Bài 6
Thống kê Bayes tính toán và Lập trình xác suất

Thống kê máy tính và ứng dụng
(Computational Statistics and Applications)

Vũ Quốc Hoàng (vqhoang@fit.hcmus.edu.vn)
FIT - HCMUS

2025

Nội dung

Nội dung

Lập trình xác suất

Lập trình xác suất (Probabilistic Programming)

- `https://en.wikipedia.org/wiki/Probabilistic_programming`.
- PPL (Probabilistic programming language): Stan, PyMC, TensorFlow Probability (TFP), Pyro, Turing.jl, ...

PyMC

- `https://www.pymc.io/welcome.html`
- Cameron Davidson-Pilon. *Bayesian Methods for Hackers*. Addision-Wesley, 2016.
- Osvaldo Martin. *Bayesian Analysis with Python*. Packt Publishing, 2024.
- Osvaldo A. Martin, Ravin Kumar and Junpeng Lao. *Bayesian Modeling and Computation in Python*. CRC Press, 2022.

PyMC - Ví dụ 1

Bài toán. Một nhà máy sản xuất bóng với 4 loại kích thước 1, 2, 3, 4. Đặt 3 quả bóng loại kích thước 2. Nhận được 3 quả bóng với các kích thước: 1.77, 2.23, 2.70. Hỏi nhà máy có sản xuất đúng loại đã đặt?

Suy diễn. Ta mô hình kích thước bóng là biến ngẫu nhiên $X \sim \mathcal{N}(\mu, \sigma^2)$ với tham số μ có thể nhận một trong 4 giá trị

$$\begin{cases} E_1 : \mu = \mu_1 = 1.0, \\ E_2 : \mu = \mu_2 = 2.0, \\ E_3 : \mu = \mu_3 = 3.0, \\ E_4 : \mu = \mu_4 = 4.0. \end{cases}$$

Xem cách làm "thủ công" trong bài Bài trước.

PyMC - Ví dụ 1 (tt)

Dùng PyMC (hay các nền tảng lập trình xác suất khác) ta chỉ cần đặc tả mô hình còn nền tảng sẽ "tự động" suy diễn. Xem Jupyter Notebook đi kèm.

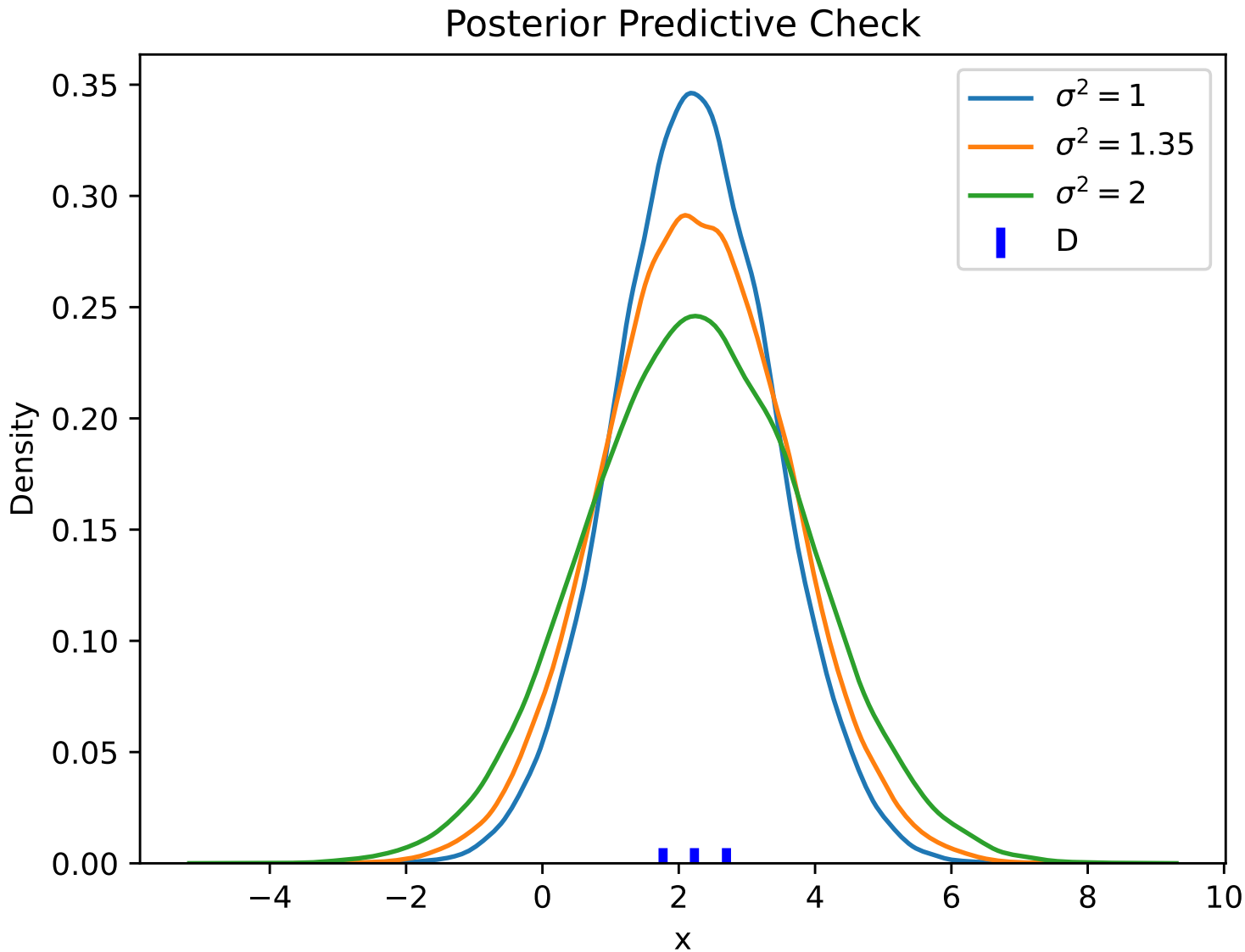
"Thử nghiệm" phân phối tiên nghiệm đều

$$p(E_i) = \frac{1}{4}, i = 1, ..., 4$$

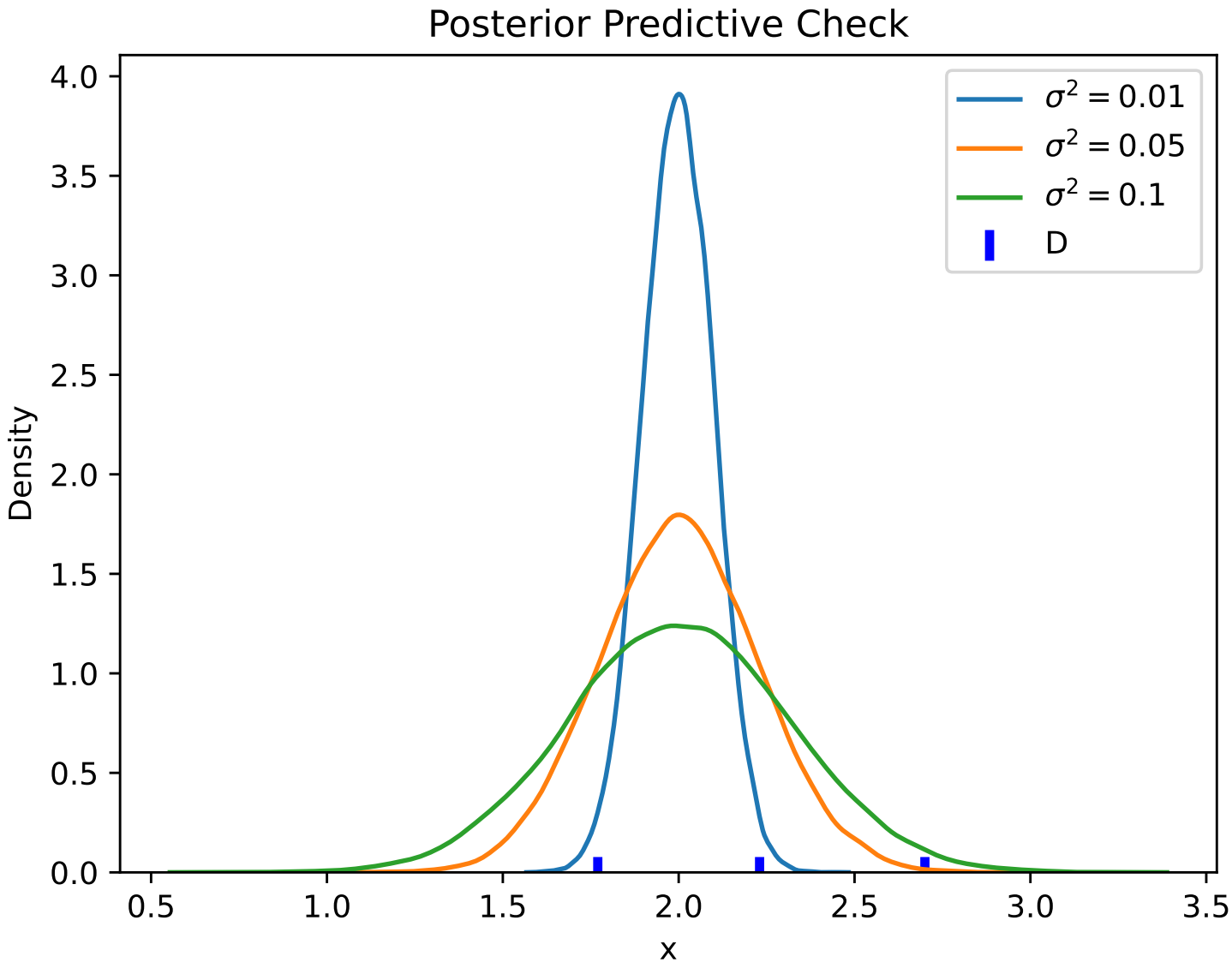
và một số giá trị của σ^2. "Chạy" PyMC và đối chiếu kết quả lấy mẫu (sampling) tự động với kết quả lý thuyết đã tính

	$\sigma^2 = 1$	$\sigma^2 = 1.35$	$\sigma^2 = 2$
$P(\mu = 1\|D)$	7%	11%	16%
$P(\mu = 2\|D)$	64%	56%	47%
$P(\mu = 3\|D)$	29%	31%	32%
$P(\mu = 4\|D)$	1%	2%	5%

PyMC - Ví dụ 1 (tt)



PyMC - Ví dụ 1 (tt)



PyMC - Ví dụ 2

Bài toán. Cho dữ liệu của dãy phép thử Bernoulli $D = \{Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n\}$, "tìm" p.

Suy diễn. Xem tham số $\theta = p$ là biến ngẫu nhiên liên tục, nhận giá trị trong $[0, 1]$. Dùng phân phối tiên nghiệm liên hợp cho θ

$$\theta \sim Beta(a, b)$$

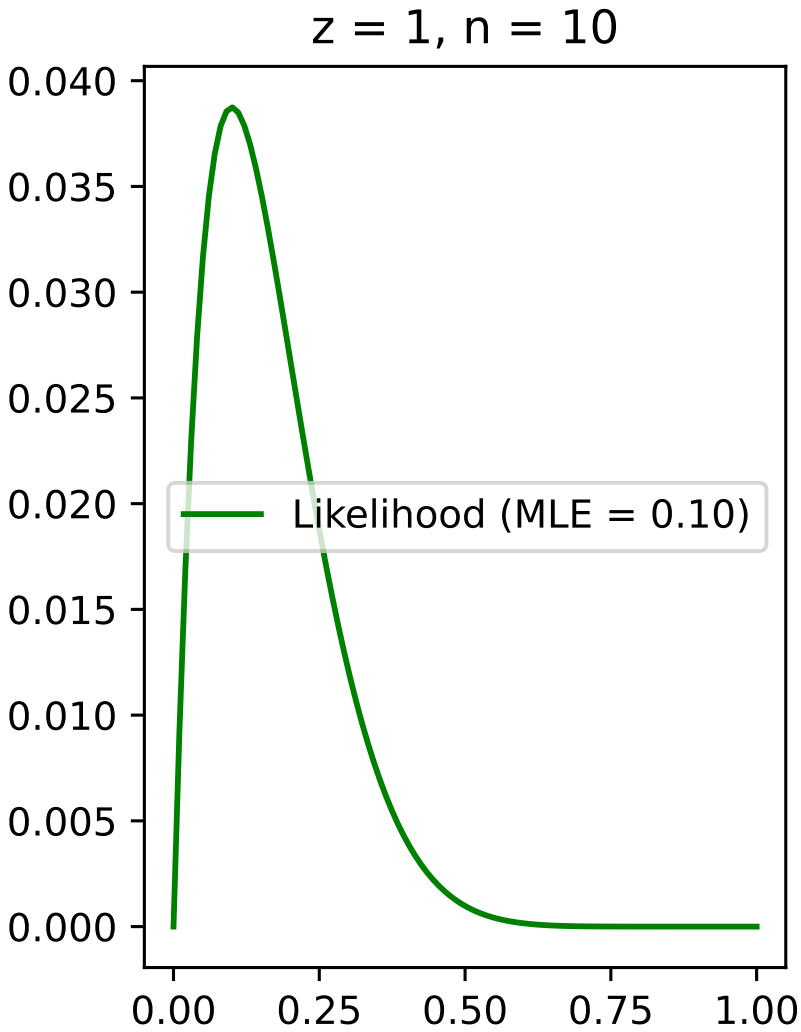
thì phân phối hậu nghiệm là

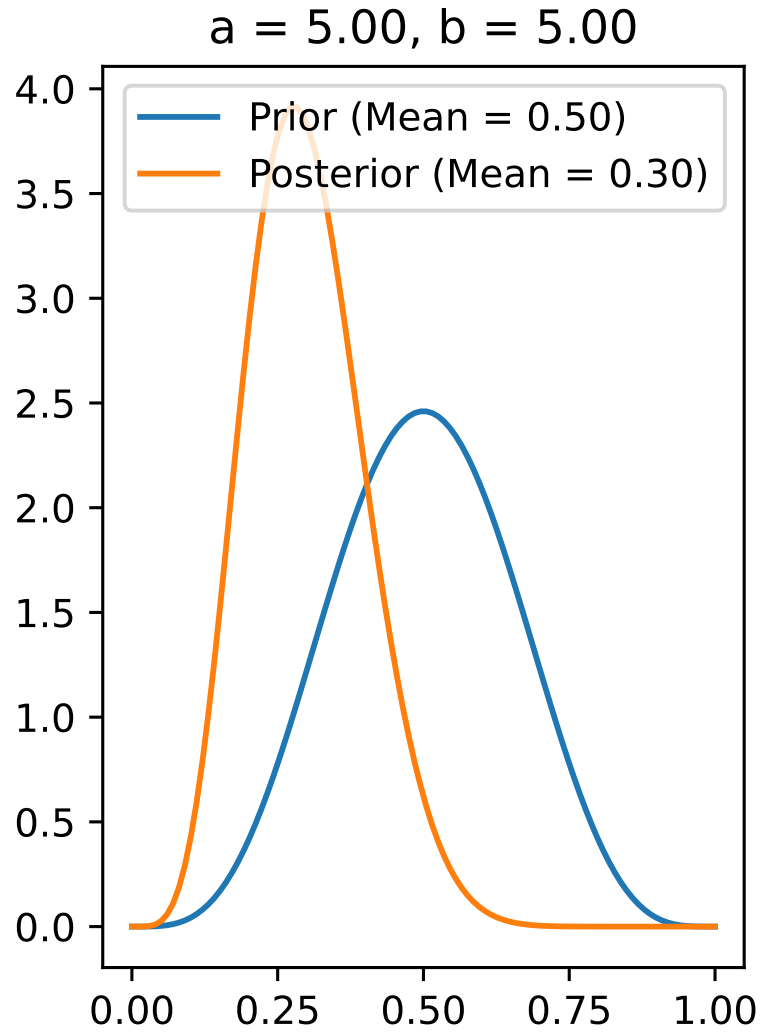
$$(\theta|D) \sim \text{Beta}(a + z, b + n - z)$$

với $z = \sum_{i=1}^{n} y_i$ là số lần thành công.

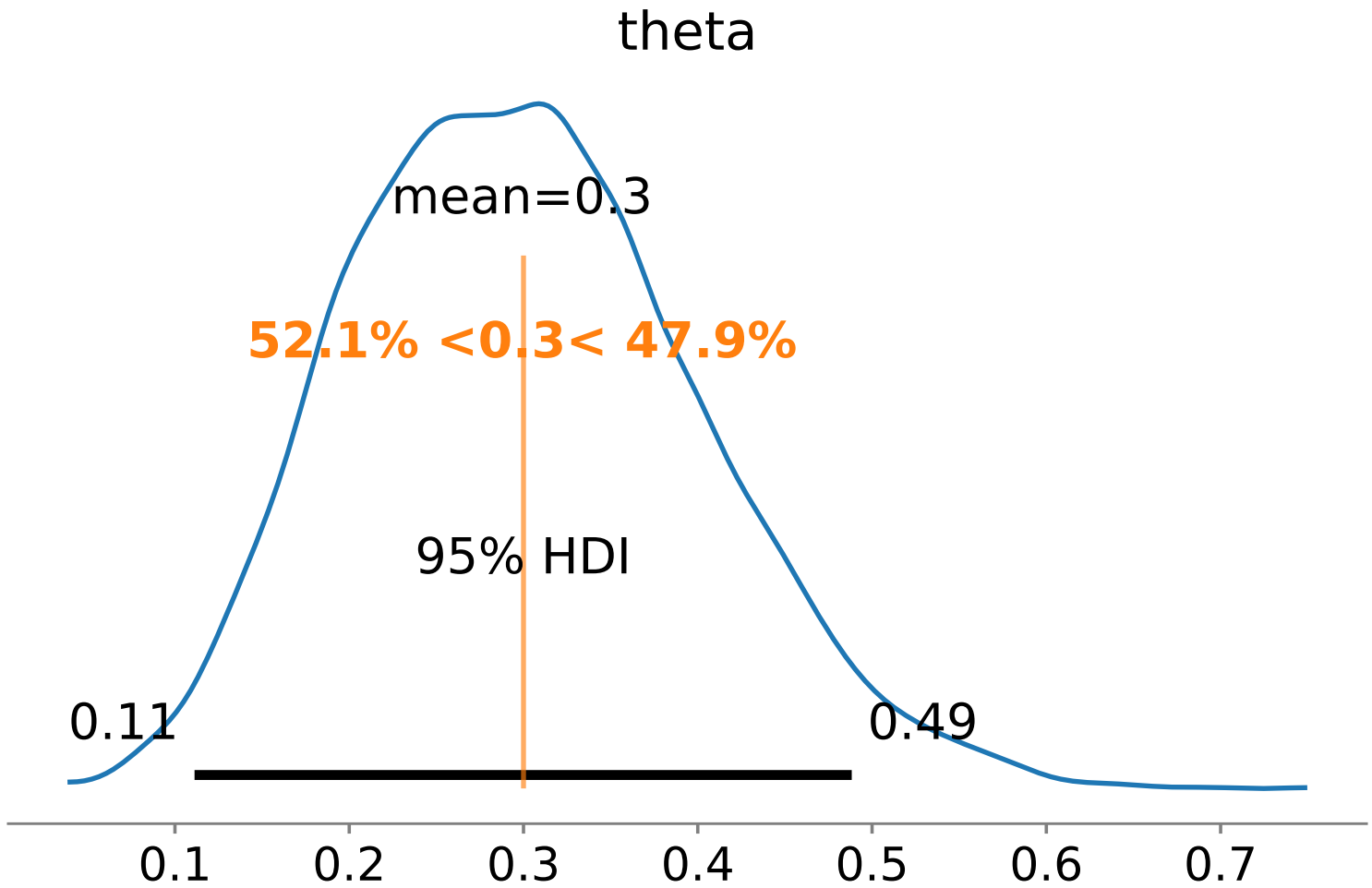
Dùng PyMC: xem Jupyter Notebook đi kèm. Lưu ý, dùng PyMC ta dễ dàng chọn các phân phối tiên nghiệm khác.

PyMC - Ví dụ 2 (tt)

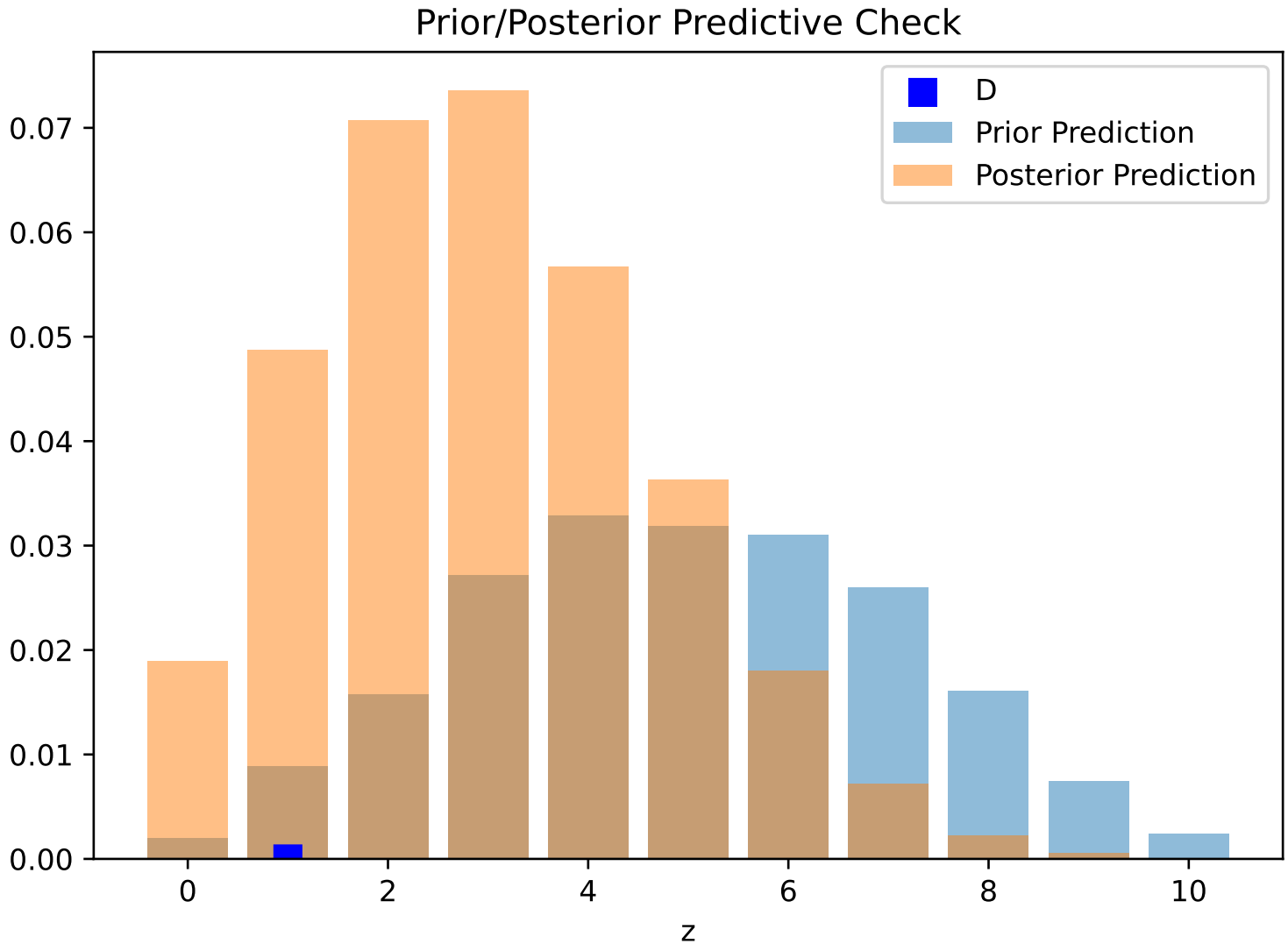




PyMC - Ví dụ 2 (tt)



PyMC - Ví dụ 2 (tt)



PyMC - Ví dụ 3

Bài toán. Cho

$$X \sim \mathcal{N}(0, 1),$$
$$Y \sim \mathcal{N}(X, 1).$$

Tính $P(X \geq 0 | Y = y)$. (Chẳng hạn, ta muốn "dự đoán" dấu của X.)

Nhận xét, xác suất tiên nghiệm

$$P(X \geq 0) = 0.5$$

nhưng vì Y được sinh dựa trên X nên nếu biết giá trị của Y là y thì xác suất hậu nghiệm

$$P(X \geq 0 | Y = y)$$

phản ánh chính xác hơn khả năng của biến cố $X \geq 0$.

PyMC - Ví dụ 3 (tt)

Giải. Hàm mật độ xác suất của X

$$f_X(x) \propto e^{-\frac{x^2}{2}}.$$

Hàm mật độ xác suất có điều kiện của Y khi biết $X = x$

$$f_{Y|X=x}(y) \propto e^{-\frac{(y-x)^2}{2}}.$$

Hàm mật độ xác suất có điều kiện của X khi biết $Y = y$

$$f_{X|Y=y}(x) \propto e^{-\frac{x^2}{2}} e^{-\frac{(y-x)^2}{2}} \propto e^{-\frac{x^2+(y-x)^2}{2}}.$$

Vì

$$-\frac{x^2 + (y-x)^2}{2} = -\left(x - \frac{y}{2}\right)^2 - \frac{y^2}{2}$$

và y là giá trị đã biết nên

$$f_{X|Y=y}(x) \propto e^{-\left(x-\frac{y}{2}\right)^2}.$$

PyMC - Ví dụ 3 (tt)

Đối chiếu với dạng của phân phối chuẩn, ta thấy

$$(X|Y=y) \sim \mathcal{N}\left(\frac{y}{2}, \frac{1}{2}\right).$$

Như vậy

$$P(X \geq 0|Y=y) = 1 - \Phi\left(\frac{-y}{\sqrt{2}}\right)$$

với Φ là hàm phân phối tích lũy của phân phối chuẩn tắc. Chẳng hạn,

$$P(X \geq 0|Y=2) = 1 - \Phi\left(-\sqrt{2}\right) \approx 0.9214$$

$$P(X \geq 0|Y=-2) = 1 - \Phi\left(\sqrt{2}\right) \approx 0.0786$$

Dùng PyMC: xem Jupyter Notebook đi kèm.

PyMC - Ví dụ 4

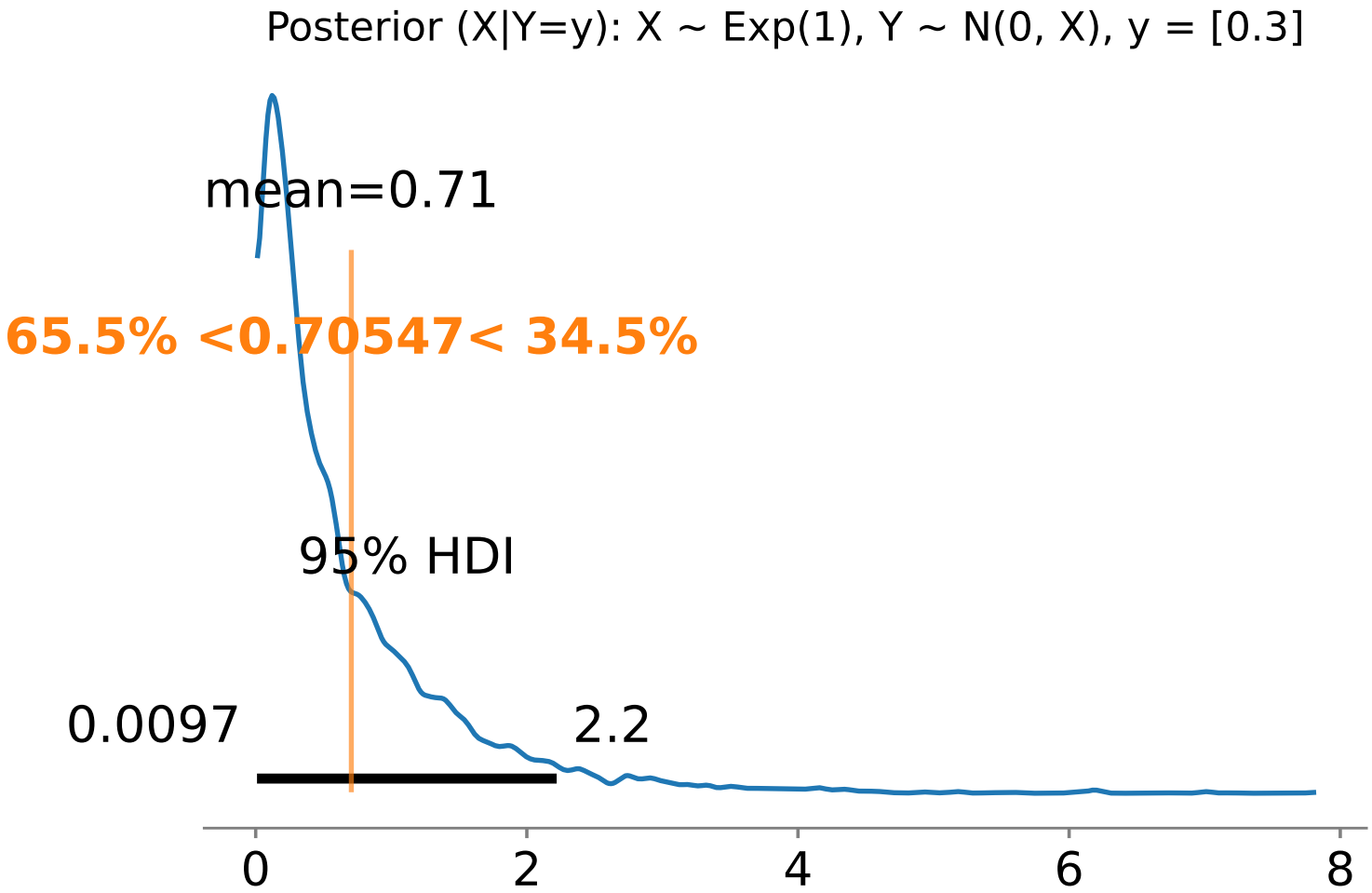
Bài toán. Cho

$$X \sim \text{Exp}(1),$$
$$Y \sim \mathcal{N}(0, X).$$

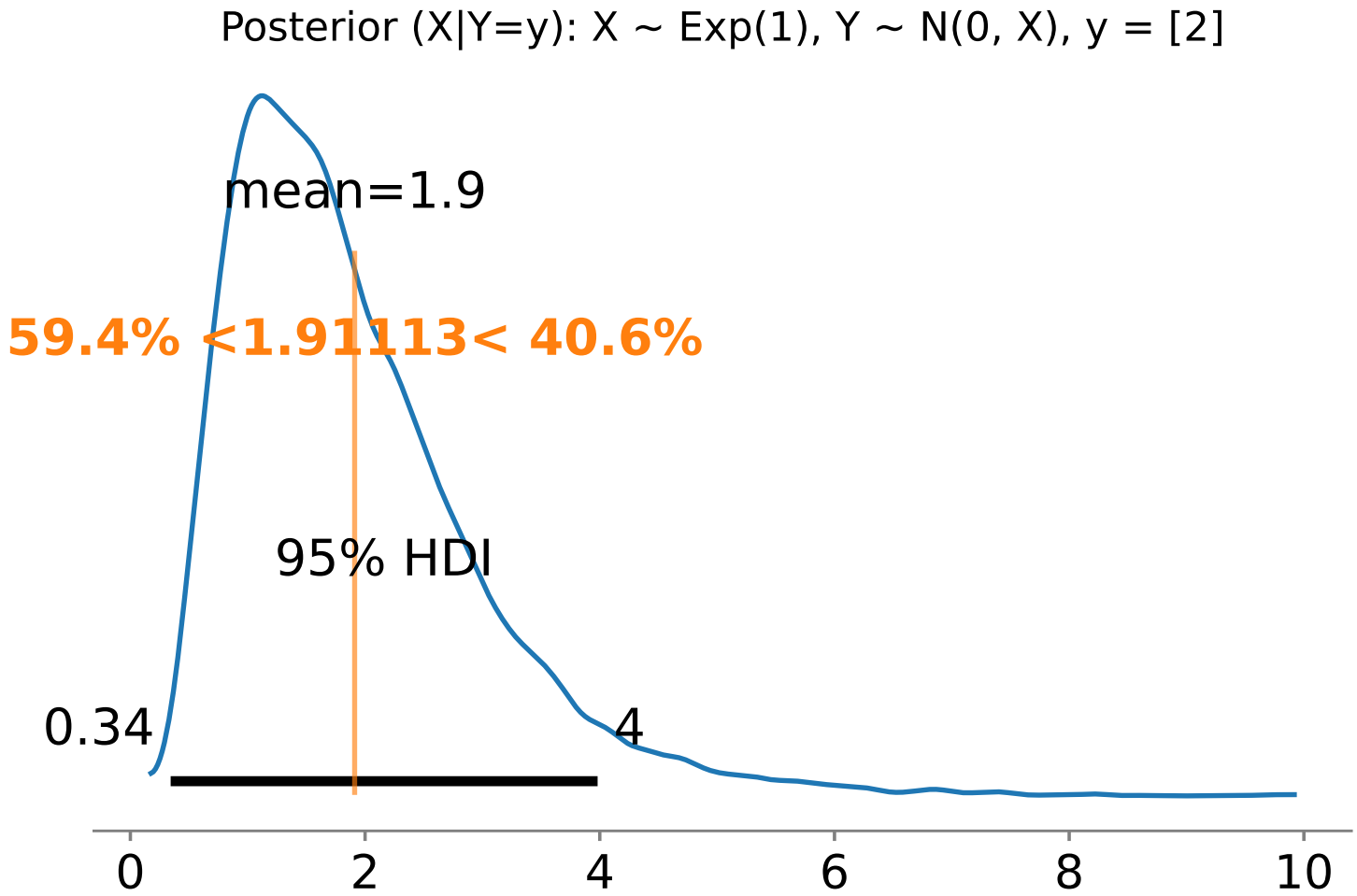
Tìm phân phối hậu nghiệm $(X|Y = y)$.

Giải. Dùng PyMC: xem Jupyter Notebook đi kèm.

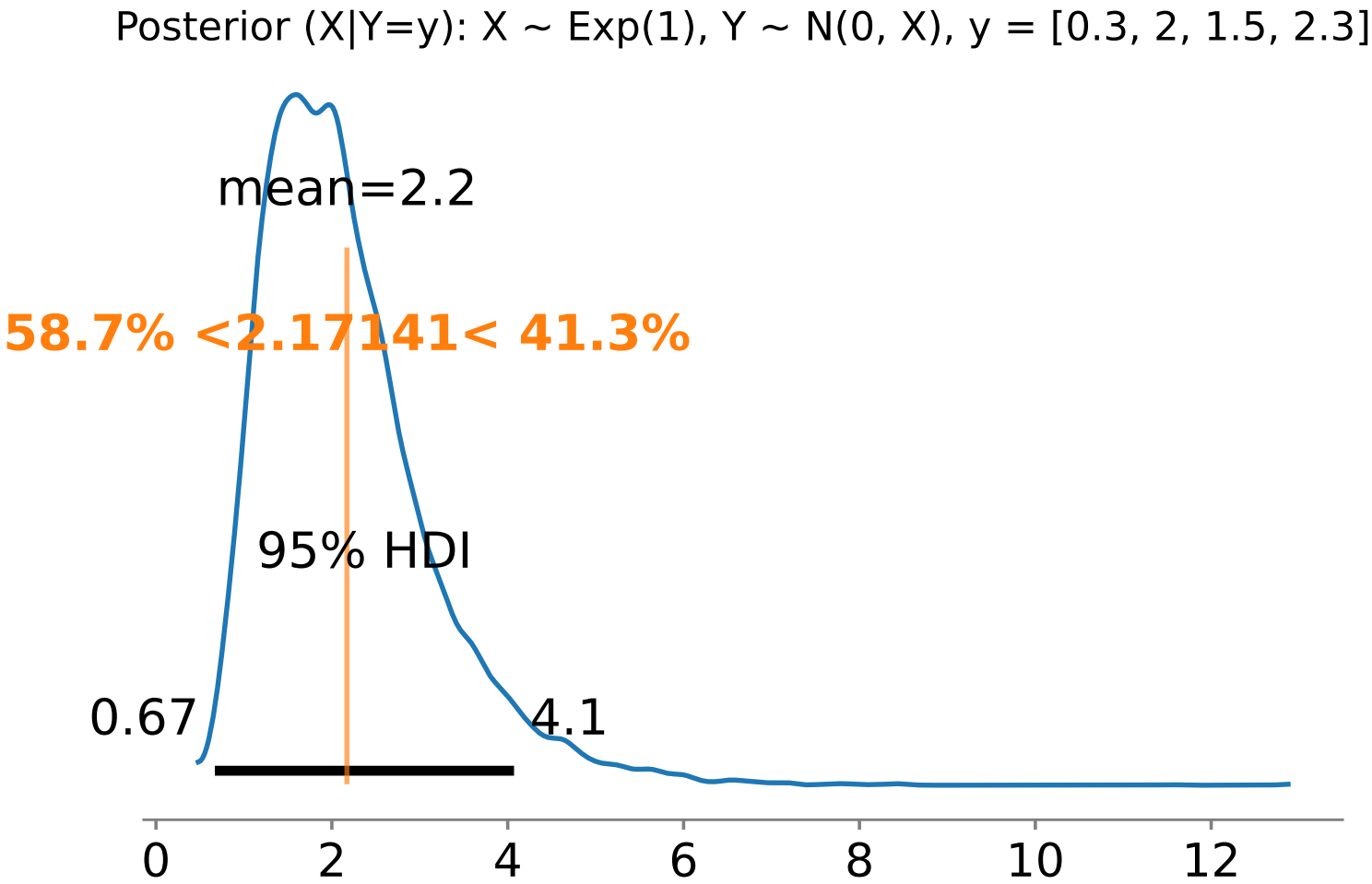
PyMC - Ví dụ 4 (tt)



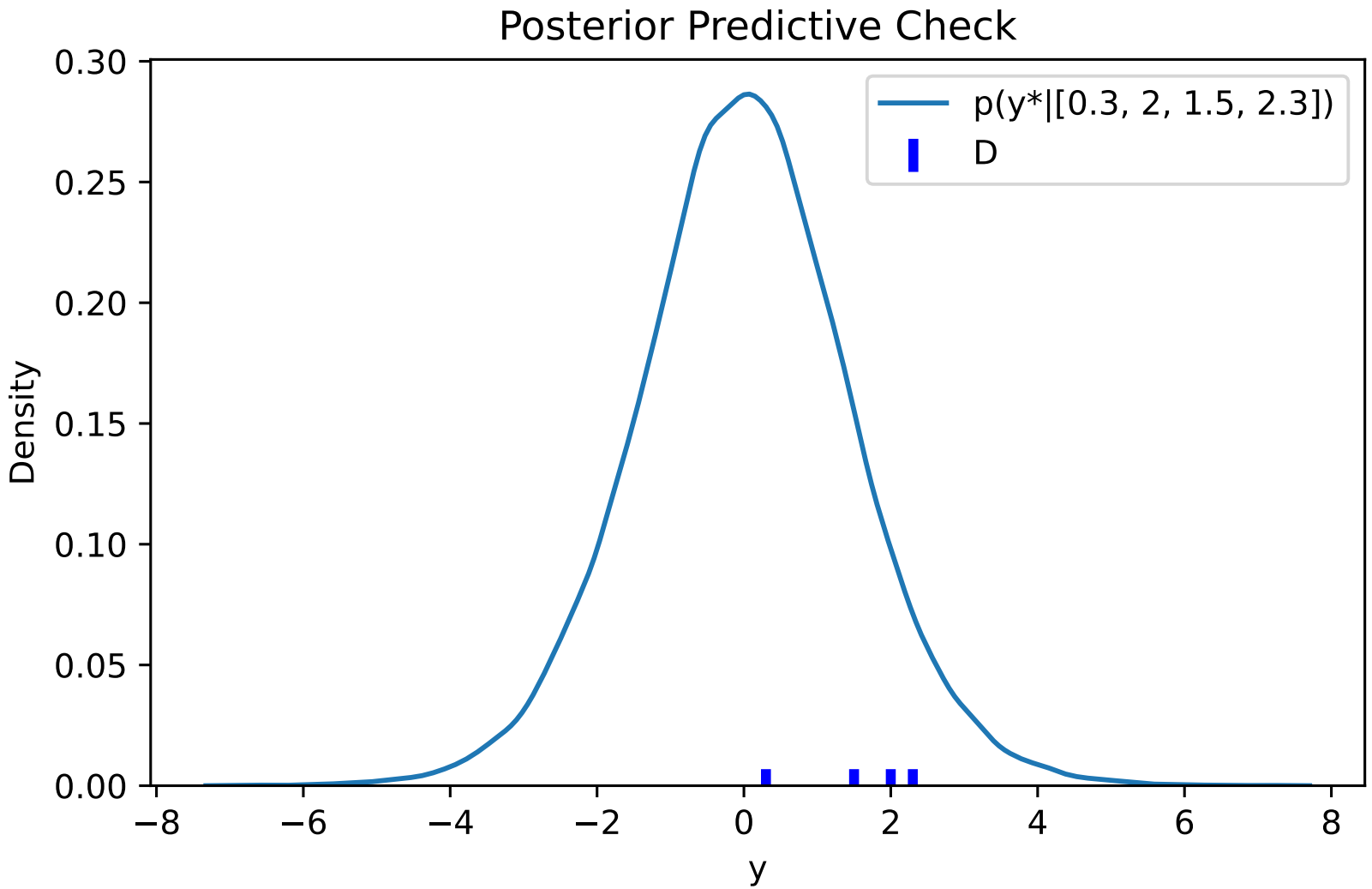
PyMC - Ví dụ 4 (tt)



PyMC - Ví dụ 4 (tt)



PyMC - Ví dụ 4 (tt)



Nội dung

So sánh nhóm

- **So sánh nhóm** (group comparision): so sánh giá trị (biến định lượng) nào đó giữa các nhóm (biến định tính) nào đó với nhau.
- Ví dụ: so sánh tác dụng thuốc giữa các nhóm bệnh nhân, so sánh hiệu quả học tập giữa các nhóm học sinh, so sánh hiệu quả thực tế của các chính sánh, ...
- So sánh nhóm thường được thực hiện bằng các thủ tục **kiểm định thống kê** (hypothesis testing) như kiểm định Z, kiểm định t, kiểm định F, ...

So sánh nhóm (tt)

- Thống kê Bayes mang lại giải pháp toàn diện cho so sánh nhóm.
- Để so sánh đại lượng y giữa các nhóm $1, 2, ..., K$, ta mô hình y cho mỗi nhóm và tìm phân phối hậu nghiệm của kì vọng μ_i và độ lệch chuẩn σ_i của y cho mỗi nhóm. Sau đó ta có thể so sánh giữa các nhóm bằng
  - Kì vọng của các μ_i
  - Hiệu kì vọng giữa các nhóm $d_{ij} = \mu_i - \mu_j$
  - Chỉ số Cohen d giữa các nhóm

$$\delta_{ij} = \frac{\mu_i - \mu_j}{\sqrt{\frac{\sigma_i^2 + \sigma_j^2}{2}}}$$

  - Xác suất "vượt trội" giữa các nhóm

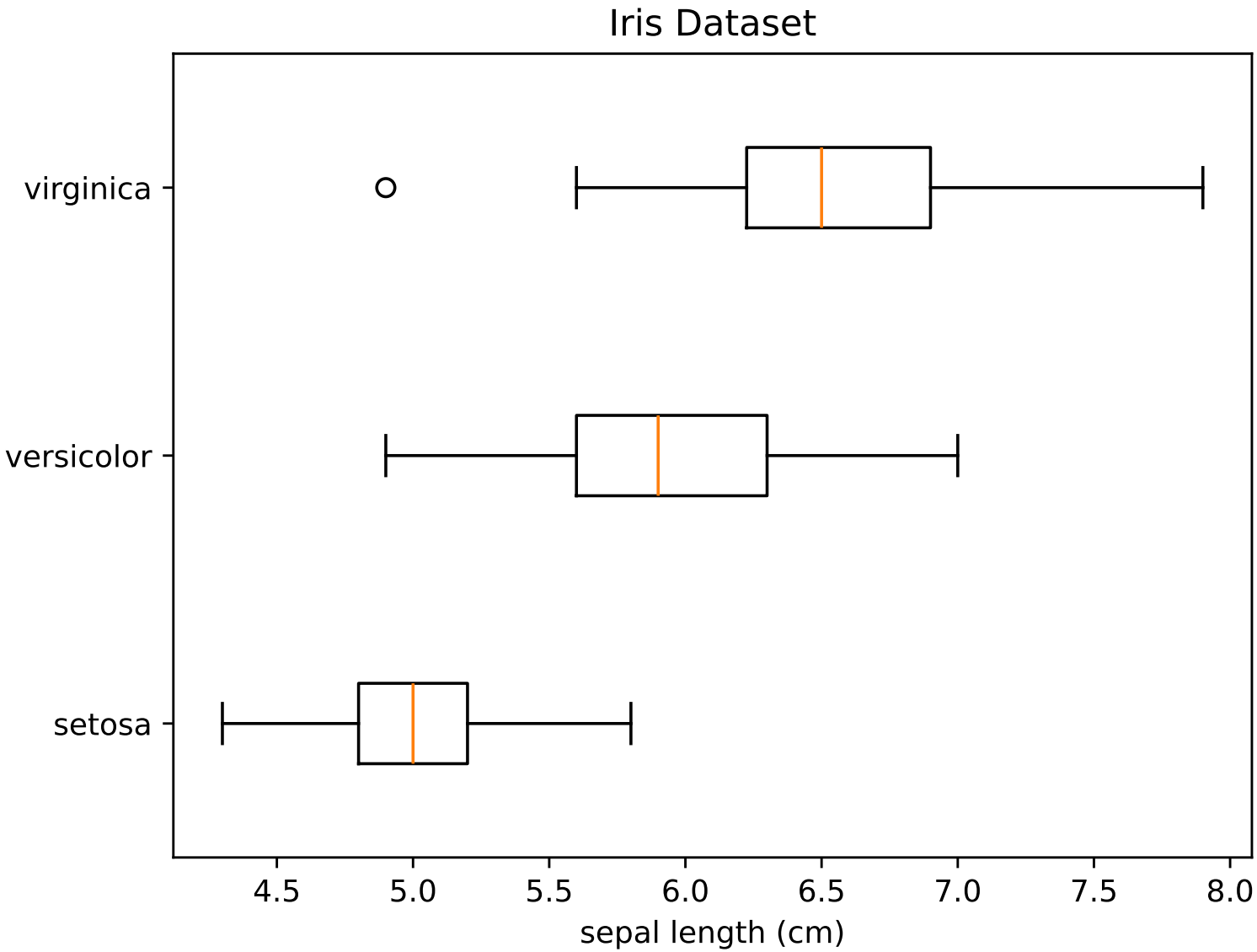
$$\Phi\left(\frac{\delta_{ij}}{\sqrt{2}}\right)$$

So sánh nhóm - Ví dụ

Iris Dataset

- Wikipedia: `https://en.wikipedia.org/wiki/Iris_flower_data_set`
- UCI: `https://archive.ics.uci.edu/ml/datasets/iris`
- Scikit-learn: `https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html`

So sánh nhóm - Ví dụ (tt)



So sánh nhóm - Ví dụ (tt)

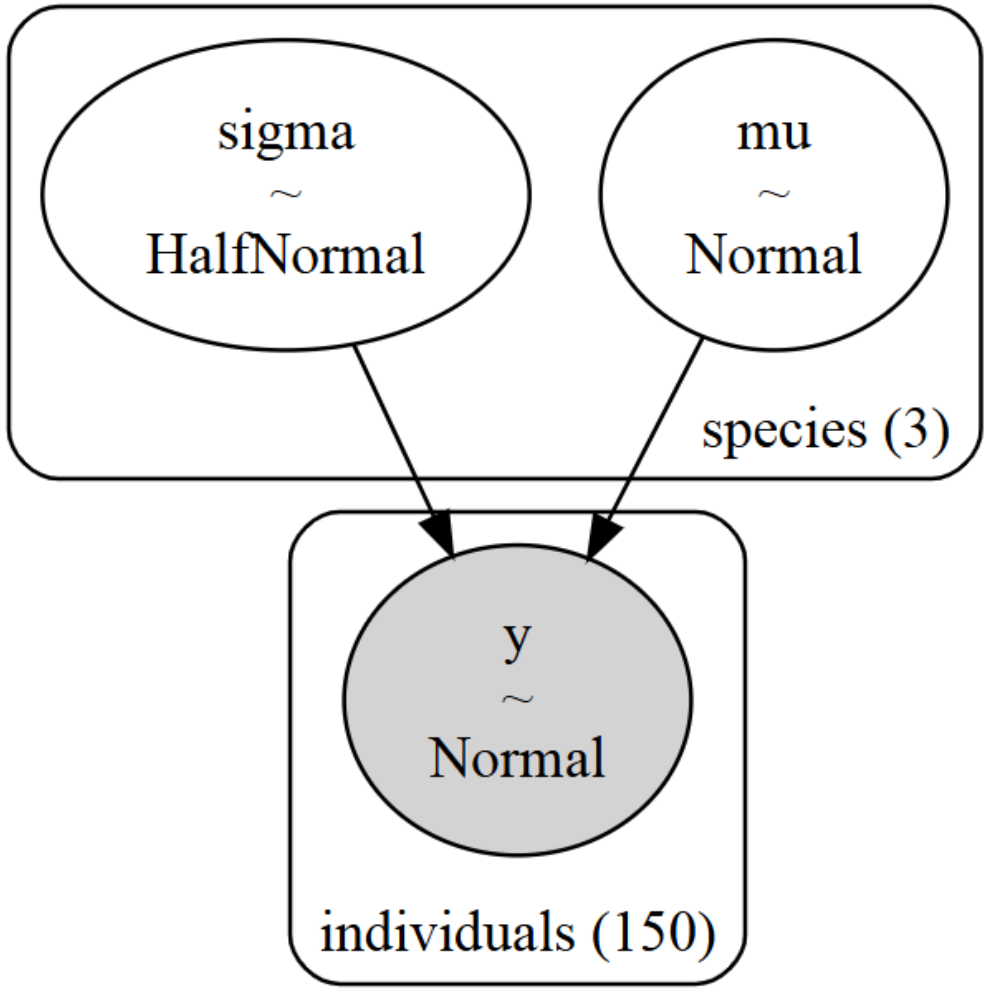
Bài toán. Từ bộ dữ liệu Iris, so sánh sepal-length giữa các giống hoa.

Giải. Ta mô hình petal-length (y) theo các giống hoa (s) như sau

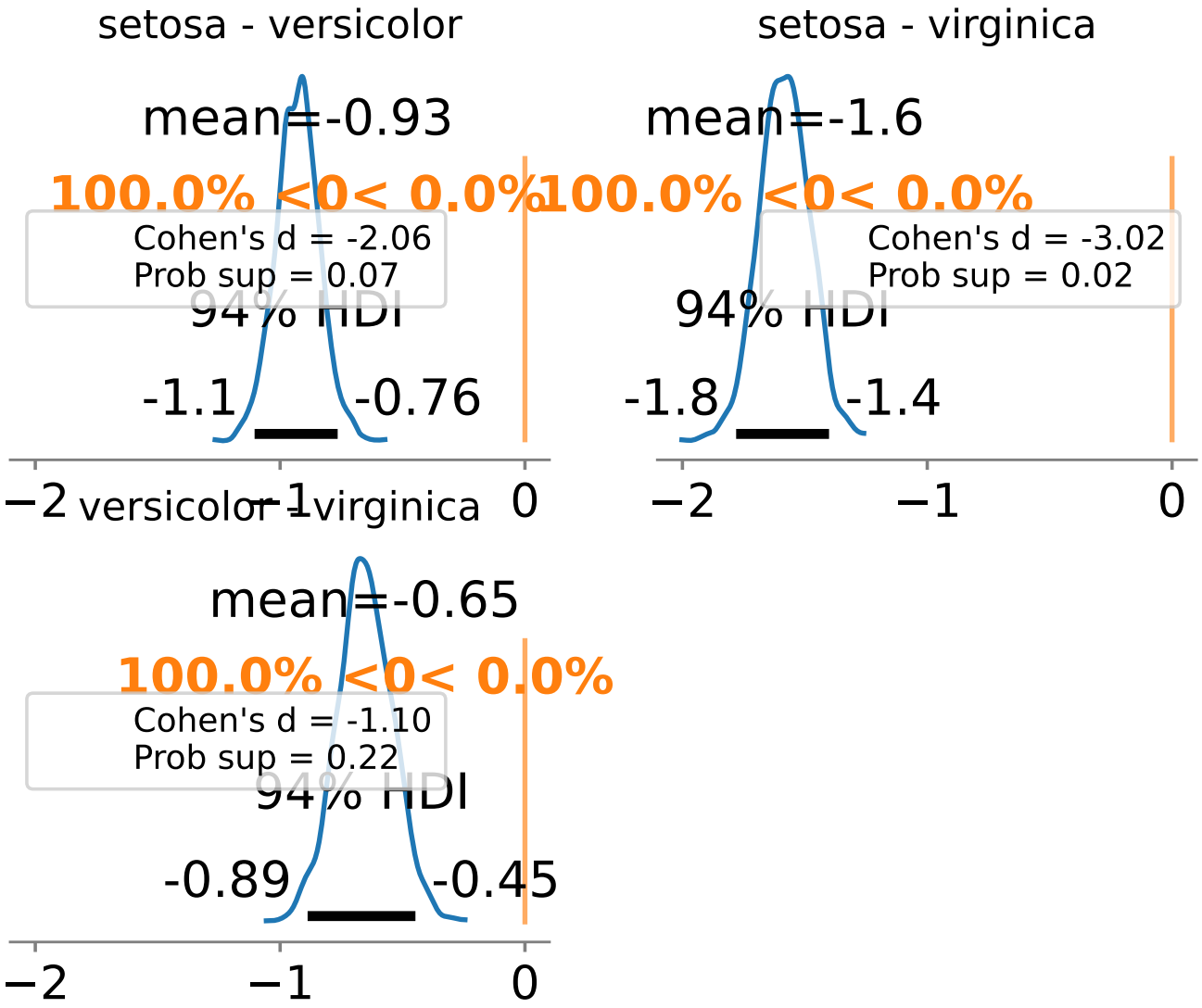
$$\mu_s \sim \mathcal{N}(0, 10^2), s = 0, 1, 2$$
$$\sigma_s \sim \mathcal{HN}(0, 10^2), s = 0, 1, 2$$
$$y_s \sim \mathcal{N}(\mu_s, \sigma_s), s = 0, 1, 2$$

Dùng PyMC: xem Jupyter Notebook đi kèm.

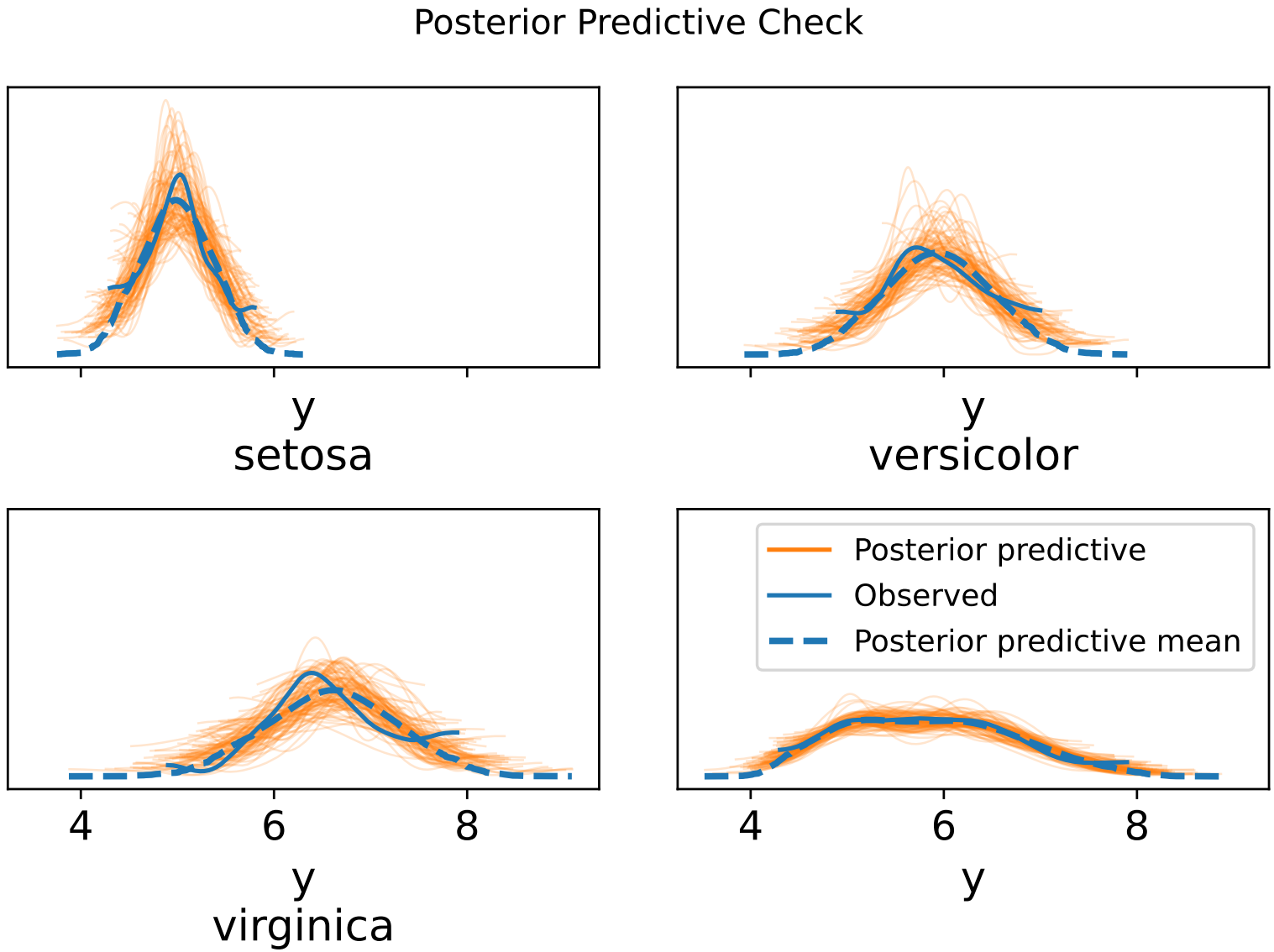
So sánh nhóm - Ví dụ (tt)



So sánh nhóm - Ví dụ (tt)



So sánh nhóm - Ví dụ (tt)



Nội dung

Suy luận kháng ngoại lai

- **Ngoại lai** (outlier, anomaly) là những dữ liệu "quá khác biệt".
- Việc phân tích dữ liệu khi có ngoại lai có thể cho kết quả không tốt.
- **Suy luận kháng ngoại lai** (robust inference) là các kỹ thuật phân tích dữ liệu cho kết quả không bị ảnh hưởng nhiều khi có ngoại lai.
- Thông thường, việc thay phân phối chuẩn bằng phân phối Student trong mô hình sẽ giúp kháng ngoại lai.

Suy luận kháng ngoại lai - Ví dụ

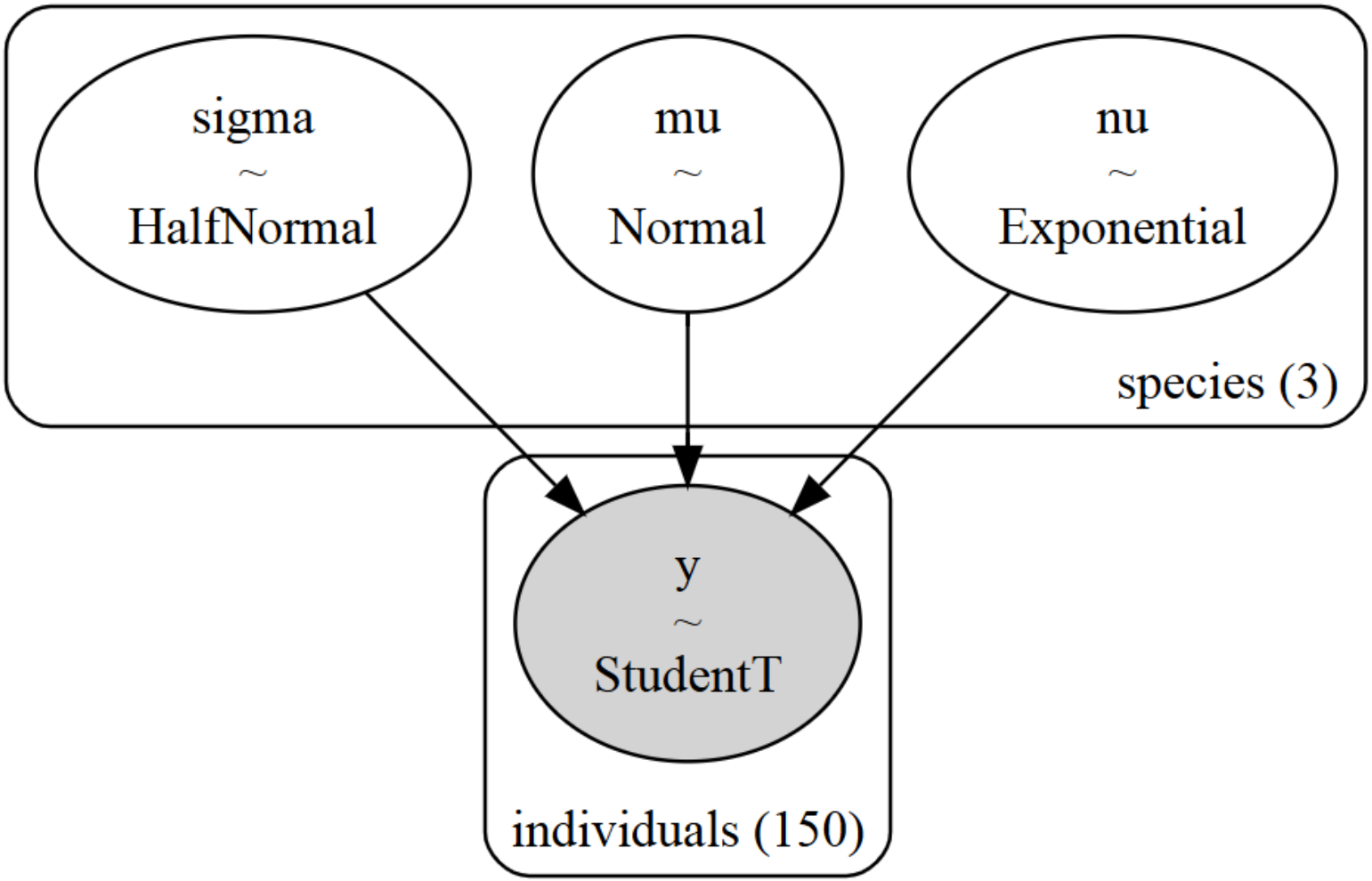
- Trong ví dụ so sánh nhóm ta thấy có ngoại lai trong dữ liệu.
- Để kháng ngoại lai, ta mô hình petal-length (y) theo các giống hoa (s) như sau

$$\mu_s \sim \mathcal{N}(0, 10^2), s = 0, 1, 2$$
$$\sigma_s \sim \mathcal{HN}(0, 10^2), s = 0, 1, 2$$
$$\nu_s \sim \text{Exp}(0, 10^{-1}), s = 0, 1, 2$$
$$y_s \sim \mathcal{T}(\nu_s, \mu_s, \sigma_s), s = 0, 1, 2$$

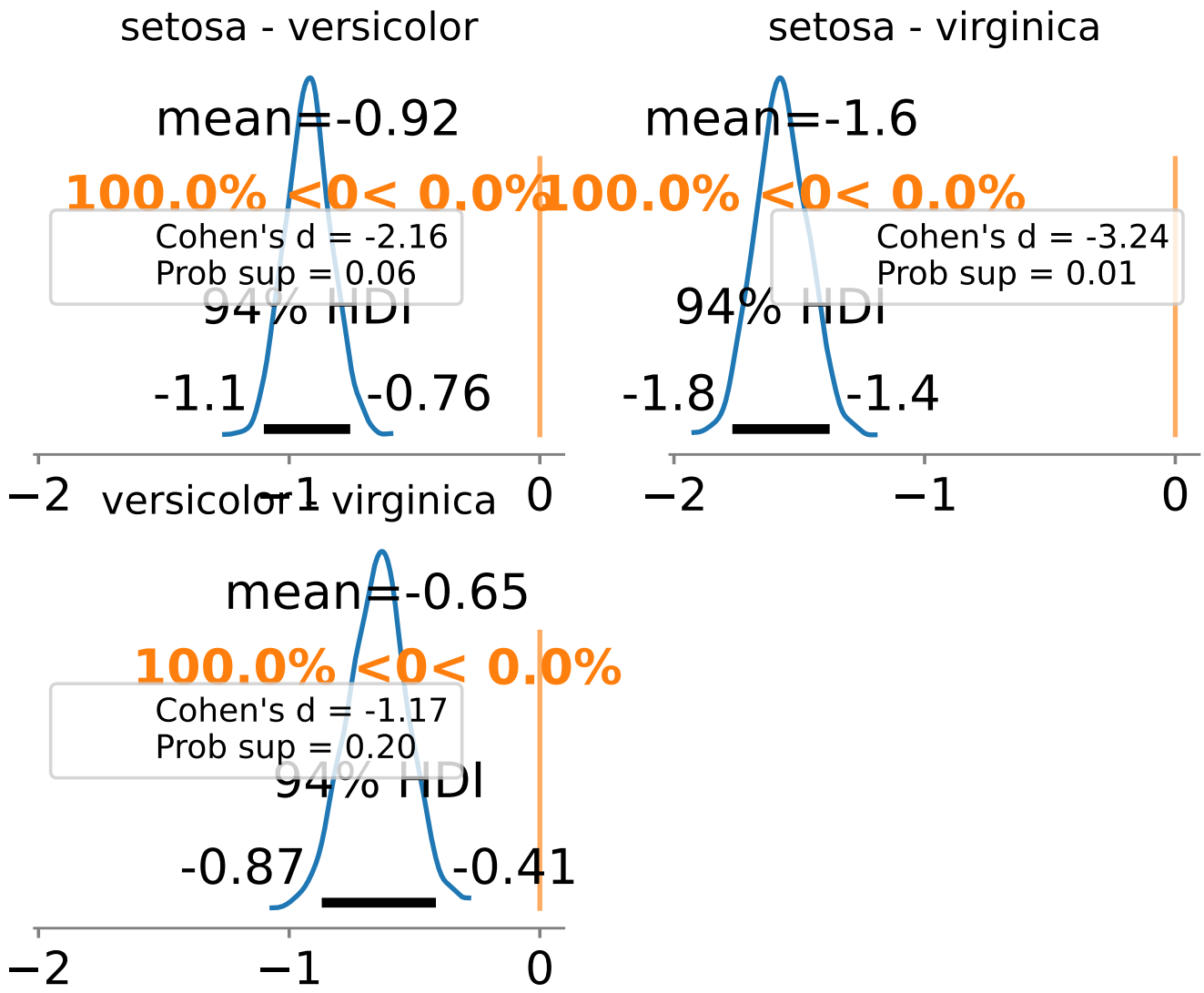
trong đó $\mathcal{T}(\nu, \mu, \sigma)$ là phân phối Student với ν bậc tự do.

Dùng PyMC: xem Jupyter Notebook đi kèm.

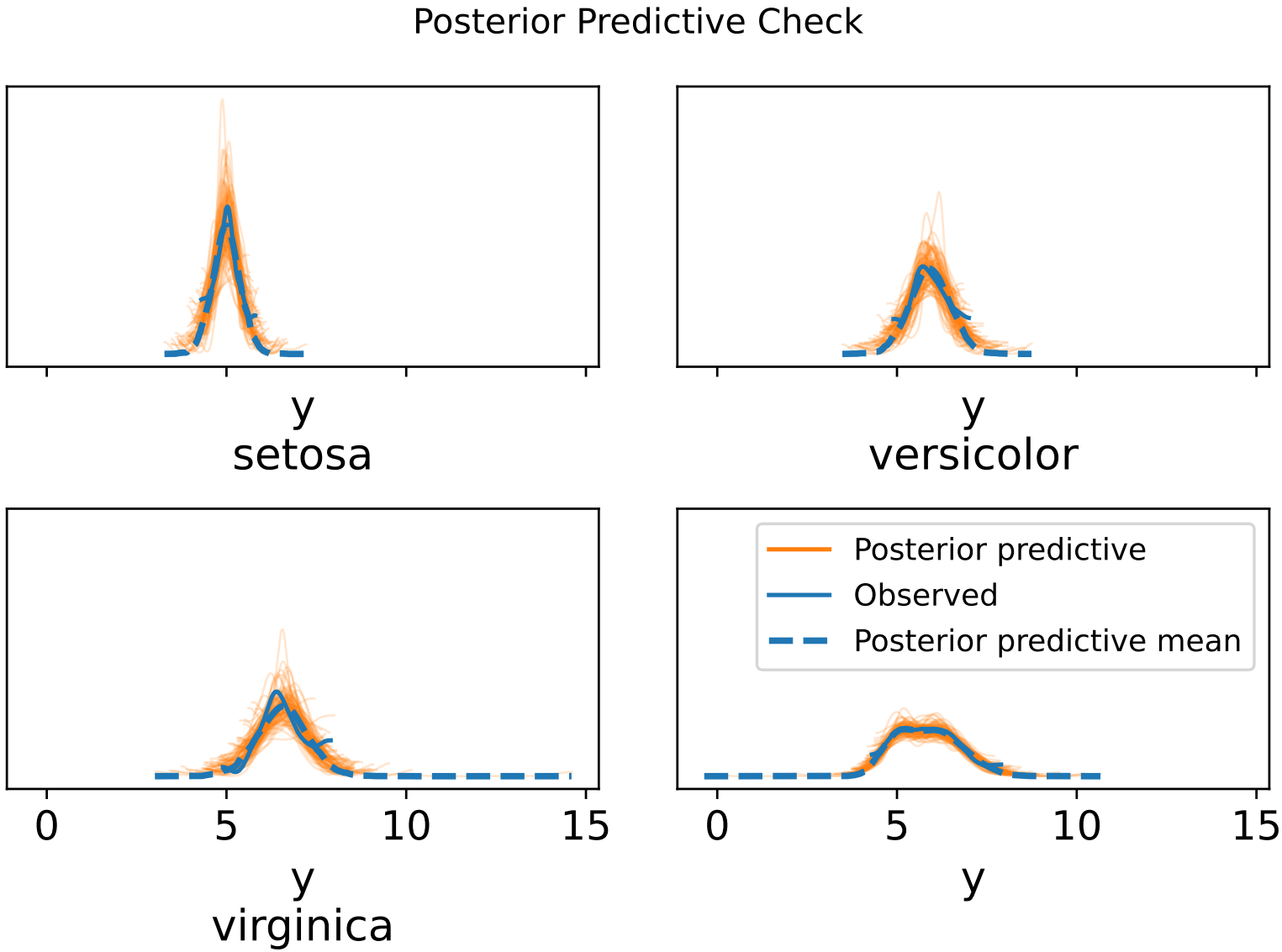
Suy luận kháng ngoại lai - Ví dụ (tt)



Suy luận kháng ngoại lai - Ví dụ (tt)



Suy luận kháng ngoại lai - Ví dụ (tt)



Nội dung

Mô hình phân cấp

- Trong ví dụ so sánh nhóm, các nhóm được mô hình riêng rẽ. Điều này tương đương với giả sử "các nhóm không liên quan gì nhau".
- Nếu muốn "chia sẻ thông tin", ta có thể mô hình sự liên quan giữa các nhóm bằng **mô hình phân cấp** (hierarchical model, multilevel model), trong đó, tham số của các phân phối tiên nghiệm được giả sử sinh từ cùng một phân phối ("tiên nghiệm của tiên nghiệm").
- Tổng quát, thống kê Bayes cho phép dễ dàng mô hình quan hệ phụ thuộc giữa các biến ngẫu nhiên theo nhiều mức, thường được gọi là **mô hình đồ thị** (graphical model).

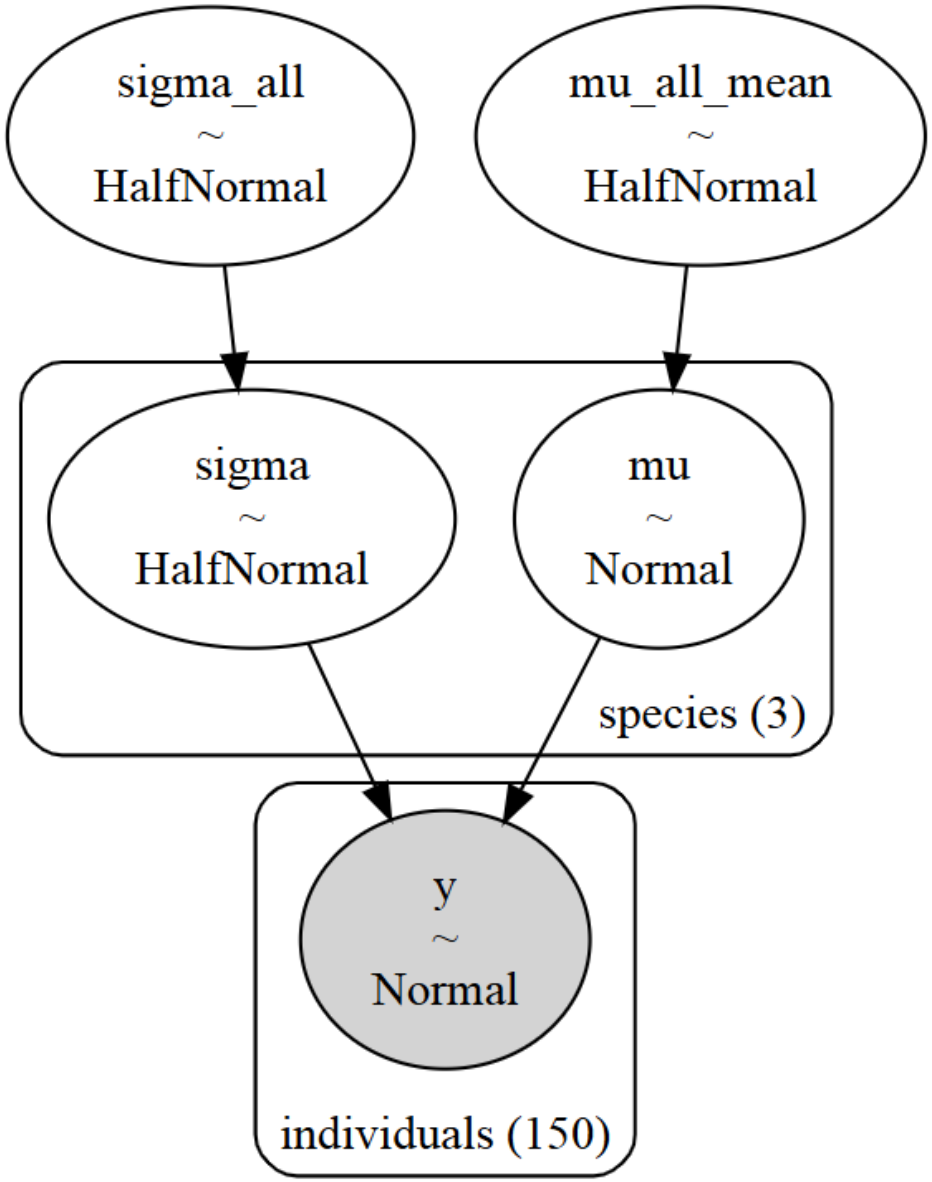
Mô hình phân cấp - Ví dụ

- Trong ví dụ so sánh nhóm, ta mô hình riêng rẽ các nhóm.
- Để "chia sẽ thông tin" giữa các nhóm, ta dùng mô hình phân cấp như sau

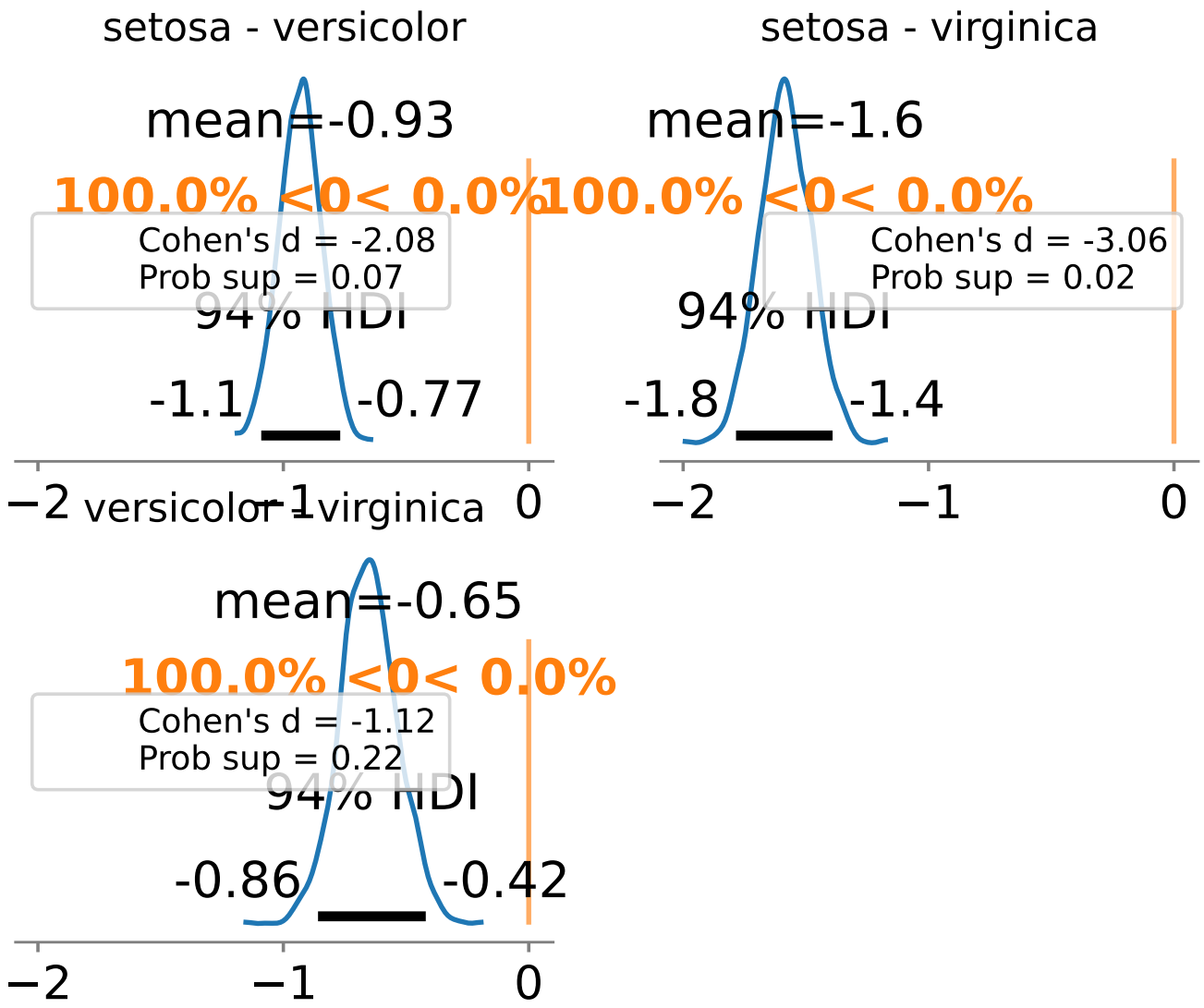
$$
\begin{aligned}
\mu_a &\sim \mathcal{N}(0, 10^2), \\
\sigma_a &\sim \mathcal{HN}(0, 10^2), \\
\mu_s &\sim \mathcal{N}(\mu_a, 10^2), s = 0, 1, 2 \\
\sigma_s &\sim \mathcal{HN}(0, \sigma_a^2), s = 0, 1, 2 \\
y_s &\sim \mathcal{N}(\mu_s, \sigma_s), s = 0, 1, 2
\end{aligned}
$$

Dùng PyMC: xem Jupyter Notebook đi kèm.

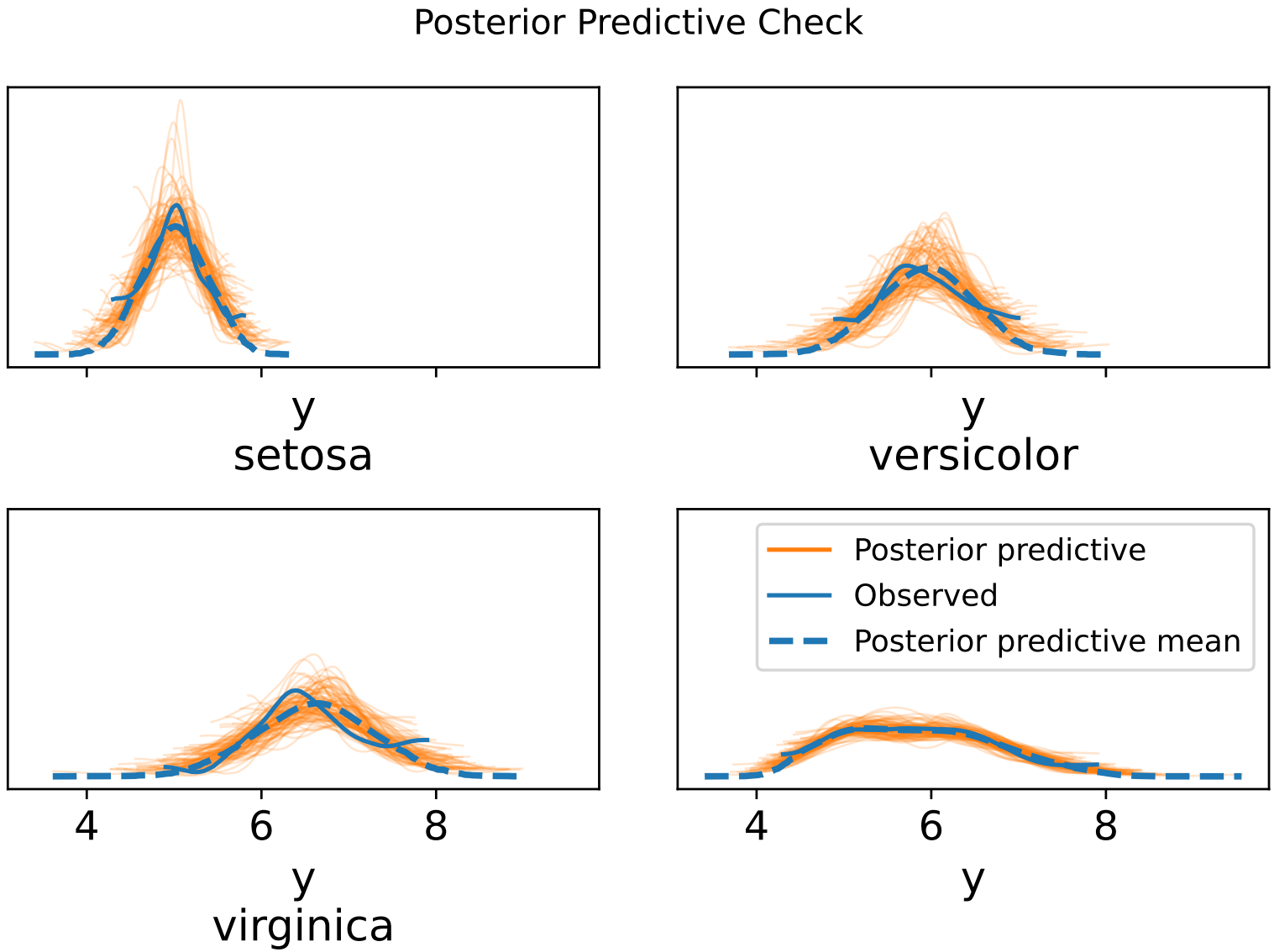
Mô hình phân cấp - Ví dụ (tt)



Mô hình phân cấp - Ví dụ (tt)



Mô hình phân cấp - Ví dụ (tt)



Nội dung

Mô hình tuyến tính tổng quát

- Ta muốn xác định sự phụ thuộc của một biến, gọi là **biến phụ thuộc** (dependent/predicted/response/output variable), vào một số biến khác, gọi là các **biến độc lập** (independent/predictor/explanatory/input variables).
- Các biến có thể được đo lường theo các **thang đo** (scale) khác nhau: metric, count, ordinal, nominal/categorical.
- Mô hình **hồi qui tuyến tính đơn** (simple linear regression)

$$y \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2),$$

trong đó, y là metric response và x là metric explanatory.

- **Mô hình tuyến tính tổng quát** (generalized linear model - GLM)

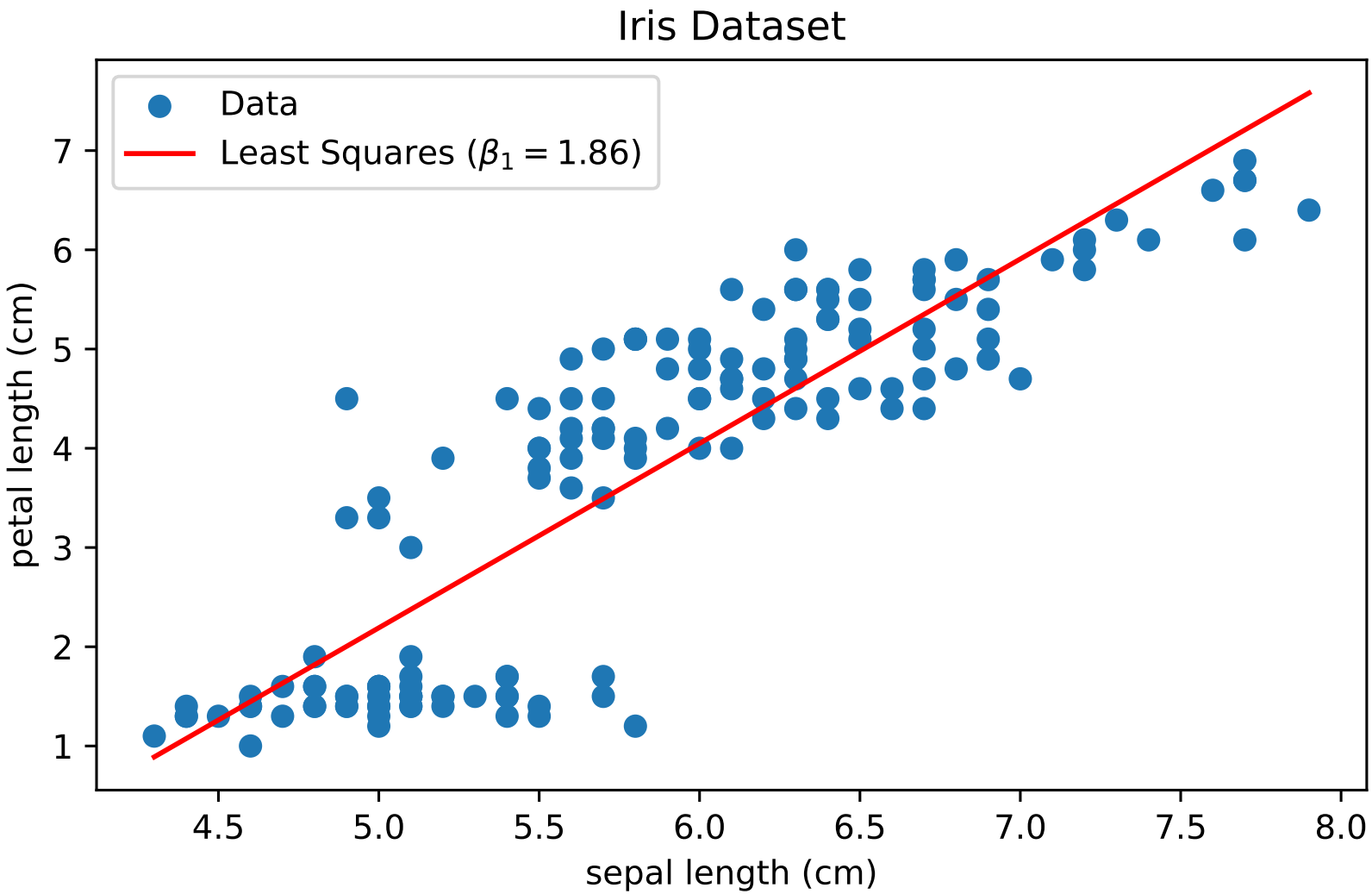
$$\begin{aligned} \text{lin}(x) &= \beta_0 + \beta_1 x_1 + ... + \beta_k x_k, \\ \mu &= f\left(\text{lin}(x), [\text{các tham số}]\right), \\ y &\sim \text{pdf}(\mu, [\text{các tham số}]). \end{aligned}$$

Mô hình tuyến tính tổng quát (tt)

Scale type of predicted y	Typical noise distribution $y \sim \text{pdf}(\mu, [\text{parameters}])$	Typical inverse link function $\mu = f(\text{lin}(x), [\text{parameters}])$
Metric	$y \sim \text{normal}(\mu, \sigma)$	$\mu = \text{lin}(x)$
Dichotomous	$y \sim \text{bernoulli}(\mu)$	$\mu = \text{logistic}(\text{lin}(x))$
Nominal	$y \sim \text{categorical}(\ldots, \mu_k, \ldots)$	$\mu_k = \dfrac{\exp(\text{lin}_k(x))}{\sum_c \exp(\text{lin}_c(x))}$
Ordinal	$y \sim \text{categorical}(\ldots, \mu_k, \ldots)$	$\mu_k = \Phi((\theta_k - \text{lin}(x))/\sigma) - \Phi((\theta_{k-1} - \text{lin}(x))/\sigma)$
Count	$y \sim \text{poisson}(\mu)$	$\mu = \exp(\text{lin}(x))$

(John K. Kruschke. *Doing Bayesian Data Analysis*.)

GLM - Ví dụ 1 (hồi qui tuyến tính đơn)



GLM - Ví dụ 1 (tt)

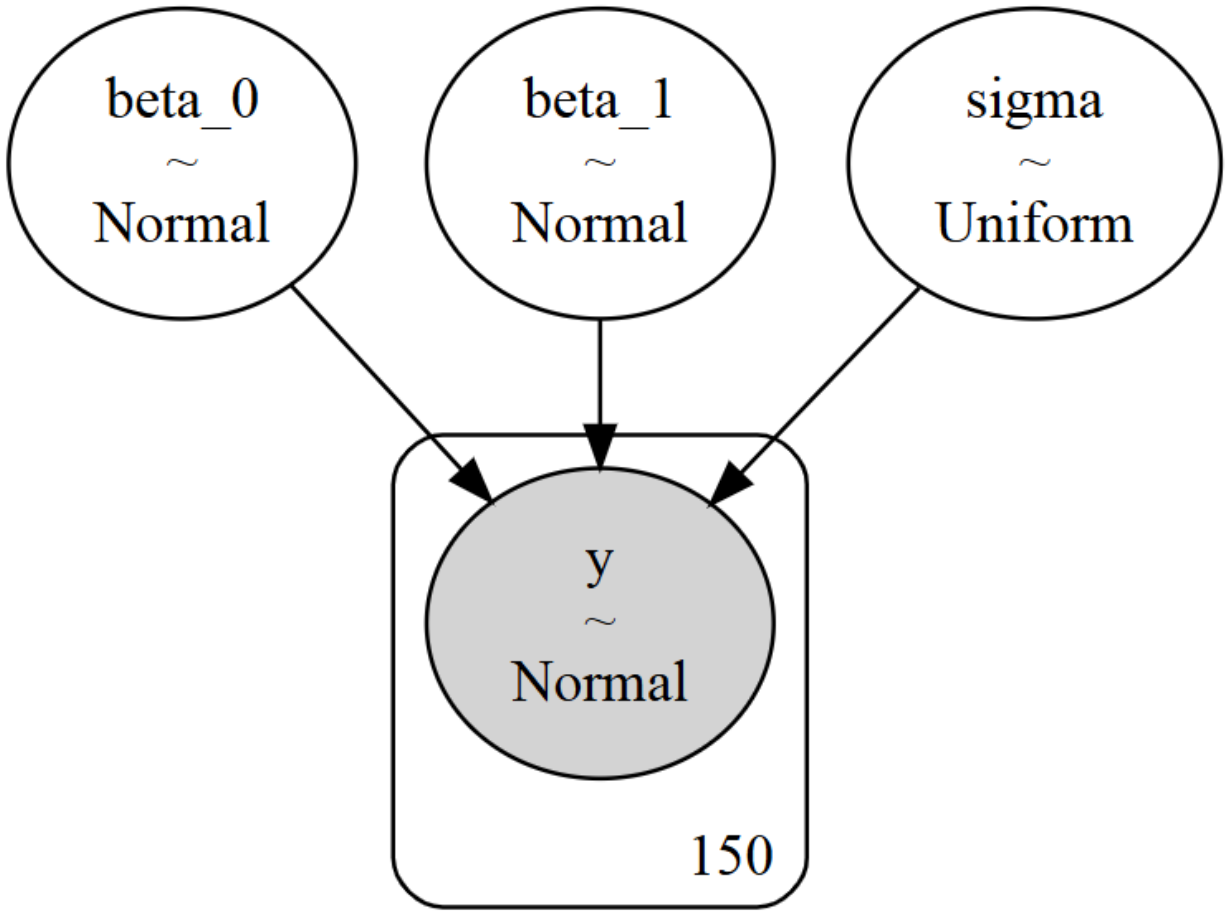
Bài toán. Từ dữ liệu Iris, xác định sự phụ thuộc của petal-length vào sepal-length.

Giải. Ta mô hình sự phụ thuộc của petal-length (y) vào sepal-length (x) như sau

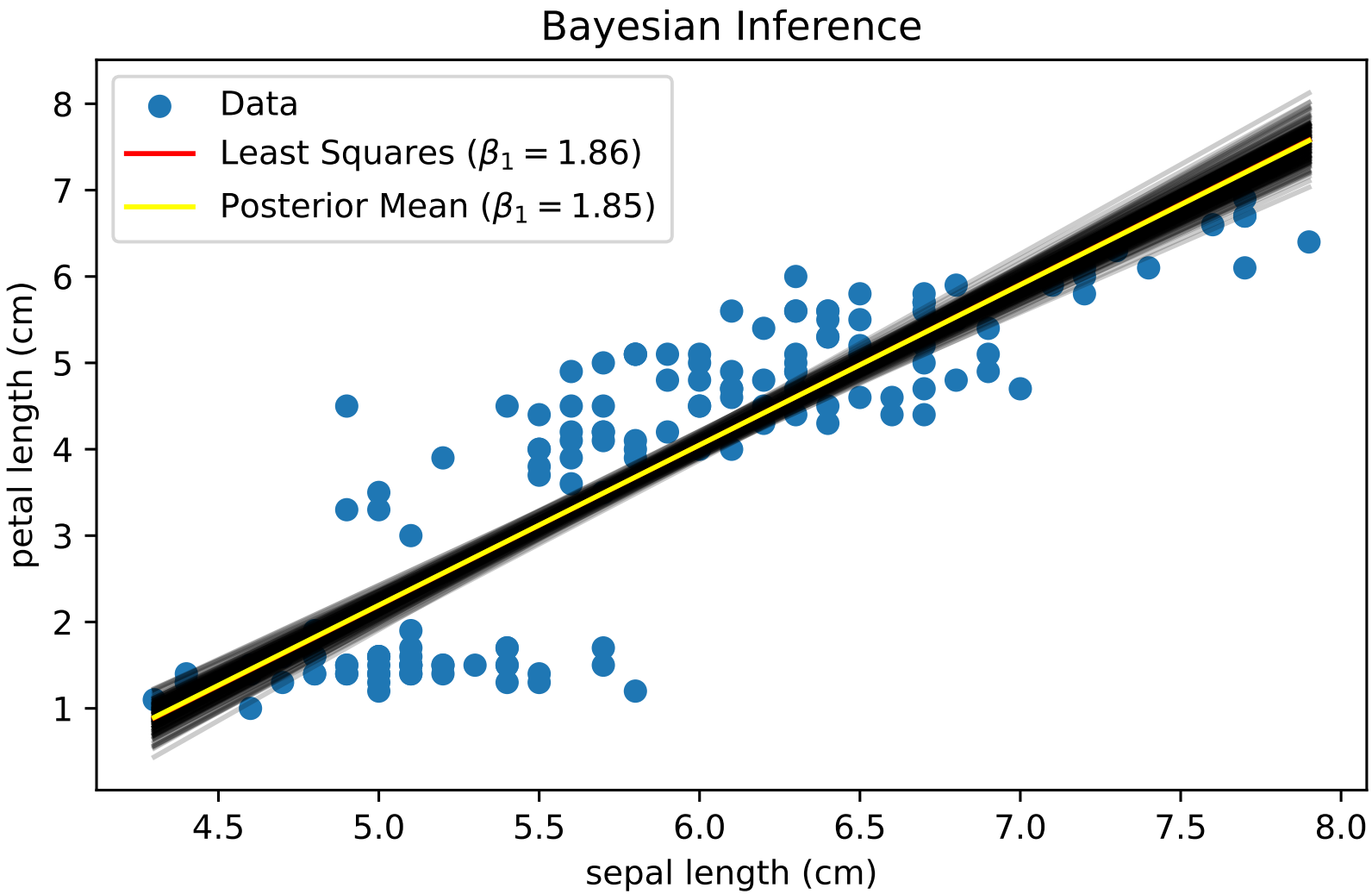
$$\begin{aligned}
\hat{y} &= \beta_0 + \beta_1 x, \\
y &\sim \mathcal{N}(\hat{y}, \sigma), \\
\beta_0 &\sim \mathcal{N}(0, 10^2), \\
\beta_1 &\sim \mathcal{N}(0, 10^2), \\
\sigma &\sim \mathcal{U}(0, 1000).
\end{aligned}$$

Dùng PyMC: xem Jupyter Notebook đi kèm.

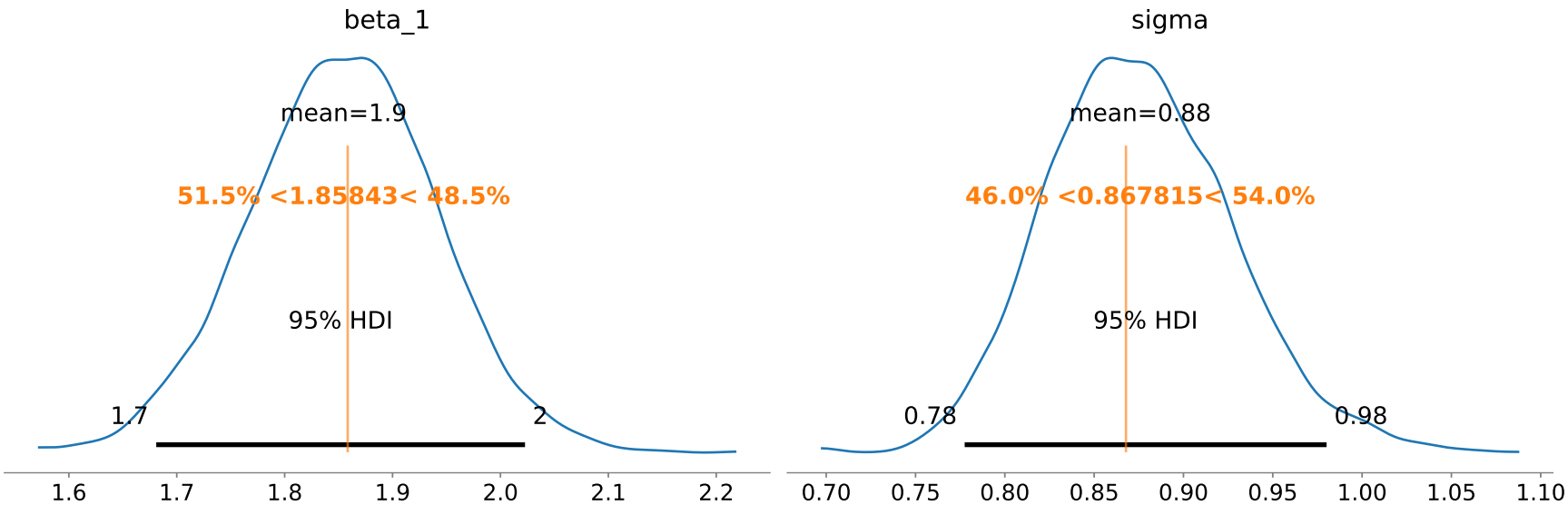
GLM - Ví dụ 1 (tt)



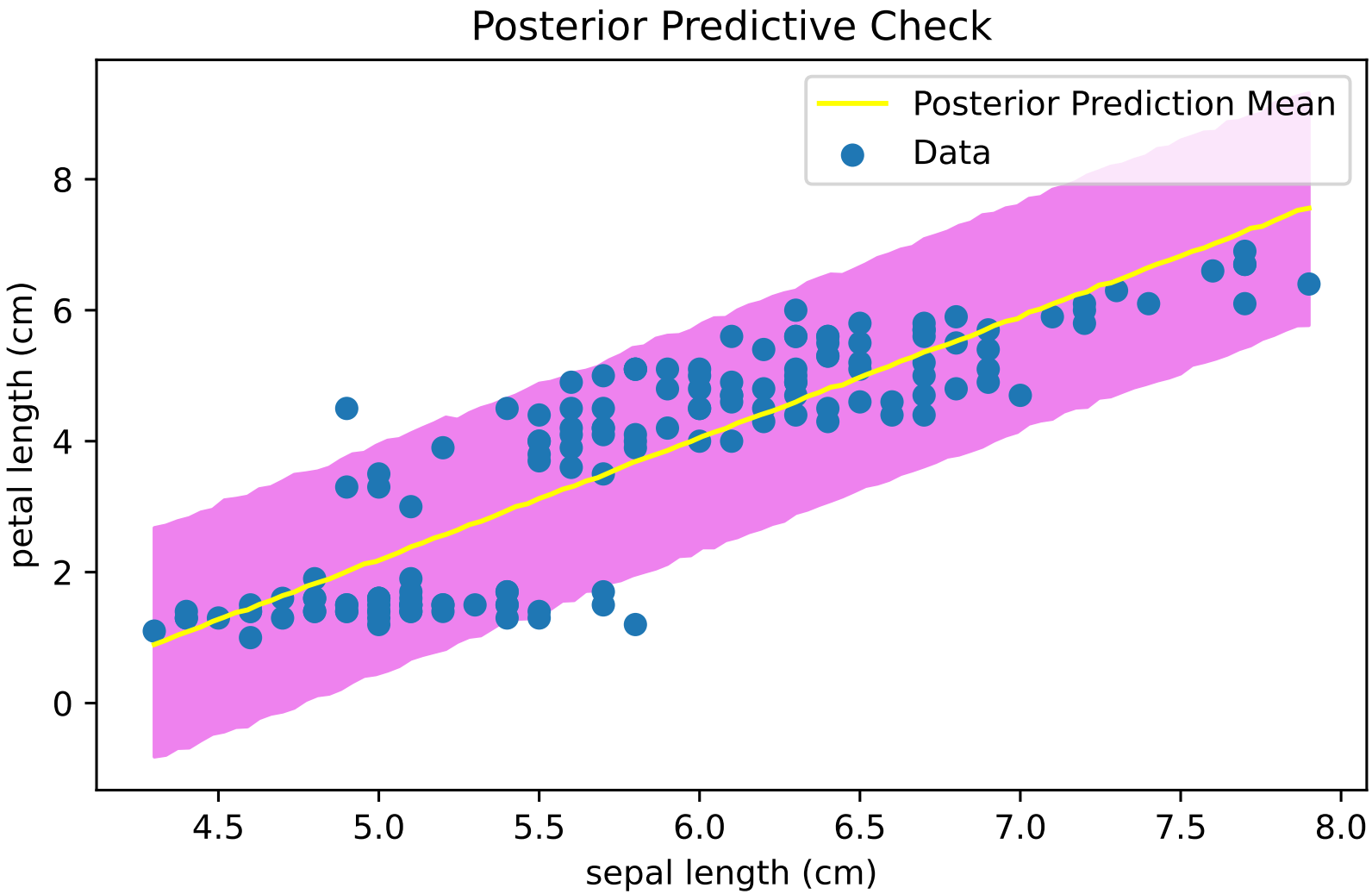
GLM - Ví dụ 1 (tt)


Bayesian Inference
Data
Least Squares ($\beta_1 = 1.86$)
Posterior Mean ($\beta_1 = 1.85$)
petal length (cm)
sepal length (cm)
1
2
3
4
5
6
7
8
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

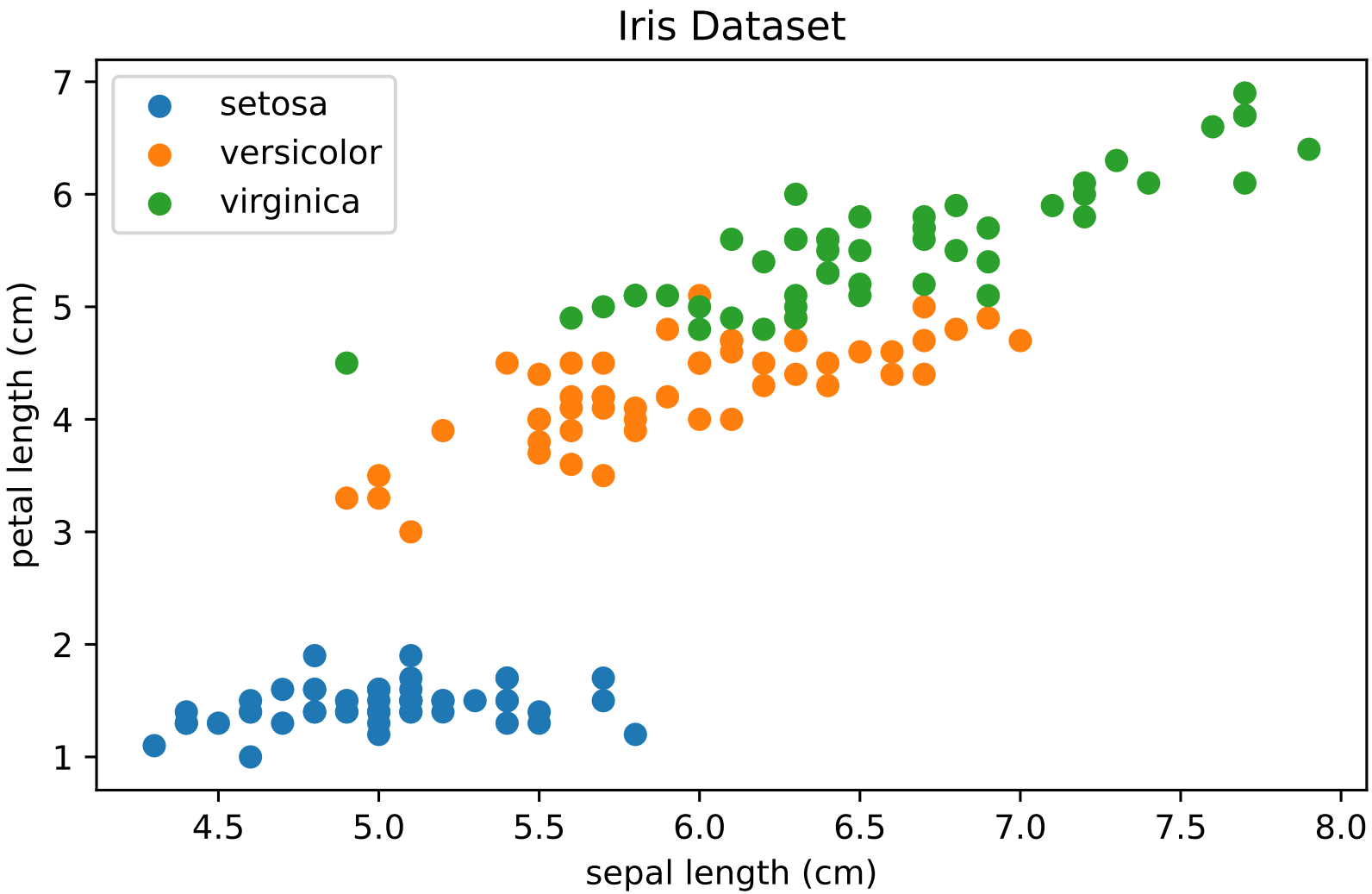
GLM - Ví dụ 1 (tt)


beta_1
mean=1.9
51.5% <1.85843< 48.5%
95% HDI
1.7
2
1.6
1.7
1.8
1.9
2.0
2.1
2.2
sigma
mean=0.88
46.0% <0.867815< 54.0%
95% HDI
0.78
0.98
0.70
0.75
0.80
0.85
0.90
0.95
1.00
1.05
1.10

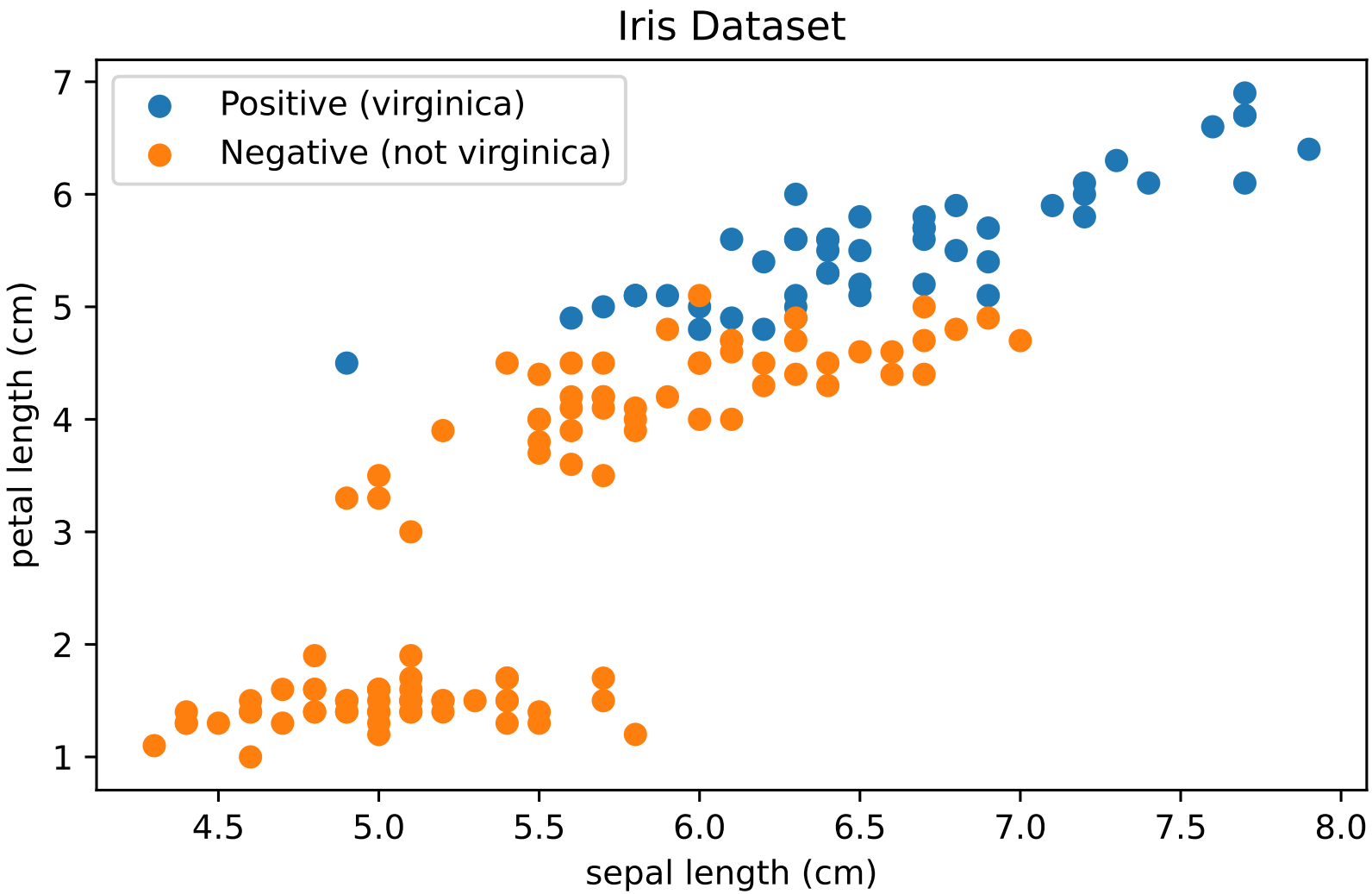
GLM - Ví dụ 1 (tt)


Posterior Predictive Check
Posterior Prediction Mean
Data
petal length (cm)
sepal length (cm)
0
2
4
6
8
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

GLM - Ví dụ 2 (hồi qui logistic)



GLM - Ví dụ 2 (tt)


Iris Dataset
Positive (virginica)
Negative (not virginica)
petal length (cm)
sepal length (cm)
1
2
3
4
5
6
7
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

GLM - Ví dụ 2 (tt)

Bài toán. Từ dữ liệu Iris, xác định sự phụ thuộc của class vào petal-length và sepal-length, cụ thể, phân biệt nhóm virginica với 2 nhóm còn lại.

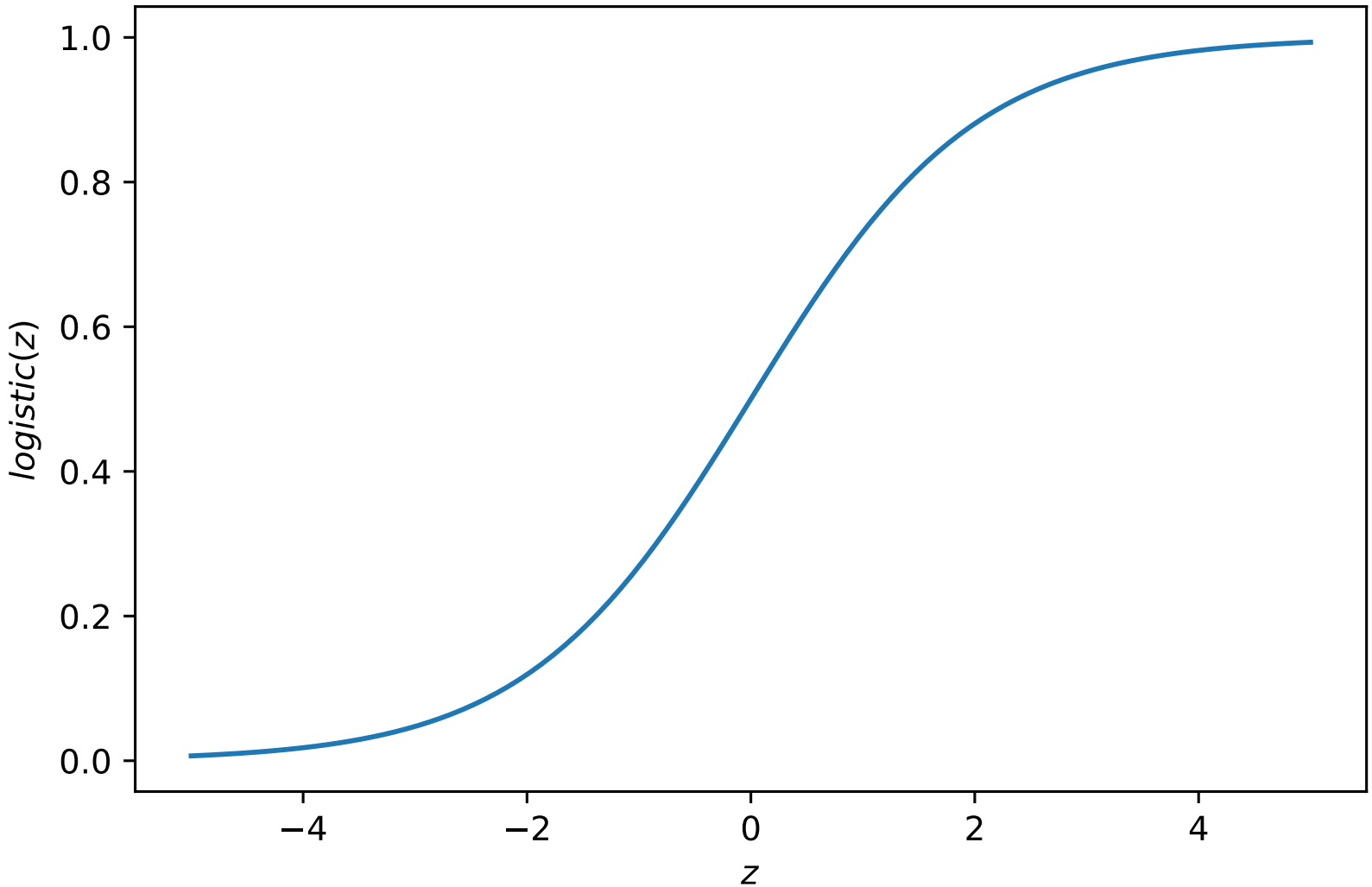
Giải. "Từ gợi ý của dữ liệu", ta mô hình sự phụ thuộc của class (y, nhận giá trị 1 nếu là virginica và 0 nếu không) vào sepal-length (x_1) và petal-length (x_2) như sau

$$\begin{aligned} \mu &= \text{logistic}(\beta_0 + \beta_1 x_1 + \beta_2 x_2), \\ y &\sim \text{Bernoulli}(\mu), \\ \beta_j &\sim \mathcal{N}(M_j, S_j^2), j = 0, 1, 2. \end{aligned}$$

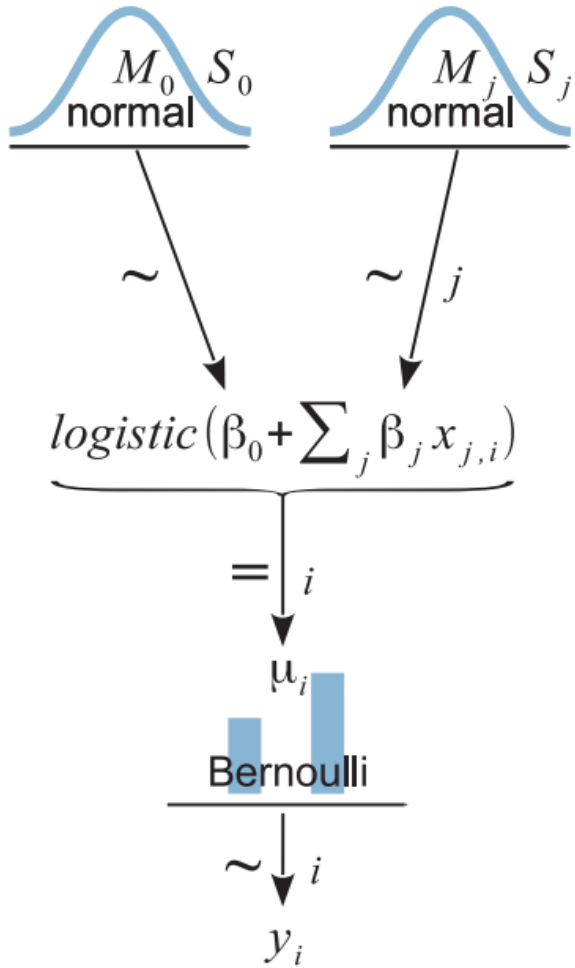
với

$$\text{logistic}(z) = \frac{1}{1 + e^{-z}}.$$

GLM - Ví dụ 2 (tt)

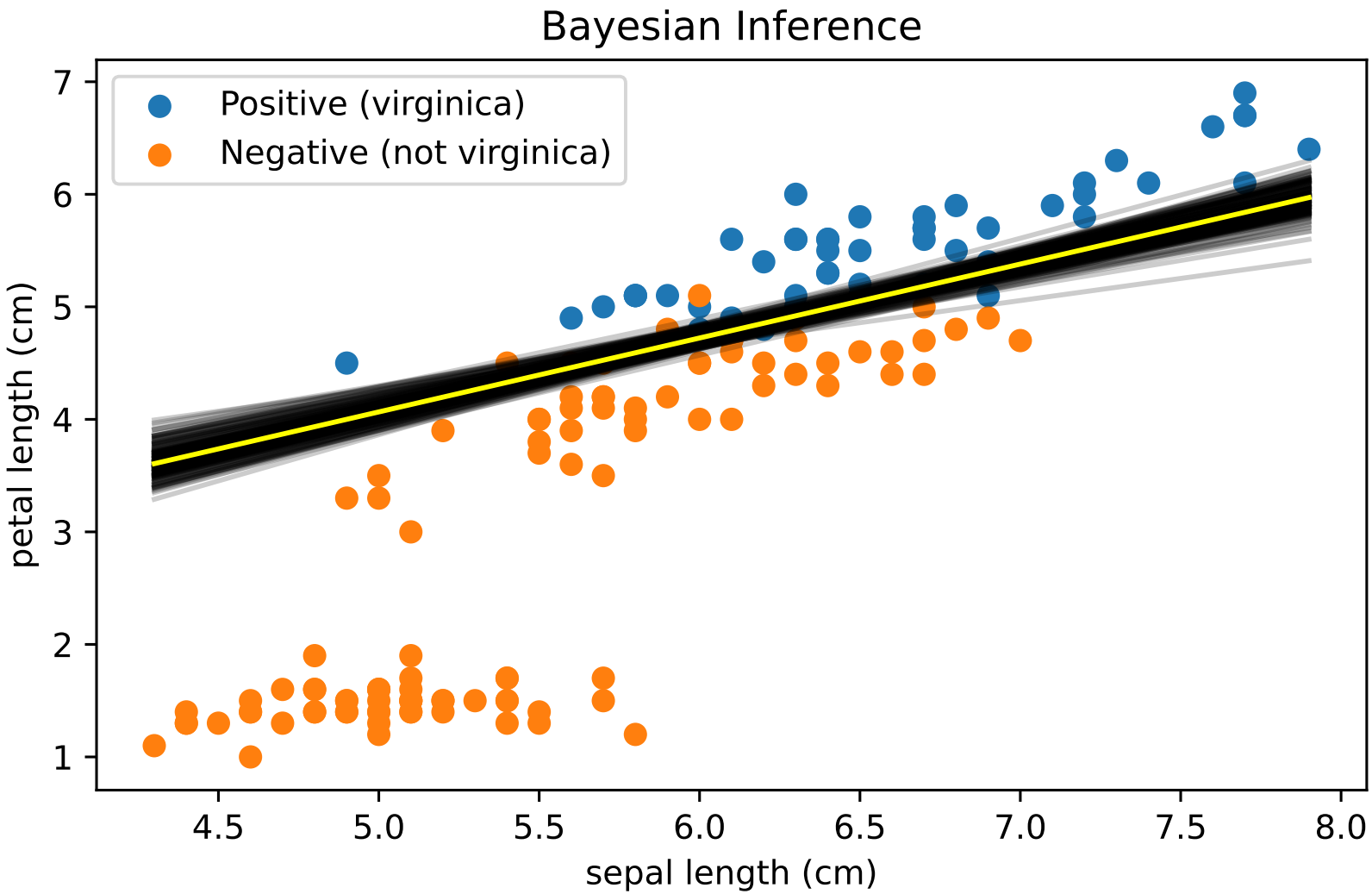


GLM - Ví dụ 2 (tt)

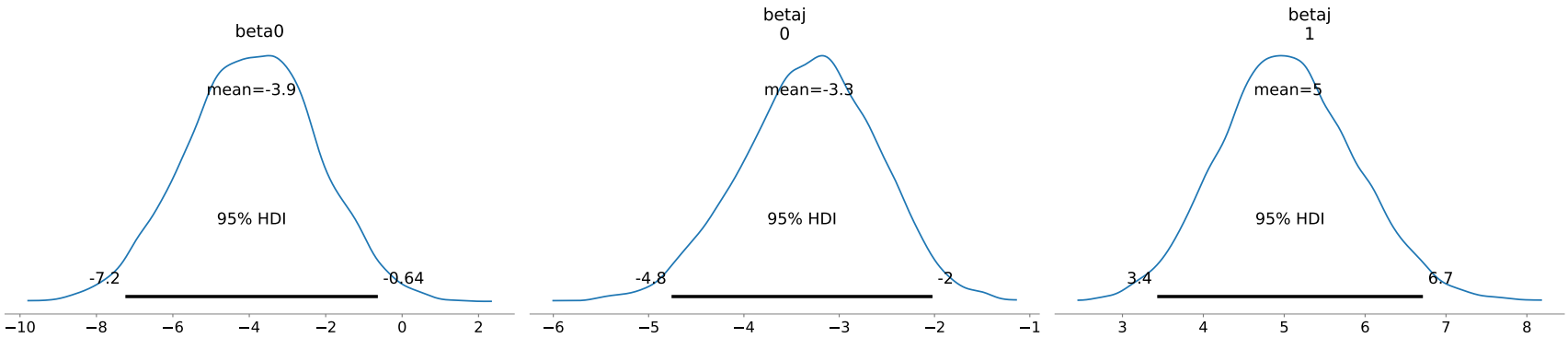


(John K. Kruschke. *Doing Bayesian Data Analysis*.)

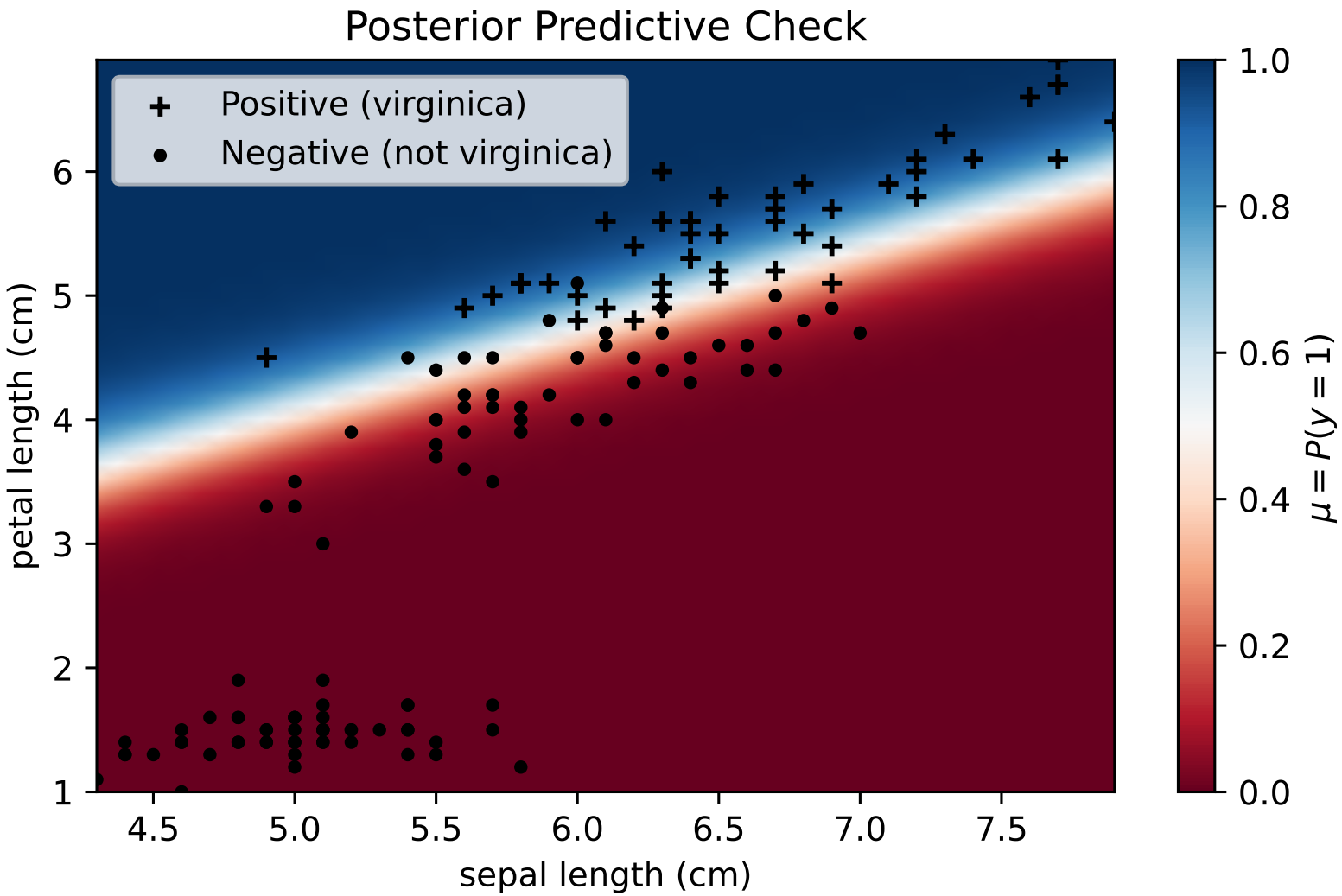
GLM - Ví dụ 2 (tt)


Bayesian Inference
Positive (virginica)
Negative (not virginica)
petal length (cm)
sepal length (cm)
1
2
3
4
5
6
7
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

GLM - Ví dụ 2 (tt)


beta0
mean=-3.9
95% HDI
-7.2
-0.64
−10
−8
−6
−4
−2
0
2
betaj
0
mean=-3.3
95% HDI
-4.8
-2
−6
−5
−4
−3
−2
−1
betaj
1
mean=5
95% HDI
3.4
6.7
3
4
5
6
7
8

GLM - Ví dụ 2 (tt)


Posterior Predictive Check
Positive (virginica)
Negative (not virginica)
petal length (cm)
sepal length (cm)
1
2
3
4
5
6
4.5
5.0
5.5
6.0
6.5
7.0
7.5
1.0
0.8
0.6
0.4
0.2
0.0
$\mu = P(y = 1)$

Tài liệu tham khảo

Jupyter Notebook đi kèm.

John K. Kruschke. *Doing Bayesian Data Analysis – A Tutorial with R, JAGS, and Stan*. Elsevier, 2015.

Osvaldo Martin. *Bayesian Analysis with Python*. Packt Publishing & Sons, 2024.

Chapter 4. Jochen Voss. *An Introduction to Statistical Computing - A Simulation-based Approach*. John Wiley & Sons, 2014.