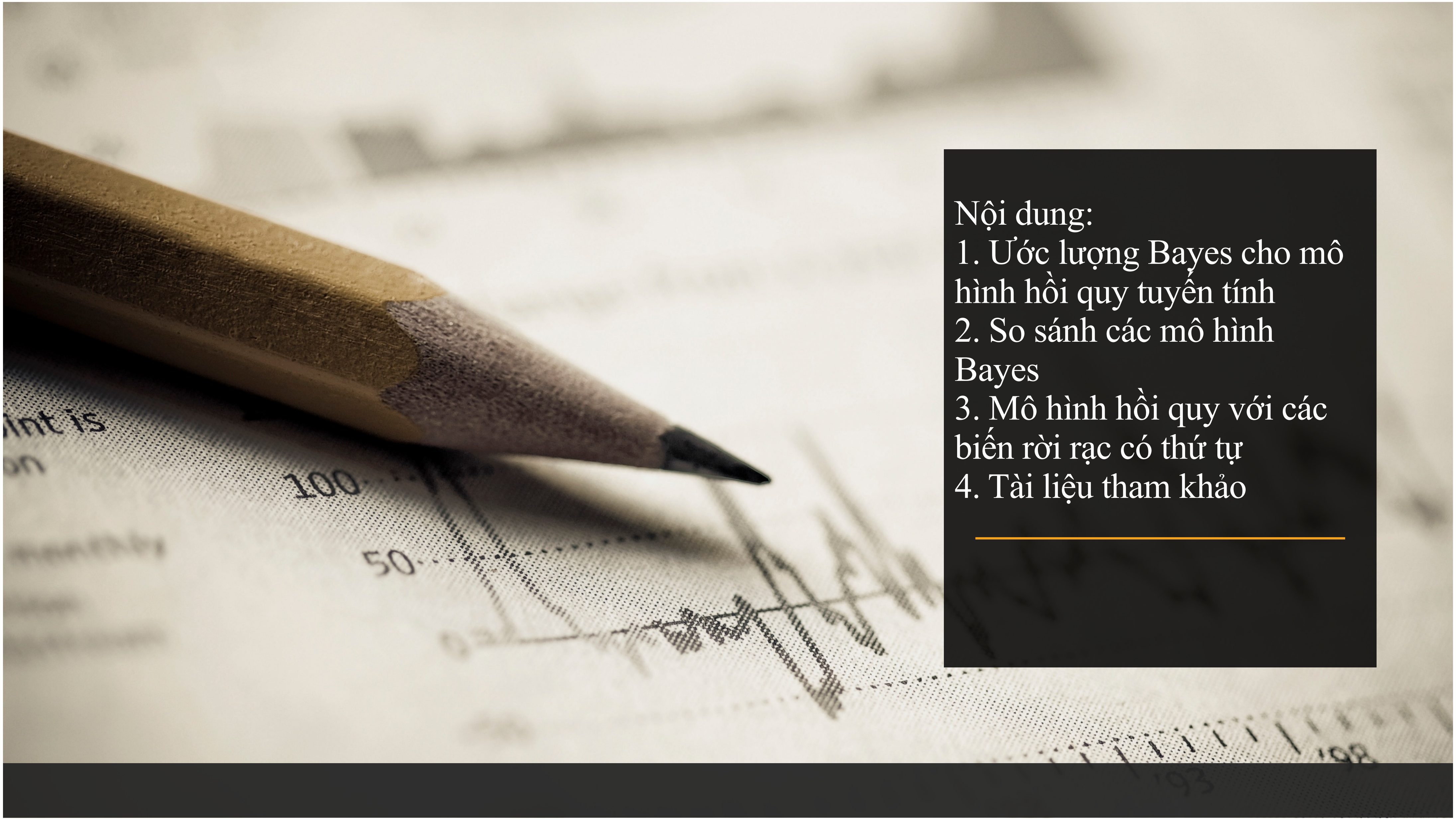


Nội dung:

1. Ước lượng Bayes cho mô hình hồi quy tuyến tính
2. So sánh các mô hình Bayes
3. Mô hình hồi quy với các biến rời rạc có thứ tự
4. Tài liệu tham khảo

Mô hình hồi quy tuyến tính

Phương trình tuyến tính:

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

Mô hình hồi quy tuyến tính chuẩn có dạng

$$\left\{ y \mid \beta, \sigma^2, X \right\} \sim \mathrm{N}_n \left(X\beta, \sigma^2 I_n \right)$$

Mật độ xác xuất chung của dữ liệu quan sát:

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

Mô hình hồi quy tuyến tính theo phương pháp bình phương tối thiểu

Tổng số dư bình phương:

$$SSR(\beta) = \sum_{i=1}^{n} \left(y_i - \beta^T x_i \right)^2 = (y - X\beta)^T (y - X\beta)$$

$$= Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta$$

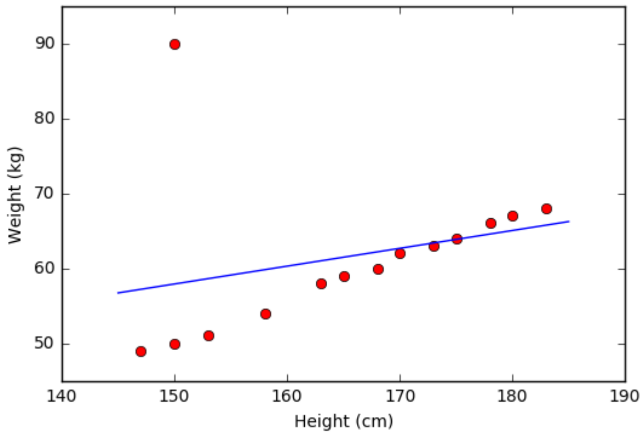
Để tổng số dư đạt giá trị nhỏ nhất:

$$\frac{d}{d\beta} SSR(\beta) = 0$$

$$\beta = \left(X^T X \right)^{-1} X^T Y$$

Hạn chế

Nhạy cảm với nhiễu



Không biểu diễn được các mô hình phức tạp

Ước lượng Bayes cho mô hình hồi quy

Mật độ xác xuất của dữ liệu quan sát:

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

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

Ta thấy β và Y có vai trò gần tương đương nhau

$$\beta \sim \text{multivariate normal}\left(\beta_0, \Sigma_0\right)$$

Phân phối hậu nghiệm của β là:

$$p\left(\beta \mid \mathrm{Y}, \mathrm{X}, \sigma^2\right)$$

$$\propto p\left(Y \mid X, \beta, \sigma^2\right) \times p(\beta)$$

$$\propto \exp\left\{\left(\Sigma_0^{-1}\beta_0 + X^T Y / \sigma^2\right) - \frac{1}{2}\beta^T\left(\Sigma_0^{-1} + X^T X / \sigma^2\right)\beta\right\}$$

$$\mathrm{Var}\left[\beta \mid \mathrm{Y}, \mathrm{X}, \sigma^2\right] = \left(\Sigma_0^{-1} + X^T X / \sigma^2\right)^{-1}$$

$$\mathrm{E}\left[\beta \mid \mathrm{Y}, \mathrm{X}, \sigma^2\right] = \left(\Sigma_0^{-1} + X^T X / \sigma^2\right)^{-1}\left(\Sigma_0^{-1}\beta_0 + X^T Y / \sigma^2\right)$$

Ước lượng Bayes cho mồi hình hồi quy

Tìm phân phối hậu nghiệm của σ^2

$$\gamma = \frac{1}{\sigma^2} \qquad \gamma \sim gamma\left(\frac{v_0}{2}, \frac{v_0 \sigma_o{}^2}{2}\right)$$

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

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

Ước lượng Bayes cho mô hình hồi quy

Dùng bộ lấy mẫu Gibbs để xấp xỉ phân phối: $p\left(\beta,\sigma^2 \mid \mathrm{Y},\mathrm{X}\right)$

Giá trị đầu vào: $\left\{\beta^{(s)},\sigma^{2(s)}\right\}$

Cập nhật β

Tính $V=\mathrm{var}\left[\beta \mid \mathrm{Y},\mathrm{X},\sigma^{2(s)}\right]$ và $m=E\left[\beta \mid Y,X,\sigma^{2(s)}\right]$

Cập nhật: $\beta^{(s+1)} \sim \text{multivariate normal}\left(m,V\right)$

Cập nhật σ^2

Tính $SSR\left(\beta^{(s+1)}\right)$

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

So sánh mô hình Bayes (Bayesian model comparison)

$\beta_j = z_j \times b_j$ Trong đó $z_j \in \{0,1\}$

$b_j \in \sim$

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

Với mỗi giá trị $z = (z_1, z_2, \ldots, z_p)$ tương ứng với mô hình hồi quy khác nhau

So sánh mô hình Bayes được thực hiện bằng cách lấy phân phôi hậu nghiệm của z.

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

So sánh mô hình Bayes (Bayesian model comparison)

So sánh hai mô hình bằng cách xét tỉ lệ hậu nghiệm:

$$odds\left(z_a, z_b \mid y, X\right) = \frac{p\left(z_a \mid y, X\right)}{p\left(z_b \mid y, X\right)} = \frac{p\left(z_a\right)}{p\left(z_b\right)} \times \frac{p\left(y \mid X, z_a\right)}{p\left(y \mid X, z_b\right)}$$

$$posterior\ odds = prior\ odds \times "Bayes\ factor"$$

"Bayes factor" thể hiện tỉ lệ dữ liệu ủng hộ mô hình z_a với mô hình z_b

Để tính được phân phối hậu nghiêm z ta cần biết $p\left(y \mid X, z\right)$

$$p\left(y \mid X, z\right) = \int\left(\int p\left(y \mid X, z, \sigma^2, \beta\right) p\left(\beta \mid X, z, \sigma^2\right) d\beta\right) p\left(\sigma^2\right) d\sigma^2$$

$$= \int p\left(y \mid X, z, \sigma^2\right) p\left(\sigma^2\right) d\sigma^2$$

So sánh mô hình Bayes (Bayesian model comparison)

$$\gamma = \frac{1}{\sigma^2} \qquad\qquad \gamma \sim gamma\left(\frac{v_0}{2}, \frac{v_0 \sigma_o{}^2}{2}\right)$$

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

So sánh mô hình Bayes (Bayesian model comparison)

$$\gamma^{(v_0+n)/2-1}\exp\left[-\gamma\times\left(v_0\sigma_0^2+SSR_g^z\right)/2\right]=$$

$$\frac{\Gamma\left(\left[v_0+n\right]/2\right)}{\left(\left[v_0\sigma_0^2+SSR_g^z\right]/2\right)^{(v_0+n)/2-1}}\times dgamma\left[\gamma,\left(v_0+n\right)/2,\left(v_0\sigma^2+SSR_g^z\right)/2\right]$$

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

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

So sánh mô hình Bayes (Bayesian model comparison)

Ví dụ: ước lượng mức hấp thụ oxy với bộ dữ liệu là 12 người đàn ông trong đó 6 người tập chạy và 6 người tập aerobic.

Phương trình hồi quy có dạng:

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

$x_{i,1} = 1$

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

$x_{i,3} =$ tuổi của người thứ i

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

So sánh mô hình Bayes (Bayesian model comparison)

Với bộ dữ liệu đầu vào:

$$x_3 = (23, 22, 22, 25, 27, 20, 31, 23, 27, 28, 22, 24)$$

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

Nghi ngờ sự ảnh hưởng của nhóm chạy hay nhảy aerobic, do đó ta sẽ xét xem liệu β_2 và β_4 có khác 0.

z	Model	$\log p(y \mid X, z)$	$p(z \mid y, X)$
$(1,0,0,0)$	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

So sánh mô hình Bayes (Bayesian model comparison)

"age" có tác động rất mạnh, vì xác suất hậu nghiệm của 3 mô hình có "age" cộng lại gần bằng 1.

"group" có xác suất kết hợp của ba mô hình là $0.00+0.63+0.19=0.82$ .Xác suất hậu nghiệm cao hơn khá nhiều so với xác suất tiên nghiệm $0.2+0.2+0.2=0.6$

Từ tính toán trên, mô hình có thể xảy ra nhất là: $z=(1,1,1,0)$

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

Mô hình hồi quy Probit có thứ tự: Liên hệ biến Y là một vector với biến độc lập X thông qua biến tiềm ẩn Z

$$\varepsilon_1, \varepsilon_2, ..., \varepsilon_n \sim i.i.d\ normal(0,1)$$

$$Z_i = \beta^T x_i + \varepsilon_i$$

$$Y_i = g(Z_i)$$ g là hàm không giảm

$$y = g(z) = 1 \qquad -\infty = g_0 < z < g_1$$

$$y = g(z) = 2 \qquad g_1 < z < g_2$$

$$y = g(z) = K \qquad g_{K-1} < z < g_K = +\infty$$

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

Phân phối có điều kiện đầy đủ của β

Với $Y = y, Z = z, g = \left(g_1, \ldots., g_{K-1}\right)$

$$p\left(\beta \mid y, z, g\right) \propto p\left(\beta\right) \times p\left(z \mid \beta\right)$$

$$\beta \sim \text{multivariate normal}\left(0, n\left(X^T X\right)^{-1}\right)$$

$$Var\left[\beta \mid z\right] = \frac{n}{n+1}\left(X^T X\right)^{-1}$$

$$E\left[\beta \mid z\right] = \frac{n}{n+1}\left(X^T X\right)^{-1} X^T Z$$

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

Phân phối có điều kiện của Z_i theo β :

$$Z_i \sim normal\left(\beta^T x_i, 1\right)$$

Phân phối có điều kiện đầy đủ của Z_i theo $\{\beta, y, g\}$

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

Với $Y_i = y_i \quad a = g_{y_i - 1} \quad b = g_{y_i}$

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

Phân phối có điều kiện đầy đủ của g

Giả sử phân phối tiên nghiệm của g là một mật độ tùy ý

$$a_k = \max\{z_i : y_i = k\}$$

$$b_k = \min\{z_i : y_i = k+1\}$$

$$\{g : a_k < g_k < b_k\}$$

Ví dụ phân phối hậu nghiệm của g tỉ lệ với kết quả:

$$\prod_{k=1}^{K-1} dnorm(g_k, \mu_k, \sigma_k)$$

$$g_1, \ldots, g_{k-1}$$

$$\text{normal}(\mu_k, \sigma_k^2)$$

$$(a_k, b_k)$$

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

Ví dụ: Phân tích về trình độ học vấn

Một số ý kiến cho rằng có con làm giảm cơ hội đạt được trình độ học vấn. Ta sẽ xem xét giả thuyết này với lực lượng nam giới. $Y_i = DEG_i \quad x_i = \left(CHILD, PDEG_i, CHILD_i \times PDEG_i\right)$

$CHILD$: số lượng trẻ em.

$PDEG$:trình độ học vấn của bố mẹ

DEG :trình độ học vấn của người lao động.

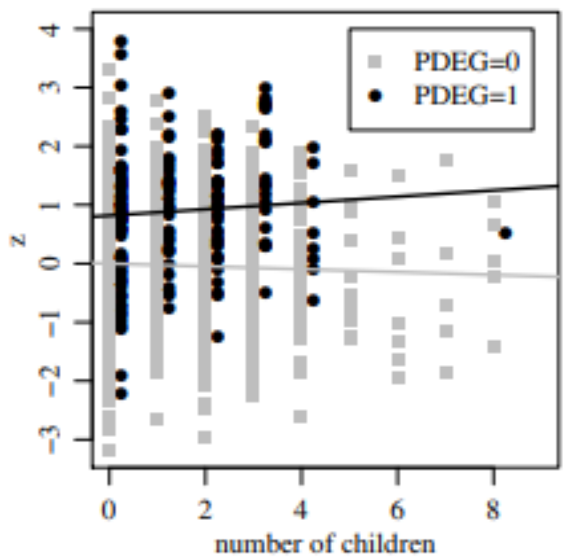
Ta sẽ ước chừng các phân phối hậu nghiệm của $\{\beta, z, g\}$ với bộ lấy mẫu Gibbs.

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

Với $x_{i,2} = 0$ ta có $E\left(z \mid y, x_1, x_2 = 0\right) = -0.224 \times x_1$

Với $x_{i,2} = 1$ ta có $E\left(z \mid y, x_1, x_2 = 1\right) = 0.818 + 0.054 \times x_1$

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 với những người có cha mẹ học đại học thì điều ngược lại dường như đúng.



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

Mô hình chuyển đổi và rank likelihood.

Ước tính β mà không thông qua giá trị g(z)

$$\left.\begin{aligned} & y_1 > y_2 \\ & y_i = g(z_i) \end{aligned}\right\} \quad \begin{aligned} & g(z_1) > g(z_2) \\ & z_1 > z_2 \end{aligned}$$

$$R(y) = \left\{ z \in \sim^{n} : z_{i_1} < z_{i_2} \, khi \, y_{i_1} < y_{i_2} \right\}$$

$$p(\beta \mid Z \in R(y)) \propto p(\beta) \times \Pr(Z \in R(y) \mid \beta)$$

$$= p(\beta) \times \int_{R(y)} \prod_{i=1}^{n} dnorm\left(z_i, \beta^T x_i, 1\right) dz_i$$

$\Pr(Z \in R(y) \mid \beta)$ được gọi là rank likelihood

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

Mô hình chuyển đổi và rank likelihood

$\Pr\left(Z \in R\left(y\right) | \beta\right)$ khó tính toán

Ước tính Z đồng thời với β

Phân phối có điều kiện đầy đủ của β

$p\left(\beta | Z = z, Z \in R\left(y\right)\right)$ giảm xuống $p\left(\beta | Z = z\right)$

Phân phối có điều kiện đầy đủ của Z_i

Có điều kiện trên β: phân phối Z_i là chuẩn $\left(\beta^T x_i, 1\right)$

Có điều kiện trên $\left\{\beta, Z \in R\left(y\right), Z_{-i}\right\}$: bị hạn chế bởi $Z \in R\left(y\right)$

$$a = \max\left\{z_j : y_j < y_i\right\} < z_i < \min\left\{z_j : y_i < y_j\right\} = b$$

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

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

Ưu điểm: có thể áp dụng cho các bộ dữ liệu rộng vì với cách tiếp cận này Y được phép là bất kì loại biến số có thứ tự, rời rạc, liên tục……

Nhược điểm: Không cung cấp cho ta biết suy luận về $g(z)$ mô tả mối quan hệ giữa biến tiềm ẩn và biến quan sát

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. Vũ Hữu Tiệp, Machine learning cơ bản. *https://machinelearningcoban.com*

Tài liệu tiếng Anh:

4. Peter D. Holf. *A First Course in Bayesian Statistical Methods,* (2009)

5. Andrew Gelman, Jonh B. Carlin, Hal S. tern, Donald B. Rubin. *Bayesian Data Analysis*, Chapman and Hall/CRC (2004)