Bài giảng 1: Giới Thiệu Mô Hình Tuyến Tính Tổng Quát

TS. Tô Đức Khánh

Khoa Toán-Tin Học, Trường Đại Học Khoa Học Tự Nhiên
Đại Học Quốc Gia Tp. HCM

Mô Hình Thống Kê Tuyến Tính Nâng Cao
–Cao học Khóa 33–

Mục lục

1 Giới thiệu

2 Hàm hợp lý

3 Ước lượng hợp lý cực đại

4 Tính chất tiệm cận của MLE

5 Thống kê suy luận

Ví dụ 1 - Dung tích phổi

Ta xét bộ dữ liệu về dung tích phổi được thu thập trên 654 người trẻ tuổi tại East Boston, Mỹ, như sau:

id	Age	FEV	Ht	Gender	Smoke
1	3	1.072	46	F	0
2	4	0.839	48	F	0
3	4	1.102	48	F	0
4	4	1.389	48	F	0
5	4	1.577	49	F	0
6	4	1.418	49	F	0
⋮	⋮	⋮	⋮	⋮	⋮
654	18	4.404	70.5	M	1

trong đó:

- Age, Ht, Gender và Smoke lần lượt là các biến đo độ tuổi, chiều cao (inch), giới tính và tình trạng hút thuốc;
- FEV (forced expiratory volume) là thể tích (đơn vị: lít) thở ra cưỡng bức được dùng để chỉ số đo dung tích phổi.

Ví dụ 1 - Dung tích phổi

id	Age	FEV	Ht	Gender	Smoke
1	3	1.072	46	F	0
2	4	0.839	48	F	0
3	4	1.102	48	F	0
4	4	1.389	48	F	0
5	4	1.577	49	F	0
6	4	1.418	49	F	0
⋮	⋮	⋮	⋮	⋮	⋮
654	18	4.404	70.5	M	1

Trong các biến trên:

- Age, FEV và Ht là các biến ngẫu nhiên liên tục;
- Gender và Smoke là các biến ngẫu nhiên định danh,

tại sao?

Ví dụ 1 - Dung tích phổi

Câu hỏi nghiên cứu

Liệu rằng dung tích phổi có thay đổi theo độ tuổi, chiều cao, giới tính hay trạng thái hút thuốc?

Ví dụ 1 - Dung tích phổi

Câu hỏi nghiên cứu

Liệu rằng dung tích phổi có thay đổi theo độ tuổi, chiều cao, giới tính hay trạng thái hút thuốc?

Để trả lời câu hỏi nghiên cứu này, trước hết ta sẽ dùng biểu đồ để biểu diễn mối liên hệ giữa khả năng dung tích phổi (FEV) và các biến: Age, Ht, Gender, Smoke.

Ví dụ 1 - Dung tích phổi

Câu hỏi nghiên cứu

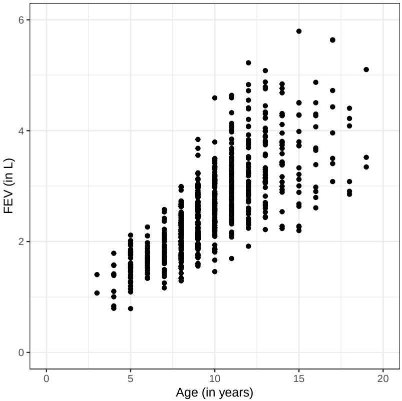
Liệu rằng dung tích phổi có thay đổi theo độ tuổi, chiều cao, giới tính hay trạng thái hút thuốc?

Để trả lời câu hỏi nghiên cứu này, trước hết ta sẽ dùng biểu đồ để biểu diễn mối liên hệ giữa khả năng dung tích phổi (FEV) và các biến: Age, Ht, Gender, Smoke.

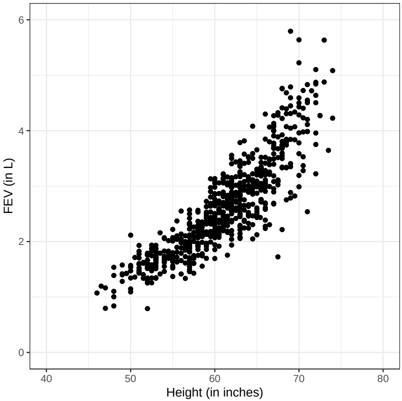
Đồ thị nào phù hợp để biểu diễn mối liên hệ của:

- **Age vs. FEV**
- **Ht vs. FEV**
- **Gender vs. FEV**
- **Smoke vs. FEV**

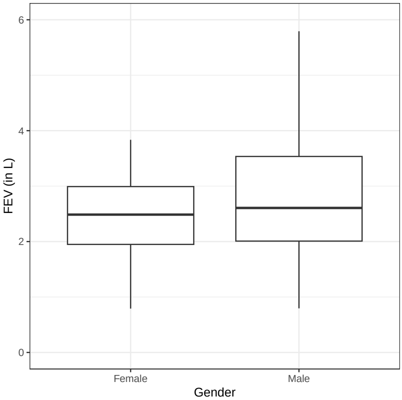
Ví dụ 1 - Dung tích phổi



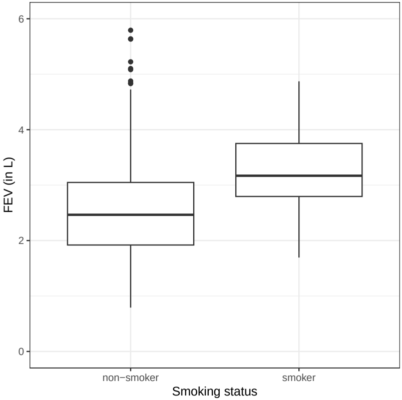
Ví dụ 1 - Dung tích phổi



Ví dụ 1 - Dung tích phổi



Ví dụ 1 - Dung tích phổi



Ví dụ 1 - Dung tích phổi

Tổng kết lại,

- độ tuổi (Age) và dung tích phổi (FEV) có sự tương quan tuyến tính (một cách tương đối, đặc biệt từ 1 tới 15 tuổi);
- sự tương quan giữa chiều cao (Ht) và dung tích phổi (FEV) là rất rõ ràng, tuy nhiên, sự tương quan có thể là phi tuyến;
- dung tích phổi (FEV) ở nam giới, nhìn chung, là cao hơn nữ giới, tuy nhiên, sự cao hơn là không thực sự rõ ràng;
- giá trị FEV ở những người hút thuốc (Smoke $= 1$) có xu hướng cao hơn ở những người không hút thuốc (Smoke $= 0$).

Ví dụ 1 - Dung tích phổi

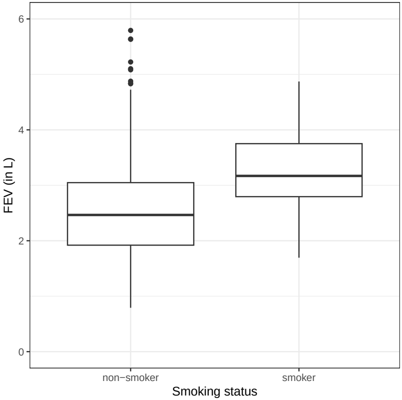
Tổng kết lại,

- độ tuổi (Age) và dung tích phổi (FEV) có sự tương quan tuyến tính (một cách tương đối, đặc biệt từ 1 tới 15 tuổi);
- sự tương quan giữa chiều cao (Ht) và dung tích phổi (FEV) là rất rõ ràng, tuy nhiên, sự tương quan có thể là phi tuyến;
- dung tích phổi (FEV) ở nam giới, nhìn chung, là cao hơn nữ giới, tuy nhiên, sự cao hơn là không thực sự rõ ràng;
- giá trị FEV ở những người hút thuốc (Smoke $= 1$) có xu hướng cao hơn ở những người không hút thuốc (Smoke $= 0$).

Tuy nhiên, những nhận xét này mới cho chúng ta góc nhìn về sự tương quan của một biến tới FEV.

Ví dụ 1 - Dung tích phổi

Giá trị FEV ở những người hút thuốc (Smoke = 1) có xu hướng cao hơn ở những người không hút thuốc (Smoke = 0). **Tại sao? Liệu rằng việc hút thuốc giúp tăng dung tích phổi?**



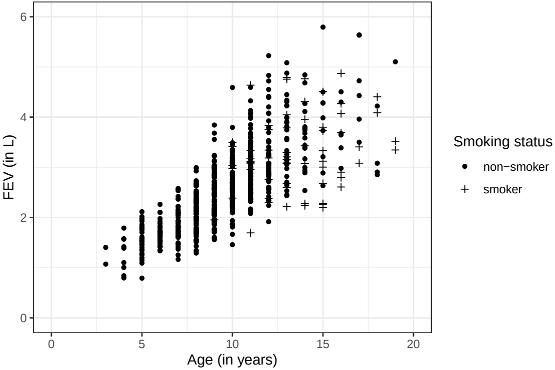
Ví dụ 1 - Dung tích phổi

Để tìm hiểu nguyên nhân tại sao giá trị FEV ở những người hút thuốc có xu hướng cao hơn ở những người không hút thuốc, ta sẽ tìm hiểu sự tương quan của các biến tới FEV dựa trên hai mẫu con:

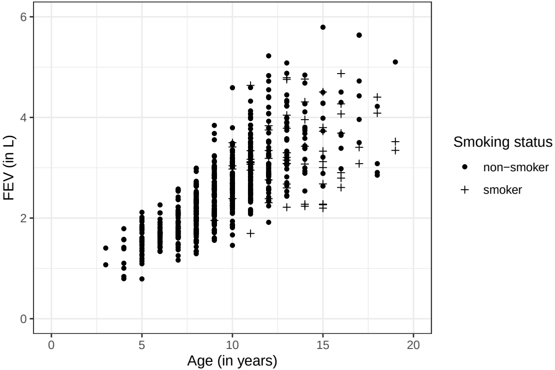
- không hút thuốc (Smoke $= 0$);
- hút thuốc (Smoke $= 1$).

Ta sẽ tập trung vào hai biến Age và Ht, bởi sự tương quan của chúng tới FEV đã được kiểm chứng.

Ví dụ 1 - Dung tích phổi

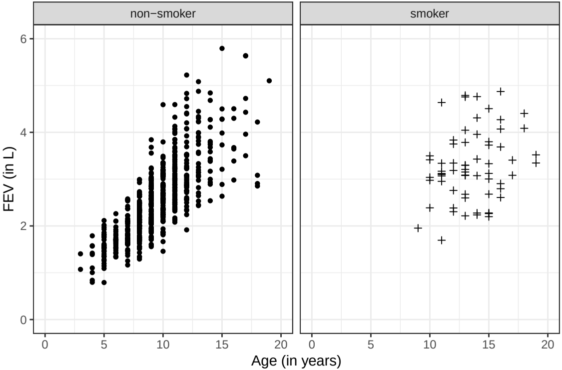


Ví dụ 1 - Dung tích phổi

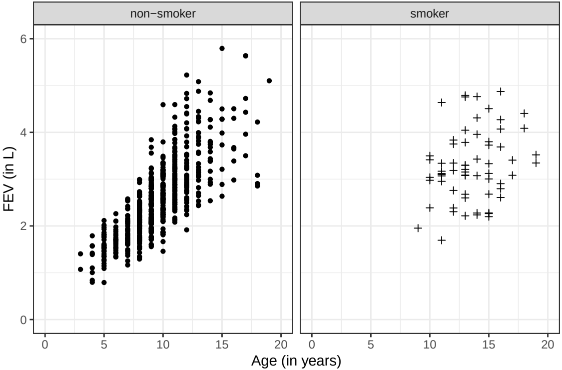


- những người hút thuốc có độ tuổi lớn (ít nhất 10);
- quan sát vùng dữ liệu từ 10 tuổi trở lên, ta thấy rằng giá trị FEV là không thực sự khác biệt giữa những người hút thuốc hoặc không hút thuốc.

Ví dụ 1 - Dung tích phổi

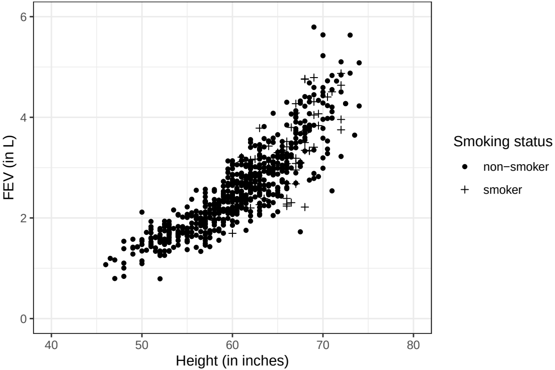


Ví dụ 1 - Dung tích phổi

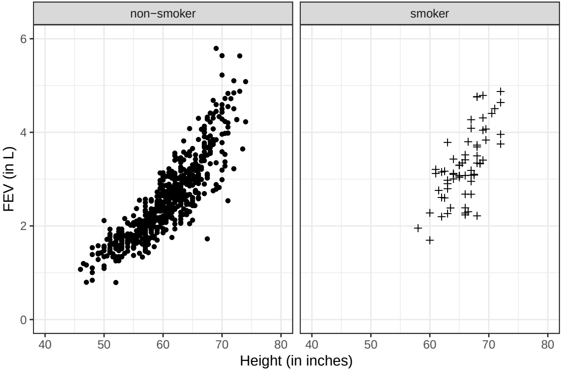


Hai đồ thị cho thấy việc giá trị FEV của những người hút thuốc cao hơn so với những người không hút thuốc là do độ tuổi!

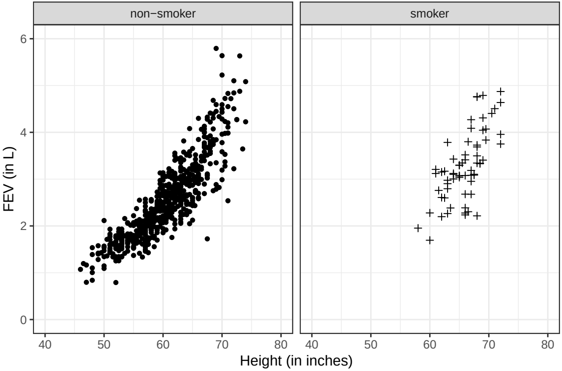
Ví dụ 1 - Dung tích phổi



Ví dụ 1 - Dung tích phổi



Ví dụ 1 - Dung tích phổi



Các đồ thị cho thấy việc giá trị FEV của những người hút thuốc cao hơn so với những người không hút thuốc là do chiều cao!

Ví dụ 1 - Dung tích phổi

Như vậy, về mặt tổng quan, các biến Age, Ht và Smoke có tương quan hay nói cách khác, là có ảnh hưởng hay tác động tới FEV.

Tuy nhiên, để có thể định lượng được sự tác động đó, về mặt toán học, ta cần thiết lập một *mô hình thống kê*.

Mô hình thống kê

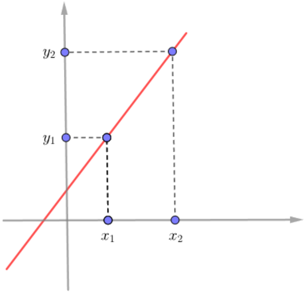
Xét phương trình đường thẳng:

$$y = a + bx$$

với a và b là các hằng số được biết trước.

- Cho trước một giá trị x_1 ta dễ dàng tính được tương ứng một giá trị y_1.

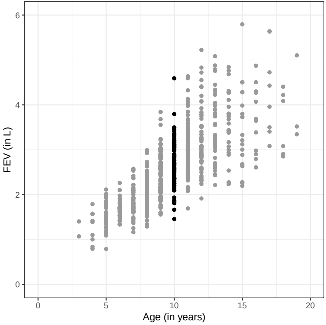
$\hookrightarrow$ Phương trình $y = a + bx$ là một dạng **mô hình toán** mô tả sự thay đổi giá trị y bởi giá trị của x.



Về mặt mô hình toán, với một giá trị của x ta sẽ quan sát được một giá trị của y, tuy nhiên, đối với dữ liệu quan sát thực tế thì không như vậy.

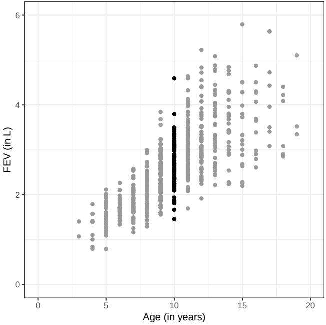
Mô hình thống kê

Một giá trị của Age (x) có thể ghi nhận nhiều giá trị khác nhau của FEV (y). Ví dụ Age $=$ 10 (các điểm màu đen).

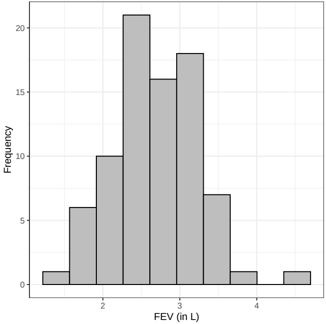


Mô hình thống kê

Một giá trị của Age (x) có thể ghi nhận nhiều giá trị khác nhau của FEV (y). Ví dụ Age $=$ 10 (các điểm màu đen).



Những quan sát FEV của các đối tượng 10 tuổi là ngẫu nhiên, và do đó, chúng có phân phối (*phân phối có điều kiện*).



Mô hình thống kê

Do đó, một mô hình toán cho FEV dựa vào Age, chẳng hạn:

$$\text{FEV} = a + b \times \text{Age},$$

sẽ chỉ có thể mô tả được giá trị trung bình của FEV tương ứng của một giá trị Age, mà không thể mô tả được phân phối của FEV.

$\hookrightarrow$ Điều này đòi hỏi phải có thêm một thành phần trong mô hình để mô tả cho phân phối của FEV.

Mô hình thống kê

Mô hình thống kê

Một mô hình thống kê (*statistical model*) là một mô hình gồm hai thành phần:

- thành phần hệ thống (*systematic component*);
- thành phần ngẫu nhiên (*random component*),

mô tả lần lượt hai đặc trưng của biến đáp ứng: trung bình và phân phối, dựa vào giá trị của một hoặc nhiều biến giải thích.

Mô hình thống kê

Ví dụ 1: Mô hình thống kê cho FEV dựa vào các biến Age, Ht, Gender và Smoke có thể là:

- thành phần hệ thống:

$$\mu = \mathbb{E}(\text{FEV}) = \beta_0 + \beta_1\text{Age} + \beta_2\text{Ht} + \beta_3\text{Gender} + \beta_4\text{Smoke},$$

- thành phần ngẫu nhiên: FEV $\sim \mathcal{N}(\mu, \sigma^2)$;

trong đó, các tham số $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ và σ là chưa biết, cần phải ước lượng từ dữ liệu.

Mô hình thống kê

Ví dụ 1: Mô hình thống kê cho FEV dựa vào các biến Age, Ht, Gender và Smoke có thể là:

- thành phần hệ thống:

$$\mu = \mathbb{E}(\mathsf{FEV}) = \beta_0 + \beta_1 \mathsf{Age} + \beta_2 \mathsf{Ht} + \beta_3 \mathsf{Gender} + \beta_4 \mathsf{Smoke},$$

- thành phần ngẫu nhiên: FEV $\sim \mathcal{N}(\mu, \sigma^2)$;

trong đó, các tham số $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ và σ là chưa biết, cần phải ước lượng từ dữ liệu.

Cách viết mô hình như trên là cách viết dạng tổng quát. Khi ta có bộ dữ liệu với n quan sát độc lập, mô hình thống kê sẽ được biểu diễn cho quan sát thứ i, ví dụ:

- thành phần hệ thống:

$$\mu_i = \mathbb{E}(\mathsf{FEV}_i) = \beta_0 + \beta_1 \mathsf{Age}_i + \beta_2 \mathsf{Ht}_i + \beta_3 \mathsf{Gender}_i + \beta_4 \mathsf{Smoke}_i,$$

- thành phần ngẫu nhiên: $\mathsf{FEV}_i \sim \mathcal{N}(\mu_i, \sigma^2)$ (chúng ta đang giả định các quan sát FEV_i có phương sai đồng nhất).

Mô hình thống kê

Ví dụ 2: Một mô hình thống kê cho FEV có thể là,

- thành phần hệ thống:
$$\mu_i = \mathbb{E}(\mathsf{FEV}_i) = \beta_0,$$
- thành phần ngẫu nhiên: $\mathsf{FEV}_i \sim \mathcal{N}(\mu_i, \sigma^2)$.

Mô hình thống kê

Ví dụ 2: Một mô hình thống kê cho FEV có thể là,

- thành phần hệ thống:
$$\mu_i = \mathbb{E}(\mathsf{FEV}_i) = \beta_0,$$
- thành phần ngẫu nhiên: $\mathsf{FEV}_i \sim \mathcal{N}(\mu_i, \sigma^2)$.

Thông thường sẽ có nhiều dạng khác nhau cho thành phần hệ thống cũng như thành phần ngẫu nhiên.

Mô hình hồi quy

Mô hình hồi quy

Nếu ta giả định rằng thành phần hệ thống, hay trung bình μ_i là một hàm f của p biến giải thích với các tham số chưa biết, tức là

$$\mu_i = \mathbb{E}(y_i) = f(x_{1i}, \ldots, x_{pi}; \beta_0, \beta_1, \ldots, \beta_p),$$

khi đó mô hình thống kê sẽ được gọi là một mô hình hồi quy (*regression model*).

Mô hình hồi quy

Mô hình hồi quy

Nếu ta giả định rằng thành phần hệ thống, hay trung bình μ_i là một hàm f của *p* biến giải thích với các tham số chưa biết, tức là

$$\mu_i = \mathbb{E}(y_i) = f(x_{1i}, \ldots, x_{pi}; \beta_0, \beta_1, \ldots, \beta_p),$$

khi đó mô hình thống kê sẽ được gọi là một mô hình hồi quy (*regression model*).

Về mặt toán học, sẽ có rất nhiều sự kết hợp khác nhau của $x_{1i}, \ldots, x_{pi}$ và $\beta_0, \beta_1, \ldots, \beta_p$. Thông thường, ta giả sử rằng sự kết hợp này là tuyến tính, tức là

$$\mu_i = \mathbb{E}(y_i) = f(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}),$$

khi này, ta có mô hình hồi quy tuyến tính theo tham số.

Mô hình hồi quy

Ta có hai dạng mô hình tuyến tính như sau:

Mô hình hồi tuyến tính - Linear regression model: là một mô hình thống kê với

- thành phần hệ thống

$$\mu_i = \mathbb{E}(y_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi},$$

- thành phần ngẫu nhiên $\mathbb{V}\text{ar}(y_i) = \sigma^2$. Chú ý, chúng ta không cần đưa ra giả định cụ thể nào về phân phối.

Mô hình hồi quy

Ta có hai dạng mô hình tuyến tính như sau:

Mô hình hồi tuyến tính - Linear regression model: là một mô hình thống kê với

- thành phần hệ thống

$$\mu_i = \mathbb{E}(y_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi},$$

- thành phần ngẫu nhiên $\mathbb{V}\mathrm{ar}(y_i) = \sigma^2$. Chú ý, chúng ta không cần đưa ra giả định cụ thể nào về phân phối.

Mô hình hồi tuyến tính tổng quát - Generalized linear model: là một mô hình thống kê với

- thành phần hệ thống

$$\mu_i = \mathbb{E}(y_i) = g^{-1}(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}),$$

  trong đó, $g(\cdot)$ là một hàm được xác định sao cho đồng biến và khả vi, và được gọi là hàm liên kết (*link function*),
- thành phần ngẫu nhiên: y_i tuân theo một phân phối xác định F với trung bình μ_i.

Một số dạng có thể của thành phần hệ thống

Giả sử ta có biến đáp ứng là y với trung bình μ, các biến giải thích lần lượt là x_1, x_2, x_3 và x_4. Các dạng có thể của thành phần hệ thống là:

$$
\begin{aligned}
\mu &= \beta_0 + \beta_1 x_1 + \beta_4 x_4 && (1)\\
\mu &= \beta_0 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_4 && (2)\\
\mu &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 && (3)\\
\mu &= \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \beta_4 x_4 && (4)\\
\mu &= \beta_0 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_4 && (5)\\
1/\mu &= \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 && (6)\\
\log(\mu) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 && (7)\\
\mu &= \beta_0 + \exp(\beta_1 x_1) - \exp(\beta_2 x_2) + \beta_4 x_4^2 && (8)
\end{aligned}
$$

- Các phương trình từ (1) - (7) đều có dạng tuyến tính theo tham số.
- Các phương trình từ (1) - (5) có thể được sử dụng để chỉ định một mô hình quy tuyến tính.

Thành phần ngẫu nhiên

Thành phần ngẫu nhiên của GLM

Thành phần ngẫu nhiên (*random component*) của GLM bao hàm 1 biến phản hồi Y với các quan sát độc lập nhau $(y_1, y_2, \dots, y_n)$ có hàm mật độ xác suất hoặc hàm xác suất của một phân phối thuộc họ phân phối mũ phân tán (*exponential dispersion family*):

$$f_Y(y_i|\theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right),$$

trong đó,

- θ_i được gọi là tham số tự nhiên (*natural parameter*);
- $\phi > 0$ được gọi là tham số phân tán (*dispersion parameter*).

Thông thường,

- $a(\phi) = 1$ và $c(y_i, \phi) = c(y_i) \Rightarrow$ họ phân phối mũ tự nhiên (*natural exponential family*);
- $a(\phi) = \phi$ hoặc $a(\phi) = \phi/\omega_i$, với ω_i là trọng số đã biết.

Thành phần ngẫu nhiên

Một số phân phối thuộc họ phân phối mũ phân tán:

- phân phối Bernoulli, $\mathcal{B}(p)$, với $p \in (0,1)$;
- phân phối nhị thức, $\mathcal{B}(n,p)$ với n cố định và $p \in (0,1)$;
- phân phối multinomial, $\mathcal{M}(n; p_1, \ldots, p_k)$ với n cố định, $p_i \in (0,1)$ và $\sum_{i=1}^{k} p_i = 1$;
- phân phối Poisson, $\mathcal{P}(\lambda)$, $\lambda > 0$;
- phân phối chuẩn, $\mathcal{N}(\mu, \sigma^2)$, $\sigma > 0$;
- phân phối Gamma, $\mathcal{G}(\alpha, \beta)$

$$f_Y(y|\alpha, \beta) = \frac{1}{\Gamma(\alpha)} y^{\alpha - 1} \exp\left(-y\beta\right) \beta^{\alpha},$$

  với $y > 0$, và $\alpha, \beta > 0$;
- phân phối Beta, $\mathcal{B}e(\alpha, \beta)$

$$f_Y(y|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha - 1} (1 - y)^{\beta - 1},$$

  với $y \in [0,1]$, và $\alpha, \beta > 0$.

Hàm liên kết - Link function

Hàm liên kết - link function

Hàm liên kết (*Link function*) là một hàm đồng biến, được sử dụng để liên kết thành phần ngẫu nhiên với thành phần tuyến tính (*linear predictor*) của GLM:

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

với $i = 1, \ldots, n$.

Cụ thể, với thành phần ngẫu nhiên, ta có $\mu_i = \mathbb{E}(y_i) \Rightarrow$ liên kết giữa η_i và μ_i được biểu diễn bởi $\eta_i = g(\mu_i) \Rightarrow g(\cdot)$ được gọi là hàm liên kết (*link function*) với tính chất:

- đồng biến (monotonic);
- khả vi (differentiable).

Hàm liên kết - Link function

Một số hàm liên kết tương ứng với thành phần ngẫu nhiên:

- phân phối chuẩn, $\eta_i = \mu_i$, hay $g(\cdot)$ là hàm đồng nhất (identity link);
- phân phối nhị thức, $\eta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right)$, logit link;
- phân phối nhị thức, $\eta_i = -\log\left(-\log(\mu_i)\right)$, log-log link;
- phân phối Poisson, $\eta_i = \log(\mu_i)$, log link;
- phân phối Gamma, $\eta_i = \mu_i^{-1}$, inverse link.

Canonical link

Trong một số trường hợp khi phân phối mũ phân tán có trung bình trùng với tham số tự nhiên (*natural parameter*) thì hàm liên kết $g(\cdot)$ được gọi là liên kết chính tắc (*canonical link*). Ví dụ:

- phân phối chuẩn, $\eta_i = \mu_i$, hay $g(\cdot)$ là hàm đồng nhất (identity link);
- phân phối nhị thức, $\eta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right)$, logit link;
- phân phối Poisson, $\eta_i = \log(\mu_i)$, log link;
- phân phối Gamma, $\eta_i = \mu_i^{-1}$, inverse link.

Ví dụ cho GLM

Ví dụ 3: *Những tác nhân nào có thể làm gia tăng nguy cơ bị bệnh tim mạch vành (coronary heart disease)?* Để tìm ra câu trả lời, một nhóm các nhà nghiên cứu đã ghi lại trạng thái bệnh tim mạch vành của 3154 người đàn ông (với độ tuổi 39 tới 59, từ khu vực San Francisco) trong vòng 8 năm và 6 tháng. Đồng thời, họ cũng ghi lại những biến có khả năng ảnh hưởng tới khả năng bị bệnh tim mạch vành.

id	age	chd	weight	chol	cigs
1	49	no	150	225	25
2	42	no	160	177	20
3	42	no	160	181	0
4	41	no	152	132	20
5	59	yes	150	255	20
6	44	no	204	182	0
⋮	⋮	⋮	⋮	⋮	⋮
3154	39	no	155	264	40

Ví dụ cho GLM

- age: độ tuổi
- chd: trạng thái bị bệnh tim mạch vành (có - yes hoặc không - no)
- weight: cân nặng
- chol: hàm lượng cholesterol
- cigs: số lượng thuốc lá hút trong một ngày

Ví dụ cho GLM

- age: độ tuổi
- chd: trạng thái bị bệnh tim mạch vành (có - yes hoặc không - no)
- weight: cân nặng
- chol: hàm lượng cholesterol
- cigs: số lượng thuốc lá hút trong một ngày

Bởi vì trọng tâm của câu hỏi là về khả năng bị bệnh tim mạch vành, do đó, biến chd sẽ là biến đáp ứng.

Một số nghi vấn cho rằng, hút thuốc là nhiều, béo phì, và hàm lượng cholesterol cao, là những nguyên nhân tăng khả năng bị bệnh tim mạch vành.

Để làm rõ nghi vấn này, ta xem xét một mô hình hồi quy tuyến tính tổng quát như sau.

Ví dụ cho GLM

Với mỗi quan sát thứ i, ta thấy rằng, chd$_i$ chỉ có hai giá trị:

- no - không bị bệnh, ký hiệu là 0;
- yes - bị bệnh, ký hiệu là 1;

do đó, chd$_i$ có thể là một biến ngẫu nhiên Bernoulli với xác suất bị bệnh p_i.

Ví dụ cho GLM

Với mỗi quan sát thứ i, ta thấy rằng, chd$_i$ chỉ có hai giá trị:

- no - không bị bệnh, ký hiệu là 0;
- yes - bị bệnh, ký hiệu là 1;

do đó, chd$_i$ có thể là một biến ngẫu nhiên Bernoulli với xác suất bị bệnh p_i.

$\hookrightarrow$ thành phần ngẫu nhiên của mô hình sẽ là: chd$_i \sim \mathcal{B}\text{er}(p_i)$, và ta có $\mu_i = \mathbb{E}(\text{chd}_i) = p_i$.

Ví dụ cho GLM

Với mỗi quan sát thứ i, ta thấy rằng, chd_i chỉ có hai giá trị:

- no - không bị bệnh, ký hiệu là 0;
- yes - bị bệnh, ký hiệu là 1;

do đó, chd_i có thể là một biến ngẫu nhiên Bernoulli với xác suất bị bệnh p_i.

$\hookrightarrow$ thành phần ngẫu nhiên của mô hình sẽ là: $\text{chd}_i \sim \mathcal{B}\text{er}(p_i)$, và ta có $\mu_i = \mathbb{E}(\text{chd}_i) = p_i$.

thành phần hệ thống có thể là một kết hợp tuyến tính bao gồm các biến: age, weight, chol và cigs, được xác định như sau:

$$\eta_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{weight}_i + \beta_3 \text{chol}_i + \beta_4 \text{cigs}_i$$

Khi đó, ta có một hàm liên kết

$$\eta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right).$$

Hay ta có thể biểu diễn

$$\mu_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}.$$

Ví dụ cho GLM

Với mỗi quan sát thứ i, ta thấy rằng, chd$_i$ chỉ có hai giá trị:

- no - không bị bệnh, ký hiệu là 0;
- yes - bị bệnh, ký hiệu là 1;

do đó, chd$_i$ có thể là một biến ngẫu nhiên Bernoulli với xác suất bị bệnh p_i.

$\hookrightarrow$ thành phần ngẫu nhiên của mô hình sẽ là: chd$_i \sim \mathcal{B}er(p_i)$, và ta có $\mu_i = \mathbb{E}(\text{chd}_i) = p_i$.

thành phần hệ thống có thể là một kết hợp tuyến tính bao gồm các biến: age, weight, chol và cigs, được xác định như sau:

$$\eta_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{weight}_i + \beta_3 \text{chol}_i + \beta_4 \text{cigs}_i$$

Khi đó, ta có một hàm liên kết

$$\eta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right).$$

Hay ta có thể biểu diễn

$$\mu_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}.$$

$\hookrightarrow$ mô hình dạng này được gọi là mô hình hồi quy logistic (*logistic regression model*).

Biến ngẫu nhiên trong mô hình

Trong một mô hình thống kê, ta thường có những định danh như sau cho các biến ngẫu nhiên (bất kể dạng biến ngẫu nhiên):

Biến ngẫu nhiên trong mô hình

Trong một mô hình thống kê, ta thường có những định danh như sau cho các biến ngẫu nhiên (bất kể dạng biến ngẫu nhiên):

biến đáp ứng (response variable) biến ngẫu nhiên được quan tâm trong nghiên cứu, giá trị và sự biến động được diễn tả bởi mô hình thông qua những biến khác;

Biến ngẫu nhiên trong mô hình

Trong một mô hình thống kê, ta thường có những định danh như sau cho các biến ngẫu nhiên (bất kể dạng biến ngẫu nhiên):

biến đáp ứng (response variable) biến ngẫu nhiên được quan tâm trong nghiên cứu, giá trị và sự biến động được diễn tả bởi mô hình thông qua những biến khác;

biến giải thích (explanatory variable) biến được dùng để giải thích sự thay đổi của biến đáp ứng bởi mô hình;

Biến ngẫu nhiên trong mô hình

Trong một mô hình thống kê, ta thường có những định danh như sau cho các biến ngẫu nhiên (bất kể dạng biến ngẫu nhiên):

biến đáp ứng (response variable) biến ngẫu nhiên được quan tâm trong nghiên cứu, giá trị và sự biến động được diễn tả bởi mô hình thông qua những biến khác;

biến giải thích (explanatory variable) biến được dùng để giải thích sự thay đổi của biến đáp ứng bởi mô hình;

biến nhân tố (factor) là biến giải thích nhưng với dạng phân loại (định danh hoặc thứ bậc), biến này có tên khác là giả biến (*dummy variable*);

Biến ngẫu nhiên trong mô hình

Trong một mô hình thống kê, ta thường có những định danh như sau cho các biến ngẫu nhiên (bất kể dạng biến ngẫu nhiên):

biến đáp ứng (response variable) biến ngẫu nhiên được quan tâm trong nghiên cứu, giá trị và sự biến động được diễn tả bởi mô hình thông qua những biến khác;

biến giải thích (explanatory variable) biến được dùng để giải thích sự thay đổi của biến đáp ứng bởi mô hình;

biến nhân tố (factor) là biến giải thích nhưng với dạng phân loại (định danh hoặc thứ bậc), biến này có tên khác là giả biến (*dummy variable*);

biến gây nhiễu (compounding variable) là một dạng đặc biệt của biến giải thích, một biến giải thích được coi là gây nhiễu nếu sự xuất hiện của nó làm thay đổi sự tác động của một hoặc nhiều biến giải thích khác.

Biến ngẫu nhiên trong mô hình

Giả sử rằng ta muốn xây dựng một mô hình thống kê để giải thích sự thay đổi của FEV bởi sự tách động của Age (độ tuổi), Ht (chiều cao), Gender (giới tính) và Smoke (trạng thái hút thuốc). Khi đó,

- FEV là biến đáp ứng;
- Age, Ht, Gender và Smoke là các biến giải thích;
- Gender và Smoke là biến nhân tố.

Diễn giải mô hình

Các mô hình hữu ích nhất khi chúng có những diễn giải hợp lý.

So sánh hai thành phần hệ thống sau:

$$\mu = \beta_0 + \beta_1 x \tag{9}$$

$$\log(\mu) = \beta_0 + \beta_1 x \tag{10}$$

Ta có nhận xét

- mô hình (9): x tăng 1 đơn vị thì μ tăng β_1 đơn vị;
- mô hình (10): x tăng 1 đơn vị thì $\log(\mu)$ tăng β_1 đơn vị;

  $\hookrightarrow$ x tăng 1 đơn vị thì μ tăng $\exp(\beta_1)$ lần. Tại sao?

Trong ứng dụng, ta cần xem xét lựa chọn mô hình (thành phần hệ thống) sao cho phù hợp với thực tế, hơn là tập trung quá nhiều vào công thức toán. Ví dụ:

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i},$$

trong đó,

- x_{1i}: số năm kinh nghiệm trong công việc;
- x_{2i}: giới tính ($1 =$ nữ, $0 =$ nam).

Diễn giải mô hình

Hai mục tiêu chính của mô hình hồi quy là

Dự đoán: để tạo ra những dự đoán chính xác từ dữ liệu mới hoặc dữ liệu trong tương lai.

Hiểu và diễn giải: để hiểu các biến có liên quan với nhau như thế nào.

Diễn giải mô hình

Hai mục tiêu chính của mô hình hồi quy là

Dự đoán: để tạo ra những dự đoán chính xác từ dữ liệu mới hoặc dữ liệu trong tương lai.

Hiểu và diễn giải: để hiểu các biến có liên quan với nhau như thế nào.

Ví dụ, hãy xem xét nghiên cứu về dung tích phổi.

- Ta bắt đầu với một câu hỏi về xác định mối liên hệ giữa trạng thái hút thuốc, giới tính, độ tuổi, chiều cao với dung tích phổi (FEV).
  $\hookrightarrow$ các hệ số trong mô hình sẽ miêu tả mối liên hệ (tuyến tính), và ta cần diễn giải triệt để các hệ số này.
- Mặt khác, trong tương lai, nếu ta thu thập được dữ liệu về trạng thái hút thuốc, giới tính, độ tuổi và chiều cao của một nhóm người, ta có thể dự đoán dung tích phổi (FEV) dựa vào mô hình, mà không cần yêu cầu thực hiện thí nghiệm.

Độ chính xác vs Tính đơn giản

Một mô hình thống kê phù hợp (adequate) cần cân bằng hai tiêu chí:

Độ chính xác: mô hình phải mô tả chính xác cả thành phần hệ thống và thành phần ngẫu nhiên của dữ liệu.

Tính đơn giản: mô hình nên đơn giản nhất có thể.

Độ chính xác vs Tính đơn giản

Một mô hình thống kê phù hợp (adequate) cần cân bằng hai tiêu chí:

Độ chính xác: mô hình phải mô tả chính xác cả thành phần hệ thống và thành phần ngẫu nhiên của dữ liệu.

Tính đơn giản: mô hình nên đơn giản nhất có thể.

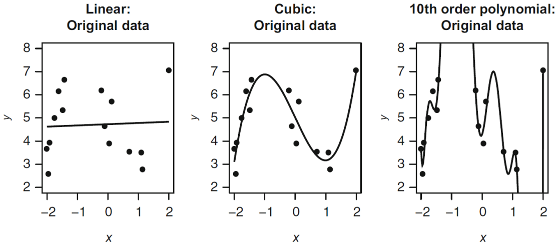
Ví dụ: Xét bộ dữ liệu mô phỏng từ mô hình hồi quy sau:

$$\begin{cases} y \sim \mathcal{N}(\mu, 0.35), \\ \mu = x^3 - 2x + 5 \end{cases}$$

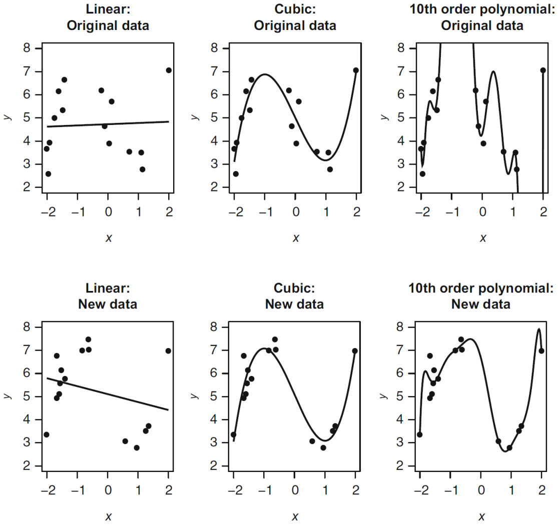
Ta xem xét ba mô hình tiềm năng sau

- tuyến tính đơn: $\mu = \beta_0 + \beta_1 x$;
- tuyến tính bậc ba: $\mu = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$;
- tuyến tính bậc mười: $\mu = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_{10} x^{10}$

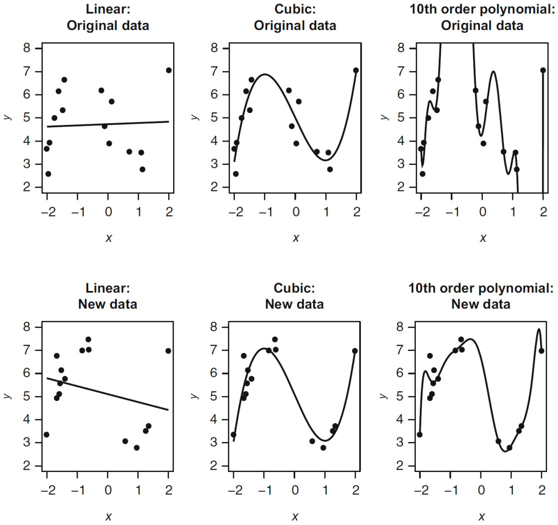
Độ chính xác vs Tính đơn giản



Độ chính xác vs Tính đơn giản



Độ chính xác vs Tính đơn giản



Một mô hình tốt sẽ giống nhau cho cả hai bộ dữ liệu.

Định nghĩa

- Giả sử, ta quan sát được giá trị y của biến ngẫu nhiên Y.
- Hàm mật độ xác suất của Y đã được biết với tham số θ, tức là $f(y;\theta)$
  $\hookrightarrow$ là một hàm của y và θ.
- Đặt $\mathcal{Y}$ là không gian mẫu $\Rightarrow y \in \mathcal{Y}$
- Đặt Θ là không gian tham số $\Rightarrow \theta \in \Theta$.
  tổng quát, y và θ có thể là hai vector.

Bài toán
Mục tiêu của chúng ta là đưa ra nhận định hoặc tuyên bố về phân phối của Y, dựa trên dữ liệu quan sát y.

Theo giả định, ta có:

- hàm mật độ xác suất f đã biết;
- quan sát y;

$\hookrightarrow$ ta cần đưa ra một nhận định về khoảng giá trị phù hợp của $\theta \in \Theta$, tương ứng với giá trị quan sát y.

Định nghĩa

Một phương pháp cơ bản là dựa trên hàm "hợp lý" (likelihood function) của θ:

$$L(\theta) = f(y; \theta),$$

với y cố định và $\theta \in \Theta$.

Diễn giải: dựa vào dữ liệu y, giá trị tham số $\theta \in \Theta$ là đáng tin hơn $\theta' \in \Theta$, như là một chỉ số của mô hình xác suất tạo ra dữ liệu, nếu $L(\theta) > L(\theta')$.

Tức là giá trị $L(\theta)$ sẽ tương đối lớn nếu như θ là gần so với giá trị thật θ_0, cái đã tạo ra dữ liệu.

- Khi Y là rời rạc, ta sử dụng hàm trọng lượng xác suất $\Pr(Y = y; \theta)$
- Khi Y là liên tục, ta sử dụng hàm mật độ xác suất $f(y; \theta)$.

Định nghĩa

Khi $y = (y_1, y_2, \ldots, y_n)$, với y_i là các quan sát độc lập nhau của Y_i, khi đó,

$$L(\theta) = f(y;\theta) = f(y_1, y_2, \ldots, y_n;\theta) = \prod_{i=1}^{n} f_{Y_i}(y_i;\theta).$$

Kết quả này có được do tính độc lập y_i.

Trong thực tế, sẽ thuận tiện hơn khi xét hàm log-likelihood:

$$\ell(\theta) = \log L(\theta) = \log f(y;\theta),$$

ta đặt $\ell(\theta) = -\infty$ nếu $L(\theta) = 0$.

Khi $y = (y_1, y_2, \ldots, y_n)$, thì

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f_{Y_i}(y_i;\theta).$$

Tính bất biến

Hai hàm likelihood được gọi là **tương đương** nếu chúng chỉ sai khác nhau một hằng số nhân (không phụ thuộc tham số).

Tính bất biến của hàm likelihood

Hàm likelihood (hoặc hàm log-likelihood) là bất biến với phép biến đổi 1-1 của dữ liệu.

Thật vậy, gọi $Z = g(Y)$, với g là một làm đơn điệu. Khi đó, hàm mật độ của Z là

$$f_Z(z;\theta) = f_Y(y;\theta)\left|\frac{\mathrm{d}y}{\mathrm{d}z}\right|$$

với $z = g(y)$, và $y = g^{-1}(z)$.

Suy ra,

$$L_Z(\theta) = \left|\frac{\mathrm{d}y}{\mathrm{d}z}\right| \times L_Y(\theta),$$

dễ thấy, $\left|\frac{\mathrm{d}y}{\mathrm{d}z}\right|$ không phụ thuộc tham số θ.

Ví dụ 1 - Phân phối Poisson

Giả sử y là một giá trị quan sát từ một phân phối Poisson:

$$\Pr(Y = y; \theta) = \frac{\theta^y \exp(-\theta)}{y!},$$

với $y \in \mathbb{Z}_+$ và $\theta > 0$.

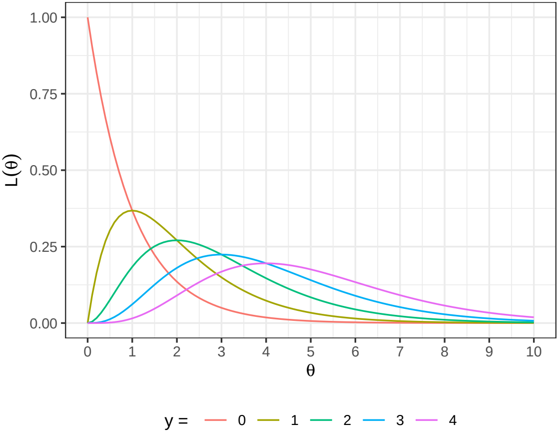
Khi đó, hàm likelihood là

$$L(\theta) = \frac{\theta^y \exp(-\theta)}{y!}.$$

Nếu

- $y = 0$, thì $L(\theta) = \exp(-\theta)$, hàm đồng điệu giảm của θ;
- $y > 0$, thì $L(\theta)$ sẽ đặt cực đại tại $\theta = y$, và có giới hạn 0 khi θ tiệm cận 0 hoặc ∞.

Ví dụ 1 - Phân phối Poisson



Ví dụ 2: phân phối mũ

Xét y là một mẫu ngẫu nhiên $y_1, y_2, \ldots, y_n$, độc lập, từ phân phối mũ với hàm mật độ xác suất

$$f(y;\theta) = \theta^{-1}\exp\left(-y/\theta\right),$$

với $y > 0$ và $\theta > 0$.

Khi đó, hàm likelihood là

$$L(\theta) = \prod_{i=1}^{n} \theta^{-1}\exp\left(-y_i/\theta\right) = \theta^{-n}\exp\left(-\frac{1}{\theta}\sum_{i=1}^{n} y_i\right)$$

và hàm log-likelihood là

$$\ell(\theta) = -n\log(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} y_i$$

Hàm likelihood và hàm log-likelihood đạt cực đại tại $\theta = \frac{1}{n}\sum_{i=1}^{n} y_i$

Ví dụ 3: phân phối chuẩn

Xét y là một mẫu ngẫu nhiên $y_1, y_2, \dots, y_n$, độc lập, từ phân phối chuẩn với hàm mật độ xác suất

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

với $y \in \mathbb{R}$, $\mu \in \mathbb{R}$ và $\sigma > 0$.

Khi đó, hàm likelihood là

$$L(\theta) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right)$$

và hàm log-likelihood là

$$\ell(\theta) = -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2,$$

với $\theta = (\mu,\sigma) \in \mathbb{R}\times\mathbb{R}_+$.

Thông tin của hàm log-likelihood

Xét hàm log-likelihood

$$\ell(\theta) = \sum_{i=1}^{n} \log f_{Y_i}(y_i; \theta)$$

Thông tin Fisher (Fisher information) được định nghĩa bởi:

$$\mathcal{I}(\theta) = -\mathbb{E}\left\{\frac{\mathrm{d}^2\ell(\theta)}{\mathrm{d}\theta^2}\right\}$$

Thông tin quan sát (observed information) được định nghĩa bởi:

$$\mathcal{J}(\theta) = -\frac{\mathrm{d}^2\ell(\theta)}{\mathrm{d}\theta^2}$$

Thông tin của hàm log-likelihood

Ví dụ 1: xét hàm log-likelihood cho phân phối Poisson

$$\ell(\theta) = \log(\theta)\sum_{i=1}^{n} y_i - n\theta,$$

với $\theta > 0$.

Ta dễ dàng tính được

- $\mathcal{J}(\theta) = \dfrac{1}{\theta^2}\sum_{i=1}^{n} y_i$
- $\mathcal{I}(\theta) = \dfrac{n}{\theta}$

Ví dụ 2: xét hàm log-likelihood cho phân phối mũ

$$\ell(\theta) = -n\log(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} y_i,$$

với $\theta > 0$. Hãy tìm $\mathcal{J}(\theta)$ và $\mathcal{I}(\theta)$.

Thông tin của hàm log-likelihood

Tổng quát, khi θ là một vecto p chiều, thì ta có

- ma trận thông tin Fisher (Fisher information matrix):

$$\mathcal{I}(\theta) = -\mathbb{E}\left\{\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^\top}\right\}$$

- ma trận thông tin quan sát (observed information matrix):

$$\mathcal{J}(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^\top}$$

Chúng đều là các ma trận cỡ $p \times p$, với các phần tử thứ (r, s) lần lượt là

$$-\mathbb{E}\left\{\frac{\partial^2 \ell(\theta)}{\partial\theta_r\partial\theta_s}\right\}, \qquad -\frac{\partial^2 \ell(\theta)}{\partial\theta_r\partial\theta_s}.$$

Nhận xét:

- ma trận thông tin Fisher $\mathcal{I}(\theta)$ có thể xác định không cần dữ liệu;
- ma trận thông tin quan sát $\mathcal{J}(\theta)$ cần dữ liệu để xác định.

Thông tin của hàm log-likelihood

Ví dụ 3: xét hàm log-likelihood cho phân phối chuẩn

$$\ell(\theta) = -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2,$$

với $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$.

Ta dễ dàng tính được

- ma trận thông tin quan sát

$$\mathcal{J}(\theta) = \begin{pmatrix} \dfrac{n}{\sigma^2} & \dfrac{2n}{\sigma^3}(\overline{y}-\mu) \\ \dfrac{2n}{\sigma^3}(\overline{y}-\mu) & -\dfrac{n}{\sigma^2} + \dfrac{3}{\sigma^4}\sum_{i=1}^{n}(y_i-\mu)^2 \end{pmatrix}$$

- ma trận thông tin Fisher

$$\mathcal{I}(\theta) = \begin{pmatrix} \dfrac{n}{\sigma^2} & 0 \\ 0 & \dfrac{2n}{\sigma^2} \end{pmatrix}$$

Ước lượng hợp lý cực đại

Như đã giới thiệu ở phần định nghĩa, một giá trị "hợp lý" cho θ là giá trị sao cho $L(\theta) > L(\theta')$ hoặc tương đương $\ell(\theta) > \ell(\theta')$.

Ta cần tìm θ sao cho $L(\theta)$ hoặc $\ell(\theta)$ đạt cực đại:

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} L(\theta),$$

hay tương đương

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta).$$

Ta gọi $\widehat{\theta}$ là **ước lượng hợp lý cực đại - maximum likelihood estimator** (**MLE**).

Để thuận tiện, ta sẽ xét phương trình thứ 2, với trường hợp tổng quát, θ là một vector p chiều.

Tính bất biến của MLE

Cho

- $\widehat{\theta}$ là MLE của tham số θ;
- $g(\cdot)$ là một hàm đơn điệu, 1-1 của θ, tức là $\psi = g(\theta)$.

Khi đó, $\widehat{\psi} = g(\widehat{\theta})$ cũng là MLE của ψ.

Điều này có được là bởi tính chất 1-1 của hàm $g(\cdot)$, tức là $\theta = g^{-1}(\psi)$, khi đó

$$\ell(\theta) = \ell\left(g^{-1}(\psi)\right) \equiv \ell^*(\psi).$$

Hơn nữa

$$\sup_{\psi} \ell^*(\psi) = \sup_{\psi} \ell\left(g^{-1}(\psi)\right) = \sup_{\theta} \ell(\theta).$$

Do đó, cực đại của $\ell^*(\psi)$ xác định tại $\psi = g(\theta) = g(\widehat{\theta})$, chứng minh rằng MLE của ψ là $g(\widehat{\theta})$.

Từ kết quả này, ta có thể viết $\widehat{\theta} = g^{-1}(\widehat{\psi})$.

Tính chất này được sử dụng trong các bài toán với miền xác định Θ của θ bị chặn.

Giải phương trình đạo hàm

Ước lượng hợp lý cực đại $\widehat{\theta}$ có thể được tìm bằng cách giải phương trình đạo hàm bậc 1. Tức là, $\widehat{\theta}$ là nghiệm của phương trình

$$U(\theta) \equiv \frac{\partial \ell(\theta)}{\partial \theta} = 0,$$

ta gọi $U(\theta)$ là hàm score (score function).

Để kiểm tra $\widehat{\theta}$ là một cực trị địa phương, ta kiểm tra điều kiện ma trận thông tin quan sát $\mathcal{J}(\theta)$ là xác định dương tại $\widehat{\theta}$.

Giải phương trình đạo hàm

Ví dụ 1 (tiếp theo): xét hàm log-likelihood cho phân phối Poisson

$$\ell(\theta) = \log(\theta)\sum_{i=1}^{n} y_i - n\theta,$$

với $\theta > 0$.

Hàm score được xác định bởi:

$$U(\theta) = \frac{1}{\theta}\sum_{i=1}^{n} y_i - n$$

Giải phương trình $U(\theta) = 0$, ta thu được:

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Ta kiểm tra được rằng $\mathcal{J}(\widehat{\theta}) = n/\widehat{\theta} > 0$.

Giải phương trình đạo hàm

Ví dụ 3 (tiếp theo): xét hàm log-likelihood cho phân phối chuẩn

$$\ell(\theta) = -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2,$$

với $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$.

Hàm score được xác định bởi:

$$U(\theta) = \begin{pmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) \\ -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-\mu)^2 \end{pmatrix}$$

Giải phương trình $U(\theta) = 0$, ta thu được:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{\mu})^2}.$$

Ta kiểm tra được rằng $\mathcal{J}(\widehat{\mu}, \widehat{\theta})$ là xác định dương.

Phương pháp giải lặp

Một phương pháp khác để xác định ước lượng hợp lý cực đại $\widehat{\theta}$ đó là giải lặp phương trình đạo hàm.

Cho trước một giá trị $\theta^\dagger$, áp dụng khai triển Taylor (bậc 1) cho hàm score tại $\theta^\dagger$, ta được

$$U(\theta) = U\left(\theta^\dagger\right) + \frac{\partial U\left(\theta^\dagger\right)}{\partial\theta}\left(\theta - \theta^\dagger\right)$$

Mặt khác, do $\widehat{\theta}$ là nghiệm của phương trình $U(\theta) = 0$, nên $U(\widehat{\theta}) = 0$ và

$$0 = U(\widehat{\theta}) = U\left(\theta^\dagger\right) + \frac{\partial U\left(\theta^\dagger\right)}{\partial\theta}\left(\widehat{\theta} - \theta^\dagger\right).$$

Suy ra,

$$\widehat{\theta} = \theta^\dagger + \mathcal{J}^{-1}\left(\theta^\dagger\right) U\left(\theta^\dagger\right),$$

với $\mathcal{J}^{-1}\left(\theta^\dagger\right)$ là ma trận nghịch đảo của $\mathcal{J}\left(\theta^\dagger\right)$.

Đây là một biến thể của phương pháp giải lặp Newton-Raphson.

Phương pháp giải lặp

Thuật toán giải lặp

1. Chọn một giá trị bắt đầu $\theta^{(0)}$
2. Với bước lặp $t = 0$, ta tính
$$\theta^{(t+1)} = \theta^{(t)} + \mathcal{J}^{-1}\left(\theta^{(t)}\right) U\left(\theta^{(t)}\right),$$
3. Đặt $t = t + 1$, lặp lại bước 2 cho tới khi nào thuật toán hội tụ, có thể là
$$\left\|\theta^{(t+1)} - \theta^{(t)}\right\| < \varepsilon.$$
4. ước lượng hợp lý cực đại $\widehat{\theta} = \theta^{(t+1)}$.

Phương pháp giải lặp

Ngoài ra, ta có thể thay thế $\mathcal{J}(\theta)$ bằng $\mathcal{I}(\theta)$, khi đó, thuật toán có tên Fisher scoring

$$\widehat{\theta} = \theta^{\dagger} + \mathcal{I}^{-1}\left(\theta^{\dagger}\right) U\left(\theta^{\dagger}\right).$$

Phương pháp thường được áp dụng khi:

- ma trận $\mathcal{J}(\theta)$ không được định nghĩa tốt;
- ma trận $\mathcal{J}(\theta)$ có công thức phức tạp.

Phương pháp giải lặp

Ví dụ 4: Xét y là một mẫu ngẫu nhiên $y_1, y_2, \ldots, y_n$, độc lập, từ phân phối Weibull với hàm mật độ xác suất:

$$f(y;\theta,\alpha) = \frac{\alpha}{\theta}\left(\frac{y}{\theta}\right)^{\alpha-1} \exp\left\{-\left(\frac{y}{\theta}\right)^{\alpha}\right\},$$

với $y > 0$ và $\theta, \alpha > 0$.

Hàm log-likelihood được xác định là

$$\ell(\theta,\alpha) = n\log(\alpha) - n\log(\theta) + (\alpha-1)\sum_{i=1}^{n}\log\left(\frac{y_i}{\theta}\right) - \sum_{i=1}^{n}\left(\frac{y_i}{\theta}\right)^{\alpha}.$$

Từ đây có xác định được hàm score là

$$U(\theta,\alpha) = \begin{pmatrix} -n\alpha/\theta + \alpha\theta^{-1}\sum_{i=1}^{n}(y_i/\theta)^{\alpha} \\ n/\alpha + \sum_{i=1}^{n}\log(y_i/\theta) - \sum_{i=1}^{n}(y_i/\theta)^{\alpha}\log(y_i/\theta) \end{pmatrix}$$

ta không thể giải phương trình này bằng phương pháp giải tích.

Phương pháp giải lặp

Từ hàm score, ta xác định ma trận thông tin quan sát $\mathcal{J}(\theta, \alpha)$:

$$\mathcal{J}(\theta, \alpha) = \begin{pmatrix} j_{\theta,\theta}(\theta, \alpha) & j_{\theta,\alpha}(\theta, \alpha) \\ j_{\theta,\alpha}(\theta, \alpha) & j_{\alpha,\alpha}(\theta, \alpha) \end{pmatrix},$$

trong đó,

$$\begin{aligned} j_{\theta,\theta}(\theta, \alpha) &= -\frac{n\alpha}{\theta^2} + \frac{\alpha(\alpha+1)}{\theta^2} \sum_{i=1}^{n} \left(\frac{y_i}{\theta}\right)^{\alpha}, \\ j_{\theta,\alpha}(\theta, \alpha) &= \frac{n}{\theta} - \sum_{i=1}^{n} \frac{y_i^{\alpha}}{\theta^{\alpha+1}} \left(1 + \alpha \log\left(\frac{y_i}{\theta}\right)\right), \\ j_{\alpha,\alpha}(\theta, \alpha) &= \frac{n}{\alpha^2} + \sum_{i=1}^{n} \left(\frac{y_i}{\theta}\right)^{\alpha} \log\left(\frac{y_i}{\theta}\right), \end{aligned}$$

với $\theta, \alpha > 0$.

Phương pháp giải lặp

Đặt $\beta = (\theta, \alpha)$, khi đó, nghiệm giải lặp là

$$\widehat{\beta} = \beta^{\dagger} + \mathcal{J}^{-1}\left(\beta^{\dagger}\right) U\left(\beta^{\dagger}\right),$$

Để đảm bảo ước lượng $\widehat{\theta}, \widehat{\alpha} > 0$, ta sử dụng biến đổi $\psi = (\log(\theta), \log(\alpha))$. Khi đó,

$$\widehat{\psi} = \psi^{\dagger} + \mathcal{J}^{-1}\left(\psi^{\dagger}\right) U\left(\psi^{\dagger}\right).$$

Sau đó, với phép biến đổi ngược, exp(), ta thu được kết quả $\widehat{\theta}, \widehat{\alpha} > 0$.

Áp dụng

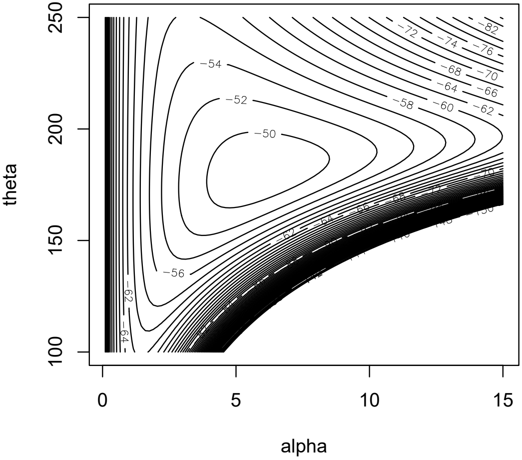
Ta áp dụng mô hình phân phối Weibull để mô hình hóa dữ liệu về thời gian hỏng của lò xo trong thí nghiệm với mức ứng suất 950 N/mm^2:

225, 171, 198, 189, 189, 135, 162, 135, 117, 162

Dựa vào công thức hàm log-likelihood của phân phối Weibull, ta có thể biểu diễn đồ thị như sau.

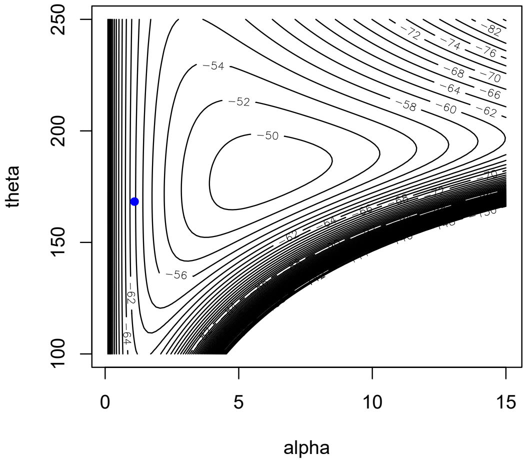
Phương pháp giải lặp

Hình chiếu của hàm log-likelihood của phân phối Weibull.



Phương pháp giải lặp

Thực hiện giải lặp, với sai số chặn là 10^{-9}.



Phương pháp giải lặp

Bảng tổng hợp kết quả

Lần lặp t	$\theta^{(t)}$	$\alpha^{(t)}$	Sai số
0	168.3000	1.1000	
1	172.1154	2.1111	6.5226×10^{-1}
2	175.1454	3.6845	5.5723×10^{-1}
3	180.4374	5.2457	3.5452×10^{-1}
4	181.1004	5.9082	1.1899×10^{-1}
5	181.4075	5.9764	1.1604×10^{-3}
6	181.4056	5.9769	8.1710×10^{-5}
7	181.4056	5.9769	3.6697×10^{-9}
8	181.4056	5.9769	2.2204×10^{-16}

$\Rightarrow$ ước lượng MLE của θ và α là $\widehat{\theta} = 181.4056$, $\widehat{\alpha} = 5.9769$.

Thông tin của hàm likelihood và MLE

Nhắc lại rằng, thông tin quan sát của hàm log-likelihood

$$\mathcal{J}(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^\top}.$$

Nhận xét:

- về mặt hình học, với $p = 1$, $\mathcal{J}(\theta)$ đo độ cong của $\ell(\theta)$;
- $\mathcal{J}(\theta)$ là một hàm tuyến tính theo n, khi $p = 1$;
- độ cong của $\ell(\theta)$ tại giá trị cực đại, sẽ tăng khi n tăng lên.
- khi thông tin quan sát tại một điểm $\theta^\dagger$, $\mathcal{J}(\theta^\dagger)$ càng lớn, thì $\theta^\dagger$ càng được gim chặt vào vùng cực trị của $\ell(\theta)$.

Điều kiện chính quy

(A1) Giá trị chính xác (*the true value*) θ_0 của θ là tồn tại và phải nằm trong không gian tham số Θ, không gian này là hữu hạn chiều và compact.

(A2) Với hai giá trị khác nhau θ_1 và θ_2, hàm log-likelihood được định nghĩa bởi hai tham số này sẽ khác nhau, tức là $\ell(\theta_1) \neq \ell(\theta_2)$.

(A3) Tồn tại một lân cận $\mathcal{N}$ của θ_0 sao cho, các đạo hàm $\dfrac{\partial \ell}{\partial \theta}, \dfrac{\partial^2 \ell}{\partial \theta \partial \theta^\top}$ và $\dfrac{\partial^3 \ell}{\partial \theta \partial \theta^\top \partial \theta}$ tồn tại hầu chắc chắn trong lân cận này. Hơn nữa, $\dfrac{1}{n}\mathbb{E}\left\{\left\|\dfrac{\partial^3 \ell}{\partial \theta \partial \theta^\top \partial \theta}\right\|\right\}$ bị chặn đều với $\theta \in \mathcal{N}$.

(A4) Ma trận thông tin Fisher $\mathcal{I}(\theta)$ là hữu hạn và xác định dương, với $\theta \in \mathcal{N}$.

Giả định (A2) ám chỉ tới tính tồn tại duy nhất của tham số.

Giả định (A4) là nhằm đảm bảo cực đại địa phương của hàm log-likelihood tồn tại trong lân cận $\mathcal{N}$ của giá trị chính xác θ_0.

Tính chất tiệm cận của MLE - I

Nếu các điều kiện chính quy được đảm bảo, ta có tính chất sau của MLE, $\widehat{\theta}$:

1. $\widehat{\theta} \xrightarrow{a.s} \theta_0$ khi $n \to \infty$ (hội tụ hầu chắc chắn), và $\ell(\widehat{\theta})$ là cực đại địa phương của $\ell(\theta)$.
2. $\widehat{\theta}$ là ước lượng vững của θ tức là $\widehat{\theta} \xrightarrow{p} \theta_0$, khi $n \to \infty$.
3. Xét score statistic $U(\theta)$. Khi đó, với θ_0, áp dụng định lý giới hạn trung tâm và luật số lớn dạng yếu, ta có

$$\mathcal{I}^{-1/2}(\theta_0)U(\theta_0) \xrightarrow{d} \mathcal{N}_p\left(0, \mathbf{I}_p\right), \quad \text{và} \quad \mathcal{I}^{-1}(\theta_0)\mathcal{J}(\theta_0) \xrightarrow{p} \mathbf{I}_p.$$

4. Ước lượng $\widehat{\theta}$ là xấp xỉ phân phối chuẩn p-chiều với vectơ trung bình là θ_0 và ma trận hiệp phương sai là nghịch đảo của ma trận thông tin Fisher, $\mathcal{I}^{-1}(\theta_0)$, tức là

$$\widehat{\theta} \xrightarrow{d} \mathcal{N}_p\left(\theta_0, \mathcal{I}^{-1}(\theta_0)\right),$$

khi $n \to +\infty$. Ta có thể viết lại thành

$$\mathcal{I}^{1/2}(\theta_0)\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}_p\left(0, \mathbf{I}_p\right).$$

Tính chất tiệm cận của MLE - II

5 Từ kết quả của tiệm cận chuẩn, ta chứng minh được rằng

$$\left(\widehat{\theta} - \theta_0\right)^\top \mathcal{I}(\theta_0) \left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \chi_p^2,$$

phân phối Chi-bình phương bậc p, khi $n \to +\infty$.

6 $\mathcal{I}(\widehat{\theta})$ là một ước lượng vững cho $\mathcal{I}(\theta_0)$.

7 Ước lượng $\widehat{\theta}$ là tiệm cận không chệch (*asymptotically unbiased*),

$$\mathbb{E}(\widehat{\theta}) \to \theta_0$$

khi $n \to \infty$. Trong một số trường hợp, độ chệch sẽ là đáng kể nếu cỡ mẫu nhỏ. Đặc biệt, đối với phân phối chuẩn, $\widehat{\mu} = \overline{Y}$ là ước lượng không chệch.

8 $\widehat{\theta}$ là một ước lượng tiệm cận hiệu quả (*asymptotically efficient*):

$$\mathbb{V}\mathrm{ar}(\widehat{\theta}) \leq \mathbb{V}\mathrm{ar}(\widetilde{\theta})$$

trong đó, $\widetilde{\theta}$ là một ước lượng tiệm cận không chệch của θ.

Nhắc lại một số tính chất

Định lý giới hạn trung tâm - nhiều chiều

Cho $X_1, X_2, \ldots, X_n$ là một dãy vector p chiều, sao cho X_i là độc lập, cùng phân phối với vector trung bình μ và ma trận hiệp phương sai Σ (với Σ xác định dương và $|\Sigma| < \infty$). Đặt

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

thì

$$\sqrt{n}\Sigma^{-1/2}\left(\overline{X} - \mu\right) \xrightarrow{d} \mathcal{N}_p\left(0, \mathbf{I}_p\right),$$

khi $n \to \infty$.

Luật số lớn dạng yếu

Cho $X_1, X_2, \ldots, X_n$ là một dãy vector p chiều, sao cho X_i là độc lập, cùng phân phối với vector trung bình μ. Khi đó

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu,$$

khi $n \to \infty$.

Chứng minh một số tính chất

Xét tính chất 3.
Nhắc lại rằng, score statistics $U(\theta)$ được xác định bởi

$$U(\theta) = \frac{\partial}{\partial\theta}\ell(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f_{Y_i}(y_i;\theta) = \sum_{i=1}^{n} u_i(\theta).$$

Do $Y_1, Y_2, \ldots, Y_n$ là i.i.d $\Rightarrow$ hàm $u_1(\theta), u_2(\theta), \ldots, u_n(\theta)$ cũng là i.i.d. với

$$\mathbb{E}\left(u_i(\theta)\right) = \int_{-\infty}^{+\infty} u_i(\theta) f_{Y_i}(y_i;\theta)\,\mathrm{d}y_i = \int_{-\infty}^{+\infty} \frac{\partial}{\partial\theta} \log\left(f_{Y_i}(y_i;\theta)\right) f_{Y_i}(y_i;\theta)\,\mathrm{d}y_i = 0$$

và

$$\mathbb{V}\mathrm{ar}(u_i(\theta)) = -\int_{-\infty}^{+\infty} \frac{\partial^2}{\partial\theta\partial\theta^\top} \log\left(f_{Y_i}(y_i;\theta)\right) f_{Y_i}(y_i;\theta)\,\mathrm{d}y_i = i(\theta),$$

với $i(\theta)$ là ma trận thông tin Fisher ứng với 1 quan sát bất kỳ.

Chứng minh một số tính chất

Áp dụng Định lý Giới hạn trung tâm cho dãy vector ngẫu nhiên i.i.d $u_1(\theta_0), u_2(\theta_0), \ldots, u_n(\theta_0)$, ta có

$$\sqrt{n} i(\theta_0)^{-1/2} \left(\frac{1}{n} \sum_{i=1}^{n} u_i(\theta_0) - 0 \right) \xrightarrow{d} \mathcal{N}_p\left(0, \mathbf{I}_p\right).$$

Mặt khác, ta chứng minh được rằng

$$\mathbb{E}(U(\theta_0)) = 0 \qquad \text{và} \qquad \mathbb{V}\mathrm{ar}(U(\theta_0)) = n i(\theta_0).$$

Do đó, ta thu được

$$\mathcal{I}^{-1/2}(\theta_0) U(\theta_0) \xrightarrow{d} \mathcal{N}_p\left(0, \mathbf{I}_p\right)$$

Ví dụ

Ví dụ 1 (tiếp theo): đối với phân phối Poisson, MLE cho θ là:

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Thông tin Fisher $\mathcal{I}(\theta) = n/\theta$. Do đó, ta xác định được

$$\mathbb{V}\mathrm{ar}(\widehat{\theta}) = \widehat{\theta}/n.$$

Ví dụ 3 (tiếp theo): đối với phân phối chuẩn, MLE cho (μ, σ) là:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\mu})^2}$$

Ma trận thông tin Fisher $\mathcal{I}(\mu, \sigma)$

$$\mathcal{I}(\mu,\sigma) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{pmatrix} \quad \Rightarrow \quad \mathbb{V}\mathrm{ar}\begin{pmatrix} \widehat{\mu} \\ \widehat{\sigma} \end{pmatrix} = \begin{pmatrix} \frac{\widehat{\sigma}^2}{n} & 0 \\ 0 & \frac{\widehat{\sigma}^2}{2n} \end{pmatrix}$$

Khoảng tin cậy thành phần

Từ kết quả tiệm cận phân phối chuẩn của $\widehat{\theta}$, ta suy ra

$$\frac{\widehat{\theta}_j - \theta_{0j}}{\sqrt{v_{jj}}} \xrightarrow{d} \mathcal{N}(0,1),$$

Với $j = 1, 2, \ldots, p$; v_{jj} là thành phần thứ j trên đường chéo của $\mathcal{I}^{-1}(\theta)$.

Từ đây, ta xây dựng được khoảng tin cậy $(1-\alpha) \times 100\%$ cho θ_j là

$$\left[\widehat{\theta}_j - z_{1-\alpha/2}\sqrt{\widehat{v}_{jj}}, \widehat{\theta}_j + z_{1-\alpha/2}\sqrt{\widehat{v}_{jj}}\right]$$

$\widehat{v}_{jj}$ là thành phần thứ j trên đường chéo của $\mathcal{I}^{-1}(\widehat{\theta})$.

Kiểm định cho từng thành phần

Ta thường quan tâm tới bài toán kiểm định

$$\begin{cases} H_0: & \theta_j = \theta_{0j} \\ H_A: & \theta_j \neq \theta_{0j} \end{cases}$$

Sử dụng kết quả tiệm cận chuẩn của MLE, ta có

$$\frac{\widehat{\theta}_j - \theta_{0j}}{\sqrt{v_{jj}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Khi này, ta định nghĩa được thống kê Z, và tính được p-value tương ứng là

$$p\text{-value} = 2\left\{1 - \Phi\left(\left|\frac{\widehat{\theta}_j - \theta_{0j}}{\sqrt{\widehat{v}_{jj}}}\right|\right)\right\}$$