Bayesian statistics

Tùy chỉnh MCMC & Prior

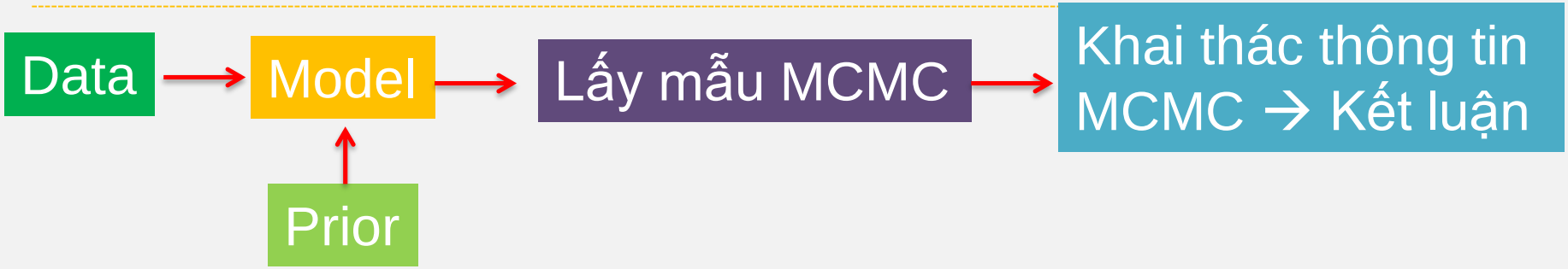
Ví dụ với mô hình log-binomial

Khương Quỳnh Long
Hà Nội, 08/2019
https://gitlab.com/LongKhuong/adhere-bayesian-statistics

Nội dung

1. Tùy chỉnh thông số MCMC
2. Các loại Prior và vai trò
3. Cách xác định Prior

Chuỗi MCMC



- Chuỗi MCMC chứa đựng toàn bộ thông tin của mô hình (bao gồm cả data (likelihood) và prior)
- Mỗi chuỗi MCMC thực chất là một dãy số (dạng time series)
- Cỡ mẫu của chuỗi MCMC phải đủ lớn để "ổn định" và đại diện ("hội tụ") cho tham số quan tâm (pp hậu nghiệm)
- Cỡ mẫu của MCMC khác cỡ mẫu của dữ liệu

Chuỗi MCMC

```
bayes, mcmcsize(#) burnin(#) thinning(#): [model]
```

- mcmcsize: số lượng mẫu MCMC cần lấy
- burnin: số lượng mẫu MCMC bỏ đi trong giai đoạn đầu
- thinning: giảm autocorrelation, thinning = 5: cứ 5 mẫu MCMC thì giữ lại 1 mẫu.
- Số iteration cần chạy = mcmcsize*thinning + burnin
- Mỗi phần mềm có quy ước tùy chỉnh khác nhau

Cỡ mẫu MCMC bao nhiêu là đủ?

- Không có con số cố định
- Thiếu thì mô hình không ổn định, quá dư thì tốn thời gian và công suất máy tính
- Tùy vào thuật toán mà số mẫu MCMC sẽ khác nhau
- Nên dựa vào Effect sample size (ESS) để xác định
- Tác giả John K. Kuschke đề nghị ESS > 10,000[1]
- Tác giả Andrew Gelman và Stan developer team đề nghị ít hơn (ESS > 2000 – 3000)[2]
- Cần chẩn đoán MCMC trước khi khai thác thông tin (đã đề cập ở bài trước)

1. John Kruschke (2014). Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan 2nd Edition. Elsevier Inc
2. Gelman A, Carlin JB, Stern HS, Rubin DB (2013). Bayesian Data Analysis. 3rd edition. Chapman & Hall/CRC

PRIOR

Prior

$$P(\theta|data) = \frac{P(data|\theta) * P(\theta)}{P(data)} = \frac{likelihood * Prior}{P(data)}$$

- Trong Bayesian, phân phối hậu định θ phụ thuộc vào Data và Prior
- Prior thể hiện ý kiến "chủ quan" về phân phối hậu định của θ quan tâm trước khi quan sát số liệu
- Có 2 kịch bản:
  - Không có thông tin gì về pp hậu định của θ trước đó
  - Có thông tin liên quan tới pp hậu định của θ trước đó

Kịch bản 1

- Xảy ra khi:
  - Không có bất kì thông tin gì về θ trước khi có số liệu
  - Không muốn xây dựng Prior
- Có 2 dạng:
  - “Non-informative prior”: uniform distribution hoặc flat
  - “Weakly-informative prior”: phân phối với thông số phân tán (scale) rất rộng → không mang hoặc mang rất ít thông tin
- Mục đích là không để Prior có ảnh hưởng tới posterior → *“let the data speak for themselves”*

Kịch bản 1

- Còn được gọi là “objective Bayesian statistics” vì dựa hoàn toàn vào dữ liệu
- Vẫn có được những ưu điểm của Bayesian trong diễn giải kết quả
- Cách tiếp cận này mất một phần thông tin quan trọng trong thống kê Bayes
- Đang được “lạm dụng”

Kịch bản 2

- Xảy ra khi có những thông tin trước đó về θ
  - Ý kiến chuyên gia, kinh nghiệm cá nhân…
  - Từ y văn (thường từ những systematic review)
- Được gọi là “informative prior” vì chứa đựng nhiều thông tin ảnh hưởng tới posterior
- Có 2 trường hợp:
  - “tương tự” với dữ liệu → mức độ chắc chắn được củng cố (95%CI nhỏ)
  - “Không tương tự” với dữ liệu → các ước lượng sẽ khác so với dữ liệu và prior

Kịch bản 2

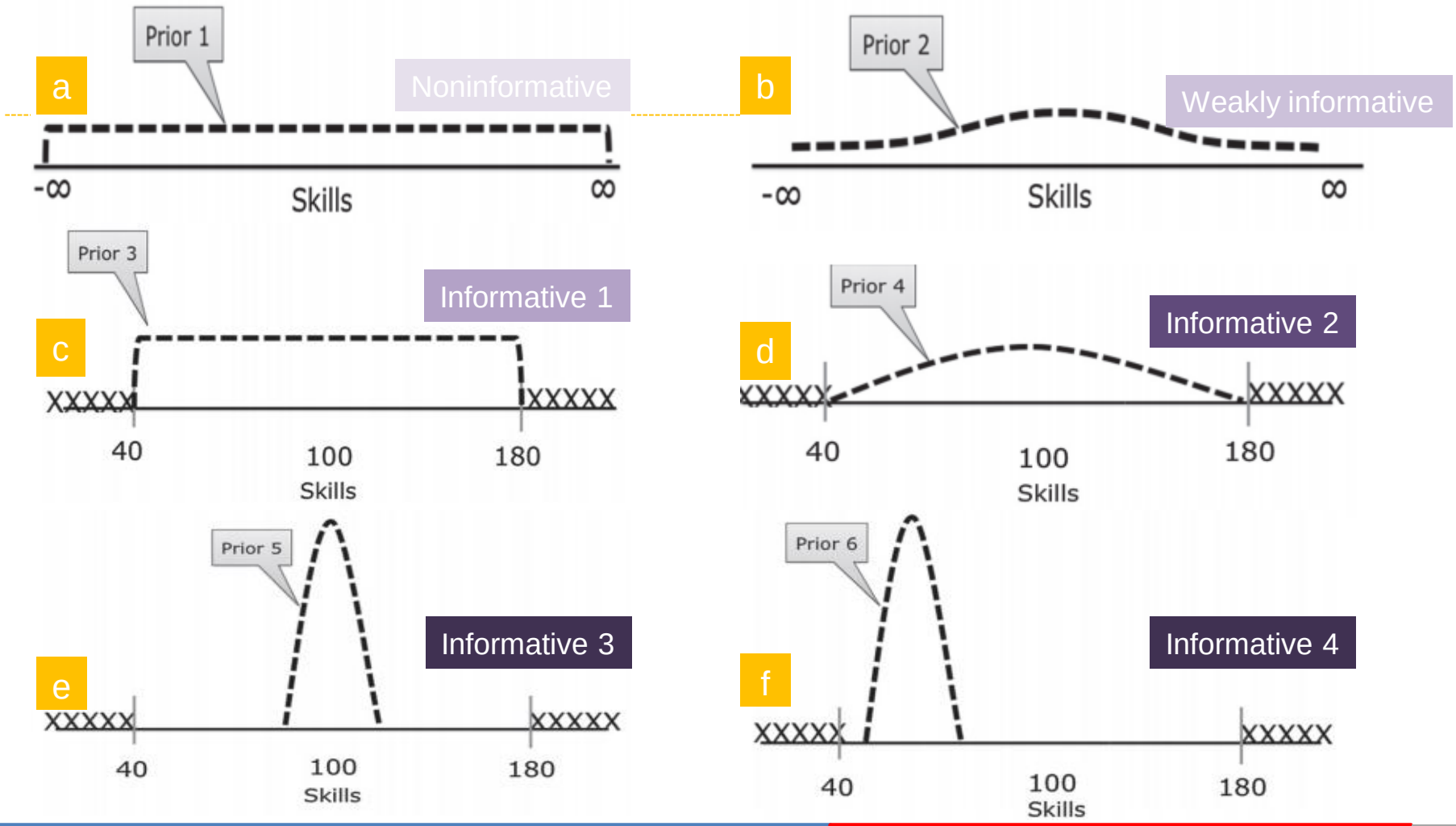
- Tận dụng được toàn bộ ưu điểm của Bayesian
- Đặc biệt quan trọng khi cỡ mẫu nhỏ
- Buộc nhà nghiên cứu phải “suy nghĩ” kỹ hơn về mỗi phép phân tích
- Informative prior là tâm điểm của sự tranh luận giữa Frequentist và Bayesian
- Đang được nhiều tác giả khuyến khích sử dụng, thay vì non-informative prior

Ví dụ về các loại Prior
và ảnh hưởng của Prior lên Posterior

Ví dụ về đánh giá điểm số kỹ năng đọc của trẻ mẫu giáo(1)

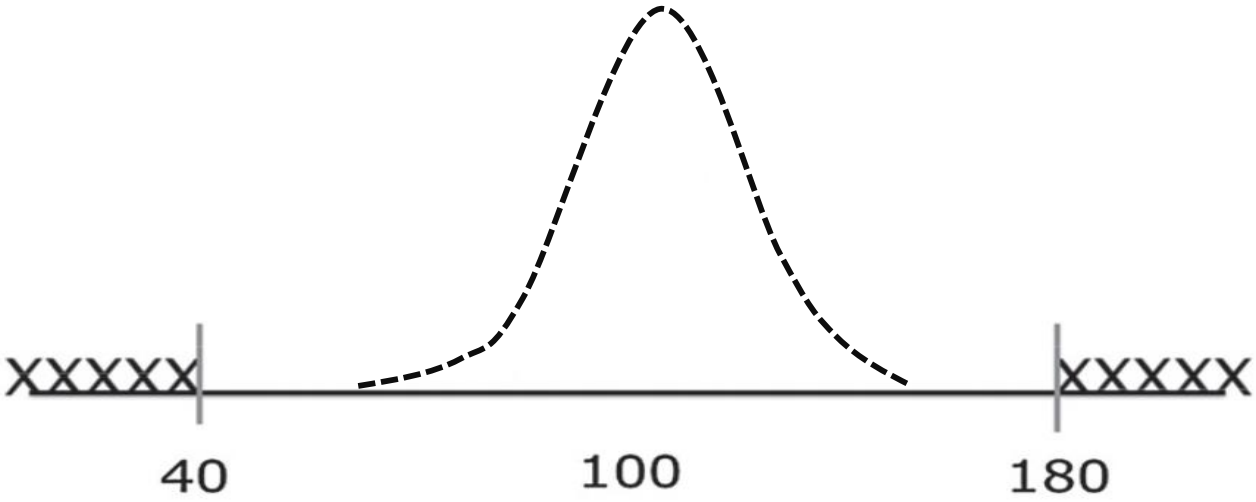
- Đánh giá 2 chỉ số: trung bình và phương sai của điểm số kỹ năng đọc
- Prior phản ánh mức độ hiểu biết về điểm số kỹ năng đọc trước khi quan sát dữ liệu
- 6 Prior khác nhau (ở 6 mức độ thông tin) được sử dụng để mô tả ảnh hưởng của Prior lên Posterior của điểm số kỹ năng đọc

1. Schoot, R. , Kaplan, D. , Denissen, J. , Asendorpf, J. B., Neyer, F. J. and Aken, M. A. (2014), A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research. Child Dev, 85: 842-860


a
Prior 1
Noninformative
-∞
Skills
∞
b
Prior 2
Weakly informative
-∞
Skills
∞
c
Prior 3
Informative 1
XXXXX
XXXXX
40
100
180
Skills
d
Prior 4
Informative 2
XXXXX
XXXXX
40
100
180
Skills
e
Prior 5
Informative 3
XXXXX
XXXXX
40
100
180
Skills
f
Prior 6
Informative 4
XXXXX
XXXXX
40
100
180
Skills

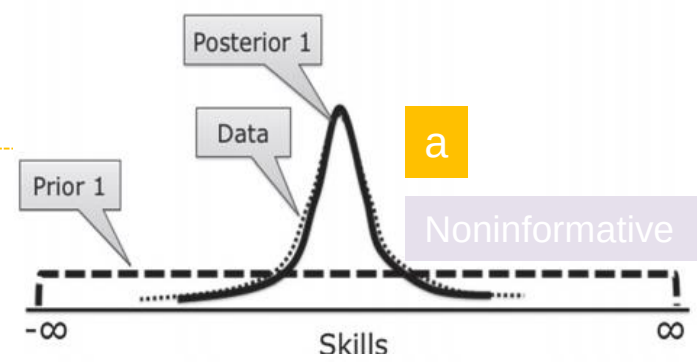
Data

- Số liệu thu thập trên 20 trẻ mẫu giáo, điểm số trung bình của kỹ năng đọc là 102



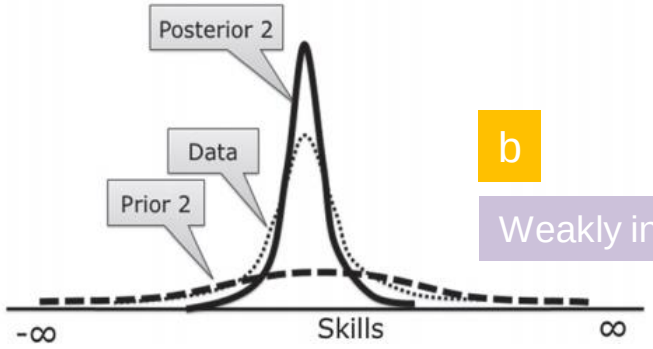
Posterior

- Mặc dù cùng 1 dữ liệu nhưng đạt được 6 Posterior tương ứng với 6 Prior



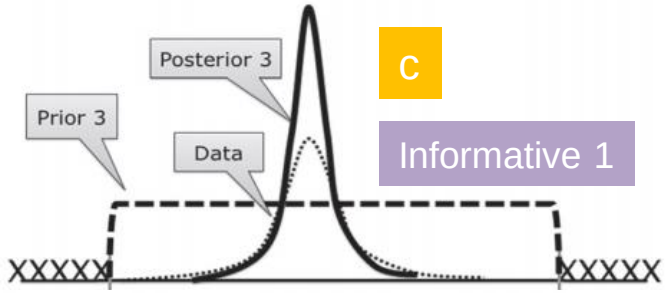
a

Noninformative



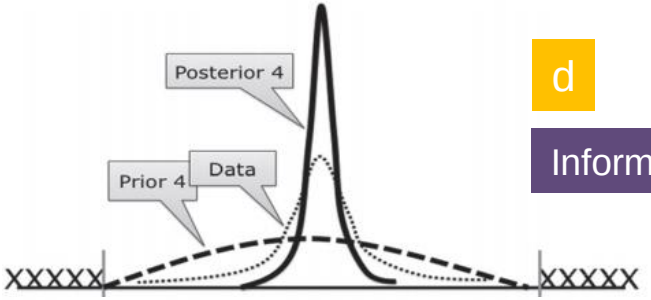
b

Weakly informative



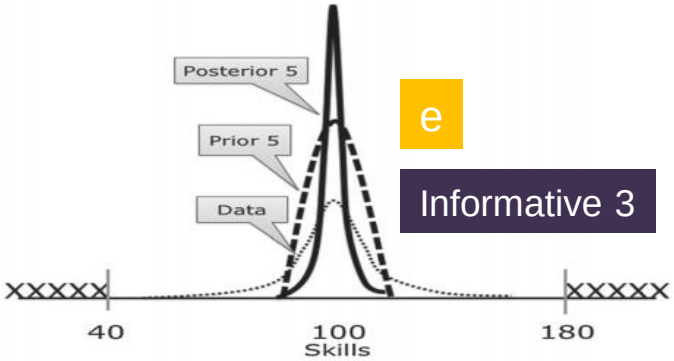
c

Informative 1



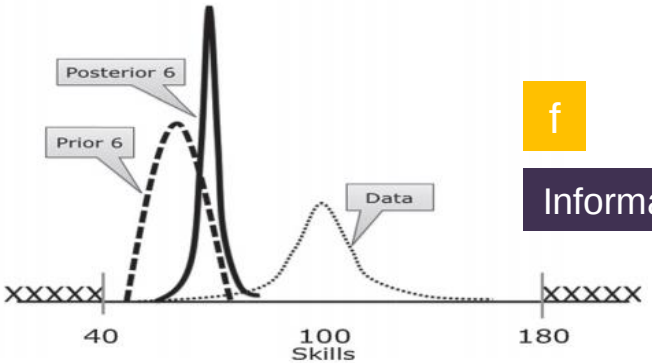
d

Informative 2



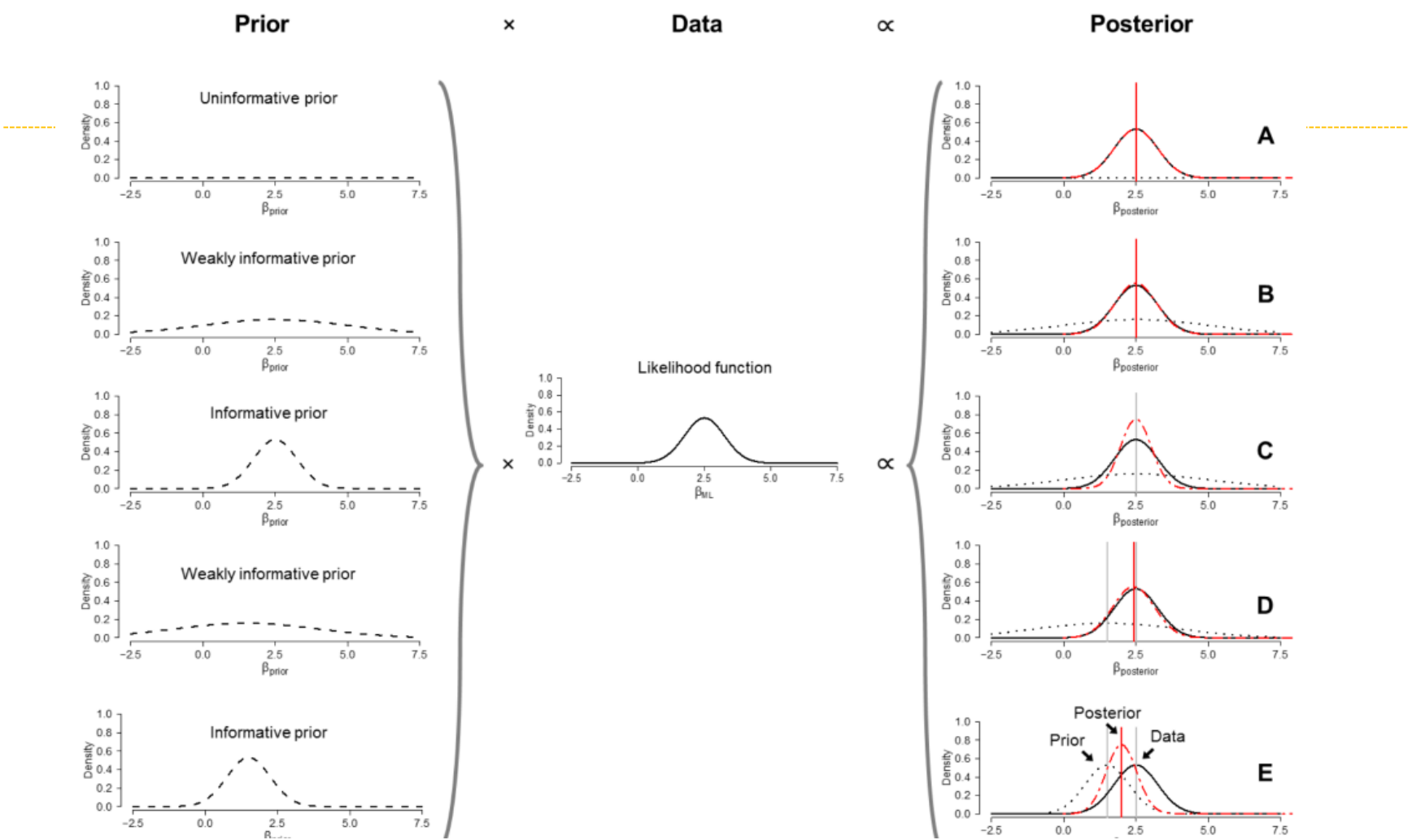
e

Informative 3



f

Informative 4


Prior
×
Data
∝
Posterior
Uninformative prior
Weakly informative prior
Informative prior
Weakly informative prior
Informative prior
Likelihood function
A
B
C
D
E
Posterior
Prior
Data
Density
β_{prior}
β_{ML}
$\beta_{posterior}$

Một số lưu ý

- Mỗi tham số trong mô hình đều phải có Prior cụ thể
- Trong thực hành, thường sử dụng Informative prior cho tham số Location. Prior cho scale, hay shape thường là weakly informative
- Mức độ “informative” không cố định cho bất cứ mô hình nào. Một Prior có thể là non-informative trong mô hình này nhưng có thể là informative trong mô hình khác
- Không được lấy kết quả của dữ liệu mới thu thập làm Prior cho chính nó

Ví dụ: Prior mặc định trong stata có thể là Informative prior

- Prior mặc định của stata:
  - Tham số beta (coefficients) ~ N(0, 10000)
  - Tham số scale ~ IG(0.01, 0.01)
- Đối với mô hình bodyfat ~ weight. Những Prior này là weakly informative

```
. sum bodyfat
```

Variable	Obs	Mean	Std. Dev.	Min	Max
bodyfat	251	18.88765	7.724121	0	45.1

Ví dụ: Prior mặc định trong stata có thể là Informative prior (tt)

- Sử dụng data “birthweight.dta”

```
. des

Contains data from C:\Users\QUYNH LONG\Desktop\Bayesian_HaNoi\data\birthweight.dta
  obs:           641
 vars:             8                          3 Oct 2018 13:04
 size:        20,512
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
maso            float   %9.0g                 ma so
tuoime          float   %9.0g                 tuoi me (nam)
tang_ha         float   %9.0g                 tang huyet ap thai ki - 1=tang ha, 0=khong tang ha
tuoithai        float   %9.0g                 tuoi thai (tuan)
gioi            float   %9.0g                 gioi tinh tre - 1=trai, 0=gai
tlsosinh        float   %9.0g                 trong luong so sinh (gram)
nghenghiep      float   %9.0g                 nghe nghiep me - 1=tu do, 2=cong nhan, 3=vien chuc
nhecan          float   %9.0g                 1= co, 0 = khong
-------------------------------------------------------------------------------
Sorted by: maso
```

```
Priors:
  {tlsosinh:tuoithai _cons} ~ normal(0,10000)
                {sigma2} ~ igamma(.01,.01)

(1) Parameters are elements of the linear form xb_tlsosin

Bayesian linear regression                    MCMC ite
Random-walk Metropolis-Hastings sampling      Burn-in
                                              MCMC sam
                                              Number o
                                              Acceptan
                                              Efficien

Log marginal likelihood = -4933.3058          max =      .1853
```

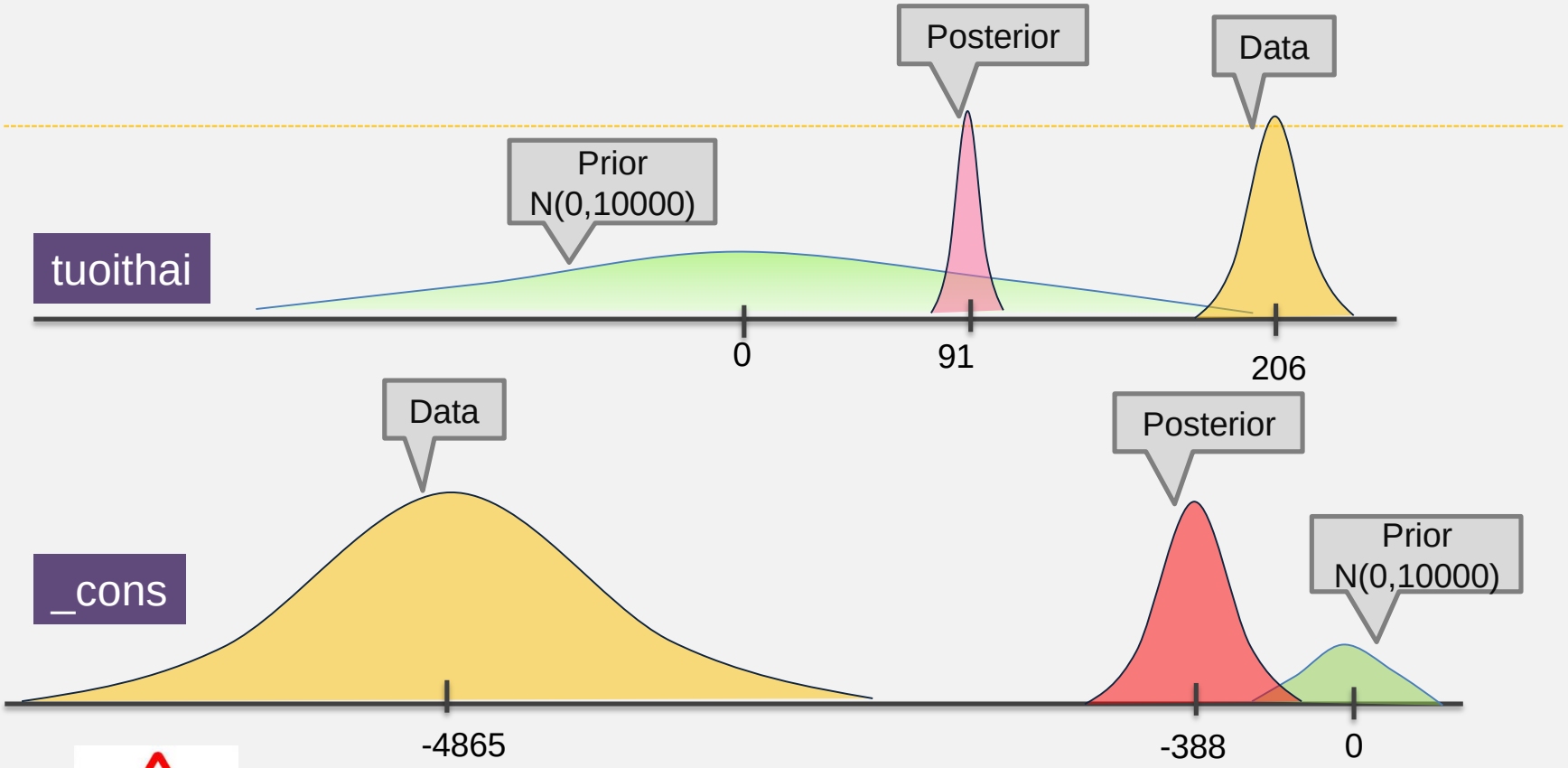
	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred. Interval]	
tlsosinh						
tuoithai	**91.34506**	**2.595914**	**.081774**	**91.33574**	**86.029**	**96.56295**
_cons	**-388.4403**	**99.08601**	**3.2015**	**-389.9419**	**-581.3275**	**-191.6009**
sigma2	**267698.9**	**15529.05**	**360.727**	**267259.2**	**237773.5**	**299820.7**

```
. reg tlsosinh tuoithai
```

Source	SS	df	MS		
Model	**148354317**	**1**	**148354317**	Prob > F	**0.0000**
Residual	**124365805**	**639**	**194625.673**	R-squared	**0.5440**
				Adj R-squared	**0.5433**
Total	**272720122**	**640**	**426125.19**	Root MSE	**441.16**

Number of obs	=	**641**				
F(1, 639)	=	**762.25**				
Prob > F	=	**0.0000**				
R-squared	=	**0.5440**				
Adj R-squared	=	**0.5433**				
Root MSE	=	**441.16**				

tlsosinh	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tuoithai	**206.6412**	**7.484572**	**27.61**	**0.000**	**191.9439**	**221.3386**
_cons	**-4865.245**	**290.0814**	**-16.77**	**0.000**	**-5434.873**	**-4295.617**



Caution

Thận trọng khi để Prior mặc định trong Stata !!!

```
Priors:
  {tlsosinh:tuoithai _cons} ~ 1 (flat)                              (1)
                       {sigma2} ~ 1 (flat)
```

(1) Parameters are elements of the linear form xb_tlsosinh.

Bayesian linear regression — Random-walk Metropolis-Hastings sampling

MCMC iterations	12,500
Burn-in	2,500
MCMC sample size	10,000
Number of obs	641
Acceptance rate	.3112
Efficiency: min	.08968
avg	.137
max	.2308

Log marginal likelihood = **-4794.8644**

	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
tlsosinh						
tuoithai	**206.7728**	**7.712887**	**.256456**	**206.7262**	**191.7608**	**221.7494**
_cons	**-4869.27**	**298.9885**	**9.98411**	**-4873.096**	**-5456.104**	**-4289.579**
sigma2	**195986**	**11181.12**	**232.735**	**195531**	**174750.9**	**218986.1**

Sử dụng Informative Prior

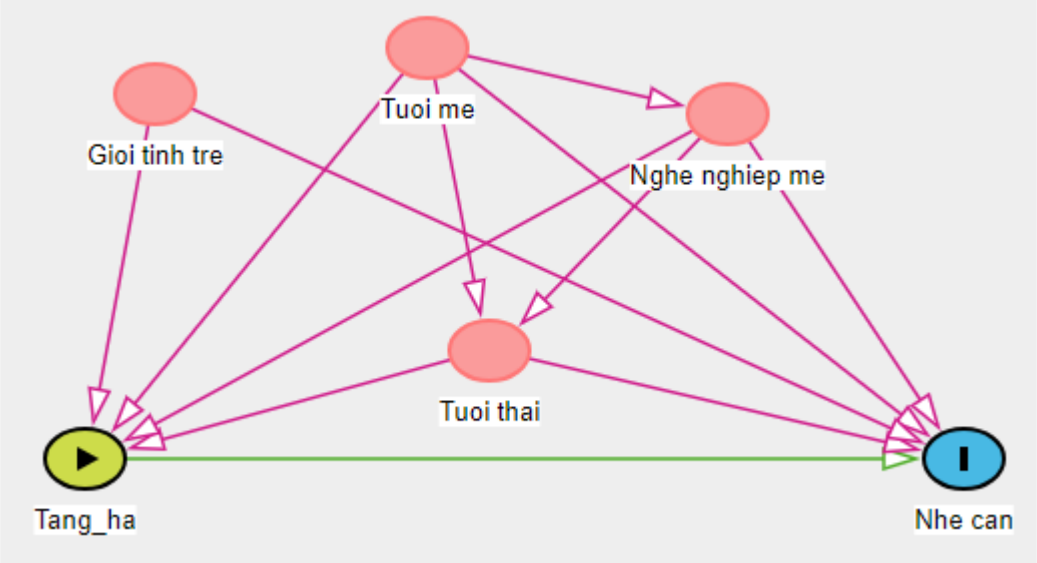
Cấu trúc câu lệnh Prior trong Stata

Bayes, prior({tham số}, phân phối tương ứng) : [model]

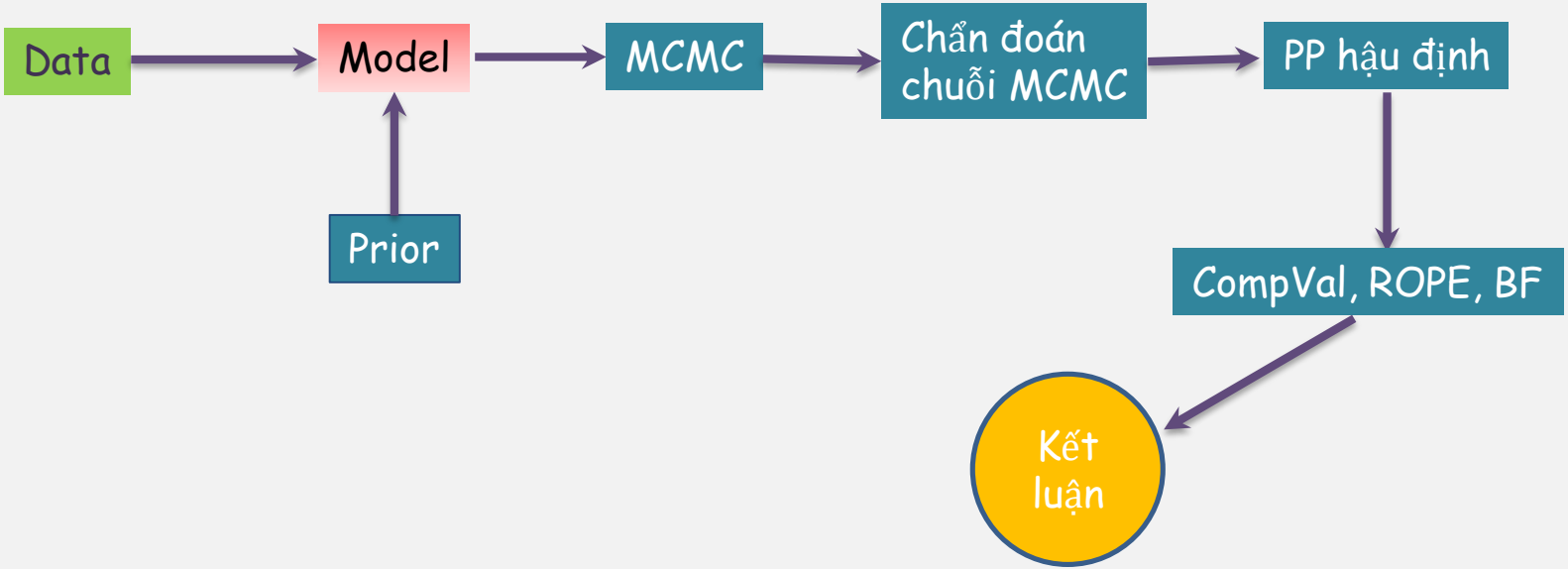
Dữ liệu birthweight

- Mục tiêu nghiên cứu: Xác định nguy cơ **nhẹ cân** của trẻ với tình trạng **tăng huyết áp thai kì** của mẹ

- Confounders:
  - Giới tính trẻ
  - Tuổi thai
  - Tuổi mẹ
  - Nghề nghiệp mẹ

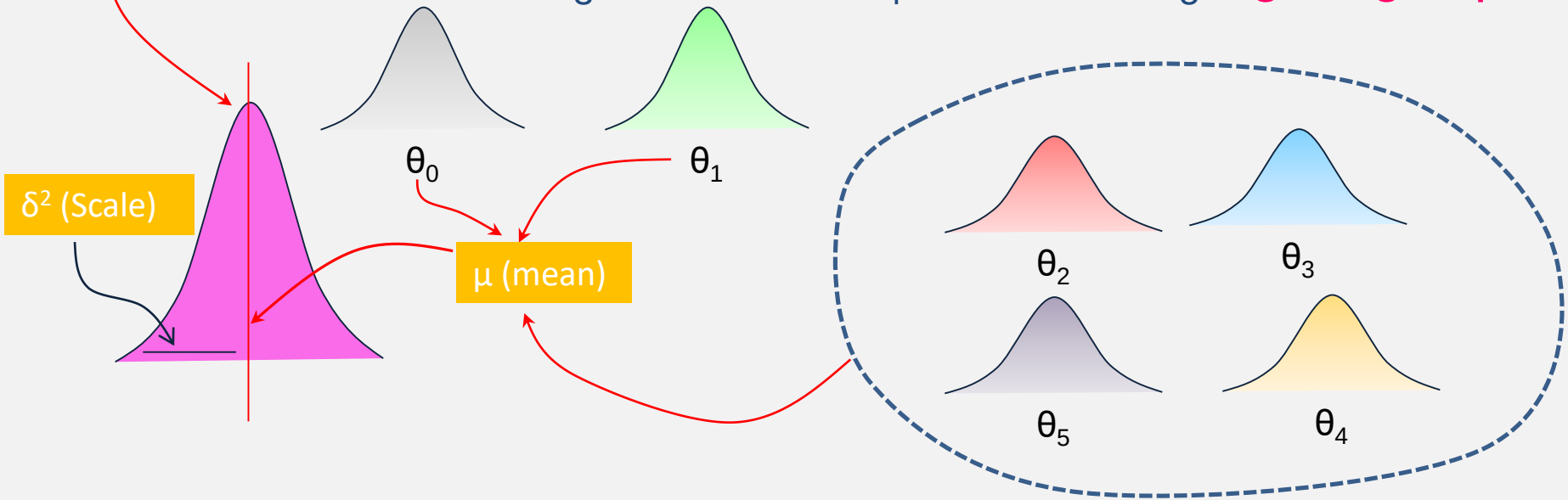


Quy trình


Data
Model
Prior
MCMC
Chẩn đoán chuỗi MCMC
PP hậu định
CompVal, ROPE, BF
Kết luận

- Sử dụng glm với phân phối Binomial và link log (hồi quy log_binomial)

$$\log(\text{risk}(\text{nhecan})) = \theta_0 + \theta_1*\text{tang\_ha} + \theta_2*\text{gioitinh} + \theta_3*\text{tuoithai} + \theta_4*\text{tuoime} + \theta_5*\text{nghenghiep}$$



Prior

- 2 kịch bản:

1. Weakly -informative prior: Sử dụng mặc định N(0,10000) (hàm ý upper 95%RR = e^{100}) cho tất cả các tham số
2. Informative prior: tham khảo y văn

✓ Nhiều nghiên cứu cho thấy tăng huyết áp thai kì là nguy cơ sinh con nhẹ cân

✓ Với RR trung bình khoảng ~ 2 và ngưỡng trên KTC 95% của nguy cơ nhẹ cân không vượt quá 5

➔ Kế hoạch:

✓ RR nhẹ cân do tăng huyết áp có tb là 2 và KTC 95% ~(0.4 – 5).

✓ Các tham số còn lại sử dụng weakly informative

Xây dựng informative prior

- Đối với tăng huyết áp thai kì
- RR = 2 (95%CI = 0.4 – 5) ➔ log(RR) = 0.693 (95% CI = -0.9162908 - 1.6094379)
- Chọn phân phối có location = 0.693 và bách phân vị 97.5^{th} = 1.6094379.
- [nhắc lại]: giá trị tại bách phân vị 97.5^{th} của pp chuẩn bằng 1.96*SD + mean, pp Cauchy là 12.706*scale + location, pp $t_{(7)}$ là 2.365*scale + location
- Có thể lựa chọn 1 trong các phân phối:
  - $N(0.693, 0.468^2)$
  - Cauchy(0.693, scale = 0.072)
  - $t_{(7)}(0.693, 0.387^2)$

Shiny apps

- Chuyển đổi ước lượng OR/RR/PR sang 3 loại phân phối hay dùng.

- https://khuongquynhlong.shinyapps.io/prior

Chuỗi MCMC

- Mô hình log_binomial thường khó hội tụ ➔ không sử dụng MLE dẫn đường (nomleinitial)
- Sử dụng chuỗi MCMC gồm:
  - Sample size: mcmcsize(50000)
  - Burnin 5000: burnin(5000)
  - Thinning 5: thinning(5)

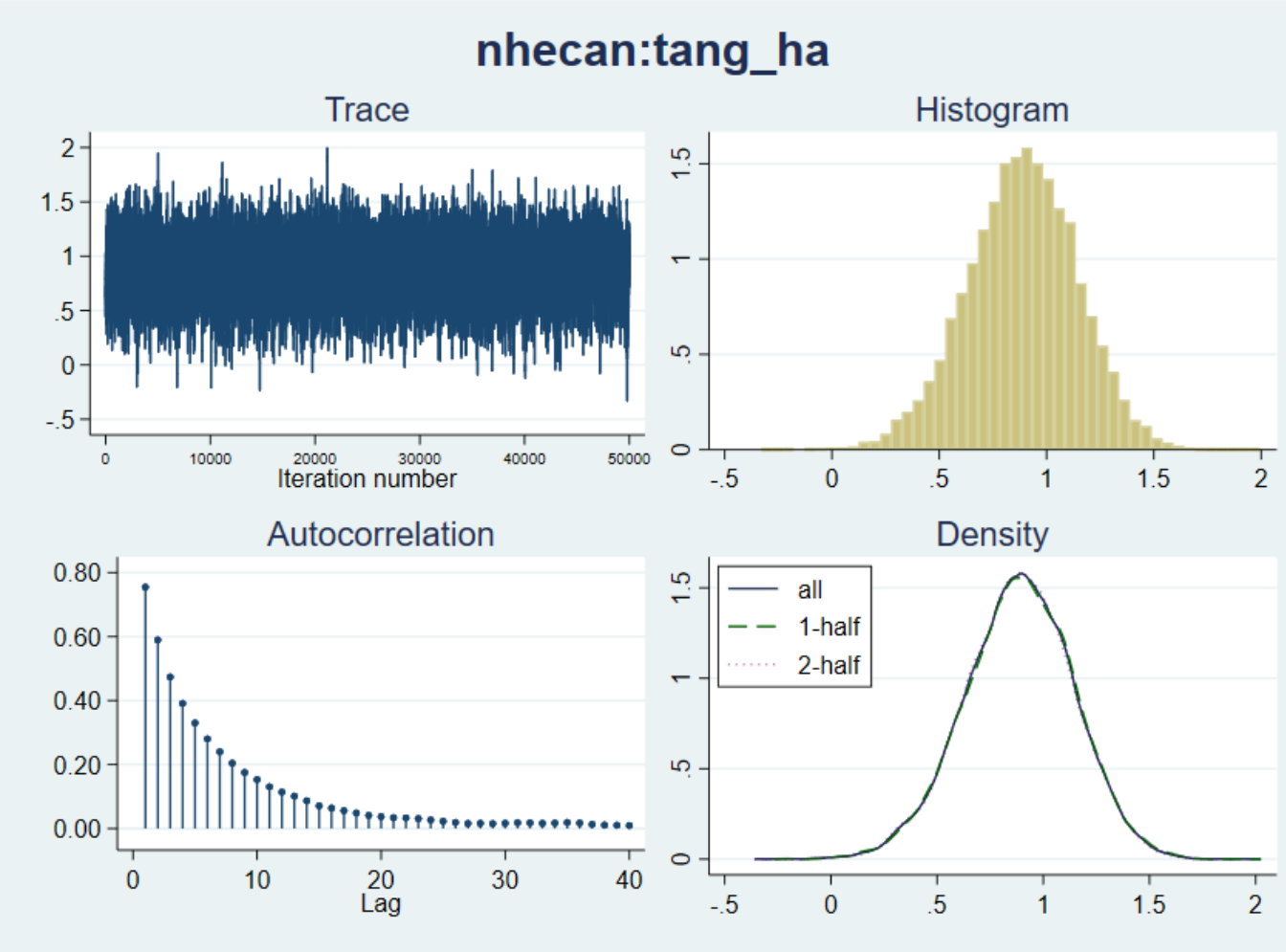
➔ cần 255000 iterations

Mô hình Weakly-informative Prior

```
set seed 1234
bayes, dots nomleinitial mcmcsize(50000) burnin(5000) thinning(5): glm
nhecan tuoime tang_ha tuoithai gioi i.nghenghiep, link(log)
family(binomial)
```

```
note: discarding every 4 sample observations; using observations 1,6,11,...

Burn-in ...
Simulation ...

Model summary
------------------------------------------------------------------------------
Likelihood:
  nhecan ~ glm(xb_nhecan)

Prior:
  {nhecan:tuoime tang_ha tuoithai gioi i.nghenghiep _cons} ~ normal(0,10000) (1)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_nhecan.

Bayesian generalized linear models                  MCMC iterations  =    254,996
Random-walk Metropolis-Hastings sampling            Burn-in          =      5,000
                                                    MCMC sample size =     50,000
Family : Bernoulli                                  Number of obs    =        641
Link   : log                                        Scale parameter  =          1
                                                    Acceptance rate  =      .1958
                                                    Efficiency:  min =     .06478
                                                                 avg =      .1217
Log marginal likelihood = -172.44566                             max =      .1938
```

```
. bayesstats ess

Efficiency summaries                MCMC sample size =     50,000

------------------------------------------------------------------------------
             nhecan |        ESS  Corr. time  Efficiency
--------------------+---------------------------------------------------------
             tuoime |    3239.13       15.44      0.0648
            tang_ha |    4796.62       10.42      0.0959
           tuoithai |    6940.33        7.20      0.1388
               gioi |    9692.42        5.16      0.1938
                    |
         nghenghiep |
                  2 |    5679.95        8.80      0.1136
                  3 |    4693.01       10.65      0.0939
                    |
              _cons |    7545.39        6.63      0.1509
------------------------------------------------------------------------------
```


nhecan:tang_ha
Trace
2
1.5
1
.5
0
-.5
0
10000
20000
30000
40000
50000
Iteration number
Histogram
1.5
1
.5
0
-.5
0
.5
1
1.5
2
Autocorrelation
0.80
0.60
0.40
0.20
0.00
0
10
20
30
40
Lag
Density
all
1-half
2-half
1.5
1
.5
0
-.5
0
.5
1
1.5
2

nhecan	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
tuoime	.0026086	.0275646	.000484	.0023706	-.0508516	.0575412
tang_ha	.8853569	.2565585	.003704	.8910112	.3580765	1.370628
tuoithai	-.5852022	.0532398	.000639	-.5836671	-.6931575	-.484523
gioi	-.2027442	.2142593	.002176	-.1968674	-.634766	.2026755
nghenghiep						
2	-.9680056	.3559847	.004723	-.9684439	-1.659925	-.2761828
3	-.917679	.3503081	.005114	-.9257425	-1.58538	-.2266529
_cons	20.28033	2.206054	.025397	20.2208	16.08815	24.70865

Note: Default priors are used for model parameters.

- Tăng huyết áp thai kì làm tăng nguy cơ sinh con nhẹ cân
- Median PP hậu định của RR = exp(0.8853569) = 2.44; 95% Credible interval = 1.43 – 3.94

Mô hình informative Prior

```
set seed 12345
bayes, dots nomleinitial mcmcsize(50000) burnin(5000) thinning(5)
prior({nhecan: tang_ha}, normal(0.693, 0.219)): glm nhecan tuoime
tang_ha tuoithai gioi i.nghenghiep, link(log) family(binomial)
```

Model summary

Likelihood:
  nhecan ~ glm(xb_nhecan)

Priors:
  {nhecan:tang_ha} ~ normal(0.693,0.468) (1)
  {nhecan:tuoime tuoithai gioi i.nghenghiep _cons} ~ normal(0,10000) (1)

(1) Parameters are elements of the linear form xb_nhecan.

Bayesian generalized linear models
Random-walk Metropolis-Hastings sampling

Family : **Bernoulli**
Link : **log**

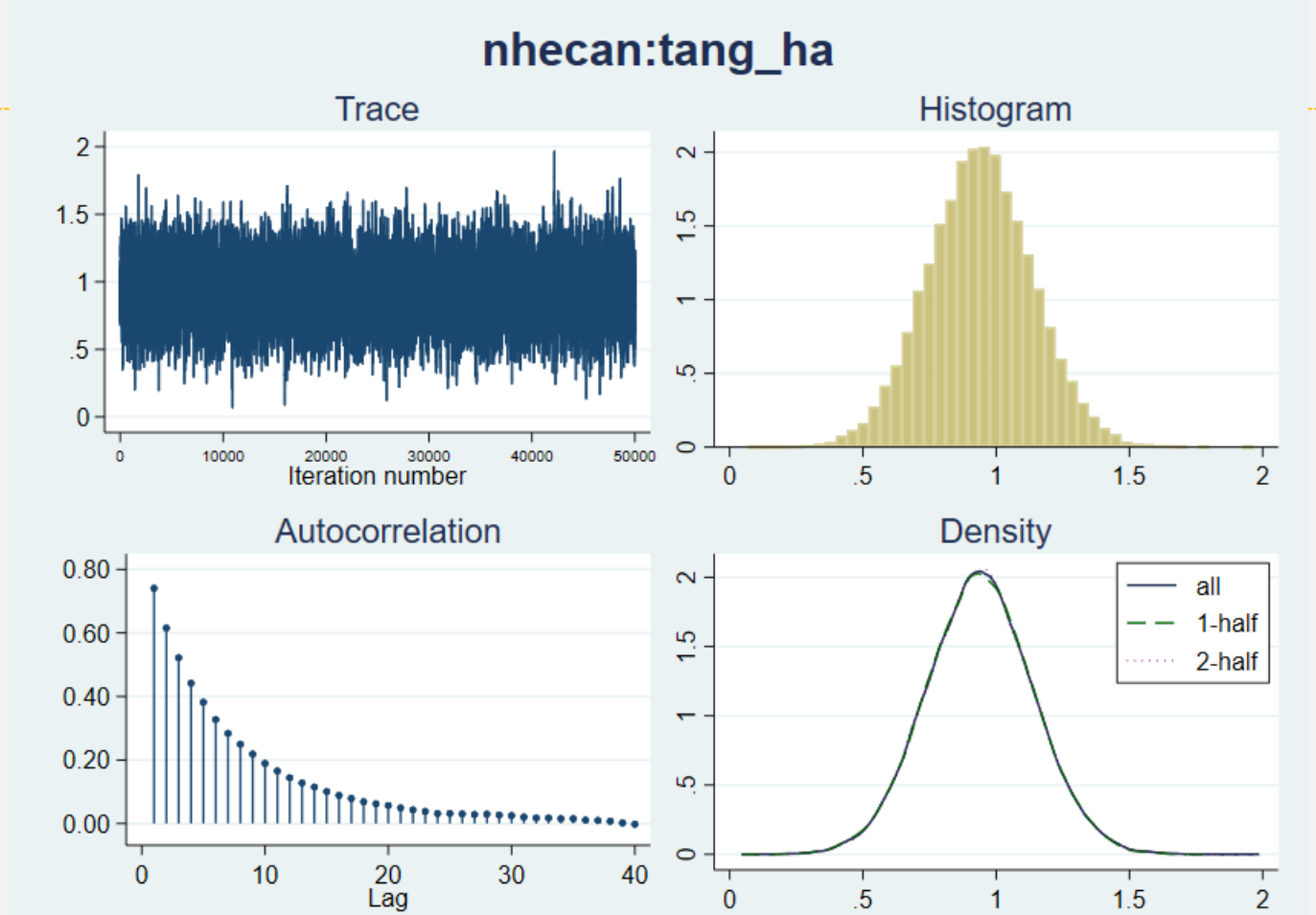
MCMC iterations	=	**254,996**
Burn-in	=	**5,000**
MCMC sample size	=	**50,000**
Number of obs	=	**641**
Scale parameter	=	**1**
Acceptance rate	=	**.3052**
Efficiency: min	=	**.02016**
avg	=	**.07852**
max	=	**.1088**

Log marginal likelihood = **-284.44385**

```
. bayesstats ess
```

Efficiency summaries MCMC sample size = **50,000**

nhecan	ESS	Corr. time	Efficiency
tuoime	**3635.34**	**13.75**	**0.0727**
tang_ha	**4235.13**	**11.81**	**0.0847**
tuoithai	**4045.25**	**12.36**	**0.0809**
gioi	**4209.89**	**11.88**	**0.0842**
nghenghiep			
2	**4908.46**	**10.19**	**0.0982**
3	**5440.64**	**9.19**	**0.1088**
_cons	**1007.94**	**49.61**	**0.0202**


nhecan:tang_ha
Trace
Histogram
Autocorrelation
Density
Iteration number
Lag
all
1-half
2-half

```
. bayesstats summary

Posterior summary statistics                     MCMC sample size =     50,000
```

nhecan	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred. Interval]	
tuoime	.156622	.0234482	.000389	.1567065	.1107574	.2022838
tang_ha	.9390539	.1993365	.003063	.9387452	.5507798	1.334615
tuoithai	-.0396384	.0213052	.000335	-.0391665	-.0825844	.0012426
gioi	-.5873098	.1855717	.00286	-.5851006	-.9631203	-.2243718
nghenghiep						
2	.3287302	.3401111	.004855	.3119734	-.2871539	1.050072
3	-.2877614	.364021	.004935	-.3041512	-.9544088	.4700508
_cons	-6.258227	.2373814	.007477	-6.252775	-6.745089	-5.810195

- Tăng huyết áp thai kì làm tăng nguy cơ sinh con nhẹ cân
- Median PP hậu định của RR = exp(0.9387452) = 2.56; 95% Credible interval = 1.73 – 3.79

Nhận xét

- Sử dụng Informative prior làm cho 95%CI hẹp hơn ➔ mức độ uncertainty giảm
- Cỡ mẫu càng lớn, ảnh hưởng của Prior càng ít lại ➔ Prior phát huy hết tác dụng khi cỡ mẫu nhỏ
- Mô hình log_binomial có thể sử dụng trong Bayesian khi phương pháp MLE không thể hội tụ

Kiểm định giả thuyết

- CompVal
- ROPE
- Bayes Factor

Thank you !