Tính toán Bayes cho mô hình hồi quy logistic

Group 1: Lecturer Dr. Do Van Cuong

Khoá bồi dưỡng thống kê nâng cao cho giảng viên phía Nam,
Trường Đại học Kinh tế UEH, TP HCM 12/06/2022

STT	Họ và tên	Đơn vị công tác	Số điện thoại	Email
1	Nguyễn Thị Như Quỳnh	ĐH Ngân hàng TP. Hồ Chí Minh	0902208422	quynhntn@buh.edu.vn
2	Đỗ Hoàng Oanh	ĐH Ngân hàng TP. Hồ Chí Minh	0942927997	Oanhdh@buh.edu.vn
3	Hoàng Thị Diễm Hương	ĐH Kinh tế TP.HCM	0987084990	diemhuonga1@ueh.edu.vn
4	Phạm Trí Cao	ĐH Kinh tế TP.HCM	0908092224	phamtricao@ueh.edu.vn
5	Nguyễn Thảo Nguyên	ĐH Kinh tế TP.HCM	0983637812	nguyennt@ueh.edu.vn
6	Trần Hà Quyên	ĐH Kinh tế TP.HCM	0979020293	quyentran@ueh.edu.vn
7	Nguyễn Lý Kiều Chinh	ĐH Kinh tế TP.HCM-PHVL	0903149497	chinhnlk@ueh.edu.vn
8	Phạm Thị Thương	ĐH Kiên giang	0907080014	ptthuong@vnkgu.edu
9	Tôn Hoàng Hồ	ĐH Kiên giang	0917244476	thho@vnkgu.edu
10	Danh Ngọc Thắm	ĐH Kiên Giang	0981422252	dntham@vnkgu.edu.vn
11	Đoàn Thiện Minh	ĐH Lạc Hồng	0938707701	dtminh@lhu.edu.vn
12	Phan Đình Khôi	ĐH Cần Thơ	0907552277	pdkhoi@ctu.edu.vn
13	Vũ Quang Mạnh	Đại học Luật TP.HCM	978441111	vqmanh@hcmulaw.edu.vn
14	Nguyễn Chí Thiện	Viện NC&ĐT Việt-Anh, ĐHĐN	358505886	thien.nguyen@vnuk.edu.vn
15	Phạm Thị Thu Hường	ĐH An Giang	0839311321	ptthuong@agu.edu.vn
16	Huỳnh Văn Hiếu	Đại học Công nghiệp TP.HCM	0988535104	huynhvanhieu@iuh.edu.vn

Mô tả các biến

Bộ dữ liệu MROZ.xlsx bao gồm thông tin về sự tham gia vào lực lượng lao động của phụ nữ. Xây dựng mô hình hồi quy logistic cho bộ số liệu.

Biến phụ thuộc là inlf: đây là biến nhị phân là cho biết một người phụ nữ có tham gia vào lực lượng lao động hay không.
Các biến độc lập:

(1) nwifeinc: thu nhập
(2) educ: số năm học
(3) exper: số năm kinh nghiệm
(4) expersq: bình phương số năm kinh nghiệm
(5) kidslt6: số trẻ em dưới 6 tuổi
(6) kidsge6: số trẻ em trên 6 tuổi

Thống kê mô tả

(1) Biến định tính: bảng tần số (frequencies table).

(2) Biến định lượng: Biểu đồ tần suất (histogram), các số đặc trưng (mean, variance, mode, median), boxplot.

(3) Tương quan giữa các biến: covariance matrix (cov), biểu đồ tương quan cặp (corplot).

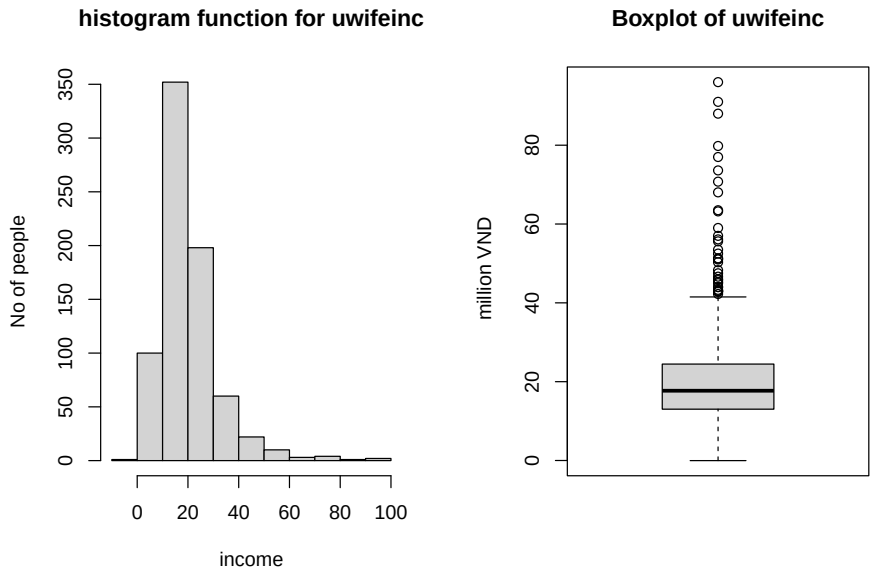
```
> table(y)
y
  0   1
325 428
```

Thống kê mô tả

(1) Biến thu nhập (nwifeinc):

Min. :-0.02906 1st Qu.:13.02504 Median :17.70000 Mean :20.12896 3rd Qu.:24.46600 Max. :96.00000

Mean (x2)= 20.12896 sd=11.6348 se= 0.4239956

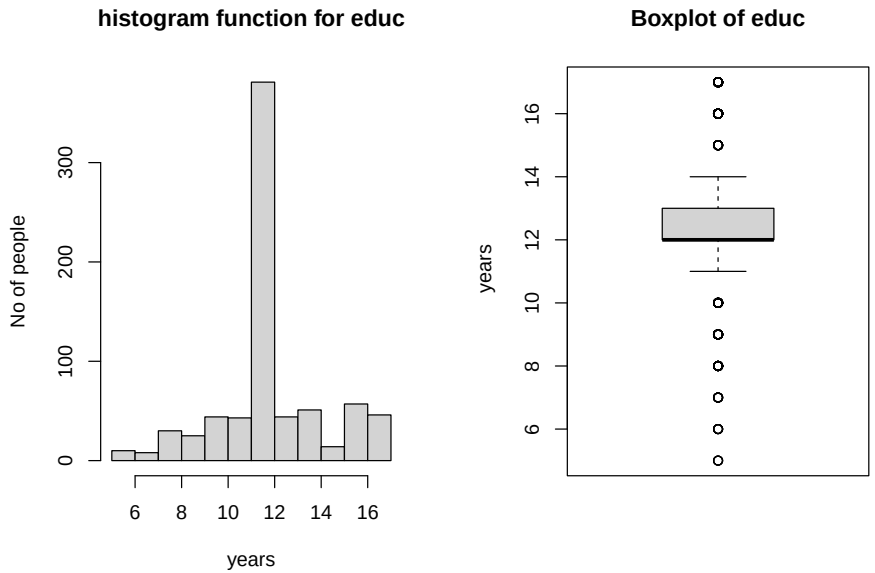


Hình 1: Histogram and Boxplot for uwifeincome

Thống kê mô tả

(2) Biến số năm học (educ):

Min. : 5.00 1st Qu.:12.00 Median :12.00 Mean :12.29 3rd Qu.:13.00
Max. :17.00
Mean(x3)= 12.28685 sd=2.280246 se= 0.08309678



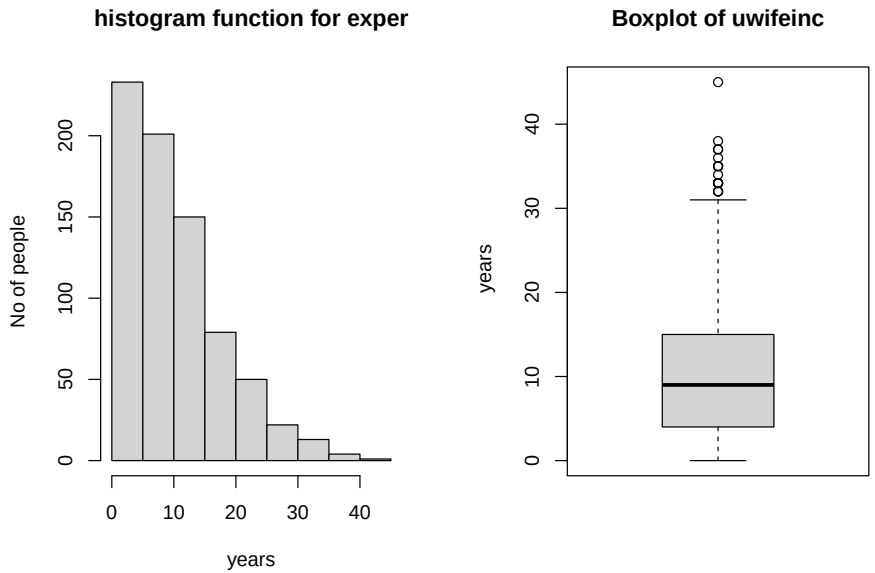
Hình 2: Histogram and Boxplot for education

Thống kê mô tả

(3) Biến số năm kinh nghiệm (exper):

Min. : 0.00 1st Qu.: 4.00 Median : 9.00 Mean :10.63 3rd Qu.:15.00 Max. :45.00

Mean(x4)= 10.63081 sd=8.06913 se= 0.2940554

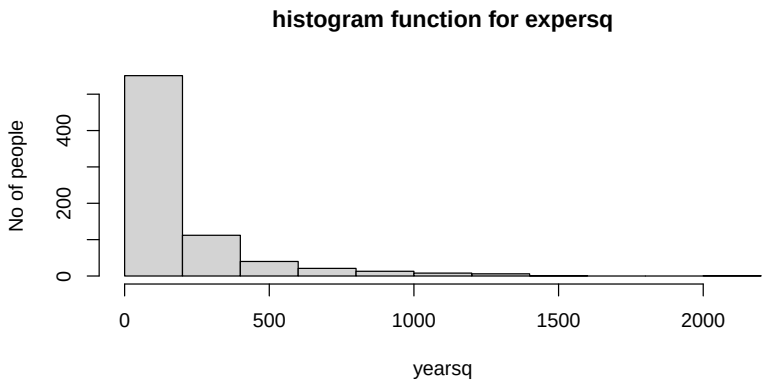


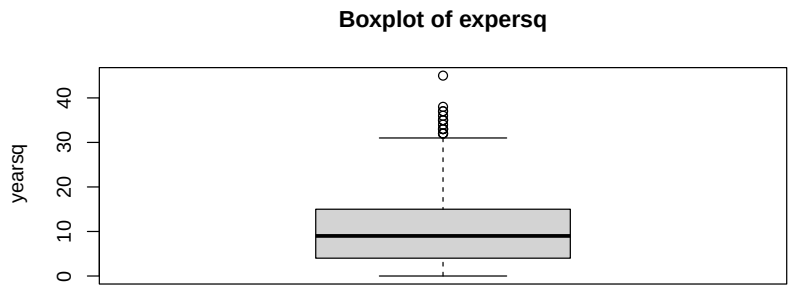
Hình 3: Histogram and Boxplot for experience

Thống kê mô tả

(4) Biến số năm kinh nghiệm bình phương (expersq):

Min. : 0 1st Qu.: 16 Median : 81 Mean : 178 3rd Qu.: 225 Max. :2025
Mean(x5)= 178 sd=249.6308 se= 9.097054



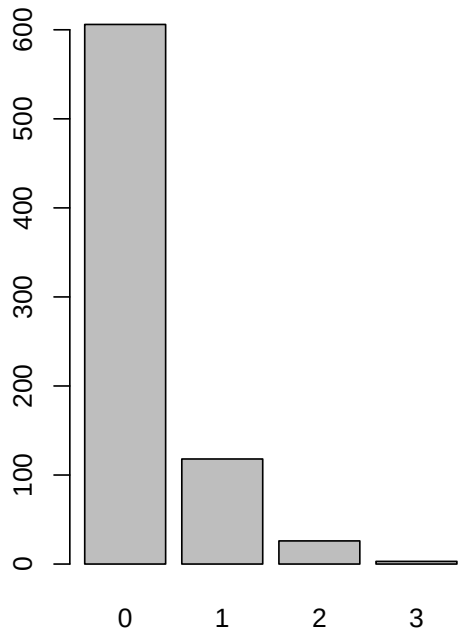


Hình 4: Histogram and Boxplot for experience square

Thống kê mô tả

(4) Biến số trẻ em dưới 6 tuổi (kidslt6):

0	1	2	3
606	118	26	3

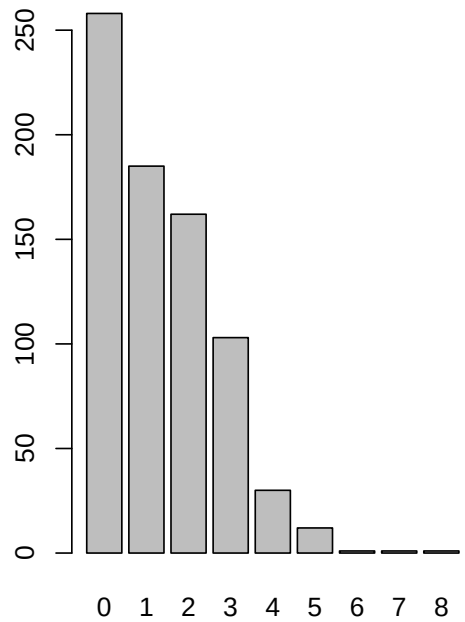


Hình 5: Barplot for kidslt6

Thống kê mô tả

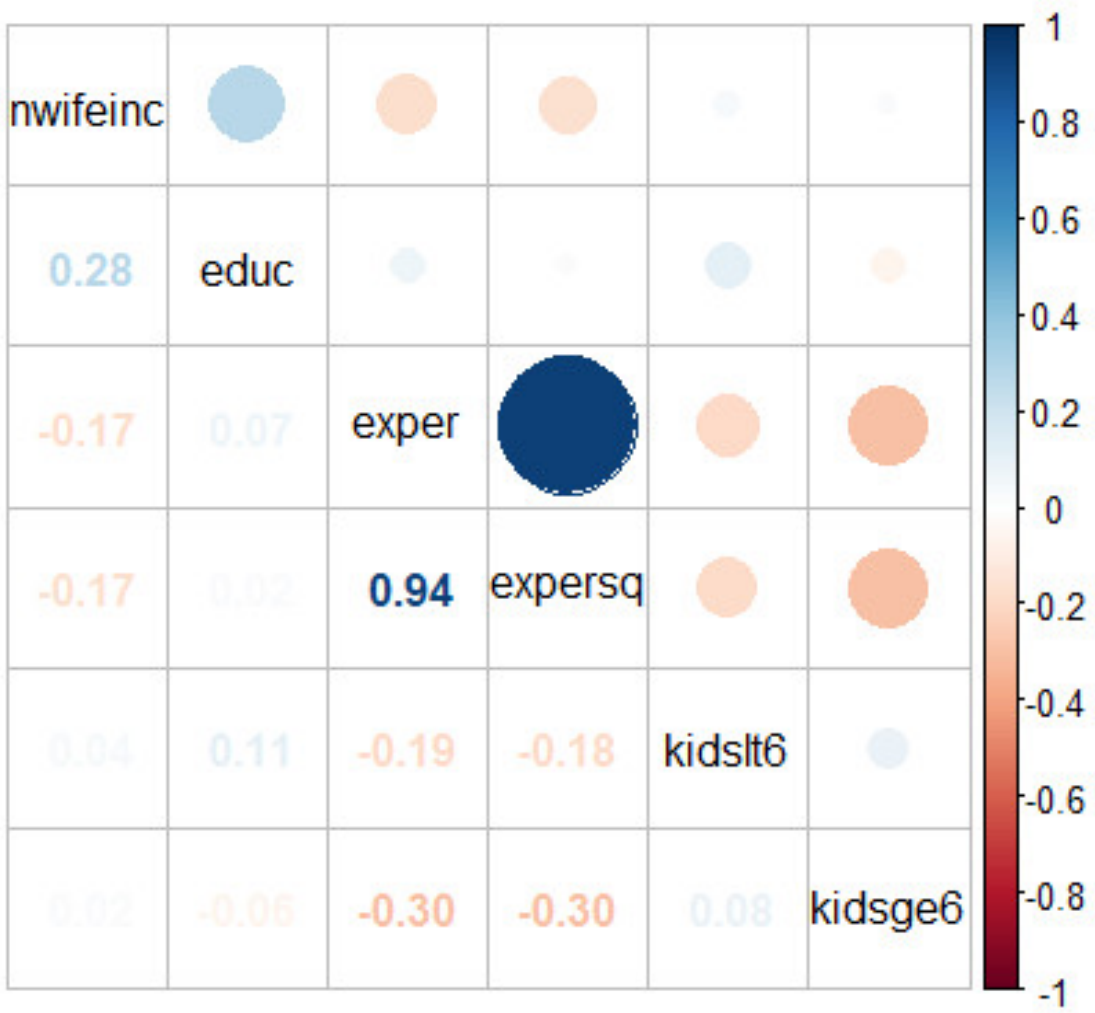
(4) Biến số trẻ em trên 6 tuổi (kidsge6):

0	1	2	3	4	5	6	7	8
258	185	162	103	30	12	1	1	1



Hình 6: Barplot for kidsge6

Thống kê mô tả



Hình 8: scatter plots

Mô hình logistic đa biến

$$p = \frac{e^{X\beta}}{1 + e^{X\beta}},$$
$$\log\left(\frac{p}{1-p}\right) = X\beta,$$
$$Y \sim \text{Ber}(p).$$

- Phương pháp ước lượng hợp lý cực đại.
- Phương pháp Bayes: thuật toán RWMH.

glm(y x2+x3+x4+x5+x6+x7, data = mroz, family = binomial)
> beta.MLE intercept: -3.739706745
beta2 (uwifeinc): -0.030117125
beta3 (educ): 0.252003837
beta4 (exper): 0.205738709
beta5 (expersq):-0.003912971
beta6 (kidslt6): -0.917512558
beta7 (kidsge6): 0.222616439
Mô hình đạt được

$$\log\left(\frac{p}{1-p}\right) = -3.7397 - 0.0301 * \text{nwifeinc} + 0.2529 * \text{educ} + 0.2057 * \text{exper} - 0.0039 * \text{expersq} - 0.9175 * \text{kidslt6} + 0.2226 * \text{kidsge6},$$

$\text{inlf} \sim \text{Ber}(p)$.

Thuật toán Random Walk Metropolis-Hastings.
Likelihood distribution

$$f(y|\beta) = \prod_{i=1}^{n} \left(\frac{exp(x_i^T \beta)}{1 + exp(x_i^T \beta)} \right)^{y_i} \left(\frac{1}{1 + exp(x_i^T \beta)} \right)^{1-y_i}$$
$$= exp \left\{ \sum_{i=1}^{n} y_i x_i^T \beta \right\} / \prod_{i=1}^{n} \left[1 + exp \left(x_i^T \beta \right) \right].$$

Posterior distribution

$$\pi(\beta|y) \propto f(y|\beta).\pi(\beta)$$

Trong đó, prior distribution $\pi(\beta) = 1$ là plat prior. Vì ở đây, chúng ta xem như mình không có thông tin về patameter beta.

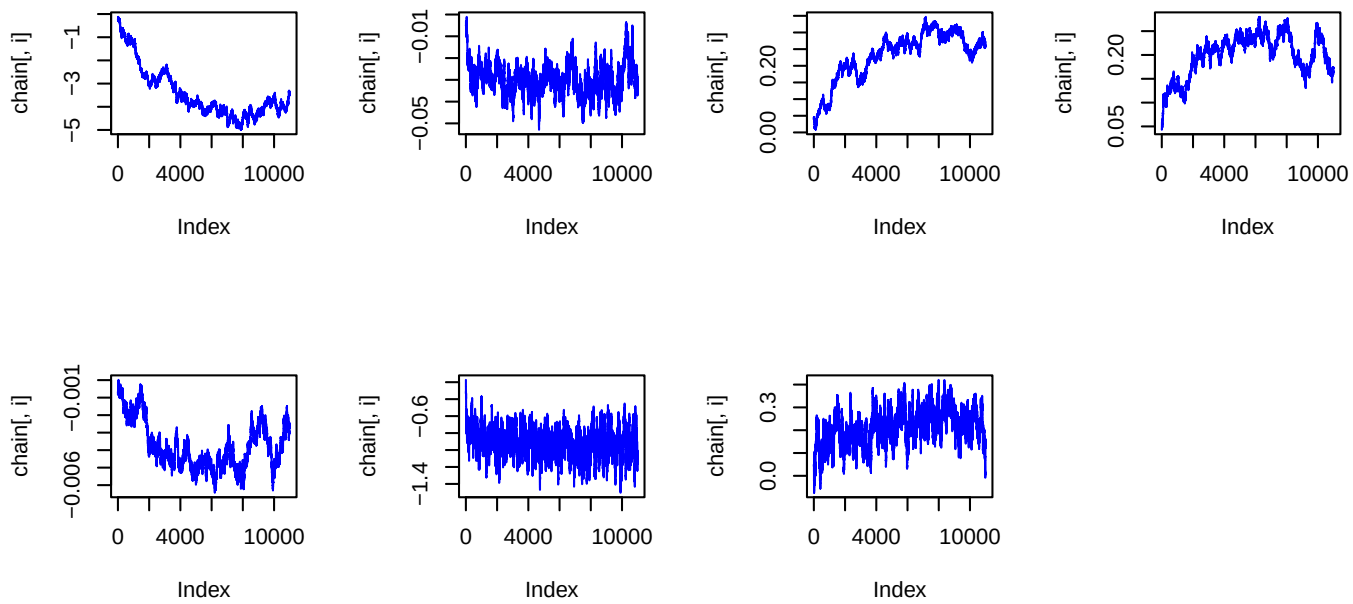
Thuật toán Random Walk Metropolis-Hastings.
Step 1. Initialise β_0: $\left(X^T X\right)^{-1}\left(X^T Y\right))$
Initialise *sig*: $diag(7)/X^T X$
Step2. For t=1 to T do
Step 3. Given the current parameter β^{cur}, propose β^{prop} from

$$\beta^{prop} \sim MultiNormal(\beta_{cur}, sig)$$

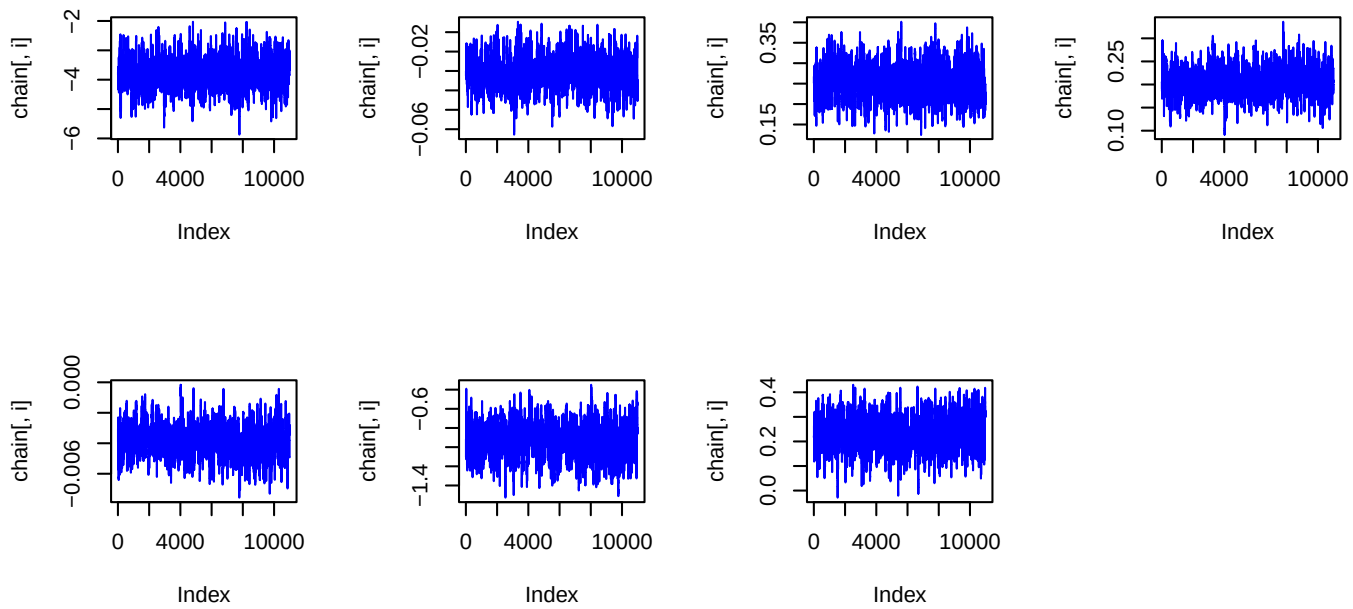
Step 4. Calculate acceptance probability

$$\alpha(\beta^{cur} \to \beta^{prop}) = min\left(1, \frac{\pi(\beta^{prop})}{\pi(\beta^{cur})}\right).$$

Step 5. Set $beta^{t+1} = \beta^{prop}$ with probability $\alpha(\beta^{cur} \to \beta^{prop})$, otherwise set $\beta^{t+1} = \beta^{cur}$.
Step 6. end for



Hình 6: Trace plot for MCMC out put with plat prior



Hình 7: Trace plot for MCMC out put with plat prior after change initial values

Kết quả ước lượng của tham số beta theo Maximum Likelihood estimation and ước lượng theo MCMC

	0	1	2	3	4	5	6
$\hat{\beta}_{\text{MLE}}$	-3.7397	-0.0301	0.2520	0.2057	-0.0039	-0.9175	0.2226
$\hat{\beta}_{\text{MSE}}$	-3.7819	-0.0308	0.2561	0.2050	-0.0038	-0.9371	0.2317

Kết quả ước lượng:

	0	1	2	3	4	5	6
$\hat{\beta}_{\text{MLE}}$	-3.7397	-0.0301	0.2520	0.2057	-0.0039	-0.9175	0.2226
$\hat{\beta}_{\text{MSE}}$	-3.2998	-0.0293	0.2195	0.1998	-0.0037	-0.9003	0.2096

Mô hình đạt được

$$\log\left(\frac{p}{1-p}\right) = -3.2998 - 0.0293 * \text{nwifeinc} + 0.2195 * \text{educ} + 0.1998 * \text{exper} - 0.0037 * \text{expersq} - 0.9003 * \text{kidslt6} + 0.2096 * \text{kidsge6},$$

$\text{inlf} \sim \text{Ber}(p)$.

Một người mẹ A có hơn 1 đứa con dưới sáu tuổi (tuổi đi học) so với người mẹ B thì tỷ số chênh (odds) của khả năng tham gia thị trường lao động của bà mẹ A và bà mẹ B là $e^{-0.9003} = 0.41$. Nói cách khác, khả năng tham thị trường lao động của người mẹ A chỉ bằng 0.41 lần khả năng tham gia lao động của người mẹ B, tức là giảm đi 59%.
Chú ý rằng khả năng ở đây không phải là xác suất p mà là $p/(1-p)$.

+ Hệ số hồi quy của biến nwifeinc là -0.0293, nghĩa là thu nhập của gia đình người mẹ A (sau khi đã trừ đi thu nhập của của người mẹ này) cao hơn 1 triệu so với thu nhập gia đình người mẹ B (sau khi đã trừ đi thu nhập của người mẹ này) thì tỷ số chênh (odds) về khả năng tham gia vào thị trường lao động của bà mẹ A và bà mẹ B là $e^{-0.0293} = 0.9711$. Hay một cách hiểu khác, khả năng tham gia thị trường lao động của người mẹ A chỉ bằng 0.9711 lần khả năng tham gia thị trường lao động của người mẹ B, Hay khả năng tham gia thị trường lao động của người mẹ A giảm 2.98% so với người mẹ B.
+ Hệ số hồi quy của biến educ là 0.2195, nghĩa là một bà mẹ A có học thức cao hơn 1 lớp so với bà mẹ B, thì tỷ số chênh về khả năng tham gia thị trường lao động của bà mẹ A và bà mẹ B là $e^{0.2195} = 1.2455$. Nghĩa là, khi một người mẹ có trình độ học cao hơn 1 bậc so với người mẹ khác thì khả năng tham gia thị trường lao động của người mẹ này tăng 24.55%.

+ Hệ số hồi quy của biến kidsge6 là 0.2096. Hàm ý rằng, một người mẹ A có hơn 1 đứa con trên 6 tuổi so với người mẹ B thì hệ số chênh về khả năng tham gia thị trường lao động của người mẹ A so với người mẹ B là $e^{0.2096} = 1.2332$, hay nói cách khác khả năng tham gia thị trường lao động của người mẹ A cao hơn $23,32\%$ so với người mẹ B.

Tài liệu tham khảo

- Mroz: https://rdrr.io/cran/wooldridge/man/mroz.html
- Package: wooldridge
- Art B. Owen. Monte Carlo theory, methods and examples. 2013 `https://artowen.su.domains/mc/`
- Andrew Gelman. Bayesian Data Analysis.
- Phương pháp Bayes: Nguyễn Văn Tuấn. Phân tích dữ liệu với R, tr. 369.
- Hướng dẫn sử dụng WINBUGS: Nguyễn Văn Tuấn. Phân tích dữ liệu với R, tr. 393.

TRÂN TRỌNG CẢM ƠN QUÝ VỊ ĐÃ LẮNG NGHE!