Bài giảng 6: Mô hình hồi quy tuyến tính và áp dụng

TS. Tô Đức Khánh

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

Học phần: Thống Kê Trong Sinh Học

Nội dung buổi học

1. *Mô hình hồi quy tuyến tính đơn*
   - Giới thiệu và ước lượng
   - Tính chất của ước lượng
   - Hồi quy đơn khi X là biến nhị phân
2. *Mô hình hồi quy tuyến tính đa biến*
   - Mô hình cơ bản
   - Mô hình với biến giải thích là định tính
   - Mở rộng mô hình hồi quy
3. *Chuẩn đoán mô hình*
4. *Một số chủ đề thêm*

1 *Mô hình hồi quy tuyến tính đơn*
- Giới thiệu và ước lượng
- Tính chất của ước lượng
- Hồi quy đơn khi X là biến nhị phân

2 *Mô hình hồi quy tuyến tính đa biến*
- Mô hình cơ bản
- Mô hình với biến giải thích là định tính
- Mở rộng mô hình hồi quy

3 *Chuẩn đoán mô hình*

4 *Một số chủ đề thêm*

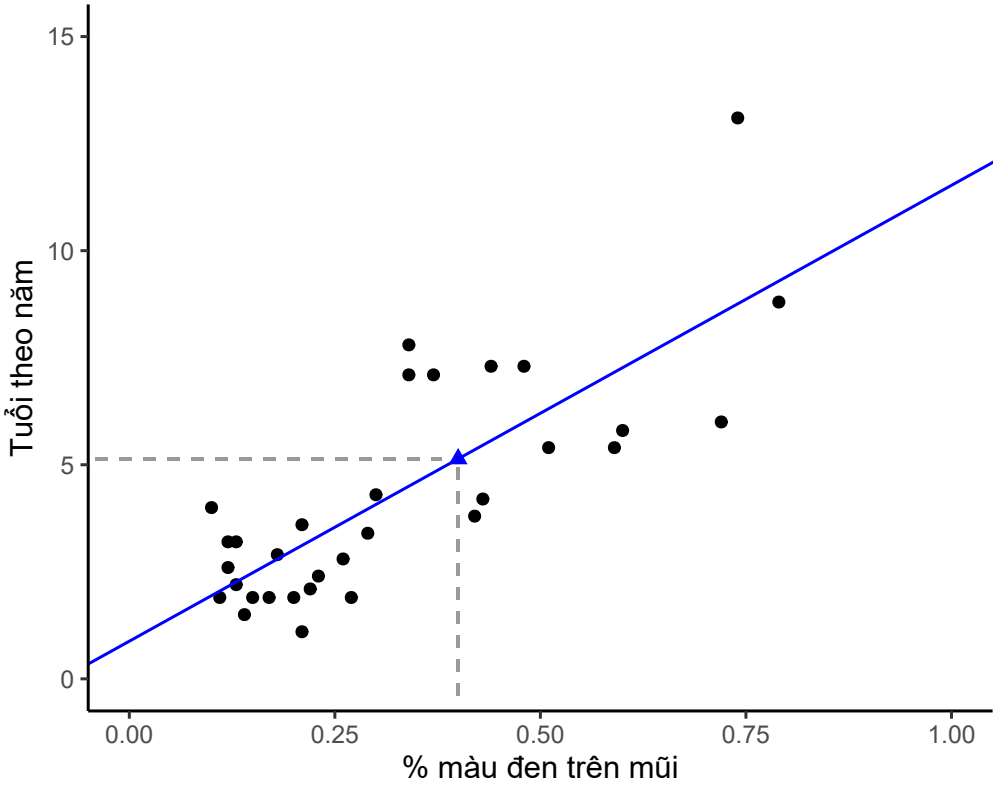
Mô hình hồi quy

Mô hình hồi quy

Mô hình hồi quy là một mô hình toán được dùng để:

- mô tả sự biến động
- dự đoán giá trị

của một biến định lượng dựa trên giá trị của một hoặc nhiều biến khác.



Mô hình hồi quy

Biến trong mô hình hồi quy

- Biến phản hồi (**response variable**) là biến mà được quan tâm trong nghiên cứu, tức là biến mà ta mong muốn tiên đoán từ mô hình. Nó cũng được biết tới với tên gọi biến phụ thuộc (**dependent variable**), thường được ký hiệu là Y.
- Biến giải thích (**explanatory variable**) là biến được dùng trong mô hình để dự đoán biến phản hồi. Nó còn được biết tới với tên gọi biến độc lập (**independent variable**) hoặc biến hồi quy **regressor**, thường được ký hiệu là X. Đặc biệt, trong thiết kết thí nghiệm, ta có thể kiểm soát các giá trị của X.

Ví dụ:

- X: % màu đen trên mũi của sư tử đực
- Y: tuổi của sư tử theo năm

ta mong muốn dự đoán được Y thông qua % màu đen trên mũi của sư tử đực, bởi vì để xác định năm tuổi của một con sử tử đực cần phải theo dõi lâu dài hoặc cần bắt và đo đạc các chỉ số sinh học.

Mô hình hồi quy

Về mặt toán học, ta có thể thiết lập

$$Y = f(X)$$

với f là một hàm số thực.

Nhưng do X và Y đều là các đại lượng ngẫu nhiên, nên

$$Y = f(X; \beta) + \varepsilon$$

trong đó, ε đại diện cho sai số giữa Y với $f(X; \beta)$, và β là tham số không biết.

Nếu $f(X; \beta) = \beta g(X)$, với $g(\cdot)$ là một hàm thực, được xác định, thì

$$Y = \beta g(X) + \varepsilon$$

được gọi là **mô hình hồi quy tuyến tính**. Ví dụ:

- $y = \beta_0 + \beta_1 x + \varepsilon$
- $y = \beta_0 + \beta_1 x^2 + \beta_3 x + \varepsilon$

Chú ý: sự tuyến tính này là tuyến tính theo hệ số β.

Hồi quy tuyến tính đơn giản

Hồi quy tuyến tính đơn giản

Mô hình hồi quy tuyến tính đơn giản là mô hình dự đoán Y theo một biến X:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

hai tham số β_0 and β_1 là không biết trước và được gọi là hệ số của mô hình:

- β_0 là hệ số chặn (**intercept coefficient**) biểu thị giá trị của Y khi biến giải thích $X = 0$;
- β_1 là hệ số góc (**slope coefficient**) cho biến giải thích X;
- nếu $\beta_1 > 0$: X tăng 1 đơn vị thì Y tăng β_1 đơn vị;
- nếu $\beta_1 < 0$: X tăng 1 đơn vị thì Y giảm $|\beta_1|$ đơn vị.

Chú ý:

- Cho giá trị của β_0 và β_1, ta có thể vẽ được một đường thẳng hồi quy biểu diễn mối quan hệ giữa X và Y.
- Trong thực tế các tham số này là không biết $\Rightarrow$ ta cần phải ước lượng từ dữ liệu.

Ước lượng mô hình hồi quy tuyến tính đơn

Giả sử ta có n cặp quan sát $(X_1, Y_1), \ldots, (X_n, Y_n)$. Xét mô hình

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

với mọi $i = 1, \ldots, n$.

Từ mô hình tuyến tính, ta có

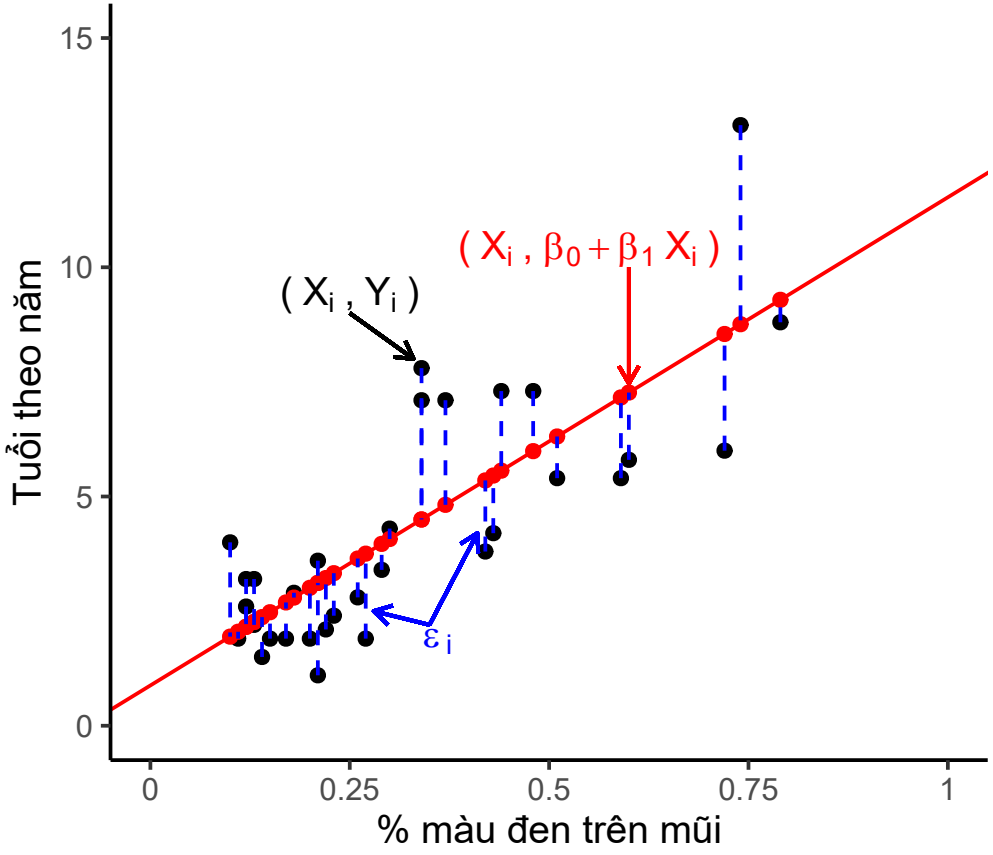
$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

Như vậy ε_i chính là sự khác biệt giữa quan sát Y_i và giá trị nằm trên đường thẳng $\beta_0 + \beta_1 X_i$.

Ước lượng mô hình hồi quy tuyến tính đơn

Xét dữ liệu với các cặp quan sát của % màu đen trên mũi của sử tử đực và tuổi (theo năm) của chúng. Giả sử

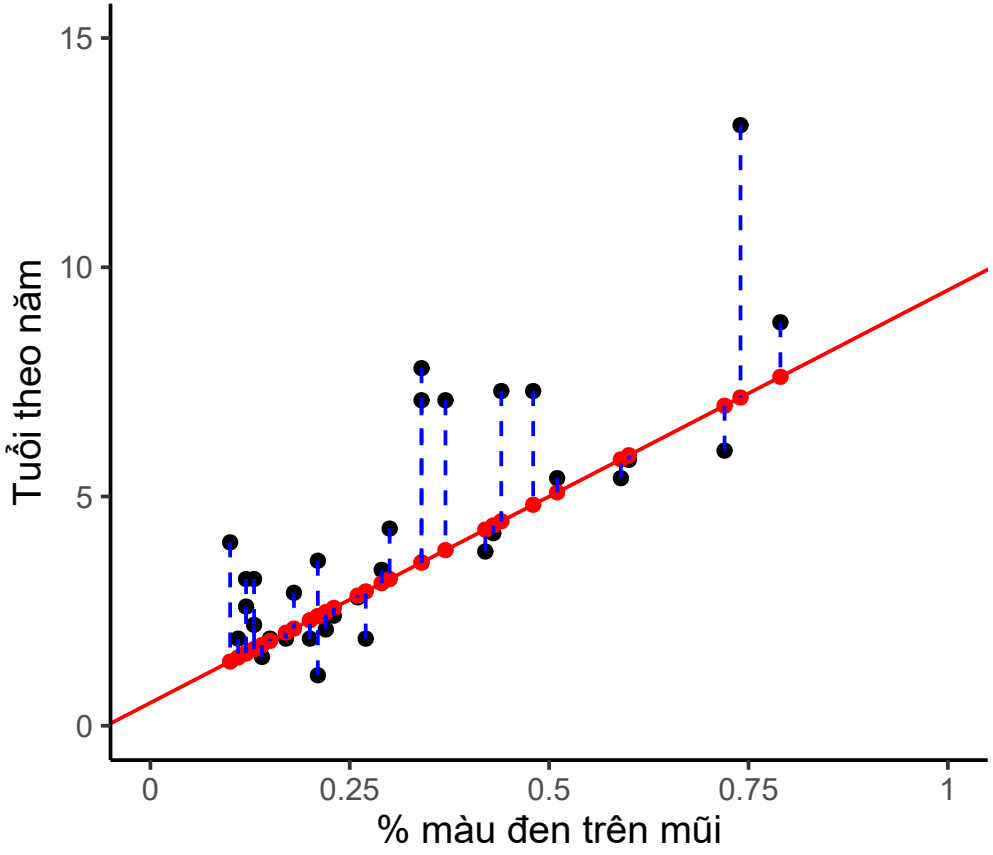
- $\beta_0 = 0.879$
- $\beta_1 = 10.647$



Ước lượng mô hình hồi quy tuyến tính đơn

Xét dữ liệu với các cặp quan sát của % màu đen trên mũi của sử tử đực và tuổi (theo năm) của chúng. Giả sử

- $\beta_0 = 0.5$
- $\beta_1 = 9$



Ước lượng mô hình hồi quy tuyến tính đơn

Do β_0 và β_1 là không biết $\Rightarrow$ ta cần tìm các ước lượng:

- $\widehat{\beta}_0$
- $\widehat{\beta}_1$

sao cho sai số giữa quan sát Y_i và $\widehat{\beta}_0 + \widehat{\beta}_1 X_i$ là nhỏ nhất, tức là

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i) \right)^2$$

là nhỏ nhất.

$\hookrightarrow$ Phương pháp tìm ước lượng $\widehat{\beta}_0$ và $\widehat{\beta}_1$ như thế này được gọi là phương pháp bình phương bé nhất (ordinary least squares - OLS).

Ước lượng mô hình hồi quy tuyến tính đơn

Xét mô hình

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Ta cần tìm ước lượng $\widehat{\beta}_0$ và $\widehat{\beta}_1$ sao cho

$$\min \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2.$$

Tức là $\widehat{\beta}_0$ và $\widehat{\beta}_1$ là hai điểm cực trị của hàm

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2.$$

Như vậy, ta tìm $\widehat{\beta}_0$ và $\widehat{\beta}_1$ bằng cách giải hệ phương trình đạo hàm riêng:

$$\begin{aligned}
\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_0} &= \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_i)\right) = 0 \\
\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_1} &= \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_i)\right) X_i = 0
\end{aligned}$$

Ước lượng mô hình hồi quy tuyến tính đơn

Giải 2 phương trình trên ta thu được

$$\begin{aligned}\widehat{\beta}_{1,n} &= \frac{\sum_{i=1}^{n} X_i Y_i - n\,\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}\\ \widehat{\beta}_{0,n} &= \overline{Y} - \widehat{\beta}_{1,n}\overline{X}\end{aligned}$$

Và với một bộ dữ liệu quan sát $(x_1, y_1), \ldots, (x_n, y_n)$, ta tính được

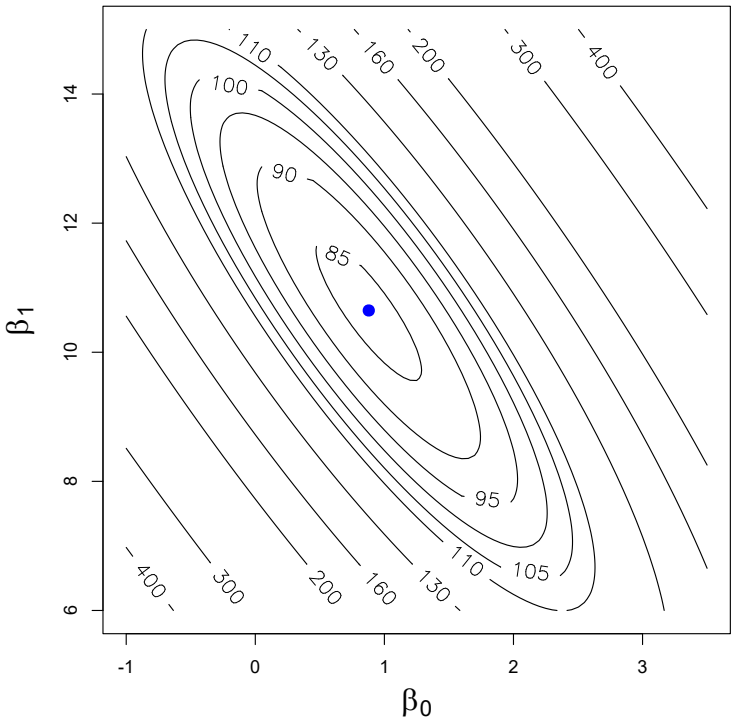
$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\\ \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1\overline{x}\end{aligned}$$

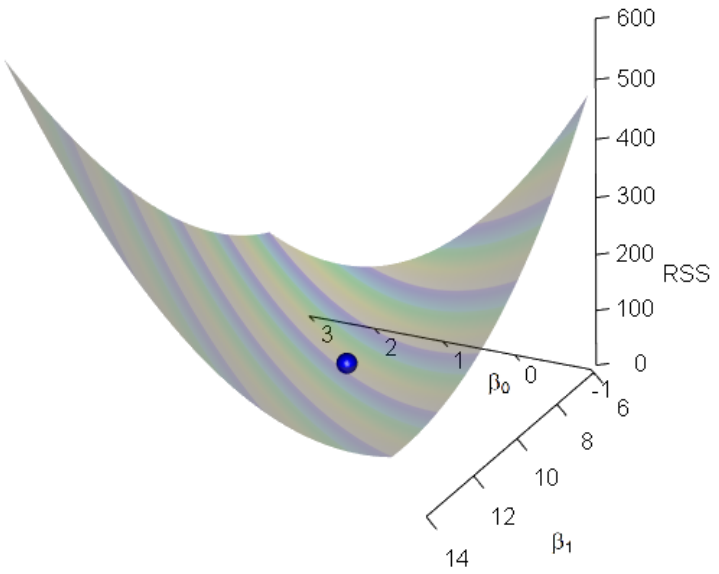
Từ đây, ta có phương trình hồi quy tuyến tính ước lượng:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x,$$

Ước lượng mô hình hồi quy tuyến tính đơn

Hình bên dưới mô tả đường đồng mức (coutour) và đồ thị 3D cho hàm $RSS(\beta_0, \beta_1)$ đối với dữ liệu trong nghiên cứu dự đoán tuổi sư tử đực từ % màu đen ở mũi.

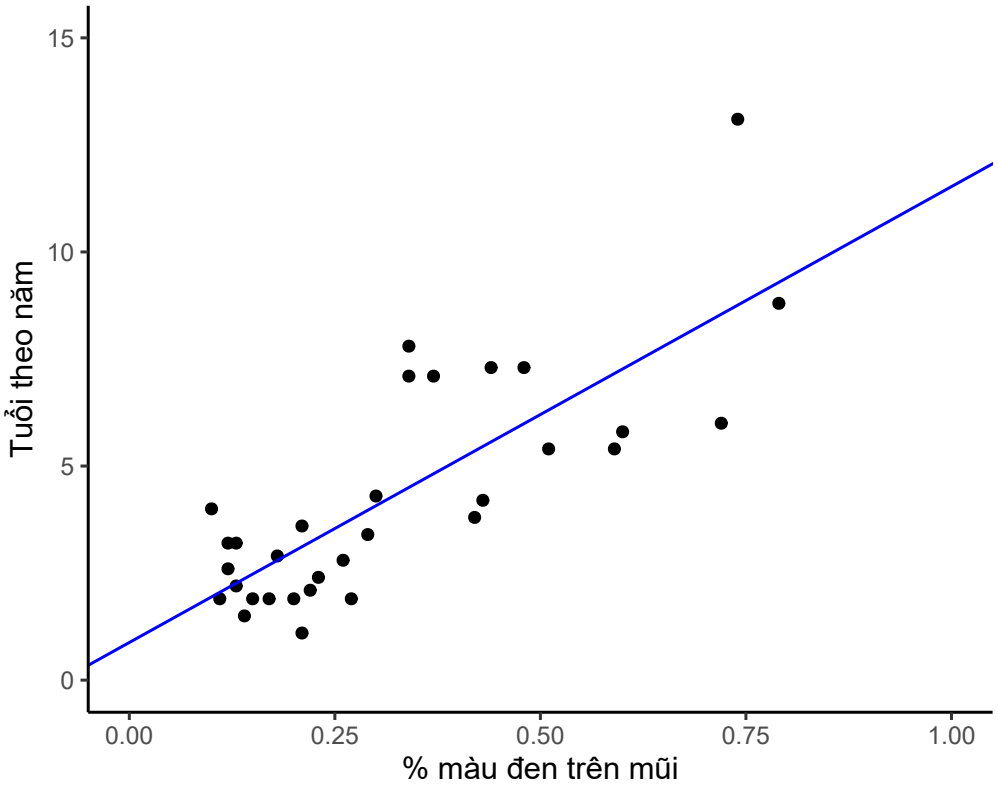




Ta ước lượng được

- $\widehat{\beta}_0 = 0.879$
- $\widehat{\beta}_1 = 10.647$

Ước lượng mô hình hồi quy tuyến tính đơn



1 *Mô hình hồi quy tuyến tính đơn*
- Giới thiệu và ước lượng
- Tính chất của ước lượng
- Hồi quy đơn khi X là biến nhị phân

2 *Mô hình hồi quy tuyến tính đa biến*
- Mô hình cơ bản
- Mô hình với biến giải thích là định tính
- Mở rộng mô hình hồi quy

3 *Chuẩn đoán mô hình*

4 *Một số chủ đề thêm*

Đánh giá tính chính xác của ước lượng

Mô hình hồi quy đơn giản

Tổng quát, mối quan hệ tuyến tính thực sự giữa X và Y được miêu tả dưới dạng:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

trong đó,

- β_0: hệ số chặn (intercept), tức là trung bình của Y khi $X = 0$;
- β_1: hệ số góc (slope), tức là mức tăng trung bình của Y tương ứng với một đơn vị tăng của X;
- ε: thành phần sai số ngẫu nhiên bao hàm tất cả các lỗi liên quan đến:
  (i) mối quan hệ thực sự không phải là tuyến tính;
  (ii) có thể có thêm khác biến khác có sự liên hệ tới Y;
  (iii) tồn tại sai số đo đạc trong Y.

Phương trình trên xác định đường hồi quy tổng thể, là xấp xỉ tuyến tính tốt nhất cho mối quan hệ thực sự giữa X và Y.

Đánh giá tính chính xác của ước lượng

Với mọi quan sát $i = 1, 2, \ldots, n$, ta có phương trình

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Áp dụng phương pháp OLS, ta có được các ước lượng:

$$\widehat{\beta}_{1,n} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}, \qquad \widehat{\beta}_{0,n} = \overline{Y} - \widehat{\beta}_{1,n}\overline{X}$$

Để đánh giá tính chính xác của các ước lượng OLS này, ta cần một số giả định như sau:

(A1) Có một quan hệ tuyến tính giữa X và Y.

(A2) Các cặp điểm dữ liệu (X_i, Y_i) với $i = 1, 2, \ldots, n$, là độc lập, cùng phân phối (i.i.d) được lấy từ phân phối của chúng.

(A3) Thành phần sai số ε_i là độc lập với X_i, và có trung bình bằng 0, tức là, $\mathbb{E}(\varepsilon_i) = 0$ với mọi i.

Tính vững và tính không chệch

Dưới các giả định (A1), (A2) và (A3), ta có thể chứng minh rằng:

- Ước lượng OLS $\widehat{\beta}_{0,n}$ và $\widehat{\beta}_{1,n}$ là các **ướng lượng vững - consistent estimators**.
- Ước lượng OLS $\widehat{\beta}_{0,n}$ và $\widehat{\beta}_{1,n}$ là các **ước lượng không chệch - unbiased estimators** của hệ số hồi quy tổng thể, tức là, $\mathbb{E}(\widehat{\beta}_{0,n}) = \beta_0$ và $\mathbb{E}(\widehat{\beta}_{1,n}) = \beta_1$.

Ví dụ, xét một mô hình hồi quy như sau

$$Y = 1 + 2X + \varepsilon$$

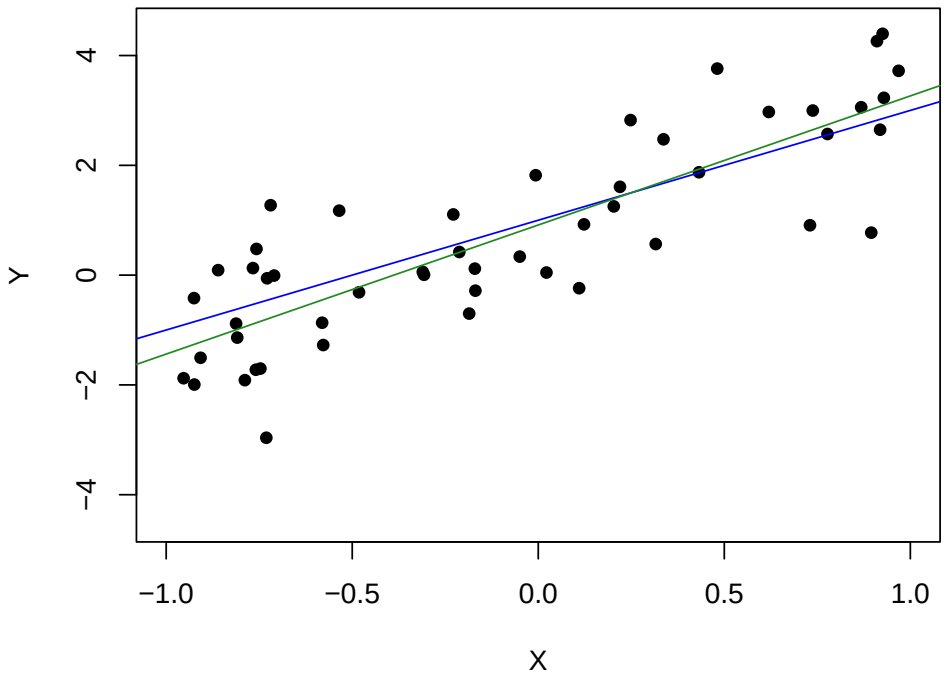
với $X \sim \mathcal{U}(-1, 1)$ và sai số $\varepsilon \sim \mathcal{N}(0, 1)$.

- Ta tạo một mẫu ngẫu nhiên gồm 50 X's và 50 ε's dựa vào phân phối của chúng;
- Dựa vào mô hình, ta dễ dàng thu được 50 quan sát tương ứng của Y's.
- Với mẫu vừa được tạo (X_i, Y_i), ta áp dụng phương pháp OLS để ước lượng $\widehat{\beta}_{0,n}$ và $\widehat{\beta}_{1,n}$.

Tính vững và tính không chệch

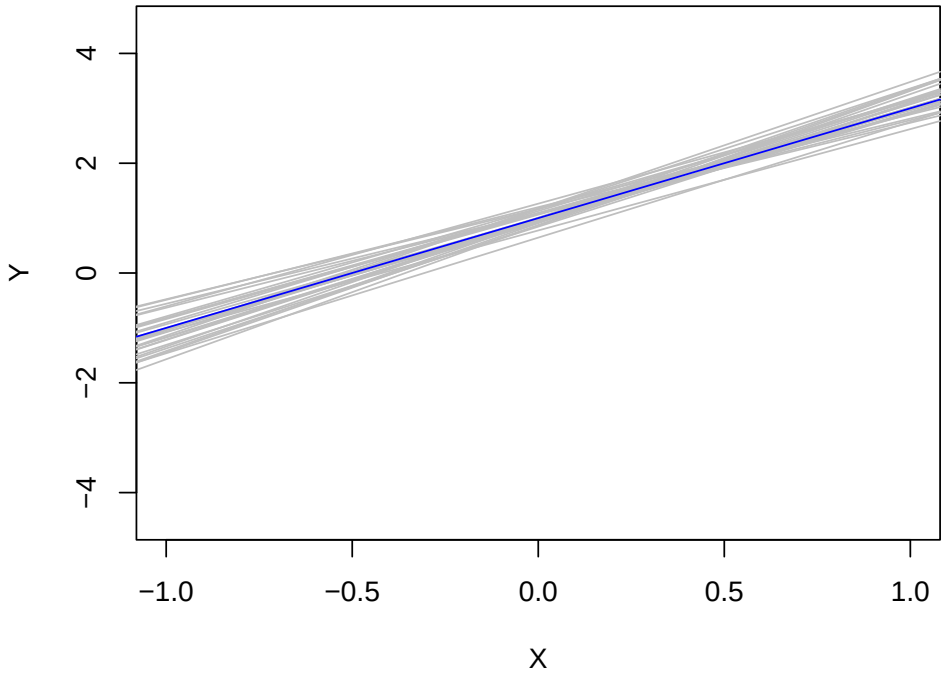
Trong hình dưới đây:

- các điểm màu đen là các mẫu được mô phỏng từ mô hình;
- đường màu xanh dương là đường hồi quy tổng thể thực sự: $Y = 1 + 2X$
- đường màu xanh lá là đường OLS ước lượng dựa trên dữ liệu



Tính vững và tính không chệch

Bây giờ, ta xem xét 30 lần mẫu được tạo dựa trên mô hình. Biểu đồ sau đây hiển thị đường tuyến tính tổng thể (màu xanh dương) và 30 đường OLS ước tính (màu xám).



$\hookrightarrow$ Tất cả các đường ước lượng OLS đều khác với đường hồi quy tổng thể, tuy nhiên chúng khá gần với đường đó.

Sai số chuẩn của ước lượng

Nếu ta thêm

giả định về tính đồng nhất (homoskedasticity)

$$(A4)\ \mathrm{Var}(\varepsilon_i) = \sigma^2 \text{ với mọi } i.$$

Nói cách khác, chúng ta đang giả định các sai số là có phương sai đồng nhất.

Thì ta có:

Sai số chuẩn của ước lượng OLS

Dưới giả định (A1)-(A3) và (A4), ta có thể chứng minh rằng sai số chuẩn của $\widehat{\beta}_{0,n}$ và $\widehat{\beta}_{1,n}$ là:

$$\begin{aligned} SE(\widehat{\beta}_{0,n}) &= \frac{\sigma\sqrt{\sum_{i=1}^{n} X_i^2}}{\sqrt{n\sum_{i=1}^{n}(X_i - \overline{X})^2}} \\ SE(\widehat{\beta}_{1,n}) &= \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}}, \end{aligned}$$

Sai số chuẩn của ước lượng

Bởi vì phương sai σ^2 là không biết, nên ta cần ước lượng nó từ dữ liệu:

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \widehat{\beta}_{0,n} - \widehat{\beta}_{1,n}X_i\right)^2$$

- Đặc biệt $\widehat{\sigma}$ được gọi là **sai số chuẩn của thặng dư - residual standard error** - RSE.
- $\widehat{\sigma}^2$ là ước lượng không chệch của σ^2.

Do đó, ta có được ước lượng của sai số chuẩn của ước lượng OLS

$$\begin{aligned}
\widehat{SE}(\widehat{\beta}_{0,n}) &= \frac{\widehat{\sigma}\sqrt{\sum_{i=1}^{n} X_i^2}}{\sqrt{n\sum_{i=1}^{n}(X_i - \overline{X})^2}} \\
\widehat{SE}(\widehat{\beta}_{1,n}) &= \frac{\widehat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}},
\end{aligned}$$

Phân phối mẫu của ước lượng OLS - sai số có phân phối chuẩn

- Để có được kết quả về phân phối mẫu của ước lượng OLS, trước hết, các giả định (A1) - (A4) cần được thỏa mãn.
- Thêm vào, nếu ta giả định rằng

Giả định phân phối chuẩn cho sai số

(A5) ε_i là độc lập cùng phân phối chuẩn với trung bình 0 và phương sai σ^2, tức là, $\mathcal{N}(0, \sigma^2)$.

Phân phối mẫu của ước lượng OLS, cỡ mẫu n hữu hạn

Nếu các giả định (A1) - (A4) và (A5) được thỏa mãn, khi đó, ta có phân phối mẫu của ước lượng OLS $\widehat{\beta}_0$ và $\widehat{\beta}_1$ là phân phối chuẩn:

$$\widehat{\beta}_{0,n} \sim \mathcal{N}(\beta_0, SE^2(\widehat{\beta}_{0,n})) \quad \text{and} \quad \widehat{\beta}_{1,n} \sim \mathcal{N}(\beta_1, SE^2(\widehat{\beta}_{1,n})).$$

Các kết quả này đúng với mọi cỡ mẫu n.

Phân phối mẫu của ước lượng OLS - sai số có phân phối chuẩn

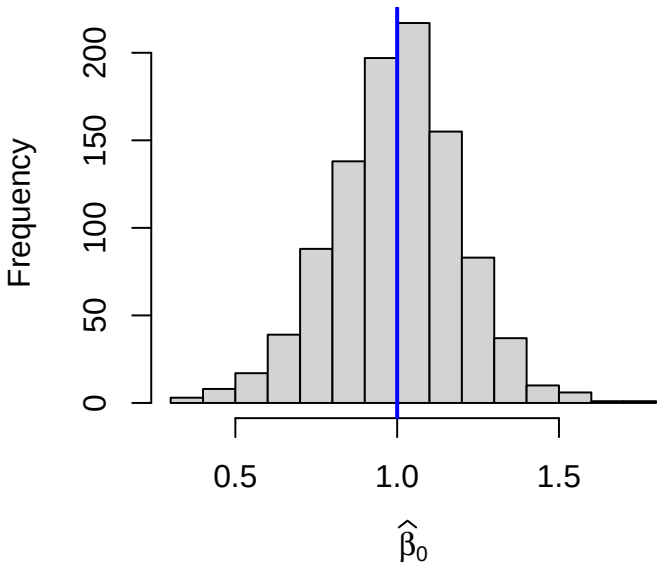
Ví dụ: Xét mô hình

$$Y = 1 + 2X + \varepsilon,$$

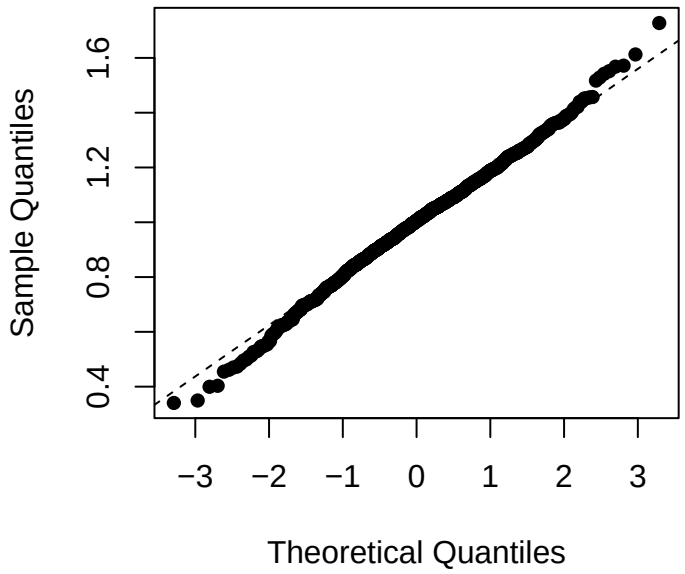
với

- $\varepsilon \sim \mathcal{N}(0, 1)$,
- $X \sim \mathcal{U}(-1, 1)$.

Tạo mẫu với $n = 25$, lặp lại 1,000 lần.

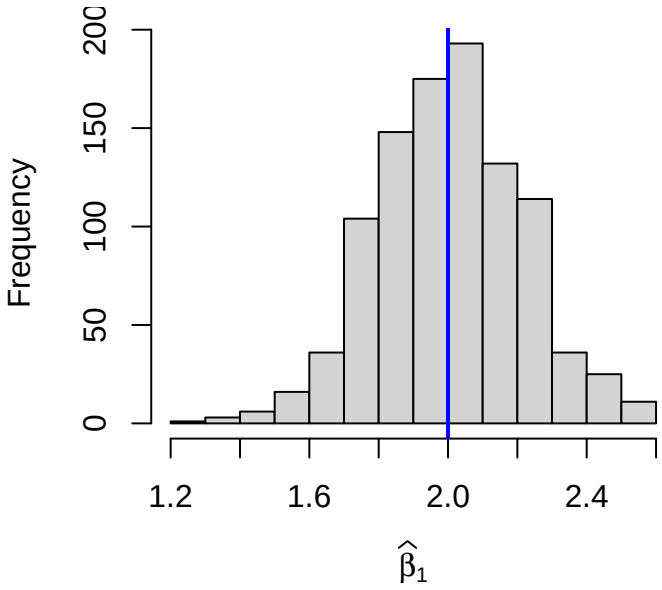


(a)

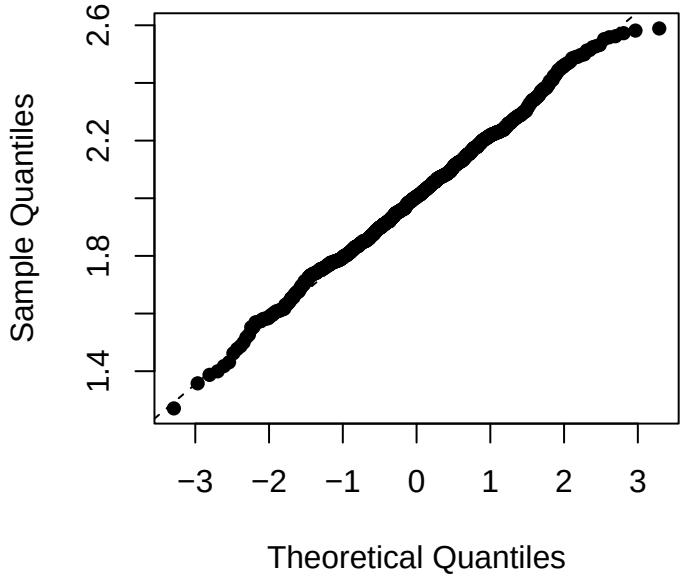


(b)

Phân phối mẫu của ước lượng OLS - sai số có phân phối chuẩn



(c)



(d)

Phân phối mẫu của ước lượng OLS - sai số không chuẩn

Trong thực tế, giả định về phân phối chuẩn của sai số thường bị vi phạm.

Trong những tình huống như thế, ta cần một giả định khác thay vì (A5), để suy ra được phân phối mẫu cho ước lượng OLS.

Giả định về điểm ngoại lại

(A6) Các giá trị ngoại lai lớn là hiếm khi xảy ra: X_i và ε_i có moment bậc 4 hữu hạn khác 0 (hoặc kurtosis hữu hạn).

Chú ý: Trong thực tế, giả định này giúp đảm bảo rằng các ngoại lai lớn không kéo đường hồi quy ra khỏi phần lớn các điểm.

Phân phối tiệm cận cho ước lượng OLS

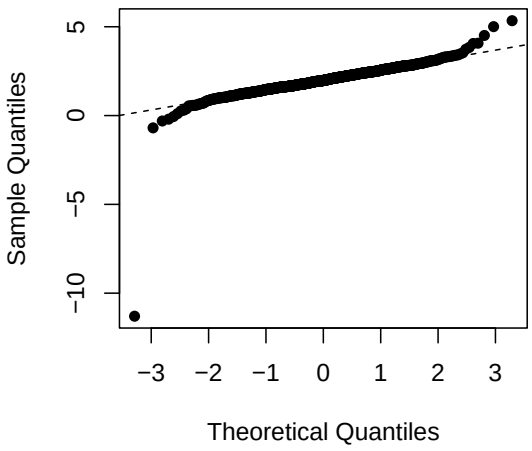
Nếu các giả định (A1) - (A4) và (A6) được thỏa mãn, bằng cách áp dụng CLT (định lý giới hạn trung tâm - Central limit theorem), ta có $\widehat{\beta}_0$ và $\widehat{\beta}_1$ là tiệm cận phân phối chuẩn, tức là,

$$\widehat{\beta}_{0,n} \xrightarrow{d} \mathcal{N}(\beta_0, SE^2(\widehat{\beta}_{0,n})) \quad \text{and} \quad \widehat{\beta}_{1,n} \xrightarrow{d} \mathcal{N}(\beta_1, SE^2(\widehat{\beta}_{1,n})),$$

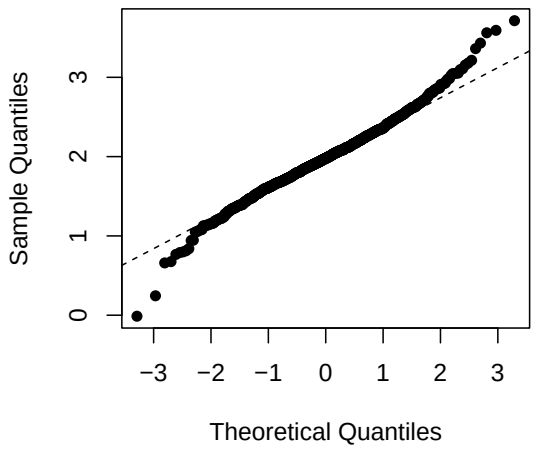
với $n \to \infty$.

Phân phối mẫu của ước lượng OLS - sai số không chuẩn

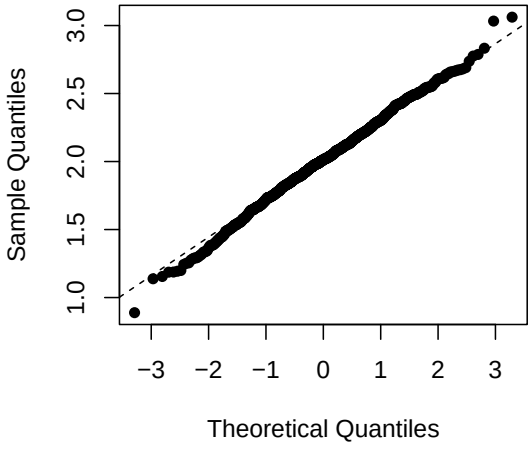
Phân phối mẫu của $\widehat{\beta}_{1,n}$ khi $\varepsilon_i \sim t_3$.



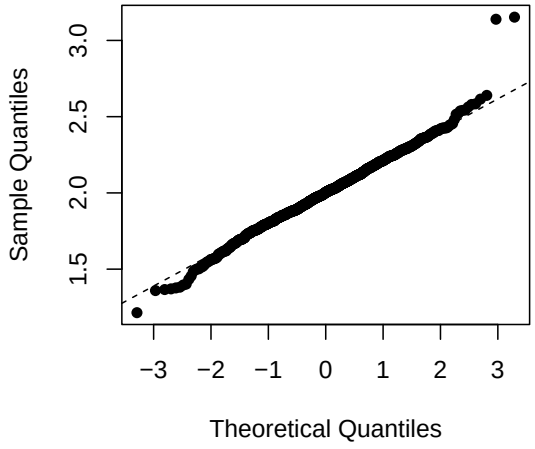
(a) $n = 25$



(b) $n = 50$



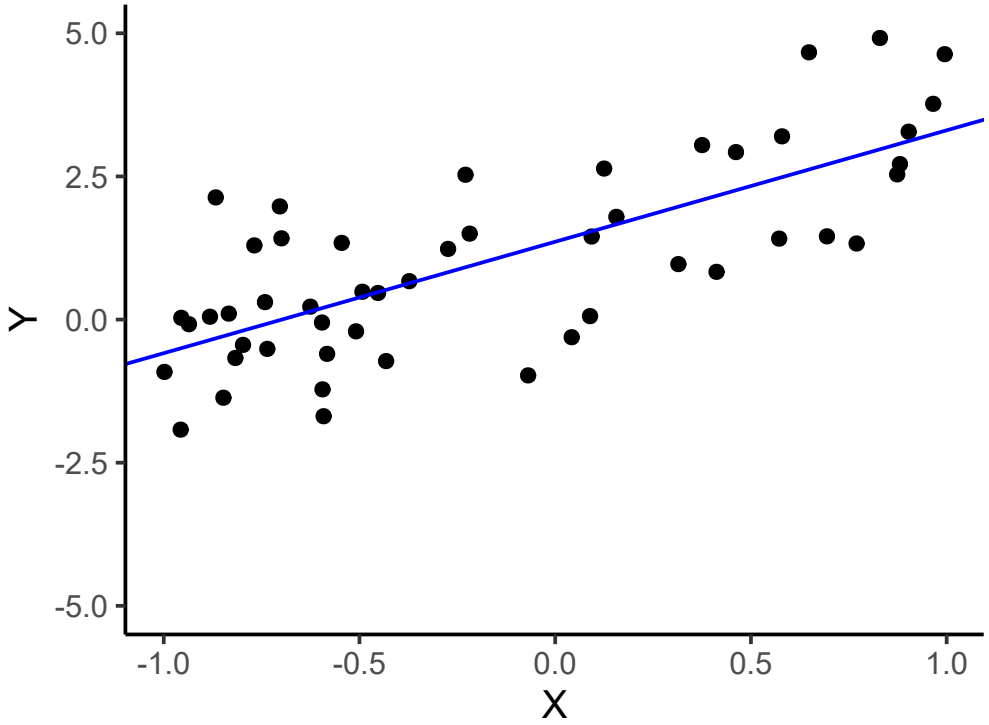
(c) $n = 100$



(d) $n = 200$

Ảnh hưởng của outliers lên ước lượng OLS

Dưới đây, ta mô phỏng sự ảnh hưởng của outliers tới ước lượng của đường hồi quy tuyến tính.



Khoảng tin cậy cho hệ số

Khoảng tin cậy cho β_0 và β_1

Dựa vào kết quả phân phối mẫu của $\widehat{\beta}_{0,n}$ và $\widehat{\beta}_{1,n}$, ta có thể xây dựng $100 \times c\%$ khoảng tin cậy cho hệ số mô hình β_0 và β_1 (tương tự như trong trung bình). Tức là,

$$\left(\widehat{\beta}_{0,n} - t_{(1+c)/2,n-2}\widehat{SE}(\widehat{\beta}_{0,n}), \widehat{\beta}_{0,n} + t_{(1+c)/2,n-2}\widehat{SE}(\widehat{\beta}_{0,n})\right)$$

và

$$\left(\widehat{\beta}_{1,n} - t_{(1+c)/2,n-2}\widehat{SE}(\widehat{\beta}_{1,n}), \widehat{\beta}_{1,n} + t_{(1+c)/2,n-2}\widehat{SE}(\widehat{\beta}_{1,n})\right),$$

trong đó, $t_{(1+c)/2,n-2}$ là phân vị thứ $(1+c)/2$ của phân phối t-student với bậc tự do $n-2$.

Khoảng tin cậy cho hệ số

Ví dụ: Xét dữ liệu với các cặp quan sát % màu đen trên mũi sư tử đực và tuổi (theo năm) của chúng.

Trong phần trước, ta đã ước lượng được

- $\widehat{\beta}_0 = 0.879$
- $\widehat{\beta}_1 = 10.647$

95% khoảng tin cậy cho

β_0: $(-0.2827, 2.0407)$

β_1: $(7.5643, 13.7299)$

Ta có thể kết luận như sau:

- Khi không quan sát thấy % màu đen trên mũi sư tử, thì tuổi trung bình của sư tử đực có thể giao động từ -0.2827 tới 2.0407 năm tuổi (không thực sự có nghĩa!!!).
- Với mỗi 0.1% tăng lên của vùng màu đen trên mũi, trung bình tuổi của sư tử sẽ thay đổi từ 0.76 tới 1.37 năm tuổi.

Tiên đoán từ mô hình

Nhắc lại mô hình hồi quy tuyến tính đơn

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Theo giả định (A3) về tính phân phối chuẩn của ε, ta có

$$Y|X = x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$$

với x cố định. Tức là phân phồi điều kiện của Y khi cố định $X = x$ là phân phối chuẩn.

Từ đây, ta có kỳ vọng điều kiện

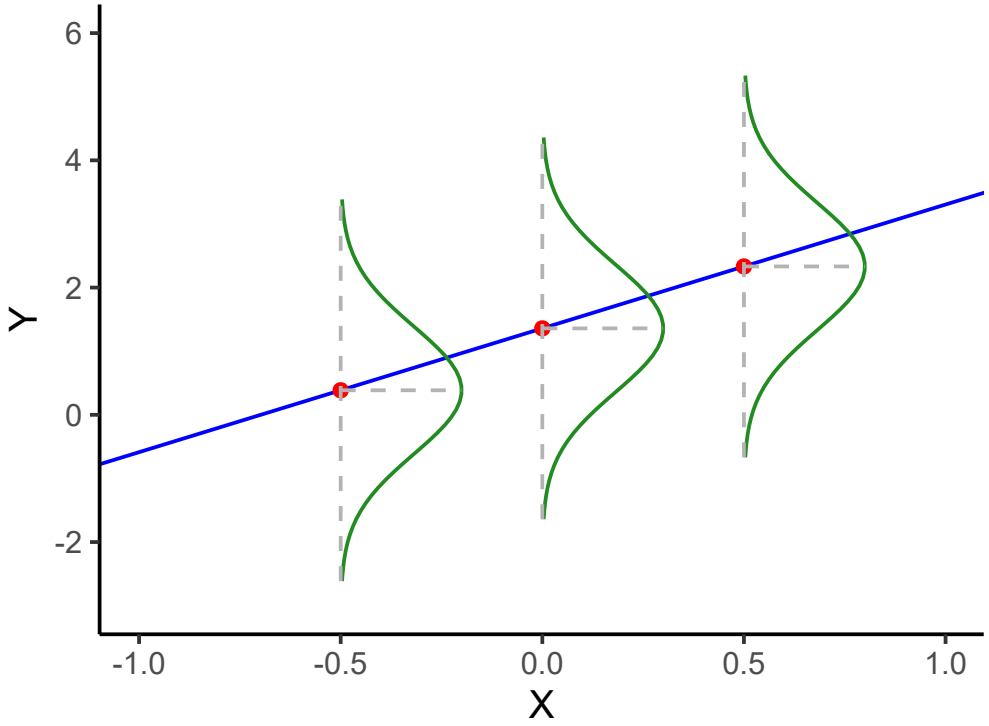
$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x$$

$\hookrightarrow$ $\widehat{Y} = \widehat{\beta}_{0,n} + \widehat{\beta}_{1,n} x$ là một ước lượng của kỳ vọng điều kiện $\mathbb{E}(Y|X = x)$.

Tiên đoán từ mô hình

Hình dưới đây, mô tả hàm mật độ điều kiện

$$Y|X = x \sim \mathcal{N}(1 + 2x, 1)$$



Tiên đoán từ mô hình

Ví dụ: Xét dữ liệu với các cặp quan sát % màu đen trên mũi sư tử đực và tuổi (theo năm) của chúng.

Các ước lượng OLS của hệ số mô hình là

- $\widehat{\beta}_0 = 0.879$
- $\widehat{\beta}_1 = 10.647$

khi đó, ta có

$$\widehat{y} = 0.879 + 10.647x$$

Cho $x = 0.4$, thì ta tính được

$$\widehat{y} = 0.879 + 10.647 \times 0.4 = 5.1378$$

Điều đó có nghĩa là, dựa trên mô hình tuyến tính, đối với một nhóm sư tử đực có 0.4% màu đen trên mũi, thì tuổi trung bình được dự đoán là 5.1378 năm tuổi.

Tiên đoán từ mô hình

Đại lượng $\widehat{Y} = \widehat{\beta}_{0,n} + \widehat{\beta}_{1,n}x$ là ngẫu nhiên.

Do đó, ta có phân phối mẫu cho $\widehat{Y}$.

- $\mathbb{E}(\widehat{Y}) = \beta_0 + \beta_1 x = \mathbb{E}(Y|X = x)$;
- $SE(\widehat{Y}) = \sqrt{\dfrac{\sigma^2}{n} + (x - \overline{X})^2 SE^2(\widehat{\beta}_{1,n})}$;
- $\widehat{SE}(\widehat{Y}) = \sqrt{\dfrac{\widehat{\sigma}^2}{n} + (x - \overline{X})^2 \widehat{SE}^2(\widehat{\beta}_{1,n})}$;
- $\widehat{Y}|x \sim \mathcal{N}\left(\beta_0 + \beta_1 x, \mathbb{V}\mathrm{ar}(\widehat{Y})\right)$

Từ đây, ta suy ra

$$\frac{\widehat{Y} - (\beta_0 + \beta_1 x)}{SE(\widehat{Y})} \sim \mathcal{N}(0, 1),$$

và

$$\frac{\widehat{Y} - (\beta_0 + \beta_1 x)}{\widehat{SE}(\widehat{Y})} \sim t_{n-2},$$

Tiên đoán từ mô hình

Từ đây, ta có thể xây dựng $100 \times c\%$ khoảng tin cậy cho trung bình điều kiện $\mathbb{E}(Y|X = x)$:

$$\left(\widehat{Y} - t_{(1+c)/2,n-2}\widehat{SE}(\widehat{Y}), \widehat{Y} + t_{(1+c)/2,n-2}\widehat{SE}(\widehat{Y})\right),$$

Khoảng tin cậy phản ánh sự không chắc chắn liên quan đến việc ước tính hệ số chặn và hệ số góc của đường hồi quy.

Ví dụ: Xét dữ liệu với các cặp quan sát % màu đen trên mũi sư tử đực và tuổi (theo năm) của chúng.

Ta có

$$\widehat{y} = 0.879 + 10.647x$$

Cho $x = 0.4$, thì ta tính được

$$\widehat{y} = 0.879 + 10.647 \times 0.4 = 5.1378$$

95% khoảng tin cậy cho trung bình độ tuổi của sử tử đực có 0.4% màu đen trên mũi là $(4.4894, 5.7863)$

Khoảng giá trị tiên đoán

- Trong thực tế, đôi khi, ta quan tâm tới ước lượng giá trị tương lai cho Y khi cố định $X = x$ hơn là $\mathbb{E}(Y|X = x)$.
  $\hookrightarrow$ theo mô hình hồi quy tuyến tính đơn, ta có

$$\widetilde{Y} = \beta_0 + \beta_1 x + \varepsilon$$

  với $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

- Như vậy, việc dự đoán một giá trị tương lai $\widetilde{Y}$ kém chính xác hơn nhiều so với việc ước lượng $\mathbb{E}(Y|X = x)$ bởi vì chúng ta cần tính thêm nguồn gây ra sự không chắc chắn từ thành phần sai số ngẫu nhiên ε.
- Từ mô hình ước lượng ta cũng có

$$\widetilde{Y} = \widehat{\beta}_{0,n} + \widehat{\beta}_{1,n} x + \widehat{\varepsilon},$$

  với $\widehat{\varepsilon} = \widetilde{Y} - \widehat{Y}$. Suy ra,

$$\begin{aligned} \mathbb{V}\mathrm{ar}(\widehat{\varepsilon}) &= \mathbb{V}\mathrm{ar}\left(\widetilde{Y}\right) + \mathbb{V}\mathrm{ar}\left(\widehat{\beta}_{0,n} + \widehat{\beta}_{1,n} x\right) \\ &= \sigma^2 + \frac{\sigma^2}{n} + (x - \overline{X})^2 SE^2(\widehat{\beta}_{1,n}) \end{aligned}$$

Khoảng giá trị tiên đoán

Theo mô hình, $\widetilde{Y}$ tuân theo phân phối chuẩn, và do đó

$$\frac{\widetilde{Y} - \widehat{Y}}{SE(\widetilde{Y})} \sim \mathcal{N}(0,1),$$

và

$$\frac{\widetilde{Y} - \widehat{Y}}{\widehat{SE}(\widetilde{Y})} \sim t_{n-2},$$

với $SE(\widetilde{Y}) = \sqrt{\mathbb{V}\mathrm{ar}\left(\widetilde{Y}\right)}$ và $\widehat{SE}(\widetilde{Y})$ là bản ước lượng.

Từ đây, ta có thể xây dựng $100 \times c\%$ khoảng tin cậy cho giá trị tiên đoán tương lai:

$$\left(\widehat{Y} - t_{(1+c)/2,n-2}\widehat{SE}(\widetilde{Y}), \widehat{Y} + t_{(1+c)/2,n-2}\widehat{SE}(\widetilde{Y})\right),$$

Khoảng giá trị tiên đoán

Ví dụ: Xét dữ liệu với các cặp quan sát % màu đen trên mũi sư tử đực và tuổi (theo năm) của chúng.

Ta có

$$\widehat{y} = 0.879 + 10.647x$$

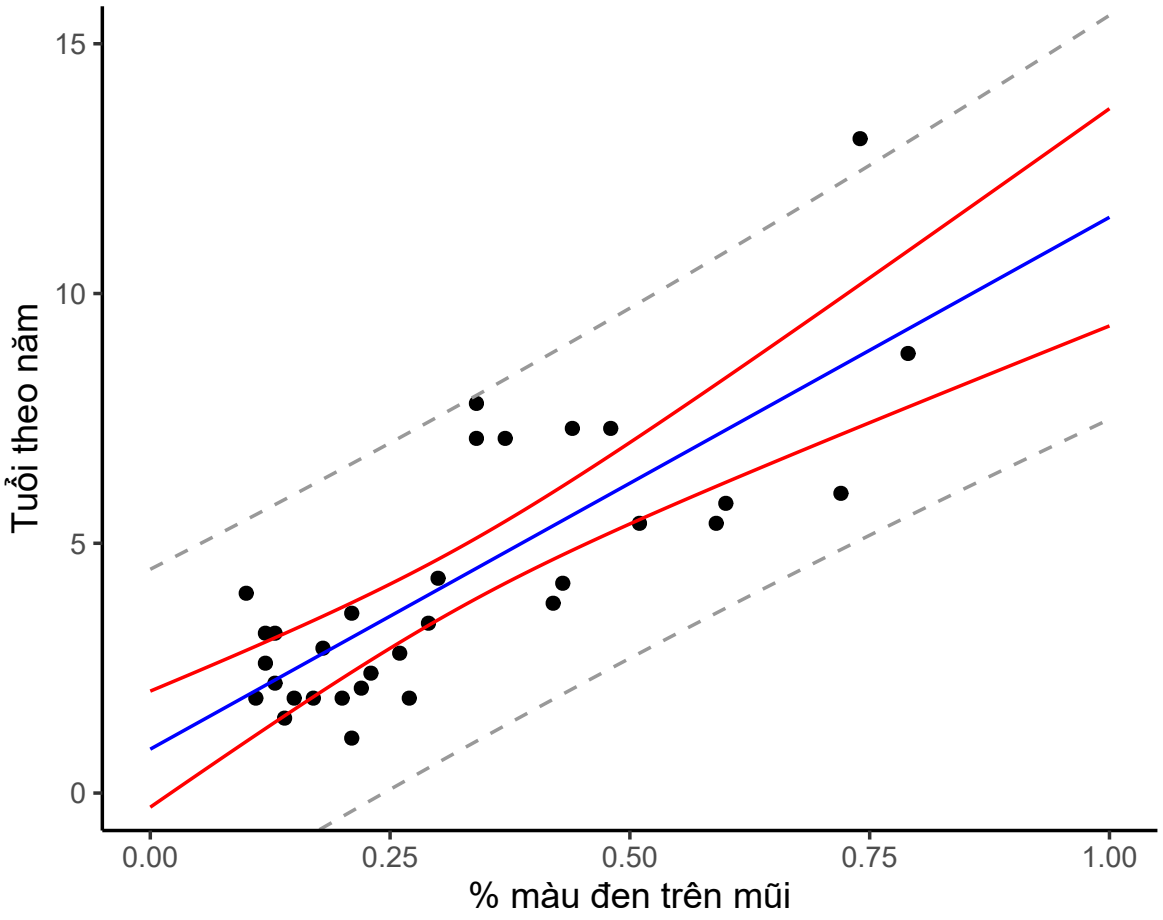
Cho $x = 0.4$, thì ta tính được

$$\widehat{y} = 0.879 + 10.647 \times 0.4 = 5.1378$$

95% khoảng tin cậy cho giá trị tiên đoán độ tuổi của sử tử đực có 0.4% màu đen trên mũi là $(1.6686, 8.6071)$.

Chú ý rằng, khoảng tin cậy cho giá trị tiên đoán chỉ thực sự hợp lý nếu như mô hình là thực sự đúng trong thực tế, điều mà ta khó có thể xác nhận.

Khoảng giá trị tiên đoán



Kiểm định giả thuyết

Sau khi đã xây dựng và ước lượng mô hình hồi quy tuyến tính đơn biến từ dữ liệu, ta thường quan tâm tới câu hỏi:

"Liệu rằng X có thực sự cần thiết cho việc dự đoán Y, trong quần thể?"

Nếu X là không cần thiết, thì điều đó tương đương với $\beta_1 = 0$.

$\hookrightarrow$ ta cần kiểm định giả thuyết

$$\begin{cases} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0 \end{cases}$$

Nhắc lại rằng, ta có

$$\frac{\widehat{\beta}_{1,n} - \beta_1}{\widehat{SE}\left(\widehat{\beta}_{1,n}\right)} \sim t_{n-2}$$

Nếu H_0 là đúng, thì

$$T = \frac{\widehat{\beta}_{1,n}}{\widehat{SE}\left(\widehat{\beta}_{1,n}\right)} \sim t_{n-2}$$

Kiểm định giả thuyết

Và khi đó, ta có giá trị quan sát t:

$$t = \frac{\widehat{\beta}_1 - 0}{\widehat{SE}(\widehat{\beta}_1)}$$

Khi này, ta xác định được p-value là $2\Pr(T \geq |t|)$.

Ta có thể bỏ H_0 khi p-value $< \alpha$,

$\hookrightarrow$ kết luận: X là cần thiết để dự đoán Y theo mô hình hồi quy tuyến tính đơn.

Áp dụng ý tưởng tương tự, ta có kiểm định giả thuyết

$$\begin{cases} H_0: & \beta_0 = 0 \\ H_1: & \beta_0 \neq 0 \end{cases}$$

tức là hệ số chặn có thực sự cần thiết trong mô hình.

Kiểm định giả thuyết

Ví dụ: Xét dữ liệu với các cặp quan sát % màu đen trên mũi sư tử đực và tuổi (theo năm) của chúng. Ta có

	Estimate	Standard error	t-statistic	p-value
β_0 - Intercept	0.8790	0.5688	1.545	0.133
β_1 - % màu đen	10.6471	1.5095	7.053	< 0.001

Như vậy, ta có thể kết luận rằng % màu đen trên mũi là cần thiết để dự đoán năm tuổi của sư tử, theo mô hình tuyến tính; trong khi đó, hệ số chặn là không cần thiết trong mô hình.

Đánh giá độ chính xác tổng thể của mô hình

Hệ số xác định

Hệ số xác định (**coefficient of determination**), được ký hiệu là R^2, là số đó cho % của sự biến động (biến thiên) trong Y mà có thể được giải thích (mô tả) bởi mô hình hồi quy tuyến tính với X.

Cụ thể, R^2 được tính theo công thức

$$R^2 = \frac{SSR(\widehat{\beta}_{0,n}, \widehat{\beta}_{1,n})}{TSS} = 1 - \frac{RSS(\widehat{\beta}_{0,n}, \widehat{\beta}_{1,n})}{TSS},$$

với

- SSR là tổng bình phương của mô hình (*sum of squares due to regression model*):

$$SSR(\widehat{\beta}_{1,n}, \widehat{\beta}_{1,n}) = \sum_{i=1}^{n} \left(\widehat{\beta}_{1,n} + \widehat{\beta}_{1,n} X_i - \overline{Y} \right)^2$$

- TSS là tổng bình phương tổng thể (*total sum of squares*) của Y

$$TSS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

Đánh giá độ chính xác tổng thể của mô hình

Công thức tính R^2 có được nhờ phân tích thành phần của TSS. Cụ thể:

$$\begin{aligned} TSS &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 \\ &= RSS(\widehat{\beta}_{0,n}, \widehat{\beta}_{1,n}) + SSR(\widehat{\beta}_{0,n}, \widehat{\beta}_{1,n}). \end{aligned}$$

R^2 luôn luôn nằm giữa 0 và 1:

- $R^2 = 1$ tức là 100% sự biến động giá trị của Y có thể được giải thích thông qua mô hình hồi quy tuyến tính (mô hình chính xác tuyệt đối).
- $R^2 = 0$ tức là 0% sự biến động giá trị của Y có thể được giải thích thông qua mô hình hồi quy tuyến tính (mô hình hoàn toàn không phù hợp).

Ví dụ, với mô hình tuyến tính dự đoán năm tuổi của sư tử, ta xác định được $R^2 = 0.6238$.

Điều đó có nghĩa là mô hình hồi quy tuyến tính đơn dựa trên % màu đen của mũi sư tử đực có thể giải thích được khoảng 62.38% độ biến động của năm tuổi của sư tử đực.

1 *Mô hình hồi quy tuyến tính đơn*
- Giới thiệu và ước lượng
- Tính chất của ước lượng
- Hồi quy đơn khi X là biến nhị phân

2 *Mô hình hồi quy tuyến tính đa biến*
- Mô hình cơ bản
- Mô hình với biến giải thích là định tính
- Mở rộng mô hình hồi quy

3 *Chuẩn đoán mô hình*

4 *Một số chủ đề thêm*

Hồi quy đơn khi X là biến nhị phân

- Trong phần trước, chúng ta đã nghiên cứu mô hình hồi quy tuyến tính đơn giản với biến giải thích X là biến định lượng.
- Tuy nhiên, trong thực tế, X có thể là một biến định tính với hai kết quả.

  X là giới tính: Nam hoặc Nữ

  X là trạng thái hút thuốc: Không hoặc Có.
- Phương pháp OLS yêu cầu tất cả các biến được trình bày dưới dạng số. Vì vậy, chúng ta phải mã hóa các yếu tố dự đoán dạng định tính thành **indicator variable** hoặc **dummy variable**.
- Ví dụ, với biến Giới tính, chúng ta có thể mã hóa Nam $= 0$ và Nữ $= 1$.

Hồi quy đơn khi X là biến nhị phân

Ví dụ - Tác động của vitamin C tới sự phát triển răng của chuột lang nhà - Guinea pigs

Crampton, E. W. (1947) đã tiến hành một nghiên cứu trên chuột lang để xem tác dụng của vitamin C đối với sự phát triển của răng. Phản ứng là chiều dài của nguyên bào ngà (tế bào chịu trách nhiệm cho sự phát triển của răng) ở 60 con chuột lang. Mỗi con vật nhận được một trong ba mức liều vitamin C (0,5, 1 và 2 mg/ngày) bằng một trong hai phương pháp phân phối, nước cam hoặc axit ascorbic (một dạng vitamin C và được mã hóa là VC).

Đặt

- Y là chiều dài của nguyên bào ngà của một con chuột lang.
- X là phương thức nhận được vitamin C.

Ta ta muốn xem liệu mô hình hồi quy tuyến tính đơn có thể dùng để tiên đoán Y theo X hay không.

Hồi quy đơn khi X là biến nhị phân

Rõ ràng, X là biến định lượng với hai kết quả: OJ (orange juice - nước cam) và VC (ascorbic acid).

$\hookrightarrow$ ta cần tạo giả biến (dummy variable):

$$D_i = \begin{cases} 0 & \text{nếu con chuột thứ } i \text{ nhận vitamin C bởi OJ} \\ 1 & \text{nếu con chuột thứ } i \text{ nhận vitamin C bởi VC} \end{cases}$$

Lúc này, mô hình trở thành

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 D_i + \varepsilon_i \\ &= \begin{cases} \beta_0 + \varepsilon_i & \text{nếu con chuột thứ } i \text{ nhận vitamin C bởi OJ} \\ \beta_0 + \beta_1 + \varepsilon_i & \text{nếu con chuột thứ } i \text{ nhận vitamin C bởi VC} \end{cases} \end{aligned}$$

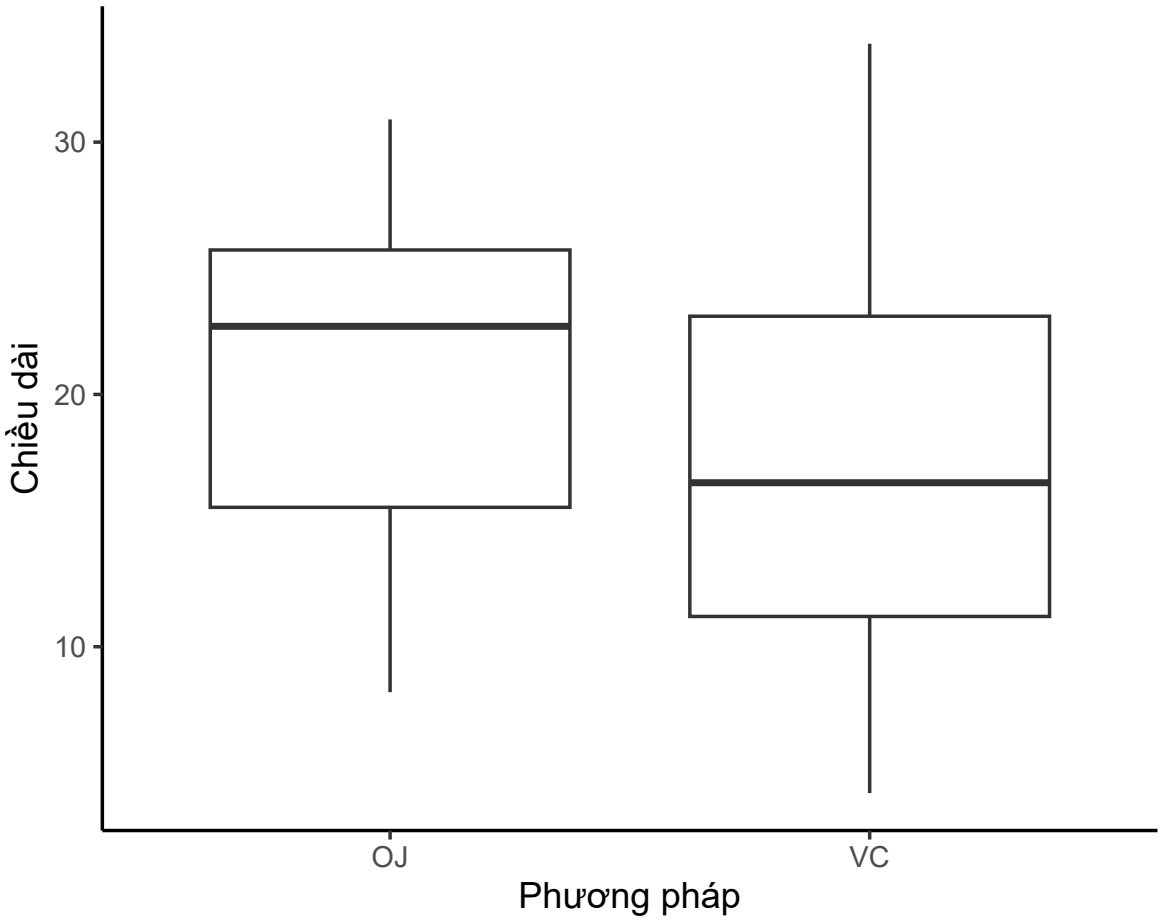
Hồi quy đơn khi X là biến nhị phân

- β_0 biểu diễn trung bình chiều dài nguyên bào ngà của những con chuột lang nhận vitamin C bởi phương pháp OJ.
- $\beta_0 + \beta_1$ biểu diễn trung bình chiều dài nguyên bào ngà của những con chuột lang nhận vitamin C bởi phương pháp VC.
- β_1 biểu diễn sự khác biệt trong trung bình chiều dài nguyên bào ngà của những con chuột lang nhận vitamin C bởi phương pháp VC so với OJ.

Chú ý: mô hình này cũng có thể được áp dụng cho bài toán so sánh hai trung bình.

Khi đó, giả định $H_0 : \beta_1 = 0$ là tương đương $H_0 : \mu_{OJ} = \mu_{VC}$.

Hồi quy đơn khi X là biến nhị phân



Hồi quy đơn khi X là biến nhị phân

Sau quá trình tính toán với phần mềm R, ta thu được:

	Estimate	Standard error	t-statistic	p-value
β_0	20.663	1.366	15.127	< 0.0001
β_1	-3.700	1.932	-1.915	0.0604

Dựa vào kết quả, ta có một nhận xét như sau:

- Chiều dài trung bình của nguyên bào ngà ở chuột lang nhận vitamin C bằng phương pháp OJ là 20.663 đơn vị.
- Chiều dài trung bình của nguyên bào ngà ở chuột lang nhận vitamin C bằng phương pháp VC là $20.663 - 3.7 = 16.963$.
- Chênh lệch trung bình chiều dài của nguyên bào ngà giữa hai phương pháp là $-3,7$, tức là phương pháp OJ có hiệu quả cao hơn phương pháp VC. Tuy nhiên, sự khác biệt chỉ có ý nghĩa thống kê ở mức ý nghĩa $\alpha = 0.1$.

1 *Mô hình hồi quy tuyến tính đơn*
- Giới thiệu và ước lượng
- Tính chất của ước lượng
- Hồi quy đơn khi X là biến nhị phân

2 *Mô hình hồi quy tuyến tính đa biến*
- Mô hình cơ bản
- Mô hình với biến giải thích là định tính
- Mở rộng mô hình hồi quy

3 *Chuẩn đoán mô hình*

4 *Một số chủ đề thêm*

Giới thiệu

- Mô hình hồi quy tuyến tính đơn giản có thể dễ sử dụng để tiên đoán Y dựa trên một biến giải thích X.
- Tuy nhiên, trong các bài toán thực tế, ta thường có nhiều hơn một biến giải thích có thể được sử dụng trong mô hình.
- Ví dụ, chiều dài thân của một loài cá (Y) có thể được tiến đoán bởi mô hình tuyến tính dựa trên tuổi và nhiệt độ nước:

$$Y = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Temp}$$

- Hoặc hàm lượng mỡ trong máu - blood fat (Y) có thể liên quan tới tuổi tác và cân nặng. Trong trường hợp này, ta có thể có một mô hình hồi quy tuyến tính

$$Y = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Weight}$$

Giới thiệu

Mô hình hồi quy tuyến tính đa biến

Về cơ bản, mô hình hồi quy tuyến tính đa biến là phiên bản mở rộng của hồi quy tuyến tính đơn giản:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon,$$

trong đó,

- Y là biến đáp ứng;
- $X_1, \ldots, X_p$ là p biến giải thích
- $\beta_0, \beta_1, \ldots, \beta_p$ là các hệ số trong mô hình.

Cụ thể:

- β_0 là hệ số chặn của mô hình (**intercept coefficient**), biểu thị giá trị của Y khi tất cả các biến giải thích $X_1 = X_2 = \ldots = X_p = 0$.
- β_j là hệ số góc (**slope coefficient**) của từng biến giải thích X_j, với $j = 1, 2, \ldots, p$.
- β_j biểu diễn trung bình mức tăng (giảm) của Y khi X_j tăng lên một đơn vị và các biến giải thích còn lại được giữ cố định giá trị.

Ước lượng mô hình

Giả sử ta có một bộ dữ liệu gồm n quan sát: $(Y_i, X_{1i}, X_{2i}, \ldots, X_{pi})$, với $i = 1, 2, \ldots, n$:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon_i,$$

Đặt

- $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$;
- $\mathbf{X}$ là ma trận thiết kế với cỡ $n \times (p+1)$:

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{21} & \ldots & X_{p1} \\ 1 & X_{12} & X_{22} & \ldots & X_{p2} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & X_{1n} & X_{2n} & \ldots & X_{pn} \end{pmatrix}$$

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^\top$ là vectơ chứa các hệ số của mô hình;
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$

Ước lượng mô hình

Khi này, ta có thể biểu diễn mô hình theo dạng ma trận

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon.$$

Khi này, hàm tổng bình phương sai số sẽ là

$$RSS(\beta) = \varepsilon^\top \varepsilon = (\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta)$$

Phương trình đạo hàm riêng có dạng là

$$\frac{\partial RSS(\beta)}{\partial \beta} = \mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\beta) = \mathbf{0}_{p+1}$$

Từ đây, suy ra,

$$\widehat{\beta}_n = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}$$

Ước lượng mô hình

Ví dụ

Chiều dài của một loài cá được biểu diễn dưới dạng mô hình tuyến tính của tuổi và nhiệt độ nước. Cá được nuôi trong bể ở nhiệt độ 25, 27, 29 và 31 độ C. Sau khi sinh, một mẫu thử nghiệm (44 con cá) được chọn ngẫu nhiên cứ sau 14 ngày và đo chiều dài của nó.

- Một mô hình tuyến tính sẽ là:

$$\text{Length} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Temp} + \varepsilon$$

- Ước lượng OLS cho các hệ số là:

	Estimate
β_0 (Intercept)	3904.27
β_1 (Age)	26.24
β_2 (Temp)	−106.41

- Ta thấy rằng nếu nhiệt độ tăng lên một đơn vị, chiều dài của cá sẽ giảm đi 106.41 đơn vị.
- Nếu xét theo độ tuổi, rõ ràng, các con cá già sẽ dài hơn những con cá trẻ.

Phân phối mẫu của ước lượng

Cũng như phần trước, trong phần này ta cần mở rộng các giả định để tương ứng với mô hình hồi quy tuyến tính đa biến.

(A1) Có sự quan hệ tuyến tính giữa Y và X_1, X_2, ..., X_p. (Linearity).

(A2) Quan sát $(X_{1i}, X_{2i}, \ldots, X_{pi}, Y_i)$ với $i = 1, 2, \ldots, n$, là độc lập, và cùng phân phối. (i.i.d data)

(A3) Thành phần sai số ε_i là độc lập với mọi biến giải thích $X_{1i}, X_{2i}, \ldots, X_{pi}$, và có trung bình 0, tức là, $\mathbb{E}(\varepsilon_i) = 0$ với mọi i.

(A4) $\mathrm{Var}(\varepsilon_i) = \sigma^2$ với mọi i. (Homoskedasticity)

(A5) ε_i là độc lập, cùng phân phối chuẩn với trung bình 0 và phương sai σ^2, tức là, $\mathcal{N}(0, \sigma^2)$. (Normality assumption)

(A6) Các giá trị ngoại lai lớn hiếm khi xuất hiện.

Ngoài ra, ta cũng cần giả định thêm rằng $p < n$, nếu không, mô hình sẽ có nhiều hơn một ước lượng OLS.

Phân phối mẫu của ước lượng

Ta có các kết quả sau:

- $\widehat{\boldsymbol{\beta}}_n$ là ước lượng không chệch, tức là $\mathbb{E}(\widehat{\boldsymbol{\beta}}_n) = \boldsymbol{\beta}$;
- $\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}}_n) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$;
- $\widehat{\boldsymbol{\beta}}_n \sim \mathcal{N}_{p+1}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right)$;
- $\dfrac{\widehat{\beta}_{j,n} - \beta_j}{\widehat{\sigma}\sqrt{v_{jj}}} \sim t_{n-p-1}$, với $\widehat{\sigma}^2$ là ước lượng không chệch của σ^2:

$$\widehat{\sigma}^2 = \frac{1}{n-p-1} RSS\left(\widehat{\boldsymbol{\beta}}_n\right)$$

và v_{jj} là thành phần thứ j trên đường chéo của $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$

$\hookrightarrow$ ta có thể xây dựng khoảng tin cậy cho các hệ số; và cũng có thể thực hiện các kiểm định thống kê để trả lời một số câu hỏi quan trọng.

Một số câu hỏi quan trọng

Ta thường quan tâm tới các câu hỏi sau khi thực hiện mô hình hồi quy tuyến tính đa biến:

1. Có ít nhất một trong các biến giải thích $X_1, X_2, \ldots, X_p$ là hữu ích trong việc dự đoán biến đáp ứng Y?
2. Tất cả các biến giải thích có giúp dự đoán Y hay chỉ một tập hợp con của các biến giải thích là hữu ích?
3. Mô hình tuyến tính này phù hợp với dữ liệu đến mức nào?
4. Với một tập hợp các giá trị của biến giải thích, giá trị dự đoán của biến đáp ứng sẽ là bao nhiêu, và dự đoán đó chính xác đến mức nào?

Câu hỏi 1

Câu hỏi này tương đương với kiểm định giả thuyết

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

Nếu H_0 bị bác bỏ có nghĩa rằng có ít nhất 1 biến giải thích là hữu ích trong việc tiên đoán Y.

Ta nhận thấy rằng, giả thuyết H_0 này tương tự như giả định của kiểm định ANOVA một nhân tố.

$\hookrightarrow$ ta sẽ áp dụng nguyên tắc của kiểm định ANOVA một nhân tố:

- tính thống kê F với bậc tự do p và $n - p - 1$;
- xác định p-value.

Ví dụ

Xem xét mô hình cho chiều dài của cá theo độ tuổi và cân nặng:

$$\text{Length} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Temp} + \varepsilon$$

Giả huyết tương ứng là $H_0 : \beta_1 = \beta_2 = 0$.

Sau quá trình tính toán, ta thu được p-value là bé hơn 0.0001.

Câu hỏi 2

Câu hỏi 2 tương đương với việc thực hiện kiểm định

$$H_0 : \beta_j = 0$$

với $j = 1, 2, \ldots, p$.

Tương tự như trong trường hợp của hồi quy đơn biến, trong trường hợp này, ta cũng sử dụng phân phối mẫu của $\widehat{\beta}_j$, tức là phân phối t_{n-p-1} để thực hiện kiểm định.

Ví dụ: Xét mô hình hồi quy dự đoán chiều dài của cá:

$$\text{Length} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Temp} + \varepsilon$$

Ta quan tâm tới giả thuyết $H_0 : \beta_j = 0$, with $j = 1, 2$. Sau quá trình tính toán, ta thu được:

	Estimate	Standard error	t-value	p-value
β_0 (Intercept)	3904.27	1149.044	3.398	0.0015
β_1 (Age)	26.24	2.055	12.769	< 0.0001
β_2 (Temp)	-106.41	40.452	-2.631	0.0120

Dựa vào p-values, ta có thể nói rằng, cả độ tuổi và nhiệt độ đều cần thiết cho mô hình dự đoán này.

Câu hỏi 3

- Câu trả lời cho câu hỏi số 3 là hệ số xác định R^2.
- Giống như trong mô hình tuyến tính đơn giản, R^2 đánh giá mức độ phù hợp của mô hình thông qua đo lường phần phương sai của Y được giải thích bởi mô hình.
- Tuy nhiên, R^2 sẽ luôn tăng khi có nhiều biến hơn được thêm vào mô hình, ngay cả khi những biến đó chỉ liên quan yếu hoặc không liên quan đến biến đáp ứng Y.

Ta dùng hệ số R^2 hiệu chỉnh (adjusted-R^2), ký hiệu là R_a^2:

$$R_a^2 = 1 - \frac{n-1}{n-p}\left(1 - R^2\right)$$

- Giống như R^2, R_a^2 là thước đo tỷ lệ biến thiên được giải thích bằng mô hình hồi quy tuyến tính.
- Khác với R^2, R_a^2 phải trả giá cho việc đưa các biến không cần thiết vào mô hình, tức là giá trị của nó sẽ giảm xuống.

Ví dụ, với mô hình hồi quy cho chiều dài cá, ta tính được R^2 là 0.8056, trong khi đó, R_a^2 là 0.7962.

Câu hỏi 4

Nhắc lại mô hình hồi quy tuyến tính đa biến:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon.$$

Theo giả định (A3) về tính phân phối chuẩn của ε, ta có

$$Y|X_1 = x_1, \ldots, X_p = x_p \sim \mathcal{N}(\beta_0 + \beta_1 x_1 + \ldots \beta_p x_p, \sigma^2)$$

với $x_1, x_2, \ldots, x_p$ cố định. Tức là phân phối điều kiện của Y khi cố định $X_1 = x_1, \ldots, X_p = x_p$, là phân phối chuẩn.

Từ đây, ta có kỳ vọng điều kiện

$$\mathbb{E}(Y|X_1 = x_1, \ldots, X_p = x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

$\hookrightarrow \widehat{Y} = \widehat{\beta}_{0,n} + \widehat{\beta}_{1,n} x_1 + \ldots + \widehat{\beta}_{p,n} x_p$ là một ước lượng của kỳ vọng điều kiện $\mathbb{E}(Y|X_1 = x_1, \ldots, X_p = x_p)$.

Tương tự như đối với mô hình hồi quy đơn biến, ta cũng xây dựng được $100 \times c\%$ khoảng tin cậy cho $\mathbb{E}(Y|X_1 = x_1, \ldots, X_p = x_p)$.

Câu hỏi 4

Ví dụ: Xét mô hình hồi quy tuyến tính dự đoán chiều dài của cá:

$$\text{Length} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Temp} + \varepsilon$$

Ta thu được

- $\widehat{\beta}_0 = 3904.27$;
- $\widehat{\beta}_1 = 26.24$;
- $\widehat{\beta}_2 = -106.41$.

Giả sử chúng ta có một nhóm cá 15 ngày sống trong bể ở nhiệt độ 25°C. Dựa vào mô hình, ta ước tính có được độ dài trung bình của nhóm này là

$$\widehat{y} = 3904.27 + 26.24 \times 15 - 106.41 \times 25 = 1637.535$$

95% khoảng tin cậy cho chiều dài trung bình của nhóm cá này: $(1220.988, 2054.083)$.

1 Mô hình hồi quy tuyến tính đơn
- Giới thiệu và ước lượng
- Tính chất của ước lượng
- Hồi quy đơn khi X là biến nhị phân

2 Mô hình hồi quy tuyến tính đa biến
- Mô hình cơ bản
- Mô hình với biến giải thích là định tính
- Mở rộng mô hình hồi quy

3 Chuẩn đoán mô hình

4 Một số chủ đề thêm

Mô hình với biến giải thích là định tính

- Khi áp dụng hồi quy tuyến tính đa biến, trong thực tế, các biến giải thích $X_1, X_2, \ldots, X_p$ không chỉ là các biến định lượng mà còn là các biến định tính.
- Trong hồi quy tuyến tính đơn giản, chúng ta đã nghiên cứu hồi quy với một biến dự đoán nhị phân (tức là biến định tính có hai kết quả).
- Bây giờ, chúng ta sẽ nghiên cứu các trường hợp tổng quát hơn trong đó biến giải thích dạng định tính có nhiều hơn hai kết quả.
- Giả sử X là một biến giải thích dạng định tính với k kết quả (hoặc mức độ), chẳng hạn như $k > 2$.
- Ý tưởng đưa X vào mô hình tuyến tính đa biến, là ta tạo ra các biến giả. $\hookrightarrow$ cho kết quả k, ta sẽ tạo $k-1$ biến giả tương ứng với $k-1$ kết quả và một kết quả sẽ được coi là "**reference**" hoặc "**baseline**" (khi tất cả các biến giả bằng 0).

Mô hình với biến giải thích là định tính

Ví dụ, nếu ta muốn tiên đoán hàm lượng mỡ trong máu dựa trên thông tin về chủng tộc: African American, Asian và Caucasian.

- Kết quả tham chiếu: African American;
- Giả biến đầu tiên tương ứng là Asian:

$$X_{1i} = \begin{cases} 1 & \text{nếu người thứ } i \text{ là Asian} \\ 0 & \text{nếu người thứ } i \text{ không phải Asian} \end{cases}$$

- Giả biến thứ hai tương ứng là Caucasian:

$$X_{2i} = \begin{cases} 1 & \text{nếu người thứ } i \text{ là Caucasian} \\ 0 & \text{nếu người thứ } i \text{ không phải Caucasian} \end{cases}$$

Mô hình với biến giải thích là định tính

Khi đó, ta có mô hình:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \\
&= \begin{cases} \beta_0 + \varepsilon_i & \text{nếu người thứ } i \text{ là African American} \\ \beta_0 + \beta_1 + \varepsilon_i & \text{nếu người thứ } i \text{ là Asian} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{nếu người thứ } i \text{ là Caucasian} \end{cases}
\end{aligned}
$$

Do đó, β_0 biểu diễn tác động của chủng tộc African American.

Mô hình với biến giải thích là định tính

Xem xét lại mô hình tuyến tính cho chiều dài của cá theo độ tuổi và nhiệt độ nước:

$$\text{Length} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Temp}$$

Trong mô hình này, ta đã xử lý nhiệt độ (Temp) như một biến định lượng.

Về mặt tổng quát, điều này là hợp lý.

Tuy nhiên, trong nghiên cứu này, nhiệt độ đã được cố định ở bốn mức: 25, 27, 29, và 31.

Do đó, ta nên coi nhiệt độ trong nghiên cứu này như một biến định tính, với nhiệt độ tham chiếu là 25.

Như vậy, ta sẽ có 3 giả biến:

$$\text{Temp}_{2i} = \begin{cases} 1 & \text{nếu con cá thứ } i \text{ sống trong nước } 27^\circ\text{C} \\ 0 & \text{nếu con cá thứ } i \text{ không sống trong nước } 27^\circ\text{C} \end{cases}$$

$$\text{Temp}_{3i} = \begin{cases} 1 & \text{nếu con cá thứ } i \text{ sống trong nước } 29^\circ\text{C} \\ 0 & \text{nếu con cá thứ } i \text{ không sống trong nước } 29^\circ\text{C} \end{cases}$$

Mô hình với biến giải thích là định tính

và

$$\text{Temp}_{4i} = \begin{cases} 1 & \text{nếu con cá thứ } i \text{ sống trong nước } 31^\circ\text{C} \\ 0 & \text{nếu con cá thứ } i \text{ không sống trong nước } 31^\circ\text{C} \end{cases}$$

Khi đó, mô hình hồi quy tuyến tính sẽ là

$$\text{Length}_i = \beta_0 + \beta_1 \text{Age}_i + \beta_{22}\text{Temp}_{2i} + \beta_{23}\text{Temp}_{3i} + \beta_{24}\text{Temp}_{4i} + \varepsilon_i$$

Chú ý rằng Temp_{2i}, Temp_{3i} và Temp_{4i} không thể bằng 1 cùng lúc.

Mô hình có thể được biểu diễn bởi:

$$\text{Length}_i = \begin{cases} \beta_0 + \beta_1 \text{Age}_i + \varepsilon_i & \text{nếu con cá thứ } i \text{ sống trong nước } 25^\circ\text{C} \\ \beta_0 + \beta_1 \text{Age}_i + \beta_{22} + \varepsilon_i & \text{nếu con cá thứ } i \text{ sống trong nước } 27^\circ\text{C} \\ \beta_0 + \beta_1 \text{Age}_i + \beta_{23} + \varepsilon_i & \text{nếu con cá thứ } i \text{ sống trong nước } 29^\circ\text{C} \\ \beta_0 + \beta_1 \text{Age}_i + \beta_{24} + \varepsilon_i & \text{nếu con cá thứ } i \text{ sống trong nước } 31^\circ\text{C} \end{cases}$$

Như vậy, về cơ bản ta có 4 mô hình hồi quy tuyến tính đơn trên cùng một tập dữ liệu.

Mô hình với biến giải thích là định tính

Kết quả ước lượng mô hình mới như sau:

	Estimate	Standard error	t-value	p-value
β_0 (Intercept)	1035.43	223.482	4.633	< 0.0001
β_1 (Age)	26.24	1.845	14.221	< 0.0001
β_{22} (Temp27)	106.36	229.716	0.463	0.6459
β_{23} (Temp29)	186.91	229.716	0.814	0.4208
β_{24} (Temp31)	-736.273	229.716	-3.205	0.0027

Hãy nhớ rằng, dựa trên mô hình đầu tiên, trong đó nhiệt độ được coi là biến biến định lượng, ta ước lượng ra rằng nhiệt độ có tác động nghịch đến chiều dài của cá.

Với mô hình mới, chúng ta có thể thấy rõ tác động nghịch của nhiệt độ, trong đó, việc giảm chiều dài cá có ý nghĩa thống kê ở mức 31°C. Ngược lại, khi nhiệt độ tăng từ 25°C lên 27°C hoặc 29°C thì sự thay đổi về chiều dài cá không có ý nghĩa thống kê.

1 Mô hình hồi quy tuyến tính đơn
- Giới thiệu và ước lượng
- Tính chất của ước lượng
- Hồi quy đơn khi X là biến nhị phân

2 Mô hình hồi quy tuyến tính đa biến
- Mô hình cơ bản
- Mô hình với biến giải thích là định tính
- Mở rộng mô hình hồi quy

3 Chuẩn đoán mô hình

4 Một số chủ đề thêm

Mở rộng mô hình hồi quy

Trong phần trước, chúng ta đã tìm hiểu về mô hình hồi quy đa biến

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon.$$

Đây thực chất là mô hình cơ bản, trong đó mối quan hệ giữa biến giải thích và biến đáp ứng là:

- **additive**: hiệu ứng cộng đồn, tức là tác động của biến giải thích X_j lên trên biến đáp ứng Y không phụ thuộc vào giá trị của các biến giải thích khác;
- **linear**: sự thay đổi trong biến đáp ứng Y do sự thay đổi một đơn vị của X_j là không đổi, bất kể giá trị của X_j.

Mở rộng mô hình hồi quy

Trong phần trước, chúng ta đã tìm hiểu về mô hình hồi quy đa biến

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon.$$

Đây thực chất là mô hình cơ bản, trong đó mối quan hệ giữa biến giải thích và biến đáp ứng là:

- **additive**: hiệu ứng cộng đồn, tức là tác động của biến giải thích X_j lên trên biến đáp ứng Y không phụ thuộc vào giá trị của các biến giải thích khác;
- **linear**: sự thay đổi trong biến đáp ứng Y do sự thay đổi một đơn vị của X_j là không đổi, bất kể giá trị của X_j.

Tuy nhiên, trong thực tế, hồi quy tuyến tính có thể được mở rộng để đáp ứng:

- **interaction** - sự tương tác giữa hai biến giải thích: định tính - định lượng, định lượng - định lượng hoặc định tính - định tính.
- **non-linear** - mối quan hệ phi tuyến của biến đáp ứng Y với một biến giải thích X_j.

Mô hình với thành phần tương tác

Như chúng ta thấy trong bài giảng về ANOVA hai chiều, có thể xảy ra việc tác động của nhân tố X_1 lên biến đáp ứng Y là bị phụ thuộc vào X_2. Và ta đã biết đó là hiệu ứng **tương tác - interaction effect**.

Trong mô hình hồi quy tuyến tính đa biến, thành phần tương tác của X_1 và X_2 được ký hiệu là X_1X_2 và thể hiện trong mô hình

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon,$$

β_{12} biểu diễn độ mạnh của hiệu ứng tương tác lên trên biến đáp ứng Y.

Về nguyên tác, ta có thể có ba dạng tương tác sau:

- **tương tác định tính - định lượng**: ám chỉ sự tương tác giữa một biến giải thích định tính với một biến giải thích định lượng;
- **tương tác định tính - định tính**: ám chỉ sự tương tác giữa hai biến giải thích định tính;
- **tương tác định lượng - định lượng**: ám chỉ sự tương tác giữa hai biến giải thích định lượng.

Mô hình với thành phần tương tác

Ví dụ:

- Có thể sự thay đổi chiều dài của cá (Y) khi độ tuổi (X_1) tăng lên là phụ thuộc vào mức nhiệt độ của nước (X_2)? Có thể cá sống ở nước lạnh (ví dụ 27°C) là dài hơn (hoặc ngắn hơn) cá sống ở nước ấm (ví dụ 31°C), vì chúng đang già đi.

  $\Rightarrow$ tương tác định tính - định lượng.

- Có lẽ sự thay đổi mức lương (Y) khi tăng chức vụ (X_1) là phụ thuộc vào giới tính (X_2)? Có thể là lương của nam giới tăng nhiều hơn (hoặc ít hơn) so với nữ giới khi họ được thăng chức.

  $\Rightarrow$ tương tác định tính - định tính.

- Ví dụ về quảng cáo: Quảng cáo trên truyền hình và đài phát thanh đều làm tăng doanh số bán hàng. Có lẽ chi tiền cho cả hai hình thức quảng cáo này có thể làm tăng doanh số bán hàng nhiều hơn là chi cùng một số tiền cho riêng một hình thức?

  $\Rightarrow$ tương tác định lượng - định lượng.

Mô hình với thành phần tương tác

Thành phần tương tác sẽ thay thế cấu trúc cộng của mô hình.

Xét mô hình hồi quy tuyến tính cơ bản sau:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

với cấu trúc cộng tính (additive).

Bây giờ, giả sử ta thêm thành phần tương tác $X_1 X_2$:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon$$

Mô hình lúc này có thể được viết lại thành

$$Y = \beta_0 + (\beta_1 + \beta_{12} X_2) X_1 + \beta_2 X_2 + \varepsilon \tag{1}$$

hoặc

$$Y = \beta_0 + \beta_1 X_1 + (\beta_2 + \beta_{12} X_1) X_2 + \varepsilon \tag{2}$$

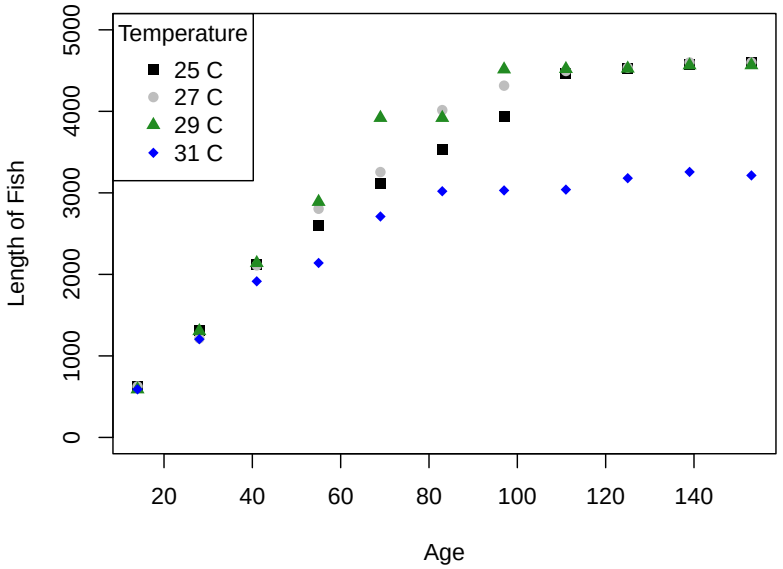
Với mô hình (1), khi X_1 tăng 1 đơn vị, biến đáp ứng Y tăng $\beta_1 + \beta_{12} X_2$ đơn vị, và sự thay đổi này là **phụ thuộc** vào giá trị của X_2.

Ta cũng có tính chất tương tự cho X_2 khi đánh giá mô hình (2).

Ví dụ cho tương tác định tính - định lượng

Ta xét lại nghiên cứu về chiều dài cá, tuổi và nhiệt độ nước.

- Hình vẽ thể hiện mối quan hệ giữa tuổi và chiều dài của cá tương ứng với 4 mức nhiệt độ nước (tức là, 25°C, 27°C, 29°C và 31°C).
- Như trong hình, khi cá còn nhỏ (dưới 80 ngày tuổi) thì nhiệt độ nước không ảnh hưởng gì đến chiều dài của cá.



Tuy nhiên, khi cá được hơn 100 ngày tuổi, ta nhận thấy cá sống ở nhiệt độ nước 31°C dường như ngừng phát triển về chiều dài.
$\hookrightarrow$ có sự tương tác giữa tuổi và nhiệt độ.

Ví dụ cho tương tác định tính - định lượng

Mô hình hồi quy tuyến tính với thành phần tương tác là:

$$\begin{aligned}\text{Length}_i &= \beta_0 + \beta_1\text{Age}_i + \beta_{22}\text{Temp}_{2i} + \beta_{23}\text{Temp}_{3i} + \beta_{24}\text{Temp}_{4i} \\ &\quad + \gamma_{22}\text{Temp}_{2i}\text{Age}_i + \gamma_{23}\text{Temp}_{3i}\text{Age}_i + \gamma_{24}\text{Temp}_{4i}\text{Age}_i + \varepsilon_i\end{aligned}$$

Và mô hình này tương đương với:

$$\begin{aligned}\text{Length}_i &= \beta_0 + (\beta_1 + \gamma_{22}\text{Temp}_{2i} + \gamma_{23}\text{Temp}_{3i} + \gamma_{24}\text{Temp}_{4i})\text{Age}_i \\ &\quad + \beta_{22}\text{Temp}_{2i} + \beta_{23}\text{Temp}_{3i} + \beta_{24}\text{Temp}_{4i} + \varepsilon_i\end{aligned}$$

Nhắc lại rằng, Temp_{2i} là biến dummy cho mức nhiệt độ thứ 2, tương tự cho Temp_{3i} và Temp_{4i}. Do đó, ta có

$$\text{Length}_i = \begin{cases} \beta_0 + \beta_1\text{Age}_i + \varepsilon_i & \text{nếu con cá thứ } i \text{ sống ở } 25^\circ\text{C} \\ \beta_0 + (\beta_1 + \gamma_{22})\text{Age}_i + \beta_{22} + \varepsilon_i & \text{nếu con cá thứ } i \text{ sống ở } 27^\circ\text{C} \\ \beta_0 + (\beta_1 + \gamma_{23})\text{Age}_i + \beta_{23} + \varepsilon_i & \text{nếu con cá thứ } i \text{ sống ở } 29^\circ\text{C} \\ \beta_0 + (\beta_1 + \gamma_{24})\text{Age}_i + \beta_{24} + \varepsilon_i & \text{nếu con cá thứ } i \text{ sống ở } 31^\circ\text{C} \end{cases}$$

Do đó, nếu độ tuổi của cá i thêm 1 ngày, thì chiều dài của nó sẽ tăng lên $\beta_1 + \gamma_{22}$, nếu nó sống trong nước có nhiệt độ 27°C.

Ví dụ cho tương tác định tính - định lượng

Kết quả phân tích mô hình là

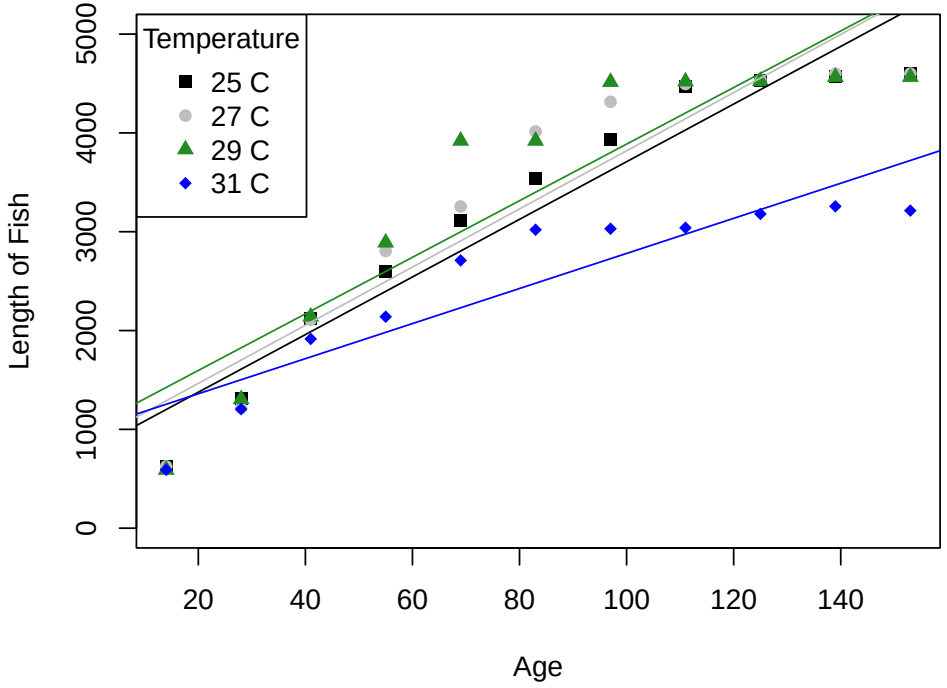
	Estimate	Standard error	t-value	p-value
β_0 (Intercept)	792.68	327.022	2.424	0.0205
β_1 (Age)	29.16	3.475	8.391	< 0.0001
β_{22} (Temp27)	84.34	462.479	0.182	0.8563
β_{23} (Temp29)	230.88	462.479	0.499	0.6207
β_{24} (Temp31)	212.80	462.479	0.460	0.6482
γ_{22} (Age:Temp27)	0.27	4.914	0.054	0.9573
γ_{23} (Age:Temp29)	-0.53	4.914	-0.108	0.9149
γ_{24} (Age:Temp31)	-11.41	4.914	-2.322	0.0260

- Ước tính $\widehat{\gamma}_{24}$ là âm có nghĩa thống kê. Điều này xác nhận đánh giá của ta về đặc tính của cá già sống ở nhiệt độ nước 31°C.

  Giả sử rằng *Age* của con cá thứ i-th tăng 1 ngày.
- Nếu $Temp_{2i} = 1$ (tức là loài cá này sống ở nhiệt độ nước 27°C), chúng ta có sự thay đổi về chiều dài là $\widehat{\beta}_1 + \widehat{\gamma}_{22} = 29.16 + 0.27 = 29.43$.
- Nếu $Temp_{4i} = 1$ (tức là loài cá này sống ở nhiệt độ nước 31°C), chúng ta có sự thay đổi về chiều dài là $\widehat{\beta}_1 + \widehat{\gamma}_{24} = 29.16 - 11.41 = 17.75$.

Ví dụ cho tương tác định tính - định lượng

Hình dưới đây thể hiện đường hồi quy tương ứng với 4 mức nhiệt độ của nước.



Nguyên tắc phân cấp

- Đôi khi một thành phần tương tác có giá trị *p*-value rất nhỏ (có ý nghĩa thống kê), nhưng các tác động chính liên quan thì không.
- Nhưng **nguyên tắc phân cấp** nói rằng: "Nếu chúng ta đưa một sự tương tác vào một mô hình, chúng ta cũng nên đưa vào các tác động chính, ngay cả khi các giá trị *p*-value liên quan đến hệ số của chúng không đáng kể."
- Lý do cơ bản cho nguyên tắc này là các tương tác rất khó diễn giải trong một mô hình mà không có tương tác chính. Việc loại bỏ các tác động chính ra khỏi mô hình sẽ làm thay đổi ý nghĩa của sự tương tác.

Mô hình với quan hệ phi tuyến

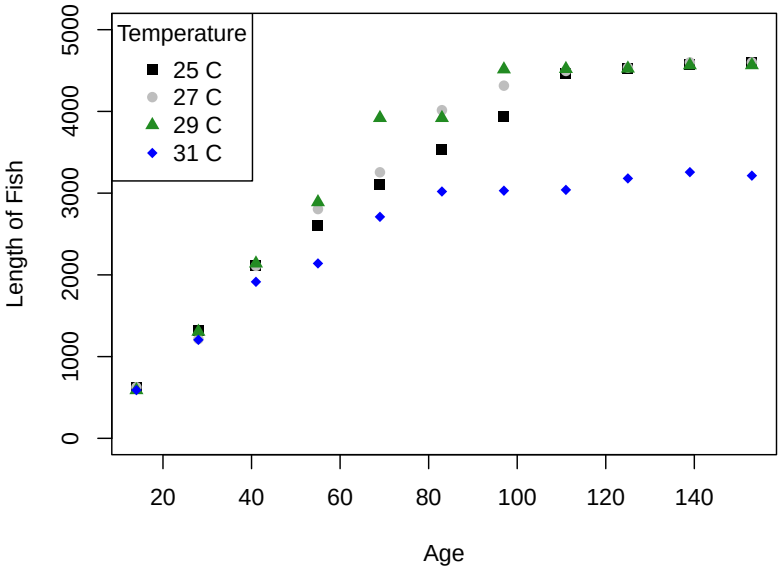
- Mô hình hồi quy tuyến tính giả định mối quan hệ tuyến tính giữa biến đáp ứng và các biến giải thích, tức là sự thay đổi trong biến đáp ứng Y do sự thay đổi một đơn vị của X_j là không đổi, β_j, bất kể giá trị của X_j.
- Tuy nhiên, trong thực tế, đôi khi mối quan hệ thực sự giữa biến đáp ứng và các biến giải thích có thể không tuyến tính.
- Một cách đơn giản để mở rộng trực tiếp mô hình tuyến tính cho phù hợp với mối quan hệ phi tuyến tính là sử dụng hồi quy đa thức.
- Ví dụ, một hồi quy đa thức có thể

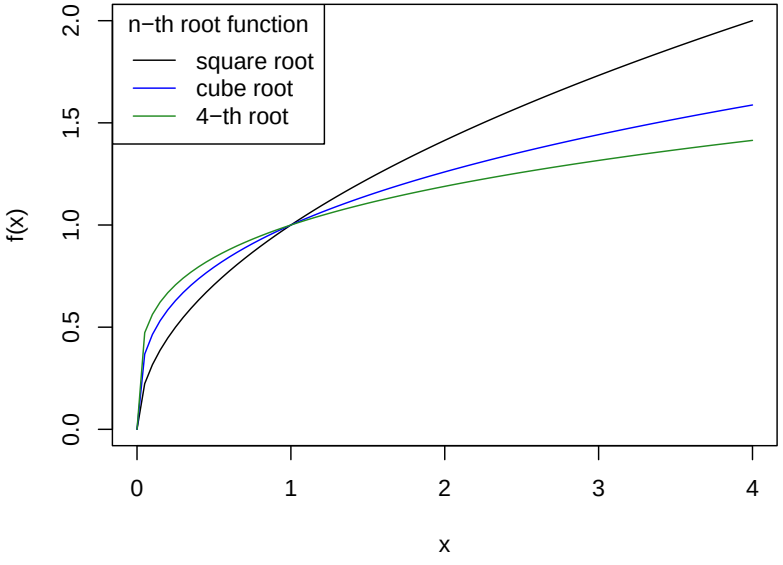
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 X_1^2 + \varepsilon$$

  Số hạng đầu tiên (màu đen) là phần tuyến tính, trong khi số hạng X_1^2 (màu xanh dương) biểu thị mối quan hệ bình phương giữa Y và X_1. Nếu X_1 tăng 1 đơn vị, sự thay đổi trong Y là $\beta_1 + \gamma_1 \left((X_1+1)^2 - X_1^2\right)$ $\Rightarrow$ không còn cố định nữa!
- Chúng ta sử dụng một số hàm toán học phổ biến để khôi phục mối quan hệ phi tuyến tính. Ví dụ: $\log(x)$, $\sqrt{x}$, $\sqrt[3]{x}$, $\exp(x)$, x^2, x^3. Sự lựa chọn phụ thuộc vào hình dạng của dữ liệu được biểu thị bằng biểu đồ phân tán.

Mô hình với quan hệ phi tuyến

- Xét lại đồ thị phân tán về độ tuổi và chiều dài của cá (hình bên trái).
- Hình bên phải thể hiện đồ thị của các hàm căn bậc n, tức là căn bậc hai ($n = 2$), căn bậc ba ($n = 3$) và hàm căn bậc 4.
- Chúng ta có thể thấy rằng hình dạng của dữ liệu có vẻ phù hợp với hình dạng của hàm căn thức.





Mô hình với quan hệ phi tuyến

Giả sử, ta sử dụng hàm căn bậc ba, $\sqrt[3]{x}$, để khôi phục mối quan hệ phi tuyến của độ tuổi và chiều dài của cá:

$$\begin{aligned}\text{Length}_i &= \beta_0 + (\beta_1 + \gamma_{22}\text{Temp}_{2i} + \gamma_{23}\text{Temp}_{3i} + \gamma_{24}\text{Temp}_{4i})\sqrt[3]{\text{Age}_i} \\ &\quad + \beta_{22}\text{Temp}_{2i} + \beta_{23}\text{Temp}_{3i} + \beta_{24}\text{Temp}_{4i} + \varepsilon_i\end{aligned}$$

Mô hình này là tương tự với mô hình trước đó, chỉ khác là, ta thay thế Age_i bởi $\sqrt[3]{\text{Age}_i}$. Ta có thể viết mô hình lại như sau:

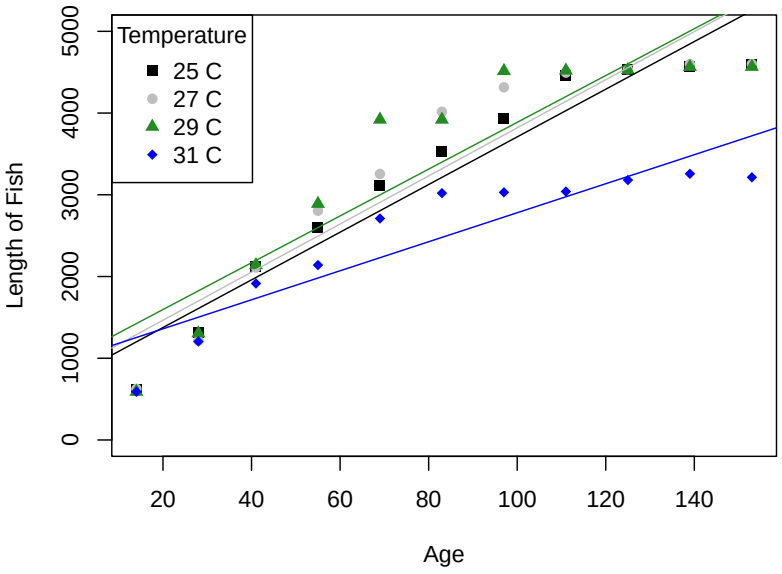
$$\text{Length}_i = \begin{cases} \beta_0 + \beta_1\sqrt[3]{\text{Age}_i} + \varepsilon_i & \text{nếu con cá thứ } i \text{ sống ở } 25^\circ\text{C} \\ \beta_0 + (\beta_1 + \gamma_{22})\sqrt[3]{\text{Age}_i} + \beta_{22} + \varepsilon_i & \text{nếu con cá thứ } i \text{ sống ở } 27^\circ\text{C} \\ \beta_0 + (\beta_1 + \gamma_{23})\sqrt[3]{\text{Age}_i} + \beta_{23} + \varepsilon_i & \text{nếu con cá thứ } i \text{ sống ở } 29^\circ\text{C} \\ \beta_0 + (\beta_1 + \gamma_{24})\sqrt[3]{\text{Age}_i} + \beta_{24} + \varepsilon_i & \text{nếu con cá thứ } i \text{ sống ở } 31^\circ\text{C} \end{cases}$$

Mô hình với quan hệ phi tuyến

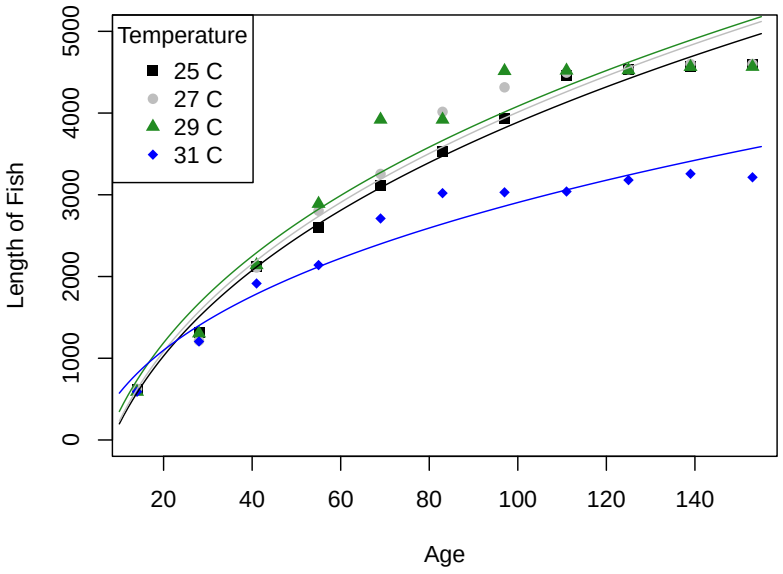
Kết quả phân tích mô hình là:

	Estimate	Standard error	t-value	p-value
β_0 (Intercept)	-3000.58	421.61	$-$ 7.117	$<$ 0.0001
β_1 ($\sqrt[3]{\text{Age}}$)	1484.17	98.42	15.081	$<$ 0.0001
β_{22} (Temp27)	-42.07	596.25	-0.071	0.9441
β_{23} (Temp29)	113.92	596.25	0.191	0.8496
β_{24} (Temp31)	1553.37	596.25	2.605	0.0133
γ_{22} ($\sqrt[3]{\text{Age}}$:Temp27)	35.42	139.18	0.255	0.8005
γ_{23} ($\sqrt[3]{\text{Age}}$:Temp29)	17.42	139.18	0.125	0.9011
γ_{24} ($\sqrt[3]{\text{Age}}$:Temp31)	-546.45	139.18	-3.926	0.0004

Mô hình với quan hệ phi tuyến



(a)



(b)

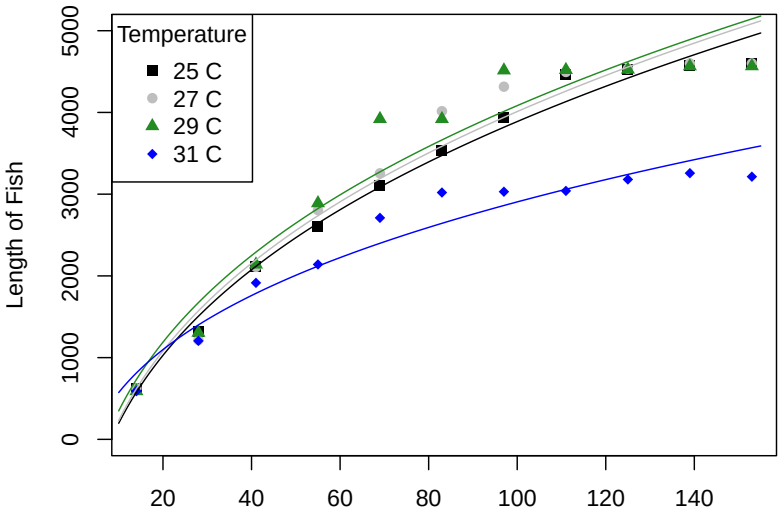
- Hình (a) thể hiện mô hình hồi quy tuyến tính với thành phần tương tác. Hệ số R_a^2 là 0.854.
- Hình (b) thể hiện mô hình hồi quy tuyến tính với thành phần tương tác và căn bậc ba của độ tuổi. Hệ số R_a^2 là 0.952.

Mô hình với quan hệ phi tuyến

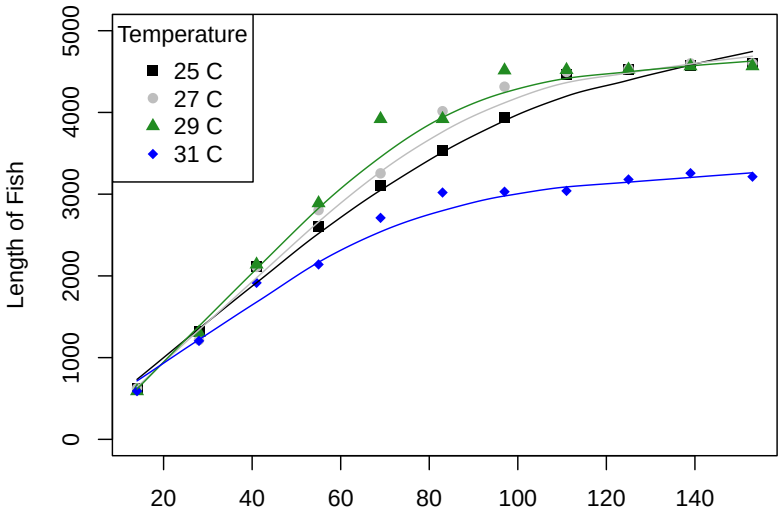
- Trong thực tế, cần phải nỗ lực rất nhiều để xác định một xấp xỉ tốt cho một mô hình tuyến tính với mối quan hệ phi tuyến giữa biến đáp ứng Y và biến giải thích X_j.
- Ngoài ra, chúng ta có thể sử dụng **mô hình hồi quy phi tham số**, đây là một cách tiếp cận dựa trên dữ liệu.
- Một mô hình hồi quy phi tham số được biểu diễn bởi:

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

với $f(\cdot)$ là một hàm không biết trước.



(a) Mô hình với thành phần phi tuyến



(b) Mô hình phi tham số

1 *Mô hình hồi quy tuyến tính đơn*
- Giới thiệu và ước lượng
- Tính chất của ước lượng
- Hồi quy đơn khi X là biến nhị phân

2 *Mô hình hồi quy tuyến tính đa biến*
- Mô hình cơ bản
- Mô hình với biến giải thích là định tính
- Mở rộng mô hình hồi quy

3 *Chuẩn đoán mô hình*

4 *Một số chủ đề thêm*

Chuẩn đoán mô hình

Mô hình hồi quy tuyến tính đa biến

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Các kết quả phân tích mô hình tuyến tính đa biến chỉ đúng nếu các giả định là được thỏa mãn

(A1) Có sự quan hệ tuyến tính giữa Y và X_1, X_2, ..., X_p. (Linearity).

(A2) Quan sát $(X_{1i}, X_{2i}, \ldots, X_{pi}, Y_i)$ với $i = 1, 2, \ldots, n$, là độc lập, và cùng phân phối. (i.i.d data)

(A3) Thành phần sai số ε_i là độc lập với mọi biến giải thích $X_{1i}, X_{2i}, \ldots, X_{pi}$, và có trung bình 0, tức là, $\mathbb{E}(\varepsilon_i) = 0$ với mọi i.

(A4) $\mathrm{Var}(\varepsilon_i) = \sigma^2$ với mọi i. (Homoskedasticity)

(A5) ε_i là độc lập, cùng phân phối chuẩn với trung bình 0 và phương sai σ^2, tức là, $\mathcal{N}(0, \sigma^2)$. (Normality assumption)

(A6) Các giá trị ngoại lai lớn hiếm khi xuất hiện.

Chuẩn đoán mô hình

- Nếu giả định (A1) không đúng, mô hình tuyến tính sẽ thực hiện **tồi tệ** trong việc lập mô hình mối quan hệ và dự đoán dữ liệu thực tế (phi tuyến tính). Vì vậy, việc sử dụng một mô hình như vậy là không phù hợp.
- Nếu giả định (A2) và (A3) không đúng thì OLS ước tính $\widehat{\beta}_j$ là **không vững**.
- Nếu giả định (A4) bị vi phạm, chúng ta sẽ có sự không nhất quán về sai số chuẩn của $\widehat{\beta}_j$ trong mô hình. Sau đó, khoảng tin cậy và kiểm định ý nghĩa sẽ bị sai lệch.
- Đối với giả định (A5), nếu các sai số trong mô hình không độc lập (tức là chúng có liên quan với nhau), chúng ta không thể ước lượng các sai số chuẩn, và cả các khoảng tin cậy và kiểm định thống kê. Giả định về tính phân phối chuẩn là quan trọng nhất khi chúng ta có cỡ mẫu nhỏ. Nếu phần dư không có phân phối chuẩn thì kết quả của các phép kiểm tra ý nghĩa sẽ bị sai lệch. Tuy nhiên, trong trường hợp cỡ mẫu lớn, chúng ta không cần lo lắng về giả định tính chuẩn (nhờ Định lý giới hạn trung tâm).

Chuẩn đoán mô hình

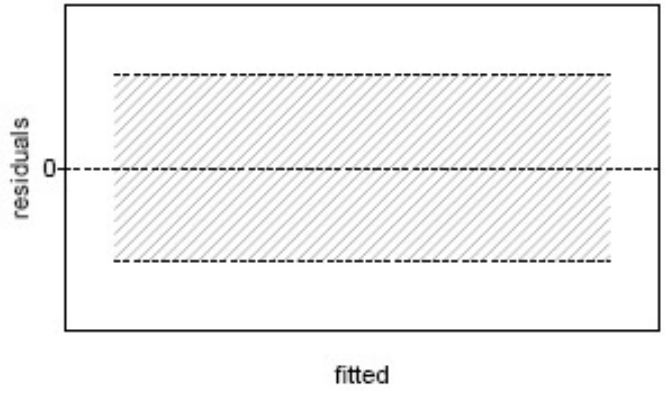
- Đối với giả định (A6), nếu chúng ta có các giá trị ngoại lai lớn: thứ nhất, Định lý giới hạn trung tâm không đúng $\Rightarrow$ Không có phân bố lấy mẫu và không có kiểm định thống kê; thứ hai, các giá trị ngoại lai có thể dẫn đến các ước tính sai lệch nghiêm trọng về hệ số hồi quy.

Các giả định được kiểm tra như sau.

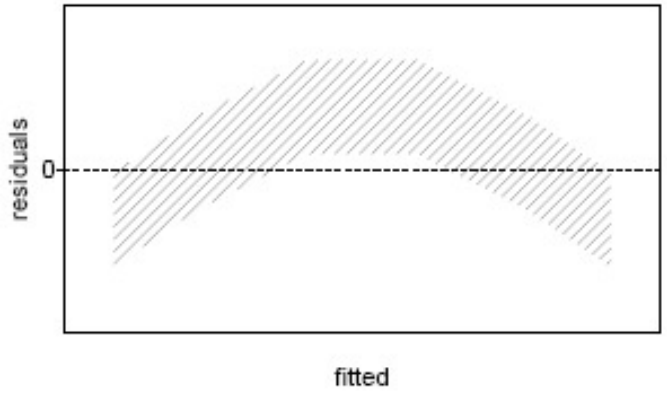
- Giả định (A2) và (A3) có thể được kiểm tra bằng cách thẩm định quá trình lấy mẫu.
- Lưu ý rằng, phương pháp OLS thường tạo ra giá trị trung bình bằng 0 đối với phần dư. Do đó, điều kiện $\mathbb{E}(\varepsilon_i) = 0$ được thỏa mãn.
- Sau khi ước lượng mô hình, chúng ta cần kiểm tra lại các giả định (A1), (A4), (A5) và (A6) bằng các đồ thị chẩn đoán của phần thặng dư.

Kiểm tra giả định (A1)

- Để kiểm tra giả định (A1), chúng ta sử dụng biểu đồ Residuals vs Fitted.
- Nếu phần dư được phân bố ngẫu nhiên (khá đối xứng) xung quanh đường nằm ngang tương ứng với phần dư = 0. Như vậy giả định (A1) được thỏa mãn. (xem Hình a)
- Sự hiện diện của một xu hướng nào đó (một loại độ cong nào đó) có thể chỉ ra một vấn đề với một số khía cạnh của mô hình tuyến tính. (ví dụ, xem Hình b)



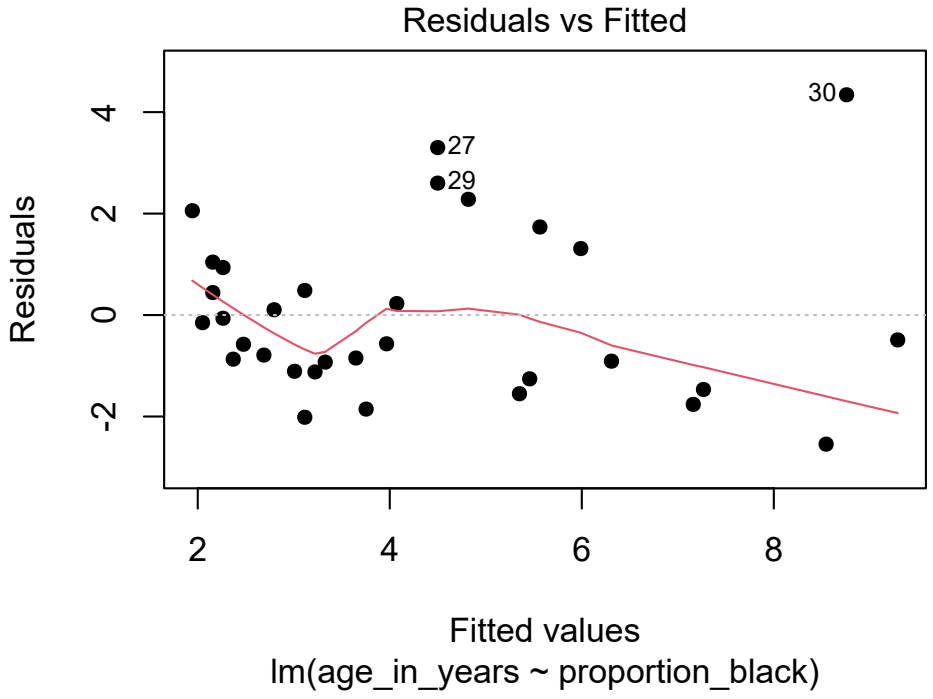
(a)



(b)

Kiểm tra giả định (A1)

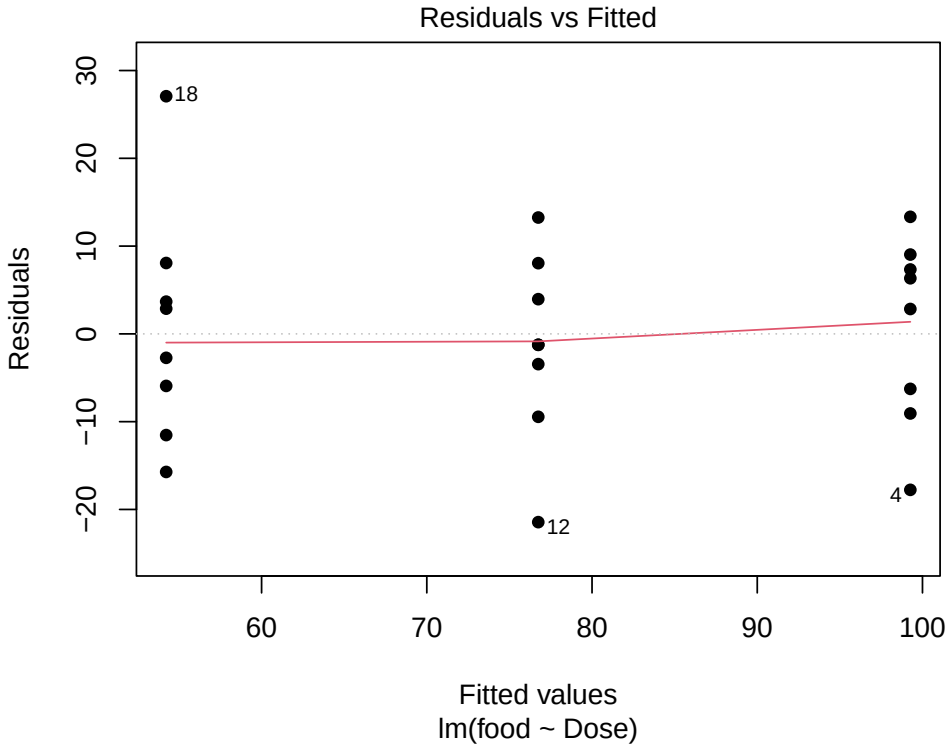
Xét biểu đồ Residuals vs Fitted của mô hình hồi quy tuyến tính cho % màu đen trên mũi và độ tuổi của sư tử đực.



Hình vẽ cho thấy một xu hướng của các điểm thặng dư. Trên thực tế, chúng ta thấy đường liền nét không xấp xỉ nằm ngang ở mức 0. Vì vậy, có thể hồi quy là phi tuyến tính.

Kiểm tra giả định (A1)

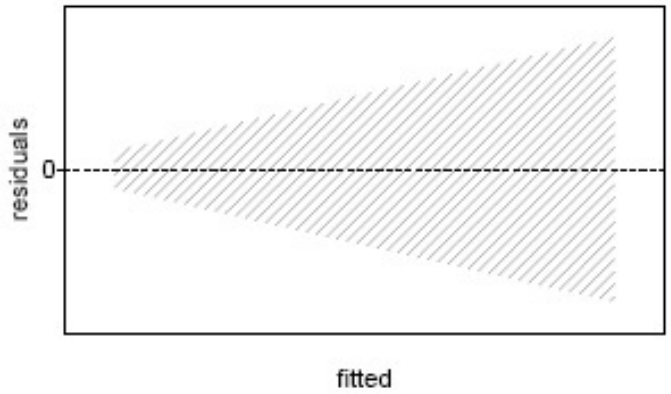
Xét biểu đồResiduals vs Fitted của mô hình hồi quy tuyến tính cho liều lượng amphetamine so với lượng thức ăn tiêu thụ.

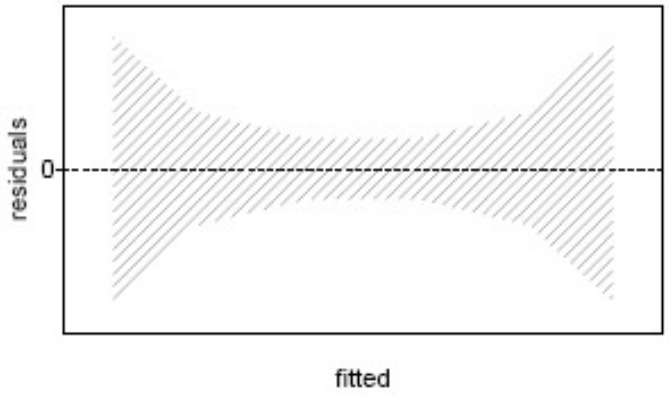


Hình vẽ không cho thấy xu hướng nào. Nghĩa là, đường liền nét gần như nằm ngang ở mức 0. Điều này cho thấy chúng ta có thể giả định mối quan hệ tuyến tính giữa liều lượng amphetamine và lượng thực phẩm tiêu thụ.

Kiểm tra giả định (A4)

- Để kiểm tra giả định (A4), chúng tôi sử dụng biểu đồ Residuals vs Fitted hoặc Scal-Location.
- Nếu độ biến thiên của phần thặng dư thay đổi khi chúng ta di chuyển từ trái sang phải. Ví dụ: nếu biểu đồ Residuals vs Fitted trông như thế này,

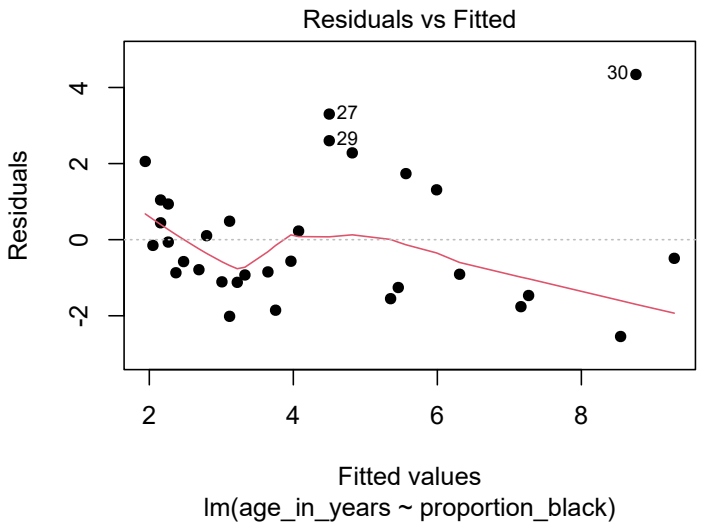


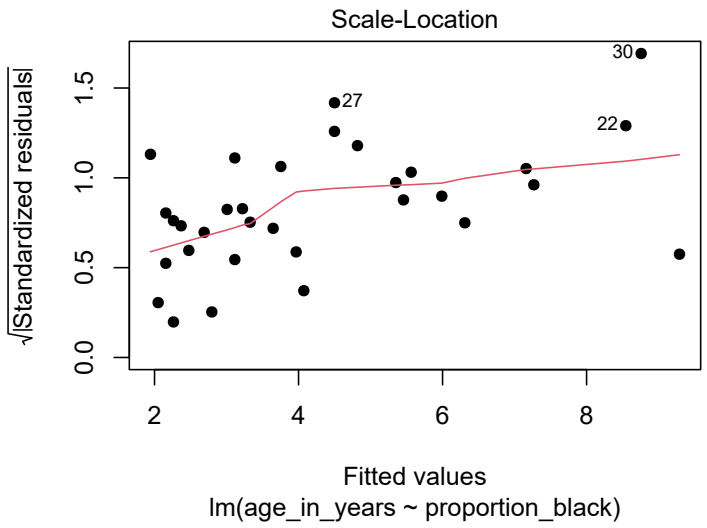


- Nếu biểu đồ Scal-Location hiển thị đường nằm ngang với các điểm trải đều bằng nhau thì đó là một dấu hiệu tốt về tính đồng nhất phương sai.

Kiểm tra giả định (A4)

Xét biểu đồ Residuals vs Fitted và Scale-Location của mô hình hồi quy tuyến tính cho % màu đen trên mũi và độ tuổi của sư tử đực.

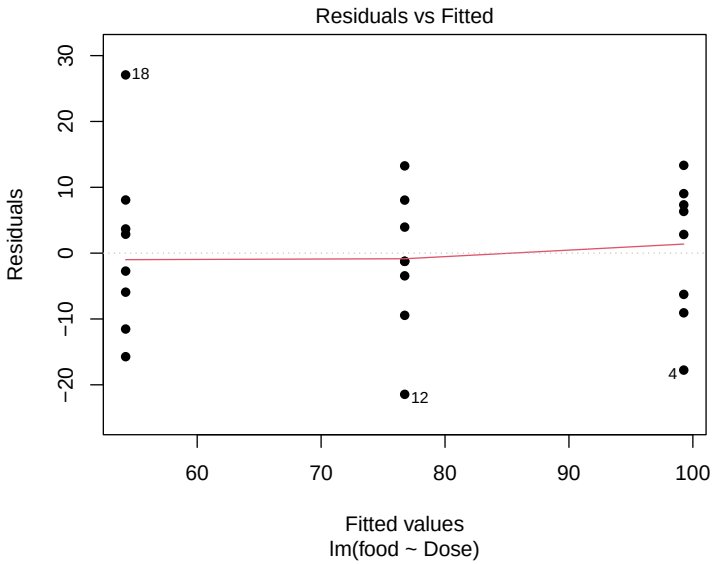


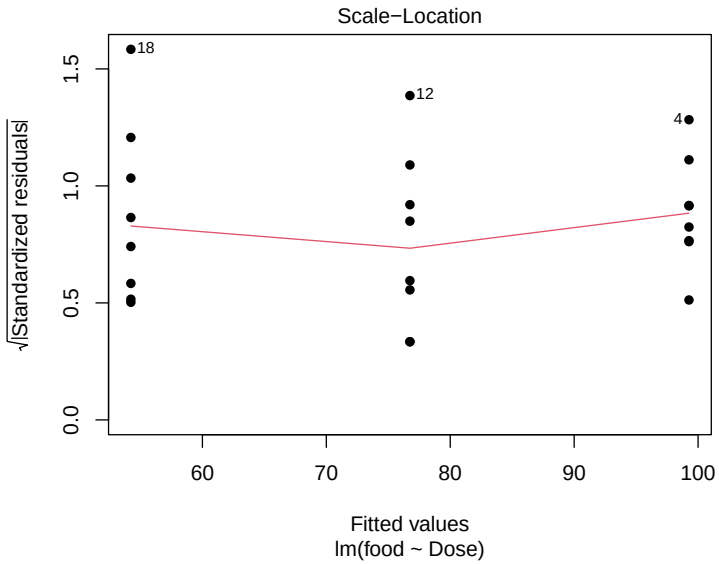


Nhìn vào hình thứ hai, có thể thấy rằng độ biến thiên (phương sai) của các điểm thặng dư tăng theo giá trị ước lượng của biến đáp ứng, cho thấy phương sai không cố định trong các phần thặng dư (hoặc phương sai thay đổi - heteroscedasticity).

Kiểm tra giả định (A4)

Xét biểu đồ Residuals vs Fitted và Scale-Location của mô hình hồi quy tuyến tính cho liều lượng amphetamine so với lượng thức ăn tiêu thụ.

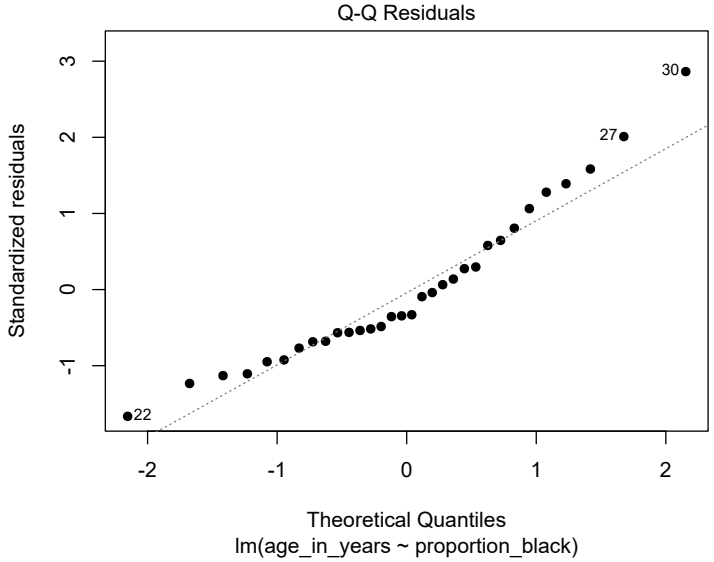


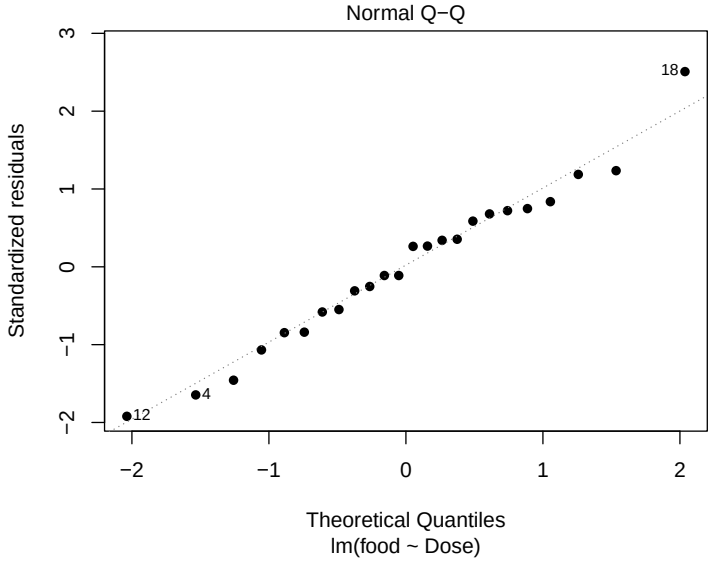


Nhìn vào hình thứ hai, có thể thấy một đường nằm ngang với các điểm trải đều bằng nhau, cho thấy phương sai không đổi trong phần thặng dư (hoặc tính đồng nhất phương sai - homoscedasticity).

Kiểm tra giả định (A5)

Ta sử dụng biểu đồ normal Q-Q để kiểm tra tính tuân theo phân phối chuẩn của sai số hay thặng dư của mô hình.

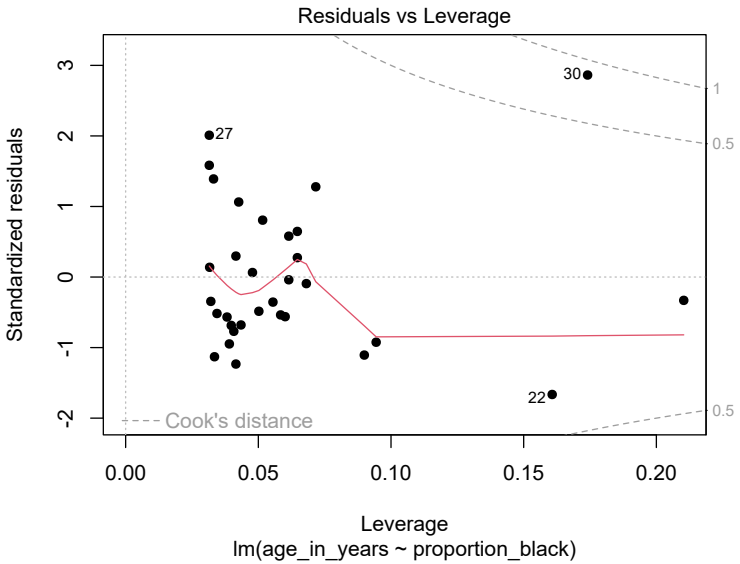


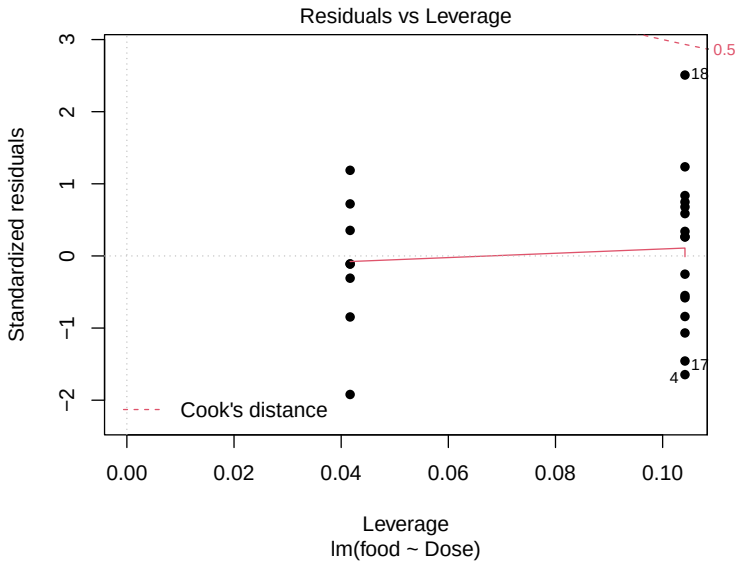


- Dựa trên hình đầu tiên cho thấy các điểm có xu hướng tạo thành một đường cong, tuy không thực sự rời khỏi đường thẳng tham chiếu. Do đó, ta có thể hoài nghi về giả định (A5) đối với mô hình tuyến tính cho % màu đen trên mũi và độ tuổi của sư tử đực.
- Hình thứ hai cho thấy hầu hết các điểm đều nằm dọc theo đường tham chiếu, do đó, giả định (A5) là thỏa mãn đối với mô hình tuyến tính cho liều amphetamine so với lượng thực phẩm tiêu thụ.

Kiểm tra giả định (A6)

- Ta sử dụng biểu đồ Residuals vs Leverage để xác định các trường hợp có ảnh hưởng tới mô hình, đó là các giá trị ngoại lai có thể ảnh hưởng đến kết quả hồi quy khi được đưa vào hoặc loại trừ khỏi phân tích.
- Trong biểu đồ Residuals vs Leverage, hãy tìm điểm dữ liệu bên ngoài đường đứt nét, tức là khoảng cách của Cook. Khi các điểm nằm ngoài khoảng cách của Cook, điều này có nghĩa là chúng có điểm khoảng cách của Cook cao và do đó, các giá trị ngoại lai có ảnh hưởng đến kết quả hồi quy. Kết quả hồi quy sẽ bị thay đổi nếu chúng ta loại trừ những trường hợp đó.





Cộng tuyến và đa cộng tuyến

- Cộng tuyến đề cập đến tình huống trong đó hai hoặc nhiều biến giải thích có liên quan chặt chẽ với nhau.
- Đa cộng tuyến đề cập đến hiện tượng cộng tuyến tồn tại giữa ba biến trở lên, ngay cả khi không có cặp biến nào có mối tương quan đặc biệt cao.
- Trong trường hợp có hiện tượng cộng tuyến hoặc đa cộng tuyến, các ước tính của các tham số hồi quy là **không ổn định** và không chính xác, và không duy nhất.
- Một thước đo phổ biến của đa cộng tuyến là **hệ số lạm phát phương sai - variance inflation factor** (VIF).
- Với mỗi biến giải thích j-th, chúng ta có một VIF_j, và bằng cách phân tích độ lớn của VIF_j, chúng ta có thể phát hiện ra sự cộng tuyến. Một nguyên tắc nhỏ:

  nếu $\text{VIF}_j > 5$ hoặc 10, thì đó là mức độ cộng tuyến có vấn đề;

  nếu $\text{VIF}_j < 5$ thì biến X_j không có liên hệ với các biến giải thích còn lại.
- Để giải quyết hiện tượng cộng tuyến:

  1. Loại bỏ một trong các biến có vấn đề khỏi mô hình.

  2. Kết hợp các biến cộng tuyến lại với nhau thành một biến giải thích duy nhất.

1 *Mô hình hồi quy tuyến tính đơn*
- Giới thiệu và ước lượng
- Tính chất của ước lượng
- Hồi quy đơn khi X là biến nhị phân

2 *Mô hình hồi quy tuyến tính đa biến*
- Mô hình cơ bản
- Mô hình với biến giải thích là định tính
- Mở rộng mô hình hồi quy

3 *Chuẩn đoán mô hình*

4 *Một số chủ đề thêm*

Heteroskedasticity - Tính không đồng nhất phương sai

- Nếu giả định (A4) - giả định về tính đồng nhất phương sai - bị vi phạm, điều đó có nghĩa là phương sai của sai số ε_i không phải là một hằng số cho mọi đối tượng i.
  $\Rightarrow$ Điều này được gọi là sự không đồng nhất phương sai (**heteroskedasticity**).
- Trong trường hợp này, các sai số chuẩn của ước tính OLS $\widehat{\beta}_j$ là không chính xác và do đó, khoảng tin cậy và kiểm định thống kê sẽ bị sai lệch.
- Để giải quyết vấn đề này, một giải pháp là sử dụng **ước lượng phương sai kẹp** (**sandwich variance estimation**) sau khi điều chỉnh mô hình hồi quy tuyến tính.
- Có nhiều phương pháp khác nhau để ước tính phương sai kẹp, hầu hết chúng được triển khai trong `R` bởi một gói có tên `sandwich`.

Sai số không theo phân phối chuẩn - Non-normal errors

- Nếu giả định (A5) - giả định về tính phân phối chuẩn - bị vi phạm, điều đó có nghĩa là sự phân bố của các sai số là lệch
  $\Rightarrow$ Vấn đề này được gọi là vấn đề **non-normal errors**.
- Đặc biệt đối với các tập dữ liệu lớn hơn, giả định về tính chuẩn là không quan trọng, vì suy luận sẽ gần đúng mặc dù không có tính phân phối chuẩn, nhờ Định lý giới hạn trung tâm.
- Tuy nhiên, trong trường hợp tập dữ liệu nhỏ hơn, chúng ta nên sử dụng một phép biến đổi có thể có cho biến đáp ứng để đạt được tính phân phối chuẩn.
- Nếu biến đáp ứng là dương và phân phối của nó bị lệch, chúng ta có thể sử dụng phép biến đổi **Box-Cox** cho mô hình hồi quy tuyến tính. Việc chuyển đổi này đã được triển khai trong `R` bởi một hàm `boxcox()` có trong thư viện `MASS`.

Các ngoại lai hoặc quan sát có ảnh hưởng

- Bằng cách xác minh giả định (A6), nếu chúng ta thấy nhiều giá trị ngoại lai hoặc giá trị cực đoan, điều đó có nghĩa là OLS hoạt động kém.
- Đặc biệt, OLS hoạt động kém đối với các sai số có phân phối với đuôi dài (nghĩa là có thể quan sát được các giá trị lớn).
- Một cách tiếp cận đơn giản là loại bỏ các giá trị ngoại lai khỏi tập dữ liệu và sau đó tiến hành OLS. Cách tiếp cận này là thỏa đáng khi chúng ta tin rằng các ngoại lai đại diện cho các quan sát **thực sự không chính xác**, nhưng ngay cả khi đó, việc phát hiện những trường hợp như vậy không phải lúc nào cũng dễ dàng vì nhiều ngoại lai có thể che lấp lẫn nhau.
- Tuy nhiên, trong những trường hợp khác, các giá trị ngoại lai lại là những quan sát thực tế. Đôi khi, việc loại bỏ những trường hợp này chỉ đơn giản là tạo ra những ngoại lai khác.
- Một cách tiếp cận nhìn chung tốt hơn là sử dụng một giải pháp thay thế **robust** cho OLS để làm giảm ảnh hưởng của các giá trị ngoại lai. $\Rightarrow$ **Hồi quy mạnh mẽ** (**Robust regression**).
- Một số phương pháp hồi quy mạnh mẽ đã được triển khai trong `R` bằng hàm `rlm` có trong gói `MASS`.

Lựa chọn biến - Variable selection

- Nhiệm vụ xác định những biến giải thích nào có liên quan đến biến đáp ứng, để phù hợp với một mô hình duy nhất chỉ liên quan đến những biến giải thích đó, được gọi là **sự lựa chọn biến**.
- Cách tiếp cận trực tiếp nhất được gọi là hồi quy **all subsets** hoặc **best subsets**: chúng ta điều chỉnh tất cả các tập hợp con có thể và sau đó chọn giữa chúng dựa trên một số tiêu chí, ví dụ: R_a^2.
- Tuy nhiên, chúng ta thường không thể kiểm tra tất cả các mô hình có thể, vì ta có 2^p số mô hình; ví dụ: khi $p = 5$ thì có 32 mô hình! Thay vào đó, chúng ta cần một cách tiếp cận tự động để tìm kiếm thông qua một tập hợp con của các biến giải thích.
- Thông thường, ta sử dụng AIC (Akaike Information Criterion) làm tiêu chí để lựa chọn mô hình. Trong `R`, chúng ta sử dụng hàm `stepAIC()` trong thư viện `MASS`.

Vấn đề khuyết dữ liệu

Trong phân tích dữ liệu, việc khuyết dữ liệu (hoặc dữ liệu không được hoàn thiện) là trường hợp phổ biến do một số lý do, có thể có tác động tiêu cực đến kết luận phân tích hồi quy. Về nguyên tắc, có ba loại dữ liệu bị khuyết:

- Khuyết hoàn toàn ngẫu nhiên (missing completely at random - MCAR): Điều này khá đơn giản. Không có mối quan hệ nào giữa việc một điểm dữ liệu bị thiếu và bất kỳ giá trị nào trong tập dữ liệu, bị thiếu hoặc được quan sát. Hoặc dữ liệu bị thiếu chỉ là một tập hợp con ngẫu nhiên của dữ liệu.
- Khuyết ngẫu nhiên (missing at random - MAR): Điều này có nghĩa là cơ chế thiếu sót không liên quan đến biến bị thiếu, nhưng nó liên quan đến một số biến được quan sát. Nói cách khác, phản hồi bị thiếu do các biến khác. Ví dụ: nếu hầu hết mọi người trong cuộc khảo sát không trả lời một câu hỏi nhất định thì tại sao họ lại làm vậy? Có phải do câu hỏi không rõ ràng?
- Khuyết không ngẫu nhiên (missing not at random - MNAR): Điều này có nghĩa là cơ chế thiếu sót không chỉ liên quan đến một số biến quan sát mà còn cả biến bị thiếu.

Vấn đề khuyết dữ liệu

Ảnh hưởng của việc khuyết dữ liệu phụ thuộc vào loại cơ chế khuyết.

- Nếu là MCAR, hiệu quả phân tích là không có bị ảnh hưởng. Bởi vì, việc phân phối biến dựa trên dữ liệu quan sát được cũng giống như phân phối biến dựa trên dữ liệu bị khuyết.

  $\Rightarrow$ Do đó, khi bạn thực hiện phân tích trên toàn bộ phần dữ liệu (bỏ qua tất cả các giá trị bị khuyết), chúng ta sẽ nhận được kết quả tương tự như khi phân tích dữ liệu đầy đủ.
- Tuy nhiên, nếu cơ chế khuyết là MAR hoặc MNAR thì việc phân tích chỉ dựa trên dữ liệu quan sát được sẽ dẫn đến kết quả sai lệch.

Để xử lý vấn đề dữ liệu, chúng tôi có một số kỹ thuật:

- Complete-case analysis: sử dụng khi cơ chế khuyết là MCAR.
- Single imputation hoặc Multiple imputation: sử dụng khi cơ chế khuyết là MAR.

Một số mô hình hồi quy khác

Trong bài giảng này, chúng ta đã nghiên cứu mô hình hồi quy tuyến tính đa biến, về cơ bản là một mô hình hồi quy.

Tuy nhiên, nhìn chung, các mô hình hồi quy không chỉ chứa các mô hình tuyến tính mà còn chứa các loại mô hình khác được liệt kê dưới đây nhưng không giới hạn:

Mô hình tuyến tính tổng quát: Mô hình tuyến tính thông thường không thể xử lý các biến đáp ứng không tuân theo phân phối chuẩn, chẳng hạn như **counts** hoặc **proportions**. Điều này thúc đẩy sự phát triển của các mô hình tuyến tính tổng quát có thể biểu diễn **biến định tính**, **nhị phân** và **các loại phản hồi khác**.

Một số mô hình hồi quy khác

Mô hình hồi quy phi tham số: Trong mô hình tuyến tính, các biến giải thích được kết hợp theo cách tuyến tính để mô hình hóa tác động lên biến đáp ứng. Đôi khi tính tuyến tính này không đủ để nắm bắt được cấu trúc của dữ liệu và cần có sự linh hoạt hơn. Các phương pháp như: **hồi quy phi tham số**, **additive models**, **trees** và **neural networls** cho phép lập mô hình hồi quy linh hoạt hơn cho biến đáp ứng mà kết hợp các biến giải thích theo mô hình phi tham số.

Mô hình hiệu ứng hỗn hợp - Mixed Effect Models: Một số dữ liệu có cấu trúc được nhóm, lồng nhau hoặc phân cấp. Các biện pháp đo lặp đi lặp lại, dữ liệu theo chiều dọc và đa cấp bao gồm một số quan sát được thực hiện trên cùng một cá nhân hoặc nhóm. Điều này gây ra một cấu trúc tương quan trong sai số. Các mô hình hiệu ứng hỗn hợp cho phép mô hình hóa dữ liệu đó.