

Mục lục	
Cơ sở lý thuyết	5
1.1 Tổng quan về thống kê Bayes	5
1.1.1 Thống kê là gì?	5
1.1.2 Thống kê Bayes là gì?.....	5
1.2 Các khái niệm cơ bản trong thống kê Bayes.....	6
1.2.1 Định lý Bayes:	6
1.2.2 Ước lượng điểm	18
1.2.3 Ước lượng khoảng:	20
1.2.4 Kiểm định giả thuyết	21
1.2.5 Phương pháp Monte Carlo.....	23
1.2.6 Xích Markov.....	25
Mô hình hồi quy tuyến tính	31
2.1. Mô hình hồi quy tuyến tính.....	31
2.1.1 Ước lượng bình phương tối thiểu	32
2.2 Ước lượng Bayes cho mô hình hồi quy	33
2.3 Chọn mô hình (model selection).....	38
2.3.1 So sánh các mô hình Bayes	42
Mô hình hồi quy với các biến rời rạc có thứ tự.	45
3.1 Hồi quy probit theo thứ tự và bậc likelihood	45
3.1.1 Hồi quy probit.....	46
3.1.2 Mô hình chuyển đổi và rank likelihood.....	49
Tài liệu tham khảo	53

Chương 1:

Cơ sở lý thuyết

1.1 Tổng quan về thống kê Bayes

1.1.1 Thống kê là gì?

Thế giới xung quanh chúng ta có rất nhiều điều, rất nhiều sự vật hiện tượng khoa học đã khám phá ra nhưng bên cạnh đó còn có rất nhiều sự bí ẩn còn tồn tại chờ khoa học giải đáp. Mỗi ngành khoa học có bộ dữ liệu khác nhau như bộ dữ liệu về bệnh nhân ung thư, tiêu đường trong y khoa, bộ dữ liệu về doanh thu trong kinh tế....

Và vấn đề hầu hết các ngành khoa học nào cũng cần phải giải quyết đó là những giữ liệu giá trị đó đang muốn cho chúng ta biết điều gì? Với mục đích đi tìm câu trả lời trên ngành khoa học thống kê đã ra đời.

Vậy thống kê là gì? Thống kê là khoa học về các phương pháp tổng quát xử lí các kết quả thực nghiệm.

1.1.2 Thống kê Bayes là gì?

Thống kê Bayes bắt đầu được ra đời từ khoảng thế kỷ thứ 18 (dựa trên một định lý toán xác suất cơ bản). Nhưng phải đến những năm cuối thế kỷ 20 khi mà các công nghệ tính toán bùng nổ, khi mà suy luận thống kê đã có những nền tảng vững chắc với những công trình của Ronald Fisher, Karl Pearson, Jerzy Neyman....thì suy luận thống kê mới ngày càng được chú ý tới và sử dụng ngày càng rộng rãi cho hầu hết các lĩnh vực của đời sống. Thống kê Bayes là phương pháp thống kê dựa trên một công cụ duy nhất là định lý Bayes để cung cấp thêm sự tin tưởng của chúng ta về các dự liệu được đưa ra. Nếu như thống kê toàn suất (thống kê cổ điển) xem tham số là một giá trị không biết nhưng không ngẫu nhiên thì thống kê Bayes coi tham số là biến ngẫu nhiên. Và chúng ta có thể gán cho tham số một phân phối xác suất để biểu thị sự tin cậy về giá trị thực của tham số. Cuối cùng bằng cách kết hợp thông tin đã có trước khi quan sát với thông tin có được khi quan sát, chúng ta thu được thông tin muốn biết. Đây chính là điểm khác biệt cơ bản giữa hai “trường phái” thống kê.

Một trong những điều thống kê Bayes làm rất tốt đó chính là việc sử dụng được các kiến thức, các nguồn thông tin tiên nghiệm và thông tin chứa trong dữ liệu. Chúng

được kết hợp trong định lý Bayes. Còn thống kê tần suất thì bỏ qua các kiến thức về tiên nghiệm.

Thống kê Bayes còn là cách đơn giản để xử lý những vấn đề xảy ra về tham số và định lý Bayes đưa ra được các cách để tìm phân phối dự đoán của các quan sát trong tương lai. Đây không phải là điều dễ thực hiện trong tần suất.

Các mô hình Bayes đa tăng tính linh hoạt và tăng sự lựa chọn cho việc phân tích dữ liệu hơn rất nhiều đối với khoa học dữ liệu nói chung.

1.2 Các khái niệm cơ bản trong thống kê Bayes

1.2.1 Định lý Bayes:

Theo xác suất có điều kiện:

Ta có hai sự kiện ngẫu nhiên xảy ra gọi là A và B

Nếu A và B là hai sự kiện độc lập, ta có xác suất để đồng thời xảy ra A và B là:

$$P(A \cap B) = P(A)P(B)$$

(1.1)

Trong đó :

$P(A)$ là xác suất xảy ra A riêng biệt.

$P(B)$ là xác suất xảy ra B riêng biệt.

Nếu A và B là hai sự kiện liên quan đến nhau, và xác suất xảy ra sự kiện B lớn hơn 0, ta có thể định nghĩa xác suất xảy ra A khi biết B như sau:

$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$

(1.2)

Ta có $P(A)$ (xác suất biên của A) được tính theo công thức:

$$P(A) = P(A \cap B) + P(A \cap \bar{B})$$

(1.3)

Trong đó: \bar{B} là phần bù của B .

Thay vào công thức (1.2) ta được:

$$P(B|A) = \frac{P(A \cap B)}{P(A \cap B) + P(A \cap \bar{B})} \quad (1.4)$$

Sử dụng công thức nhân xác suất ta có:

$$P(B|A) = \frac{P(B) \times P(A|B)}{P(B) \times P(A|B) + P(\bar{B}) \times P(A|\bar{B})} \quad (1.5)$$

Công thức xác suất đầy đủ: B_1, B_2, \dots, B_n là nhóm đầy đủ các biến cốt. Giả sử biến cốt A có thể xảy ra đồng thời với một trong các biến cốt B_1, B_2, \dots, B_n thì ta có:

$$P(A) = \sum_{j=1}^n P(A \cap B_j)$$

Sử dụng công thức nhân xác suất ta có:

$$P(A) = \sum_{j=1}^n P(B_j) \times P(A|B_j)$$

Vậy ta có công thức Bayes:

$$P(B_j|A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(B_j) \times P(A|B_j)}{\sum_{j=1}^n P(B_j) \times P(A|B_j)} \quad (1.6)$$

Các biến cốt B_1, B_2, \dots, B_n gọi là các giả thiết. Các xác suất $P(B_1), P(B_2), \dots, P(B_n)$ được xác định trước khi phép thử được tiến hành, do đó gọi là xác suất biến duyên hay xác suất tiên nghiệm (priors)

$P(A|B_i)$ là xác suất có điều kiện của A nếu biến B_i xảy ra còn được gọi là hàm hợp lí (likelihood).

Các xác suất $P(B_1|A), \dots, P(B_n|A)$ được xác định sau khi các phép thử đã được tiến hành và biến cốt A đã xảy ra, do đó gọi là xác suất hậu nghiệm (posterior).

Định lí Bayes cho biến ngẫu nhiên rời rạc.

Biến ngẫu nhiên là gì?

Trong suy luận Bayes, một biến ngẫu nhiên được định nghĩa là một đại lượng số chưa biết mà chúng ta sẽ đưa ra các báo cáo xác suất.

Ví dụ: kết quả định lượng của một cuộc khảo sát, thí nghiệm hoặc nghiên cứu là một biến ngẫu nhiên trước khi nghiên cứu được thực hiện.

Biến ngẫu nhiên rời rạc là gì?

Đặt Y là một biến ngẫu nhiên rời rạc và Y là tập hợp tất cả các giá trị có thể có của Y . Ta nói rằng biến Y ngẫu nhiên rời rạc nếu tập hợp các kết quả có thể đếm được, có nghĩa là Y có thể được biểu thị bằng:

$$Y = \{y_1, y_2, \dots\}$$

Xét tham số là biến ngẫu nhiên X bao gồm các giá trị x_1, \dots, x_n . Y là một biến ngẫu nhiên phụ thuộc vào tham số, với các giá trị y_1, y_2, \dots, y_J . Chúng ta sẽ suy luận về biến ngẫu nhiên X dựa trên điều kiện $Y = y_m$ bằng việc sử dụng định lý Bayes.

Gọi f là phân phối xác suất (có điều kiện hoặc không có điều kiện) của quan sát ngẫu nhiên Y , g là xác suất của biến ngẫu nhiên X . Không gian Bayes bao gồm các cặp (x_i, y_j) với $i = 1, \dots, n$ và $j = 1, \dots, J$. Và xác suất của từng cặp trong không gian Bayes được tìm bởi công thức:

$$f(x_i, y_j) = g(x_i) \times f(y_j | x_i)$$

Trong đó:

$g(x_i)$ với $i = 1, \dots, n$ là xác suất tiên nghiệm của tham số X

$f(y_j | x_i)$ với $i = 1, \dots, n$ là hàm hợp lí (likelihood). Và đây chính là xác suất có điều kiện của Y khi biết giá trị $X = x_i$.

Vậy xác suất hậu nghiệm $g(x_i | y_j)$

$$g(x_i | y_j) = \frac{g(x_i) \times f(y_j | x_i)}{\sum_{i=1}^n g(x_i) \times f(y_j | x_i)}$$

Ví dụ 1.1: Một chủ tiệm chim có 5 con chim cu gáy ở trong một cái lồng rộng được phủ vải che kín để chim mau “chịu lòng”. Có hai loại trong lồng là chim cu lừa và chim cu luồng. Ta không biết có chính xác bao nhiêu con chim cu lừa, bao nhiêu con cu luồng. Gọi X là số con chim cu lừa có trong lồng. Vậy giá trị của X có thể là

x_i với $i = 0, 1, 2, 3, 4, 5$. Và các giá trị này có khả năng xảy ra là nhau nhau nên xác suất tiên nghiệm của X là:

$$g(0) = g(1) = g(2) = g(3) = g(4) = g(5) = \frac{1}{6}$$

Người mua chim lấy ngẫu nhiên một con trong lồng, gọi $Y = 1$ là biến ngẫu nhiên lấy được con chim cu lứa và $Y = 0$ là biến ngẫu nhiên lấy được con chim cu luồng.

Tìm $P(X = x_i | Y = 1)$.

Ta đi tìm hàm hợp lí (likelihood): $P(Y = 1 | X = x_i) = \frac{i}{5}$ và $P(Y = 0 | X = x_i) = \frac{(5-i)}{5}$

Giả sử con chim lấy ra đầu tiên là con chim cu lứa.

Ta có bảng sau:

x_i	Xác suất tiên nghiệm	$P(x_i)P(y_j = 0 X = x_i)$	$P(x_i)P(y_j = 1 X = x_i)$
0	$\frac{1}{6}$	$\frac{1}{6} \times \frac{5}{5} = \frac{1}{6}$	$\frac{1}{6} \times \frac{0}{5} = 0$
1	$\frac{1}{6}$	$\frac{1}{6} \times \frac{4}{5} = \frac{2}{15}$	$\frac{1}{6} \times \frac{1}{5} = \frac{1}{30}$
2	$\frac{1}{6}$	$\frac{1}{6} \times \frac{3}{5} = \frac{1}{10}$	$\frac{1}{6} \times \frac{2}{5} = \frac{1}{15}$
3	$\frac{1}{6}$	$\frac{1}{6} \times \frac{2}{5} = \frac{1}{15}$	$\frac{1}{6} \times \frac{3}{5} = \frac{1}{10}$
4	$\frac{1}{6}$	$\frac{1}{6} \times \frac{1}{5} = \frac{1}{30}$	$\frac{1}{6} \times \frac{4}{5} = \frac{2}{15}$
5	$\frac{1}{6}$	$\frac{1}{6} \times \frac{0}{5} = 0$	$\frac{1}{6} \times \frac{5}{5} = \frac{1}{6}$
$f(y_j)$		$\frac{1}{2}$	$\frac{1}{2}$

Bảng 1.1

Vậy xác suất hậu nghiệm $P(X | Y = 1)$ là:

x_i	Xác suất tiên nghiệm	tnx hhl	Hậu nghiệm
0	$\frac{1}{6}$	$\frac{1}{6} \times \frac{0}{5} = 0$	0
1	$\frac{1}{6}$	$\frac{1}{6} \times \frac{1}{5} = \frac{1}{30}$	$\frac{1}{30} / \frac{1}{2} = \frac{1}{15}$
2	$\frac{1}{6}$	$\frac{1}{6} \times \frac{2}{5} = \frac{1}{15}$	$\frac{2}{30} / \frac{1}{2} = \frac{2}{15}$

3	$\frac{1}{6}$	$\frac{1}{6} \times \frac{3}{5} = \frac{1}{10}$	$\frac{3}{30} / \frac{1}{2} = \frac{3}{15}$
4	$\frac{1}{6}$	$\frac{1}{6} \times \frac{4}{5} = \frac{2}{15}$	$\frac{4}{30} / \frac{1}{2} = \frac{4}{15}$
5	$\frac{1}{6}$	$\frac{1}{6} \times \frac{5}{5} = \frac{1}{6}$	$\frac{5}{30} / \frac{1}{2} = \frac{5}{15}$
$f(y=1)$		$\frac{1}{2}$	

Bảng 1.2

Cũng với đề bài trên, giả sử sau khi lấy con chim cu lừa ở lần thứ nhất mà không trả lại vào lồng, ta tiếp tục lấy một con chim thứ hai được con chim cu luồng. Và ta muốn tính hậu nghiệm dựa trên hai kết quả đó. Ta sẽ phân tích theo trình tự của từng giai đoạn.

Kết quả xác suất hậu nghiệm của lần lấy đầu tiên sẽ được làm xác xuất tiên nghiệm cho kết quả thứ hai.

Ta có bảng sau:

x_i	Xác suất tiên nghiệm	Hàm hợp lý	$tn \times hhl$	Hậu nghiệm
0	0	1	0	0
1	$\frac{1}{15}$	$\frac{4}{4}$	$\frac{1}{15}$	$\frac{1}{15} / \frac{1}{3} = \frac{1}{5}$
2	$\frac{2}{15}$	$\frac{3}{4}$	$\frac{1}{10}$	$\frac{1}{10} / \frac{1}{3} = \frac{3}{10}$
3	$\frac{1}{5}$	$\frac{1}{2}$	$\frac{1}{10}$	$\frac{1}{10} / \frac{1}{3} = \frac{3}{10}$
4	$\frac{4}{15}$	$\frac{1}{4}$	$\frac{1}{15}$	$\frac{1}{15} / \frac{1}{3} = \frac{1}{5}$
5	$\frac{1}{3}$	$\frac{0}{4}$	0	$0 / \frac{1}{3} = 0$
$f(y=0)$			$\frac{1}{3}$	

Bảng 1.3

Định lí Bayes cho biến ngẫu nhiên liên tục

Giả sử không gian mẫu của Y tương đương vs \sim . Chúng ta không thể định nghĩa $P(Y \leq 9) = \sum_{y \leq 9} p(y)$ vì tổng này không có nghĩa (tập hợp các số thực nhỏ hơn hoặc bằng 9 là không đếm được). Vì vậy thay vì việc xác định xác suất của các sự kiện theo pdf $p(y)$ chúng ta sẽ xác định theo cdf (hàm phân phối tích lũy):

$$F(y) = P(Y \leq y)$$

Lưu ý: $F(\infty) = 1$; $F(-\infty) = 0$ và $F(b) \leq F(a)$ nếu $b \leq a$. Xác suất của các biến cố khác nhau có thể được lấy từ cdf:

$$\begin{aligned} P(Y > a) &= 1 - F(a) \\ P(a < Y \leq b) &= F(b) - F(a) \end{aligned}$$

Nếu F là liên tục chúng ta nói rằng Y là một biến ngẫu nhiên liên tục. Và ta có:

$$F(a) = \int_{-\infty}^a p(y) dy$$

Hàm này được gọi là hàm mật độ xác suất của Y và có các thuộc tính:

$$\begin{aligned} p(y) &\geq 0 \quad \forall y \in \mathbb{Y} \\ \int_{-\infty}^{+\infty} p(y) dy &= 1 \end{aligned}$$

Xác xuất của biến ngẫu nhiên trong khoảng (a, b) được tính bởi tích phân của hàm mật độ xác suất trên khoảng đó:

$$P(a < Y < b) = \int_a^b p(y) dy$$

Công thức hậu nghiệm với biến ngẫu nhiên rời rạc:

$$g(x|y) = \frac{g(x) \times f(y|x)}{\sum_{x \in X} g(x) \times f(y|x)}$$

Và công thức với biến ngẫu nhiên liên tục là:

$$g(x|y) = \frac{g(x) \times f(y|x)}{\int_{-\infty}^{+\infty} g(x) \times f(y|x)} \quad (1.7)$$

Một số phân phối liên tục:

Phân phối Gamma:

Phân phối Gamma được sử dụng cho biến ngẫu nhiên liên tục với giá trị $0 \leq x \leq \infty$.
Hàm mật độ xác suất cho bởi:

$$g(x;r,v) = k \times x^{r-1} \times e^{-vx}$$

Trong đó $k = \frac{v^r}{\Gamma(r)}$.

Kì vọng phân phối Gamma là:

$$E(X) = \frac{r}{v}$$

Phương sai của phân phối Gamma là:

$$Var(x) = \frac{r}{v^2}$$

X có phân phối $Gamma(r, v)$ thì xác suất của X là:

$$P(X \leq x_0) = \int_0^{x_0} g(x; r, v) dx$$

Phân phối chuẩn (μ, σ^2)

$$g(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$$

Phân phối đều:

Biến ngẫu nhiên có phân phối đều $(0, 1)$ nếu hàm mật độ của nó là hằng số trên đoạn $[0, 1]$ và nhận giá trị 0 nếu ngược lại.

$$g(x) = \begin{cases} 1 & (0 \leq x \leq 1) \\ 0 & (x \notin [0, 1]) \end{cases}$$

Phân phối Beta

Phân bố Beta(a, b) là phân phối cho biến ngẫu nhiên liên tục với $0 \leq x \leq 1$.

Hàm mật độ xác suất cho bởi:

$$g(x; a, b) = \begin{cases} k \times x^{a-1} (1-x)^{b-1} & (0 \leq x \leq 1) \\ 0 & (x \notin [0, 1]) \end{cases}$$

Trong đó $k = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$. Phân phối đều $(0, 1)$ là dạng đặc biệt của phân phối Beta (a, b) với $a = b = 1$.

Giá trị kì vọng:

$$E(X) = \frac{a}{a+b}$$

Giá trị phương sai:

$$Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

Khi X có phân phối Beta, ta tính được xác suất như sau:

$$P(X \leq x_0) = \int_0^{x_0} g(x; a, b) dx$$

Định lí Bayes cho phân phối chuẩn với tiên nghiệm rời rạc.

Sau đây chúng ta sẽ cùng tìm hiểu về cách áp dụng định lí Bayes vào vấn đề cần giải quyết trong phân bố chuẩn.

Phân phối có điều kiện $y | \mu$ là phân phối chuẩn với giá trị trung bình là μ và phương sai là σ^2 . Hàm mật độ là:

$$f(y | \mu) = \frac{1}{2\pi\mu} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

Do vậy ta có hình dạng hàm hợp lí:

$$f(y | \mu) \propto e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

Phân phối hậu nghiệm là:

$$g(\mu | y) = \frac{g(\mu) \times f(y_1, \dots, y_n | \mu)}{\sum g(\mu) \times f(y_1, \dots, y_n | \mu)}$$

(1.8)

Tìm hàm hợp lí từ hàm mật độ chuẩn

Ví dụ 1.2 : Giả sử $y | \mu$ có phân bố chuẩn với giá trị trung bình là μ và độ lệch tiêu chuẩn $\sigma^2 = 1$. Và ta biết trước μ có năm giá trị là 2; 2.5; 3; 3.5; 4. Khả năng của các tiên nghiệm là 0.3; 0.3; 0.1; 0.1; 0.2. Quan sát y nhận được có giá trị $y = 3.3$. Sử dụng công thức hàm mật độ chuẩn

$$e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

Ta có bảng sau:

μ	Tiên nghiệm	Hàm hợp lí	tn× hhl	Hậu nghiệm
2.0	0.3	$e^{-\frac{1}{2}(3.3-2.0)^2} = .4295$.1288	.1848
2.5	0.3	$e^{-\frac{1}{2}(3.3-2.5)^2} = .7261$.2178	.3126
3.0	0.1	$e^{-\frac{1}{2}(3.3-3.0)^2} = .9559$.0956	.1372
3.5	0.1	$e^{-\frac{1}{2}(3.3-3.5)^2} = .9801$.0980	.1407
4.0	0.2	$e^{-\frac{1}{2}(3.3-4.0)^2} = .7827$.1565	.2247
			.6967	1.0000

Bảng 1.4

Tìm xác suất hậu nghiệm:

Ví dụ 1.3: giả sử ta có mẫu ngẫu nhiên của bốn quan sát từ phân bố chuẩn có giá trị trung bình μ và phương sai $\sigma^2 = 1$. Các giá trị quan sát được là: 3.3; 2.2; 3.6; 4.1, áp dụng tính toán ở bảng trên, lấy xác suất hậu nghiệm của lần quan sát trước làm xác suất tiên nghiệm cho lần quan sát tiếp theo ta có:

μ	Tiên nghiệm 2	Hàm hợp lí 2	tn \times hhl	Hậu nghiệm
2.0	.1848	$e^{-\frac{1}{2}(2.2-2.0)^2} = .9802$.1811	.2646
2.5	.3126	$e^{-\frac{1}{2}(2.2-2.5)^2} = .9560$.2988	.4366
3.0	.1372	$e^{-\frac{1}{2}(2.2-3.0)^2} = .7261$.0996	.1455
3.5	.1407	$e^{-\frac{1}{2}(2.2-3.5)^2} = .4296$.0605	.0884
4.0	.2247	$e^{-\frac{1}{2}(2.2-4.0)^2} = .1979$.0444	.0649
			.6844	1.0000
μ	Tiên nghiệm 3	Hàm hợp lí 3	tn \times hhl	Hậu nghiệm
2.0	.2646	$e^{-\frac{1}{2}(3.6-2.0)^2} = .2780$.0735	.1264
2.5	.4366	$e^{-\frac{1}{2}(3.6-2.5)^2} = .5461$.2384	.4102
3.0	.1455	$e^{-\frac{1}{2}(3.6-3.0)^2} = .8353$.1215	.2091
3.5	.0884	$e^{-\frac{1}{2}(3.6-3.5)^2} = .9950$.0879	.1512
4.0	.0649	$e^{-\frac{1}{2}(3.6-4.0)^2} = .9231$.0599	.1031
			.5812	1.0000
μ	Tiên nghiệm 4	Hàm hợp lí 4	tn \times hhl	Hậu nghiệm
2.0	.1264	$e^{-\frac{1}{2}(4.1-2.0)^2} = .1103$.0139	.0295
2.5	.4102	$e^{-\frac{1}{2}(4.1-2.5)^2} = .2780$.1140	.2421
3.0	.2091	$e^{-\frac{1}{2}(4.1-3.0)^2} = .5461$.1142	.2425
3.5	.1512	$e^{-\frac{1}{2}(4.1-3.5)^2} = .8352$.1262	.2681
4.0	.1031	$e^{-\frac{1}{2}(4.1-4)^2} = .9950$.1025	.2178
			.4708	1.0000

Bảng 1.5

Định lí Bayes cho phân phối chuẩn với tiên nghiệm đều

Quan sát y là biến ngẫu nhiên được rút ra từ phân phối chuẩn với giá trị trung bình μ và phương sai là σ^2 . Nếu ta sử dụng tiên nghiệm liên tục cho μ thì xác suất hậu nghiệm được tính bởi:

$$g(\mu | y_1, y_2, \dots, y_n) = \frac{g(\mu) \times f(y_1, y_2, \dots, y_n | \mu)}{\int g(\mu) \times f(y_1, y_2, \dots, y_n | \mu) d\mu} \quad (1.9)$$

Sử dụng tiên nghiệm đều ta có:

$$g(\mu) = 1 \text{ khi } \mu \geq 0$$

Khi y có n quan sát, cỡ mẫu trung bình là \bar{y} ta có: $\bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ và hàm hợp lí có dạng:

$$f(\bar{y} | \mu) \propto e^{-\frac{1}{2\sigma^2/n}(\bar{y}-\mu)^2}$$

Do đó phân phối hậu nghiệm cũng là phân phối chuẩn:

$$g(\mu | \bar{y}) \propto e^{-\frac{1}{2\sigma^2/n}(\mu-\bar{y})^2}$$

Định lí Bayes cho phân phối chuẩn với tiên nghiệm liên tục

Nếu phân phối tiên nghiệm là phân phối chuẩn với giá trị trung bình là m và phương sai là s^2 ta có:

$$g(\mu) \propto e^{-\frac{1}{2s^2}(\mu-m)^2}$$

Hàm hợp lí :

$$f(y | \mu) \propto e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

Từ đó ta có phân phối hậu nghiệm:

$$\begin{aligned} g(\mu | y) &\propto g(\mu) f(y | \mu) \propto e^{-\frac{1}{2} \left[\frac{(\mu-m)^2}{s^2} + \frac{(y-\mu)^2}{\sigma^2} \right]} \\ &\propto \exp \left\{ -\frac{1}{2} \left[\frac{\mu^2 (\sigma^2 + s^2) - 2\mu (\sigma^2 m + s^2 y) + \sigma^2 m^2 + s^2 y^2}{s^2 \sigma^2} \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2 \left(\frac{s^2 \sigma^2}{s^2 + \sigma^2} \right)} \left[\mu - \frac{\sigma^2 m + s^2 y}{s^2 + \sigma^2} \right] \right\} \end{aligned} \quad (1.10)$$

Phân phối hậu nghiệm là phân phối chuẩn với giá trị trung bình là m' và phương sai là $(s')^2$

Trong đó:

$$m' = \frac{\sigma^2 m + s^2 \bar{y}}{s^2 + \sigma^2} \quad (s')^2 = \frac{s^2 \sigma^2}{s^2 + \sigma^2}$$

(1.11)

Nếu sử dụng hàm hợp lí của cỡ mẫu trung bình \bar{y} thì phân phối có giá trị trung bình μ và phương sai $\frac{\sigma^2}{n}$, ta có hậu nghiệm:

$$\begin{aligned} \frac{1}{s'^2} &= \frac{1}{s^2} + \frac{n}{\sigma^2} \\ m' &= \frac{1/s^2}{n/\sigma^2 + 1/s^2} \times m + \frac{n/\sigma^2}{n/\sigma^2 + 1/s^2} \times \bar{y} \end{aligned}$$

(1.12)

Ví dụ 1.4: Ba nhà khoa học về động vật là W, I, N đang nghiên cứu ước lượng chiều dài của con chuột trong một năm tuổi. Các nghiên cứu trước đó đã chỉ ra rằng chiều dài của chuột tuân theo phân phối chuẩn với độ lệch chuẩn là $2cm$.

Nhà khoa học W nghĩ rằng giá trị trung bình tiên nghiệm $\mu = 30$. W nghĩ rằng chiều dài của con chuột trong một năm sẽ không nhỏ hơn $18cm$ và không lớn hơn $42cm$, do đó độ lệch tiêu chuẩn là $4cm$. Nhà khoa học này đã sử dụng tiên nghiệm chuẩn $(30, 4^2)$.

I là nhà khoa học không chuyên về những loài bò sát và gặm nhấm nên cậu ấy quyết định sử dụng tiên nghiệm đều(phẳng).

Còn nhà khoa học N sử dụng công thức nội suy tuyến tính giữa các giá trị và quyết định rằng tiên nghiệm của ông có dạng hình thang với trọng số 0 tại $18cm$, 1 tại $24cm$ và kéo dài cho đến $40cm$ sau đó lại xuống 0 tại $46cm$.

Họ lấy mẫu ngẫu nhiên $n = 12$ và tìm được trung bình mẫu là $\bar{y} = 32cm$. Hậu nghiệm của W có phân phối chuẩn $(\mu | \bar{y}) \sim N(m'; s'^2)$ do đó:

$$\begin{aligned} \frac{1}{s'^2} &= \frac{1}{s^2} + \frac{n}{\sigma^2} = \frac{1}{4^2} + \frac{12}{2^2} = 3.0625 \\ \Rightarrow s' &= 0.5714 \end{aligned}$$

Giá trị trung bình là:

$$\frac{1/4^2}{1/0.5714^2} \times 30 + \frac{12/2^2}{1/0.5714^2} \times 32 = 31.96$$

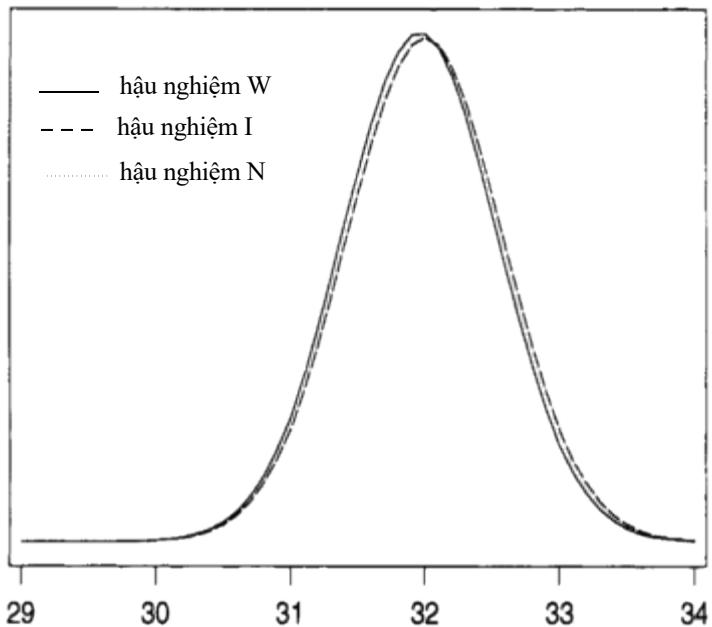
I sử dụng tiên nghiệm đều, vì vậy phương sai là:

$$s'^2 = \frac{\sigma^2}{n} = \frac{2^2}{12} = \frac{1}{3}$$

Vậy độ lệch chuẩn của I là $s' = 0.5744$, $m' = 32\text{cm}$. Vậy cả 2 nhà khoa học là W và I đều có hậu nghiệm chuẩn. Nhà khoa học C sử dụng công thức:

$$g(\mu | y_1, y_2, \dots, y_n) = \frac{g(\mu) \times f(y_1, y_2, \dots, y_n | \mu)}{\int g(\mu) \times f(y_1, y_2, \dots, y_n | \mu) d\mu}$$

Ta có hình dạng hậu nghiệm



Hình 1.1: Hình dạng hậu nghiệm của W, I, N

Ta có thể thấy hình dạng hậu nghiệm là tương tự nhau trong cả ba trường hợp.

1.2.2 Ước lượng điểm

Một ước lượng là một giá trị $\hat{\theta}$ được tính toán trên một mẫu được lấy một cách ngẫu nhiên, do đó giá trị của $\hat{\theta}$ là một biến ngẫu nhiên với kì vọng $E(\hat{\theta})$ và phương sai $V(\hat{\theta})$. Điều đó có nghĩa là giá trị $\hat{\theta}$ có thể giao động tùy theo mẫu thử, nó có ít cơ hội để có thể bằng đúng chính xác giá trị θ mà nó đang ước lượng. Mục đích ở đây là ta muốn kiểm soát sự sai lệch giá trị θ và $\hat{\theta}$.

Ước lượng điểm là cách thức tính toán một giá trị đơn lẻ của tham số tổng thể dựa trên dữ liệu mẫu.

Như ta đã biết một biến ngẫu nhiên luôn dao động xung quanh giá trị kì vọng của nó. Ta muốn kì vọng của $\hat{\theta}$ phải bằng θ . Khi đó ta nói ước lượng này là ước lượng không chêch.

Giả sử sai số ước lượng của $\hat{\theta}$ là $bias(\hat{\theta}) = E(\hat{\theta}) - \theta$. Ước lượng $\hat{\theta}$ là không chêch khi và chỉ khi:

$$E(\hat{\theta}) = \int \hat{\theta} f(\hat{\theta} | \theta) d\hat{\theta} = \theta$$

$f(\hat{\theta} | \theta)$ là phân phối mẫu của ước lượng.

Sai số trung bình của một ước lượng là:

$$MS(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \int (\hat{\theta} - \theta)^2 f(\hat{\theta} | \theta) d\hat{\theta} = Var(\hat{\theta}) + bias(\hat{\theta})^2 \quad (1.13)$$

Để ước lượng giá trị trung bình, ta thường sử dụng định lí giới hạn trung tâm: Một tổng thể có trung bình là μ và phương sai là σ^2 . Ta thu thập nhiều mẫu có cùng kích thước n , thì ta thu được nhiều giá trị \bar{y} . Khi số lượng mẫu thu thập đủ lớn thì các giá trị \bar{y} này có phân phối chuẩn với giá trị trung bình là μ và độ lệch chuẩn là $\frac{\sigma}{\sqrt{n}}$.

Vì vậy trung bình của mẫu \bar{y} có thể dùng để làm ước lượng không chêch cho trung bình của tổng thể μ .

Cho (y_1, y_2, \dots, y_n) là một mẫu ngẫu nhiên rút ra từ phân phối chuẩn $Y \sim N(\mu; \sigma^2)$.

Phân phối mẫu tương ứng của Y là $\bar{y} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$.

Theo tần suất:

Sử dụng \bar{y} để ước lượng không chêch cho μ . Khi đó $\bar{y} = \hat{\mu}_F$ là ước lượng có sai số $bias(\hat{\mu}) = 0$.

$$\begin{aligned} Var(\hat{\mu}_F) &= \frac{\sigma^2}{n} \\ MS(\hat{\mu}_F) &= Var(\hat{\theta}) + bias(\hat{\theta})^2 = Var(\hat{\mu}_F) = \frac{\sigma^2}{n} \end{aligned} \quad (1.14)$$

Theo Bayes:

Sử dụng kì vọng μ trong phân tích phân phối hậu nghiệm để ước lượng cho μ

$$\hat{\mu}_B = E(\mu | y_1, y_2, \dots, y_n) = \frac{1/s^2}{1/s^2 + n/\sigma^2} m + \frac{n/\sigma^2}{1/s^2 + n/\sigma^2} \bar{y}$$

(1.15)

Ta có:

$$\begin{aligned} E(\hat{\mu}_B) &= \frac{1/s^2}{1/s^2 + n/\sigma^2} m + \frac{n/\sigma^2}{1/s^2 + n/\sigma^2} \mu \\ Var(\hat{\mu}_B) &= \left[\frac{n/\sigma^2}{1/s^2 + n/\sigma^2} \right]^2 \frac{\sigma^2}{n} \\ &= \left(\frac{ns^2}{ns^2 + \sigma^2} \right)^2 \frac{\sigma^2}{n} \\ MS(\hat{\mu}_B) &= bias^2 + Var(\hat{\mu}_B) \\ &= \left(\frac{1/s^2}{1/s^2 + n/\sigma^2} m + \frac{n/\sigma^2}{1/s^2 + n/\sigma^2} \mu - \mu \right)^2 + \left(\frac{ns^2}{ns^2 + \sigma^2} \right)^2 \frac{\sigma^2}{n} \end{aligned} \quad (1.16)$$

Ví dụ 1.5:

Giả sử $\mu = 31, n = 12, s = 4, \sigma = 2, m = 30$ ta có sai số trung bình:

$$\begin{aligned} MS(\hat{\mu}_F) &= 0.3333 \\ MS(\hat{\mu}_B) &= 0.320 \end{aligned}$$

Ta thấy ước lượng điểm Bayes có sai số trung bình bình phương nhỏ hơn so với ước lượng tần suất. Vì vậy ước lượng Bayes là tốt hơn.

1.2.3 Ước lượng khoảng:

Ước lượng khoảng là cách sử dụng dữ liệu mẫu để tính toán (hoặc dự đoán) một khoảng giá trị có thể của một biến tổng thể chưa biết, trái ngược với ước lượng điểm, vì ước lượng điểm sẽ đưa ra một số duy nhất.

Theo tần suất:

$$\mu = \bar{y} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (1.17)$$

Trong đó:

\bar{y} là giá trị trung bình

$z_{\alpha/2}$ là hệ số tin cậy

α là mức tin cậy

σ là độ lệch chuẩn

n là kích thước mẫu

Khoảng tin cậy $(1-\alpha)100\%$ cho μ là:

$$\begin{aligned} P\left(\mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{y} < \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) &= 1 - \alpha \\ \Leftrightarrow P\left(\bar{y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) &= 1 - \alpha \end{aligned} \quad (1.18)$$

Theo Bayes:

Nếu phương sai đã biết: nếu sử dụng tiên nghiệm là phân phối đều hoặc phân phối chuẩn $N(m, s^2)$ thì phân phối hậu nghiệm của $N(m', s'^2)$. Một khoảng tin cậy Bayes $(1-\alpha)100\%$ cho μ là:

$$m' \pm z_{\alpha/2} s' \quad (1.19)$$

Nếu phương sai chưa biết thì ta cần tính phương sai mẫu $\hat{\sigma}^2$ từ dữ liệu. Ta mở rộng khoảng tin cậy bằng cách lấy giá trị bảng phân phối Student's thay cho phân phối chuẩn tắc. Khoảng tin cậy Bayes là:

$$m' \pm t_{\alpha/2} s' \quad (1.20)$$

1.2.4 Kiểm định giả thuyết

Giả thuyết thống kê: là một giả sử hay một phát biểu có thể đúng, có thể sai liên quan đến tham số của một hay nhiều tập hợp chính.

Giả thuyết không (giả thuyết đơn): là sự giả sử mà chúng ta muốn kiểm định, thường được kí hiệu là H_0 .

Giả thuyết ngược lại (đối thuyết): việc bác bỏ giả thuyết không sẽ dẫn đến việc chấp nhận giả thuyết ngược lại. Đối thuyết thường được kí hiệu là H_1 .

Miền bác bỏ và miền chấp nhận.

Tất cả các giá trị có thể có của đại lượng thống kê trong kiểm định có thể chia làm hai miền: miền bác bỏ và miền chấp nhận.

Miền bác bỏ là miền chứa các giá trị làm cho giả thuyết H_0 bị bác bỏ.

Miền chấp nhận là miền chứa các giá trị giúp cho giả thuyết H_0 không bị bác bỏ.

Trong thực tế khi H_0 không bị bác bỏ cùng nghĩa với việc nó được chấp nhận.

Giá trị chia đôi hai miền được gọi là giá trị giới hạn (critical value)

Kiểm định một phía và kiểm định hai phía

Kiểm định giả thuyết một phía tức là ta chỉ quan tâm đến việc xác định kết quả trong một hướng (H_1 có tính chất một phía). Ví dụ:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_0 \end{cases} \text{ hoặc } \begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}$$

Kiểm định hai phía là khi đối thuyết H_1 có tính chất hai phía.

Ví dụ:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

Kiểm định giả thuyết một phía theo phương pháp Bayes:

Đặt:

$$\begin{aligned} H_0 : \pi \leq \pi_0 \\ H_1 : \pi > \pi_0 \end{aligned}$$

Với mức ý nghĩa α sử dụng phương pháp Bayes ta có thể tính toán xác suất hậu nghiệm của “giả thuyết không”:

$$P(H_0 : \pi \leq \pi_0 | y) = \int_0^{\pi_0} g(\pi | y) d\pi$$

Ta bác bỏ H_0 nếu xác suất hậu nghiệm nhỏ hơn mức ý nghĩa α .

Kiểm định giả thuyết hai phía bằng phương pháp Bayes

Nếu chúng ta sử dụng tiên nghiệm liên tục thì sẽ được hậu nghiệm liên tục. Như đã biết xác suất tại một điểm chính xác của giả thuyết không sẽ bằng 0. Ta sẽ sử dụng khoảng tin cậy Bayes, tính toán $(1-\alpha)100\%$ cho π . Nếu π_0 nằm trong khoảng tin cậy Bayes thì ta chấp nhận giả thuyết không. Nếu không thì bác bỏ.

Khoảng tin cậy Bayes

Ta có tiên nghiệm $Beta(a, b)$ thì phân phối hậu nghiệm là (a', b') . Giá trị trung bình và phương sai của hậu nghiệm được cho bởi:

$$\begin{aligned} m' &= \frac{a'}{a' + b'} \\ (s')^2 &= \frac{a'b'}{(a' + b')^2 (a' + b' + 1)} \end{aligned} \tag{1.21}$$

Với độ tin cậy $(1-\alpha)100\%$ ta có khoảng tin cậy:

$$m' \pm z_{\alpha/2} \times s'$$

$z_{\alpha/2}$ là giá trị được lấy từ bảng phân phối chuẩn.

Ví dụ 1.6: Có hai loại bóng xanh và đỏ trong một hộp, người ta lấy ngẫu nhiên 15 quả bóng và nhận được 10 quả đỏ. Với π là xác suất nhận được quả đỏ, ta có thể khẳng định rằng xác suất nhận được quả đỏ là khác 0.5

Đặt giả thuyết không $H_0 : \pi = 0.5$

Đặt giả thuyết đối: $H_1 : \pi \neq 0.5$

Giả sử ta sử dụng tiên nghiệm $Beta(a, b) = (1, 1)$ cho π . Ta có $y = 10$ và $n = 15$. Với $a' = a + y; b' = n - y + 1$ nên mật độ hậu nghiệm là $Beta(a', b') = (11, 6)$

Áp dụng công thức khoảng tin cậy với độ tin cậy 95% ta có khoảng tin cậy cho π là:

$$\frac{11}{17} + 1.96 \times \sqrt{\frac{11 \times 6}{(11+6)^2} (11+6+1)} = (0.647 \pm 0.221) = (0.426, 0.848).$$

Giá trị giả thuyết không $\pi = 0.5$ nằm trong khoảng tin cậy, vì vậy ta không thể bác bỏ giả thuyết không.

1.2.5 Phương pháp Monte Carlo

Thông thường để tính tích phân hàm $f(x)$ trên miền D ta tiến hành theo các bước:

Bước 1: Chia miền lấy tích phân D thành n đoạn x_1, \dots, x_n

Bước 2: Tính giá trị $f(x)$ với độ dài đoạn tương ứng

Bước 3: Nhân $f(x_i)$ với độ dài đoạn tương ứng.

Bước 4: Tích phân được xấp xỉ bằng tổng các tích. Khi n tăng lên thì sai số của xấp xỉ được giảm đi.

Tuy nhiên không phải lúc nào ta cũng có thể tính dễ dàng như vậy. Ví dụ: với biến ngẫu nhiên liên tục thì tích phân $\int_{-\infty}^{+\infty} f(\theta) \times f(y|\theta) d(\theta)$ trong trường hợp nhiều chiều, nhiều biến số rất khó thực hiện. Vì vậy để giải quyết vấn đề này chúng ta đưa ra kĩ thuật để xấp xỉ những tích phân khó.

Kĩ thuật được sử dụng phổ biến nhất cho xấp xỉ tích phân trong thống kê chính là phương pháp Monte Carlo, nó dựa trên việc tính toán mô phỏng của biến ngẫu nhiên để tạo ra một xấp xỉ của tích phân hội tụ. Theo phương pháp Monte Carlo, để tính tích phân, thay vì chọn x_1, \dots, x_n là các điểm cố định, chúng ta lấy x_1, \dots, x_n ngẫu nhiên từ một phân phối $\pi(x)$ trên miền lấy tích phân D. Sau đó tính $f(x_i)$ cho mỗi x_i .

Trung bình các giá trị $f(x_i)$ cho ta xấp xỉ tích phân cần tính.

Cơ sở toán học của phương pháp Monte Carlo là luật số lớn.

Định nghĩa: Nếu x_1, \dots, x_n được phân phối từ π thì trung bình thực nghiệm

$\hat{J}_n = (f(x_1) + \dots + f(x_n))/n$ hội tụ hầu như chắc chắn đến tích phân:

$$J = \int f(x) \pi(x) dx \quad (1.22)$$

Ví dụ 1.7:

$$\text{Tính tích phân } I = \int_0^1 f(x) dx = \int_0^1 x^2 dx$$

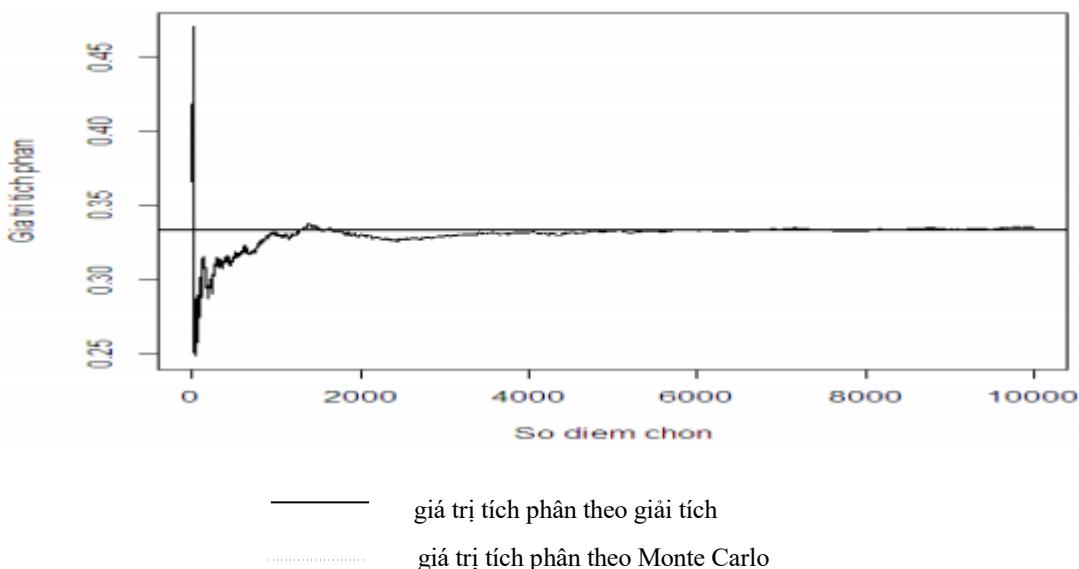
$$\text{Theo kết quả giải tích ta có: } I = \int_0^1 x^2 dx = \frac{1}{3}$$

Sử dụng mô phỏng Monte Carlo:

Miền lấy tích phân bị chặt $D = [0,1]$. Chọn $\pi(x)$ có phân phối đều trên D, do vậy $\pi(x) \sim U(0,1)$

Sinh ngẫu nhiên n giá trị x_i từ phân phối đều $U(0,1)$. Tính $f(x_i)$.

$$\text{Theo công thức luật số lớn ta có: } I \approx \frac{1}{n} \sum_{i=1}^n f(x_i)$$



Hình 1.2. Giá trị tích phân khi tính theo hai phương pháp

Từ biểu đồ ta có thể thấy: Khi n càng lớn hai đường biểu diễn càng gần nhau, xác suất tích phân càng chính xác. Nhưng tốc độ để tiến tới sự chính xác tương đối chậm, do đó phương pháp Monte Carlo tỏ ra kém hiệu quả khi giải quyết vấn đề một hoặc hai chiều. Tuy nhiên khi kích thước bài toán tăng lên, số tiến hành các phép thử lặp n không tăng theo kích cỡ của bài toán. Vì vậy phương pháp Monte Carlo được ưa chuộng sử dụng đối với bài toán nhiều chiều.

Nhược điểm của phương pháp Monte Carlo là không phải bài toán nào phương pháp Monte Carlo cũng có thể giải quyết được vì với những phân phối cơ bản thì việc

sinh mẫu là đơn giản nhưng đối với những phân phối không quen thuộc, một phân phối bất kì thì áp dụng Monte Carlo sẽ gặp khó khăn. Khi đó, yếu tố Markov được đưa vào Monte Carlo như một cải tiến quan trọng

1.2.6 Xích Markov

Thế nào là một quá trình ngẫu nhiên?

Xét một hệ nào đó tiến triển theo thời gian, quan sát tại các thời điểm rời rạc $0, 1, 2, \dots$. Giả sử các quan sát đó là $X_0, X_1, \dots, X_n, \dots$. Khi đó ta có một dãy các đại lượng ngẫu nhiên ứng với mỗi thời điểm n , giả sử $X^{(n)}$ là một biến ngẫu nhiên mô tả vị trí (tình trạng) của hệ. Quá trình $\{X^{(n)}\}_{n \geq 0}$ được gọi là một quá trình ngẫu nhiên. Tập hợp tất cả các trạng thái có thể có của hệ được gọi là không gian trạng thái, ký hiệu S .

Thế nào là một quá trình Markov?

Nếu không gian trạng thái S gồm một số hữu hạn hoặc vô hạn đếm được các trạng thái thì quá trình Markov X_n được gọi là xích Markov trạng thái rời rạc. Lúc này, có thể ký hiệu $S = 1, 2, 3, \dots$ tức là các trạng thái được đánh số. Nếu tập giá trị n không quá đếm được chẳng hạn $n = 0, 1, 2, 3, \dots$ thì ta có xích Markov thời gian rời rạc. Nếu $n \in [0, \infty)$ thì ta có xích Markov thời gian liên tục.

Giả sử trước thời điểm s , hệ đã ở trạng thái nào đó, còn tại thời điểm s , hệ ở trạng thái i . Chúng ta muốn đánh giá xác suất để tại thời điểm $t (t > s)$, hệ sẽ ở trạng thái j . Nếu xác suất này chỉ phụ thuộc vào bộ bốn (s, i, t, j) , nghĩa là

$p(X^{(s)} = i, X^{(t)} = j) = p(s, i, t, j)$ đúng với $\forall i, \forall j, \forall s, \forall t$ điều này thể hiện: xác suất có điều kiện của một sự kiện nào đó trong tương lai nếu biết hiện tại và quá khứ chỉ phụ thuộc vào trạng thái hiện tại và độc lập với quá khứ. Đây chính là tính Markov của hệ.

Định nghĩa: Ta nói rằng dãy các đại lượng ngẫu nhiên (X_n) là một xích Markov nếu với mọi $n_1 < \dots < n_k < n_{k+1}$ và với mọi $i_1, i_2, \dots, i_{k+1} \in S$

$$\begin{aligned} P\{X_{n_{k+1}} = i_{k+1} | X_{n_1} = i_1, X_{n_2} = i_2, \dots, X_{n_k} = i_k\} \\ = P\{X_{n_{k+1}} = i_{k+1} | X_{n_k} = i_k\} \end{aligned}$$

Quá trình Markov thuần nhất theo thời gian

Xét một chuỗi Markov. Nếu xác suất chuyển trạng thái:

$$p(s, i, t, j) = p(s+h, i, t+h, j) \quad \forall i, \forall j, \forall s, \forall t, h > 0$$

Ta nói chuỗi Markov thuần nhất theo thời gian.

Trong chương này và toàn bộ khóa luận chúng ta chỉ xét quá trình Markov thời gian rời rạc và thuần nhất. Vì vậy những phần sau nếu không nói gì thêm thì quá trình Markov được hiểu là thời gian rời rạc và thuần nhất.

Xét một chuỗi Markov rời rạc và thuần nhất theo thời gian $\{X^{(t)}\}$, $t = 0, 1, 2, \dots$ có không gian trạng thái S gồm N phần tử $S = \{1, 2, 3, \dots, N\}$.

Ma trận xác suất chuyển trạng thái

Kí hiệu:

$$p_{ij} = p(X^{(t+1)} = j | X^{(t)} = i) \quad \forall t \quad (1.23)$$

Là xác suất chuyển trạng thái từ vị trí i sang vị trí j sau một bước.

Ma trận xác suất chuyển trạng thái có được bằng cách liệt kê danh sách tất cả các trạng thái theo hàng và theo cột rồi điền vào đó xác suất chuyển trạng thái tương ứng.

Ma trận $P = [p_{ij}]_{N \times N}$ có kích thước $N \times N$ được gọi là ma trận xác suất chuyển trạng thái sau một bước.

Ví dụ 1.8:

Một người mẹ có 3 người con là A, B, C mỗi người con đều có một căn nhà riêng. Tại 1 trong 3 nhà của 3 người con người mẹ có thể đi sang 2 nhà khác hoặc đi lại chính ngôi nhà mà mình đang ở với đứa con hiện tại với xác suất được cho trong ma trận như sau:

$$P = \begin{pmatrix} 0.5 & 0.4 & 0.1 \\ 0.24 & 0.53 & 0.23 \\ 0.15 & 0.575 & 0.2275 \end{pmatrix}$$

Chỉ số hàng là vị trí hiện tại của người mẹ. Chỉ số cột là vị trí tiếp theo của người mẹ sẽ tới. Ví dụ, xét hàng thứ hai người mẹ đang ở vị trí người con B, khi đó người mẹ có thể qua người con A với xác suất 0.24, qua người con C với xác suất 0.23 hoặc ở lại nhà người con B với xác suất 0.53.

Ta gọi ma trận P là ma trận chuyển trạng thái sau một bước.

Ma trận xác suất chuyển trạng thái sau n bước.

P^n được gọi là ma trận chuyển trạng thái sau n bước.

Ví dụ 1.9:

Tiếp tục với ví dụ bên trên ta có:

Với $n = 2$ ma trận chuyển 2 bước là: $P^{(2)} = \begin{pmatrix} 0.362 & 0.469 & 0.169 \\ 0.282 & 0.509 & 0.209 \\ 0.254 & 0.523 & 0.223 \end{pmatrix}$

Với $n = 3$ ma trận chuyển 3 bước là: $P^{(3)} = \begin{pmatrix} 0.319 & 0.491 & 0.190 \\ 0.294 & 0.503 & 0.203 \\ 0.286 & 0.507 & 0.207 \end{pmatrix}$

Với $n = 4$ ma trận chuyển 4 bước là: $P^{(4)} = \begin{pmatrix} 0.306 & 0.497 & 0.197 \\ 0.298 & 0.501 & 0.201 \\ 0.295 & 0.502 & 0.203 \end{pmatrix}$

Với $n = 5$ ma trận chuyển 5 bước là: $P^{(5)} = \begin{pmatrix} 0.302 & 0.499 & 0.199 \\ 0.299 & 0.501 & 0.200 \\ 0.298 & 0.501 & 0.201 \end{pmatrix}$

Vector phân phối.

Giả sử tại thời điểm $t = m$, $X^{(m)}$ có thể nhận một trong N giá trị từ $1, 2, 3, \dots, N$ với các xác suất tương ứng là $\pi_1^{(m)}, \pi_2^{(m)}, \dots, \pi_N^{(m)}$ thỏa mãn điều kiện:

$$\pi_1^{(m)} + \pi_2^{(m)} + \dots + \pi_N^{(m)} = 1$$

Lúc đó vector $\pi^{(m)} = (\pi_1^{(m)}, \pi_2^{(m)}, \dots, \pi_N^{(m)})$ được gọi là vector phân phối tại thời điểm $t = m$

Với $t = 0$ ta có phân phối ban đầu $\pi^{(0)} = (\pi_1^{(0)}, \pi_2^{(0)}, \dots, \pi_N^{(0)})$.

Ta có P là ma trận xác suất chuyển trạng thái của chuỗi Markov và $\pi^{(0)}$ là vector biểu thị phân phối ban đầu. Khi đó ta có:

Xác suất để chuỗi ở trạng thái i sau n bước là phần thứ i trong vector:

$$\pi^{(n)} = \pi^{(0)} P^n \quad (1.24)$$

Hay

$$\pi^{(m+n)} = \pi^{(m)} P^n \quad (1.25)$$

Ví dụ 1.10:

Trở lại với ví dụ trên:

Tại thời điểm $t = 0$ xác suất để người mẹ tới nhà của ba người con lần lượt là : $(0.6, 0.3, 0.1)$. Khi đó ta có vector phân phối ban đầu là :

$$\pi^{(0)} = (0.6, 0.3, 0.1)$$

Tại thời điểm $t = 1$ ta có:

$$\pi^{(1)} = \pi^{(0)} P = (0.6, 0.3, 0.1) \begin{pmatrix} 0.5 & 0.4 & 0.1 \\ 0.24 & 0.53 & 0.23 \\ 0.15 & 0.575 & 0.275 \end{pmatrix} = (0.387, 0.456, 0.157)$$

Tại thời điểm $t = 2$ ta có:

$$\pi^{(2)} = \pi^{(0)} P^2 = (0.6, 0.3, 0.1) \begin{pmatrix} 0.362 & 0.469 & 0.169 \\ 0.282 & 0.509 & 0.209 \\ 0.254 & 0.523 & 0.223 \end{pmatrix} = (0.326, 0.486, 0.187)$$

Tại thời điểm $t = 3$ ta có:

$$\pi^{(3)} = \pi^{(0)} P^3 = (0.6, 0.3, 0.1) \begin{pmatrix} 0.319 & 0.491 & 0.190 \\ 0.294 & 0.503 & 0.203 \\ 0.286 & 0.507 & 0.207 \end{pmatrix} = (0.308, 0.495, 0.197)$$

Tương tự tại thời điểm $t = 4$ ta có:

$$\pi^{(4)} = (0.302, 0.498, 0.199)$$

Tại thời điểm $t = 5$ ta có:

$$\pi^{(5)} = (0.301, 0.499, 0.200)$$

Các tính chất cơ bản của quá trình Makov

Phương trình C-K (Chapman-Kolmogorov):

$$p_{ij}^{(n+m)} = \sum_{k \in S} p_{ik}^{(n)} p_{kj}^{(m)} \quad (1.26)$$

$$\pi^{(m+n)} = \pi^{(m)} P^n \quad (1.27)$$

$$P^{(n+m)} = P^{(n)} P^{(m)}$$

$$P^{(m)} = P^m \quad (1.28)$$

$$P^{(2)} = P \times P = P^2$$

Các phân phối cơ bản và cần thiết của quá trình Markov

Phân phối dừng:

Cho (X_0, X_1, \dots) là một xích Markov rời rạc và thuận nhất với không gian trạng thái S . Với ma trận xác suất chuyển trạng thái P , lúc đó vector phân phối $\pi = (\pi_1, \pi_2, \dots, \pi_N)$ được gọi là phân phối dừng của xích Markov nếu nó thỏa mãn:

$$(1) \quad \pi_i \geq 0 \text{ với } i = 1, \dots, N \text{ và } \sum_{i=1}^N \pi_i = 1$$

$$(2) \quad \pi = \pi P$$

Phân phối dừng π không phụ thuộc vào $\pi^{(0)}$ mà chỉ phụ thuộc vào ma trận P

Phân phối giới hạn:

Giả sử (X_n) là xích Markov với không gian trạng thái S , với ma trận xác suất chuyển P và ma trận xác suất chuyển sau n bước là $P^{(n)}$. Phân phối $\pi = (\pi_1, \pi_2, \dots, \pi_N)$ thỏa mãn điều kiện:

$$(1) \quad \pi_1 + \pi_2 + \dots + \pi_N = 1$$

$$(2) \quad \lim_{n \rightarrow \infty} p_{ij}^{(n)} = \pi_j \text{ không phụ thuộc vào } i$$

Phân phối π được gọi là phân phối giới hạn.

Nếu phân phối giới hạn tồn tại thì phân phối dừng cũng tồn tại và duy nhất. Hơn nữa hai phân phối này trùng nhau. Tuy nhiên điều ngược lại không đúng tức là có những xích Markov có tồn tại phân phối dừng nhưng không tồn tại phân phối giới hạn.

Phân phối Ergodic

Phân phối $\pi = (\pi_1, \pi_2, \dots, \pi_N)$ thỏa mãn điều kiện:

$$(1) \quad \pi_1 + \pi_2 + \dots + \pi_N = 1$$

$$(2) \quad \lim_{n \rightarrow \infty} p_{ij}^{(n)} = \pi_j$$

$$(3) \quad \pi_j > 0, \forall j$$

Được gọi là phân phối Ergodic

Từ đó ta có:

Định lý 1: giả sử X_n là xích Markov Ergodic thỏa mãn điều kiện tối giản không có chu kỳ với không gian trạng thái hữu hạn $S = \{1, 2, \dots, N\}$. Khi đó mọi trạng thái đều hồi quy dương và xích có phân phối giới hạn $\pi = (\pi_1, \pi_2, \dots, \pi_N)$. Phân phối này cũng là phân phối dừng duy nhất của xích.

Định lý 2: (Định lý Ergodic) Đối với bất kì xích Markov Ergodic với phân phối dừng $\pi = (\pi_1, \pi_2, \dots, \pi_N)$ thì:

$$\frac{1}{n} \sum_{i=1}^n f(X_i) \rightarrow E_\pi [f(X)] \text{ khi } n \rightarrow \infty \quad (1.29)$$

Như vậy khi ta sinh một xích Markov từ phân phối mục tiêu $\pi(x)$ cho trước thì khi chạy đủ lâu, xích sẽ hội tụ về một phân phối xác suất cố định, độc lập với trạng thái ban đầu. Phân phối xác suất này chính là phân phối dừng và giá trị trung bình $f(X)$ sẽ hội tụ tới kì vọng của nó.

Phương pháp xích Markov Monte Carlo

Như chúng ta đã biết hầu hết các mô hình ta gặp phải trong Bayes khá là phức tạp, các phương pháp mô phỏng tiêu chuẩn không phải là một giải pháp đầy đủ và linh hoạt để giải quyết bài toán. Nên ta sẽ trình bày các nguyên lý cơ bản của một kĩ thuật xuất hiện vào cuối những năm 1980 như là bản chất của tính toán Bayes. Để giải quyết bài toán sinh mẫu từ một phân phối thì phương pháp MCMC hoạt động như sau:

Vai trò của phương pháp Monte Carlo tạo ra một chuỗi Markov, còn gọi là mô phỏng.

Chuỗi Markov thỏa mãn điều kiện Ergodic có phân phối dừng là phân phối cần sinh mẫu.

Chương 2:

Mô hình hồi quy tuyến tính

2.1. Mô hình hồi quy tuyến tính

Ta có một biến hay một tập hợp các biến độc lập $X = (x_1, \dots, x_p)$ là các biến giải thích. Các biến phụ thuộc Y .

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (2.1)$$

Và các tham số $\beta = (\beta_1, \dots, \beta_p)^T$ được ước lượng từ dữ liệu. Các dữ liệu được tạo thành từ các vector $Y = (y_1, y_2, \dots, y_n)$ và ma trận cấp $n \times (p+1)$ của X

$$X = [1_n, x_1, \dots, x_p] = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}$$

Điều kiện: $p+1 < n$ ta có:

$$\begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Trong nhiều mô hình và cấu trúc phụ thuộc ta tập trung vào các mô hình hồi quy tuyến tính chuẩn (Gauss), cụ thể $E(y|x, \theta)$ là tuyến tính trong x và nhiều là chuẩn.

Mô hình hồi quy tuyến tính chuẩn có dạng:

$$\{y | \beta, \sigma^2, X\} \sim N_n(X\beta, \sigma^2 I_n) \quad (2.2)$$

Với I là ma trận đơn vị.

Do đó:

$$\begin{aligned} E(y_i | \beta, X) &= \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} \\ Var(y_i | \sigma^2, X) &= \sigma^2 \end{aligned} \quad (2.3)$$

Mô hình hồi quy tuyến tính chuẩn xác định rằng độ biến thiên của sai số là:

$$\varepsilon_1, \dots, \varepsilon_n \sim i.i.d. normal(0, \sigma^2) \quad (2.4)$$

Một cách khác để viết mô hình này cung cấp mật độ xác suất chung của dữ liệu quan sát y_1, \dots, y_n dựa trên điều kiện x_1, \dots, x_n và giá trị β, σ^2 :

$$\begin{aligned} & p(y_1, \dots, y_n | x_1, \dots, x_n, \beta, \sigma^2) \\ &= \prod_{i=1}^n p(y_i | x_i, \beta, \sigma^2) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta^T x_i)^2\right\} \end{aligned} \quad (2.5)$$

Từ công thức trên ta có thể thấy biểu thức trong số mũ đạt giá trị lớn nhất khi tổng số dư bình phương $SSR(\beta) = \sum_{i=1}^n (y_i - \beta^T x_i)^2$ đạt giá trị nhỏ nhất.

$$\begin{aligned} SSR(\beta) &= \sum_{i=1}^n (y_i - \beta^T x_i)^2 = (y - X\beta)^T (y - X\beta) \\ &= Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta \end{aligned}$$

Ta có $SSR(\beta)$ đạt giá trị nhỏ nhất khi $\frac{d}{d\beta} SSR(\beta) = 0$

Do đó:

$$\begin{aligned} \frac{d}{d\beta} SSR(\beta) &= \frac{d}{d\beta} (Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta) \\ &= -2X^T Y + 2X^T X\beta \\ \frac{d}{d\beta} SSR(\beta) = 0 &\Leftrightarrow -2X^T Y + 2X^T X\beta = 0 \\ &\Leftrightarrow X^T X\beta = X^T Y \\ &\Leftrightarrow \beta = (X^T X)^{-1} X^T Y \end{aligned} \quad (2.6)$$

Ước lượng không chêch của σ^2 có thể được tính bằng công thức

$$\hat{\sigma}_{ols}^2 = SSR(\hat{\beta}_{ols}) / (n - p) \quad (2.7)$$

Phương sai mẫu của vector $\hat{\beta}_{ols}$ có thể được tính bằng $(X^T X)^{-1} \sigma^2$. Chúng ta chưa biết giá trị thực của σ^2 nhưng ta có thể thay thế bằng $\hat{\sigma}_{ols}^2$.

2.1.1 Ước lượng bình phương tối thiểu

Ví dụ 2.1: Mười hai người đàn ông khỏe mạnh không tập thể dục thường xuyên đã được tuyển dụng để tham gia vào một nghiên cứu về tác động của hai chế độ tập thể dục khác nhau với sự hấp thụ oxy. Sáu trong số mười hai người đàn ông được phân công vào việc chạy trong 12 tuần mức độ hấp thụ oxy tối đa của từng đối tượng

được đo (tính bằng lít mỗi phút) trong khi chạy trên máy chạy bộ cả trước và sau chương trình. Sự thay đổi trong việc hấp thụ oxy tối đa có thể phụ thuộc vào chương trình mà họ được chỉ định, tuổi tác, giới tính. Vậy làm thế nào chúng ta có thể ước tính sự phân phối có điều kiện của sự hấp thụ oxy cho một chương trình tập thể dục và độ tuổi nhất định với dữ liệu đã cho dưới đây:

$$x_3 = (23, 22, 22, 25, 27, 20, 31, 23, 27, 28, 22, 24) \text{ (tuổi của người tham gia)}$$

$$Y = (-0.87, -10.74, -3.27, -1.97, 7.5, -7.25, 17.05, 4.96, 10.4, 11.05, 0.26, 2.51)$$

$$\text{(mức độ hấp thụ oxy)}$$

Theo bài ra ta có:

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \varepsilon_i \text{ với}$$

$$x_{i,1} = 1 \text{ với mỗi giá trị } i$$

$$x_{i,2} = 0 \text{ nếu người thứ } i \text{ tập chạy, } 1 \text{ nếu tập aerobic}$$

$$x_{i,3} = \text{tuổi người thứ } i$$

$$x_{i,4} = x_{i,2} \times x_{i,3}$$

$$X^T X = \begin{pmatrix} 12 & 6 & 294 & 155 \\ 6 & 6 & 155 & 155 \\ 294 & 155 & 7314 & 4063 \\ 155 & 155 & 4063 & 4063 \end{pmatrix} \quad X^T Y = \begin{pmatrix} 29.63 \\ 46.23 \\ 978.81 \\ 1298.79 \end{pmatrix}$$

$$\text{Vậy } \hat{\beta}_{ols} = (-51.29, 13.11, 2.09, -0.32)^T$$

Hạn chế của mô hình hồi quy tuyến tính theo phương pháp bình phương tối thiểu:

Nó rất nhạy cảm với nhiễu. Chỉ cần có một cặp dữ liệu nhiễu là kết quả sẽ khác đi rất nhiều.

Không biểu diễn được các mô hình phức tạp.

2.2 Ước lượng Bayes cho mô hình hồi quy

Ta có:

$$p(Y | X, \beta, \sigma^2) \propto \exp \left\{ -\frac{1}{2\sigma^2} SSR(\beta) \right\}$$

$$= \exp \left\{ -\frac{1}{2\sigma^2} [Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta] \right\}.$$

Ta thấy vai trò của Y và β gần tương đương nhau và phân phối của Y là phân phối chuẩn đa biến. Điều này cho thấy rằng phân phối chuẩn đa biến tiên nghiệm của β là liên hợp. Nếu điều này đúng: nếu $\beta \sim \text{multivariate normal}(\beta_0, \Sigma_0)$ khi đó ta có:

$$\begin{aligned}
& p(\beta | Y, X, \sigma^2) \\
& \propto p(Y | X, \beta, \sigma^2) \times p(\beta) \\
& \propto \exp \left\{ -\frac{1}{2} \left(-2\beta^T X^T Y / \sigma^2 + \beta^T X^T X \beta / \sigma^2 \right) - \frac{1}{2} \left(-2\beta^T \sum_0^{-1} \beta_0 + \beta^T \sum_0^{-1} \beta \right) \right\} \\
& = \exp \left\{ \left(\sum_0^{-1} \beta_0 + X^T Y / \sigma^2 \right) - \frac{1}{2} \beta^T \left(\sum_0^{-1} + X^T X / \sigma^2 \right) \beta \right\}
\end{aligned} \tag{2.8}$$

Từ đó ta có:

$$\begin{aligned}
\text{Var}[\beta | Y, X, \sigma^2] &= \left(\sum_0^{-1} + X^T X / \sigma^2 \right)^{-1} \\
\text{E}[\beta | Y, X, \sigma^2] &= \left(\sum_0^{-1} + X^T X / \sigma^2 \right)^{-1} \left(\sum_0^{-1} \beta_0 + X^T Y / \sigma^2 \right)
\end{aligned} \tag{2.9}$$

Như trong hầu hết các vấn đề lấy mẫu chuẩn, phân phối tiên nghiệm bán liên hợp cho σ^2 là một phân phối inverse-gamma. Đặt $\gamma = \frac{1}{\sigma^2}$, nếu

$$\gamma \sim \text{gamma} \left(\frac{v_0}{2}, \frac{v_0 \sigma_0^2}{2} \right) \text{ thì}$$

$$\begin{aligned}
p(\gamma | Y, X, \beta) &\propto p(\gamma) p(Y | X, \beta, \gamma) \\
&\propto \left[\gamma^{v_0/2-1} \exp(-\gamma \times v_0 \sigma_0^2 / 2) \right] \times \left[\gamma^{n/2} \exp(-\gamma \times \text{SSR}(\beta) / 2) \right] \\
&= \gamma^{(v_0+n)/2-1} \exp(-\gamma [v_0 \sigma_0^2 + \text{SSR}(\beta)] / 2)
\end{aligned}$$

Ta thấy đó là mật độ gamma, do vậy:

$$\{\sigma^2 | Y, X, \beta\} \sim \text{inverse-gamma} \left([v_0 + n] / 2, [v_0 \sigma_0^2 + \text{SSR}(\beta)] / 2 \right). \tag{2.10}$$

Dùng bộ lấy mẫu Gibbs để xấp xỉ phân phối $p(\beta, \sigma^2 | Y, X)$. Cho giá trị đầu vào là $\{\beta^{(s)}, \sigma^{2(s)}\}$, giá trị mới có thể được cập nhật:

1. Cập nhật β :

- a) Tính $V = \text{var}[\beta | Y, X, \sigma^{2(s)}]$ và $m = E[\beta | Y, X, \sigma^{2(s)}]$
- b) Cập nhật: $\beta^{(s+1)} \sim \text{multivariate normal}(m, V)$

2. Cập nhật σ^2 :

a) Tính $SSR(\beta^{(s+1)})$

b) Cập nhật $\sigma^{2(s+1)} \sim inverse-gamma([v_0 + n]/2, [v_0\sigma_0^2 + SSR(\beta^{(s+1)})]/2)$

Một nguyên tắc khác để xây dựng phân phối tiên nghiệm cho B dựa trên ý tưởng rằng ước lượng tham số nên bất biến đối với những thay đổi trong thang đo của biến hồi quy. Ví dụ: giả sử ai đó phân tích dữ liệu hấp thụ oxy bằng cách sử dụng $\tilde{x}_{i,3} =$ tuổi theo tháng thay vì $x_{i,3} =$ tuổi tính theo năm. Phân phối hậu nghiệm cho $12 \times \tilde{\beta}_3$ trong mô hình có $\tilde{x}_{i,3}$ phải giống với phân phối hậu nghiệm cho β_3 dựa trên mô hình có $\tilde{x}_{i,3}$. Điều này đòi hỏi sự thay đổi dự kiến của Y trong một năm thay đổi là như nhau cho dù tuổi được ghi theo tháng hay năm. Tóm lại, giả sử X là một tập các biến hồi quy đã cho và $\tilde{X} = XH$ (H là ma trận $p \times p$). Nếu chúng ta thu được phân phối hậu nghiệm của β từ Y và X , và phân phối hậu nghiệm của $\tilde{\beta}$ từ Y và \tilde{X} thì theo nguyên tắc bất biến này các phân phối hậu nghiệm của β và $H\tilde{\beta}$ phải giống nhau. Điều kiện này sẽ được đáp ứng nếu:

$$\beta_0 = 0 \text{ và } \Sigma_0 = k(X^T X)^{-1} \text{ với mọi giá trị dương } k.$$

Một đặc điểm kỹ thuật phổ biến của k là liên kết nó với một phuong sai σ^2 , vì vậy ta đặt: $k = g\sigma^2$ với mỗi giá trị dương g . Những lựa chọn tham số ở trên được gọi là một phiên bản “g-prior”, một nghiên cứu được sử dụng rộng rãi cho các tham số hồi quy của phân phối tiên nghiệm (bản gốc g-prior của Zellner cho β_0 khác 0). Ta có phân phối có điều kiện của β cho với điều kiện (Y, X, σ^2) vẫn là multivariate normal (chuẩn đa biến) do đó công thức (4) và (5) được viết gọn như sau:

$$\begin{aligned} \text{var}[\beta | Y, X, \sigma^2] &= \left[X^T X / (g\sigma^2) + X^T X / \sigma^2 \right]^{-1} \\ &= \frac{g}{g+1} \sigma^2 (X^T X)^{-1} \end{aligned} \tag{2.11}$$

$$\begin{aligned} E[\beta | Y, X, \sigma^2] &= \left[X^T X / (g\sigma^2) + X^T X / \sigma^2 \right]^{-1} X^T Y / \sigma^2 \\ &= \frac{g}{g+1} (X^T X)^{-1} X^T Y \end{aligned} \tag{2.12}$$

Ước tính tham số theo g-prior cũng được đơn giản hóa. Bởi vì nếu ước tính tham số theo g-prior, theo phân phối tiên nghiệm ta có $p(\sigma^2 | Y, X)$ là một phân phối

inverse-gamma, điều đó có nghĩa là chúng ta có thể lấy mẫu trực tiếp (σ^2, β) từ phân phối hậu nghiệm của chúng bằng cách lấy mẫu đầu tiên từ $p(\sigma^2 | Y, X)$ và tiếp theo là từ $p(\beta | \sigma^2, Y, X)$

Chứng minh rằng việc ước lượng tham số σ^2 không thông qua β .

Phân bố hậu nghiệm của σ^2 tỉ lệ thuận với $p(\sigma^2) \times p(Y | X, \sigma^2)$.

Sử dụng quy tắc xác suất biên ta có:

$$p(Y | X, \sigma^2) = \int p(Y | X, \beta, \sigma^2) p(\beta | X, \sigma^2) d\beta$$

Viết theo phân bố của hai mật độ trong tích phân ta có:

$$\begin{aligned} p(Y | X, \sigma^2) &= (2\pi)^{-n/2} (1+g)^{-p/2} (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} SSR_g\right) \\ p(Y | X, \beta, \sigma^2) p(\beta | X, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} (Y - X\beta)^T (Y - X\beta)\right] \\ &\quad \times \left|2\pi g\sigma^2 (X^T X)^{-1}\right|^{-1} \exp\left[-\frac{1}{2g\sigma^2} \beta^T X^T X \beta\right]. \end{aligned}$$

Kết hợp tất cả các phép tính trên số mũ lại ta được:

$$\begin{aligned} &-\frac{1}{2\sigma^2} \left[(Y - X\beta)^T (Y - X\beta) + \beta^T X^T X \beta / g \right] \\ &= -\frac{1}{2\sigma^2} \left[Y^T Y - 2Y^T X\beta + \beta^T X^T X \beta (1+1/g) \right] \\ &= -\frac{1}{2\sigma^2} Y^T Y - \frac{1}{2} (\beta - m)^T V^{-1} (\beta - m) + \frac{1}{2} m^T V^{-1} m \end{aligned}$$

Với

$$\begin{aligned} V &= \frac{g}{g+1} \sigma^2 (X^T X)^{-1} \\ m &= \frac{g}{g+1} (X^T X)^{-1} X^T Y \end{aligned}$$

Chúng ta có thể viết lại $p(Y | X, \beta, \sigma^2) p(\beta | X, \sigma^2)$ như sau:

$$\begin{aligned} p(Y | X, \beta, \sigma^2) p(\beta | X, \sigma^2) &= \left[(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} Y^T Y\right) \right] \times \left[(1+g)^{-p/2} \exp\left(\frac{1}{2} m^T V^{-1} m\right) \right] \\ &\quad \times \left[|2\pi V|^{-1/2} \exp\left[-\frac{1}{2} (\beta - m)^T V^{-1} (\beta - m)\right] \right] \end{aligned}$$

Đặt $A = |2\pi V|^{-1/2} \exp\left[-\frac{1}{2}(\beta - m)^T V^{-1}(\beta - m)\right]$, ta thấy A là biểu thức duy nhất ở đẳng thức trên phụ thuộc vào β . Và ta có thể nhận ra

$|2\pi V|^{-1/2} \exp\left[-\frac{1}{2}(\beta - m)^T V^{-1}(\beta - m)\right]$ chính xác là hàm mật độ chuẩn đa biến với giá trị kỳ vọng là m và phương sai V. và kết hợp với tích phân theo β ta có:

$$\int |2\pi V|^{-1/2} \exp\left[-\frac{1}{2}(\beta - m)^T V^{-1}(\beta - m)\right] d\beta = 1$$

Do vậy:

$$\begin{aligned} p(Y|X, \sigma^2) &= \int p(Y|X, \beta, \sigma^2) p(\beta|X, \sigma^2) d\beta \\ &= \left[(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} Y^T Y\right) \right] \times \left[(1+g)^{-p/2} \exp\left(\frac{1}{2} m^T V^{-1} m\right) \right] \end{aligned} \quad (2.13)$$

$$\text{Với } SSR_g = Y^T Y - m^T V^{-1} m = Y^T \left(I - \frac{g}{g+1} X (X^T X)^{-1} X^T \right) Y$$

Thuật ngữ SSR_g sẽ giảm xuống $SSR_{ols} = \sum (y_i - \hat{\beta}_{ols} x_i)$ khi $g \rightarrow \infty$. g giúp chúng ta thu nhỏ độ lớn của các hệ số hồi quy và có thể ngăn chặn tình trạng overfitting của dữ liệu.

Bước cuối cùng trong việc xác định $p(\sigma^2 | Y, X)$ là nhân $p(Y|X, \sigma^2)$ với phân bố tiên nghiệm. Đặt $\gamma = 1/\sigma^2 \sim \text{gamma}(v_0/2, v_0\sigma_0^2/2)$ ta có:

$$\begin{aligned} p(\gamma|Y, X) &\propto p(\gamma) p(Y|X, \gamma) \\ &\propto \left[\gamma^{v_0/2-1} \exp(-\gamma \times v_0\sigma_0^2/2) \right] \times \left[\gamma^{n/2} \exp(-\gamma \times SSR_g/2) \right] \\ &= \gamma^{(v_0+n)/2-1} \exp\left[-\gamma \times (v_0\sigma_0^2 + SSR_g)/2\right] \\ &\propto dgamma\left(\gamma, [v_0+n]/2, [v_0\sigma_0^2 + SSR_g]/2\right) \end{aligned}$$

Và $\{\sigma^2 | Y, X\} \sim \text{inverse-gamma}\left([v_0+n]/2, [v_0\sigma_0^2 + SSR_g]/2\right)$.

$p(\sigma^2 | Y, X)$ và $p(\beta | Y, X, \sigma^2)$ là phân phối inverse-gamma và multivariate normal tương ứng. Vì ta có thể lấy mẫu từ cả hai phân phối này nên các mẫu từ phân phối hậu nghiệm $p(\sigma^2, \beta | Y, X)$ có thể được thực hiện với xấp xỉ Monte Carlo. Và lấy giá trị mẫu Gibbs là không cần thiết. Giá trị mẫu của (σ^2, β) từ $p(\sigma^2, \beta | Y, X)$ có thể được lấy bằng cách:

1. Mẫu $1/\sigma^2 \sim \text{gamma}([v_0 + n]/2, [v_0\sigma_o^2 + SSR_g]/2)$;
2. Mẫu $\beta \sim \text{multivariate normal} \left(\frac{g}{g+1}\hat{\beta}_{ols}, \frac{g}{g+1}\sigma^2 [X^T X]^{-1} \right)$.

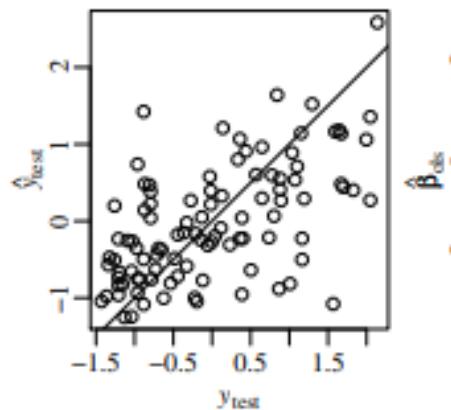
2.3 Chọn mô hình (model selection)

Thông thường trong phân tích hồi quy, chúng ta phải đối mặt với một số lượng lớn các biến hồi quy không có mối quan hệ thực sự với biến Y. Trong các tình huống này, chúng ta chỉ nên đưa vào mô hình hồi quy của mình những biến có mối liên hệ thực sự với Y. Làm như vậy không chỉ tạo ra các phân tích dữ liệu đơn giản hơn, thẩm mĩ hơn mà còn cung cấp các mô hình với các thuộc tính thống kê tốt hơn về mặt dự đoán và ước tính.

Ví dụ 2.2:

Ta có dữ liệu cơ sở cho mười biến x_1, \dots, x_{10} trong một nhóm gồm 442 bệnh nhân tiểu đường đã được thu thập, cũng như Y là sự tiến triển của bệnh. Từ những dữ liệu này ta hi vọng rằng sẽ tạo ra một mô hình dự đoán cho Y dựa trên các phép đo cơ sở. Mặc dù mô hình hồi quy có mười biến sẽ không quá phức tạp nhưng người ta nghĩ ngờ rằng mối quan hệ giữa Y và X có thể không tuyến tính và bao gồm các biến bậc hai như x_j^2 hoặc là $x_j x_k$ trong mô hình hồi quy hỗ trợ việc dự đoán. Do đó các biến hồi quy bao gồm mười biến chính là $x_1, x_2, \dots, x_9, x_{10}$, 45 biến có dạng $x_j x_k$ và 9 biến bậc hai có dạng x_j^2 , (một trong những biến $x_2 = \text{sex}$ có giá trị nhị phân nên vậy nên $x_2 = x_2^2$, do đó không cần thiết phải bao gồm x_2^2). Tổng cộng ta có 64 biến hồi quy.

Trong phần này ta sẽ xây dựng các mô hình hồi quy dự báo cho Y dựa trên 64 biến hồi quy. Để đánh giá các mô hình, ta sẽ phân chia ngẫu nhiên 442 bệnh nhân tiểu đường vào 342 mẫu huấn luyện và kiểm tra 100 mẫu, khi cung cấp 1 dữ liệu huấn luyện (Y, X) và một bộ dữ liệu test (Y_{test}, X_{test}) . Ta sẽ điều chỉnh mô hình hồi quy bằng cách sử dụng dữ liệu huấn luyện và sau đó sử dụng các hệ số hồi quy ước tính để tạo $\hat{Y}_{test} = X_{test} \hat{\beta}$. Hiệu suất của mô hình dự đoán sau đó có thể được đánh giá bằng cách so sánh Y_{test} và \hat{Y}_{test} . Chúng ta sẽ bắt đầu bằng cách xây dựng một mô hình phân bố chuẩn với sai số bình phương nhỏ nhất cho 64 biến.

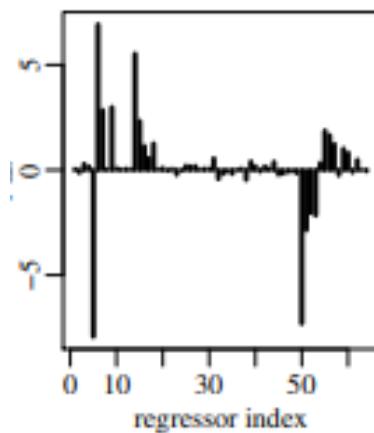


Hình 2.1

Qua biểu đồ bên trên ta có thể thấy được giá trị của 100 mẫu thử \hat{Y}_{test} so với giá trị dự đoán ban đầu của chúng. $\hat{Y}_{test} = X_{test}\hat{\beta}$, với $\hat{\beta}$ đã được ước tính bằng cách sử dụng 342 mẫu đào tạo.

Mặc dù rõ ràng có một mối quan hệ giữa giá trị thực và dự đoán, nhưng vẫn có khá nhiều lỗi. Sai số dự đoán trung bình bình phương là:

$$\frac{1}{100} \sum (y_{test,i} - \hat{y}_{test,i})^2 = 0.67$$



Hình 2.2

Biểu đồ tiếp theo cho thấy các giá trị ước tính của 64 tham số. Như trên hình biểu diễn, điều đáng lưu ý ta có thể thấy đó là phần lớn các giá trị được ước tính khá nhỏ.

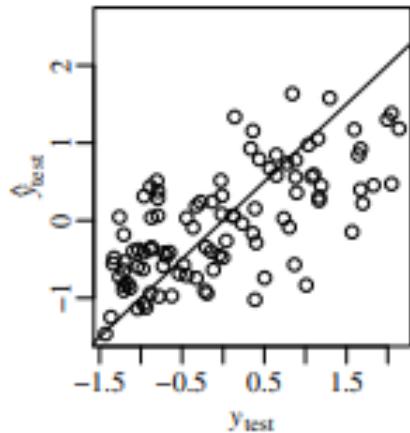
Có lẽ dự đoán có thể được cải thiện bằng cách loại bỏ khỏi mô hình hồi quy những biến cho thấy giá trị rất gần với 0. Bằng cách đó, ta hi vọng loại bỏ khỏi mô hình dự đoán bất kỳ biến hồi quy nào có liên kết giả với Y chỉ để lại những biến hồi quy có liên kết với bất kỳ nhóm đối tượng nào. Có một cách tiêu chuẩn cho thấy giá trị thực của hệ số hồi quy $\hat{\beta}_j$ không bằng 0 với t-statistic, đó là cách chia ước lượng $\hat{\beta}_j$ cho sai số chuẩn của nó.

$$t_j = \frac{\hat{\beta}_j}{\left[\hat{\sigma}^2 (X^T X)_{j,j}^{-1} \right]^{1/2}} \quad (2.14)$$

Sau đó chúng ta có thể xem xét loại bỏ khỏi mô hình các biến hồi quy có giá trị tuyệt đối nhỏ của t_j . Ví dụ, hãy cùng xem các bước sau:

1. Lấy công thức ước tính $\hat{\beta}_{ols} = (X^T X)^{-1} X^T Y$ và lấy t-statistics của nó.
2. Nếu có bất kì biến hồi quy j nào có $|t_j| < t_{cutoff}$
 - a. Tìm biến hồi quy j_{min} có giá trị nhỏ nhất là $|t_j|$ và xóa cột j_{min} từ X
 - b. Trở lại bước 1.
3. Nếu $|t_j| > t_{cutoff}$ với mọi giá trị của j còn lại trong mô hình, thì kết thúc.

Các bước thực hiện như vậy trong một tập hợp hồi quy có khả năng cao sẽ được giảm xuống một tập nhỏ hơn được gọi là thủ tục lựa chọn mô hình. Quy trình được thể hiện trong các bước 1, 2 và 3 ở trên mô tả một loại quy trình loại bỏ ngược, trong đó tất cả các biến hồi quy ban đầu được bao gồm nhưng sau đó sẽ được loại bỏ lặp đi lặp lại cho đến khi các biến hồi quy còn lại thỏa mãn một số tiêu chí. Một lựa chọn tiêu chuẩn cho t_{cutoff} là trung vị phần trên (upper quantile) của một t hoặc phân phối chuẩn tắc. Nếu chúng ta áp dụng quy trình trên cho dữ liệu bệnh tiêu đường với $t_{cutoff} = 1.65$ (tương ứng với giá trị $p = 0.10$), sau đó 44 trong 64 biến được loại bỏ, để lại 20 biến trong mô hình hồi quy. Biểu đồ dưới đây cho thấy các giá trị của Y_{test} so với dự đoán dựa trên các hệ số hồi quy của mô hình rút gọn:



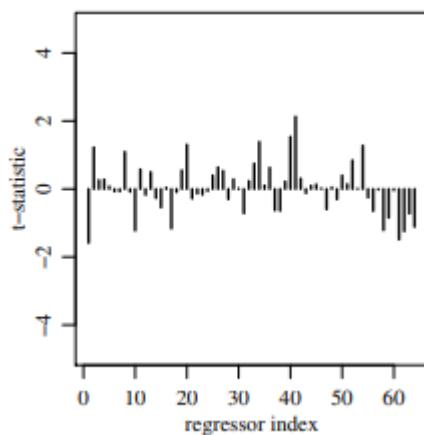
Hình 2.3

Biểu đồ chỉ ra rằng các giá trị dự đoán từ mô hình này chính xác hơn các giá trị từ mô hình đầy đủ và sai số dự đoán trung bình bình phương là:

$$\frac{1}{100} \sum (y_{test,i} - \hat{y}_{test,i})^2 = 0.53$$

Ta có thể thấy nhỏ hơn hẳn so với sai số của mô hình 64 biến.

Tuy nhiên lựa chọn ngược (backwards selection) không phải là không có nhược điểm của nó. Quy trình này sẽ tạo ra mô hình nào nếu không có mối liên hệ nào giữa Y và bất kì một biến hồi quy nào? Chúng ta có thể đánh giá điều này thông qua việc tạo một vector dữ liệu mới \tilde{Y} bằng cách hoán vị ngẫu nhiên các giá trị của Y. Vì trong trường hợp này giá trị của x_i không ảnh hưởng tới giá trị \tilde{y}_i nên mối liên hệ trực tiếp giữa \tilde{Y} và các cột bằng 0. Tuy nhiên mô hình hồi quy OLS vẫn chọn các mối liên kết giả. Biểu đồ dưới đây cho thấy t-statistics cho một hoán vị được tạo ngẫu nhiên là \tilde{Y} của Y trước khi loại bỏ ngược (backwards elimination.)



Hình 2.4

2.3.1 So sánh các mô hình Bayes

Một cách thuận tiện để biểu diễn điều này là viết hệ số hồi quy cho biến j là $\beta_j = z_j \times b_j$ trong đó $z_j \in \{0,1\}$ và b_j là một số thực. Với việc tham số hóa này phương trình hồi quy của chúng ta sẽ trở thành:

$$y_i = z_1 b_1 x_{i,1} + \dots + z_p b_p x_{i,p} + \varepsilon_i$$

Các giá trị của z_j cho biết hệ số hồi quy nào khác 0.

Với mỗi giá trị của $z = (z_1, z_2, \dots, z_p)$ tương ứng với một mô hình khác nhau, hay cụ thể hơn là một tập hợp các biến $x_{i,j}$ khác nhau có hệ số hồi quy khác không.

Ví dụ như, mô hình với $z = (1, 0, 1, 0)$ là một mô hình hồi quy tuyến tính với y như là một hàm về tuổi. Mô hình với $z = (1, 1, 1, 0)$ được gọi là mô hình hồi quy với y như là hàm của tuổi, nhưng được phân loại theo nhóm nhảy aerobic hoặc chạy. Với mô hình tham số hóa này, việc chọn biến nào để đưa vào mô hình hồi quy tương đương với việc chọn z_j nào là 0 z_j nào là 1.

Lựa chọn mô hình theo Bayes được thực hiện bằng cách lấy phân phối hậu nghiệm của z . Để làm được như vậy ta phải có phân phối tiên nghiệm chung trên (z, β, σ^2) .

Nó chính là một phiên bản của g-prior cho phép ta đánh giá $p(y|X, z)$ cho từng mô hình z cụ thể. Đưa ra phân phối tiên nghiệm $p(z)$ trên các mô hình, điều này giúp ta tính được xác suất hậu nghiệm cho mỗi mô hình hồi quy

$$p(z_i | y, X) = \frac{p(z_i) p(y | X, z_i)}{\sum p(z_k) p(y | X, z_k)}$$

Chúng ta cũng có thể so sánh hai mô hình bất kỳ bằng cách xét tỉ lệ hậu nghiệm:

$$\text{odds}(z_a, z_b | y, X) = \frac{p(z_a | y, X)}{p(z_b | y, X)} = \frac{p(z_a)}{p(z_b)} \times \frac{p(y | X, z_a)}{p(y | X, z_b)} \quad (2.15)$$

posterior odds = prior odds × "Bayes factor"

"Bayes factor" được hiểu là bao nhiêu dữ liệu ủng hộ mô hình z_a hơn z_b . Để có được phân phối hậu nghiệm trên các mô hình chúng ta sẽ phải tính toán $p(y | X, z)$.

Tính toán xác suất biến.

Giá trị xác suất biến được tính bởi công thức:

$$\begin{aligned} p(y | X, z) &= \iint p(y, \beta, \sigma^2 | X, z) d\beta d\sigma^2 \\ &= \iint p(y | \beta, X) p(\beta | X, z, \sigma^2) p(\sigma^2) d\beta d\sigma^2 \end{aligned}$$

Áp dụng phân phối g-prior cho β . Bất kì giá trị z nào với số mục nhập p_z khác 0, đặt X_z là ma trận $n \times p_z$ với các biến j mà $z_j = 1$, và β_z là vector $p_z \times 1$ bao gồm các thành phần của β với $z_j = 1$. Phân phối g-prior đã sửa đổi cho β là $\beta_j = 0$ với giá trị j sao cho $z_j = 0$ và ta có:

$$\{\beta_z | X_z, \sigma^2\} \sim \text{multivariate normal}\left(0, g\sigma^2 [X_z^T X_z]^{-1}\right).$$

Kết hợp với công thức tính giá trị xác suất trên ta có:

$$\begin{aligned} p(y | X, z) &= \int \left(\int p(y | X, z, \sigma^2, \beta) p(\beta | X, z, \sigma^2) d\beta \right) p(\sigma^2) d\sigma^2 \\ &= \int p(y | X, z, \sigma^2) p(\sigma^2) d\sigma^2 \end{aligned}$$

Sử dụng công thức tính giá trị $p(y | X, z, \sigma^2)$ ở phần trước ta đặt $\gamma = \frac{1}{\sigma^2}$ và đặt $p(\gamma)$ là mật độ gamma với tham số $(v_0/2, v_0\sigma_0^2/2)$, chúng ta có thể thấy rằng mật độ có điều kiện của (y, γ) khi biết (X, z) là:

$$p(y | X, z, \gamma) \times p(\gamma) = (2\pi)^{-n/2} (1+g)^{-p_z/2} \times \left[\gamma^{n/2} e^{-\gamma SSR_g^z/2} \right] \times \left(v_0 \sigma_0^2 / 2 \right)^{v_0/2} \Gamma(v_0/2)^{-1} \times \left[\gamma^{v_0/2-1} e^{-\gamma v_0 \sigma_0^2/2} \right]$$

(2.16)

Với SSR_g^z tương tự như ở phần trên ngoại trừ việc dựa trên ma trận với hệ số hồi quy X_z

$$SSR_g^z = y^T \left(I - \frac{g}{g+1} X_z \left(X_z^T X_z \right)^{-1} X_z \right) y$$

Như đã tính ở bên trên ta có công thức :

$$\begin{aligned} \gamma^{(v_0+n)/2-1} \exp \left[-\gamma \times (v_0 \sigma_0^2 + SSR_g^z) / 2 \right] &= \\ \frac{\Gamma([v_0+n]/2)}{([v_0 \sigma_0^2 + SSR_g^z]/2)^{(v_0+n)/2-1}} \times dgamma &\left[\gamma, (v_0+n)/2, (v_0 \sigma_0^2 + SSR_g^z)/2 \right] \end{aligned}$$

Mật độ gamma sẽ tích hợp thành giá trị 1 trong tích phân do đó thay vào công thức tích phân để tính $p(y | X, z)$ là:

$$p(y | X, z) = \pi^{-n/2} \frac{\Gamma([v_0+n]/2)}{\Gamma(v_0/2)} (1+g)^{-p_z/2} \frac{(v_0 \sigma_0^2)^{v_0/2}}{(v_0 \sigma_0^2 + SSR_g^z)^{(v_0+n)/2}}$$

(2.17)

Ví dụ 2.3: Tiếp tục với ví dụ 2.1 ta có:

Mô hình hồi quy trong ví dụ về lượng hấp thụ oxy:

$$\begin{aligned} E[y_i | \beta, x_i] &= \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} \\ &= \beta_1 + \beta_2 \times group_i + \beta_3 \times age_i + \beta_4 \times group_i \times age_i \end{aligned}$$

Từ câu hỏi có hay không việc ảnh hưởng của nhóm chạy hay nhảy aerobic chuyển thành câu hỏi liệu β_2 và β_4 có khác 0. Từ công thức tính $p(y|X,z)$ và $p(z|y,X)$ ở trên ta có bảng sau:

z	Model	$\log p(y X,z)$	$p(z y,X)$
(1,0,0,0)	β_1	-44,33	0.00
(1,1,0,0)	$\beta_1 + \beta_2 \times group_i$	-42,35	0.00
(1,0,1,0)	$\beta_1 + \beta_3 \times age_i$	-37,66	0.18
(1,1,1,0)	$\beta_1 + \beta_2 \times group_i + \beta_3 \times age_i$	-36,42	0.63
(1,1,1,1)	$\beta_1 + \beta_2 \times group_i + \beta_3 \times age_i + \beta_4 \times group_i \times age_i$	-37,60	0.19

Bảng 2.1

Chúng ta có thể trực tiếp đánh giá liệu rằng β_2 hay β_4 sẽ bằng 0 bằng cách tính xác suất của dữ liệu theo nhiều mô hình cạnh tranh. Bảng bên trên đã liệt kê năm mô hình hồi quy khác nhau. Sử dụng g-prior cho β với $g = n$ và một đơn vị thông tin phân phối tiên nghiệm cho σ^2 cho mỗi giá trị của z , giá trị của $\log p(y|X,z)$ có thể được tính với từng giá trị của z trên bảng. Nếu ta cho mỗi mô hình có xác suất tiên nghiệm bằng nhau thì xác suất hậu nghiệm cho mỗi mô hình có thể được tính toán. Các tính toán chỉ ra rằng, trong số năm mô hình ở trên, mô hình có thể xảy ra nhất là $z = (1,1,1,0)$ vì mô hình này có độ dốc theo tuổi và có vùng giao thoa riêng cho mỗi nhóm. Xác suất hậu nghiệm của ba mô hình có tuổi cộng lại gần bằng 1. Do đó ta có thể thấy nhóm tuổi có tác động rất mạnh. Bằng chứng về tác động của $group$ thì yếu hơn vì xác suất kết hợp của ba mô hình có $group$ là: $0.00 + 0.63 + 0.19 = 0.82$ Nhưng ta có thể thấy xác suất hậu nghiệm cao hơn khá nhiều so với xác suất tiên nghiệm cho 3 mô hình tương ứng: $0.2 + 0.2 + 0.2 = 0.6$.

Chương 3:

Mô hình hồi quy với các biến rời rạc có thứ tự.

3.1 Hồi quy probit theo thứ tự và bậc likelihood

Giả sử ta quan tâm đến việc mô tả mối quan hệ giữa trình độ học vấn và số trẻ em của các cá nhân trong nơi trú ngụ. Ngoài ra, ta có thể nghi ngờ rằng trình độ học vấn của một cá nhân có thể bị ảnh hưởng bởi trình độ học vấn của cha mẹ họ. General Social Survey đã cung cấp dữ liệu về các biến DEG, CHILD, PDEG cho mỗi cá nhân nước Anh với:

DEG_i : chỉ trình độ đạt được cao nhất của cá nhân thứ i

$CHILD_i$: là số lượng trẻ em của cá nhân đó.

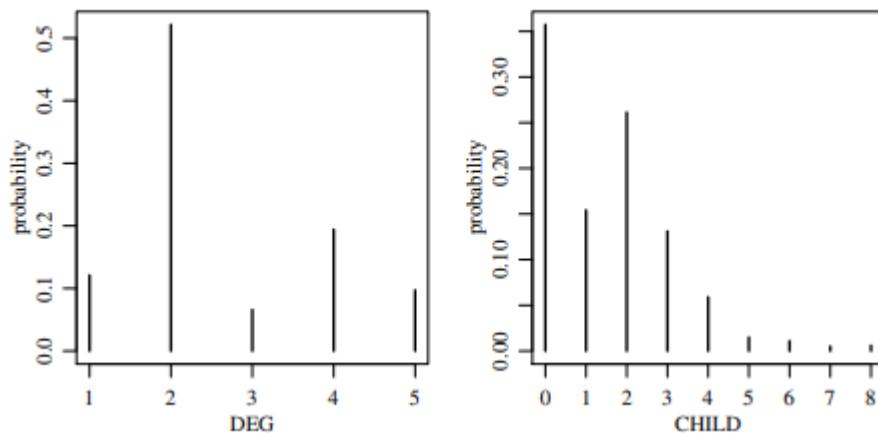
$PDEG_i$: là chỉ số nhị phân thể hiện cha mẹ của người i có được bằng đại học hay không.

Sử dụng những dữ liệu này, chúng ta có thể muốn điều tra mối quan hệ giữa các biến với mô hình hồi quy tuyến tính như sau:

$$DEG_i = \beta_1 + \beta_2 \times CHILD_i + \beta_3 \times PDEG_i + \beta_4 \times CHILD_i \times PDEG_i + \varepsilon_i$$

Với $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \sim i.i.d. normal(0, \sigma^2)$

Tuy nhiên mô hình như vậy sẽ không phù hợp vì một vài lí do sau:



Hình 3.1

Phân bố của DEG và CHILD cho một mẫu có 1002 người đàn ông trong lực lượng lao động vào năm 1994 được thể hiện trong hình trên. Giá trị của DEG được ghi chép là 1 giá trị trong tập $\{1, 2, 3, 4, 5\}$ tương ứng với bằng cấp cao nhất của người đó:

không có bằng cấp, bằng tốt nghiệp trung học, bằng cao đẳng, bằng cử nhân, bằng sau đại học. Ta thấy rằng biến DEG chỉ đảm nhận một tập hợp nhỏ các giá trị rời rạc vậy nên giả định tính chuẩn của phần dư chắc chắn sẽ bị vi phạm. Nhưng có lẽ quan trọng hơn là mô hình hồi quy áp đặt thang đo số cho dữ liệu không thực sự hiện diện. Điều hình là một bằng cử nhân không có giá trị gấp đôi một bằng tốt nghiệp trung học, một bằng cao đẳng không bằng một nửa so với bằng sau đại học. Các giá trị này chỉ có một thứ tự theo nghĩa là một tấm bằng sau đại học cao hơn một tấm bằng của nhân.

Các biến có thứ tự logic của không gian mẫu được biết là giá trị ordinal. Với định nghĩa này DEG và CHILD là giá trị ordinal cũng như các biến chiều cao, cân nặng. Tuy nhiên CHILD, chiều cao và cân nặng là các biến được đo trên thang số có ý nghĩa, nhưng DEG thì không. Chương này ta sẽ dùng thuật ngữ “ordinal” để chỉ bất kỳ biến mà có một trật tự logic của không gian mẫu. Sử dụng thuật ngữ “numeric” để chỉ các biến có thang đo số ý nghĩa và “continuous” nếu một biến có thể có một giá trị bất kì số thực nào trong một khoảng.

3.1.1 Hồi quy probit

Mô hình hồi quy tuyết tính hoặc tuyến tính tổng quát, giả sử dữ liệu có thang đo số ý nghĩa, có thể phù hợp với các biến như CHILD, cân nặng, chiều cao.., nhưng nó không áp dụng được cho các biến có thứ tự không gian mẫu “ordinal” mà là “non-numeric” như DEG. Tuy nhiên, ta thấy rằng các giá trị non-numeric là sự phát sinh từ một số quy trình numeric cơ bản. Ví dụ như mức độ nghiêm trọng của một bệnh nhân có thể được mô tả qua các mốc: nhẹ, bình thường, nặng, mặc dù tình trạng bệnh nhân là biến liên tục. Tương tự như vậy trình độ học vấn của một người có thể là một biến liên tục nhưng cuộc khảo sát chỉ có thể ghi nhận một cách khá là “thô”. Ý tưởng này đã thúc đẩy một kỹ thuật mô hình được gọi là “ordered probit regression”(hồi quy probit có thứ tự). Trong đó chúng ta liên hệ một biến Y là một vector của biến độc lập x thông qua một hồi quy về biến tiềm ẩn Z. Chính xác hơn, ta có mô hình:

$$\begin{aligned}\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n &\sim i.i.d normal(0,1) \\ Z_i &= \beta^T x_i + \varepsilon_i \\ Y_i &= g(Z_i)\end{aligned}$$

Trong đó β và g là tham số chưa biết. Ví dụ như, mô hình phân phối có điều kiện của DEG cho CHILD và PEDG. Ta đặt Y_i là DEG_i và đặt

$$x_i = (CHILD_i, PDEG_i, CHILD_i \times PDEG_i)$$

Hệ số hồi quy β mô tả mối quan hệ giữa biến giải thích (biến X) và biến tiềm ẩn không quan sát được (biến Z). Hàm g liên quan tới giá trị của Z với biến quan sát Y.

Hàm g được coi là hàm không giảm, do đó ta có thể hiểu các giá trị Z nhỏ và lớn tương ứng với các giá trị nhỏ và lớn của Y .

Lưu ý rằng trong mô hình hồi quy probit này, chúng tôi đã lấy phuong sai của $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ bằng 1. Điều này là do quy mô phân phối của Y có thể đã được đại diện bởi g , vì g được phép là bất kì hàm không giảm nào. Tương tự, g có thể biểu thị vị trí phân phối của Y và vì vậy chúng ta không cần bao gồm thuật ngữ chặn trong mô hình.

Nếu không gian mẫu của Y lấy giá trị $K \{1, \dots, K\}$ sau đó hàm g có thể được diễn tả với $K - 1$ tham số được sắp xếp $g_1 < g_2 < \dots < g_{K-1}$ như sau:

$$\begin{aligned} y = g(z) = 1 &\text{ nếu } -\infty = g_0 < z < g_1 \\ y = g(z) = 2 &\text{ nếu } g_1 < z < g_2 \\ &\dots \\ y = g(z) = K &\text{ nếu } g_{K-1} < z < g_K \end{aligned}$$

Giá trị $\{g_1, \dots, g_{K-1}\}$ có thể hiểu là “thresholds” (ngưỡng), vì vậy khi di chuyển z qua một ngưỡng thì y sẽ vào vị trí tiếp theo. Những tham số chưa biết trong mô hình bao gồm hệ số hồi quy β và các ngưỡng g_1, \dots, g_{K-1} . Nếu ta sử dụng phân phối tiên nghiệm chuẩn cho các số này phân phối hậu nghiệm chung của $\{\beta, g_1, \dots, g_{K-1}, Z_1, \dots, Z_n\}$ cho $Y = y = (y_1, \dots, y_n)$ có thể được xấp xỉ bằng cách sử dụng bộ lấy mẫu Gibbs.

Phân phối có điều kiện đầy đủ của β

Cho $Y = y, Z = z, g = (g_1, \dots, g_{K-1})$, phân phối có điều kiện đầy đủ của β chỉ phụ thuộc vào z và thỏa mãn $p(\beta | y, z, g) \propto p(\beta) \times p(z | \beta)$. Cũng giống như trong hồi quy thông thường, phân phối chuẩn đa biến tiên nghiệm cho β cung cấp một phân phối chuẩn đa biến cho hậu nghiệm. Ví dụ ta sử dụng

$\beta \sim \text{multivariate normal}\left(0, n(X^T X)^{-1}\right)$ khi đó $p(\beta | z)$ là phân phối chuẩn đa biến với:

$$\begin{aligned} Var[\beta | z] &= \frac{n}{n+1} (X^T X)^{-1} \\ E[\beta | z] &= \frac{n}{n+1} (X^T X)^{-1} X^T Z \end{aligned}$$

Phân phối có điều kiện đầy đủ của Z.

Phân bố có điều kiện của Z_i phức tạp hơn. Theo mô hình lấy mẫu phân phối có điều kiện của Z_i theo β là $Z_i \sim \text{normal}(\beta^T x_i, 1)$. Cho g, quan sát $Y_i = y_i$ cho chúng ta biết Z_i phải dao động trong khoảng (g_{y_i-1}, g_{y_i}) . Đặt $a = g_{y_i-1}$ và $b = g_{y_i}$, phân phối có điều kiện của Z_i khi biết $\{\beta, y, g\}$ là:

$$p(z_i | \beta, y, g) \propto dnorm(z_i, \beta^T x_i, 1) \times \delta_{(a,b)}(z_i)$$

Đây là mật độ của một phân phối chuẩn bị ràng buộc với $\delta_{(a,b)}(z_i)$ là hàm dirac. Nếu z_i trong khoảng (a, b) thì hàm bằng 1, với các trường hợp còn lại thì bằng 0. Để lấy mẫu một giá trị x từ một phân phối chuẩn (μ, σ^2) bị ràng buộc trong khoảng (a, b) , ta thực hiện hai bước sau:

1. mẫu $u \sim \text{uniform}(\Phi[(a-u)/\sigma], \Phi[(b-\mu)/\sigma])$
2. Đặt $x = \mu + \sigma \times \Phi^{-1}(u)$

Trong đó Φ và Φ^{-1} là cdf và inverse-cdf của phân phối chuẩn.

Phân phối có điều kiện đầy đủ của g

Giả sử phân phối tiên nghiệm của g là một vài mật độ tùy ý $p(g)$. Cho $Y = y, Z = z$, chúng ta biết từ phần 3.3 rằng g_k phải cao hơn tất cả các z_i với $y_i = k$ và nhỏ hơn tất cả các giá trị z_i với $y_i = k+1$. Đặt $a_k = \max\{z_i : y_i = k\}$ và $b_k = \min\{z_i : y_i = k+1\}$ phân phối có điều kiện của g tỉ lệ với $p(g)$ nhưng bị ràng buộc bởi tập hợp

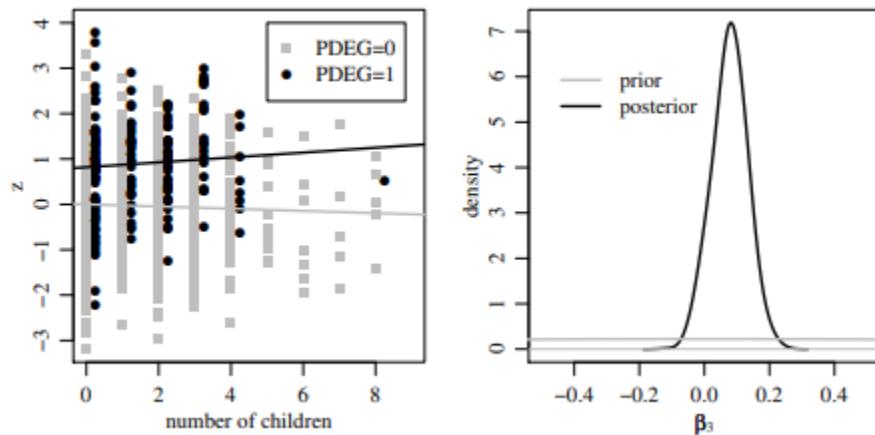
$\{g : a_k < g_k < b_k\}$. Ví dụ, nếu $p(g)$ tỉ lệ thuận với kết quả $\prod_{k=1}^{K-1} dnorm(g_k, \mu_k, \sigma_k)$, nhưng bị ràng buộc bởi g_1, \dots, g_{k-1} , do đó mật độ có điều kiện đầy đủ của g_k là $\text{normal}(\mu_k, \sigma_k^2)$ bị giới hạn trong khoảng (a_k, b_k) .

Ví dụ 3.1.1 : Phân tích về trình độ học vấn

Một số ý kiến cho rằng việc có con làm giảm cơ hội đạt được trình độ học vấn. Ta sẽ xem xét giả thuyết này trong một mẫu nam giới trong lực lượng lao động. Đối với 959 người trả lời khảo sát chúng tôi có dữ liệu đầy đủ về các biến DEG, CHILD, PDEG được mô tả ở trên. Đặt $Y_i = DEG_i$ và $x_i = (CHILD, PDEG_i, CHILD_i \times PDEG_i)$.

Chúng ta ước tính các tham số trong mô hình hồi quy probit bằng cách sử dụng phân phối tiên nghiệm của β là phân phối chuẩn đa biến với phương sai là $n(X^T X)^{-1}$, giá trị kì vọng bằng 0 và $p(g) \propto \prod_{k=1}^{K-1} dnorm(y_k, 0, 100)$ nhưng bị hạn chế bởi

$g_1 < g_2 < \dots < g_{K-1}$. Chúng ta sẽ ước chừng tương ứng các phân phối hậu nghiệm của $\{\beta, z, g\}$ với bộ lấy mẫu Gibbs bao gồm 25000 giá trị. Việc lưu các giá trị tham số sau 25 lần quét sẽ tạo ra 1000 giá trị cho mỗi tham số để ước tính phân phối hậu nghiệm. Đường hồi quy trung bình hậu nghiệm cho những người không có cha mẹ học đại học, $x_{i,2} = 0$ là $E(z|y, x_1, x_2 = 0) = -0.224 \times x_1$ trong đó đường hồi quy cho những người có bố mẹ học đại học là: $E(z|y, x_1, x_2 = 1) = 0.818 + 0.054 \times x_1$. Những dòng này được hiển thị trong bảng sau, cùng với giá trị của z thu được trong lần quét cuối cùng của bộ lấy mẫu Gibbs. Các dòng vẽ gợi ý rằng đối với những người mà cha mẹ không học đại học, số lượng trẻ em thực sự liên quan tiêu cực đến kết quả giáo dục của họ. Tuy nhiên điều ngược lại dường như đúng với những người có cha mẹ học đại học. Phân phối hậu nghiệm của β_3 được đưa ra trong bảng thứ hai của hình cùng với phân phối tiên nghiệm để so sánh. Khoảng tin cậy 95% cho β_3 là $(-0.020, 0.178)$, dù chưa 0 nhưng vẫn thể hiện một lượng bằng chứng hợp lý rằng độ dốc của nhóm $x_2 = 1$ lớn hơn $x_2 = 0$



3.1.1: Kết quả từ phân tích hồi quy probit

3.1.2 Mô hình chuyển đổi và rank likelihood.

Việc phân tích dữ liệu trình độ học vấn ở trên yêu cầu chúng ta chỉ định phân phối tiên nghiệm cho β và chuyển đổi $g(z)$ với ngưỡng $K-1$. Trong khi các phân phối tiên nghiệm đơn giản mặc định cho β tồn tại (như g-prior), điều này không đúng với g. Tiếp theo là phân phối tiên nghiệm cho g đại diện cho thông tin tiên nghiệm thực tế là một nhiệm vụ khó khăn khi K là một số lượng danh mục lớn. Ví dụ như thu nhập (INC) của các đối tượng lao động được ghi nhận là một trong 21 loại được sắp xếp, vậy nên một vòng hồi quy trong đó $Y_i = INC_i$ sẽ yêu cầu g bao gồm 20 tham

số. Vậy nên ước lượng và đặc điểm kĩ thuật tiên nghiệm khi cho một số lượng lớn các tham số như vậy có thể khó khăn.

Do đó, các nhà khoa học đã tìm ra một cách tiếp cận khác để ước tính mà không yêu cầu ta ước tính hàm $g(z)$. Vì ta biết rằng g là hàm không giảm, ta biết điều gì đó về thứ tự của z_i . Ví dụ nếu dữ liệu quan sát là $y_1 > y_2, y_i = g(z_i)$, ta đều biết $g(z_1) > g(z_2)$. Mà g là hàm không giảm, điều này có nghĩa là $z_1 > z_2$. Nói cách khác, khi quan sát $Y = y$, ta biết rằng z_i nằm trong tập hợp:

$$R(y) = \{z \in \mathbb{R}^n : z_{i_1} < z_{i_2} \text{ khi } y_{i_1} < y_{i_2}\}$$

Vì sự phân phối của z_i không phụ thuộc vào g , xác suất mà $Z \in R(y)$ khi cho trước y không phụ thuộc vào hàm g . Điều này cho thấy rằng chúng ta sẽ dựa trên suy luận hậu nghiệm của chúng ta về $Z \in R(y)$. Phân phối hậu nghiệm của chúng ta cho β trong trường hợp này:

$$\begin{aligned} p(\beta | Z \in R(y)) &\propto p(\beta) \times \Pr(Z \in R(y) | \beta) \\ &= p(\beta) \times \int_{R(y)} \prod_{i=1}^n d\text{norm}(z_i, \beta^T x_i, 1) dz_i \end{aligned}$$

Xác suất $\Pr(Z \in R(y) | \beta)$ được biết như rank likelihood. Nó được gọi là rank likelihood vì với dữ liệu liên tục nó chứa thông tin tương tự về y khi biết thứ hạng của $\{y_1, y_2, \dots, y_n\}$ cái nào có giá trị lớn nhất cái nào có giá trị nhỏ nhất... Và điều quan trọng là với bất kì giá trị đầu ra được sắp xếp Y (non-numeric, numeric, discrete hoặc continuous) thông tin về β có thể được lấy từ $\Pr(Z \in R(y) | \beta)$ mà không cần phải xác định $g(z)$.

Với bất kì giá trị nào của β , giá trị của $\Pr(Z \in R(y) | \beta)$ liên quan đến một tích phân rất phức tạp, khó tính toán. Tuy nhiên bằng cách ước tính Z đồng thời với β , ta có thể có được ước tính của β mà không cần phải tính toán bằng $\Pr(Z \in R(y) | \beta)$.

Phân phối chung hậu nghiệm của $\{\beta, Z\}$ có thể được xấp xỉ bằng cách sử dụng lấy mẫu Gibb, lấy mẫu luôn phiên từ các bản phân phối có điều kiện đầy đủ, phân phối có điều kiện đầy đủ của β rất dễ dàng. Cho một giá trị hiện tại z của Z , mật độ có điều kiện đầy đủ $p(\beta | Z = z, Z \in R(y))$ giảm xuống $p(\beta | Z = z)$ bởi vì ta biết giá trị của Z có nhiều thông tin hơn là chỉ có Z nằm trong tập $R(y)$. Một phân phối chuẩn đa biến tiên nghiệm cho β sau đó dẫn tới một phân phối chuẩn đa biến có điều kiện đầy đủ như trước. Các phân phối có điều kiện đầy đủ của z_i cũng rất đơn giản để rút

ra. Hãy xem xét phân phối đầy đủ có điều kiện của z_i cho $\{\beta, Z \in R(y), Z_{-i}\}$ với Z_{-i} biểu thị tất cả các giá trị của Z trừ Z_i . Có điều kiện trên β , phân phối z_i là chuẩn $(\beta^T x_i, 1)$. Có điều kiện trên $\{\beta, Z \in R(y), Z_{-i}\}$, mật độ của z_i là tỉ lệ thuận với mật độ bình thường nhưng bị hạn chế bởi $Z \in R(y)$. Cùng nhớ lại bản chất của ràng buộc này: $y_i < y_j$ dẫn đến $z_i < z_j$ và $y_i > y_j$ dẫn đến $z_i > z_j$. Điều này có nghĩa là z_i phải nằm trong khoảng:

$\max\{z_j : y_j < y_i\} < z_i < \min\{z_j : y_i < y_j\}$. Đặt a, b biểu thị các giá trị số của các điểm cuối dưới và trên của khoảng này thì phân phối có điều kiện đầy đủ của z_i là:

$$p(z_i | \beta, Z \in R(y), Z_{-i}) \propto dnorm(z_i, \beta^T x_i, 1) \times \delta_{(a,b)}(z_i)$$

Phân phối có điều kiện đầy đủ này hoàn toàn giống với phân phối của z_i trong mô hình probit có thứ tự, ngoại trừ việc các ràng buộc đối với z_i được xác định trực tiếp bởi các giá trị z_{-i} thay vì các biến threshold(ngưỡng). Như vậy, lấy mẫu từ phân phối có điều kiện đầy đủ này rất giống với lấy mẫu từ phân phối tương tự trong mô hình hồi quy probit. Cách tiếp cận rank likelihood có thể áp dụng cho một loạt các bộ dữ liệu rộng vì với cách tiếp cận này Y được phép là bất kì loại biến số thứ tự, rời rạc, liên tục... Hạn chế của việc sử dụng rank likelihood là nó không cung cấp cho ta suy luận về $g(z)$ mô tả mối quan hệ giữa các biến tiềm ẩn và các biến quan sát. Nếu tham số này được quan tâm thì rank likelihood không phù hợp, nhưng nếu chỉ quan tâm đến β , thì rank likelihood cung cấp một sự thay thế đơn giản cho mô hình probit được sắp xếp.

Kết luận:

“Tìm hiểu về thống kê Bayes trong mô hình hồi quy tuyến tính” là một đề tài nghiên cứu rất hay và thú vị. Thông qua việc tổng hợp các kiến thức có trong tài liệu tiếng Anh và tài liệu tiếng Việt, khóa luận đã trình bày xuyên suốt các kiến thức từ phần cơ bản là giới thiệu chung về thống kê Bayes cho tới ước lượng Bayes trong mô hình hồi quy tuyến tính, chọn mô hình Bayes có tỉ lệ tốt nhất.

Dưới sự hướng dẫn của thầy giáo TS. Trịnh Quốc Anh em đã được tìm hiểu các kiến thức thống kê Bayes mà em chưa được học trong chương trình đại học. Nội dung chính của khóa luận là:

1. Cung cấp các kiến thức cơ bản về thống kê Bayes
2. Ứng dụng của Bayes trong việc giải bài toán có mô hình hồi quy tuyến tính, trình bày cụ thể về phân phối hậu nghiệm và cách cập nhật cho các biến hồi quy. Và cách chọn mô hình Bayes có tỉ lệ dữ liệu ủng hộ, xác xuất hậu nghiệm cao nhất.
3. Ứng dụng thống kê Bayes trong việc giải quyết bài toán khi dữ liệu được đo trên thang số không có ý nghĩa bằng cách sử dụng mô hình hồi quy Probit.

Đối với bản thân em, nội dung luận văn là những kiến thức thực sự hữu dụng trong cuộc sống. Vì thế sau này nếu có thêm nhiều cơ hội, em nhất định sẽ tìm hiểu thật sâu sắc về lý thuyết thống kê Bayes cũng như hướng tới ứng dụng của nó trong các bài toán thực tế.

Dù đã cố gắng nhưng do hạn chế của bản thân và thời gian có hạn nên khóa luận này khó tránh khỏi những thiếu sót. Em rất mong nhận được những góp ý nhận xét của các thầy cô.

Tài liệu tham khảo

Tài liệu Tiếng Việt

1. Nguyễn Văn Hữu, Nguyễn Hữu Dư, *Phân tích thống kê và dự báo* (2003)
2. Đặng Hùng Thắng, *Quá trình ngẫu nhiên và tính toán ngẫu nhiên*, NXB Đại học Quốc Gia Hà Nội (2007)
3. Website về Machine learning của TS. Vũ Hữu Tiệp:
<https://machinelearningcoban.com>

Tài liệu Tiếng Anh

4. Peter D.Hoff, *A First Course in Bayesian Statistical Methods*, (2009)
5. Andrew Gelman, Jonh B. Carlin, Hal S. Stern, Donald B. Rubin, *Bayesian Data Analysis*, Chapman and Hall/CRC (2004)