COURSE

Hồi quy tuyến tính tổng quát

Logistic & Poisson

Lớp phân tích thống kê cơ bản

Khương Quỳnh Long
Hà Nội, 06-08/06/2020

Nội dung

- Tóm tắt hồi quy tuyến tính tổng quát
- Hồi quy Logistic
- Hồi quy Poisson
- Giải thích kết quả

Tóm tắt

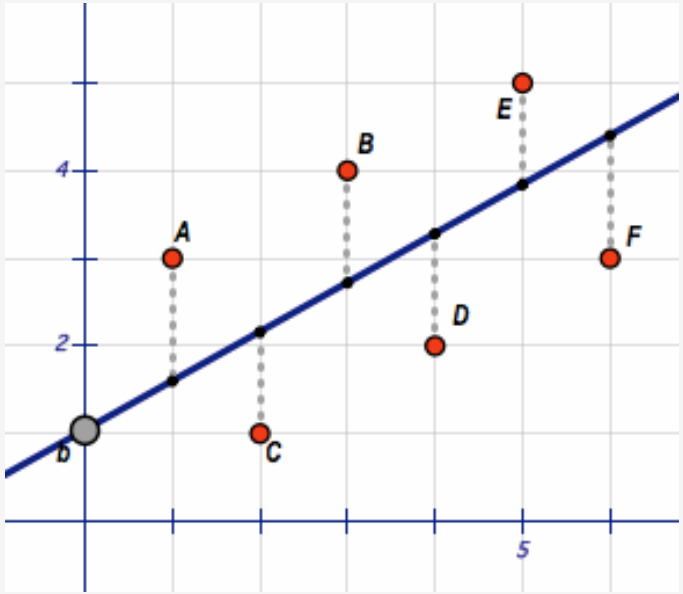
- Khi biến kết cuộc là

1. Định lượng liên tục (y = continuous): hồi quy tuyến tính
2. Nhị giá (y = binary): hồi quy logistic
3. Đếm (y = count/rate): hồi quy poisson
4. Sống còn (y = time to event): phân tích sống còn

Hồi quy tuyến tính tổng quát (GLM)

Nhắc lại hồi quy tuyến tính

- Mô tả mối liên hệ tuyến tính giữa biến kết cuộc là biến định lượng và các biến độc lập (định lượng, nhị giá, danh định…)
- $y = \alpha + \beta x$



Hồi quy tuyến tính tổng quát

- Hồi quy logistic thuộc "gia đình" hồi quy tuyến tính tổng quát (generalized linear model - GLM)

Hồi quy tuyến tính vs. tuyến tính tổng quát

- Hồi quy tuyến tính: $y = \alpha + \beta x$
- Tuyến tính tổng quát: $\eta = \alpha + \beta x$

$$\eta = g(\mu)$$

- g là hàm liên kết (link function) giữa thành phần không tuyến tính (μ) với phần tuyến tính $\alpha + \beta x$
- Cần thêm thông tin về phân phối ("*family*") của y

Hồi quy tuyến tính tổng quát

- Các hàm link (g): $\eta = \alpha + \beta x$

✓ Identity ($\eta = \mu$)
✓ log ($\eta = \log(\mu)$)
✓ logit ($\eta = \log(\mu/(1- \mu))$)
✓ Inverse ($\eta = \mu^{-1}$)

- Phân phối thuộc họ phân phối mũ “Exponential family”

✓ gaussian (pp chuẩn)
✓ binomial
✓ poisson
✓ gamma
✓ Nhị thức âm (negative binomial) & Weibull

Hồi quy Logistic

Hồi quy logistic

- Sử dụng khi biến kết cuộc là biến nhị giá

✓ Tử vong (có/không)

✓ CHD (có/không)

✓ Dung tích sống giảm (có/không) – phân biệt với dung tích sống (FEV) sử dụng trong bài hồi quy tuyến tính trước

- Thành phần của GLM

✓ link = logit

✓ family = binomial

Hồi quy logistic

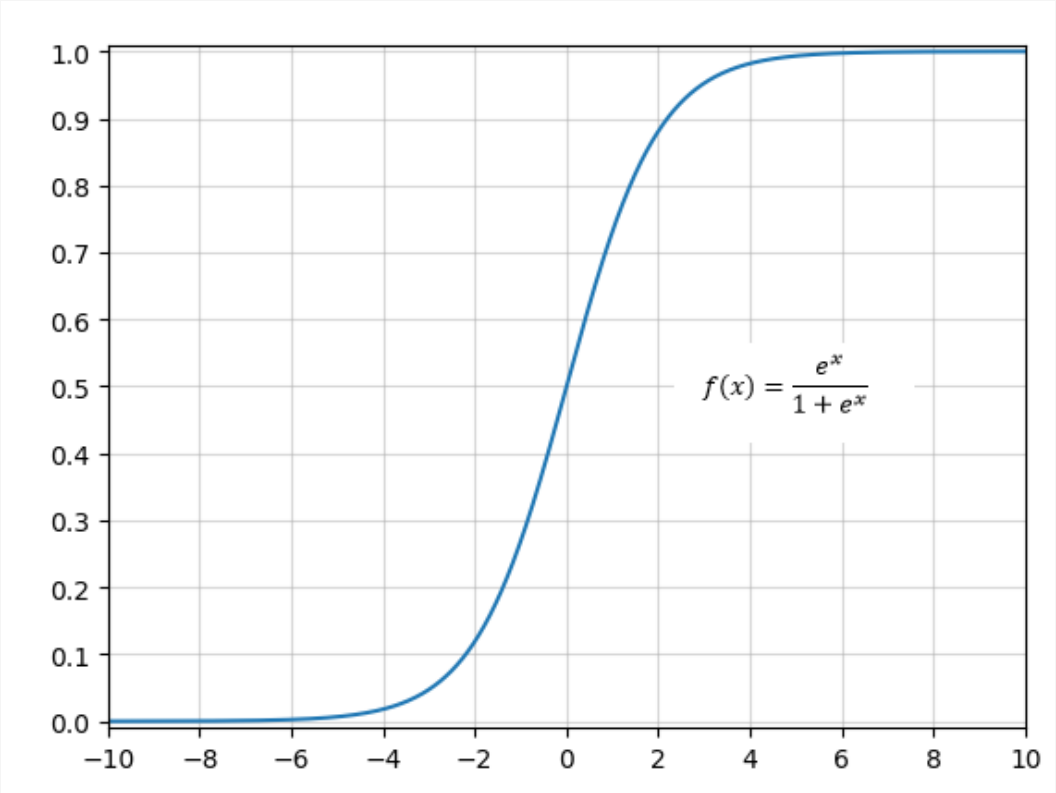
- Biến kết cuộc (0, 1)

Mô hình mở rộng của hồi quy logistic (i.e., multinomial logistic regression) có thể sử dụng cho trường hợp ≥ 2 nhóm

- Không thể sử dụng hồi quy tuyến tính cho biến kết cuộc là nhị giá vì:

✓ Xác suất dự đoán từ hồi quy tuyến tính vượt quá ngưỡng 0 – 1

✓ Xác suất dự đoán và các biến độc lập có mối liên hệ không tuyến tính


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−10
−8
−6
−4
−2
0
2
4
6
8
10
$f(x) = \frac{e^x}{1 + e^x}$

Hồi quy logistic

- Hồi quy tuyến tính:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

- Hồi quy logistic:

$$p = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

$$\log(p/(1-p)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Hồi quy logistic

- Hồi quy tuyến tính:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

- Hồi quy logistic:

$$\log(p/(1-p)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

✓ log = logarit tự nhiên (natural logarithm)

✓ p/(1-p) = Odds

✓ Xây dựng mô hình và giải thích dựa vào “log-odds”

Hồi quy logistic

$$\text{Logit}(p) = \text{Log}(\text{odds}) = \log(p/(1-p)) = \alpha + \beta * x$$

- Log(odds) của x = 0: $\text{Log}(\text{odds}_0) = \alpha + \beta * 0 = \alpha$
- Log(odds) của x = 1: $\text{Log}(\text{odds}_1) = \alpha + \beta * 1 = \alpha + \beta$
- So sánh 2 giá trị của x {0, 1}
- $\text{Log}(\text{odds}_1) - \text{Log}(\text{odds}_0) = \text{Log}(\text{odds}_1 / \text{odds}_0) = \alpha + \beta - \alpha = \beta$
- Log(OR) = β, hay OR = e^{β}

Stata

- **Cách 1**

#1: Trình bày OR

```
logistic biếnphụthuộc biếnđộclập
```

#2: Trình bày hệ số β [log(OR)]

```
logit biếnphụthuộc biếnđộclập
```

- **Cách 2**

#1: Trình bày OR

```
glm biếnphụthuộc biếnđộclập, link(logit) family(binomial) eform
```

#2: Trình bày hệ số β [log(OR)]

```
glm biếnphụthuộc biếnđộclập, link(logit) family(binomial)
```

WCGS data

- Nghiên cứu Western Collaborative Group Study (Rosenman et al. 1964) là một nghiên cứu dịch tế học nghiên cứu về mối liên qian giữa các hành vi và bệnh mạch vành (Coronary heart disease – CHD)
- Mục đích bài thực hành:

Xác định mối liên hệ của biến CHD với biến độc lập là

- ✓ Biến nhị giá: Nhóm cân nặng (wt2)
- ✓ Biến liên tục: cân nặng (weight)
- ✓ Biến phân nhóm (≥2 nhóm): nhóm cân nặng (4 nhóm) (wt4)

CHD vs. wt2

- Tạo biến nhóm cân nặng (wt2) với 2 giá trị: ≥170 lb và <170 lb
- Nhóm <170 lb làm nhóm chứng
- Xây dựng hồi quy logistic xác định mối liên hệ giữa CHD và nhóm cân nặng

CHD vs. wt2

- Hồi quy tuyến tính

$y = \alpha + \beta x$ ➔ y là mean của biến liên tục

- Hồi quy logistic

$\log(p/(1-p)) = \alpha + \beta x$ ➔ y là log(odds)

Hay:

$\log(odds_{CHD}) = \log(p_{CHD}/(1 - p_{CHD}) = \alpha + \beta * wt2$

CHD vs. wt2 – hệ số β

```
logit chd wt2
```

```
Iteration 0:   log likelihood = -890.62187
Iteration 1:   log likelihood = -882.41775
Iteration 2:   log likelihood = -882.31075
Iteration 3:   log likelihood = -882.31073

Logistic regression                             Number of obs     =      3,154
                                                LR chi2(1)        =      16.62
                                                Prob > chi2       =     0.0000
Log likelihood = -882.31073                     Pseudo R2         =     0.0093

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wt2 |   .5323229   .1311638     4.06   0.000     .2752465    .7893993
       _cons |  -2.685376   .0972268   -27.62   0.000    -2.875937   -2.494815
------------------------------------------------------------------------------
```

$\log(\text{odds}_{CHD}) = -2.69 + 0.53*\text{wt2}$

(với wt2, 1 ~ >170 lb, 0 ~ <170lb)

CHD vs. wt2 – Odds ratio (OR)

```
logistic chd wt2
```

```
Logistic regression                             Number of obs     =      3,154
                                                LR chi2(1)        =      16.62
                                                Prob > chi2       =     0.0000
Log likelihood = -882.31073                     Pseudo R2         =     0.0093

------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wt2 |   1.702883   .2233567     4.06   0.000     1.316855    2.202073
       _cons |   .0681955   .0066304   -27.62   0.000     .0563633    .0825117
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

- Odds CHD của nhóm >170 lb **gấp 1.70 lần** so với nhóm <170 lb
- Odds CHD của nhóm >170 lb **cao hơn 70%** so với nhóm <170 lb

Đối với biến đôc lập là biến liên tục (β)

- Xây dựng hồi quy logistic xác định mối liên hệ giữa CHD và cân nặng weight

logit chd weight

```
Iteration 0:   log likelihood = -890.62187
Iteration 1:   log likelihood = -884.55455
Iteration 2:   log likelihood = -884.46878
Iteration 3:   log likelihood = -884.46876

Logistic regression                             Number of obs     =      3,154
                                                LR chi2(1)        =      12.31
                                                Prob > chi2       =     0.0005
Log likelihood = -884.46876                     Pseudo R2         =     0.0069

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   .0104242   .0029196     3.57   0.000     .0047019    .0161465
       _cons |  -4.214706   .5120636    -8.23   0.000    -5.218333   -3.21108
------------------------------------------------------------------------------
```

$\log(\text{odds}_{CHD}) = -4.21 + 0.01*\text{weight}$

Cân nặng tăng 1 lb ➔ log-odds CHD tăng thêm 0.01 "**đơn vị**"

CHD vs. weight (OR)

```
logistic chd weight
```

```
Logistic regression                               Number of obs     =      3,154
                                                  LR chi2(1)        =      12.31
                                                  Prob > chi2       =     0.0005
Log likelihood = -884.46876                       Pseudo R2         =     0.0069

------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   1.010479   .0029502     3.57   0.000     1.004713    1.016278
       _cons |   .0147767   .0075666    -8.23   0.000     .0054164     .040313
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

$OR = e^{0.01} = 1.01$

"Với mỗi lb tăng thêm ➔ Odds CHD tăng thêm 1.01 **lần** (1%)"

Đối với biến độc lập là biến phân nhóm

- Nhóm cân nặng (wt4)

✓ 1 ~ <160 lb

✓ 2 ~ 160- <180 lb

✓ 3 ~ 180- < 200 lb

✓ 4 ~ ≥ 200 lb

CHD vs. wt4 (β)

```
logit chd i.wt4
```

```
Logistic regression                             Number of obs     =      3,154
                                                LR chi2(3)        =      14.08
                                                Prob > chi2       =     0.0028
Log likelihood = -883.58309                     Pseudo R2         =     0.0079

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wt4 |
160-<180 lb  |   .4891742   .1764166     2.77   0.006      .143404    .8349444
180-<200 lb  |   .5870319    .193432     3.03   0.002     .2079121    .9661518
    ≥200 lb  |   .7448789   .244718     3.04   0.002     .2652404    1.224517
             |
       _cons |  -2.841415   .1454884   -19.53   0.000    -3.126567   -2.556263
------------------------------------------------------------------------------
```

- Nhóm 160-180 lb có log-odds CHD cao hơn nhóm <160 lb 0.489 “đơn vị”
- Nhóm 180-200 lb có log-odds CHD cao hơn nhóm <160 lb 0.587 “đơn vị”
- Nhóm ≥200 lb có log-odds CHD cao hơn nhóm <160 lb 0.745 “đơn vị”

CHD vs. wt4 (OR)

```
logistic chd i.wt4
```

```
Logistic regression                             Number of obs     =      3,154
                                                LR chi2(3)        =      14.08
                                                Prob > chi2       =     0.0028
Log likelihood = -883.58309                     Pseudo R2         =     0.0079

------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          wt4 |
160-<180 lb  |   1.630969   .28773     2.77   0.006     1.154196    2.304686
180-<200 lb  |   1.798642   .347915     3.03   0.002     1.231105    2.627812
    ≥200 lb  |   2.106186   .5154218    3.04   0.002     1.303744    3.402524
             |
       _cons |   .0583431   .0084882   -19.53  0.000     .0438681    .0775942
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

$OR_{160\text{-}180 \text{ vs. } <160} = e^{0.489} = 1.63$

$OR_{180\text{-}200 \text{ vs. } <160} = e^{0.587} = 1.80$

$OR_{\geq 200 \text{ vs. } <160} = e^{0.745} = 2.10$

Tóm tắt

- Giới thiệu hồi quy tuyến tính tổng quát và logistic
- Diễn giải kết quả hồi quy logistic với biến độc lập là

✓ Biến nhị giá: Nhóm cân nặng (wt2)

✓ Biến liên tục: cân nặng (weight)

✓ Biến phân nhóm (≥2 nhóm): nhóm cân nặng (4 nhóm)

Hồi quy Poisson

- Khi biến kết cuộc là

1. Định lượng liên tục (y = continuous): hồi quy tuyến tính
2. Nhị giá (y = binary): hồi quy logistic
3. Đếm (y = count/rate): hồi quy poisson
4. Sống còn (y = time to event): phân tích sống còn

Hồi quy Poisson

- Sử dụng khi biến kết cuộc là biến đếm

✓ Số người tử vong

✓ Số người mắc/tháng

- Thành phần của GLM

✓ link = log

✓ family = poisson

Hồi quy Poisson

Số người tử vong ~ Poisson(λ)

λ [0; +∞)

➔ log(λ) (-∞; +∞)

$$\log(\lambda) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Hồi quy Poisson cho biến nhị phân

- Sử dụng hồi quy Poisson cho biến nhị phân (modified poisson regression)

$$Log(\pi) = \alpha + \beta * x$$

- π chính là p (xác suất/nguy cơ) nhưng được "tuân theo" phân phối Poisson [0; +∞) ➔ log(π) (-∞; +∞)
- Với x = 0 ➔ $log(risk_0) = \alpha + \beta*0 = \alpha$
- Với x = 1 ➔ $log(risk_1) = \alpha + \beta*1 = \alpha + \beta$
- So sánh các giá trị của x

$log(risk_1) - log(risk_0) = log(risk_1/risk_0) = log(RR) = \alpha + \beta - \alpha = \beta$

Hay $RR = e^{\beta}$

Hồi quy Poisson cho biến nhị phân

- Vấn đề!
- Phân phối Poisson:

$E(X) = \lambda$; $Var(X) = \sigma^2 = \lambda$ (trung bình = phương sai)

→ Var càng tăng khi λ tăng

- Phân phối Binomial:

$E(X) = n*p$; $Var(X) = n*p*(1 - p)$ ➔ p

→ Var lớn nhất khi p = 0.5

➔ Khi áp dụng hồi quy poisson cho biến nhị phân, sai số thường bị overestimated ➔ cần áp dụng các phương pháp hiệu chỉnh phương sai

Hồi quy Poisson cho biến nhị phân

- Một số phương pháp hiệu chỉnh phương sai
  - ✓ **Robust (Sandwich) variance**
  - ✓ Scale theo χ^2 → quasi-poisson
  - ✓ Scale theo deviance

Hồi quy Poisson (tóm tắt)

- Biến kết cuộc là biến count/rate
- Áp dụng hồi quy poisson cho biến nhị phân
  - ✓ Biến kết cuộc là biến nhị phân (nhưng “tuân theo” poisson)
  - ✓ Dùng để tính RR hoặc PR
  - ✓ Sai số bị overestimated → hiệu chỉnh phương sai
    - ✓ **Robust**
    - ✓ Chi^2
    - ✓ Deviance

Hồi quy Poisson cho biến nhị phân

- Stata

✓ Hiệu chỉnh phương sai Robust

Cách #1: **poisson biếnphụthuộc biếnđộclập, robust irr**

Cách #2: **glm biếnphụthuộc biếnđộclập, link(log) family(poisson) eform robust**

✓ Hiệu chỉnh phương sai theo Chi^2

Cách #1: **poisson biếnphụthuộc biếnđộclập, scale(x2)irr**

Cách #2: **glm biếnphụthuộc biếnđộclập, link(log) family(poisson) eform scale(x2)**

WCGS data

- Mục đích bài thực hành:

Xác định mối liên hệ của biến CHD với biến độc lập là

✓ Biến nhị giá: Nhóm cân nặng (wt2)

✓ Biến liên tục: cân nặng (weight)

✓ Biến phân nhóm (≥2 nhóm): nhóm cân nặng (4 nhóm) (wt4)

CHD vs. wt2

- Tạo biến nhóm cân nặng (wt2) với 2 giá trị: ≥170 lb và <170 lb
- Nhóm <170 lb làm nhóm chứng
- Xây dựng hồi quy poisson xác định mối liên hệ giữa CHD và nhóm cân nặng

CHD vs weight

CHD vs wt4

Bài tập

- Sử dụng data “tlsosinh.dta”
- Tìm các yếu tố ảnh hưởng tới tình trạng nhẹ cân của trẻ (phân tích đơn biến)