HỒI QUY TUYẾN TÍNH
VÀ SỰ TƯƠNG QUAN TUYẾN TÍNH

Trong thực tế người ta thường phải giải các bài toán liên quan đến tập hợp các biến khi đã biết giữa các biến có một mối quan hệ cố hữu. Chẳng hạn, trong một tình huống công nghiệp, người ta có thể biết hàm lượng nhựa đường ở đầu ra của một quá trình hóa học có liên quan đến nhiệt độ đầu vào. Ta có thể quan tâm đến việc phát triển một phương pháp dự đoán, nghĩa là một quy trình có thể ước lượng hàm lượng nhựa đường cho các mức nhiệt độ đầu vào khác nhau từ thông tin thực nghiệm. Khi đó phương pháp thống kê trở thành phương pháp tốt nhất để ước lượng mối quan hệ giữa các biến.

Thông thường sẽ có một biến phụ thuộc, hoặc biến ngẫu nhiên Y không được kiểm soát trong thực nghiệm. Biến ngẫu nhiên này phụ thuộc vào một hoặc nhiều biến hồi quy độc lập, ví dụ như $x_1, x_2, ..., x_k$, các biến này trên có thể kiểm soát được trong thực nghiệm, và được đo với sai số không đáng kể. Do đó các biến độc lập $x_1, x_2, ..., x_k$ không phải các biến ngẫu nhiên. Trong ví dụ trên, nhiệt độ đầu vào là biến độc lập (biến hồi quy), kí hiệu *x*, hàm lượng nhựa đường ở đầu ra là biến ngẫu nhiên Y. Mối liên hệ giữa biến phụ thuộc Y theo các biến độc lập $x_1, x_2, ..., x_k$ trong một tập hợp dữ liệu được đặc trưng bởi một phương trình dự đoán gọi là phương trình hồi quy.

Trong bài này chúng ta sẽ chỉ xem xét đến trường hợp có 1 biến hồi quy, khi đó ta gọi là bài toán hồi quy tuyến tính đơn.

Bài toán dẫn đến khái niệm Y|*x* – nghĩa là giá trị của biến ngẫu nhiên Y tương ứng với một giá trị *x* cố định. Rõ ràng, trong ví dụ trên, mỗi một giá trị *x* cố định, có các giá trị Y|*x* khác nhau, vì thế rất tự nhiên ta sẽ quan tâm đến các giá trị trung bình, phương sai $\mu_{Y|x}, \sigma^2_{Y|x}$. Như vậy, cứ có 1 giá trị x_i ta sẽ có một biến ngẫu nhiên $Y \mid x_i$, với trung bình và phương sai $\mu_{Y|x_i}, \sigma^2_{Y|x_i}$.

Khi nối các điểm $\left(\mu_{Y|x_i}, x_i\right)$ ta sẽ được một đường thẳng, đó chính là đường hồi quy, cụ thể đó là ***đường hồi quy tuyến tính đơn.***

Ở đây gọi là hồi quy tuyến tính vì $\mu_{Y|x}$ có quan hệ tuyến tính với *x* theo phương trình hồi quy tổng thể

$$\mu_{Y|x} = \alpha + \beta x$$

Trong đó α, β gọi là hệ số hồi quy được ước lượng từ dữ liệu mẫu tương ứng là *a, b*. Nghĩa là ta có thể ước lượng $\mu_{Y|x}$ theo $\hat{y}$ từ đường hồi quy thực nghiệm:

$$\hat{y} = a + bx$$

I. HỒI QUY TUYẾN TÍNH ĐƠN

Như ta đã biết $\mu_{Y|x} = \alpha + \beta x$ gọi là đường hồi quy tuyến tính đơn. Nếu ta kí hiệu E là biến ngẫu nhiên chỉ độ lệch giữa giá trị của biến ngẫu nhiên Y và $\mu_{Y|x}$ thì ta có :

$$Y = \mu_{Y|x} + E = \alpha + \beta x + E$$

Công thức này gọi là mô hình hồi quy tuyến tính đơn.

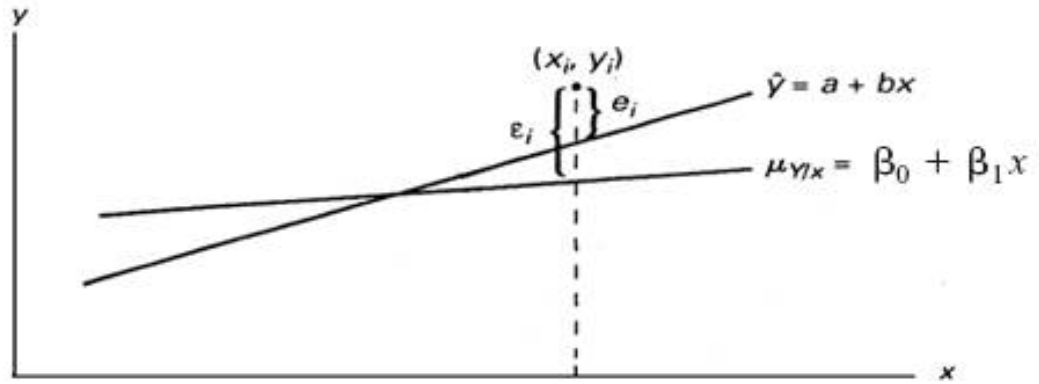
Vấn đề đặt ra là ta chưa biết các giá trị hệ số hồi quy, vì thế phải đi ước lượng nó từ số liệu mẫu, tức là ta sẽ có các cặp điểm: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.

Sai số E sẽ nhận những giá trị cụ thể ε_i khi (*x*,*y*) nhận giá trị (x_i, y_i) cụ thể:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Tương tự với đường hồi quy ước lượng $\hat{y} = a + bx$, ta cũng có mối liên hệ:

$$y_i = a + bx_i + e_i$$



Mô tả điểm khác biệt giữa ε_i, e_i

Công việc chính của chúng ta là tìm *a, b* để ước lượng cho α, β. Ta sẽ tìm *a, b* dựa vào phương pháp bình phương tối thiểu. Các số *a, b* sẽ được chọn làm ước lượng cho α, β nếu làm cực tiểu hóa hàm hai biến sau:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - a - bx_i\right)^2$$

Lấy đạo hàm của SSE đối với *a* và *b*, ta có:

$$\frac{\partial(SSE)}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i), \quad \frac{\partial(SSE)}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i$$

Cho đạo hàm từng phần bằng 0 và sắp xếp lại các số hạng, ta được phương trình:

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \quad a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i,$$

phương trình này có thể được giải đồng thời để tìm ra công thức tính cho *a* và *b*.

1. Ước lượng điểm cho các hệ số hồi quy theo phương pháp bình phương tối thiểu

Cho mẫu $\{(x_i, y_i) = 1,2, ...,n\}$, các ước lượng bình phương tối thiểu *a* và *b* của hệ số hồi quy α, β được tính từ công thức sau:

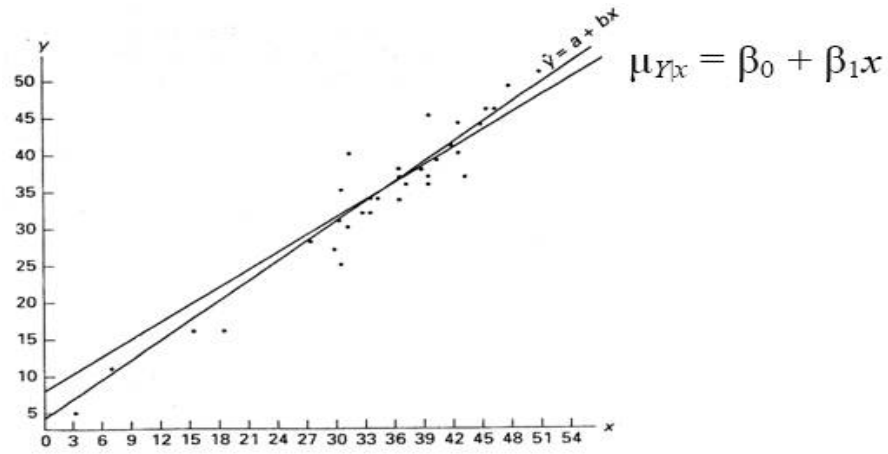
$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Và

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \overline{y} - b\overline{x}$$

Ví dụ 1: Một trong những bài toán khó giải trong việc kiểm soát ô nhiễm nước đang phải đối mặt được ngành công nghiệp thuộc da đưa ra. Chất thải từ các xưởng thuộc gia rất phức tạp về mặt hóa học. Chúng được đặc trưng bởi nhu cầu ô xy sinh hóa có giá trị cao, chất rắn dễ bay hơi, và các chỉ số ô nhiễm khác. Xét số liệu thực nghiệm được trình bày trong Bảng dưới đây, các số liệu được lấy từ 33 mẫu chất thải đã qua xử lý hóa chất trong nghiên cứu tiến hành tại Học viện Bách khoa và Đại học Tiểu bang Virginia. Các chỉ số về *X*, mức giảm tính theo phần trăm tổng lượng chất rắn, và *Y*, mức giảm tính theo phần trăm về nhu cầu ô xy sinh hóa của 33 mẫu, đã được ghi lại. Hãy ước lượng hàm hồi quy của *Y* đối với *X*.

Phần trăm chất rắn, *x* (%)	Nhu cầu oxy sinh hoá, *y* (%)	Phần trăm chất rắn, *x* (%)	Nhu cầu oxy sinh hoá, *y* (%)
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		



Biểu đồ phân bố với đường hồi quy

Ta có $n = 33$, $\sum_{i=1}^{33} x_i = 1104$, $\sum_{i=1}^{33} y_i = 1124$, $\sum_{i=1}^{33} x_i y_i = 41355$, $\sum_{i=1}^{33} x_I^2 = 41086$

Nên

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{33\times 41355 - 1104\times 1124}{33\times 41086 - 1104^2} = 0,903643$$

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \frac{1124 - 0,903643\times 1104}{33} = 3,829633$$

Do đó, đường hồi quy mẫu được cho bởi công thức:

$$\hat{y} = 3,8296 + 0,9036x$$

2. Ước lượng không chệch của phương sai

Phương sai σ^2 cho ta biết mức độ phân tán của biến ngẫu nhiên Y xung quanh đường hồi quy. Người ta chứng minh được rằng ước lượng không chệch của σ^2 là

$$s^2 = \frac{SSE}{n-2} = \sum_{i=1}^{n} \frac{(y_i - \hat{y})^2}{n-2} = \frac{S_{yy} - bS_{xy}}{n-2}$$

Kí hiệu

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 \qquad S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 \qquad S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Để tính toán những số liệu trên, chúng ta thường dùng máy tính, vì thế ta có thể sử dụng một trong các công thức tương đương sau đây:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = (n-1)S_x^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = (n-1)S_y^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i\, y_i - n.\overline{x}.\overline{y}\,; \qquad b = \frac{S_{xy}}{S_{xx}}$$

Ví dụ 3: Một cuộc nghiên cứu về lượng mưa và lượng ô nhiễm không khí thải ra đã cho các số liệu sau:

Lượng mưa hàng ngày, *x* (0,01 cm)	**Lượng hạt ô nhiễm thải ra, *y* (mcg/cum)**
4,3	126
4,5	121
5,9	116
5,6	118
6,1	114
5,2	118
3,8	132
2,1	141
7,5	108

a) Tìm phương trình đường hồi quy để dự đoán trước lượng hạt ô nhiễm thoát ra từ lượng mưa hàng ngày.

b) Tính lượng hạt ô nhiễm thoát ra khi lượng mưa hàng ngày là x = 4,8 đơn vị.

HỒI QUY TUYẾN TÍNH
VÀ SỰ TƯƠNG QUAN TUYẾN TÍNH

Các nội dung chính:

- Khoảng tin cậy cho các hệ số hồi quy
- Kiểm định giả thiết cho các hệ số hồi quy.
- Dự đoán dựa vào đường hồi quy.
- Hệ số tương quan

I. KHOẢNG TIN CẬY CHO CÁC HỆ SỐ CỦA ĐƯỜNG HỒI QUY

1. Khoảng tin cậy cho hệ số β

Khoảng tin cậy (1– α)100% đối với tham số β trong đường hồi quy $\mu_{Y|x} = \alpha + \beta x$ là

$$b - \frac{t_{\alpha/2}s}{\sqrt{S_{xx}}} < \beta < b + \frac{t_{\alpha/2}s}{\sqrt{S_{xx}}}$$

trong đó $t_{\alpha/2}$ là giá trị của phân phối t với $n-2$ bậc tự do.

Ví dụ 1: Dùng bảng số liệu ví dụ 1, bài 13.

Tính s^2. Từ đó tìm khoảng tin cậy 95% đối với tham số β trong đường hồi quy.

GIẢI:

Ta có: $n = 33$, $\sum_{i=1}^{33} x_i = 1104$, $\sum_{i=1}^{33} y_i = 1124$, $\sum_{i=1}^{33} x_i y_i = 41355$,

$$\sum_{i=1}^{33} x_i^2 = 41086$$

Suy ra

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 4152,8$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = 3713,88$$

$$S_{xy} = \sum_{i=1}^{n} x_i\, y_i - n.\bar{x}.\bar{y} = 3752,09$$

Ta đã có: $b = 0,903643$

Nên ta được:

$$s^2 = \frac{S_{yy} - bS_{xy}}{n-2} = 10,4299 \Rightarrow s = 3,2295$$

Tra bảng A4, với 31 bậc tự do: $t_{0,025} = 2,045$

Vậy, khoảng tin cậy 95% cho β là:

$$b - \frac{t_{\alpha/2}s}{\sqrt{S_{xx}}} < \beta < b + \frac{t_{\alpha/2}s}{\sqrt{S_{xx}}}$$

Thay số

$$0,8011 < \beta < 1,0061$$

2. Khoảng tin cậy cho hệ số α

Khoảng tin cậy (1– α)100% đối với α trong đường hồi quy $\mu_{Y|x} = \alpha + \beta x$ là :

$$a - \frac{t_{\alpha/2}s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{nS_{xx}}} < \alpha < a + \frac{t_{\alpha/2}s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{nS_{xx}}}$$

trong đó $t_{\alpha/2}$ là giá trị của phân phối t với $n-2$ bậc tự do.

Ví dụ 2: Dùng bảng số liệu ví dụ 1, bài 13. Tính khoảng tin cậy 95% đối với α trong đường hồi quy $\mu_{Y|x} = \alpha + \beta x$.

GIẢI:

Tương tự ví dụ 1, với $a = 3,829633$, $t_{0,025} = 2,045$ với 31 bậc tự do.

Khoảng tin cậy 95% đối với α là

$$a - \frac{t_{\alpha/2} s \sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{nS_{xx}}} < \alpha < a + \frac{t_{\alpha/2} s \sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{nS_{xx}}}$$

Thay số: $0{,}2131 < \alpha < 7{,}4461$.

II. KIỂM ĐỊNH GIẢ THUYẾT VỀ HỆ SỐ HỒI QUY

1. Kiểm định giả thuyết hệ số β

Các bước của bài toán:

1. Đặt bài toán: H_0: $\beta = \beta^*$

H_1: $\beta \neq \beta^*$

2. Chỉ tiêu kiểm định $T = \dfrac{b - \beta^*}{S / \sqrt{s_{xx}}}$

Từ đó tính

$$t = \frac{b - \beta^*}{s / \sqrt{s_{xx}}}$$

3. Chọn mức ý nghĩa α. Suy ra miền bác bỏ

$$D = (-\infty, -t_{\alpha/2, n-2}] \ [t_{\alpha/2, n-2}, +\infty)$$

4. Kiểm tra: $t \in D$ hay $t \notin D$?

5. Kết luận: Với mức ý nghĩa α, bác bỏ hay chấp nhận giả thuyết

Ví dụ 3: Hãy kiểm định giả thuyết

H_0: $\beta = 0$

H_1: $\beta \neq 0$

Với mức ý nghĩa 0,03.

GIẢI:

1. Đặt bài toán: $\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases}$

2. Chỉ tiêu kiểm định $T = \dfrac{b}{S / \sqrt{s_{xx}}}$

Từ đó tính

$$t = \frac{b}{s / \sqrt{s_{xx}}} = \frac{0{,}903643}{3{,}2295 / \sqrt{4152{,}18}} \approx 18{,}0302$$

3. Chọn mức ý nghĩa $\alpha = 0{,}03$. Suy ra miền bác bỏ

$$D = (-\infty; -2,278] \cup [2,278; +\infty)$$

4. Kiểm tra: $t \in D$.

5. Kết luận: Với mức ý nghĩa α, bác bỏ giả thuyết H_0.

2. Kiểm định giả thuyết hệ số α

Các bước làm bài toán:

1. H_0: $\alpha = \beta^*$

   H_1: $\alpha \neq \beta^*$

2. Chỉ tiêu kiểm định: $T = \dfrac{a - \beta^*}{S\sqrt{\sum_{i=1}^{n} x_i^2 / nS_{xx}}}$

Từ đó tính

$$t = \frac{a - \beta^*}{s\sqrt{\sum_{i=1}^{n} x_i^2 / nS_{xx}}}$$

3. Chọn mức ý nghĩa α. Miền bác bỏ

$$D = (-\infty, -t_{\alpha/2,\, n-2}] \cup [t_{\alpha/2,\, n-2}, +\infty)$$

4. Kiểm tra: $t \in D$ hay $t \notin D$.

5. Kết luận: Với mức ý nghĩa α, bác bỏ hay chấp nhận giả thuyết.

Ví dụ 5: Sử dụng bài 13, ví dụ 1 với giá trị ước lượng $a = 3,829640$. Hãy kiểm định giả thiết rằng $\alpha = 0$ với đối thuyết $\alpha = 0$, chọn mức ý nghĩa 0,05.

GIẢI:

1. Đặt bài toán: $\begin{cases} H_0 : \alpha = 0 \\ H_1 : \alpha \neq 0 \end{cases}$

2. Chỉ tiêu kiểm định: $T = \dfrac{a}{S\sqrt{\sum_{i=1}^{n} x_i^2 / nS_{xx}}}$

Từ đó tính được:

$$t = \frac{a}{s\sqrt{\sum_{i=1}^{n} x_i^2 / nS_{xx}}} = \frac{3,829633 - 0}{3,2295 / \sqrt{41086/(33) \times (4152,18)}} = 2,17$$

3. Với mức ý nghĩa α= 0,05. Miền bác bỏ là

$$D = (-\infty, -2,042] \cup [2,042, +\infty)$$

4. Kiểm tra: $t \in D$

5. Kết luận: Với mức ý nghĩa α, bác bỏ giả thuyết H_0.

III. DỰ ĐOÁN DỰA VÀO ĐƯỜNG HỒI QUY

Một trong những ứng dụng quan trọng của đường hồi quy là ước lượng giá trị trung bình và dự báo giá trị của biến phụ thuộc khi đã biết giá trị của biến độc lập.

1. Khoảng tin cậy đối với $\mu_{Y|x_0}$

Khi $X = x_0$, ước lượng điểm cho giá trị trung bình của Y là

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

Khoảng tin cậy (1- α)100% cho $\mu_{Y|x_0}$ là

$$\hat{y}_0 - t_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < \mu_{Y|x_0} < \hat{y}_0 + t_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Với $t_{\alpha/2}$ là của giá trị của phân phối t với $n-2$ bậc tự do.

2. Khoảng dự đoán cho giá trị của biến *Y*

Khi $X = x_0$, khoảng dự đoán (1- α)100% cho giá trị tương ứng y_0 của Y là

$$\hat{y}_0 - t_{\alpha/2} s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < y_0 < \hat{y}_0 + t_{\alpha/2} s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

trong đó $t_{\alpha/2}$ là giá trị của phân phối t với $n-2$ bậc tự do.

Ví dụ 3: Dùng bảng số liệu ví dụ 1, bài 13.

a) Tìm khoảng tin cậy 95% cho kỳ vọng có điều kiện $\mu_{Y|x_0}$, với $x_0 = 20$;

b) Hãy xây dựng khoảng dự đoán 95% cho y_0 khi $x_0 = 20$.

GIẢI:

- Từ phương trình hồi quy, với $x_0 = 20$, ta tìm

$$\widehat{y_0} = 3,829633 + (0,903643)(20) = 21,9025.$$

- Ta đã tính được

$\bar{x} = 33,4545$; $S_{xx} = 4152,18$; $s = 3,2295$,

$t_{0,025} \cong 2,042$, với số bậc tự do là 31.

a) Khoảng tin cậy 95% cho $\mu_{Y|20}$ là :

$$\hat{y}_0 - t_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} < \mu_{Y|x_0} < \hat{y}_0 + t_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

Thay số,

$$21,9025 - 2,042 \times 3,2295 \times \sqrt{\frac{1}{33} + \frac{(20 - 33,4545)^2}{4152,18}} < \mu_{Y|20} <$$

$$21,9025 + 2,042 \times 3,2295 \times \sqrt{\frac{1}{33} + \frac{(20 - 33,4545)^2}{4152,18}}$$

Tính toán ta được: $20,467 < \mu_{Y|20} < 23,3380$.

b) Khoảng dự đoán 95% cho y_0 là

$$\hat{y}_0 - t_{\alpha/2} s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} < y_0 < \hat{y}_0 + t_{\alpha/2} s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

Thay số:

$$21,9025 - 2,042 \times 3,2295 \times \sqrt{1 + \frac{1}{33} + \frac{(20 - 33,4545)^2}{4152,18}} < y_0 <$$

$$21,9025 + 2,042 \times 3,2295 \times \sqrt{1 + \frac{1}{33} + \frac{(20 - 33,4545)^2}{4152,18}}$$

Tính toán ta được: $15,2340 < y_0 < 28,5710$

IV. TƯƠNG QUAN TUYẾN TÍNH

Như ta đã biết, khái niệm hệ số tương quan dùng để đo mức độ phụ thuộc tuyến tính giữa hai biến ngẫu nhiên

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

Người ta đã chứng minh được rằng:

- $\rho_{XY} \in [-1, 1]$
- Khi $\rho_{XY} = 0$, thì không có tương quan tuyến tính giữa *X* và *Y*;
- Khi $|\rho_{XY}|$ càng gần 1, thì sự phụ thuộc tuyến tính giữa *X* và *Y* càng mạnh;
- Khi $|\rho_{XY}| = 1$, sự phụ thuộc tuyến tính là mạnh nhất. Khi đó, *Y* là một hàm tuyến tính của *X* tức là tồn tại hai số *a* và *b* sao cho

$$Y = aX + b.$$

Muốn biết được ρ_{XY}, ta phải có phân phối xác suất của (*X*, *Y*). Khi ta không biết phân phối xác suất của (*X*, *Y*), thì đặt ra bài toán là ước lượng ρ_{XY} và kiểm định giả thuyết về giá trị của ρ_{XY} dựa vào một mẫu cụ thể (x_1, y_1), (x_2, y_2), (x_3, y_3),..., (x_n, y_n).

Khi đó ta quan tâm đến việc ước lượng cho ρ_{XY} bởi giá trị từ một mẫu:

$$r = b\sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

r được gọi là hệ số tương quan tuyến tính mẫu.

Lưu ý rằng r cũng là một số nằm trong [-1, 1], nên khi tính toán mà thu được r với $|r| > 1$ thì nghĩa là ta đã tính toán sai.

Ví dụ 6: Việc nghiên cứu sự tương quan giữa kết cấu và các thuộc tính cơ học của gỗ là việc làm quan trọng đối với các nhà khoa học nghiên cứu về lâm sản.

Khối lượng riêng, x (g/cm^3)	Cường độ chịu kéo, y (kPa)	Khối lượng riêng, x (g/cm^3)	Cường độ chịu kéo, y (kPa)
0,414	29,186	0,581	85,156
0,383	29,266	0,557	69,571
0,399	26,215	0,550	84,160
0,402	30,162	0,531	73,466