƯL điểm cho giá trị quan sát tương lai của BNN là giá trị trung bình của nó, ở đây là $y_0 = a + bx_0$. Các tham số a, b lại chưa biết, ta phải dùng ƯL của chúng. Vậy, ƯL điểm cho Y_0 là

$$\hat{y}_0 = \hat{a} + \hat{b}x_0. \qquad (5.1.28)$$

Chú ý rằng BNN Y_0 là quan sát tương lai, nó độc lập với các quan sát quá khứ $y_1, ..., y_n$. Cùng với các giả thiết độc lập, cùng phân bố chuẩn của các sai số, sai số dự đoán $e_0 = Y_0 - \hat{y}_0$ có phân bố chuẩn quy tâm, phương sai

$$V[e_0] = V[Y_0] + V[\hat{y}_0] = \sigma^2\left[1 + \frac{1}{n} + \frac{(x - x_0)^2}{S_{XX}}\right].$$

Giống như trên, ta tìm được khoảng tin cậy (còn gọi là khoảng dự đoán) $100(1-\alpha)\%$ cho quan sát tương lại Y_0 tại x_0 là $(\hat{y}_0 \pm \omega^*)$ với

$$\begin{cases} \omega^* = t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right]}, \\ \hat{y}_0 = \hat{a} + \hat{b}x_0. \end{cases} \quad (5.1.29)$$

Nhận xét. Cả hai khoảng (5.1.27) và (5.1.29) đều đạt cực tiểu tại $x_0 = \bar{x}$ và rộng dần khi x_0 đi ra xa $\bar{x}$. Mặt khác, với cùng mức ý nghĩa, cùng xét tại điểm x_0, khoảng dự đoán luôn luôn rộng hơn khoảng tin cậy. Chúng ta sẽ thấy rõ hơn hiện tượng này ở ví dụ sau.

Ví dụ 5.1. Thông thường, người ta vẫn nghĩ mức tiêu thụ nhiên liệu không phụ thuộc vào việc lái xe nhanh hay chậm. Để kiểm tra người ta cho chạy thử một chiếc xe con ở nhiều vận tốc khác nhau từ 45 đến 70 dặm/giờ. Kết quả ghi thành bảng

Vận tốc	45	50	55	60	65	70	75
Mức tiêu thụ (ml/gal)	24,2	25,0	23,3	22,0	21,5	20,6	19,8

Liệu có thể thay đổi cách nghĩ rằng mức tiêu thụ nhiên liệu không phụ thuộc vào vận tốc xe? Tìm các khoảng tin cậy 95% cho giá trị trung bình và của quan sát tương lai của mức tiêu thụ nhiên liệu khi xe ở vận tốc 50 ml/h .

Giải. Chúng ta xét mô hình HQTT đơn $Y = a + bx + \varepsilon$, trong đó Y là mức tiêu thụ nhiên liệu, x là vận tốc xe. Cần phải xét xem hệ số b có bằng không hay không. Muốn thế ta xét bài toán kiểm định:

$$H_0: \ b = 0 \,/\, H_1: \ b \neq 0.$$

Tính toán các thống kê liên quan ta được

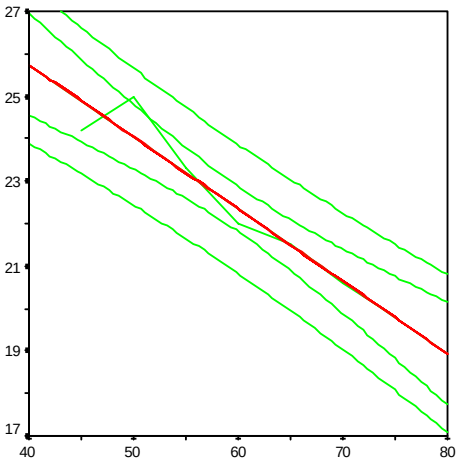
$\bar{x} = 60;\quad S_{XX} = 700;\quad \bar{y} = 22,343;\quad S_{YY} = 21,757;\quad S_{XY} = -119$

$\hat{a} = 32,543;\qquad \hat{b} = -0.17;\qquad SS_R = 1.527$

Mô hình thực nghiệm: $y = 32,54 - 0,17x$.

Tra bảng ta thấy $t_{0.025}(5) = 2,571$. Theo (5.1.26), khoảng tin cậy 95% của b là $(-0,170 \pm 2.571\sqrt{\frac{1.527}{3500}}) = (-0,224; -0,116)$. Khoảng này không chứa điểm 0, vậy ta bác bỏ giả thuyết $b = 0$ với mức ý nghĩa 5%; coi $b \neq 0$, tức là mức tiêu thụ nhiên liệu phụ thuộc vào vận tốc xe. Cũng có thể tính trực tiếp để bác bỏ $b = 0$:

$$T_b = \frac{|\hat{b} - b_0|}{\sqrt{\hat{\sigma}^2 / S_{XX}}} = \frac{|-0,17|}{\sqrt{\frac{0,305426}{700}}} = 8,13 > 2,571 = t_{0.025}(5).$$



Hình 5.3. Khoảng tin cậy (2 đường Hyperbol giữa) và khoảng dự đoán (2 đường hyperbol ngoài) cho mức tiêu thụ nhiên liệu

Dùng (5.1.27) và (5.1.29), khoảng tin cậy và khoảng dự đoán 95% tại vận tốc 50ml/h là

$$\left(24,04 \pm 2,571\left[\sqrt{\frac{1}{7}+\frac{(50-60)^2}{700}}\right]\right) = (24,04 \pm 1,37) = (22,67;\ 24,41)$$

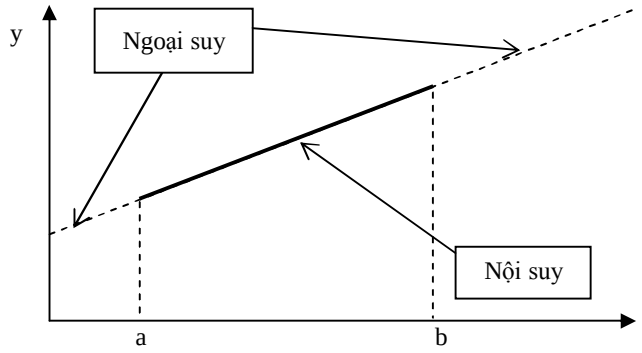
$$\left(24,04 \pm 2,571\sqrt{1+\frac{1}{7}+\frac{(50-60)^2}{700}}\right) = (24,04 \pm 2,92) = (21,12;\ 26,96)$$

Khi x_0 thay đổi, mút trên và mút dưới của khoảng tin cậy tạo thành 2 đường hyperbol giữa, của khoảng dự báo tạo thành 2 đường hyperbol ngoài ở Hình 5.3. Một lần nữa ta thấy khoảng tin cậy cho giá trị trung bình của quan sát là hẹp hơn. #

d) Lưu ý khi sử dụng MHHQ

• *Trường hợp nội suy.* Nói chung, sau những kiểm định cần thiết, chúng ta có thể sử dụng MHHQ thực nghiệm (5.1.10) để làm một số dự đoán "nội suy". Cụ thể là, khi X nhận giá trị x_0 nằm trong dải biến thiên $[a;b]$ của số liệu, giá trị dự đoán của trung bình, cũng như giá trị quan sát tương lai của biến đầu ra sẽ là $\hat{a}+\hat{b}x_0$ … Sự chính xác của các công thức này đã chỉ ra ở phần b) và c).

• *Trường hợp ngoại suy.* Sử dụng phương trình hồi quy để dự đoán giá trị của biến Y ứng với những giá trị của biến đầu vào X nằm ngoài dải biến thiên của số liệu gọi là dự đoán ngoại suy. Tuy nhiên, ở ngoài dải biến thiên của số liệu, các giả thiết về mô hình, thậm chí là quan hệ $E[Y \mid X = x] = ax + b$ có thể không còn đúng. Vì thế, dự đoán với sai lầm đáng kể có thể gây ra từ ngoại suy.



Một cách khắc phục là lấy thêm quan sát (làm thêm thí nghiệm) để dải biến thiên rộng ra, chứa điểm ta quan tâm. Tuy nhiên trong kỹ thuật, nhiều khi ngoại suy là cách duy nhất mà ta có thể tiệm cận vấn đề. Cần lưu ý rằng ta nên áp dụng nó một cách mềm mỏng, với x_0 không xa dải biến thiên [a; b], ta vẫn có thể có kết quả khả dĩ. Tóm lại, ta chỉ áp dụng ngoại suy một cách hãn hữu khi rất cần thiết, chưa thể có đủ số liệu và không còn cách nào khác.

5.1.6. Tính phù hợp của mô hình

a) Phân tích phần dư

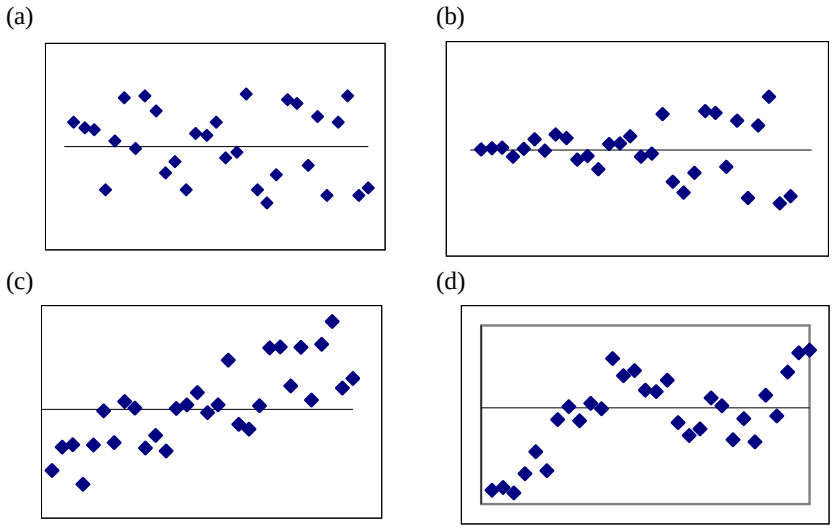
Bước khởi đầu của phân tích hồi quy là dựng đồ thị rải điểm của quan sát. Nếu dáng điệu đồ thị tập trung quanh 1 đường thẳng nào đó, chúng ta sẽ đi tìm các hệ số a, b. Tuy nhiên, việc lập mô hình phải dựa vào các giả thiết chuẩn (5.1.11). Vì các phần dư $e_i = y_i - \hat{y}_i$ đại diện tốt cho các sai số ε_i, người ta thường dùng phân tích phần dư để kiểm tra xem mô hình có phù hợp hay không.

Các phần dư phải tuân theo phân bố chuẩn. Một phương pháp kiểm tra xấp xỉ tính chuẩn là lập tổ chức đồ khi số quan sát n lớn, hoặc lập đồ thị P - P chuẩn khi n nhỏ (xem mục 4.7.1d).

Người ta cũng hay dùng các phần dư chuẩn hóa $d_i = e_i / \sqrt{\hat{\sigma}^2}$, $i = 1,...,n$. Nếu các sai số có phân bố chuẩn, có khoảng 95% các phần dư chuẩn hóa rơi vào khoảng (-2; 2) (nếu $Z \sim N(0;1)$ thì $P\{-2 < Z < 2\} = 0,95$). Hơn nữa, đồ thị d_i phải có dạng bình thường, tập trung "đều đặn" trong dải (-2; 2) quanh trục hoành như dạng (a) ở Hình 5.5. Vi phạm điều đó, chẳng hạn nếu nó có dạng (b), (c), (d) thì phải sửa chữa mô hình, hay tìm mô hình khác và phân tích lại.

Bởi vì $\{\varepsilon_i, i = 1,...,n\}$ là dãy các BNN độc lập thì khi sắp xếp chúng theo thứ tự bất kỳ vẫn được dãy các BNN độc lập. Chúng ta vừa nói đến dãy phần dư d_i theo chiều tăng của chỉ số thời gian i. Người ta cũng lập dãy phần dư theo chiều tăng của x_i hay của $\hat{y}_i$. Nếu một trong các đồ thị đó có dạng (b) thì phương sai của sai số tăng lên theo thời gian (theo chiều tăng của x_i hay của $\hat{y}_i$), xảy ra (c)

thì phương sai của sai số thay đổi, xảy ra (d) thì cần thêm một số hạng bậc cao hơn vào mô hình đa thức hay phải tìm mô hình khác.



Hình 5.5. Dáng điệu phần dư

b) Hệ số xác định (coefficient of determination)

Hệ số xác định ký hiệu bởi R^2 được tính theo công thức sau:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}. \qquad (5.1.30)$$

Theo (5.1.23'), tính chất của hệ số xác định là

$$0 \le R^2 \le 1.$$

Đại thể, ta thường coi R^2 như là độ biến động trong biến đầu ra được giải thích bởi các giá trị đầu vào khác nhau. Khi R^2 lớn, gần bằng 1, thì có nghĩa rằng hầu như độ biến động của các biến đầu ra được giải thích bởi sự khác biệt của các biến đầu vào. Chẳng hạn, với số liệu mức tiêu thụ xăng, vì $R^2 = 0,9298$ nên ta nói mô hình chứa đựng 92,98 % độ biến động trong số liệu.

Gọi r_{XY} là hệ số tương quan mẫu của các cặp điểm (x_i, y_i) (xem mục 4.1.2e) thì ta có thể thấy

$$R^2 = r_{XY}^2 . \qquad (5.1.30')$$

Như vậy, nếu coi X là BNN thì hệ số xác định R^2 chính bằng bình phương của hệ số tương quan mẫu giữa X và Y. Tuy nhiên chúng ta vẫn viết hệ số xác định là R^2 mà không phải r_{XY}^2 vì X không là BNN.

Giá trị R^2 thường được xem như một chỉ thị cho tính "tốt" của mô hình: Khi giá trị này gần bằng 1, mô hình phù hợp tốt; khi giá trị này nhỏ, gần bằng 0, mô hình không phù hợp với số liệu, cần tìm mô hình khác. Tuy nhiên, cần thận trọng, ngưỡng nào cho một mô hình cụ thể lại là điều ta chưa biết, ít ra là đến thời điểm này.

Lưu ý. Liên quan đến máy tính bỏ túi CASIO, ta có thể tính $\hat{\sigma}^2$ như sau:

$$R^2 = 1 - \frac{n-2}{n}\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \Big/ \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 1 - \frac{n-2}{n}\frac{\hat{\sigma}^2}{(y\sigma n)^2}$$

$$\Rightarrow \hat{\sigma}^2 = \frac{n}{n-2}(1-R^2)(y\sigma n)^2 \qquad (5.1.31)$$

với $(y\sigma n)^2 = S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Ví dụ 5.2. Trong nhà máy sản xuất các linh kiện bán dẫn, linh kiện hoàn chỉnh là dây được bó xếp lại thành một cái khung. Người ta quan tâm đến 3 biến: lực kéo (số đo của lực làm cho khung bị hỏng), độ dài của dây, và chiều cao của khuôn đúc. Số liệu có 25 quan sát thể hiện ở 4 cột đầu Bảng 5.5.

Trước hết ta quan tâm đến mối quan hệ giữa lực kéo y và độ dài x_1 của dây, ở đây để tiện ta vẫn ký hiệu là x. Thể hiện số liệu lên đồ thị, dường như đây là quan hệ tuyến tính. Chúng ta dùng mô hình $Y = ax + b + \varepsilon$ để lọc số liệu. Ta tính được:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 8,24; \qquad S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = 698,56;$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 29,0328; \qquad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i = 320,3388;$$

$$(\sigma yn)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 224,237.$$

Từ đó ƯL của các hệ số là

$$\hat{b} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{S_{XX}/n} = 2,9027; \quad \hat{a} = \bar{y} - \hat{b}\bar{x} = 5,115.$$

Ta thu được phương trình :

$$Y = 5,115 + 2,9027x. \qquad (5.1.32)$$

ƯL của σ^2 có thể tính theo $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Tuy nhiên trước hết ta tìm hệ số xác định:

$$R^2 = \frac{SS_R}{SS_T} = \left(\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\right) / \left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right) = 0.964.$$

Đây là giá trị khá lớn. Ta nói có 96,4% số liệu được giải thích bởi mô hình. Theo (5.1.31) thì

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{n}{n-2}(1-R^2)(y\sigma n)^2 = 9,5696 = 3,0934^2.$$

Bây giờ ta kiểm định hệ số $b = 0$. Theo (5.1.15),

$$se(\hat{b}) = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}} = 0.1179 \Rightarrow T_b = \frac{|\hat{b} - 0|}{se(\hat{b})} = \frac{2,9027}{0,1179} = 24,80.$$

P – giá trị của phân bố Student 23 bậc tự do ứng với giá trị 24,80 là 0,000. Vậy ta chấp nhận giả thuyết $b \neq 0$.

Bây giờ ta xét phân tích phương sai.

$$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = 5885,9 \Rightarrow SS_R/1 = 5885,9,$$

$$SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 220,1 \Rightarrow \hat{\sigma}^2 = \frac{SS_E}{n-2} = 9,569$$

$$SS_T = \sum_{i=1}^{n}(y_i - \overline{y})^2 = 6105,9.$$

$$\Rightarrow F = \frac{SS_R/1}{SS_E/(n-2)} = 615,08$$

P - giá trị của phân bố F(1, 23) ứng với giá trị 615,08 bằng 0,000 nên ta cũng kết luận $b \neq 0$.

Các kết quả tính toán trên được cô đọng lại vào trong bảng phân tích hệ số và phân tích phương sai. Thông thường các phần mềm thống kê đều đưa ra các bảng này (xem Bảng 5.3).

Bảng 5.3. Phân tích hệ số và phân tích phương sai cho Ví dụ 5.2

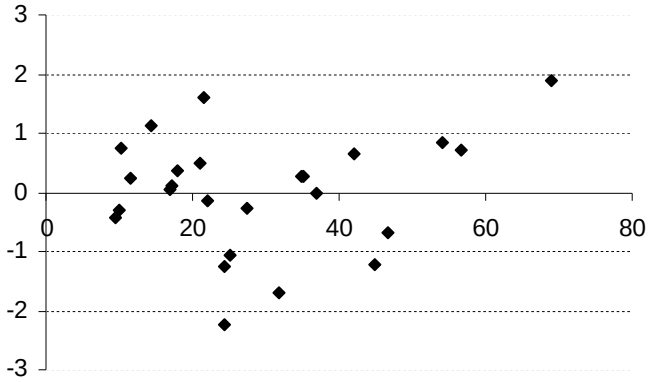
R	R Square	Adjusted R Square	Std. Error of the Estimate
.982	.964	.962	3.0934

	Sum of Squares	df	Mean Square	F	Sig.
Regression	5885.852	1	5885.852	615.080	.000
Residual	220.093	23	9.569		
Total	6105.945	24			

	Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
	B	Std. Error			Lower Bound	Upper Bound
Constant	5.115	1.146	4.464	.000	2.744	7.485
X1	2.903	.117	24.801	.000	2.661	3.145

Ta lập đồ thị rải điểm của các phần dư chuẩn hóa $d_i = e_i/\hat{\sigma}$ theo $\hat{y}_i$ như Hình 5.6. Nhìn vào đồ thị ta thấy có 1 số liệu nằm ngoài dải [-2; 2]. Phần dư chuẩn hóa phân bố khá đều đặn trong dải [-2; 2], duy chỉ có 1 giá trị nằm ngoài dải này (tỷ lệ vi phạm là 1/25, nhỏ hơn 5% nên chấp nhận được).

Tóm lại, chúng ta chấp nhận mô hình (5.1.32).



Hình 5.6. Đồ thị phần dư chuẩn hóa cho số liệu độ kéo

Bây giờ một dây có độ dài 8 sẽ có sức kéo trung bình là

$$\hat{y}(8) = 5,115 + 2,9027.8 = 28,336$$

Theo (5.1.27), khoảng tin cậy 90% của UL này là $(28,336 \pm 1.062) = (27,274;\ 29,398)$. Theo (5.1.29), khoảng tin cậy 90% cho quan sát tương lai khi dây có độ dài 8 là $(28,336 \pm 5,407) = (22.929;\ 33.743)$. #

5.1.7. Tuyến tính hóa một số mô hình

Dùng phép biến đổi loga với biến hồi quy hay biến phản hồi, hoặc với cả hai, dùng phép nghịch đảo với biến hồi quy ..., ta có thể đưa một số mô hình về dạng tuyến tính.

Hồi quy logarith	$y = a + b.\ln x$
Hồi quy mũ	$y = a.e^{b.x}$ $(\Leftrightarrow \ln y = \ln a + b \ln x)$
Hồi quy lũy thừa	$y = a.x^b$ $(\Leftrightarrow \ln y = \ln a + b \ln x)$
Hồi quy nghịch đảo	$y = a + b.(1/x)$
Hồi quy tam thức	$y = a + bx + cx^2$

Chẳng hạn, khi cần dùng hồi quy mũ, trong phần chọn mô hình ta ấn Exp(3); mọi thao tác khác tương tự.

• ***Sử dụng máy tính bỏ túi.*** Chúng ta mô tả ngắn gọn cách sử dụng máy tính bỏ túi CASIO fx-500MS để tính toán hồi qui. Dầu rằng những kết quả còn sơ lược so với các phần mềm chuyên dụng, song chúng cũng giúp ta nhất định trong công việc.

Xoá nhớ thống kê [SHIFT] [MODE] [1] [=]

Gọi chương trình tính [MODE] [REG[3]]

Chọn mô hình [Lin[1]]

Nhập dữ liệu. Chẳng hạn, cần nhập dữ liệu ở Ví dụ 5.1 ta ấn

[45] [,] [24.2] [M+]

Cứ thế ta nhập cho hết dữ liệu.

Gọi kết quả. Nhập dữ liệu xong thì gọi kết quả. Việc gọi kết quả với biến x hoặc y: $\sum x_i^2$, $\sum x_i$, $\bar{x}$, s_X, $\tilde{s}_X$, $\sum y_i^2$, $\sum y_i$, $\bar{y}$, s_Y, $\tilde{s}_Y$ vẫn tiến hành như với thống kê 1 biến đã nêu ở cuối mục 4.1.2. Bảng 5.4 đưa ra vài tính toán như vậy cũng như một số tính toán khác.

Bảng 5.4. Một số thao tác phân tích hồi quy trên máy tính bỏ túi

Lượng cần tính	**Ấn**	**Kết quả**
$\sum x_i y_i$	[SHIFT] [S–SUM] [▷] [$\sum$xy [3]] [=]	9,265
s_Y	[SHIFT] [S–VAR] [▷] [yσn [2]] [=]	1,762
$\tilde{s}_Y$	[SHIFT] [S–VAR] [▷] [yσn–1 [3]] [=]	1,904
$\hat{a}$	[SHIFT] [S–VAR] [▷] [▷] [A [1]] [=]	32,543
$\hat{b}$	[SHIFT] [S–VAR] [▷] [▷] [B [2]] [=]	-0.170
r_{XY}	[SHIFT] [S–VAR] [▷] [▷] [r [3]] [=]	-0.964
$\hat{x}(20)$	[20] [SHIFT] [S–VAR] [▷] [▷] [▷] [$\hat{x}$ [1]] [=]	73.78
$\hat{y}(70)$	[70] [SHIFT] [S–VAR] [▷] [▷] [▷] [$\hat{y}$ [2]] [=]	20.64

Sau khi có giá trị r_{XY}, dùng (5.1.31) ta tính được ƯL cho sai số chung $\hat{\sigma}^2$; tiếp theo ta có thể tính được T_a, T_b, ...

§ 5.2. MÔ HÌNH HỒI QUY TUYẾN TÍNH BỘI

MHHQ tuyến tính bội là sự mở rộng tự nhiên của MHHQ tuyến tính đơn. Chúng ta ghi ra dưới đây những kết quả tóm tắt.

5.2.1. Phương trình hồi quy

a) Dạng quan sát và dạng ma trận

Giả sử mối quan hệ giữa biến phụ thuộc (biến phản hồi) Y và k biến độc lập (biến hồi quy) $x_1,...,x_k$ cho bởi mô hình

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon \qquad (5.2.1)$$

trong đó $\beta_0, \beta_1,...,\beta_k$ là các tham số chưa biết, gọi là các hệ số hồi quy, β_0 gọi là hệ số chặn, $\beta_1,...,\beta_k$ là các hệ số góc; ε là sai số ngẫu nhiên có kỳ vọng 0 và phương sai σ^2.

Khi không sợ nhầm lẫn, ta viết ngắn gọn (5.2.1) dưới dạng

$$E[Y \mid x_1,..., x_k] = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \qquad (5.2.2)$$

hay đơn giản hơn nữa

$$E[Y] = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \qquad (5.2.3)$$

Để tìm hiểu mô hình (5.2.1) chúng ta tiến hành n quan sát và ghi lại kết quả dưới dạng bảng như Bảng 5.5.

Bảng 5.5. Số liệu cho mô hình hồi quy bội

y	x_1	x_2	.	x_k
y_1	x_{11}	x_{12}	.	x_{1k}
.	.	.	.	
y_n	x_{n1}	x_{n2}	.	x_{nk}

Như vậy, dưới dạng quan sát, mô hình (5.2.1) viết lại dưới dạng

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + ... + \beta_k x_{1k} + \varepsilon_1 \\ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \\ y_n = \beta_0 + \beta_1 x_{n1} + ... + \beta_k x_{nk} + \varepsilon_n \end{cases} \qquad (5.2.4)$$

Để thuận lợi cho ký hiệu và các phân tích tiếp theo, chúng ta sử dụng các ký hiệu ma trận sau đây.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ . \\ y_n \end{bmatrix}; \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1k} \\ . & . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{nk} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ . \\ \beta_k \end{bmatrix}; \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ . \\ \varepsilon_n \end{bmatrix}$$

Khi đó, phương trình (5.2.4) được viết lại dưới dạng ma trận

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (5.2.5)$$

trong đó $\mathbf{y}$ là n - véc tơ quan sát, $\mathbf{X}$ là ma trận cấp $n \times p$ của các biến độc lập ($p = k+1$) - còn gọi là ma trận kế hoạch - $\boldsymbol{\beta}$ là p - véc tơ các hệ số hồi quy, $\boldsymbol{\varepsilon}$ là n - véc tơ sai số ngẫu nhiên.

b) Tuyến tính hóa một số mô hình

Mô hình (5.2.3) là tuyến tính vì nó tuyến tính với các tham số β_i. Trong ứng dụng chúng ta thường gặp mô hình dạng

$$E[Y] = \beta_1 g_1(x_1, ..., x_\ell) + ... + \beta_p g_p(x_1, ..., x_\ell) \quad (5.2.6)$$

trong đó $g_1, ..., g_p$ là các hàm nào đó của các biến hồi quy $x_1, ..., x_\ell$.

Đây là mô hình tuyến tính với các tham số β_i, phi tuyến với các biến $x_1, ..., x_\ell$. Xét phép đổi biến

$$z_1 = g_1(x_1, ..., x_\ell); ...; \; z_p = g_p(x_1, ..., x_\ell).$$

Ta có thể đưa (5.2.5) về dạng thông thường

$$E[Y] = \beta_1 z_1 + ... + \beta_p z_p \qquad (5.2.7)$$

là mô hình tuyến tính với cả tham số lẫn các biến hồi quy. Như vậy từ nay ta vẫn gọi mô hình (5.2.6) là tuyến tính. Xét một số trường hợp đặc biệt.

b1. Hồi quy đa thức. Xét mô hình

$$E[Y] = a_0 + a_1 x + ... + a_k x^k .$$

Đặt $z_1 = x; ...; z_k = x^k$, ta đưa mô hình này về dạng

$$E[Y] = a_0 + a_1 z_1 + ... + a_k z_k .$$

Đặc biệt, người ta hay xét mô hình tam thức và đa thức bậc ba:

$$E[Y] = a + cx + cx^2,$$

$$E[Y] = a + cx + cx^2 + dx^3.$$

b2. Mô hình đa thức bậc 2 của hai biến. Đó là mô hình

$$E[Z] = a + bx + cy + dx^2 + exy + fy^2 .$$

Đây là mô hình tuyến tính với 6 tham số a, b, c, d, e, f. Trường hợp giả thuyết $e = 0$ bị bác bỏ, ta nói hai biến hồi quy x và y là tương tác với nhau, mô hình có chứa số hạng tích chéo xy. Trái lại, nếu $e = 0$, ta nói mô hình không chứa số hạng tích chéo xy, 2 biến x và y là không tương tác với nhau.

b3. Dùng phép biến đổi loga với biến phản hồi

Giả sử biến phản hồi Y biểu diễn dưới dạng hồi quy mũ:

$$Y = Ae^{\beta_1 x_1 + ... + \beta_k x_k} . \delta ,$$

trong đó $A, \beta_1, ..., \beta_k$ là các tham số, δ là sai số ngẫu nhiên dạng nhân.

Logarit hóa ta được

$$Z = \ln Y = \ln A + \beta_1 x_1 + ... + \beta_k x_k + \ln \delta$$
$$= \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon, \qquad (\beta_0 = \ln A;\ \varepsilon = \ln \delta)$$

là mô hình tuyến tính thông thường.

Người ta cũng dùng phép biến đổi loga với các biến hồi quy, hoặc với cả biến phản hồi lẫn các biến hồi quy để được các mô hình tuyến tính hóa (xem [1], [9],…).

b4. Hồi quy có chứa sin, cos.

Giả sử biến phụ thuộc có dạng

$$Y(t) = a + bt + c\sin t + d\cos t + \varepsilon .$$

Bằng cách đặt $x_1 = t;\ x_2 = \sin t;\ x_3 = \cos t$, ta đưa mô hình về dạng tuyến tính thông thường.

5.2.2. Ước lượng hệ số hồi quy và tính chất của ƯL

Giả thiết đầu tiên cần có là ma trận **X** có số hàng ít nhất bằng số cột, $p = k + 1 \le n$, và hạng của nó bằng số cột:

$$\text{Rank}(\mathbf{X}) = p. \qquad (5.2.8)$$

Khi đó, ƯL làm cực tiểu tổng bình phương các sai số

$$L(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta)$$

gọi là ƯL bình phương cực tiểu, ký biệu là $\hat{\beta}$, cho bởi:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \qquad (5.2.9)$$

Giống như (5.1.14), ƯL cho sai số chung của mô hình là

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^{\,2} = \frac{1}{n-p}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (5.2.10)$$

với $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik}$: dự báo của quan sát thứ i

$e_i = y_i - \hat{y}_i$: phần dư thứ i.

Nhận thấy vế phải của (5.2.10) có chứa mẫu số $n - p$. Vậy, khi số biến hồi quy p tăng lên, (chẳng hạn với hồi quy đa thức, khi số bậc của đa thức tăng) có thể sai số mô hình tăng lên. Ta sẽ có mô hình cực tồi nếu $p \cong n$.

Để nghiên cứu các tính chất của ƯL tham số, giống với trường hợp có 1 biến hồi quy, cần có giả thiết:

$\varepsilon_1, ..., \varepsilon_n$ độc lập, cùng phân bố chuẩn $N(0; \sigma^2)$. (5.2.11)

Định lý 5.2. Với các giả thiết (5.2.8), (5.2.11) thì:

i) $\hat{\boldsymbol{\beta}}$ là ƯL không chệch của véc tơ tham số $\boldsymbol{\beta}$: $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.

ii) Ma trận covarian của $\hat{\boldsymbol{\beta}}$ cho bởi:

$$Cov(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 .$$

iii) $\hat{\sigma}^2$ theo (5.2.10) là ƯL không chệch của σ^2:

$$E[\hat{\sigma}^2] = \sigma^2 .$$

5.2.3. Kiểm định giả thuyết

a) Kiểm định ý nghĩa của hồi quy. Đó là kiểm tra xem có một quan hệ tuyến tính nào đó giữa biến phản hồi Y với một tập con nào đó của các biến hồi quy $x_1,...,x_k$ hay không. Cụ thể là xét bài toán kiểm định:

$H_0: \beta_1 = \beta_2 = ... = \beta_n = 0$ / $H_1: \beta_j \neq 0$ với ít nhất một $j \in \{1,...,k\}$.

Nếu H_0 bị bác bỏ thì có nghĩa là ít ra một trong các biến hồi quy $x_1,...,x_k$ có ý nghĩa đối với mô hình.

Dưới giả thuyết H_0 có thể chứng minh tổng bình phương hồi quy và tổng bình phương các sai số theo (5.1.24) là những BNN độc lập và có bậc tự do tương ứng là k và $n-p$. Thế thì (xem Định lý 3.23)

$$F_0 = \frac{SS_R / k}{SS_E / (n-p)} = \frac{MS_R}{MS_E} \sim F(k; n-p). \qquad (5.2.12)$$

Từ đó giả thuyết bị bác bỏ ở mức α nếu $F_0 \geq f_\alpha(k; n-p)$.

Các phần mềm thường dùng P–giá trị và đưa ra bảng phân tích phương sai cho thủ tục vừa nêu.

Người ta cũng xét kiểm định cho một tập con của các hệ số $\beta_0, \beta_1,...,\beta_k$ bằng 0. Chi tiết xem [1], [9].

b) Hệ số xác định bội R^2 *và hệ số xác định hiệu chỉnh* R^2_{adj}

Với mô hình hồi quy nhiều biến định nghĩa hệ số xác định bội R^2 và các tính chất của nó như với trường hợp hồi quy đơn:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

Tính chất đặc biệt của hệ số xác định là nó không giảm khi tăng số biến hồi quy. Từ đó, hệ số xác định khó nói cho ta biết việc tăng biến có lợi gì hay không, nhất là khi sự gia tăng hệ số xác định là nhỏ. Vì thế nhiều nhà phân tích lại thích dùng hệ số xác định hiệu chỉnh (adjusted R^2):

$$R^2_{adj} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}. \qquad (5.2.13)$$

Mẫu ở vế phải là hằng số, còn tử là ước lượng của sai số; nó bé nhất khi và chỉ khi hệ số xác định hiệu chỉnh R^2_{adj} lớn nhất. Từ đó, một quy tắc lựa chọn biến hồi quy là:

Chọn một số trong các biến hồi quy $x_1,..,x_k$ để R^2_{adj} lớn nhất.

c) Kiểm định một tham số triệt tiêu (kiểm định T).

Xét bài toán kiểm định một tham số đơn lẻ nào đó triệt tiêu:

$$H_0: \beta_j = 0 / \ H_1: \beta_j \neq 0 \quad (j = 0,1,...,k).$$

Nếu giả thuyết không bị bác bỏ thì có nghĩa rằng biến hồi quy tương ứng không bị loại khỏi mô hình. Thống kê kiểm định là

$$T_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \qquad (5.2.14)$$

trong đó C_{jj} là phần tử thứ j của đường chéo chính của ma trận $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ ứng với $\hat{\beta}_j$.

Vì $T_j \sim T(n-p)$ nên giả thuyết bị bác bỏ nếu $|T_j| > t_{\alpha/2}(n-p)$.

5.2.4. Ước lượng và dự đoán

a) Khoảng tin cậy cho tham số đơn lẻ

Khoảng tin cậy $100(1-\alpha)\%$ cho tham số β_j cho bởi

$$\hat{\beta}_j \pm t_{\alpha/2}(n-p)\,se(\hat{\beta}_j), \quad (se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}). \quad (5.2.15)$$

b) Khoảng tin cậy cho đáp ứng trung bình.

Giả sử quan sát tương lai thực hiện tại mức $x_{01},...,x_{0k}$ của các biến hồi quy $x_1,...,x_k$. Đặt $\mathbf{x}_0 = (1, x_{01}, ..., x_{0k})^T$. Đáp ứng trung bình tại điểm này là $E[\mathbf{Y}|\mathbf{x}_0] = \mathbf{x}_0^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{01} + ... + \beta_k x_{0k}$, ƯL điểm của nó là

$$\hat{y}_0 = \mathbf{x}_0^T\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + ... + \hat{\beta}_k x_{0k}.$$

Đối với MHHQ tuyến tính bội, khoảng tin cậy $100(1-\alpha)\%$ cho đáp ứng trung bình tại điểm $x_{01},...,x_{0k}$ là

$$\hat{y}_0 \pm t_{\alpha/2}(n-p)\sqrt{\hat{\sigma}^2 \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0} \quad (5.2.16)$$

c) Dự đoán cho quan sát mới. ƯL điểm của dự đoán cho quan sát tương lai tại mức $x_{01},...,x_{0k}$ của các biến độc lập là

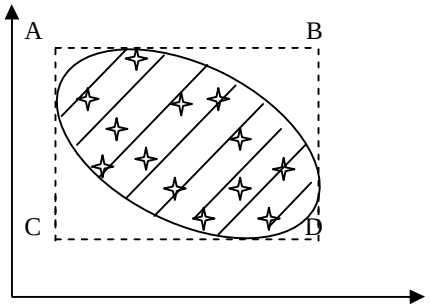
$$\hat{y}_0 = \mathbf{x}_0^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{01} + ... + \beta_k x_{0k}.$$

Khoảng dự đoán $100(1-\alpha)\%$ cho quan sát tương lai này là

$$\hat{y}_0 \pm t_{\alpha/2}(n-p)\sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)}. \quad (5.2.17)$$

d) Vấn đề ngoại suy với mô hình hồi quy bội

Vẫn có những chú ý tương tự như với hồi quy đơn, song vấn đề cần thận trọng hơn. Chẳng hạn, với mô hình có hai biến hồi quy x, y miền biến thiên của các biến hồi quy ở Hình 5.7 phải hiểu là elip chứ không phải hình chữ nhật ABCD. Tình hình sẽ khó khăn hơn khi số biến hồi quy tăng lên.



Hình 5.7. Miền biến thiên của các biến hồi quy

5.2.5. Phân tích phần dư

Với mô hình bội, người ta cũng tiến hành lập đồ thị phần dư chuẩn hóa $d_i = e_i / \hat{\sigma}$ như với mô hình đơn. Nếu có không quá 95% các giá trị d_i nằm trong dải (-2; 2) và phần dư có dáng điệu tương đối đều đặn quanh trục hoành như ở Hình 5.5a thì chấp nhận mô hình. Trái lại, phải tiến hành phân tích lại. Người ta cũng kiểm tra tính chuẩn của phần dư bằng tổ chức đồ hay đồ thị P-P chuẩn như ở mục 4.7.1.

Tuy nhiên, có hai điểm khác biệt. Thứ nhất, ngoài lập đồ thị phần dư chuẩn hóa theo thời gian (theo chỉ số i), theo chiều tăng của một vài biến hồi quy x_i nào đó, theo chiều tăng của dự báo $\hat{y}_i$, khi xét mô hình với một nhóm con các biến hồi quy, người ta còn lập đồ thị phần dư theo biến hồi quy chưa tham gia vào mô hình. Nếu phát hiện ra đồ thị phần dư chuẩn hóa theo biến này không đạt yêu cầu thì có nhiều khả năng biến hồi quy đó cần phải tham gia vào mô hình.

Thứ hai, thay cho đồ thị phần dư chuẩn hóa d_i, người ta thấy rằng đồ thị phần dư điều chỉnh r_i (còn gọi là phần dư student hóa – (studentized residual)) ưu việt hơn, trong đó

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}, \qquad (5.2.18)$$

với h_{ii} là phần tử chéo thứ i của ma trận

$$H = X(X^TX)^{-1}X^T.$$

(Lưu ý rằng $0 < h_{ii} \le 1 \Rightarrow d_i < r_i$).

5.2.6. Sử dụng phần mềm

Các phần mềm thống kê ngày nay cho phép phân tích mô hình với số biến hồi quy lên đến hàng ngàn và số quan sát lên đến hàng chục vạn. Chúng ta cần có những kiến thức cơ bản để tận dụng những lợi thế của các phần mềm này. Mỗi phần mềm có những thế mạnh của nó, song chúng đều có phần phân tích hệ số và phân tích phương sai. Chúng ta tìm hiểu sơ bộ qua một vài ví dụ.

Ví dụ 5.3 (*Phân tích số liệu lực kéo).* Chúng ta lấy lại ví dụ lực kéo ở Ví dụ 5.2. Giả sử chúng ta đã nhập số liệu vào cửa sổ biên tập dữ liệu. Sau đây là một số thao tác cơ bản.

Bảng 5.5. Kết quả xử lý với số liệu lực kéo dây dẫn

TT	Lực kéo y_i	Độ dài x_1	Độ cao x_2	Dự báo $\hat{y}_i$	Phần dư e_i	Phần dư chuẩn hóa d_i
1	9.95	2	50	8.38	1.57	.687
2	24.45	8	110	25.60	-1.15	-.501
3	31.75	11	120	33.95	-2.20	-.963
4	35.00	10	550	36.60	-1.60	-.698
5	25.02	8	295	27.91	-2.89	-1.265
6	16.86	4	200	15.75	1.11	.487
7	14.38	2	375	12.45	1.93	.843
8	9.60	2	52	8.40	1.20	.523
9	24.35	9	100	28.21	-3.86	-1.689
10	27.50	8	300	27.98	-.48	-.208
11	17.08	4	412	18.40	-1.32	-.578
12	37.00	11	400	37.46	-.46	-.202
13	41.95	12	500	41.46	.49	.215
14	11.66	2	360	12.26	-.60	-.263
15	21.65	4	205	15.81	5.84	2.553
16	17.89	4	400	18.25	-.36	-.158
17	69.00	20	600	64.67	4.33	1.894
18	10.30	1	585	12.34	-2.04	-.890
19	34.93	10	540	36.47	-1.54	-.674
20	46.59	15	250	46.56	.03	.013
21	44.88	15	290	47.06	-2.18	-.953
22	54.12	16	510	52.56	1.56	.681
23	56.63	17	590	56.31	.32	.141
24	22.13	6	100	19.98	2.15	.939
25	21.15	5	400	21.00	.15	.067

Chọn chương trình phân tích: Analyze / regression / linear.

Chọn biến: Từ danh sách các biến, đẩy biến y sang ô biến phụ thuộc (Dependent), đẩy hai biến x1, x2 sang ô biến độc lập (Independent(s)).

Phương pháp lọc mô hình: Trong ô phương pháp (Method) ta chọn enter. Phần mềm sẽ lọc mô hình có tất cả các biến.

Tìm ƯL cho tham số và khoảng tin cậy của chúng: Statistics / chọn Estimates và Confidence intervals / Continue.

Lập đồ thị phần dư chuẩn theo y_i: Plots / Đẩy biến phụ thuộc (DEPENDENT) sang ô X, đẩy biến phần dư chuẩn hóa (ZRESID) sang ô Y; muốn có đồ thị xác suất chuẩn chọn thêm Normal probability plot / Continue.

Lưu dự báo $\hat{y}_i$, phần dư e_i, phần dư chuẩn hóa d_i vào danh sách các biến: Save/ trong mục giá trị dự báo (Predicted Values) chọn Unstandardized, trong mục phần dư (Residuals) chọn Unstandardized và Standardized / Continue.

Chạy chương trình: OK

Chi tiết về phần mềm SPSS có thể xem ở [4]. Với các thao tác trên, dự báo cho lực kéo, phần dư, phần dư chuẩn được ghi ở các cột 5,6,7 của Bảng 5.5. Các bảng tóm tắt, phân tích phương sai, phân tích hệ số cho ở Bảng 5.6.

Bảng 5.6. Tóm tắt, phân tích phương sai và phân tích hệ số cho Ví dụ 5.3

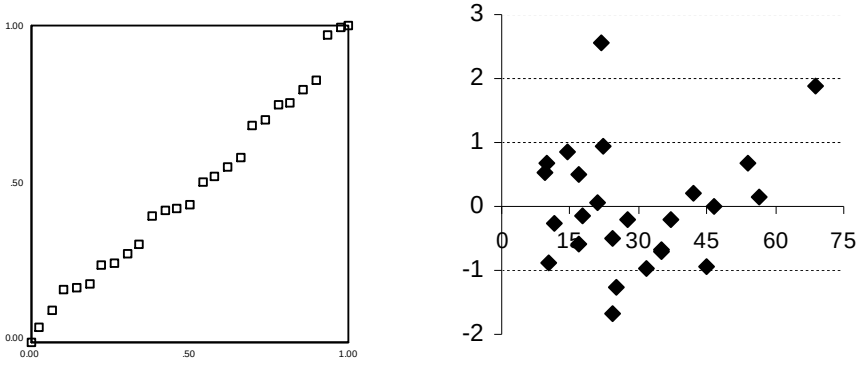
R	R Square	Adjusted R Square	Std. Error of the Estimate
.991	.981	.979	2.28805

	Sum of Squares	df	Mean Square	F	Sig.
Regression	5990.771	2	2995.386	572.167	.000
Residual	115.173	22	5.235		
Total	6105.945	24			

	Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
	B	Std. Error			Lower Bound	Upper Bound
Constant	2.264	1.060	2.136	.044	.065	4.462
X1	2.744	.094	29.343	.000	2.550	2.938
X2	.013	.003	4.477	.000	.007	.018

Ta thấy hệ số xác định $R^2 = 0,981$, vậy có 98,1% số liệu được giải thích bởi mô hình; đây là một tỷ lệ khá lớn. UL cho phương sai chung của mô hình là $\hat{\sigma}^2 = 2,2881^2$. Mức ý nghĩa của thống kê F là 0,000, rất nhỏ so với 0,01: Mô hình có tác dụng tốt để giải thích số liệu. Tất cả các mức ý nghĩa của thống kê T của các tham số đều nhỏ hơn 0,05 (giá trị cực đại 0,044 ứng với biến hằng số). Hậu quả là khoảng tin cậy của tất cả các hệ số đều không chứa gốc tọa độ. Như vậy, các kiểm định T không bác bỏ mô hình. Mô hình dự tuyển là

$$Y = 2,264 + 2,744x_1 + 0,013x_2 + \varepsilon \qquad (*)$$



Hình 5.8. Đồ thị xác suất chuẩn và phần dư chuẩn hóa của số liệu lực kéo

Đồ thị xác suất chuẩn và đồ thị phần dư chuẩn thể hiện ở Hình 5.8. Mặc dầu không phải rất sát, song sai lệch của đồ thị xác suất chuẩn với đường thẳng $y = x$ là có thể chấp nhận được. Đồ thị phần dư chuẩn hóa bố trí khá đều dặn, đối xứng trong dài [-2; 2]. Tuy nhiên 1 quan sát (thứ 15) có trị tuyệt đối phần dư chuẩn vượt quá 2. Dù sao, tỷ lệ $1/25$ là nhỏ hơn 5% và có thể chấp nhận được. Tóm lại, các kiểm định đều không bác bỏ mô hình (*).

Như vậy, với số liệu lực kéo ta có tới 2 mô hình được chấp nhận: mô hình (5.1.32) ở Ví dụ 5.2 và mô hình (*) vừa nêu. Do sử dụng nhiều biến hơn, hệ số xác định của mô hình (*) lớn hơn. Vả lại, mô hình (*) không phải là quá phức tạp, chúng ta chọn nó làm mô hình cuối cùng. #

§ 5.3. LỰA CHỌN BIẾN VÀ XÂY DỰNG MÔ HÌNH

5.3.1. Lựa chọn biến

Vấn đề quan trọng trong ứng dụng của phân tích hồi quy là lựa chọn tập hợp các biến hồi quy để xây dựng mô hình. Đôi khi những kinh nghiệm hay những hiểu biết về mặt lý thuyết có thể giúp nhà phân tích định ra được tập các biến hồi quy sử dụng trong những tình huống cụ thể. Nhiều khi vấn đề lại ở chỗ, người ta biết rất rõ các biến quan trọng, nhưng lại không chắc rằng có phải tất cả các biến dự tuyển đều là cần thiết cho một mô hình thỏa đáng hay không.

Như vậy xuất hiện vấn đề lựa chọn biến hồi quy: Lựa chọn ra trong các biến dự tuyển một tập con các biến "tốt nhất" theo các nghĩa sau đây.

+ Khả năng ứng dụng: Chọn đủ biến hồi quy để việc sử dụng đa dạng của mô hình (dự đoán, ước lượng…) cho kết quả thỏa đáng.

+ Tính kiệm: Để mô hình với giá thấp chấp nhận được và dễ sử dụng, người ta muốn mô hình ít biến hồi quy nhất có thể.

Tuy nhiên, hầu như chẳng có mô hình nào "tốt nhất" theo nghĩa đáp ứng đồng thời nhiều tiêu chuẩn như trên. Những đánh giá, những kinh nghiệm từ xử lý hệ thống đang xem xét thường là trợ lực cần thiết cho việc lựa chọn tập biến hồi quy.

Không có thuật toán nào luôn luôn đưa ra lời giải tốt cho vấn đề lựa chọn biến. Mặc dầu người ta đã đưa ra rất nhiều thuật toán lựa chọn, song chúng chỉ để ý đến khía cạnh kỹ thuật, cần có sự liên kết chặt chẽ với nhà phân tích. Chúng ta sẽ mô tả ngắn gọn một số kỹ thuật thông dụng nhất với vấn đề chọn biến.

Giả sử có K biến dự tuyển $x_1,...,x_K$ và một biến phản hồi y. Tất cả các mô hình đều có hệ số chặn β_0, vậy mô hình có cả thảy $K+1$ số hạng. Chúng ta cũng giả sử dạng hàm của các biến dự tuyển (chẳng hạn $x_1 = 1/x, x_2 = \ln x...$) là đúng.

a) Thủ tục cân nhắc tất cả. Toàn bộ các mô hình có thể đều được xem xét. Ta sẽ cần so sánh 2^K mô hình hồi quy. Mặc dầu việc phân tích 1 mô hình không là vấn đề với các phần mềm ngày nay, song khi

K tương đối lớn, số phương trình cần cân nhắc sẽ tăng lên nhanh chóng (với $K = 10,\ 2^K = 1024$).

b) Dựa vào R^2 *hoặc* R^2_{adj}. Người ta xuất phát từ mô hình có ít biến đến mô hình có nhiều biến hơn. Nếu sự gia tăng R^2 không đáng kể thì dừng lại và lựa chọn mô hình tương ứng.

Tiêu chuẩn dựa vào R^2_{adj} thường tốt hơn. Chọn mô hình có R^2_{adj} cực đại hoặc gần cực đại (nếu muốn số biến hồi quy nhỏ).

c) Tiêu chuẩn PRESS. Gọi $\hat{y}_{(i)}$ là dự đoán tại quan sát thứ i dựa vào mô hình chỉ có n - 1 quan sát còn lại. Đặt

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

trong đó $e_i = y_i - \hat{y}_i$ là phần dư thông thường.

Mô hình có PRESS nhỏ là mô hình được đề nghị.

d) Thủ tục cân nhắc từng bước (stepwise procedure)

Sau đây chúng ta dùng ký hiệu f_{in} (tương ứng f_{aut}) để chỉ giá trị cụ thể của thống kê F riêng phần sau khi bỏ đi (tương ứng thêm vào) một biến hồi quy khỏi mô hình.

Đầu tiên chọn mô hình một biến hồi quy mà có hệ số tương quan cao nhất với biến phản hồi Y. Đây cũng là biến có thống kê f lớn nhất. Chẳng hạn chọn được biến x_1 ở bước thứ nhất.

Giả sử ở bước nào đó đã lựa chọn được m biến, chẳng hạn $x_1, ..., x_m$. Ở bước tiếp theo, xét các mô hình với m biến đã lựa chọn $x_1, ..., x_m$ và 1 biến trong các biến còn lại. Nếu thống kê f riêng phần tăng lên, quay lại xét xem nếu bỏ một trong m biến $x_1, ..., x_m$ thì thống kê f riêng phần có tiếp tục được tăng lên hay không. Như vậy ta tăng thêm hoặc tăng thêm và bỏ đi biến nếu $f_{in} > f_{aut}$. Thủ tục dừng lại đến khi không có biến nào được thêm vào hoặc bỏ đi.

e) Thủ tục tiến (forward procedure). Tại một bước nào đó đưa thêm vào tập biến lựa chọn trong các biến còn lại một biến làm tăng thống kê F riêng phần nhiều nhất. Nếu không có biến nào như vậy thì dừng quá trình lựa chọn biến.

Như vậy, khác với thủ tục cân nhắc từng bước, thủ tục tiến mặc nhiên công nhận các biến lựa chọn ở các bước trước là "tốt". Thực ra, khi có biến mới thêm vào tập chọn, các biến cũ có thể trở nên tồi và cần phải loại bỏ như ở thủ tục cân nhắc từng bước; thủ tục cân nhắc từng bước là ưu việt hơn. Tuy nhiên, nhiều ví dụ chỉ ra rằng, hai thủ tục vừa nêu cho ra cùng một tập chọn các biến hồi quy.

f) Thủ tục lùi (backward procedure). Thủ tục bắt đầu với toàn bộ K biến hồi quy. Biến hồi quy với thống kê f riêng phần nhỏ nhất sẽ bị loại bỏ nếu thống kê f riêng phần này có ý nghĩa, tức là $f < f_{aut}(\ )$. Tiếp tục đến khi không có biến hồi quy nào bị loại.

g) Vài nhận xét về lựa chọn mô hình cuối cùng. Tiêu chuẩn chủ yếu để lựa chọn biến là cân nhắc từng bước. Có thể có một vài mô hình tốt như nhau. Khi đó ta có thể cân nhắc thêm các tiêu chuẩn khác. Nếu số biến hồi quy không lớn, có thể dùng thủ tục cân nhắc tất cả.

Sau khi lựa chọn được biến hồi quy, vẫn phải tiến hành các kiểm tra thông thường: phân tích phần dư, kiểm tra sự phù hợp …, xem xét về mặt lý thuyết như có nhất thiết phải chứa tích chéo, nhất thiết phải chứa biến hồi quy nào đó, dấu của biến nào đó nhất thiết phải dương (hay âm) … hay không.

5.3.2. Những khía cạnh khác của kiểm định mô hình.

a) Đa cộng tuyến. Chúng ta nhớ rằng giả thiết (5.2.8) rằng hạng của ma trận kế hoạch X phải bằng số tham số p. Điều này tương đương với $\det(\mathbf{X}^T\mathbf{X}) \neq 0$. Tuy nhiên điều gì xảy ra nếu $\det(\mathbf{X}^T\mathbf{X}) \approx 0$.

Nếu xảy ra $\det(\mathbf{X}^T\mathbf{X}) \approx 0$ thì có quan hệ tuyến tính mạnh giữa các cột của ma trận X, tức là có sự phụ thuộc tuyến tính mạnh giữa các biến hồi quy $1, x_1, ..., x_k$. Ta nói đã xảy ra hiện tượng đa cộng tuyến (multicollinearity). Đa cộng tuyến có thể gây ra những hậu quả tai hại về ƯL các hệ số hồi quy như phương sai, hiệp phương sai của các ƯL tham số trở nên lớn, tỷ số T mất ý nghĩa trong khi R^2 có thể

cao, dấu của hệ số hồi quy có thể sai… cũng như sai lầm trong sử dụng mô hình nói chung. Nhiều tài liệu đã nêu ra cách phát hiện đa cộng tuyến cũng như cách khắc phục (xem [1], [9], …).

b) Phương sai của sai số thay đổi

Cho đến giờ, trừ trường hợp tìm ƯL cho các tham số, tất cả các thủ tục phân tích đều dựa vào giả thiết (5.2.11). Tuy nhiên, nếu giả thiết này không thỏa mãn; đặc biệt, giả thiết cùng phương sai σ^2 của các sai số bị vi phạm, ta nói đã xảy ra trường hợp phương sai của sai số thay đổi. Nếu ta vẫn sử dụng các phương pháp xử lý thông thường thì có thể chứng minh được ƯL thu được là chệch và không hiệu quả.

Có thể phát hiện phương sai sai số thay đổi bằng đồ thị: Đồ thị phần dư chuẩn hóa theo một biến nào đó (theo chỉ số i, theo biến hồi quy x_i nào đó hoặc theo $\hat{y}_i$) có dạng (b) (c) hay (d) ở Hình 5.5. Cũng có thể dụng một số tiêu chuẩn về lượng như tiêu chuẩn tương quan hạng Spearman, kiểm định Gleiser … Khắc phục hiện tượng phương sai thay đổi chủ yếu dùng phương pháp bình phương tối thiểu trọng lượng, dùng phép biến đổi loga …(xem thêm ở [1], [9]).

c) Có sự tương quan chuỗi giữa các sai số. Xem [1], [9].

Ví dụ 5.4. Một bài báo trên Tạp chí Dược học (Journal of Pharmaceuticals Sciences - 1991) đưa ra dữ liệu về độ hòa tan tỷ số mol quan sát của một chất tan tại nhiệt độ không đổi với các tham số tan riêng phần phân tán, lưỡng cực và liên kết hydro Hansen. Số liệu ở Bảng 5.7, trong đó Y là logarit âm của độ hòa tan tỷ số mol, x_1 là độ hòa tan riêng phần khuyếch tán, x_2 là độ hòa tan riêng phần lưỡng cực, x_3 là độ hòa tan riêng phần liên kết hydro.

Trước hết chúng ta lọc mô hình đa thức bậc hai đầy đủ

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{11}x_1^2 + b_{22}x_2^2 + b_{33}x_3^2 + \varepsilon$$

Các kết quả phân tích sau dựa vào phần mềm SPSS. Hệ số xác định $R^2 = 0.917$ và hệ số xác định hiệu chỉnh $R^2_{Adj} = 0,870$ khá lớn, sai số chung của mô hình $\sigma^2 = 0,06923^2$ khá nhỏ, mức ý nghĩa của

Bảng 5. 7. Số liệu độ tan

TT	Y	x_1	x_2	x_3
1	0.222	7.3	0	0
2	0.395	8.7	0	0.3
3	0.422	8.8	0.7	1
4	0.437	8.1	4	0.2
5	**0.428**	**9**	**0.5**	**1**
6	0.467	8.7	1.5	2.8
7	0.444	9.3	2.1	1
8	0.378	7.6	5.1	3.4
9	0.494	10	0	0.3
10	**0.456**	**8.4**	**3.7**	**4.1**
11	0.452	9.3	3.6	2
12	0.112	7.7	2.8	7.1
13	0.432	9.8	4.2	2
14	0.101	7.3	2.5	6.8
15	**0.232**	**8.5**	**2**	**6.6**
16	0.306	9.5	2.5	5
17	0.0923	7.4	2.8	7.8
18	0.116	7.8	2.8	7.7
19	0.0764	7.7	3	8
20	**0.439**	**10.3**	**1.7**	**4.2**
21	0.0944	7.8	3.3	8.5
22	0.117	7.1	3.9	6.6
23	0.0726	7.7	4.3	9.5
24	0.0412	7.4	6	10.9
25	**0.251**	**7.3**	**2**	**5.2**
26	0.00002	7.6	7.8	20.7

thống kê F là 0,000. Vậy mô hình giải thích tốt dữ liệu. Tuy nhiên, tất cả mức ý nghĩa (P-giá trị) của các hệ số đều lớn hơn 0,05 (giá trị nhỏ nhất là 0,087 ứng với biến x_3, giá trị lớn nhất là 0,719 ứng với biến x_2x_3). Hậu quả là, tất cả các khoảng tin cậy 95% của các hệ số đều chứa gốc tọa độ. (Xem Bảng 5.8). Ta phải tìm mô hình khác.

Bây giờ chúng ta dùng thủ tục cân nhắc từng bước (stepwise prosedure) để lựa chọn biến. Phần mềm dừng lại 3 biến lựa chọn, đó là x_3, x_1 và x_2^2 (tất nhiên có biến hằng số). Bảng 5.9 sau đây chỉ ra tóm tắt bết quả, phân tích phương sai, phân tích hệ số của mô hình lựa chọn.

Nhận thấy rằng hệ số xác định $R^2 = 0,886$ tuy thua kém trường hợp có đầy đủ các biến là 0,917, song hệ số xác định hiệu chỉnh (quan trọng hơn) là $R^2_{Adj} = 0,870$ lại không thua kém trường hợp có đầy đủ các biến. Sai số chuẩn hóa ($\approx 0,0609$) cũng như mức ý nghĩa của thống kê F ($\approx 0,000$) xem là như nhau với 2 mô hình. Tuy nhiên, đối với mô hình sau, tất cả các mức ý nghĩa của thống kê T ứng với các biến lựa chọn đều nhỏ hơn 0,05 (cực đại bằng 0,0320, tất cả các khoảng tin cậy 95% đều không chứa gốc tọa độ.

Bảng 5. 8. Tóm tắt, phân tích phương sai và phân tích các hệ số cho mô hình đầy đủ của số liệu độ tan

R	R Square	Adjusted R Square	Std. Error of the Estimate
0.958	.917	.870	.060923263

	Sum of Squares	df	Mean Square	F	Sig.
Regression	.656	9	.073	19.628	.000
Residual	.059	16	.004		
Total	.715	25			

	Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
	B	Std. Error			Lower Bound	Upper Bound
Constant	-1.769	1.287	-1.375	.188	-4.498	.959
X1	.421	.294	1.430	.172	-.203	1.044
X2	.222	.131	1.701	.108	-.055	.500
X3	-.128	.070	-1.822	.087	-.277	.021
X1X2	-.020	.012	-1.651	.118	-.045	.006
X1X3	.009	.008	1.201	.247	-.007	.025
X2X3	.003	.007	.366	.719	-.012	.017
X1B	-.019	.017	-1.150	.267	-.055	.016
X2B	-.007	.012	-.618	.545	-.033	.018
X3B	.001	.001	.572	.575	-.002	.004

Lưu ý. Dùng thủ tục tiến (forward prosedure) cho kết quả trùng với kết quả từ thủ tục cân nhắc từng bước. Nếu dùng thủ tục lùi (backward prosedure), khoảng tin cậy của hệ số của mô hình cuối cùng có chứa gốc tọa độ. Nếu dùng thủ tục loại biến từng bước (remove prosedure) mô hình cuối cùng chỉ chứa biến hằng số, không thể dùng để dự báo được.

Như vậy, qua khâu lựa chọn biến chúng ta được

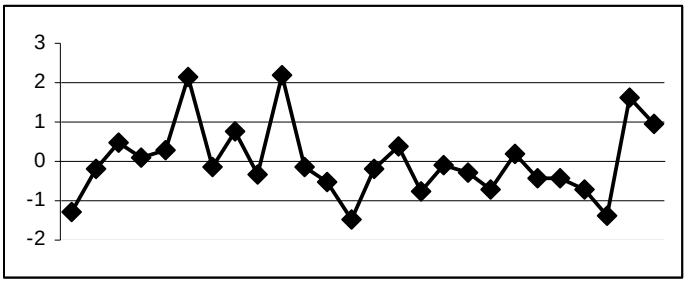
$$Y = -0,304 + 0,083x_1 - 0,031x_3 + 0,004x_2^2 + \varepsilon . \qquad (*)$$

Bảng 5.9. Tóm tắt, phân tích phương sai, phân tích hệ số của mô hình cuối cùng theo phương pháp cân nhắc từng bước của số liệu độ tan

R	R Square	Adjusted R Square	Std. Error of the Estimate
0.941	.886	.870	.060973528

	Sum of Squares	df	Mean Square	F	Sig.
Regression	.633	3	.211	56.778	.000
Residual	.082	22	.004		
Total	.715	25			

	B	Std. Error	t	Sig.	95% Confidence Interval for B	
					Lower Bound	Lower Bound
Constant	-.304	.132	-2.292	.032	-.578	-.029
X3	-.031	.004	-7.156	.000	-.041	-.022
X1	.083	.015	5.564	.000	.052	.113
X2B	.004	.001	3.205	.004	.002	.007



Hình 5.9. Phần dư chuẩn hóa theo quan sát của số liệu độ tan

Kiểm tra phần dư của mô hình này. Chẳng hạn theo chỉ số i ta thấy có 2 giá trị phần dư chuẩn hóa (ứng với quan sát thứ 6 và thứ 10) vượt quá 2; vi phạm thứ hai là d_i khá nhỏ tại các quan sát 11 – 24. Dù sao 2 vi phạm này cũng không đến nỗi nào. Phần dư chuẩn hoá xếp theo x_1, x_2 hay $\hat{y}$ đều không có vi phạm đáng kể. Chúng ta lựa chọn (*) làm mô hình cuối cùng. #