Chương 4: HỒI QUI BAYES

4.1 Mô hình hồi quy đơn biến

Giả sử chúng ta muốn mô hình biểu diễn mối quan hệ giữa hai biến x và y, thông thường chúng ta muốn sử dụng giá trị của x giúp dự báo giá trị y thông qua sử dụng mối quan hệ giữa hai biến x và y nêu ở trên.

Dữ liệu bao gồm n quan sát $(x_i, y_i), i = \overline{1,n}$. Trong trường hợp đơn giản nhất, mô hình hồi quy có dạng

$$y = \alpha + \beta x + \epsilon,$$

Trong đó ϵ là nhiễu (sai số), được giả định tuân theo phân phối chuẩn với trung bình bằng 0 và phương sai σ^2.

Trong thống kê tần suất, các giá trị α, β được xem xét như các hằng số. Do đó, pương pháp bình phương cực tiểu nhằm cực tiểu hóa bình phương các sai lệch.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2 \Rightarrow min$$

Trong đó các tham số được ước lượng theo công thức

$$\begin{cases} \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} \\ \hat{\beta} = \dfrac{S_{xy}}{S_{xx}} = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \end{cases}$$

Dựa vào các giả thuyết trong mô hình hồi quy tuyến tính về nhiễu ϵ tuân theo phân phối chuẩn đồng thời sai số của tất cả các quan sát độc lập với nhau. Khi đó, các suy luận về dạng phân phối xác suất của các tham số

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right), \hat{\alpha} \sim N\left(\alpha, \frac{\sum_{i=1}^{n} x_i^2}{n} \frac{\sigma^2}{S_{xx}}\right)$$

Suy ra các bài toán về khoảng tin cậy cho các tham số của mô hình hồi quy với độ tin cậy $(1 - \alpha)$ là

$$\alpha \in \left(\hat{\alpha} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} \frac{\sigma^2}{S_{xx}}}\right), \beta \in \left(\hat{\beta} \pm \sqrt{\frac{\sigma^2}{S_{xx}}}\right)$$

Bài toán kiểm định các tham số hồi quy, cũng như dự báo cho quan sát tiếp theo đều dựa

vào dạng phân phối xác suất của các tham số.

Ví dụ 1. Các kết quả của một cuộc khảo sát được tổng kết lại và tính toán các kết quả bằng phần mềm như bảng sau

Code R

```
x <- c(14.36, 14.48, 14.53, 14.52, 14.35, 14.31, 14.44, 14.23, 14.32, 14.57, 14.28, 14.36, 14.50,
    14.52, 14.28, 14.13, 14.54, 14.60, 14.86, 14.28, 14.09, 14.20, 14.50, 14.02, 14.45)

y <- c(13.84, 14.41, 14.22, 14.63, 13.95, 14.37, 14.41, 13.99, 13.89, 14.59, 14.32, 14.31, 14.43,
    14.44, 14.14, 13.90, 14.37, 14.34, 14.78, 13.76, 13.85, 13.89, 14.22, 13.80, 14.67)

c(mean(x), mean(y), sum(x^2)/length(x), sum(y^2)/length(y), sum(x*y)/length(x))
```

Khi đó, các giá trị tương ứng là

$\overline{x}$ =14.3888, $\overline{y}$ = 14.2208, $\overline{x^2}$ =207.0703, $\overline{y^2}$ =202.3186, $\overline{xy}$ =204.6628

Các tham số ước lượng được như sau

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - (\overline{x})^2} = 1.299635$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} = -4.479383$$

Hoặc đơn giản chỉ cần một lệnh trong R để tính ra các giá trị ước lượng cho α và β là

Code R

```
lm(y~x)
```

Kết quả như sau

```
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     -4.479        1.300
```

Ví dụ 2. Các kết quả ước lượng về phân phối xác suất của các tham số

Code R

```
linearModel <- lm(y~x)
modelSummary <- summary(linearModel)
modelSummary
```

```
> modelSummary

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34337 -0.14532  0.01753  0.12266  0.36966

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.4794     2.9692  -1.509    0.145
x             1.2996     0.2063   6.298 1.99e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1867 on 23 degrees of freedom
Multiple R-squared:  0.633,     Adjusted R-squared:  0.617
F-statistic: 39.67 on 1 and 23 DF,  p-value: 1.994e-06
```

Ngoài ra, chúng ta có thể tính được một số các tham số trong lựa chọn mô hình như AIC, BIC:

Code R
c(AIC(linearModel), BIC(linearModel))

```
> c(AIC(linearModel), BIC(linearModel))
[1] -9.044825 -5.388198
```

4.2 Hồi qui tuyến tính đơn Bayes

Hàm hợp lý cho các quan sát

Hàm hợp lý cho quan sát thứ i là

$$l(x_i|\alpha,\beta) \propto \exp\left[-\frac{1}{2\sigma^2}(y_i-\alpha-\beta x_i)^2\right]$$

Hàm hợp lý cho mẫu gồm n quan sát là

$$l(x_1,x_2,\cdots,x_n|\alpha,\beta) \propto \prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}(y_i-\alpha-\beta x_i)^2\right]$$

Hay công thức tương đương

$$l(x_1,x_2,\cdots,x_n|\alpha,\beta) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2\right]$$

Phân phối tiên nghiệm cho các tham số

$$\pi(\alpha,\beta) = \pi(\alpha) \times \pi(\beta)$$

Chúng ta có thể chọn phân phối tiên nghiệm cho các tham số α và β trong trường hợp đơn giản là phân phối đều hoặc phân phối chuẩn.

Phân phối hậu nghiệm cho các tham số

$$\pi(\alpha, \beta | x_1, x_2, \cdots, x_n) \propto \pi(\alpha, \beta) \times l(x_1, x_2, \cdots, x_n | \alpha, \beta)$$

Do đó, nếu thông tin tiên nghiệm giả sử các tham số tuân theo phân phối đều hoặc phân phối chuẩn thì phân phối hậu nghiệm của các tham số cũng tuân theo phân phối chuẩn.

Cụ thể, giả sử phân phối tiên nghiệm cho tham số β là $\pi(\beta) \sim N(m_\beta, s_\beta^2)$ và hàm hợp lý cho tham số β là $N\left(\hat{\beta}, \frac{\sigma^2}{S_{xx}}\right)$. Khi đó, phân phối hậu nghiệm cho tham số β là $\pi(\beta | x_1, x_2, \cdots, x_n) \sim N(m'_\beta, s'^2_\beta)$ với các tham số được xác đinh như sau

$$\frac{1}{s'^2_\beta} = \frac{1}{s_\beta^2} + \frac{S_{xx}}{\sigma^2}$$

$$m'_\beta = \frac{\frac{1}{s_\beta^2}}{\frac{1}{s'^2_\beta}} m_\beta + \frac{\frac{S_{xx}}{\sigma^2}}{\frac{1}{s'^2_\beta}} \hat{\beta}$$

Tương tự, đối với phân phối tiên nghiệm cho tham số α là $\pi(\alpha) \sim N(m_\alpha, s_\alpha^2)$ và hàm hợp lý cho tham số α là $N\left(\hat{\alpha}, \frac{\sum_{i=1}^n x_i^2}{n} \frac{\sigma^2}{S_{xx}}\right)$. Khi đó, phân phối hậu nghiệm cho tham số α là $\pi(\alpha | x_1, x_2, \cdots, x_n) \sim N(m'_\alpha, s'^2_\alpha)$ với các tham số được xác định như sau

$$\frac{1}{s'^2_\alpha} = \frac{1}{s_\alpha^2} + \frac{n}{\sum_{i=1}^n x_i^2} \frac{S_{xx}}{\sigma^2}$$

$$m'_\alpha = \frac{\frac{1}{s_\alpha^2}}{\frac{1}{s'^2_\alpha}} m_\alpha + \frac{\frac{n}{\sum_{i=1}^n x_i^2} \frac{S_{xx}}{\sigma^2}}{\frac{1}{s'^2_\alpha}} \hat{\alpha}$$

Trong đó σ^2 chưa biết được ước lượng bằng $\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}$

Code R

```
coef <- c(linearModel$coefficients)
err <- y - coef[1]- coef[2]*x
sighatsq <- sum(err^2)/(length(x)-2)
sighatsq
```

Kết quả thu được giá trị ước lượng σ^2 là

```
> sighatsq
[1] 0.03486442
```

Hoặc sử dụng code R tương đương

Code R

```
sum(linearModel$residuals^2)/(length(x)-2)
```

Với kết quả tương tự

```
> sum(linearModel$residuals^2)/(length(x)-2)
[1] 0.03486442
```

Khoảng ước lượng cho các tham số

Do $\pi(\beta|x_1, x_2, \cdots, x_n) \sim N(m'_\beta, s'^2_\beta)$ nên khoảng ước lượng cho tham số β với độ tin cậy $(1-\alpha)$ trong trường hợp đã biết σ^2 là

$$m'_\beta \pm z_{\frac{\alpha}{2}}\sqrt{s'^2_\beta}$$

Và trong trường hợp chưa biết σ^2 được tính toán thông qua $\hat{\sigma}^2$ là

$$m'_\beta \pm t_{n-2,\frac{\alpha}{2}}\sqrt{s'^2_\beta}$$

4.3 Thuật toán Metropolis-Hastings

Gibbs sampler

Giả sử vectơ tham số $\theta = (\theta_1, \theta_2, \cdots, \theta_d)$. Các xác suất thành phần của θ là

$$p(\theta_j|\theta_{-j}^{t-1}, y)$$

Trong đó θ_{-j}^{t-1} tương ứng là các thành phần của θ, ngoại trừ cho θ_j:

$$\theta_{-j}^{t-1} = (\theta_1^t, \cdots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \cdots, \theta_d^{t-1})$$

Thuật toán Metropolish

Thuật toán Metropolish là một hiệu chỉnh của bước ngẫu nhiên với quy tắc chấp nhận/bác bỏ hội tụ tới phân phối xuất phát cụ thể. Các bước của thuật toán như sau:

- Điểm xuất phát θ^0, trong đó $p(\theta^0|y) > 0$ với phân phối xuấy phát $p_0(0)$.
- Với $t = 1,2, \cdots$
    - Phân phối đề nghị θ^* với phân phối nhảy tại thời điểm t là $J_t(\theta^*|\theta^{t-1})$, sao cho phân phối nhảy có tính chất đối xứng $J_t(\theta_a|\theta_b) = J_t(\theta_b|\theta_a)$ với mọi θ_a, θ_b, t.
    - Tính tỷ lệ các hàm mật độ

$$r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)}$$

    - Đặt $\theta^t = \begin{cases} \theta^* & với\ xác\ suất\ \min(r,1) \\ \theta^{t-1} & trong\ các\ trường\ hợp\ khác \end{cases}$

- Với giá trị hiện tại θ^{t-1}, phân phối chuyển $T_t(\theta^t|\theta^{t-1})$ của chuỗi Markov là hỗn hợp của các điểm rời rạc tại $\theta^t = \theta^{t-1}$. Khi $\theta^t = \theta^{t-1}$ có nghĩa là bước nhảy không được chấp nhận.

Thuật toán Metropolis-Hastings

Thuật toán Metropolis-Hastings là trường hợp tổng quát của thuật toán Metropolis trong đó phân phối nhảy mở rộng không cần phân phối đối xứng, với tỷ lệ r là tỷ lệ của các tỷ lệ

$$r = \frac{\frac{p(\theta^*|y)}{J_t(\theta^*|\theta^{t-1})}}{\frac{p(\theta^{t-1}|y)}{J_t(\theta^{t-1}|\theta^*)}}$$

4.4 Mô hình Bayes trung bình

Mô hình Bayes trung bình nhằm chọn mô hình đơn tốt nhất trong số các mô hình có thể có phù hợp với dữ liệu.

Giả sử có r mô hình $M_1, M_2, \cdots, M_r$. Chúng ta tính xác suất hậu nghiệm của mô hình M_k với tập dữ liệu $X_n = \{x_1, x_2, \cdots, x_n\}$ được xác định bởi công thức

$$P(M_k|X_n) = \frac{P(M_k)\int f_k(X_n|\theta_k)\pi_k(\theta_k)d\theta_k}{\sum_{j=1}^{r} P(M_j)\int f_j(X_n|\theta_j)\pi_j(\theta_j)d\theta_j}.$$

Phân phối dự báo cho quan sát tương lai z là $f(z|X_n)$ được xác định bởi công thức

$$f(z|X_n) = \sum_{j=1}^{r} P(M_j|X_n) f_j(z|X_n),$$

Trong đó $f_j(z|X_n) = \int f_j(z|\theta_j)\pi_j(\theta_j|X_n)d\theta_j, j = 1,2,\cdots,r.$

Phân phối dự báo $f(z|X_n)$ là trung bình của các phân phối dự báo dựa vào tất cả các mô hình được xem xét, với trọng là xác suất các mô hình hậu nghiệm tương ứng (Ando, 2010).

Mô hình Bayes trung bình cho mô hình hồi quy tuyến tính

$$y_n = X_{jn}\beta_j + \epsilon_{jn},$$

Trong đó y_n là vectơ $n \times 1$ của biến quan sát mà chúng tôi muốn dự báo, X_{jn} là các ma trận $n \times p_j$ các quan sát ảnh hưởng trong dự báo, β_j là vectơ $p_j \times 1$ các tham số, ϵ_{jn} là vectơ các sai số, trong đó các sai số độc lập, có phân phối giống hệt nhau với trung bình 0 và phương sai σ^2.

Tài liệu tham khảo

Ando, T. (2010). *Bayesian model selection and statistical modeling.* Chapman and Hall/CRC.

Bolstad, W. M., & Curran, J. M. (2016). *Introduction to Bayesian statistics.* John Wiley & SonsJohn Wiley & Sons.

Dorfman, J. H. (1997). *Bayesian economics through numerical methods: a guide to econometrics and decision-making with prior information.* Springer Science & Business Media.

Ghosh, J. K., Delampady, M., & Samanta, T. (2007). *An introduction to Bayesian analysis: theory and methods.* Springer Science & Business Media.

Lê Hồng Nhật, Phạm Văn Chững, Võ Thị Lệ Uyển & Lê Thanh Hoa. (2017). *Giáo trình Kinh tế lượng.* NXB Đại học Quốc gia Thành phố Hồ Chí Minh.

Lindley, D. V. (2011). *Introduction to probability and statistics from a Bayesian viewpoint, Part 2, Inference.* CUP Archive.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists.* Springer Science & Business Media.

Phạm Hoàng Uyên, Lê Thị Thiên Hương, Huỳnh Văn Sáu, Nguyễn Phúc Sơn & Huỳnh Tố Uyên. (2016). *Lý thuyết xác suất.* Nhà xuất bản Đại học Quốc gia Thành phố Hồ Chí Minh.

Phạm Văn Chững, Nguyễn Đình Uông & Lê Thanh Hoa. (2016). *Thống kê ứng dụng.* Nhà xuất bản Đại học Quốc gia Thành phố Hồ Chí Minh.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation.* Springer Science & Business Media.

Singpurwalla, N. D. (2006). *Reliability and risk: a Bayesian perspective.* John Wiley & Sons.