Bài 13: HỒI QUY VÀ TƯƠNG QUAN TUYẾN TÍNH

Vũ Mạnh Tới

Bộ môn Toán-Trường Đại học Thủy lợi

Ngày 30 tháng 5 năm 2024

- 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.

- 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. 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 $Y|x_i$, với trung bình và phương sai $\mu_{Y|x_i}, \sigma^2_{Y|x_i}$.
- Khi nối các điểm $(\mu_{Y|x_i}, x_i)$ 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: $\hat{y} = a + bx$.

- 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: $y_i = a + bx_i + e_i$.

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

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

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

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

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

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

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

giải để tìm ra công thức tính cho a và b.

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

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

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

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

13.2. Ước lượng không chệch của phương sai

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

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

Để tính toán những số liệu trên, ta sử dụng:

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

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y},\ b = \frac{S_{xy}}{S_{xx}}$$

Ví dụ

Một cuộc nghiên cứu về lượng mưa hàng ngày x (0,01 cm) và lượng ô nhiễm không khí thải ra y (mcg/cum) cho bởi số liệu sau:

x	4,3	4,5	5,9	5,6	6,1	5,2	3,8	2,1	7,5
y	126	121	116	118	114	118	132	141	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.

Bài tập về nhà

Trang 388-391: 2, 3, 4, 6, 7.