Bài giảng 5: Regression và Prediction - I

TS. Tô Đức Khánh

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

Học phần: Xử lý Số liệu Thống kê

Nội dung buổi học

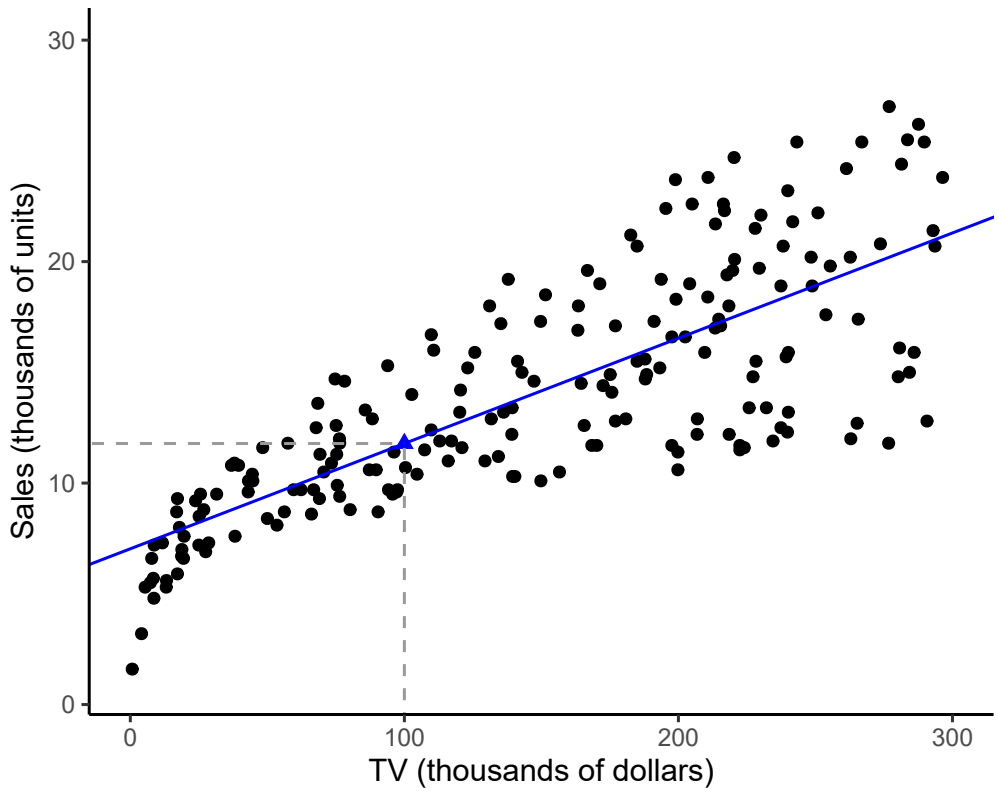
Mô hình hồi quy

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: số tiền đầu tư cho quảng cáo trên TV
- Y: lượng sản phẩm được bán

ta mong muốn dự đoán được Y thông qua số tiền đầu tư cho quảng cáo trên TV.

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 số, 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

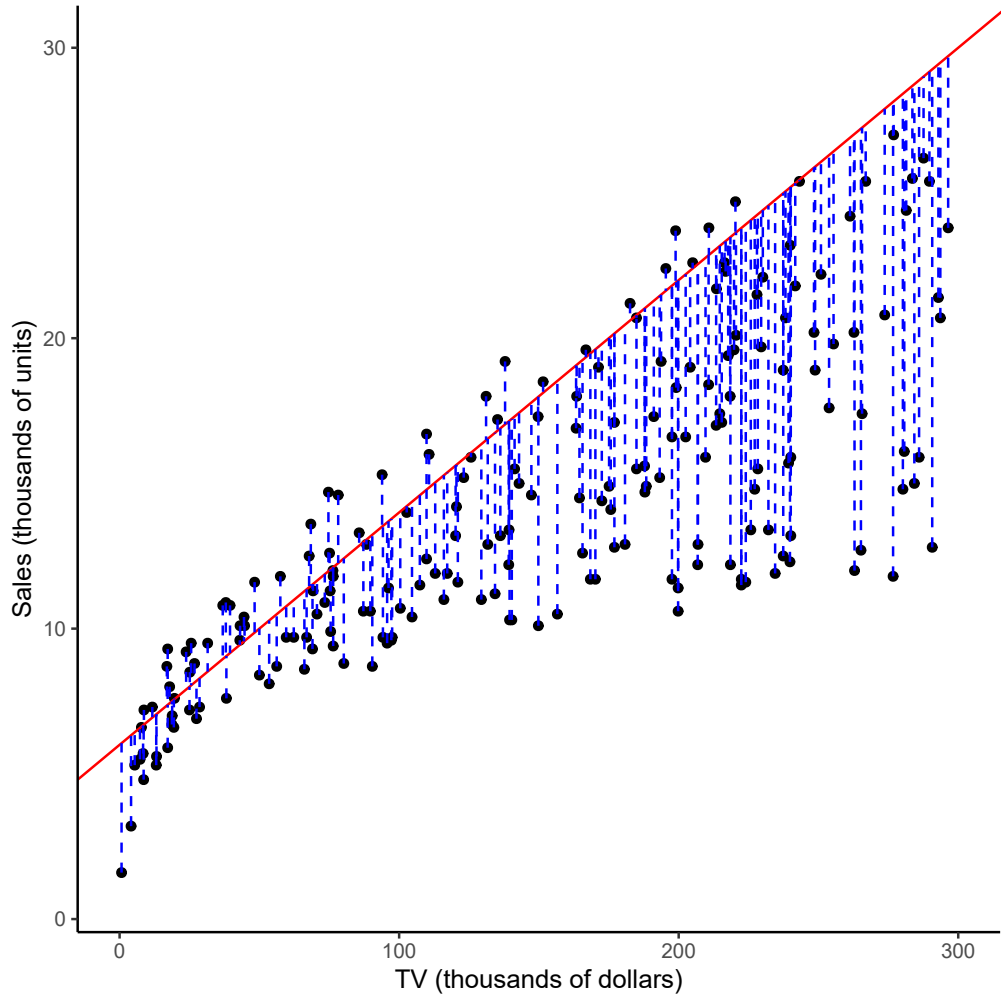
$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

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 số tiền đầu tư quảng cáo trên TV và số lượng sản phẩm được bán. Giả sử

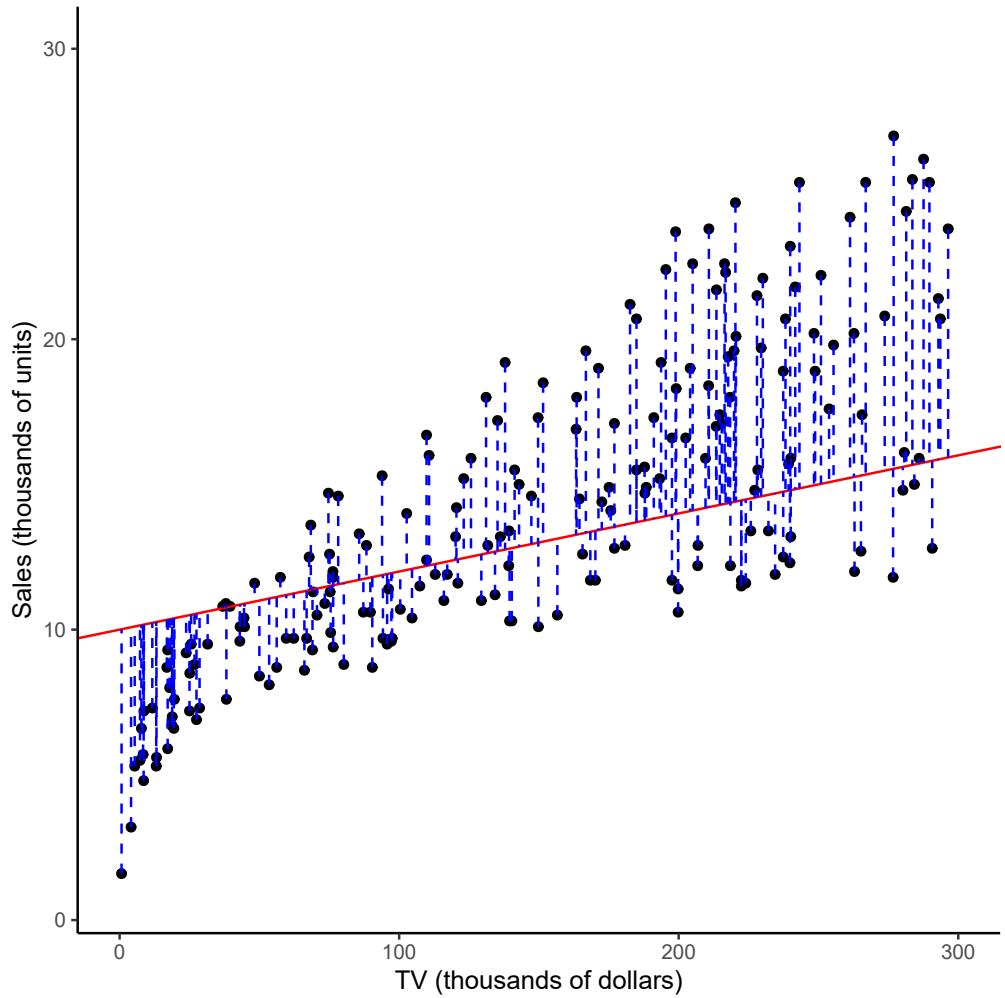
- $\beta_0 = 6$
- $\beta_1 = 0.08$



Ước lượng mô hình hồi quy tuyến tính đơn

Xét dữ liệu với các cặp quan sát của số tiền đầu tư quảng cáo trên TV và số lượng sản phẩm được bán. Giả sử

- $\beta_0 = 10$
- $\beta_1 = 0.02$



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

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

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

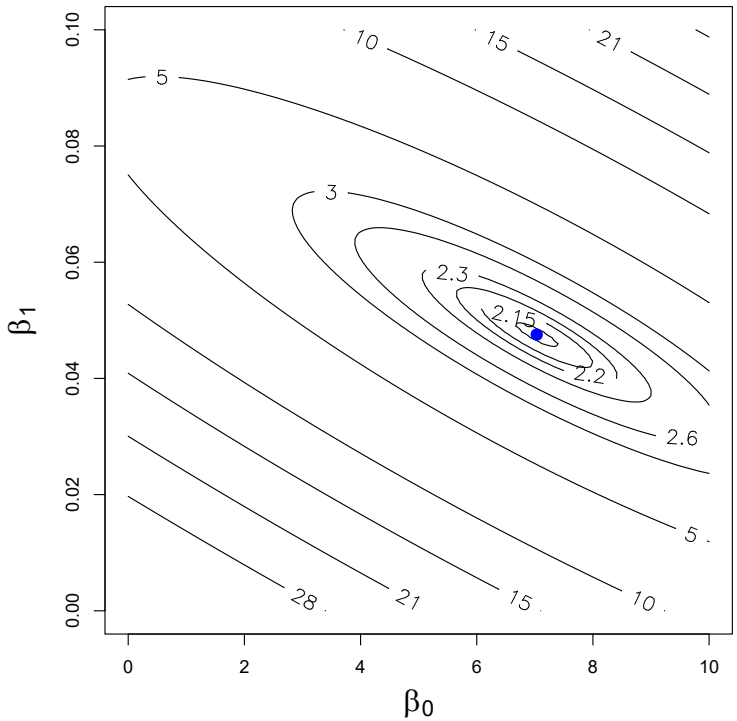
$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \\ \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1\overline{x}\end{aligned}$$

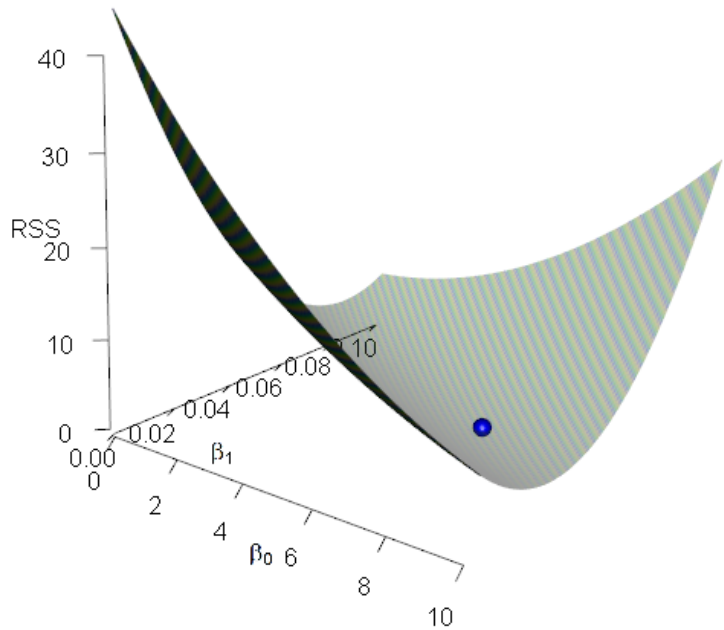
Từ đây, ta có phương trình hồi quy tuyến tính ước lượng:

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

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 về tiền đầu tư quảng cáo và doanh số bán hàng.

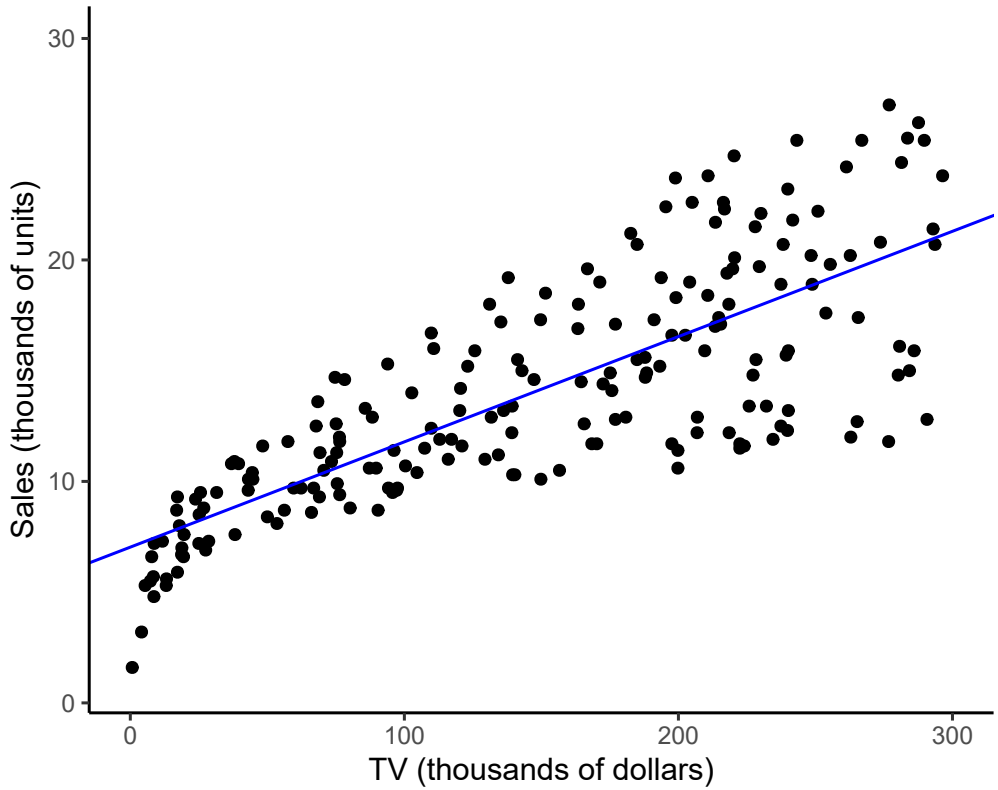




Ta ước lượng được

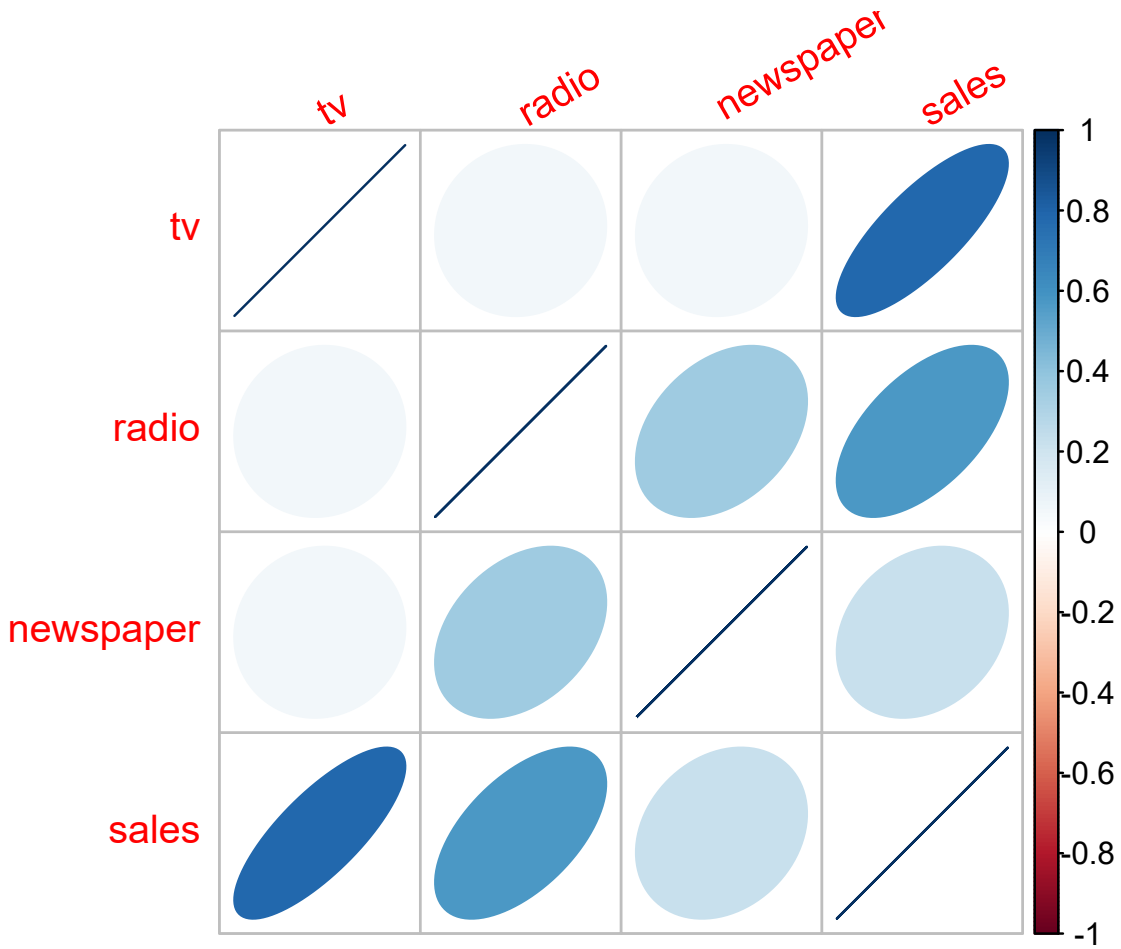
- $\widehat{\beta_0} = 7.033$
- $\widehat{\beta_1} = 0.048$

Ước lượng mô hình hồi quy tuyến tính đơn



Ví dụ: Doanh số và chi phí quảng cáo

Trước hết, ta xét biểu đồ hệ số tương quan tuyến tính giữa các biến:



Ví dụ: Doanh số và chi phí quảng cáo

Tất cả các biến `tv`, `radio` và `newspaper` đều có quan hệ tuyến tính (thuận) với `sales`.

Một cách làm thô, đó là xét từng mô hình hồi quy đơn lẻ:

- MH1: $\texttt{sales} = \beta_0 + \beta_1 \text{tv} + \varepsilon_i$
- MH2: $\texttt{sales} = \beta_0 + \beta_1 \text{radio} + \varepsilon_i$
- MH3: $\texttt{sales} = \beta_0 + \beta_1 \text{newspaper} + \varepsilon_i$

Nhưng cách làm này có các nhược điểm về mặt kỹ thuật:

- Không thể tạo ra chung một giá trị dự đoán cho `sales` khi có thông tin về số tiền đầu tư cho cả 3 phương tiện truyền thông.
- Bỏ qua sự liên hệ giữa các ngân sách quảng cáo khi dự đoán `sales`.

$\hookrightarrow$ ta cần một mô hình hồi quy (tuyến tính) mới để xem xét cùng lúc các tác động của `tv`, `radio` và `newspaper` tới `sales`.

Mô hình hồi quy tuyến tính đa biến

- 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, doanh số bán hàng (Y) có thể được tiến đoán bởi mô hình tuyến tính dựa trên số tiền đầu tư quảng cáo trên TV, đài Radio, và báo in:

$$Y = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{Radio} + \beta_3 \text{Newspaper}$$

- 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

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

Mô hình hồi quy tuyến tính đa biến

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 - Doanh số bán hàng vs. Chi phí quảng cáo

- Một mô hình tuyến tính sẽ là:

$$\texttt{sales} = \beta_0 + \beta_1 \texttt{tv} + \beta_2 \texttt{radio} + \beta_3 \texttt{newspaper} + \varepsilon$$

- Ước lượng OLS cho các hệ số là:

	Estimate
β_0 (Intercept)	2.9389
β_1 (tv)	0.0458
β_2 (radio)	0.1885
β_3 (newspaper)	−0.0010

- Ta thấy rằng nếu tăng chi phí quảng cáo TV lên một đơn vị (1000$), doanh số bán hàng tăng 45.8 đơn vị sản phẩm.
- Nếu xét theo chi phí quảng cáo trên Radio, mức tăng doanh số khi tăng chi phí quảng cáo lên một đơn vị (1000$) là 188.5 đơn vị sản phẩm.
- Trong khi đó, tăng 1000$ cho chi phí quảng cáo trên báo in thì mức tăng doanh số bán hàng lại là −1 đơn vị sản phẩm! Tại sao lại như vậy? Theo biểu đồ tương quan, mức tăng phải dương.

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.
- 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).
- Biến X như thế này được gọi là biến nhân tố - **factor**.

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 độ tuổi và 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:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 \text{Age}_i + \varepsilon_i \\
&= \begin{cases} \beta_0 + \beta_3 \text{Age}_i + \varepsilon_i & \text{nếu người thứ } i \text{ là African American} \\ \beta_0 + \beta_1 + \beta_3 \text{Age}_i + \varepsilon_i & \text{nếu người thứ } i \text{ là Asian} \\ \beta_0 + \beta_2 + \beta_3 \text{Age}_i + \varepsilon_i & \text{nếu người thứ } i \text{ là Caucasian} \end{cases}
\end{aligned}
$$

Do đó,

- β_0 biểu diễn tác động của chủng tộc African American;
- β_1 biểu diễn sự khác biệt trong Y giữa hai chủng tộc African American và Asian khi cố định tuổi;
- β_2 biểu diễn sự khác biệt trong Y giữa hai chủng tộc African American và Caucasian khi cố định tuổi.

Mô hình với thành phần tương tác

Trong mô hình hồi quy đa biến, 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 gọi 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 tắc, 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

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

hoặc

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

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 - Tương tác giữa quảng cáo TV và Radio

Ta xem xét mô hình hồi quy dự đoán doanh số bán hàng theo chi phí quảng cáo trên TV, Radio, và sự tương tác giữa hai kênh quảng cáo này:

$$\texttt{sales} = \beta_0 + \beta_1\texttt{tv} + \beta_2\texttt{radio} + \beta_3\texttt{tv} \times \texttt{radio} + \varepsilon$$

Mô hình này, tương đương với

$$\texttt{sales} = \beta_0 + (\beta_1 + \beta_3\texttt{radio}) \times \texttt{tv} + \beta_2\texttt{radio} + \varepsilon$$

hoặc

$$\texttt{sales} = \beta_0 + \beta_1\texttt{tv} + (\beta_2 + \beta_3\texttt{tv}) \times \texttt{radio} + \varepsilon$$

Như vậy:

- nếu tăng chi phí quảng cáo trên TV lên 1000\$, thì tương ứng mức tăng

$$\left(\widehat{\beta}_1 + \widehat{\beta}_3\texttt{radio}\right) \times 1000 = 19.1 + 1.1 \times \texttt{radio}$$

  trong doanh số bán hàng; hoặc
- nếu tăng chi phí quảng cáo trên Radio lên 1000\$, thì tương ứng mức tăng

$$\left(\widehat{\beta}_2 + \widehat{\beta}_3\texttt{tv}\right) \times 1000 = 28.9 + 1.1 \times \texttt{tv}$$

  trong doanh số bán hàng.

Ví dụ - Tương tác giữa quảng cáo TV và Radio

Kết quả ước lượng cho ta

	Estimate
β_0 (Intercept)	6.7502
β_1 (tv)	0.0191
β_2 (radio)	0.0289
β_3 (tv $\times$ `radio`)	0.0011

Như vậy:

- nếu tăng chi phí quảng cáo trên TV lên 1000\$, thì tương ứng mức tăng

$$\left(\widehat{\beta}_1 + \widehat{\beta}_3 \texttt{radio}\right) \times 1000 = 19.1 + 1.1 \times \texttt{radio}$$

  trong doanh số bán hàng; hoặc

- nếu tăng chi phí quảng cáo trên Radio lên 1000\$, thì tương ứng mức tăng

$$\left(\widehat{\beta}_2 + \widehat{\beta}_3 \texttt{tv}\right) \times 1000 = 28.9 + 1.1 \times \texttt{tv}$$

  trong doanh số bán hàng.

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 $X_1, X_2, \ldots, X_p$. (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:

- với các điều kiện (A1)-(A3) thì $\widehat{\beta}_n$ là ước lượng không chệch, tức là $\mathbb{E}(\widehat{\beta}_n) = \beta$;
- nếu có thêm điều kiện (A4) thì $\mathbb{V}\mathrm{ar}(\widehat{\beta}_n) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$;
- với các điều kiện (A1)-(A5) thì $\widehat{\beta}_n \sim \mathcal{N}_{p+1}\left(\beta, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right)$;
- với các điều kiện (A1)-(A4) và (A6) thì $\widehat{\beta}_n \xrightarrow{d} \mathcal{N}_{p+1}\left(\beta, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right)$, khi $n \to \infty$;
- $\dfrac{\widehat{\beta}_{j,n} - \beta_j}{\widehat{\sigma}\sqrt{v_{jj}}} \sim t_{n-p-1}$, với $\widehat{\sigma}^2$ là ước lượng không chệch của σ^2:

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

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, cho giá trị tiên đoán; và cũng có thể thực hiện các kiểm định thống kê để trả lời một số câu hỏi quan trọng.

Bootstrap cho mô hình hồi quy

Trong thực tế, điều kiện (A4) và (A5) thường hay bị vi phạm.

$\hookrightarrow$ Do đó, để tìm được phân phối mẫu và sai số chuẩn cho $\widehat{\beta}_n$, ta sử dụng phương pháp bootstrap.

1. Tạo một mẫu ngẫu nhiên cỡ n từ dữ liệu gốc, có lặp lại (with replacement);
2. Ước lượng mô hình hồi quy với mẫu vừa tạo, lưu lại các ước lượng hệ số;
3. Lặp lại bước 1 và 2 trong R lần (ít nhất 500 lần), và lưu kết quả lại.

Ta có thể sử dụng R giá trị của thống kê để:

- xác định độ lệch chuẩn của các ước lượng hệ số $\Longrightarrow$ sai số chuẩn

$$\text{SE}(\widehat{\beta}_n) = \sqrt{\frac{1}{R-1}\sum_{j=1}^{R}\left(\widehat{\beta}_{jn} - \frac{1}{n}\sum_{j=1}^{R}\widehat{\beta}_{jn}\right)^2}.$$

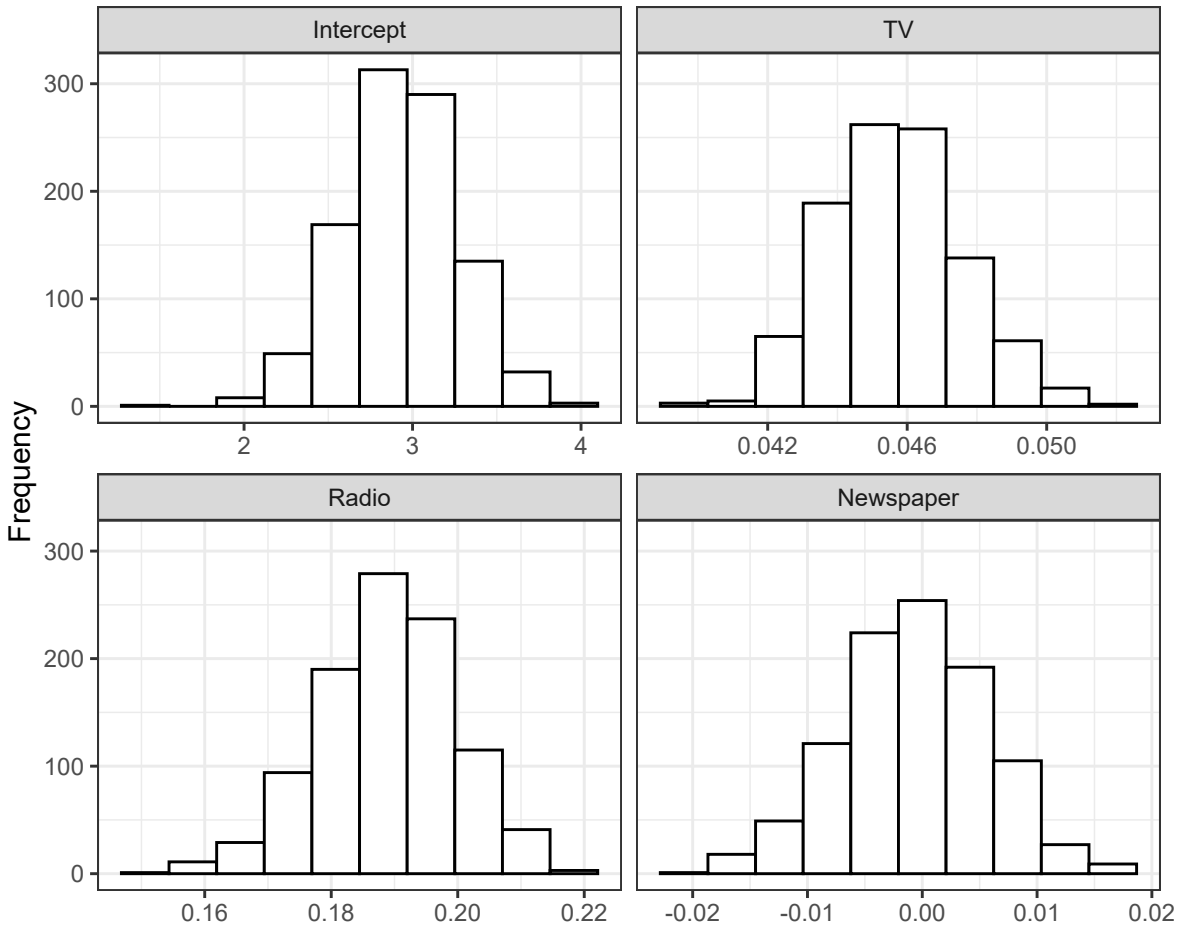
- biểu diễn histogram để mô tả phân phối mẫu của các ước lượng hệ số;
- xác định khoảng tin cậy cho ước lượng hệ số.

Bootstrap cho mô hình hồi quy

Ta xét lại mô hình hồi quy:

$$\texttt{sales} = \beta_0 + \beta_1 \texttt{tv} + \beta_2 \texttt{radio} + \beta_3 \texttt{newspaper} + \varepsilon$$

Áp dụng phương pháp bootstrap với 1000 lần lặp:



Bootstrap cho mô hình hồi quy

Ta thu được kết quả:

	Est	SE	95% CI
Intercept	2.9389	0.3333	(2.279, 3.592)
TV	0.0458	0.0019	(0.0422, 0.0496)
Radio	0.1885	0.0108	(0.1655, 0.2093)
Newspaper	−0.0010	0.0063	(−0.0136, 0.0113)

Kiểm định cho hệ số

Một trong những kiểm định được quan tâm trong mô hình hồi quy đó là:

$$\text{Giả thuyết: } \beta_j = 0$$
$$\text{Đối thuyết: } \beta_j \neq 0$$

Kiểm định Giả thuyết này tương đương với việc trả lời câu hỏi

"Có thực sự tồn tại mối liên hệ giữa biến X_j và Y".

Để thực hiện kiểm định này, theo lý thuyết, ta sẽ tính t-value

$$t = \frac{\widehat{\beta}_j}{\mathrm{SE}(\widehat{\beta}_j)}$$

dưới giả định (A4) và (A5). Sau đó, tính p-value:

$$p\text{-value} = \Pr(T > |t|),$$

với $T \sim t_{n-p-1}$, phân phối t-student với $n - p - 1$ bậc tự do.

Tuy nhiên, do điều kiện (A4) và (A5) thường hay bị vi phạm, nên ta tránh dùng cách kiểm định theo lý thuyết $\Longrightarrow$ sử dụng phương pháp bootstrap.

Kiểm định cho hệ số

Từ kết quả của bootstrap để xác định sai số chuẩn của ước lượng hệ số, ta có tính p-value bởi:

$$p\text{-value} = \begin{cases} 2q_j, & \text{nếu } q_j < 0.5, \\ 2(1-q_j), & \text{nếu } q_j \geq 0.5, \end{cases}$$

với

$$q_j = \frac{1}{R}\sum_{j=1}^{R} \mathrm{I}(\widehat{\beta}_{jn} \leq 0)$$

Áp dụng cách phương pháp này, ta có kết quả sau:

	Est	SE	95% CI	p-value
Intercept	2.9389	0.3333	(2.279, 3.592)	< 0.0001
TV	0.0458	0.0019	(0.0422, 0.0496)	< 0.0001
Radio	0.1885	0.0108	(0.1655, 0.2093)	< 0.0001
Newspaper	-0.0010	0.0063	$(-0.0136, 0.0113)$	0.916

Tiên đoán

Xét lại mô hình hồi quy đa biến

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

Trong mô hình này,

- thành phần biến hồi quy $X_1, X_2, \ldots, X_p$ được coi là cố định;
- thành phần sai số ε là ngẫu nhiên, với $\mathbb{E}(\varepsilon) = 0$.

Cho trước dữ liệu mới của các biến hồi quy, ta tính được

$$\mathbb{E}\left(Y | X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Do đó, với các ước lượng OLS, ta dễ dàng tính được giá trị ước lượng cho trung bình của Y khi biết giá trị của $X_1, X_2, \ldots, X_p$:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_p x_p.$$

Chú ý: Ta chỉ có thể tính giá trị tiên đoán với các dữ liệu mới nằm trong vùng được quan sát của các biến hồi quy trong dữ liệu gốc.

Tiên đoán

Ví dụ: ta có mô hình hồi quy đa biến cho dữ liệu doanh số bán hàng

$$\texttt{sales} = \beta_0 + \beta_1 \texttt{tv} + \beta_2 \texttt{radio} + \beta_3 \texttt{newspaper} + \varepsilon,$$

Mô hình ước lượng là

$$\texttt{sales} = 2.9389 + 0.0458 \times \texttt{tv} + 0.1885 \times \texttt{radio} - 0.001 \times \texttt{newspaper},$$

Giả sử ta biết rằng, chi phí dành cho quảng cáo lần lượt là

- `tv`: 150
- `radio`: 40
- `newspaper`: 55

khi đó, ước lượng trung bình doanh số bán hàng là

$$\texttt{sales} = 2.9389 + 0.0458 \times 150 + 0.1885 \times 40 - 0.001 \times 55 = 17.2939,$$

Khoảng tin cậy cho trung bình

Nhắc lại rằng, $\widehat{y}$ là ước lượng của $\mathbb{E}(Y|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$ - tức là tham số trung bình.

$\hookrightarrow$ xây dựng khoảng tin cậy cho $\mathbb{E}(Y|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$.

Về mặt lý thuyết, ta hoàn toàn có thể xác định khoảng tin cậy dựa trên công thức

$$\left(\widehat{y} - t_{1-\alpha/2,n-p-1}\mathrm{SE}(\widehat{y}), \widehat{y} + t_{1-\alpha/2,n-p-1}\mathrm{SE}(\widehat{y})\right),$$

vốn được xây dựng dựa trên giả định (A4) - (A5).

Tuy nhiên, như những phần trước, hai giả định này thường bị vi phạm.

$\hookrightarrow$ ta sử dụng phương pháp bootstrap để xác định khoảng tin cậy cho $\mathbb{E}(Y|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$.

Khoảng tin cậy cho trung bình

1. Tạo một mẫu ngẫu nhiên cỡ n từ dữ liệu gốc, có lặp lại (with replacement);
2. Ước lượng mô hình hồi quy với mẫu vừa tạo, lưu lại các ước lượng hệ số;
3. Tính $\widehat{y}$ dựa trên dữ liệu mới của biến hồi quy $x_1, x_2, \ldots, x_p$;
4. Lặp lại bước từ 1 tới 3 trong R lần (ít nhất 1000 lần), và lưu kết quả lại.
5. Xác định khoảng tin cậy bootstrap percentile.

Ví dụ: giả sử ta biết rằng, chi phí dành cho quảng cáo lần lượt là

- `tv`: 150
- `radio`: 40
- `newspaper`: 55

Với 1000 lần lặp bootstrap, ta xác định được khoảng tin cậy 95% cho trung bình của doanh số bán hàng là (16.8658, 17.7543) ngàn đơn vị sản phẩm.

Khoảng dự đoán

Khoảng tin cậy cho $\mathbb{E}(Y|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$ không xét tới sự biến động của sai số ε.
$\hookrightarrow$ thiếu tính thực tiễn cho việc dự đoán.

Khoảng dự đoán

Khoảng dự đoán cung cấp một vùng tin cậy cho sự thay đổi giá trị dự đoán của biến phản hồi Y dựa trên kết quả ước lượng mô hình:

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

Để xác định khoảng dự đoán, ta cần

- ước lượng OLS của các hệ số;
- ước lượng thặng dư của mô hình ước lượng:

$$\widehat{\varepsilon_i} = Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + \ldots + \widehat{\beta}_p X_{pi}\right)$$

Khoảng dự đoán

1. Tạo một mẫu ngẫu nhiên cỡ n từ dữ liệu gốc, có lặp lại (with replacement);
2. Ước lượng mô hình hồi quy với mẫu vừa tạo, lưu lại các ước lượng hệ số;
3. Tính $\widehat{y}$ dựa trên dữ liệu mới của biến hồi quy $x_1, x_2, \ldots, x_p$;
4. Lấy mẫu ngẫu nhiên 1 giá trị thặng dư $\widehat{\varepsilon}$ của mô hình, và tính $\widetilde{y} = \widehat{y} + \widehat{\varepsilon}$.
5. Lặp lại bước từ 1 tới 4 trong R lần (ít nhất 1000 lần), và lưu kết quả lại.
6. Xác định khoảng tin cậy bootstrap percentile.

Ví dụ: giả sử ta biết rằng, chi phí dành cho quảng cáo lần lượt là

- `tv`: 150
- `radio`: 40
- `newspaper`: 55

Với 1000 lần lặp bootstrap, ta xác định được khoảng dự đoán 95% cho doanh số bán hàng là (13.8388, 19.6606) ngàn đơn vị sản phẩm.

Khoảng dự đoán lớn hơn khoảng tin cậy, do nó có tính tới sự biến động của sai số trong mô hình.

Các chỉ số đánh giá tổng quát

- Một trong những chỉ số quan trọng, thường được dùng data science (machine learning) là **root mean square error** - căn bậc hai trung bình bình phương sai số:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}$$

  RMSE đo lường độ chính xác tổng thể của mô hình và là cơ sở để so sánh với các mô hình khác.

- Trong lý thuyết mô hình hồi quy, một chỉ số đo lường khác, tương tự với RMSE là **residual standard error** - sai số chuẩn thặng dư:

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}$$

- Về mặt ý tưởng, mô hình tốt, mô tả khớp với dữ liệu, là mô hình có RMSE và RSE nhỏ.

Các chỉ số đánh giá tổng quát

Ví dụ: Với dữ liệu về doanh số bán hàng và chi phí quảng cáo, ta có ba mô hình đã được xét:

- MH1: $\texttt{sales} = \beta_0 + \beta_1 \texttt{tv} + \varepsilon$
- MH2: $\texttt{sales} = \beta_0 + \beta_1 \texttt{tv} + \beta_2 \texttt{radio} + \beta_3 \texttt{newspaper} + \varepsilon$
- MH3: $\texttt{sales} = \beta_0 + \beta_1 \texttt{tv} + \beta_2 \texttt{radio} + \beta_3 \texttt{tv} \times \texttt{radio} + \varepsilon$

Chỉ số RMSE và RSE cho từng mô hình là

- MH1: RMSE $= 3.242$, RSE $= 3.259$;
- MH2: RMSE $= 1.669$, RSE $= 1.686$;
- MH3: RMSE $= 0.934$, RSE $= 0.944$.

Sự khác biệt giữa RMSE và RSE là không đáng kể.

Hệ số xác định

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 $X_1, X_2, \ldots, X_p$.

Cụ thể, R^2 được tính theo công thức

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2},$$

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).

Hệ số xác định

Ví dụ: với mô hình ước lượng

$$\texttt{sales} = 2.9389 + 0.0458 \times \texttt{tv} + 0.1885 \times \texttt{radio} - 0.001 \times \texttt{newspaper},$$

chỉ số xác định R^2 là 0.8972. Có nghĩa là mô hình này có thể giải thích được khoảng 89.72% độ biến động của doanh số bán sản phẩm.

Về mặt ý tưởng, chỉ số xác định R^2 càng cao thì mô hình càng tốt.

Ví dụ: với mô hình ước lượng

$$\texttt{sales} = 6.7502 + 0.0191 \times \texttt{tv} + 0.0289 \times \texttt{radio} + 0.0011 \times \texttt{tv} \times \texttt{radio},$$

chỉ số xác định R^2 là 0.9678. Có nghĩa là mô hình này có thể giải thích được khoảng 96.78% độ biến động của doanh số bán sản phẩm.

Mô hình với sự tương tác giữa chi phí quảng cáo TV và quảng cáo Radio là tốt hơn so với mô hình không có sự tương tác.

Hệ số xác định được hiệu chỉnh

Hệ số xác định 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 Y.

Ta dùng hệ số R^2 được 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.
- Đối với mô hình tuyến tính đa biến, ta ưu tiên sử dụng hệ số xác định được điều chỉnh R_a^2.

Ví dụ: với mô hình ước lượng

$$\texttt{sales} = 6.7502 + 0.0191 \times \texttt{tv} + 0.0289 \times \texttt{radio} + 0.0011 \times \texttt{tv} \times \texttt{radio},$$

chỉ số xác định R^2 là 0.9673. Có nghĩa là mô hình này có thể giải thích được khoảng 96.73% độ biến động của doanh số bán sản phẩm.

Phương pháp tập xác thực - the validation set approach

Một trong những cách thường được dùng để đánh giá mô hình là sử dụng phương pháp tập xác thực.

Trong đó, tập dữ liệu được chia là hai phần (theo tỷ lệ cụ thể):

- training set: dùng để ước lượng mô hình;
- test set: dùng để so sánh giá trị Y gốc so với giá trị $\widehat{Y}$ từ mô hình ước lượng bởi training set.

Một trong những chỉ số so sánh được dùng là RMSE.

Phương pháp tập xác thực - the validation set approach

Tuy nhiên, phương pháp này có một số nhược điểm là:

- Tỷ lệ training set và test set là bao nhiêu thì phù hợp? (Ta thường chọn tỷ lệ 7-3).
- Cỡ mẫu nhỏ, việc chia thành 2 tập dữ liệu, là không hợp lý.
- Việc chia dữ liệu là ngẫu nhiên, do đó, nếu ta lặp lại nhiều lần, sai số trong test set sẽ thay đổi, có thể biến động cao.
- Nếu training set có ít quan sát, sẽ dẫn tới ước lượng "xấu" của mô hình, hệ quả là đánh giá sai sai số trong test set.

Phương pháp tập xác thực - the validation set approach

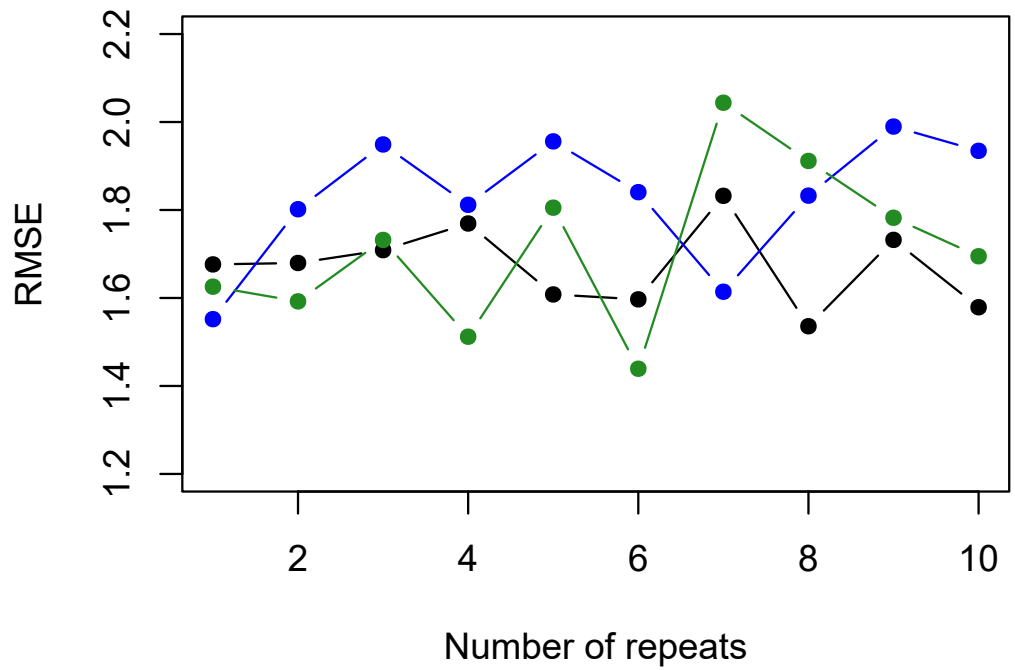
Ví dụ: xét mô hình

$$\texttt{sales} = \beta_0 + \beta_1 \texttt{tv} + \beta_2 \texttt{radio} + \beta_3 \texttt{newspaper} + \varepsilon$$

Ta tiến hành chia training set và test set theo ba tỷ lệ

- 5-5 (đường màu đen)
- 6-4 (đường màu xanh dương)
- 7-3 (đường màu xanh lá)

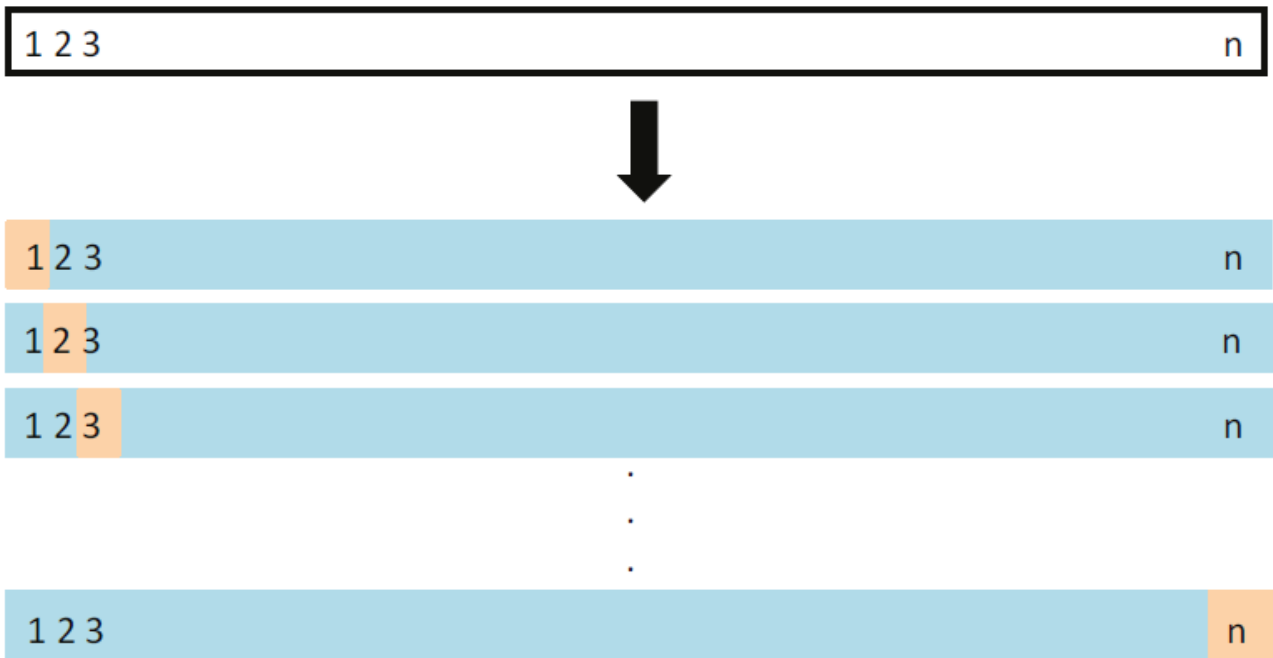
và lặp lại 10 lần.



Leave-one-out-cross-validation

Leave-one-out-cross-validation (LOOCV) là một phương pháp đánh giá với ý tưởng gần với phương pháp tập xác thực.

Trong đó, tập dữ liệu được chia là hai phần:



- training set: dùng để ước lượng mô hình, gồm $n-1$ quan sát;
- test set: dùng để so sánh giá trị Y gốc so với giá trị $\widehat{Y}$ từ mô hình ước lượng bởi training set, gồm 1 quan sát.

Leave-one-out-cross-validation

Khi đó, chỉ số đánh giá mô hình sẽ là trung bình cộng của chỉ số được sử dụng trong mỗi đánh giá mô hình khi bỏ một quan sát trong mô hình.

Chẳng hạn, nếu ta dùng RMSE thì chỉ số được quan tâm là

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{RMSE}_i$$

Ví dụ: xét mô hình

$$\texttt{sales} = \beta_0 + \beta_1\texttt{tv} + \beta_2\texttt{radio} + \beta_3\texttt{newspaper} + \varepsilon$$

Ta có 200 quan sát trong dữ liệu.

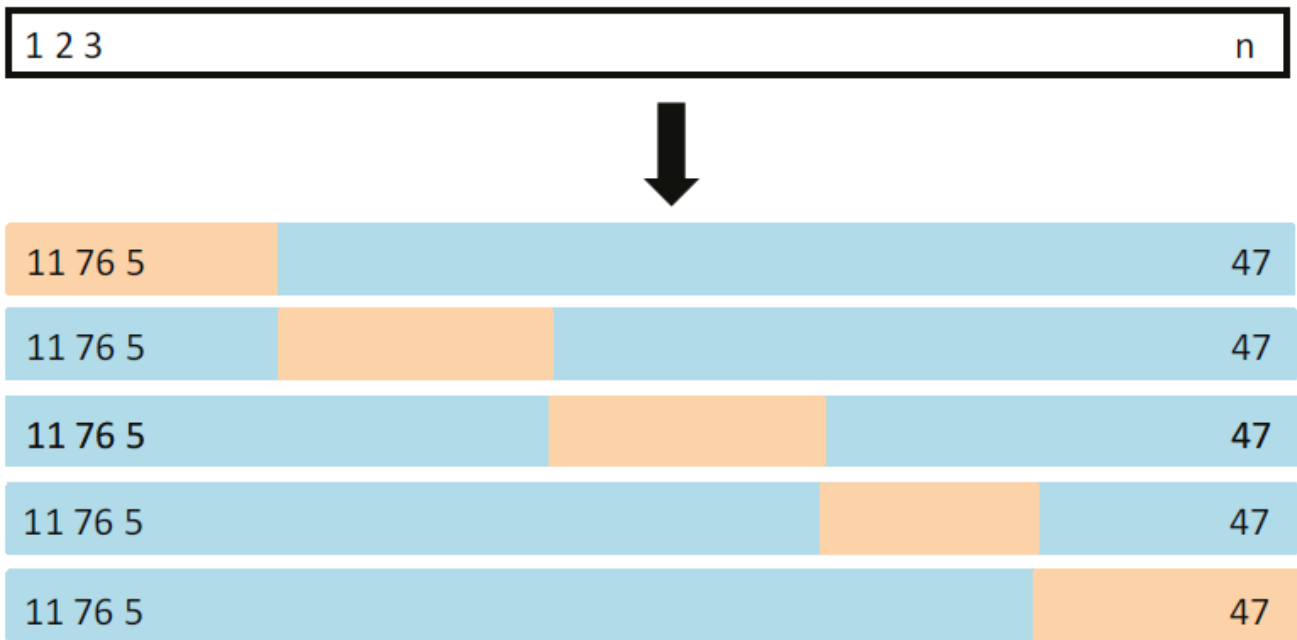
Áp dụng LOOCV ta tính được $\mathrm{CV}_{(n)} = 1.2829$.

Chú ý: Nếu cỡ mẫu lớn (trên 1000) thì thời gian tính $\mathrm{CV}_{(n)}$ sẽ lớn.

k-fold cross-validation

k-fold cross-validation là một phương pháp thay thế LOOCV, để giải quyết vấn đề thời gian tính toán khi cỡ mẫu lớn.

Trong đó, tập dữ liệu được chia là k phần với cỡ mẫu xấp xỉ bằng nhau:



- training set: dùng để ước lượng mô hình, gồm $(k-1)n/k$ quan sát;
- test set: dùng để so sánh giá trị Y gốc so với giá trị $\widehat{Y}$ từ mô hình ước lượng bởi training set, gồm n/k quan sát.

k-fold cross-validation

Thông thường

- $k = 5$, hoặc
- $k = 10$

được sử dụng rộng rãi.

Chỉ số đánh giá mô hình sẽ là trung bình cộng của chỉ số được sử dụng trong mỗi đánh giá mô hình.

Chẳng hạn, nếu ta dùng RMSE thì chỉ số được quan tâm là

$$\mathrm{CV}_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \mathrm{RMSE}_i$$

Ví dụ: xét mô hình

$$\texttt{sales} = \beta_0 + \beta_1 \texttt{tv} + \beta_2 \texttt{radio} + \beta_3 \texttt{newspaper} + \varepsilon$$

Áp dụng k-fold CV, ta tính được $\mathrm{CV}_{(5)} = 1.7098$ và $\mathrm{CV}_{(10)} = 1.6913$.