

# Generalized Linear Models

Nguyễn Tư Thành Nhân, Bùi Nguyễn Đức Tân  
Nguyễn Thị Mai Anh, Châu Thanh Văn

PiMA 2021



Trình bày: Nhóm 7, GLM

August 8, 2021

## 1 Linear Regression

- Định nghĩa
- Ước lượng tham số

## 2 Generalized Linear Models

- Motivation
- Component
- Ước lượng tham số
- Đánh giá mô hình

## 3 Count data

- Component
- Đánh giá mô hình
- Ví dụ minh họa

# Contents

## 1 Linear Regression

- Định nghĩa
- Ước lượng tham số

## 2 Generalized Linear Models

- Motivation
- Component
- Ước lượng tham số
- Đánh giá mô hình

## 3 Count data

- Component
- Đánh giá mô hình
- Ví dụ minh họa



# Bài toán mở đầu

**Bài toán.** Mối quan hệ giữa điểm trung bình với số học sinh, quỹ lớp và số tiết học của một lớp là gì?

$Y$ : điểm trung bình của lớp học.

$X_1$ : số học sinh,  $X_2$ : quỹ lớp,  $X_3$ : số tiết học.

Mô hình:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, i = 1, 2, \dots, n$$

$y_i, x_{i1}, x_{i2}, x_{i3}$  là giá trị của  $Y, X_1, X_2, X_3$  khi xét lớp thứ  $i$ .



# Định nghĩa

$Y$  là biến phản hồi,  $X_1, X_2, \dots, X_p$  là các biến giải thích.

Trong Linear Regression, ta cần xác định quan hệ tuyến tính giữa  $Y$  và  $X_1, X_2, \dots, X_p$  dựa vào bộ dữ liệu  $\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^n$ .

## Mô hình Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$



# Contents

## 1 Linear Regression

- Định nghĩa
- Ước lượng tham số

## 2 Generalized Linear Models

- Motivation
- Component
- Ước lượng tham số
- Đánh giá mô hình

## 3 Count data

- Component
- Đánh giá mô hình
- Ví dụ minh họa



# Ước lượng tham số

Ta quan tâm tới việc chọn  $\beta$  như thế nào thì **hợp lí** ?

- các biến  $\epsilon_i$  nhận giá trị càng nhỏ càng tốt  $\rightarrow$  Least Square.
- biết phân phối  $\rightarrow$  tối ưu hàm Likelihood (MLE)

Giả sử trong Linear Regression

$$Y \mid X = \mathbf{x} \sim \mathcal{N}(\beta^\top \mathbf{x}, \sigma^2)$$



# Likelihood

Hàm Likelihood:

$$L(\beta) = \prod_n p(y_n | \mathbf{x}_n, \beta, \sigma^2)$$

Likelihood của mô hình Linear Regression

$$L(\beta) = \prod_n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_n - \beta^\top \mathbf{x}_n)^2\right)$$





# Log-likelihood

Hàm Log-likelihood:

$$\text{LL}(\beta) = \sum_n \left[ \left( \log \frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2} (y_n - \beta^\top \mathbf{x}_n)^2 \right]$$

Log-likelihood của mô hình Linear Regression

$$\text{LL}(\beta) = \frac{1}{2\sigma^2} \sum_n (y_n - \beta^\top \mathbf{x}_n)^2 - \frac{n}{2} \log(2\pi) - n \log \sigma$$



# Maximum Likelihood Estimation

Ta cần tính

$$\nabla_{\beta} LL(\hat{\beta}) = 0$$

Mặt khác

$$\nabla_{\beta} LL(\beta) = -\frac{1}{\sigma^2} (\mathbf{X}^{\top} \mathbf{X} \beta - \mathbf{X}^{\top} \mathbf{y})$$

MLE trong bài toán Linear Regression

$$\hat{\beta} = (\mathbf{X}^{\top} \mathbf{X})^{-1} \mathbf{X}^{\top} \mathbf{y}$$



# Contents

## 1 Linear Regression

- Định nghĩa
- Ước lượng tham số

## 2 Generalized Linear Models

- **Motivation**
- Component
- Ước lượng tham số
- Đánh giá mô hình

## 3 Count data

- Component
- Đánh giá mô hình
- Ví dụ minh họa



# Motivation

## Ví dụ 1

Dự đoán mong muốn học tiếp sau đại học của sinh viên Việt Nam



# Motivation

## Ví dụ 1

Dự đoán mong muốn học tiếp sau đại học của sinh viên Việt Nam

## Ví dụ 2

Dự đoán số ca dương tính với Covid-19 tại thành phố Hồ Chí Minh



# Contents

## 1 Linear Regression

- Định nghĩa
- Ước lượng tham số

## 2 Generalized Linear Models

- Motivation
- **Component**
- Ước lượng tham số
- Đánh giá mô hình

## 3 Count data

- Component
- Đánh giá mô hình
- Ví dụ minh họa



# Giới thiệu GLM

- Là mô hình thống kê.
- Dùng trong bài toán hồi quy lẫn phân loại.

Mối quan hệ giữa biến phản hồi và biến giải thích

- 1 Phân phối ? (LR : phân phối chuẩn  $\rightarrow$  GLM : ?)
- 2 Biểu thức ? (LR : tuyến tính  $\rightarrow$  GLM : ?)



# Thành phần

- 1 Random Component (Thành phần ngẫu nhiên)
- 2 Linear Predictor (Dự đoán tuyến tính)
- 3 Link Function (Hàm liên kết)





# Exponential family

## Hàm xác suất

$$f(y; \theta, \phi) = \exp \left( \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right)$$

Trong đó,

$\theta$ : natural parameter

$\phi$ : dispersion parameter

$a(\phi)$ ,  $b(\theta)$ ,  $c(y, \phi)$  là các hàm cố định.



# Exponential family

## Hàm xác suất

$$f(y; \theta, \phi) = \exp \left( \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right)$$

Trong đó,

$\theta$ : natural parameter

$\phi$ : dispersion parameter

$a(\phi)$ ,  $b(\theta)$ ,  $c(y, \phi)$  là các hàm cố định.

## Phân phối chuẩn thuộc họ Exponential

$$f(y; \mu, \sigma^2) = \exp \left( \frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right)$$



# Random Component

Họ exponential chứa các phân phối quen thuộc

- Phân phối liên tục gồm Phân phối chuẩn, Gamma, ...
- Phân phối rời rạc gồm Phân phối Bernoulli, Poisson, ...

Trong GLM,  $Y|X = \mathbf{x}$  có phân phối thuộc họ exponential.



# Linear Predictor

Dự đoán tuyến tính

$$\eta_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}, i = 1 \dots n$$



# Link function

## Hàm liên kết

$$\eta_i = g(\mu_i)$$

thỏa mãn hai điều kiện:

- Hàm  $g$  đơn điệu
- Hàm  $g$  khả vi



# Tổng kết

Gọi  $Y$  là biến phản hồi,  $X = (X_1, X_2, \dots, X_p)$  là các biến giải thích.

## Mô hình

- 1  $Y \mid X = \mathbf{x}$  có phân phối thuộc Exponential family.
- 2  $Y \mid X = \mathbf{x}_i$  có kỳ vọng có liên hệ với biểu thức tuyến tính của các biến giải thích thông qua hàm liên kết

$$g(\mu_i) = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} = \boldsymbol{\beta}^T \mathbf{x}_i, i = \overline{1, n}.$$



# Contents

## 1 Linear Regression

- Định nghĩa
- Ước lượng tham số

## 2 Generalized Linear Models

- Motivation
- Component
- Ước lượng tham số
- Đánh giá mô hình

## 3 Count data

- Component
- Đánh giá mô hình
- Ví dụ minh họa



# Kết quả quan trọng

## General Likelihood

Gọi LL là hàm log-likelihood của phân phối với tham số  $\theta$ , khi đó

$$\mathbb{E} \left[ \frac{\partial \text{LL}}{\partial \theta} \right] = 0, \quad \mathbb{E} \left[ \frac{\partial^2 \text{LL}}{\partial \theta^2} \right] + \mathbb{E} \left[ \left( \frac{\partial \text{LL}}{\partial \theta} \right)^2 \right] = 0.$$

## Hệ quả (Đặc trưng trong thành phần ngẫu nhiên)

Với  $\mathbb{E}[y] := \mathbb{E}[Y \mid X = x]$  và  $\text{var}(y) := \text{var}(Y \mid X = x)$  thì

$$b'(\theta) = \mu = \mathbb{E}[y], \quad \text{var}(y) = b''(\theta)a(\phi).$$





# Likelihood

Hàm phân phối  $Y$  khi biết điều kiện  $X = \mathbf{x}_n$  có dạng như sau

$$f(y_n; \theta_n, \phi) = \exp \left( \frac{y_n \theta_n - b(\theta_n)}{a_n(\phi)} + c(y_n, \phi) \right)$$

Likelihood của mô hình Generalized Linear Models

$$L(\beta) = \prod \left( \frac{y_n \theta_n - b(\theta_n)}{a_n(\phi)} + c(y_n, \phi) \right)$$

Log-likelihood của mô hình Generalized Linear Models

$$LL(\beta) = \sum \left( \frac{y_n \theta_n - b(\theta_n)}{a_n(\phi)} + c(y_n, \phi) \right) = \sum LL_n(\beta)$$



# Đẳng thức Likelihood

Ta tính đạo hàm của Log-likelihood:

$$\frac{\partial \text{LL}_n}{\partial \beta_m} = \frac{(y_n - \mu_n)x_{nm}}{\text{var}[y_n]} \frac{\partial \mu_n}{\partial \eta_n}$$

## Đẳng thức Likelihood

$$\frac{\partial \text{LL}(\beta)}{\partial \beta_m} = \sum_n \frac{(y_n - \mu_n)x_{nm}}{\text{var}[y_n]} \frac{\partial \mu_n}{\partial \eta_n} = 0, \forall m$$



# Đẳng thức Likelihood

Gọi  $\mathbf{V}$  là ma trận đường chéo của các phương sai quan sát được,  
 $\mathbf{D}$  là ma trận đường chéo của các phần tử  $\frac{\partial \mu_n}{\partial \eta_n}$

## Đẳng thức Likelihood

$$\mathbf{X}^\top \mathbf{D} \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}$$

Trong đó,  $\boldsymbol{\beta}$  được nằm trong công thức  $\mu_n = g^{-1}(\boldsymbol{\beta}^\top \mathbf{x}_n)$



# Newton-Raphson method

Gọi

$$\mathbf{g}(\beta) = \nabla_{\beta} LL(\beta)$$

là gradient của  $LL(\beta)$ ,

$$\mathbf{H}(\beta) = \nabla_{\beta}^2 LL(\beta)$$

là ma trận Hessian của  $LL(\beta)$

Newton-Raphson method

$$\beta^{(t+1)} = \beta^{(t)} + \left( \mathbf{H}(\beta^{(t)}) \right)^{-1} \mathbf{g}^{(t)}$$



# Fisher-scoring method

Gọi  $\mathcal{J}$  là ma trận có các phần tử là

$$-\mathbb{E} \left[ \frac{\partial^2 \text{LL}_n}{\partial \beta_m \partial \beta_p} \right]$$

. Hay

$$\mathcal{J} = -\mathbb{E}[\mathbf{H}]$$

## Fisher-scoring method

$$\beta^{(t+1)} = \beta^{(t)} - \left( \mathcal{J}(\beta^{(t)}) \right)^{-1} \mathbf{g}^{(t)}$$



# Contents

## 1 Linear Regression

- Định nghĩa
- Ước lượng tham số

## 2 Generalized Linear Models

- Motivation
- Component
- Ước lượng tham số
- **Đánh giá mô hình**

## 3 Count data

- Component
- Đánh giá mô hình
- Ví dụ minh họa



# Đánh giá mô hình

Công thức tính (tổng) độ lệch của mô hình  $M_0$  với kì vọng  $\hat{\mu} = E[Y|\hat{\beta}_0]$

## Deviance

$$D(y, \hat{\mu}) = 2(\log(p(y|\hat{\beta}_s)) - \log(p(y|\hat{\beta}_0)))$$

trong đó,  $\hat{\beta}_0$  kí hiệu là fitted values của tham số trong mô hình  $M_0$ , còn  $\hat{\beta}_s$  là fitted tham số trong mô hình saturated.



# Đánh giá mô hình

## Ví dụ

Công thức tính sự sai lệch của Poisson distribution.

$$p(Y = y) = \frac{e^{-\mu} \mu^y}{y!}.$$

$$LL(\beta) = \sum_i ((y_i \log(\mu_i) - \mu_i))$$

trong đó  $\log(\mu_i) = \mathbf{x}_i^\top \beta$

Ta có công thức tính độ lệch là:

$$D = 2 \sum_i \left[ y_i \log \left( \frac{y_i}{\mu_i} \right) - (y_i - \mu_i) \right].$$





# Đánh giá mô hình

Công thức tính sự sai lệch của một số loại phân bố có thể xem ở bảng sau, trong đó  $i = 1, 2, \dots, n$ .

Phân phối	Công thức tính sự lệch
Chuẩn	$\sum (y_i - \hat{\mu}_i)^2$
Poisson	$2 \sum \{y \log(\frac{y_i}{\hat{\mu}_i}) - (y_i - \hat{\mu}_i)\}$
Nhị thức	$2 \sum \{y \log(\frac{y_i}{\hat{\mu}_i}) + (m - y_i) \log[\frac{m - y_i}{m - \hat{\mu}_i}]\}$
Gamma	$2 \sum \{-\log(\frac{y_i}{\hat{\mu}_i}) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\}$



# Đánh giá mô hình

The generalized Pearson  $\chi^2$  statistic

$$\chi^2 = \sum \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

trong đó,  $V(\hat{\mu}_i)$  là hàm ước lượng phương sai cho phân phối được dùng đến.



# Contents

## 1 Linear Regression

- Định nghĩa
- Ước lượng tham số

## 2 Generalized Linear Models

- Motivation
- Component
- Ước lượng tham số
- Đánh giá mô hình

## 3 Count data

- Component
- Đánh giá mô hình
- Ví dụ minh họa



# Thành phần

## Random Component

$Y|X = \mathbf{x}$  tuân theo phân phối Poisson

## Linear Predictor

$$\eta_i = \theta_i = \log(\mu_i) = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}$$

## Link function

Hàm liên kết được sử dụng ở đây còn gọi là Log Function vì  $\log(\mu) = \eta$ . Hàm nghịch đảo của nó là:  $\mu = \exp(\eta)$



# Contents

## 1 Linear Regression

- Định nghĩa
- Ước lượng tham số

## 2 Generalized Linear Models

- Motivation
- Component
- Ước lượng tham số
- Đánh giá mô hình

## 3 Count data

- Component
- Đánh giá mô hình
- Ví dụ minh họa



## Deviance

$$D = 2 \sum \left[ y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right].$$

## Pearson's chi-squared

$$\chi_p = \sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$



# Contents

## 1 Linear Regression

- Định nghĩa
- Ước lượng tham số

## 2 Generalized Linear Models

- Motivation
- Component
- Ước lượng tham số
- Đánh giá mô hình

## 3 Count data

- Component
- Đánh giá mô hình
- Ví dụ minh họa



Dataset: Australian Health Service Utilization Data 1977 - 1978

## Bài toán

Từ các thông số về tuổi tác (age), bệnh lý (illness), năng suất hoạt động (reduced), v.v. của đối tượng được khảo sát, dự đoán số lần khám bác sĩ của đối tượng trong 2 tuần vừa qua.





Biến  $y = \text{visits}$  trong dataset trên biểu thị số lần khám bác sĩ trong 2 tuần vừa qua.

Khảo sát số liệu của biến này, ta thu được:

$$E(y) = 0.3, \text{var}(y) = 0.64$$



Biến  $y = \text{visits}$  trong dataset trên biểu thị số lần khám bác sĩ trong 2 tuần vừa qua.

Khảo sát số liệu của biến này, ta thu được:

$$E(y) = 0.3, \text{var}(y) = 0.64$$

Chênh lệch giữa mean và variance đủ nhỏ nên mô hình Poisson Regression sẽ không bị ảnh hưởng quá nhiều bởi overdispersion.



Áp dụng mô hình Linear Regression (scikit-learn), Poisson Regression (scikit-learn) và GLM (statsmodels) trên dữ liệu có được, thực nghiệm cho kết quả sau:

Mô hình	Deviance
Linear Regression	0.824381
Poisson Regression	0.903828
GLM	0.879162



# References

- Lecture Statistics for Applications fall 2016 from MIT
- Peter K. Dunn · Gordon K. Smyth Generalized Linear Models With Examples in R
- P. Mccullagh J.A. Nelder (1989) Generalized linear models chapman hall

