Bài giảng 5: Regression và Prediction - II

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

1 Chuẩn đoán mô hình

2 Lựa chọn mô hình

3 Mở rộng mô hình

4 Generalized Linear Models

1 Chuẩn đoán mô hình

2 Lựa chọn mô hình

3 Mở rộng mô hình

4 Generalized Linear Models

Chuẩn đoán mô hình

Chuẩn đoán mô hình là một trong những bước quan trọng của xây dựng và sử dụng mô hình hồi quy tuyến tính.

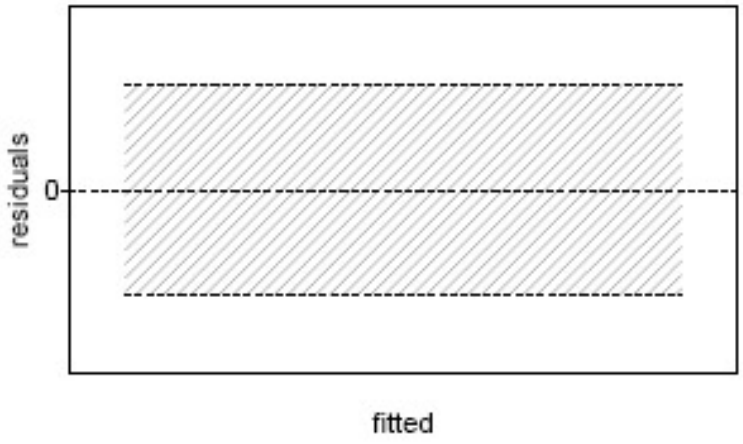
Chuẩn đoán mô hình bao gồm:

- Kiểm tra sự quan hệ tuyến tính giữa Y và các biến hồi quy $X_1, X_2, \ldots X_p$ (giả định A1).
- Mối quan hệ tuyến tính giữa Y và một biến hồi quy X_j.
- Tính đồng nhất phương sai của thặng dư (giả định A4) - Homoskedasticity.
- Tính độc lập của thặng dư.
- Các giá trị ngoại lai xuất hiện trong (giả định A6).

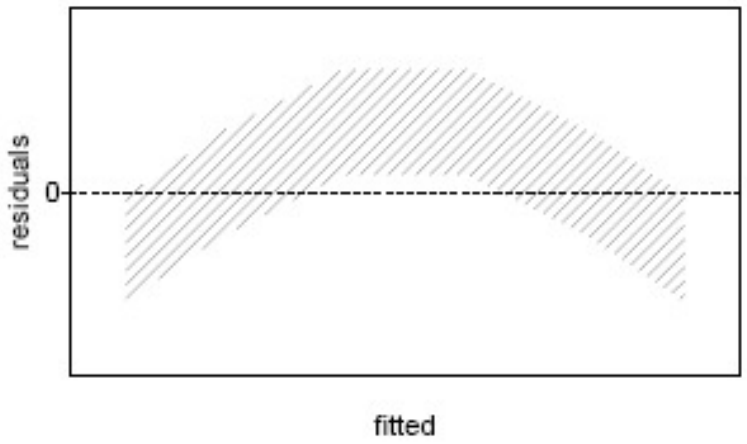
$\hookrightarrow$ các chuẩn đoán mô hình này giúp ta biết được mô hình được xây dựng đã thực sự hợp lý với dữ liệu, và mô hình có vững.

Kiểm tra tính tuyến tính

- Để kiểm tra sự quan hệ tuyến tính giữa Y và các biến hồi quy $X_1, X_2, \ldots X_p$ (giả định A1), 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 thặng dư $= 0$.
  $\hookrightarrow$ giả định 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 vấn đề với một số khía cạnh của mô hình tuyến tính.
  $\hookrightarrow$ giả định không được thỏa mãn. (xem Hình b)



(a)

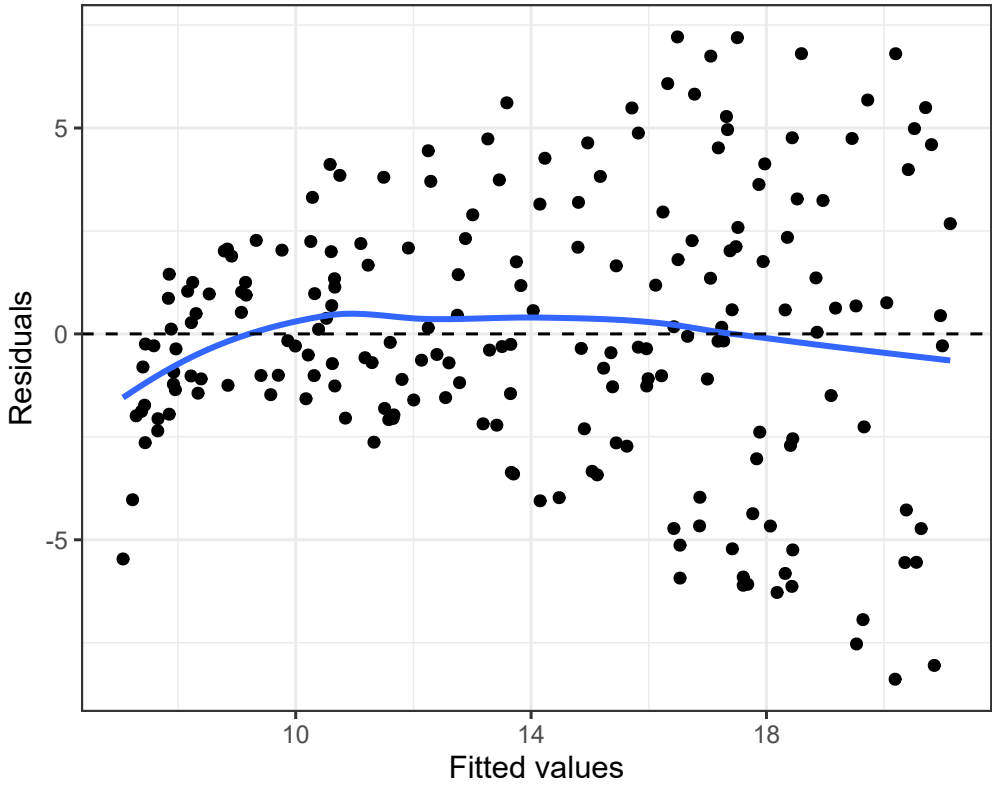


(b)

Kiểm tra tính tuyến tính

Xét biểu đồ Residuals vs Fitted của mô hình hồi quy tuyến tính:

$$\texttt{sales} = \beta_0 + \beta_1 \texttt{tv} + \varepsilon.$$



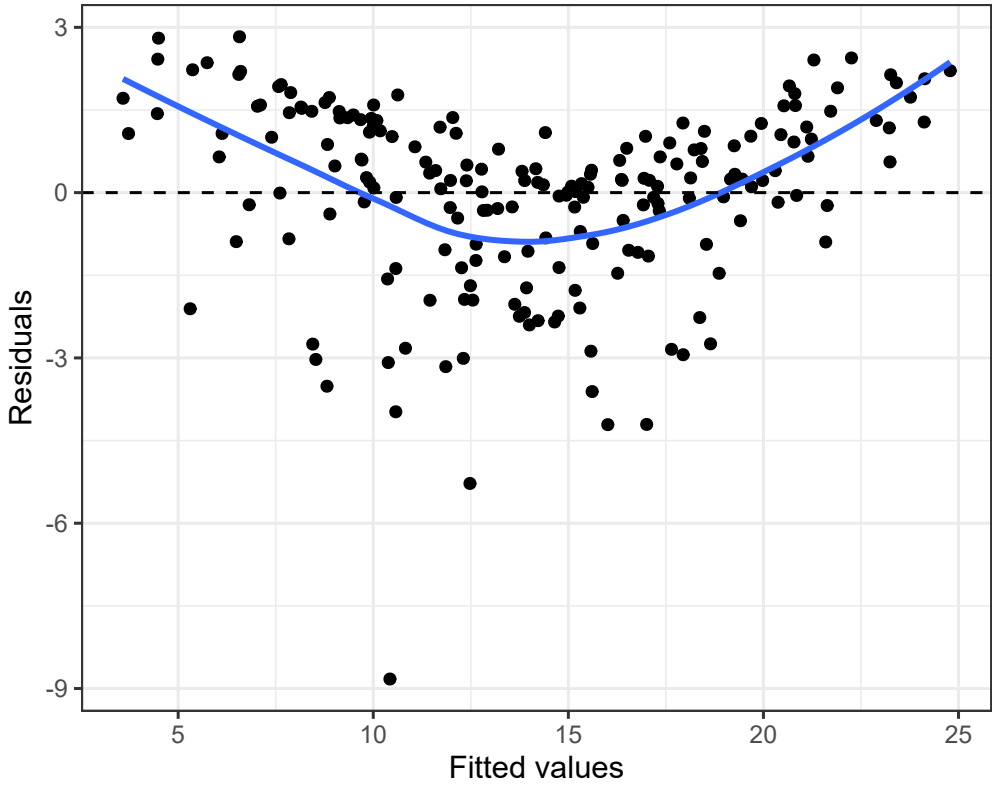
Hình vẽ không cho thấy xu hướng đường cong nào đáng kể.

$\hookrightarrow$ Giả định về tính tuyến tính của mô hình là phù hợp.

Kiểm tra tính tuyến tính

Xét biểu đồ Residuals vs Fitted của mô hình hồi quy tuyến tính:

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



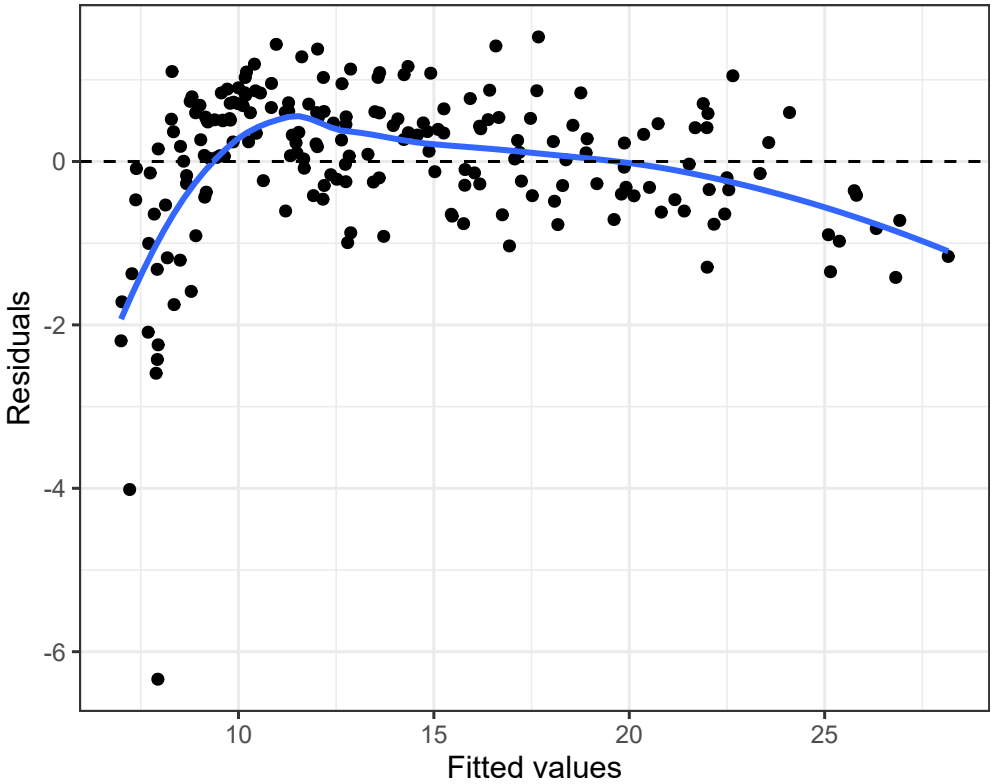
Hình vẽ cho thấy một xu hướng đường cong rõ ràng.

$\hookrightarrow$ Giả định về tính tuyến tính của mô hình là không phù hợp.

Kiểm tra tính tuyến tính

Xét biểu đồ Residuals vs Fitted của mô hình hồi quy tuyến tính:

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



Hình vẽ cho thấy một xu hướng đường cong tương đối.

$\hookrightarrow$ Giả định về tính tuyến tính của mô hình là không phù hợp.

Sự tuyến tính từng phần

Khi sự quan hệ tuyến tính giữa Y và các biến hồi quy $X_1, X_2, \ldots, X_p$ không hợp lý.

$\hookrightarrow$ có lẽ có một vài biến không hỗ trợ quan hệ tuyến tính.

Để kiểm tra sự tuyến tính của Y đối với từng biến hồi quy X_j, ta sử dụng biểu đồ **thặng dư từng phần** (**partial residual plots**):

$$\widehat{\varepsilon}_{ji} = \widehat{\varepsilon}_i + \widehat{\beta}_j X_{ji}.$$

Ý tưởng của biểu đồ thặng dư từng phần là để cô lập mối liên hệ giữa một biến hồi quy X_j và Y, nhưng tính đến mối liên hệ của tất cả các biến hồi quy còn lại.

Thông qua biểu đồ:

- nếu các điểm thặng dư từng phần $\widehat{\varepsilon}_{ji}$ có xu hướng theo một đường thẳng
  $\hookrightarrow$ mối liên hệ tuyến tính giữa Y và X_j là hợp lý;
- nếu các điểm thặng dư từng phần $\widehat{\varepsilon}_{ji}$ có xu hướng theo một đường cong
  $\hookrightarrow$ mối liên hệ tuyến tính giữa Y và X_j là không hợp lý.

Sự tuyến tính từng phần

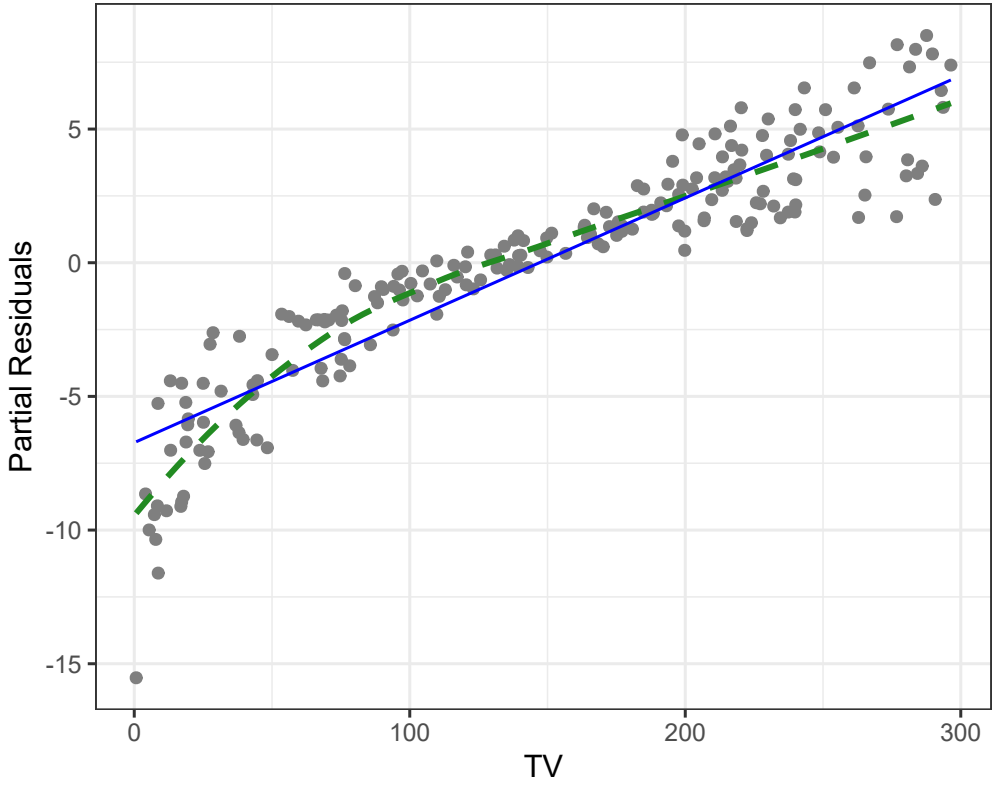
Xét mô hình hồi quy tuyến tính:

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

Ta xét ba biểu đồ thặng dư từng phần cho:

- `tv`
- `radio`
- `newspaper`

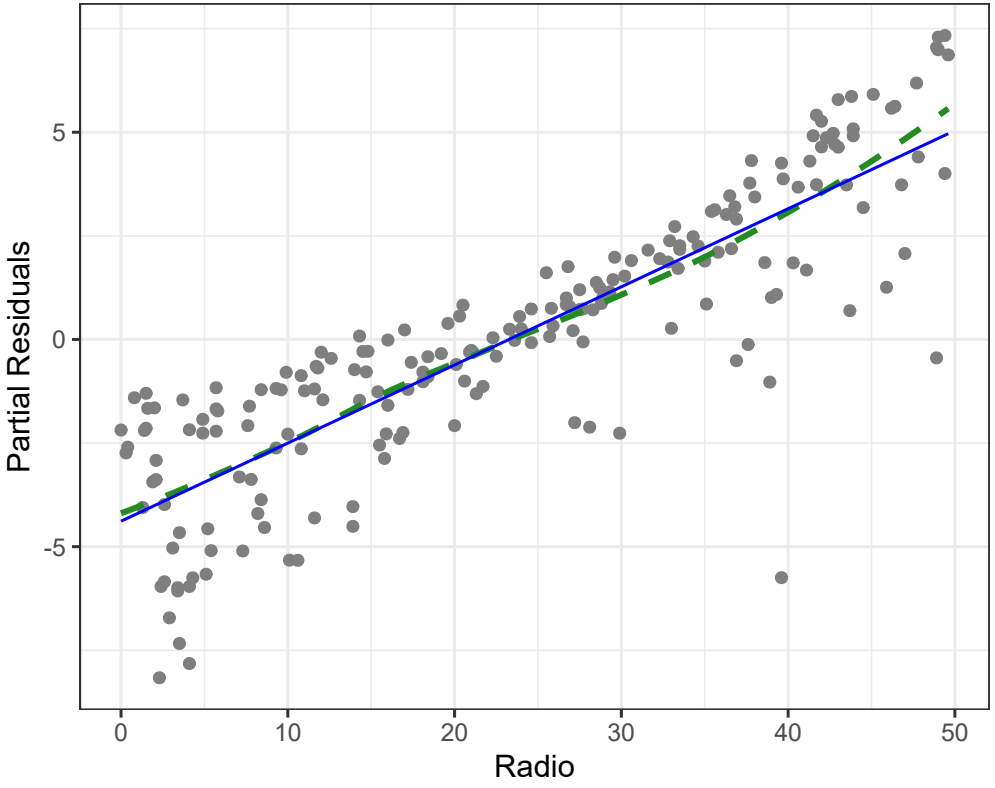
Sự tuyến tính từng phần



Kết quả cho thấy

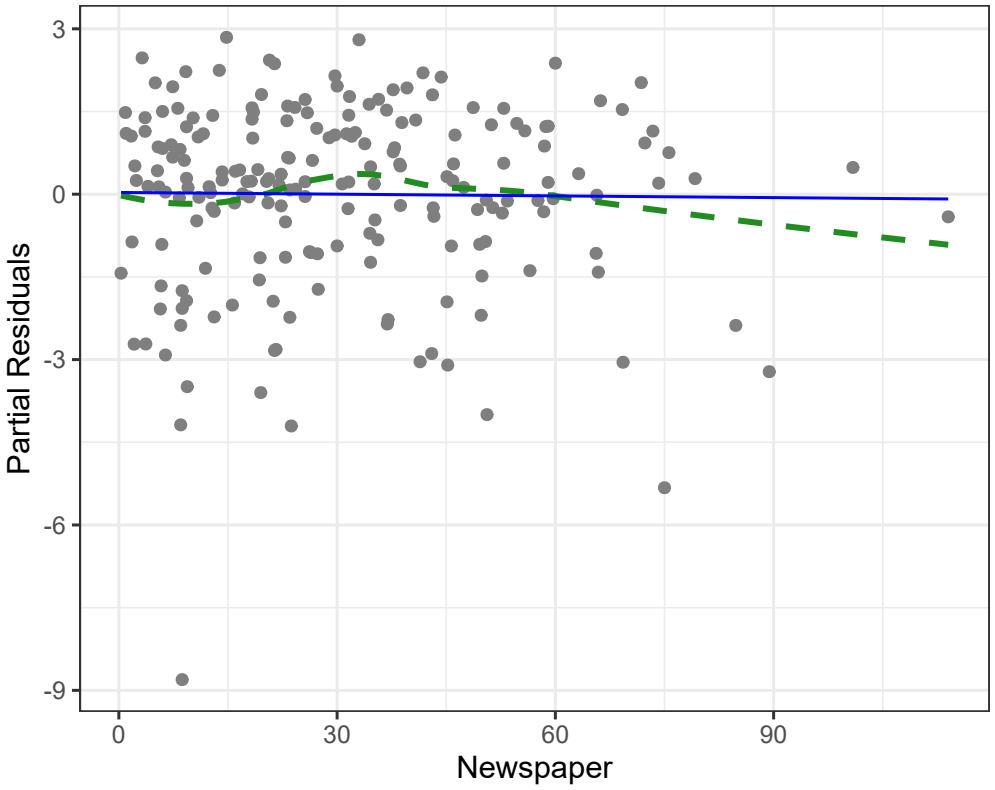
- đường thẳng tuyến tính (màu xanh dương) ước lượng không khớp với dữ liệu. Underestimate trong khoảng từ 80 tới 200;
- đường cong nét đứt (màu xanh lá) cho thấy xu hướng mối quan hệ phi tuyến tính của `tv` và `sales`.

Sự tuyến tính từng phần



Kết quả cho thấy đường thẳng tuyến tính (màu xanh dương) ước lượng tương đối khớp với dữ liệu.

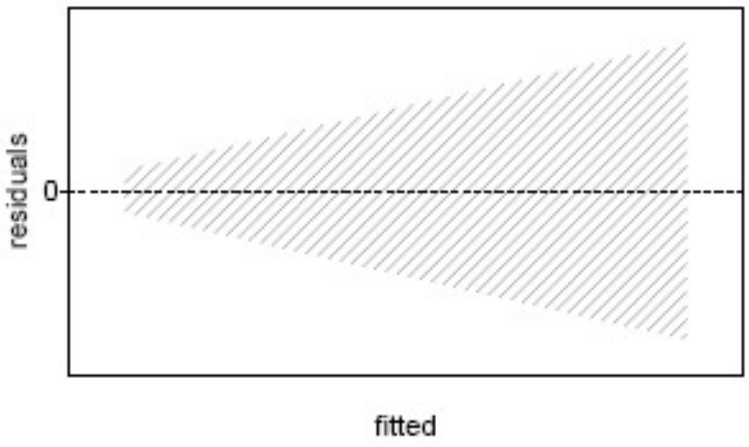
Sự tuyến tính từng phần

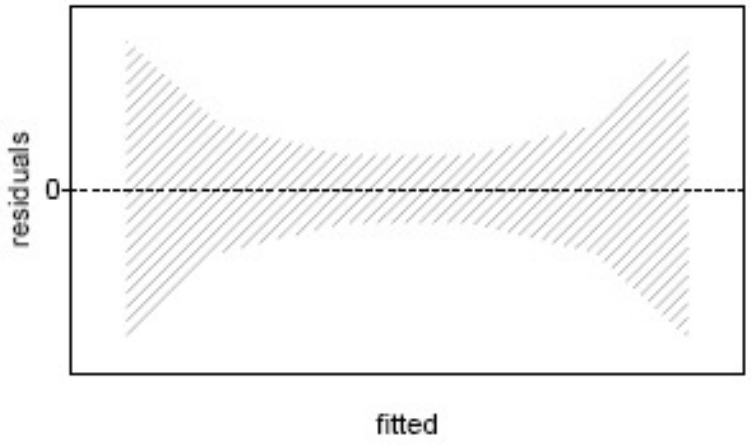


Kết quả cho thấy không có quan hệ tuyến tính giữa `sales` và `newspaper` trong mô hình.

Kiểm tra đồng nhất phương sai - Homoskedasticity

- Để kiểm tra giả định đồng nhất phương sai - Homoskedasticity, ta sử dụng biểu đồ Residuals vs Fitted hoặc Scale-Location.
- Đối với biểu đồ Residuals vs Fitted, nếu độ biến thiên của phần thặng dư thay đổi từ trái sang phải, như thế này:





thì điều kiện Homoskedasticity bị vi phạm, khi đó ta nói thặng dư của mô hình bị **heteroskedasticity**.

Kiểm tra đồng nhất phương sai - Homoskedasticity

Nhắc lại mô hình hồi quy tuyến tính

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

Ước lượng OLS $\widehat{\beta}_n = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}$. Do đó, giá trị fitted là

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

Do đó, ta thặng dư của mô hình được viết thành

$$\widehat{\varepsilon} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\,\mathbf{Y}$$

Khi đó, ta tính được (dưới giả định đồng phương sai)

$$\mathbb{V}\mathrm{ar}(\widehat{\varepsilon}) = \sigma^2\,(\mathbf{I} - \mathbf{H})$$

Hay $\mathbb{V}\mathrm{ar}(\widehat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$, với h_{ii} là thành phần đường chéo của ma trận $\mathbf{H}$.

Ta định nghĩa standardized residuals là

$$r_i = \frac{\widehat{\varepsilon}_i}{\sqrt{\mathbb{V}\mathrm{ar}(\widehat{\varepsilon}_i)}} = \frac{y_i - \widehat{y}_i}{s\sqrt{1 - h_{ii}}},$$

s là ước lượng vững của σ.

Kiểm tra đồng nhất phương sai - Homoskedasticity

Một số tính chất thống kê của thặng dư chuẩn hóa (standardized residuals):

- r_i của độ lệch chuẩn bằng 1, nếu giả định về đồng nhất phương sai là đúng;
- các điểm r_i sẽ phân bố khá đồng đều theo chiều ngang khi so sánh với giá trị fitted.

Biểu đồ Scale-Location có trục

- x là giá trị fitted;
- trục y là $\sqrt{|\text{standardized residuals}|}$.

Nếu biểu đồ 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.

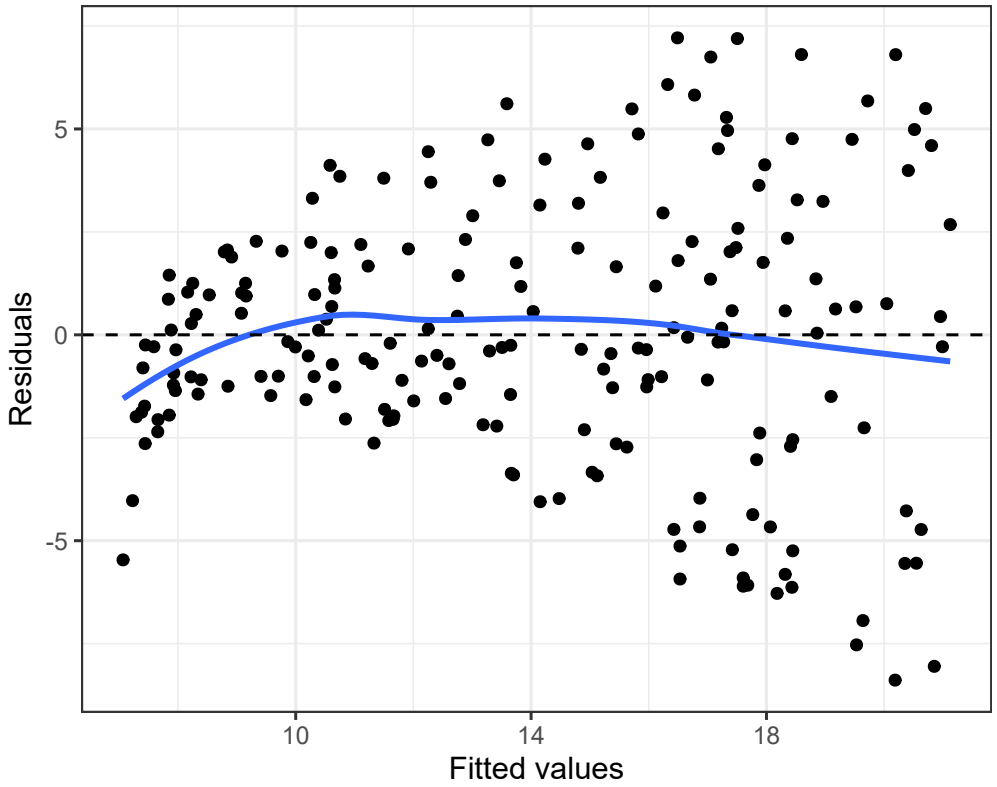
Chú ý:

Sự không đồng nhất phương sai trong mô hình (Heteroskedasticity) chỉ ra rằng các sai số dự đoán khác nhau đối với các phạm vi khác nhau của giá trị dự đoán và có thể gợi ý một mô hình không đầy đủ.

Kiểm tra đồng nhất phương sai - Homoskedasticity

Xét biểu đồ Residuals vs Fitted của mô hình

$$\texttt{sales} = \beta_0 + \beta_1 \texttt{tv} + \varepsilon.$$

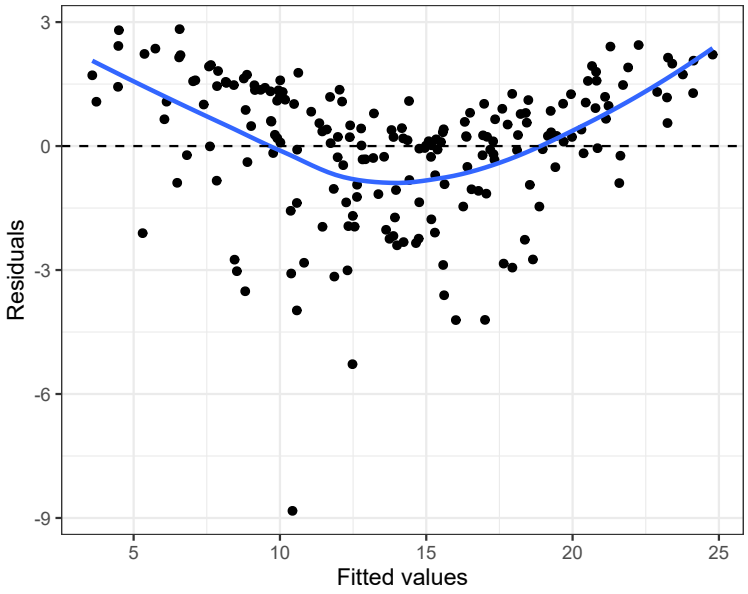


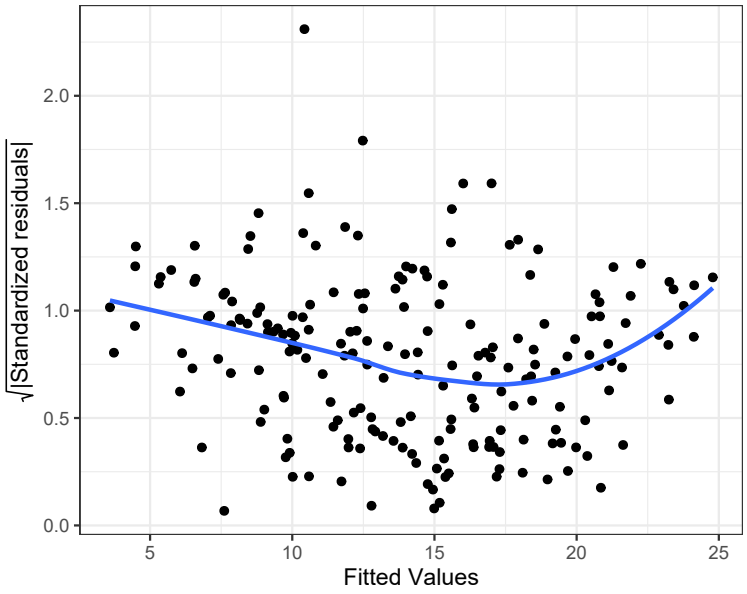
Các điểm thặng dư tăng theo giá trị ước lượng của biến đáp ứng, cho thấy phương sai không cố định trong các phần thặng dư (hoặc phương sai thay đổi - heteroscedasticity).

Kiểm tra đồng nhất phương sai - Homoskedasticity

Xét biểu đồ Residuals vs Fitted và Scale-Location của mô hình hồi quy tuyến tính:

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





Nhìn vào hình thứ hai, có thể thấy xu hướng của dữ liệu không xấp xỉ đường thẳng 1, ám chỉ phương sai thặng dư là không đồng nhất.

Tính độc lập của thặng dư

Khi tính độc lập của thặng dư không được thỏa mãn, thì

- mô hình được xây dựng là không phù hợp với dữ liệu,
- ước lượng của hệ số không đảm bảo tính vững,
- các kết quả thống kê suy luận không còn chính xác.

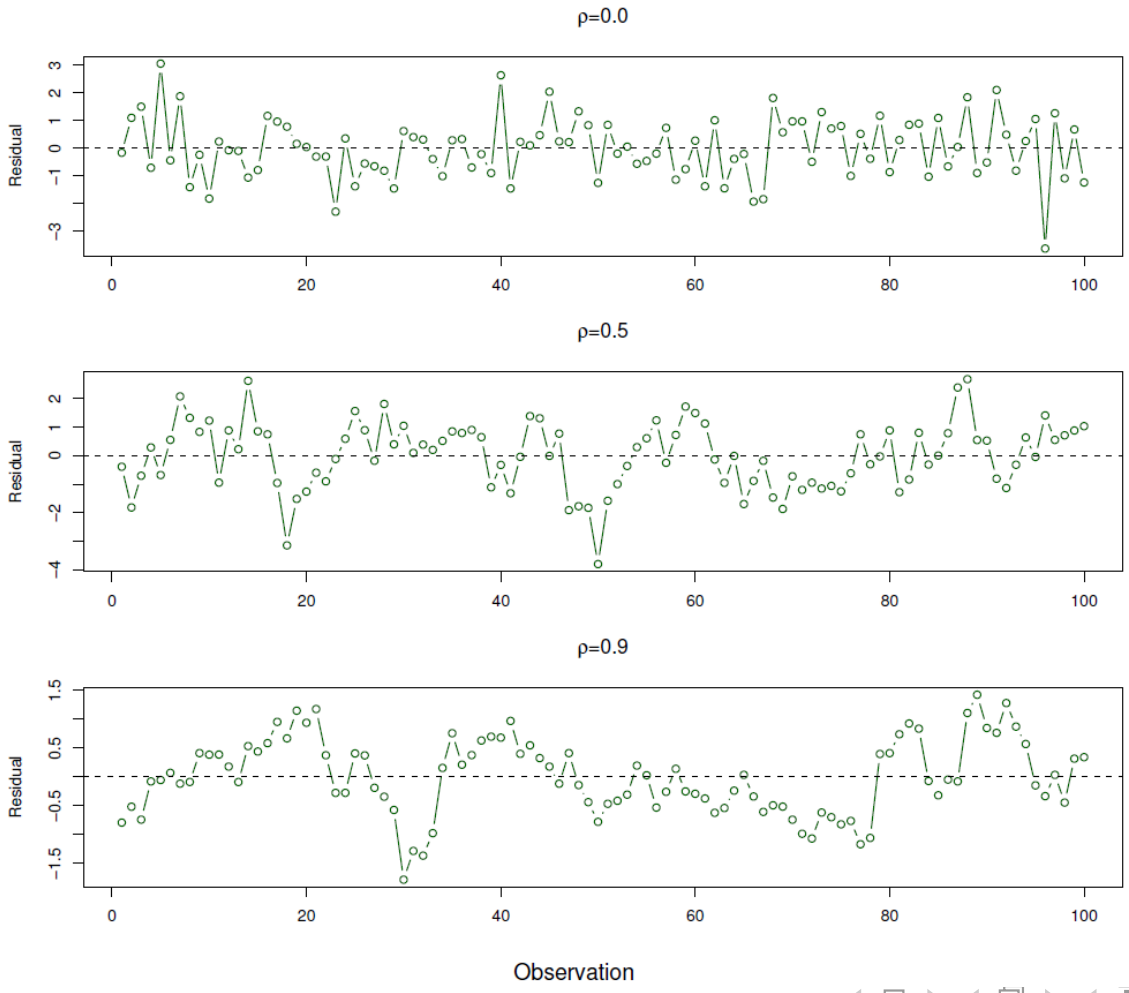
Điều kiện này thường sai khi dữ liệu được thu thập theo thời gian:

- thời gian máy bay khởi hành trễ theo các ngày trong năm;
- số lượng sinh viên theo năm học;
- doanh số bán hàng theo tháng.

Tính độc lập của thặng dư

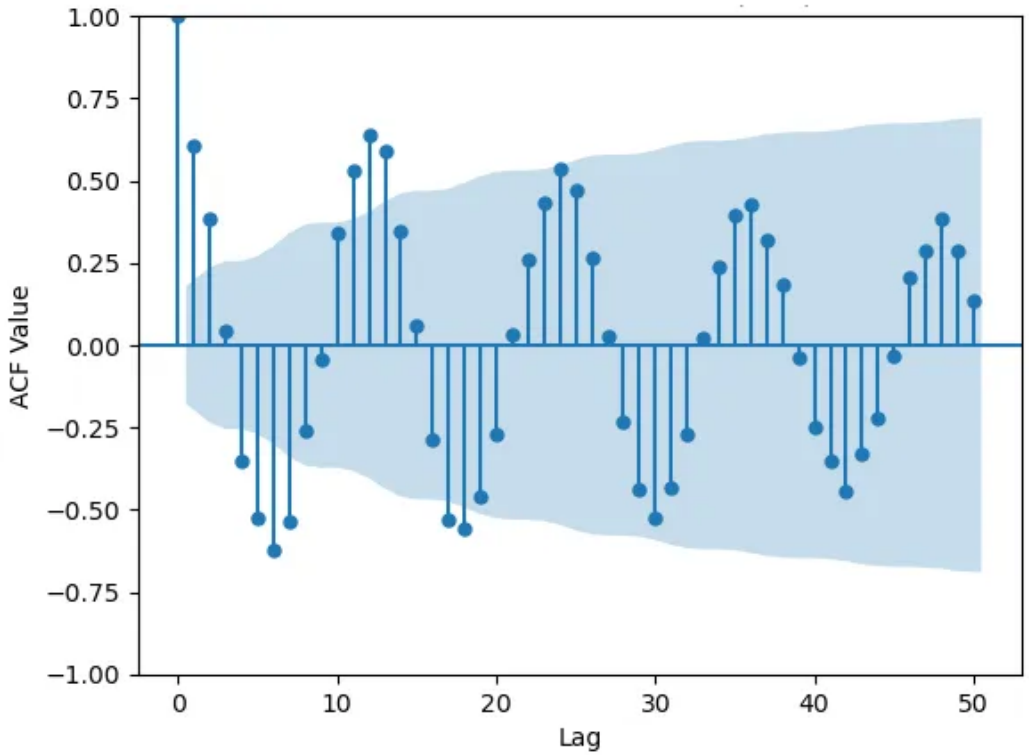
Sự khác biệt của thặng dư trong các tình huống

- độc lập ($\rho = 0$),
- có tương quan ($\rho = 0.5$ và $\rho = 0.8$).



Tính độc lập của thặng dư

Trong thực nghiệm, ta có thể dùng đồ thị Auto Correlation Function (ACF) của các thặng dư để kiểm ta sự tương quan giữa các thặng dư (có ý nghĩa hay không):

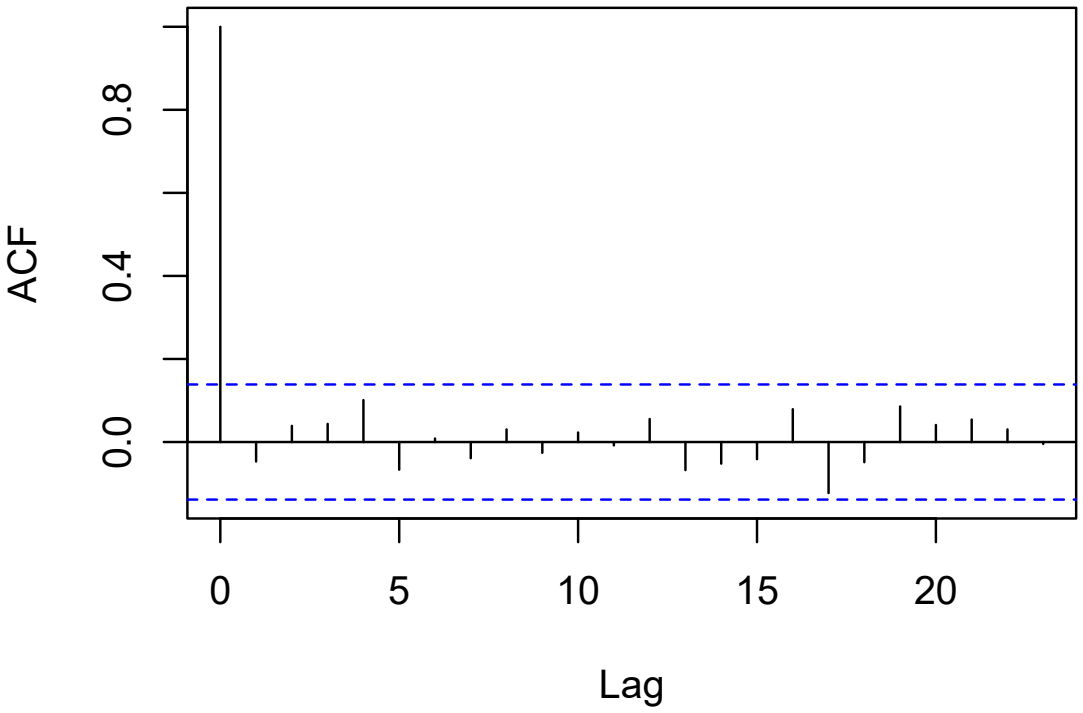


Đồ thị trên cho thấy sự tương quan giữa các thặng dư là có ý nghĩa.

Tính độc lập của thặng dư

Đồ thị Auto Correlation Function (ACF) cho thặng dư của mô hình

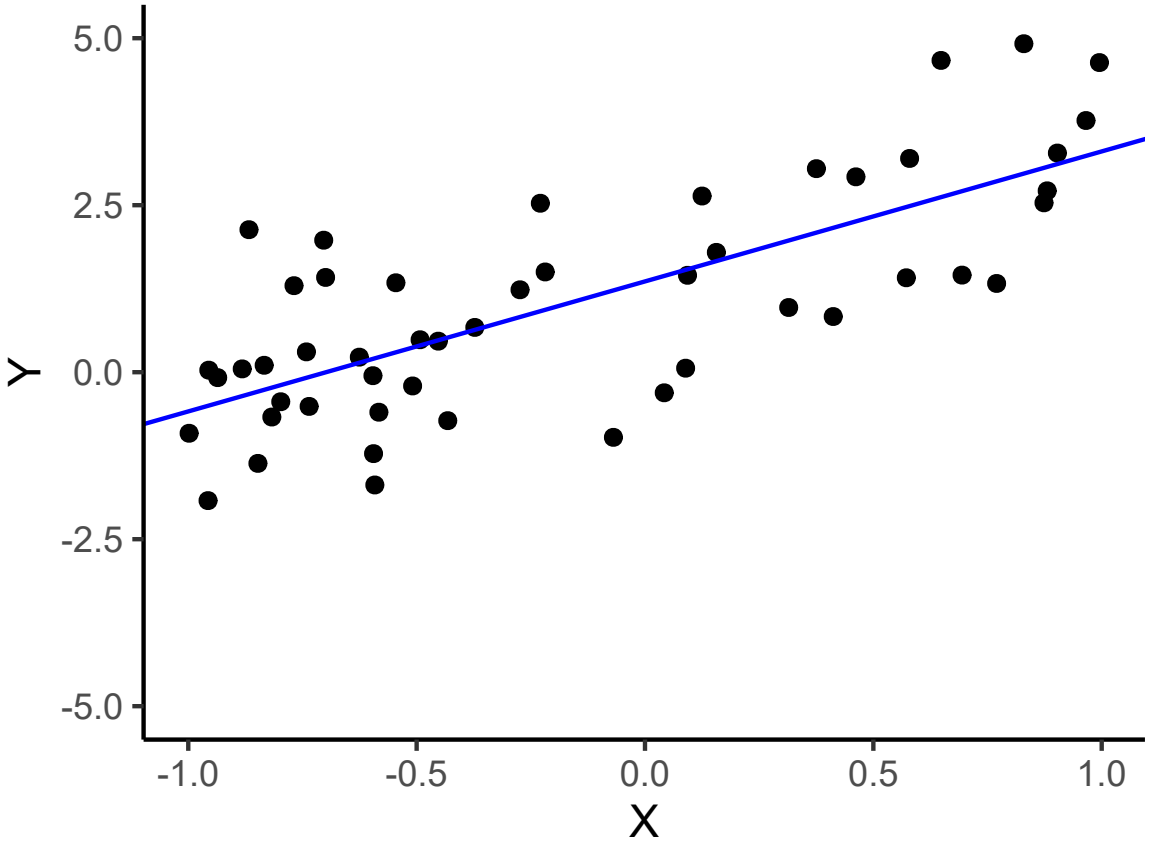
$$\texttt{sales} = \beta_0 + \beta_1\texttt{tv} + \beta_2\texttt{radio} + \beta_3\texttt{newspaper} + \varepsilon.$$



Kết quả cho thấy sự độc lập của các thặng dư.

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.



Phát hiện giá trị ngoại lai

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.

Leverage là giá trị h_{ii} trên đường chéo của ma trận **H**.

Vì $\mathbb{V}\mathrm{ar}(\widehat{y}_i) = \sigma^2 h_{ii}$, $\mathbb{V}\mathrm{ar}(\widehat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$, và $0 \leq h_{ii} \leq 1$, nên

- giá trị h_{ii} lớn gần 1, thì $\mathbb{V}\mathrm{ar}(\widehat{y}_i) \approx \sigma^2 = \mathbb{V}\mathrm{ar}(y_i)$ và $\mathbb{V}\mathrm{ar}(\widehat{\varepsilon}_i) \approx 0$

  $\hookrightarrow$ y_i có ảnh hưởng lớn tới $\widehat{y}_i$.
- giá trị h_{ii} nhỏ gần 0, thì $\mathbb{V}\mathrm{ar}(\widehat{y}_i)$ là tương đối nhỏ

  $\hookrightarrow$ $\widehat{y}_i$ là được dựa trên đóng góp của nhiều quan sát, và y_i không có ảnh hưởng lớn tới $\widehat{y}_i$.

Phát hiện giá trị ngoại lai

Một độ đo khác là khoảng cách Cook (Cook's distance), có thể dùng để xác định điểm ngoại lai trong mô hình tuyến tính:

$$D_i = r_i^2 \frac{h_{ii}}{p(1 - h_{ii})},$$

trong đó,

- p là số hệ số trong mô hình,
- r_i là thặng dư được chỉnh hóa,
- h_{ii} là giá trị leverage.

Những quan sát có giá trị khoảng cách Cook, D_i, lớn (thường là lớn hơn 0.5 - 1), tương ứng với trường hợp cả r_i và h_{ii} là lớn.

$\hookrightarrow$ những quan sát đó là các điểm ngoại lai (có ảnh hưởng lớn) trong mô hình.

$\hookrightarrow$ kết quả hồi quy sẽ bị thay đổi nếu chúng ta loại trừ những quan sát này.

Ta có thể tích hợp khoảng cách Cook vào trong biểu đồ Residuals vs Leverage, trong đó, khoảng cách Cook biểu thị kích cỡ của các điểm quan sát:

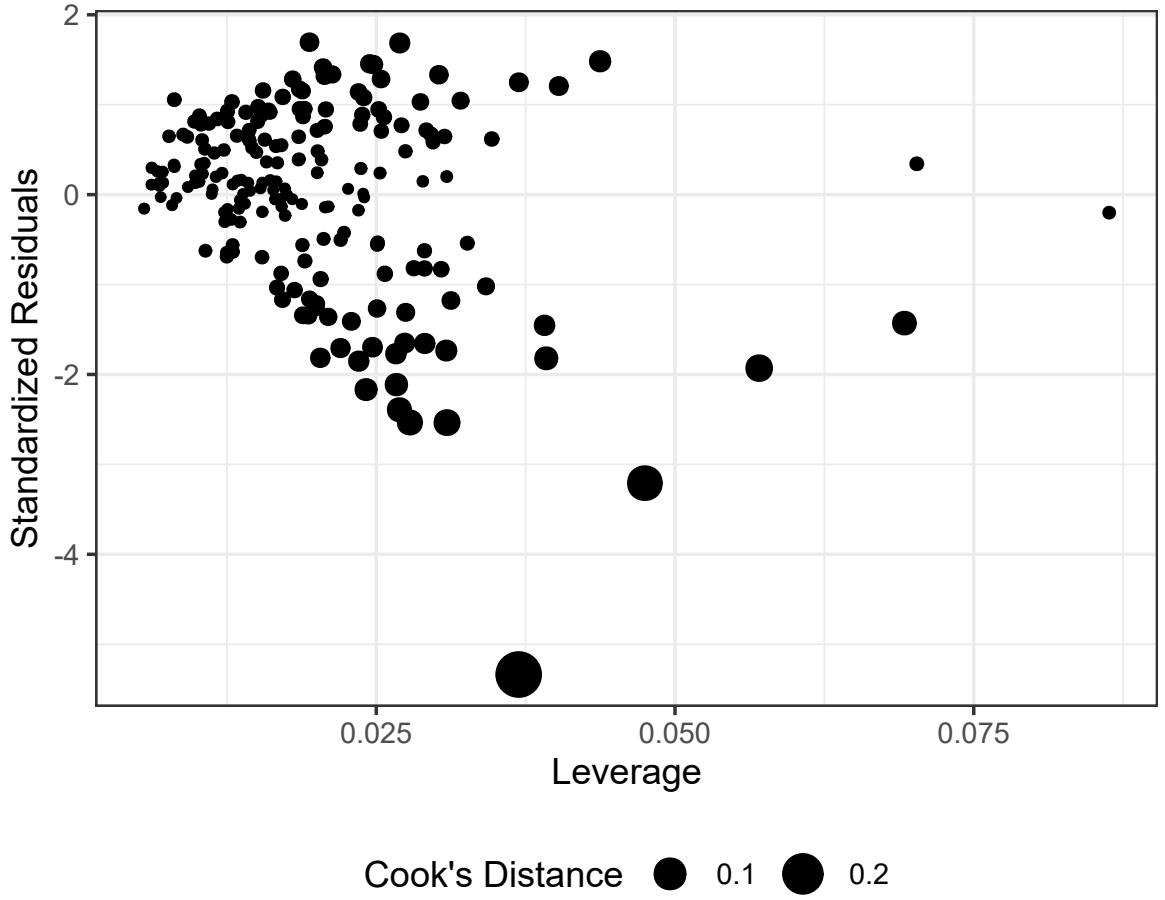
- điểm có kích cỡ nhỏ tương ứng khoảng cách Cook nhỏ;
- điểm có kích cỡ lớn tương ứng khoảng cách Cook lớn.

Phát hiện giá trị ngoại lai

Với mô hình

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

biểu đồ Residuals vs Leverage:



Multicollinearity - Đa cộng tuyến

Multicollinearity - Đa cộng tuyến

Multicollinearity - Đa cộng tuyến là hiện tượng xảy ra trong mô hình hồi quy tuyến khi 1 biến hồi quy có tương quan (*correlated*) với 1 hoặc nhiều biến hồi quy khác có mặt trong mô hình.

Multicollinearity thường xuất hiện khi:

- 1 biến hồi quy được đưa vào nhiều lần, do lỗi trong xử lý;
- k giả biến (dummy variable), thay vì $k-1$ giả biến, được tạo ra từ một biến nhân tố (biến định tính có k danh mục - categories);
- hai hoặc nhiều biến hồi quy có tương quan gần như hoàn hảo với nhau.

Multicollinearity - Đa cộng tuyến

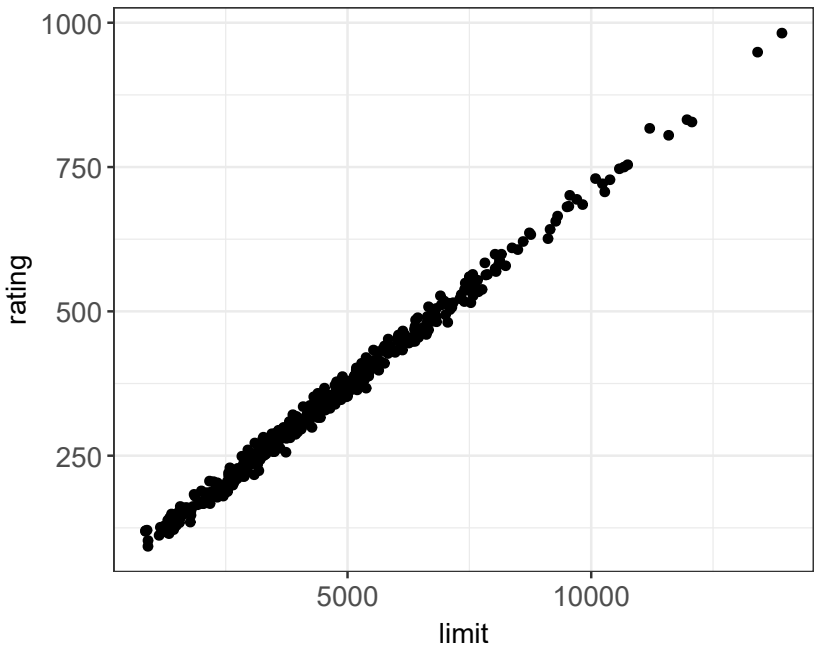
Ví dụ: Xét mô hình

$$\texttt{balance} = \beta_0 + \beta_1\texttt{rating} + \beta_2\texttt{limit} + \varepsilon,$$

trong đó,

- `balance`: số tiền cân bằng trong tài khoản;
- `rating`: xếp hạng tín dụng;
- `limit`: giới hạn tín dụng.

Biểu đồ phân tán của `rating` và `limit` cho thấy sự tương quan gần như hoàn hảo.



$\Rightarrow$ chắc chắn có hiện tượng multicollinearity trong mô hình.

Multicollinearity - Đa cộng tuyến

Tác động của multicollinearity

- làm giảm độ chính xác trong ước lượng hệ số của mô hình;
- làm sai số chuẩn của $\widehat{\beta}_j$ tăng lên;
- làm giảm độ mạnh (power) của kiểm định giả thuyết $H_0 : \beta_j = 0$ (tức là, khả phát hiện đúng hệ số là khác không, bị giảm đi).

Ví dụ: Xét 2 mô hình

$$\begin{aligned} \text{(MH1)} \quad &: \quad \texttt{balance} = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{limit} + \varepsilon, \\ \text{(MH2)} \quad &: \quad \texttt{balance} = \beta_0 + \beta_1 \texttt{rating} + \beta_2 \texttt{limit} + \varepsilon. \end{aligned}$$

Trong đó, (MH2) là mô hình có sự xuất hiện của multicollinearity (giữa `limit` và `rating`).

Multicollinearity - Đa cộng tuyến

Có kết quả như sau:

		Coefficient	Std. error	t-statistic	p-value
(MH1)	`Intercept`	-173.411	43.828	-3.957	< 0.0001
	`age`	-2.292	0.672	-3.407	0.0007
	`limit`	0.173	0.005	34.496	<0.0001
(MH2)	`Intercept`	-377.537	45.254	-8.343	<0.0001
	`rating`	2.202	0.952	2.312	0.0213
	`limit`	0.025	0.064	0.384	0.7012

$\Rightarrow$ sai số chuẩn của `limit` trong (MH2) tăng lên 12 lần so với trong (MH1).

Multicollinearity - Đa cộng tuyến

⇒ Cần phải giải quyết tình trạng đa cộng tuyến trong hồi quy.

⇒ Cần loại bỏ các biến cho đến khi tình trạng đa cộng tuyến biến mất.

Cách phát hiện đa cộng tuyến

- quan sát ma trận hệ số tương quan của các biến hồi quy trước khi thực hiện mô hình;

  ⇒ các giá trị lớn chỉ ra các cặp biến có sự tương quan cao ⇒ đa cộng tuyến có thể xuất hiện nếu sử dụng các cặp biến này;
- tính hệ số lạm phát phương sai (*variance inflation factor*) - VIF của từng biến hồi quy trong mô hình:

$$\text{VIF}_j = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

  trong đó, $R^2_{X_j|X_{-j}}$ là hệ số R^2 của mô hình hồi quy của biến X_j theo các biến hồi quy còn lại trong mô hình;

  ⇒ VIF_j nhỏ nhất là 1 ⇒ không có đa cộng tuyến;

  ⇒ VIF_j vượt quá 5 hoặc 10 ⇒ vấn đề lớn với đa cộng tuyến;

Multicollinearity - Đa cộng tuyến

Ví dụ: Xét mô hình

$$\texttt{balance} = \beta_0 + \beta_1\texttt{rating} + \beta_2\texttt{limit} + \beta_3\texttt{age} + \varepsilon$$

Khi đó, hệ số VIF được xác định như sau:

Biến	Mô hình	$R^2_{X_j \mid X_{-j}}$	VIF_j
`rating`	$\texttt{rating} = \beta_0 + \beta_1\texttt{limit} + \beta_2\texttt{age} + \varepsilon$	0.9938	160.6683
`limit`	$\texttt{limit} = \beta_0 + \beta_1\texttt{rating} + \beta_2\texttt{age} + \varepsilon$	0.9937	160.5929
`age`	$\texttt{age} = \beta_0 + \beta_1\texttt{limit} + \beta_2\texttt{rating} + \varepsilon$	0.0113	1.0114

$\Rightarrow$ hai biến `rating` và `limit` là nguồn của đa cộng tuyến.

$\Rightarrow$ cần loại bỏ 1 trong hai biến này khỏi mô hình.

Multicollinearity - Đa cộng tuyến

Loại bỏ `rating` $\Rightarrow$ mô hình mới

$$\texttt{balance} = \beta_0 + \beta_1\texttt{limit} + \beta_2\texttt{age} + \varepsilon$$

Khi đó, hệ số VIF được xác định như sau:

Biến	Mô hình	$R^2_{X_j \mid X_{-j}}$	VIF_j
`limit`	$\texttt{limit} = \beta_0 + \beta_1\texttt{age} + \varepsilon$	0.0102	1.0103
`age`	$\texttt{age} = \beta_0 + \beta_1\texttt{limit} + \varepsilon$	0.0102	1.0103

$\Rightarrow$ không còn đa cộng tuyến.

Loại bỏ `limit` $\Rightarrow$ mô hình mới

$$\texttt{balance} = \beta_0 + \beta_1\texttt{rating} + \beta_2\texttt{age} + \varepsilon$$

Khi đó, hệ số VIF được xác định như sau:

Biến	Mô hình	$R^2_{X_j \mid X_{-j}}$	VIF_j
`rating`	$\texttt{rating} = \beta_0 + \beta_1\texttt{age} + \varepsilon$	0.0106	1.0108
`age`	$\texttt{age} = \beta_0 + \beta_1\texttt{rating} + \varepsilon$	0.0106	1.0108

$\Rightarrow$ không còn đa cộng tuyến.

1 Chuẩn đoán mô hình

2 Lựa chọn mô hình

3 Mở rộng mô hình

4 Generalized Linear Models

Lựa chọn mô hình

Trong xây dựng mô hình hồi quy đa biến

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

về mặt kỹ thuật, ta có thể có

- p mô hình hồi quy tuyến tính đơn;
- C_p^2 mô hình hồi quy tuyến tính hai biến;
- C_p^3 mô hình hồi quy tuyến tính ba biến;
- C_p^q mô hình hồi quy tuyến tính q biến, $(q \leq p)$.

Khi này, một câu hỏi quan trọng nảy sinh:

"Mô hình nào là tốt nhất trong số các mô hình có thể có?"

$\hookrightarrow$ ta cần có một phương pháp để chọn ra mô hình tốt nhất.

Lựa chọn mô hình

Trong xây dựng mô hình hồi quy đa biến

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

về mặt kỹ thuật, ta có thể có

- p mô hình hồi quy tuyến tính đơn;
- C_p^2 mô hình hồi quy tuyến tính hai biến;
- C_p^3 mô hình hồi quy tuyến tính ba biến;
- C_p^q mô hình hồi quy tuyến tính q biến, $(q \leq p)$.

Khi này, một câu hỏi quan trọng nảy sinh:

"Mô hình nào là tốt nhất trong số các mô hình có thể có?"

$\hookrightarrow$ ta cần có một phương pháp để chọn ra mô hình tốt nhất.

Hai vấn đề kỹ thuật:

- thế nào là một mô hình tốt?
- phương pháp lựa chọn?

Mô hình tốt?

Trong phần "Đánh giá mô hình", ta đã biết rằng, một mô hình có

- RMSE nhỏ, hoặc
- RSE nhỏ, hoặc
- R^2 hoặc R_a^2 lớn, hoặc
- cross-validation error nhỏ,

sẽ là một mô hình tốt.

Tuy nhiên, khi ta thêm biến vào mô hình hồi quy

- RMSE và RSE sẽ luôn giảm (tối thiểu 0);
- R^2 luôn tăng (tối đa 1).

$\hookrightarrow$ không thực tế nếu dùng các chỉ số này để so sánh các mô hình có số lượng biến khác nhau.

Mô hình tốt?

Để so sánh các mô hình với số lượng biến khác nhau, ta dùng các chỉ số sau.

- AIC (Akaike's Information Criteria)[1] cho mô hình hồi quy tuyến tính:

$$\mathrm{AIC} = 2p + n\log\left(\frac{RSS}{n}\right).$$

  AIC được xây dựng dựa trên lý thuyết likelihood và giả định phân phối chuẩn.

- BIC (Bayesian Information Criteria) cho mô hình hồi quy tuyến tính:

$$\mathrm{BIC} = p\log(n) + n\log\left(\frac{RSS}{n}\right).$$

  BIC được xây dựng dựa trên lý thuyết likelihood, thống kê Bayes và giả định phân phối chuẩn.

- Mallow's C_p cho mô hình hồi quy tuyến tính

$$C_p = \frac{1}{n}\left(RSS + 2p\widehat{\sigma}^2\right).$$

- Adjusted R-squares R_a^2.

[1] do nhà thống kê học Hirotugu Akaike, người Nhật bản, đề xuất vào thập niên 70

Mô hình tốt?

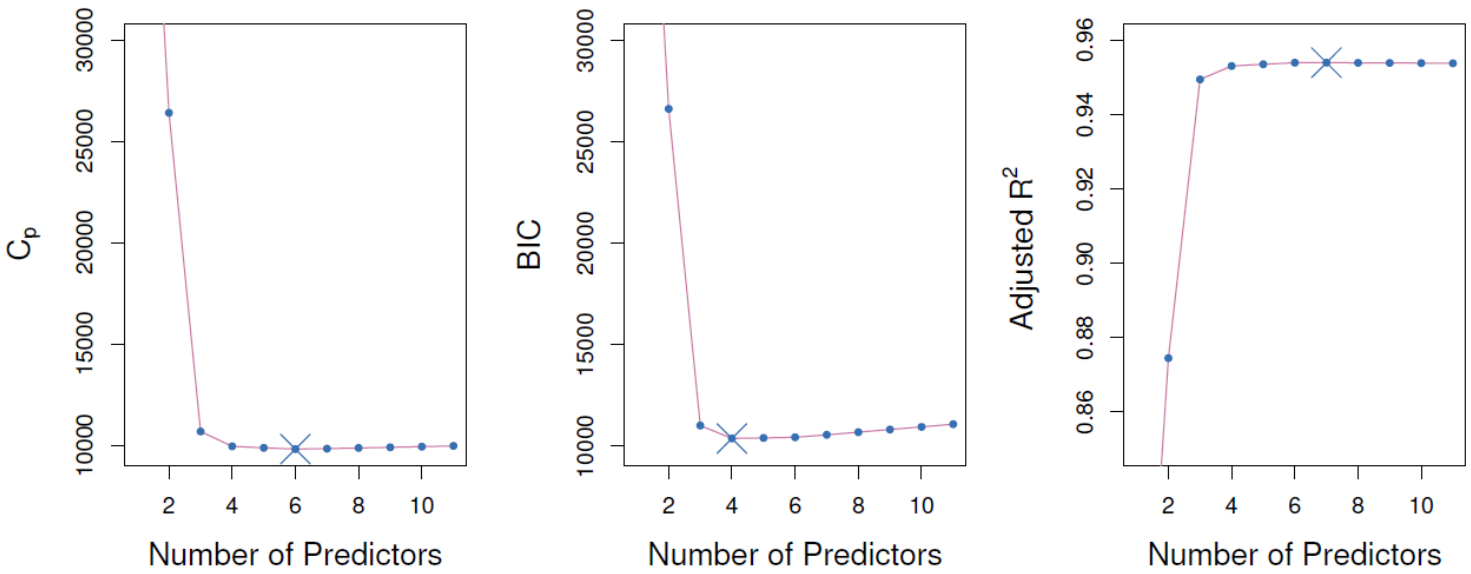
Nhận xét:

- AIC, BIC và C_p sẽ có giá trị nhỏ tại các mô hình có sai số kiểm tra nhỏ.

  $\hookrightarrow$ ta cần tìm mô hình có giá trị AIC, BIC hoặc C_p nhỏ (tức là sai số kiểm tra nhỏ).
- Trong khuôn khổ mô hình hồi quy tuyến tính, AIC và BIC chỉ có nghĩa nếu giả định về phân phối chuẩn là hợp lý.
- Adjusted R-squares R_a^2 sẽ đạt giá trị lớn ở các mô hình có sai số kiểm tra nhỏ.

  $\hookrightarrow$ ta cần tìm mô hình có giá trị R_a^2 lớn.
- AIC sẽ tỷ lệ với C_p trong trường hợp giả định phân phối chuẩn là đúng.
- BIC sẽ cho kết quả mô hình đơn giản hơn (số biến nhỏ hơn) so với mô hình được tìm thấy bởi C_p.
- Adjusted R-squares R_a^2 thường ít được sử dụng trong thực tế hơn các chỉ số AIC, BIC và C_p.

Mô hình tốt?



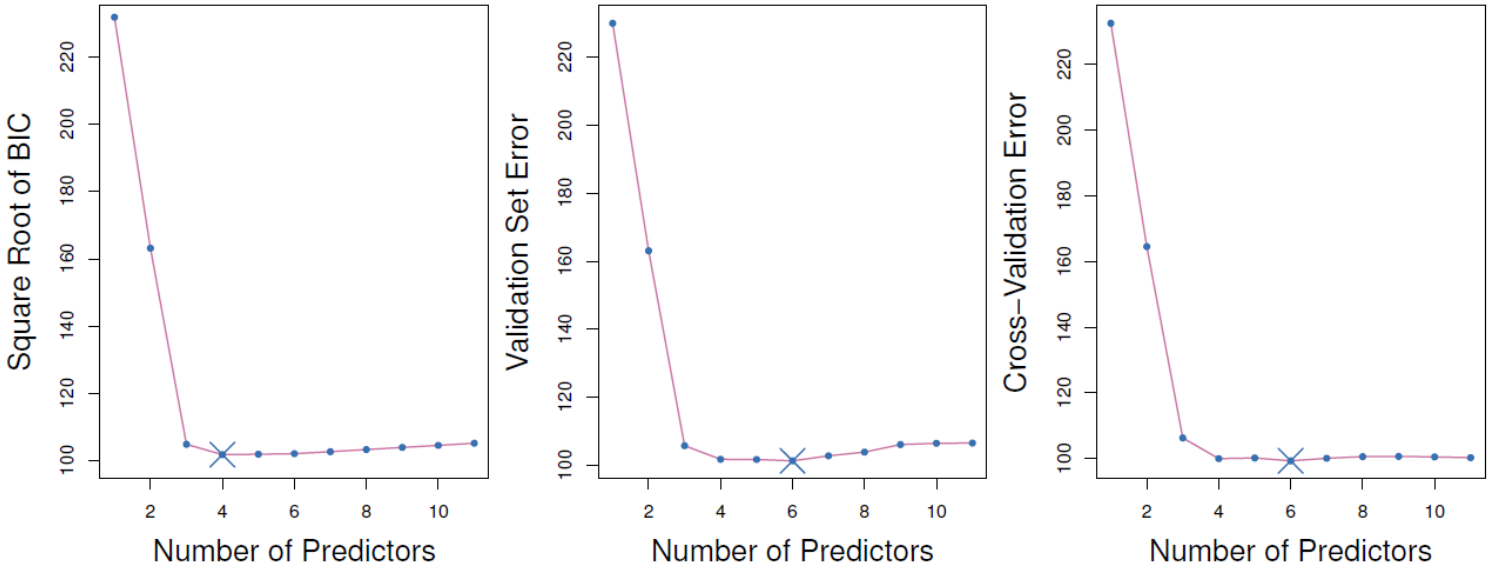
Hình trên, minh họa giá trị của C_p, BIC và R_a^2 trong việc lựa chọn mô hình hồi quy với số lượng biến hồi quy thay đổi (từ 1 tới 11).

- Tiêu chuẩn C_p cho kết quả là mô hình với 6 biến hồi quy.
- Tiêu chuẩn BIC cho kết quả là mô hình với 4 biến hồi quy.
- Tiêu chuẩn R_a^2 cho kết quả là mô hình với 7 biến hồi quy.

Quy tắc: Khi hiệu suất, sai số của các mô hình là tương đương nhau, thì mô hình đơn giản hơn là mô hình được ưu tiên.

Mô hình tốt?

Ngoài các tiêu chí trên, ta có thể sử dụng validation set hoặc cross-validation để tìm mô hình hồi quy với số biến q ($q < p$), sao cho sai số kiểm tra là nhỏ nhất.



Lựa chọn mô hình con tốt nhất

Xét mô hình hồi quy đa biến

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

ta cần áp dụng các chỉ số AIC, BIC, C_p, R_a^2, cross-validation cho:

- p mô hình hồi quy tuyến tính đơn;
- C_p^2 mô hình hồi quy tuyến tính hai biến;
- C_p^3 mô hình hồi quy tuyến tính ba biến;
- C_p^q mô hình hồi quy tuyến tính q biến, $(q \leq p)$.

Tổng quát, ta cần áp dụng các chỉ số đánh giá cho 2^p mô hình hồi quy có thể.

Việc này dẫn tới việc tính toán khá nặng.

Ví dụ: với $p = 200$ ta cần thực hiện $2^{20} = 1,048,576$ mô hình!

$\hookrightarrow$ ta sử dụng cách tiếp cận khác, có tên là hồi quy từng bước - stepwise regression.

Stepwise regression

Hồi quy từng bước - Stepwise regression là một phương pháp được dùng để tìm mô hình con tốt nhất của mô hình chứa p biến hồi quy.

Trong mỗi một bước, hồi quy từng bước sẽ thêm hoặc bớt một biến hồi quy vào trong mô hình tốt nhất hiện tại, và tìm ra mô hình tốt nhất mới.

Có hai cách tiếp cận cơ bản:

- hồi quy tiến từng bước - forward stepwise regression
- hồi quy lùi từng bước - backward stepwise regression

trong cả hai cách tiếp cận, số mô hình cần đánh giá là $1 + p(p+1)/2$.

Ví dụ: với $p = 20$, số mô hình cần đánh giá là 211 mô hình.

Stepwise regression

Hồi quy tiến từng bước - Forward stepwise regression

1. Bắt đầu với mô hình $\mathcal{M}_0$, chỉ chứa hệ số chặn: $Y = \beta_0 + \varepsilon$
2. Với $k = 0, 1, \ldots, p-1$:

   (a) Xem xét tất cả $p-k$ mô hình làm tăng các yếu tố dự đoán trong $\mathcal{M}_k$ bằng một yếu tố dự đoán bổ sung.

   (b) Chọn mô hình tốt nhất trong số $p-k$ mô hình này và gọi nó là $\mathcal{M}_{k+1}$. Ở đây mô hình "tốt nhất" được xác định là mô hình có RSS nhỏ nhất hoặc R^2 cao nhất.
3. Chọn một mô hình tốt nhất trong số $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ bằng cách sử dụng C_p, AIC, BIC, R_a^2 hoặc cross-validation.

Stepwise regression

Hồi quy lùi từng bước - Backward stepwise regression

1. Bắt đầu với mô hình $\mathcal{M}_p$, chứa tất cả các biến hồi quy:

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

2. Với $k = p, p-1, \ldots, 1$:

   (a) Xem xét tất cả k mô hình chứa tất cả ngoại trừ một trong các yếu tố dự đoán trong $\mathcal{M}_k$, để có tổng số $k-1$ yếu tố dự đoán.

   (b) Chọn mô hình tốt nhất trong số k mô hình này và gọi nó là $\mathcal{M}_{k-1}$. Ở đây mô hình "tốt nhất" được xác định là mô hình có RSS nhỏ nhất hoặc R^2 cao nhất.

3. Chọn một mô hình tốt nhất trong số $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ bằng cách sử dụng C_p, AIC, BIC, R_a^2 hoặc cross-validation.

Stepwise regression

- Các phương pháp hồi quy tiến từng bước và lùi từng bước tốt nhất thường cho các mô hình tương tự nhưng không giống nhau.
- Để cho thuận tiện cho việc áp dụng và diễn giải, ta có thể sử dụng phương pháp hỗn hợp (kết hợp cả tiến và lùi).

Hồi quy từng bước hỗn hợp - Hybrid stepwise regression

1. Bắt đầu với mô hình $\mathcal{M}_0$, chỉ chứa hệ số chặn: $Y = \beta_0 + \varepsilon$
2. Với $k = 0, 1, \ldots, p-1$:

   (a) Xem xét tất cả $p-k$ mô hình làm tăng các yếu tố dự đoán trong $\mathcal{M}_k$ bằng một yếu tố dự đoán bổ sung.

   (b) Loại bỏ bất kỳ một biến dự đoán (hồi quy) trong mô hình $\mathcal{M}_k$.

   (c) Chọn mô hình tốt nhất trong số p mô hình này và gọi nó là $\mathcal{M}_{k+1}$. Ở đây mô hình "tốt nhất" được xác định là mô hình có RSS nhỏ nhất hoặc R^2 cao nhất.
3. Chọn một mô hình tốt nhất trong số $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ bằng cách sử dụng C_p, AIC, BIC, R_a^2 hoặc cross-validation.

Phương pháp co hệ số - Shrinkage methods

Xét mô hình hồi quy:

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

Một cách tiếp cận khác cho việc lựa chọn biến hồi quy, là ước lượng hệ số của mô hình với một ràng buộc sao cho các hệ số không cần thiết sẽ "co lại" về 0.

- Ridge regression

$$\widehat{\beta}_R = \arg\min \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ji} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Lasso regression

$$\widehat{\beta}_L = \arg\min \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ji} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Phương pháp co hệ số - Shrinkage methods

- SCAD (Smoothly Clipped Absolute Deviation)

$$\widehat{\beta}_S = \arg\min \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ji} \right)^2 + n \sum_{j=1}^{p} p_\lambda(|\beta_j|)$$

với

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 \mathrm{I}(|\theta| < \lambda).$$

Tham số λ được gọi là turning parameter.

- $\lambda = 0$, ước lượng tương đương với ước lượng OLS.
- $\lambda \to \infty$, các hệ số ước lượng sẽ xấp xỉ 0.

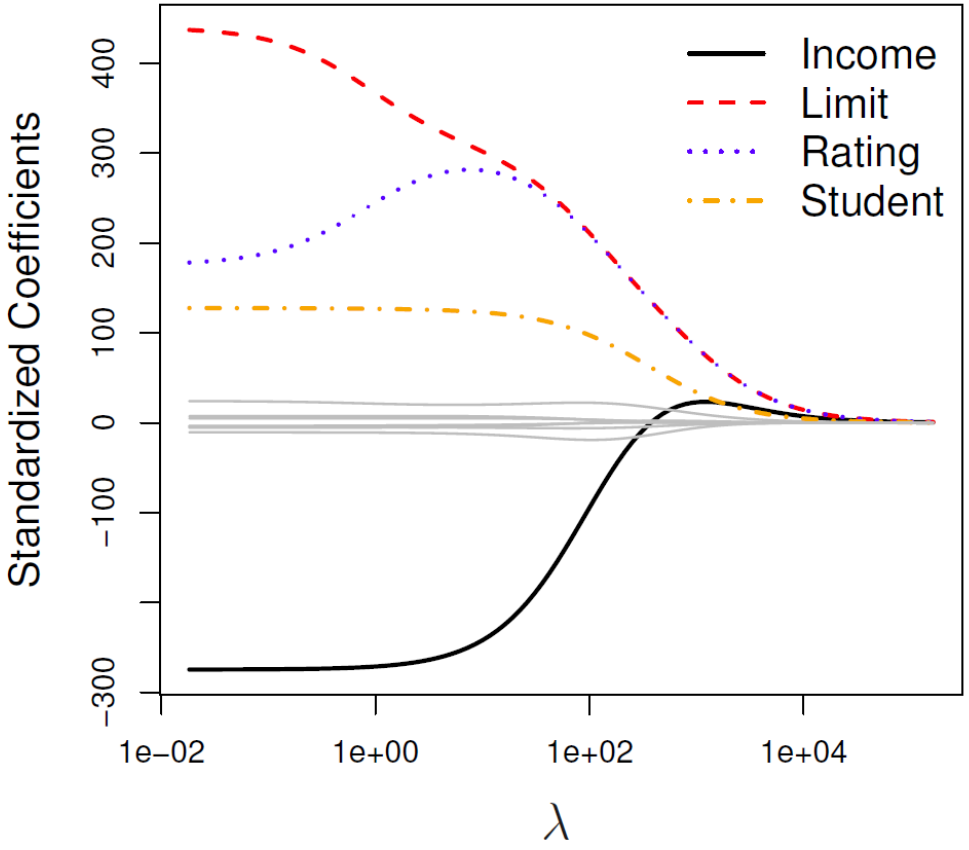
$\hookrightarrow$ một trong những vấn đề ứng dụng là xác định được λ hợp lý.

$\hookrightarrow$ cross-validation được áp dụng để tìm λ. Ví dụ:

$$\mathrm{CV}_{(k)}(\lambda) = \frac{1}{k} \sum_{j=1}^{k} \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i(\lambda) \right)^2,$$

tìm λ sao cho cực tiểu hóa $\mathrm{CV}_{(k)}(\lambda)$.

Phương pháp co hệ số - Shrinkage methods



Sự thay đổi của hệ số ước lượng tương ứng với giá trị λ của ridge regression.

1 Chuẩn đoán mô hình

2 Lựa chọn mô hình

3 Mở rộng mô hình

4 Generalized Linear Models

Weighted Regression

Hồi quy có trọng số - Weighted regression, là một biến thể của mô hình hồi quy tuyến tính:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{V}^{1/2}\varepsilon,$$

trong đó, $\mathbf{V} = \mathbf{W}\mathbf{W}^\top$, với $\mathbf{W} = (W_1, \ldots, W_n)$ vectơ trọng số của n quan sát.

Ước lượng trọng số OLS (weighted OLS):

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

Hồi quy có trọng số được dùng trong các trường hợp:

- Trọng số phương sai nghịch đảo khi các quan sát khác nhau được đo với độ chính xác khác nhau.
- Phân tích dữ liệu ở dạng tổng hợp sao cho biến trọng số mã hóa số lượng quan sát ban đầu mà mỗi hàng trong dữ liệu tổng hợp đại diện.

Đặc biệt, khi mô hình hồi quy thông thường bị heteroskedasticity (không đồng nhất phương sai) thì hồi quy có trọng số là một giải pháp thay thế.

Hồi quy đa thức

Trong các phần trước, ta đã giới thiệu mô hình hồi quy tuyến tính:

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

Thuật ngữ "tuyến tính" ám chỉ sự tuyến tính theo hệ số β_j.

Do đó, X_j có thể được sử dụng dưới dạng một biến đổi: bình phương, log, căn bậc hai.

Đặc biệt, nếu ta sử dụng một đa thức bậc k của X_j, thì khi đó, ta có mô hình hồi quy đa thức (polynomial regression).

Ví dụ: với $k = 2$,

$$Y = \beta_0 + \alpha_1 X_1 + \alpha_2 X_1^2 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon.$$

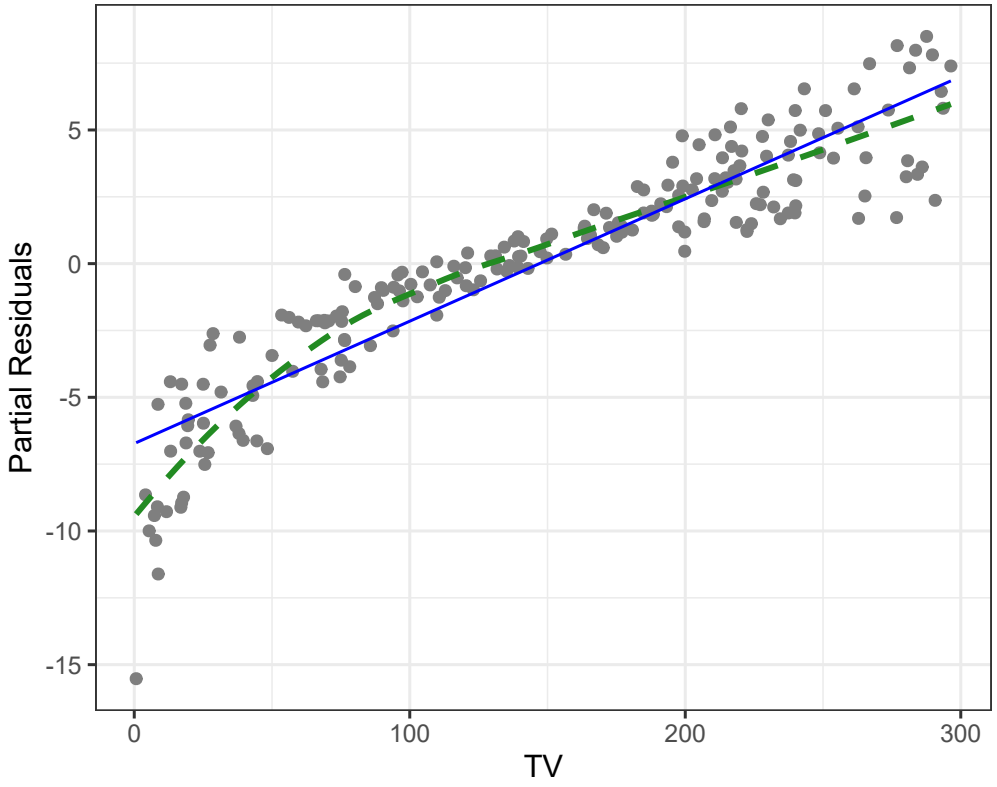
Hồi quy đa thức thường được dùng trong các trường hợp mà mối quan hệ giữa Y và X_j là phi tuyến.

Hồi quy đa thức

Ví dụ: Trong mô hình tuyến tính:

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

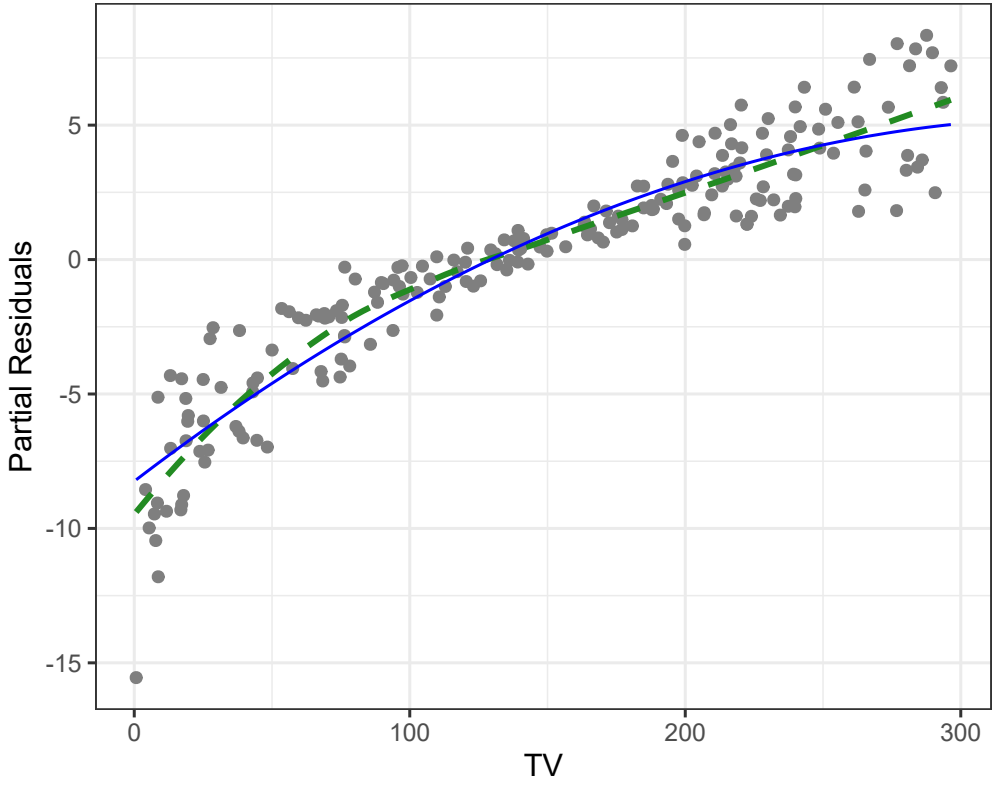
thông qua biểu đồ thặng dư từng phần cho biến `tv`, ước lượng tuyến tính là không phù hợp với dữ liệu



Hồi quy đa thức

Thay vì sử dụng `tv`, ta sẽ áp dụng dạng đa thức bậc k cho `tv`. Ta chọn $k = 2$

$$\texttt{sales} = \beta_0 + \alpha_1 \texttt{tv} + \alpha_2 \texttt{tv}^2 + \beta_2 \texttt{radio} + \beta_3 \texttt{newspaper} + \varepsilon.$$

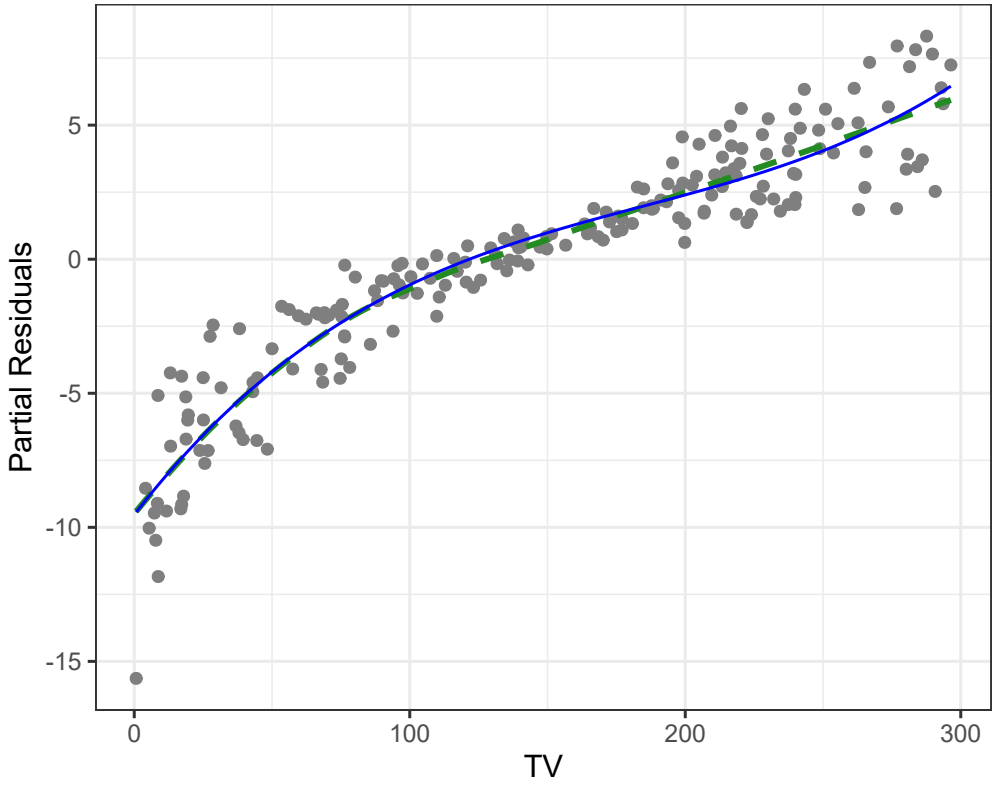


Biểu đồ thặng dư từng phần cho thấy mô hình đã được cải thiện đáng kể.

Hồi quy đa thức

Ta chọn $k = 3$

$$\texttt{sales} = \beta_0 + \alpha_1 \texttt{tv} + \alpha_2 \texttt{tv}^2 + \alpha_3 \texttt{tv}^3 + \beta_2 \texttt{radio} + \beta_3 \texttt{newspaper} + \varepsilon.$$



Biểu đồ thặng dư từng phần cho thấy mô hình đã được cải thiện đáng kể hơn $k = 2$.

Hồi quy đa thức

Như vậy, để áp dụng hồi quy đa thức:

- Xây dựng mô hình hồi quy tuyến tính (không đa thức).
- Kiểm tra mối quan hệ giữa Y và X_j thông qua biểu đồ thặng dư từng phần.
- Nếu mối quan hệ giữa Y và X_j (cố định j) không phải là tuyến tính thì ta có sử dụng dạng đa thức bậc k của X_j để cải thiện mô hình.
- Để tìm k tối ưu, ta có thể áp dụng thuật toán cross-validation.
- Điểm hạn chế của hồi quy đa thức, đó là ta không thể làm một cách tự động.
- Mô hình có thể trở nên cồng kềnh, với k lớn.

Hồi quy đa thức từng phần

Trong phần trước, ta đã xây dựng mô hình hồi quy đa thức trên toàn miền giá trị của X.

Tuy nhiên, trong thực tế, ta có thể xây dựng và ước lượng các dạng đa thức khác nhau trên từng đoạn giá trị của X:

$$Y = \begin{cases} \beta_{01} + \beta_{11}X + \beta_{21}X^2 + \ldots + \beta_{k1}X^k + \varepsilon, & \text{nếu } X \leq c \\ \beta_{02} + \beta_{12}X + \beta_{22}X^2 + \ldots + \beta_{k2}X^k + \varepsilon, & \text{nếu } X > c \end{cases}$$

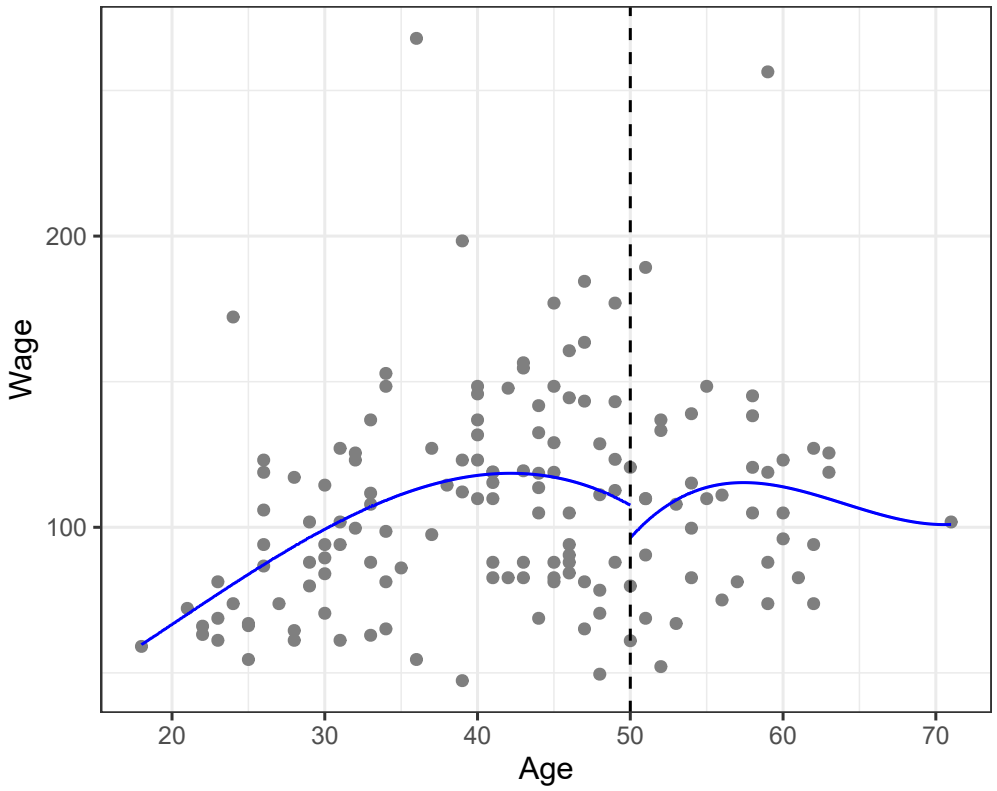
với c là một điểm nút (knot) được xác định trước.

Trên thực tế, ta có thể có nhiều hơn một điểm nút, tùy thuộc vào sự thay đổi của dữ liệu theo vùng giá trị của X.

Hồi quy đa thức từng phần

Ví dụ: Xét mô hình dự đoán tiền công của công nhân làm việc ở khu vực Mid-Atlantic, theo độ tuổi.

$$\texttt{wage} = \begin{cases} \beta_{01} + \beta_{11}\text{age} + \beta_{21}\text{age}^2 + \beta_{31}\text{age}^3 + \varepsilon, & \text{nếu age} \leq 50 \\ \beta_{02} + \beta_{12}\text{age} + \beta_{22}\text{age}^2 + \beta_{32}\text{age}^3 + \varepsilon, & \text{nếu age} > 50 \end{cases}$$



Hồi quy đa thức từng phần

Một số nhận xét:

- Có thể tương thích tốt hơn với sự thay đổi của Y trên các miền của X.
- Tuy nhiên, đường ước lượng là không liên tục tại điểm nút.
- Phải xác định bằng tay (manually) các nút và khoảng của X tương ứng với sự thay đổi dáng điệu của dữ liệu.
- Phải xác định dạng cụ thể của mô hình trên mỗi vùng. Không nhất thiết là phải dạng đa thức.
- Khi có nhiều vùng, mô hình trở nên cồng kềnh.

Hàm cơ sở - Basic functions

Xét mô hình hồi quy đa thức

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k + \varepsilon,$$

Mô hình này là dạng đặc biệt của mô hình hồi quy với các hàm cơ sở

$$Y = \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \ldots + \beta_k b_k(X) + \varepsilon,$$

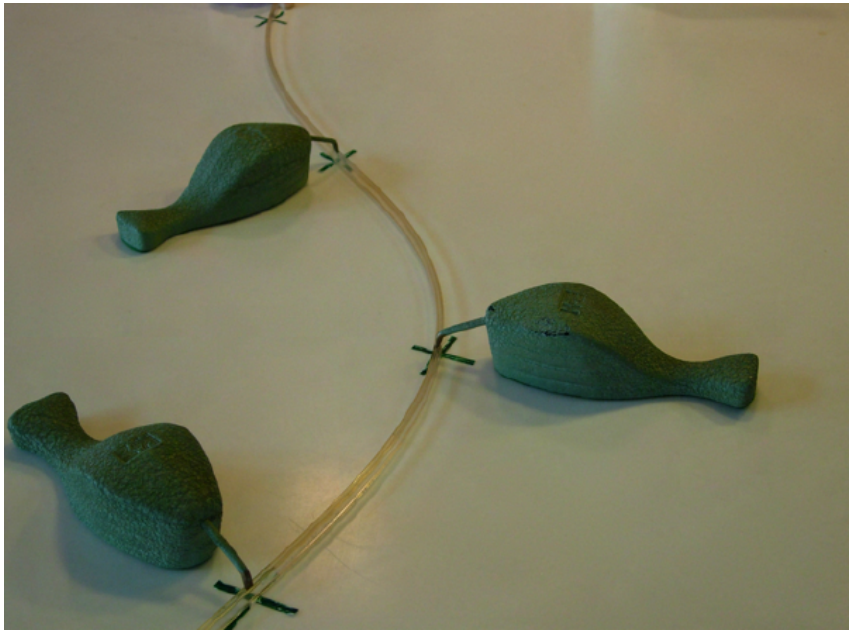
với $b_j(X)$ là các hàm đã được cố định, biết trước của X.

Trên thực tế, ta có nhiều định nghĩa và các xây dựng hàm cơ cở b:

- $b_j(x) = x^j$
- $b_j(x) = \mathrm{I}(c_j \leq x_j \leq c_{j+1})$
- $b(x) = \log(x)$
- wavelets
- Fourier series

Hồi quy splines

Splines là những dải linh hoạt (gỗ hoặc cao su) được sử dụng để vẽ các đường cong.



Hồi quy splines

Về mặt toán học, **Splines** là các hàm đa thức từng phần, được kết nối trơn tru.

Một cách chính thức hơn, một đường spline có bậc r với k điểm nút (knots) $\xi_1, \xi_2, \ldots, \xi_k$ là hàm bất kỳ:

$$s(x) = \sum_{j=0}^{r+k} \alpha_j b_j(x),$$

với $b_j(x)$ là các hàm cơ sở của x.

- $r = 3$, ta có cubic spline
- nếu ta thêm điều kiện rằng $s(x)$ là hàm tuyến tính ngoài khoảng giá trị được quan sát của X, thì ta có natural spline.

Hồi quy splines

Ta có các cách chọn hàm cơ sở cơ bản:

- hàm cơ sở mũ chặt cụt (truncated power basis):

$$s(x) = \sum_{j=0}^{r} \alpha_j x^j + \sum_{j=1}^{k} \delta_j (x - \xi_j)_+^r,$$

với

$$(x - \xi_j)_+^r = \begin{cases} (x - \xi_j)^r & \text{nếu } x > \xi_j, \\ 0 & \text{nếu } x \leq \xi_j, \end{cases}$$

với $r = 3$ và sử dụng natural spline, ta có natural cubic spline;

Hồi quy splines

- hàm cơ sở B-spline:

$$s(x) = \sum_{j=0}^{r+k} \alpha_j B_{j,r}(x),$$

với $B_{j,r}(x)$ là hàm cơ sở xác định bởi:

(+) $r = 1$, các hàm cơ sở $B_{j,1}(x)$, $j = 0, \ldots, r, r+1, \ldots, k$:

$$B_{j,1}(x) = \mathrm{I}(\xi_j \leq x < \xi_{j+1})$$

(+) $r > 1$,

$$B_{j,r}(x) = \frac{x - \xi_j}{\xi_{j+r} - \xi_j} B_{j,r-1}(x) + \frac{\xi_{j+r+1} - x}{\xi_{j+r+1} - \xi_{j+1}} B_{j+1,r-1}(x),$$

với $0/0 = 0$ và $\mathrm{Inf}/0 = 0$, và các điểm nút cố định

$$a = \xi_0 < \xi_1 < \xi_2 < \ldots < \xi_k < b = \xi_{k+1}$$

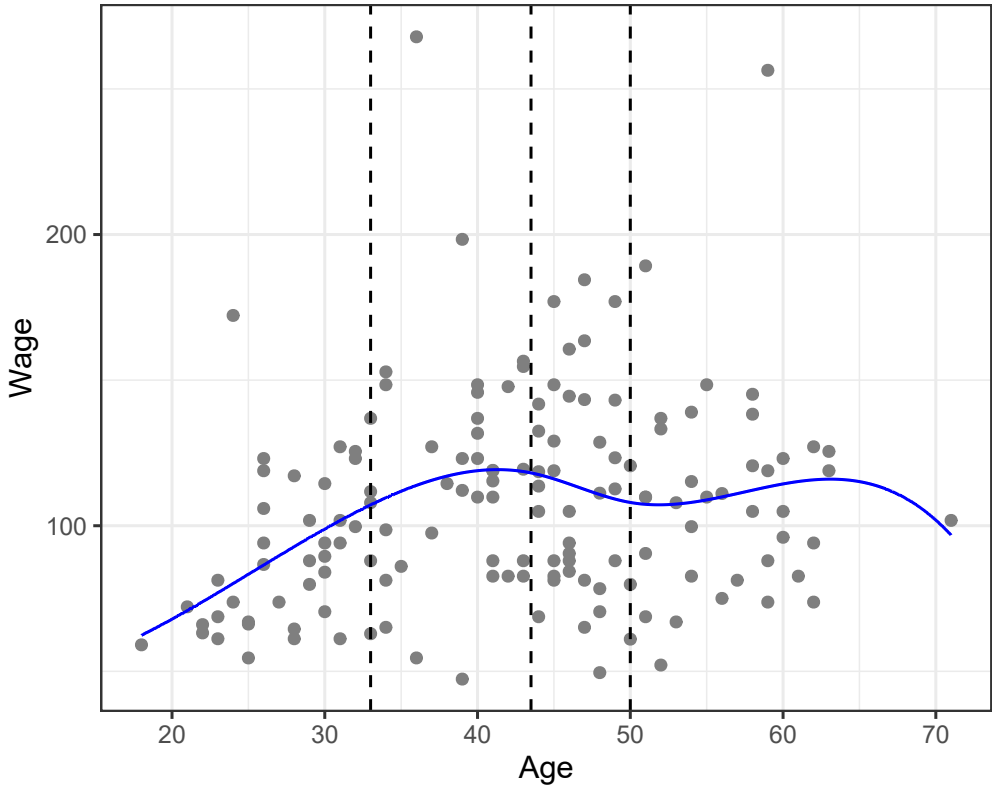
Phương pháp B-spline thường được dùng phổ biến hơn natural spline.

Một cách mặc định, $r = 3$ thường được sử dụng, bởi vì hàm ước lượng sẽ "đủ" trơn khi nhìn bằng mắt.

Hồi quy splines

Ta áp dụng B-spline cho mô hình dựa đoán tiền công của công nhân dựa vào tuổi của họ:

- $r = 3$
- $k = 3$
- ξ_j là các điểm tương ứng với phân vị mức 0.25, 0.5 và 0.75 của age.

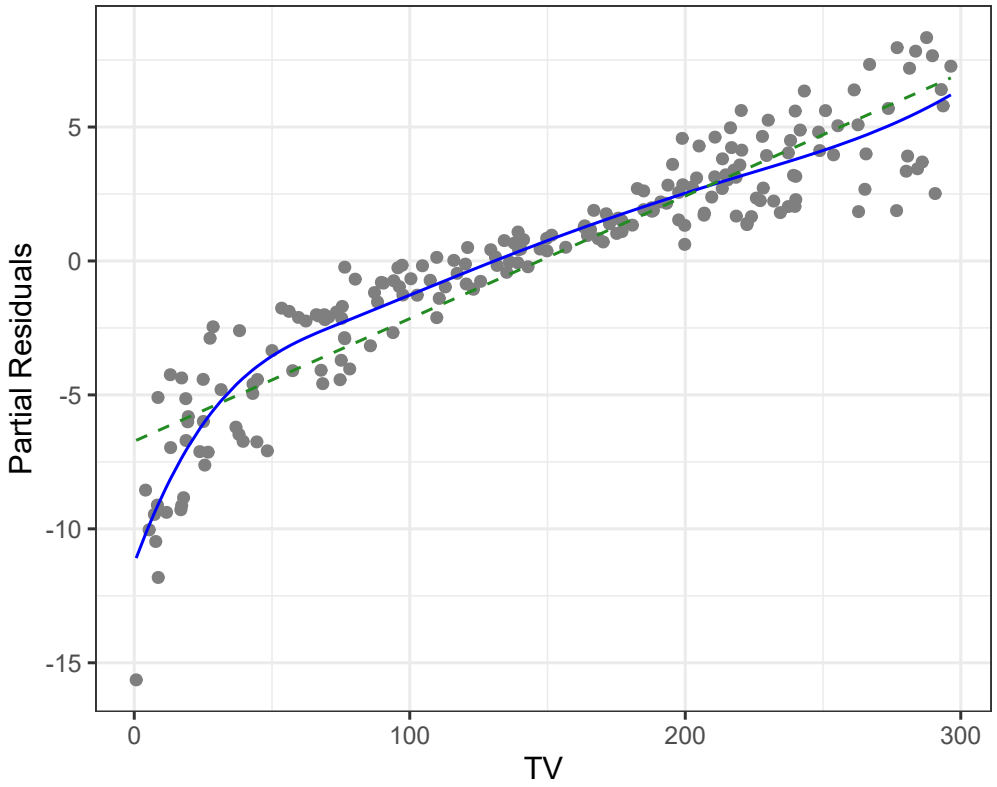


Hồi quy splines

Ví dụ: xét lại mô hình dự đoán doanh số bán hàng, với thành phần B-spline:

$$\texttt{sales} = \beta_0 + \mathrm{bs}(\texttt{tv}, 3) + \beta_2\texttt{radio} + \beta_3\texttt{newspaper} + \varepsilon.$$

với $\mathrm{bs}(\texttt{tv}, 3)$ là thành phần được xác định bởi B-spline bậc 3.



Đường màu xanh dương tương ứng với ước lượng dựa trên B-spline bậc 3, đường nét đứt là đường ước lượng tuyến tính (mô hình ban đầu).

Hồi quy splines

So sánh với các mô hình tiên đoán `sale` đã ước lượng trước đây:

Mô hình	RSE	AIC	R_a^2
Tuyến tính	1.686	782.362	0.896
Đa thức bậc 2 cho `tv`	1.521	742.288	0.915
Đa thức bậc 3 cho `tv`	1.431	718.823	0.925
B-spline bậc 3 cho `tv`	1.399	712.626	0.928

- Mô hình với B-spline bậc 3 là mô hình tốt nhất.
- So với mô hình đa thức bậc 3, thì mô hình với B-spline bậc 3 là linh hoạt hơn, do ta không cần phải xác định số bậc của đa thức.

Trong áp dụng, ta cần quan tâm:

- vị trí đặt có điểm nút: có thể đặt theo điểm phân vị của dữ liệu;
- số lượng điểm nút: có thể sử dụng cross-validation để xác định;
- bậc của B-spline: có thể sử dụng cross-validation để xác định.

Chú ý: Ta có thể áp dụng B-spline cho nhiều hơn một biến hồi quy trong mô hình.

Generalized Additive Models

Một cách tự nhiên để mở rộng mô hình hồi quy tuyến tính

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

để cho phép biểu diễn linh hoạt quan hệ

- tuyến tính
- phi tuyến tính

giữa Y và X_j, là thay thế $\beta_j X_j$ bằng một làm trơn $f_j(X_j)$ chưa biết của X_j:

$$Y = \beta_0 + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p) + \varepsilon.$$

Mô hình này được gọi là Mô hình cộng tính tổng quát (Generalized Additive Models), gọi tắt là GAM.

Ta gọi là "cộng tính" vì ta ước lượng riêng biệt các f_j cho mỗi X_j, sau đó, cộng chúng lại với nhau, để ước đoán biến phản hồi.

Generalized Additive Models

Trong áp dụng, ta có thể:

- sử dụng GAM cho toàn bộ biến trong mô hình tuyến tính;
- sử dụng GAM cho một hoặc một vài biến trong mô hình tuyến tính.

Đối với cách sử dụng thứ 2, sẽ hữu dụng khi ta phát hiện ra mối quan hệ giữa Y và một X_j là không tuyến tính (thông qua biểu đồ của thặng dư từng phần).

Cách sử dụng thứ 1, là cách ước lượng mô hình tự động, hữu ích cho trường hợp

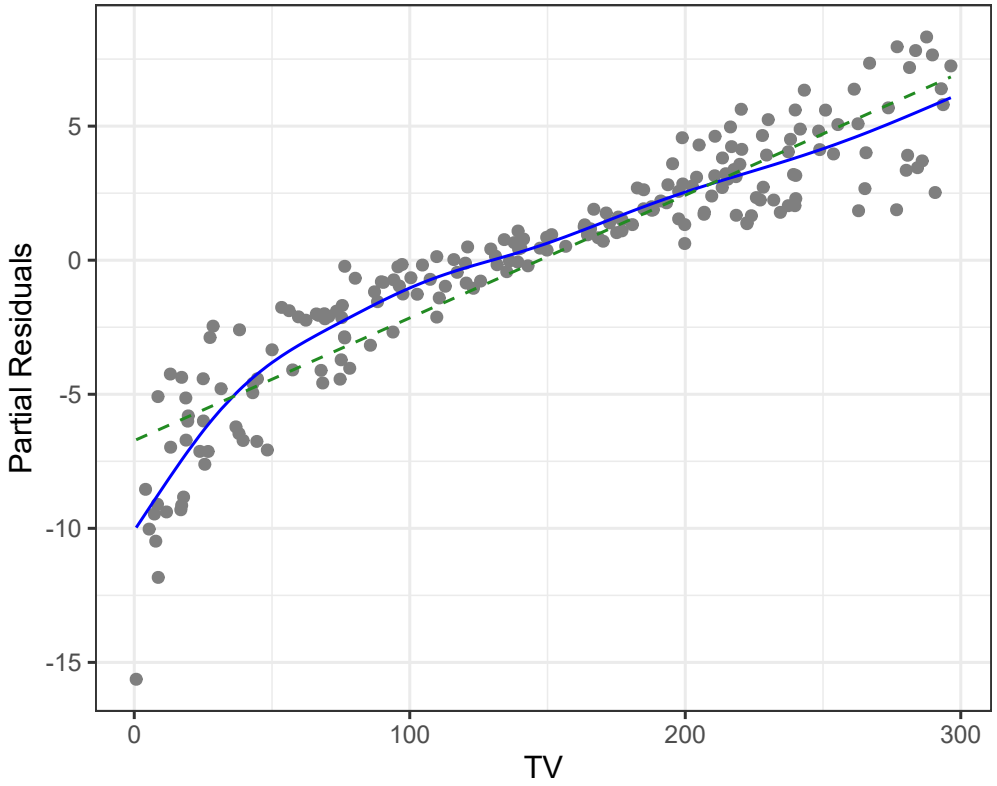
- có nhiều biến hồi quy,
- ta không thể kiểm tra quan hệ tuyến tính cho tất cả các biến.

Nếu mối quan hệ giữa Y và X_j thực sự là tuyến tính, f_j sẽ có dạng tuyến tính, và ngược lại.

Generalized Additive Models

Ví dụ: xét lại mô hình dự đoán doanh số bán hàng, với thành phần $f(\texttt{tv})$:

$$\texttt{sales} = \beta_0 + f(\texttt{tv}) + \beta_2\texttt{radio} + \beta_3\texttt{newspaper} + \varepsilon.$$

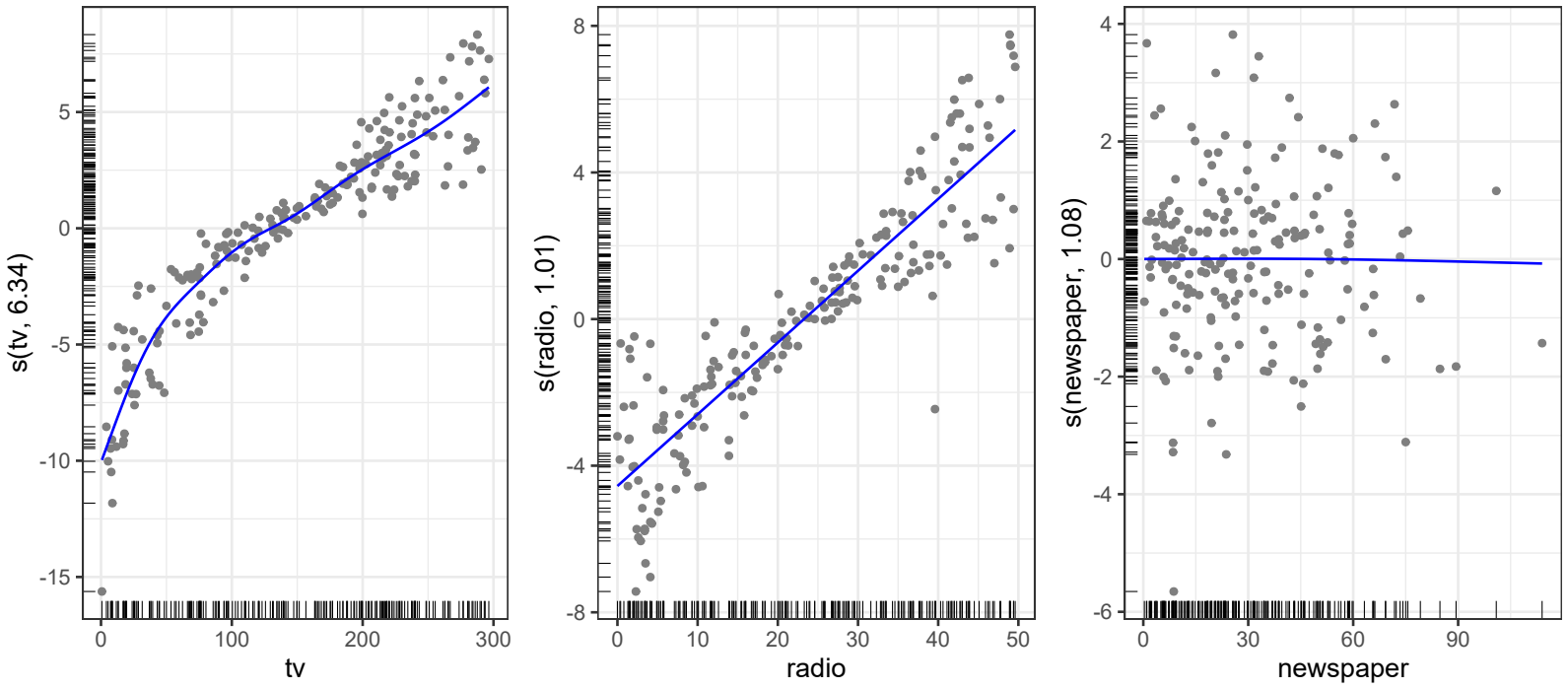


Đường màu xanh dương tương ứng với ước lượng dựa trên $f(\texttt{tv})$, đường nét đứt là đường ước lượng tuyến tính (mô hình ban đầu).

Generalized Additive Models

Ví dụ: xét mô hình GAM dự đoán doanh số bán hàng:

$$\texttt{sales} = \beta_0 + f_1(\texttt{tv}) + f_2(\texttt{radio}) + f_3(\texttt{newspaper}) + \varepsilon.$$



Chú ý: ta có thể thêm thành phần tương tác vào mô hình GAM, ví dụ

$$\texttt{sales} = \beta_0 + f_1(\texttt{tv}) + f_2(\texttt{radio}) + f_3(\texttt{tv}, \texttt{radio}) + \varepsilon.$$

1 Chuẩn đoán mô hình

2 Lựa chọn mô hình

3 Mở rộng mô hình

4 Generalized Linear Models

Vấn đề của Linear Models

Mô hình tuyến tính (Linear Models) không hiệu quả trong các trường hợp:

- biến phản hồi không liên tục: nhị thức, số đếm, dạng định tính;
- biến phản hồi có phân phối bất đối xứng: lệch phải, có nhiều giá trị lớn (extreme values);
- phương sai của biến phản hồi trong mô hình không phải hằng số (đồng nhất phương sai - homoskedasticity).

Mô hình tuyến tính tổng quát (**Generalized Linear Models** - GLM) là sự tổng quát mô hình tuyến tính để tương thích với các dạng dữ liệu:

- biến phản hồi có dạng: nhị thức, số đếm, dạng định tính;
- biến phản hồi có phân phối bất đối xứng: lệch phải, có nhiều giá trị lớn (extreme values);
- phương sai của biến phản hồi trong mô hình thay đổi theo giá trị trung bình.

Cấu trúc của Generalized Linear Models

GLM có ba thành phần chính:

- **biến phản hồi**: thành phần này chỉ định biến phản hồi Y và phân phối xác suất của nó. Bộ n quan sát $(y_1, \ldots, y_n)$ của Y được coi là quan sát thực tế của các biến ngẫu nhiên độc lập.
- **biến giải thích**: thành phần này chỉ định p biến giải thích cho một *linear predictor*, $\beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}$, trong đó X_{ji} là giá trị của biến giải thích j cho quan sát thứ i.
- **hàm liên kết**: thành phần này là một hàm g được áp dụng cho kỳ vọng có điều kiện $\mu_i = \mathbb{E}(Y_i | X_{1i} = x_{1i}, \ldots, X_{pi} = x_{pi})$ của p biến phản hồi tại các giá trị biến giải thích $(x_{1i}, \ldots, x_{pi})$, liên hệ nó với *linear predictor*,

$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} \equiv \eta_i.$$

Hàm $g(\cdot)$ là hàm đơn điệu, 1-1, khả vi, và do đó

$$\mu_i = g^{-1}(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}) \equiv g^{-1}(\eta_i).$$

Nếu $g(\mu_i) = \mu_i$, thì ta có hàm liên kết đồng nhất (*identity link function*).

Cấu trúc của Generalized Linear Models

Họ phân phối mũ phân tán

Biến phản hồi Y trong GLM được giả định rằng có hàm phân phối thuộc vào **họ phân phối mũ phân tán** (**exponential-dispersion family**):

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

trong đó,

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

Thông thường,

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

Cấu trúc của Generalized Linear Models

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

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

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

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

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

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

Cấu trúc của Generalized Linear Models

Một số hàm liên kết tương ứng với phân phối của biến phản hồi:

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

Canonical link

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

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

GLM cho dữ liệu nhị phân

Cho dữ liệu, trong đó, biến phản hồi Y có dạng nhị phân (0/1), ví dụ:

- Y là trạng thái mua hàng của khách hàng: Mua hoặc Không mua;
- Y là trạng thái của thị trường trong 1 ngày: Lên hoặc Xuống;
- Y là trạng thái tín dụng của một người: Vỡ nợ hoặc Bình thường;
- Y là tình trạng công việc của một người: Thất nghiệp hoặc Có việc.

Với n quan sát, khi đó, quan sát thứ i, Y_i có giá trị 0 hoặc 1.

$\Rightarrow$ Y_i là biến ngẫu nhiên Bernoulli;

$\Rightarrow$ $\mu_i = \mathbb{E}(Y_i) = \Pr(Y_i = 1)$.

Với p biến giải thích, ta thiết lập được linear predictor:

$$\eta_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}.$$

$\Rightarrow$ mục tiêu dự đoán $\pi_i = \mathbb{E}(Y_i = 1 | X_{1i}, \ldots, X_{pi}) = \Pr(Y_i = 1 | X_{1i}, \ldots, X_{pi})$.

GLM cho dữ liệu nhị phân

Sử dụng liên kết chính tắc, logit link, ta có GLM cho biến nhị phân

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}$$

hay tương đương với

$$\Pr(Y_i = 1 | X_{1i}, \ldots, X_{pi}) = \pi_i = \frac{\exp\left(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}\right)}{1 + \exp\left(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}\right)}.$$

Hàm $\dfrac{\exp(u)}{1+\exp(u)}$ được gọi là **hàm logistic**.

$\Rightarrow$ GLM cho biến nhị phân với logit link được gọi là **mô hình hồi quy logistic**.

Ngoài ra ta còn có **mô hình hồi quy probit**:

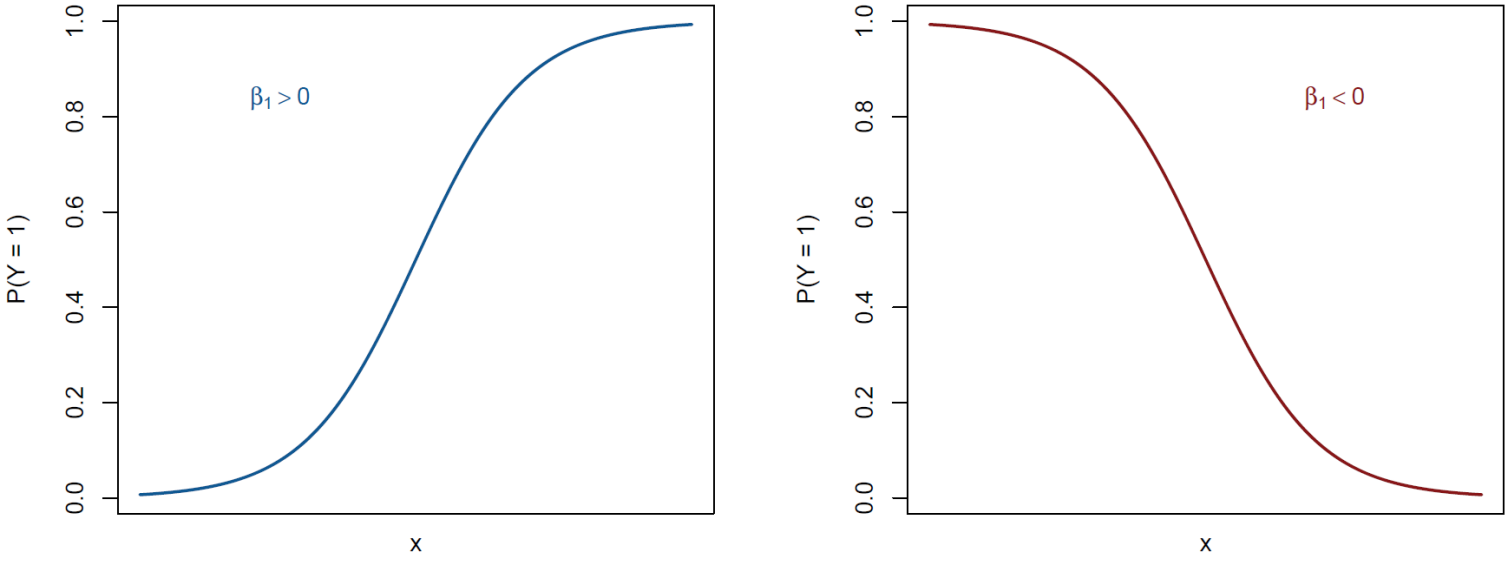
$$\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi},$$

với $\Phi^{-1}(\cdot)$ là hàm ngược của hàm xác suất tích lũy phân phối chuẩn $\mathcal{N}(0,1)$.

GLM cho dữ liệu nhị phân

Xét mô hình

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$$



Đường cong hồi quy logistic với một biến giải thích định lượng duy nhất là đường đơn điệu tăng hoặc đơn điệu giảm theo dấu của β_1.

GLM cho dữ liệu nhị phân - Diễn giải mô hình

Tỷ lệ cược - odds

Nếu một sự kiện A có xác suất xảy ra là p, thì tỷ lệ cược (*odds*) cho sự xuất hiện của A được xác định bởi $\frac{p}{1-p}$, tức là tỷ lệ giữa xác suất mà A xuất hiện và xác suất mà A không xuất hiện.

Ví dụ: A là sự kiện một người bị di căn tế bào ung thư:

- xác suất A xuất hiện là $p = 0.7$;
- tỷ lệ cược cho một người bị di căn tế bào ung thư là $\frac{0.7}{1-0.7} = 2.33$

Điều này có nghĩ là: xác suất quan sát được một người bệnh bị di căn tế bào ung thư là cao hơn 2.33 lần so với xác suất không quan sát được sự di căn tế bào ung thư.

GLM cho dữ liệu nhị phân - Diễn giải mô hình

Theo thiết lập mô hình logistic, ta có

$$\frac{\pi}{1-\pi} = \exp\left(\beta_0 + \beta_1 X_1 + \ldots + \beta_q X_q\right);$$

như vậy, ta có được biểu diễn của tỷ lệ cược theo $X_1, X_2, \ldots, X_q$. Hay dạng biểu diễn tuyến tính của log-odds

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_q X_q.$$

Từ dạng biểu diễn tuyến tính này, ta có quy tắc sau:

- nếu $\beta_j > 0$, thì khi giá trị X_j tăng một đơn vị thì log-odds tăng β_j đơn vị;
- nếu $\beta_j < 0$, thì khi giá trị X_j tăng một đơn vị thì log-odds giảm $|\beta_j|$ đơn vị.

Quy tắc này cũng tương đương với

- nếu $\beta_j > 0$, thì khi giá trị X_j tăng một đơn vị thì tỷ lệ cược tăng $\exp(\beta_j)$ lần;
- nếu $\beta_j < 0$, thì khi giá trị X_j tăng một đơn vị thì tỷ lệ cược giảm $\exp(\beta_j)$ lần.

GLM cho dữ liệu nhị phân - Ví dụ

Xét dữ liệu `Default` về tình trạng tín dụng của 10,000 khách hàng

	default	student	balance	income
1	No	No	729.52650	44361.625
2	No	Yes	817.18041	12106.135
3	No	No	1073.54916	31767.139
4	No	No	529.25060	35704.494
5	No	No	785.65588	38463.496
6	No	Yes	919.58853	7491.559
7	No	No	825.51333	24905.227
8	No	Yes	808.66750	17600.451
9	No	No	1161.05785	37468.529
10	No	No	0.00000	29275.268

- `default`: tình trạng vỡ nợ tín dụng (`Yes`/`No`);
- `student`: khách hàng là sinh viên (`Yes`/`No`);
- `balance`: số dư trung bình mà khách hàng còn lại trên thẻ tín dụng sau khi thực hiện thanh toán hàng tháng;
- `income`: thu nhập của khách hàng.

GLM cho dữ liệu nhị phân - Ví dụ

Dựa vào dữ liệu này, công ty tín dụng muốn xây dựng mô hình dự đoán khả năng một người sẽ vỡ nợ, $\Pr(\texttt{default} = 1)$, dựa trên các thông tin cơ bản:

- khách hàng là sinh viên - `student`
- số dư trung bình - `balance`
- thu nhập - `income`

Xét mô hình hồi quy logistic

$$\log\left(\frac{\Pr(\texttt{default} = 1)}{1 - \Pr(\texttt{default} = 1)}\right) = \beta_0 + \beta_1 \times \texttt{balance} + \beta_2 \times \texttt{student} + \beta_3 \times \texttt{income}.$$

GLM cho dữ liệu nhị phân - Ví dụ

Kết quả ước lượng mô hình như sau:

	Coefficient	Std. error	z-value	p-value
`Intercept`	-10.8691	0.4923	-22.080	< 0.0001
`balance`	0.0057	0.0002	24.738	< 0.0001
`student[Yes]`	-0.6467	0.2363	-2.738	0.0062
`income`	0.0303×10^{-4}	0.082×10^{-4}	0.370	0.7115

- Hệ số dương của `balance` cho thấy rằng việc tăng số dư có liên quan đến việc tăng khả năng vỡ nợ.
- Nói một cách chính xác, việc tăng một đơn vị số dư có liên quan đến việc tăng tỷ lệ cược vỡ nợ là $\exp(0.0057) \approx 1.0057$ lần.
- Hệ số âm đối với `student` cho biết rằng đối với một giá trị cố định của `balance` và `income`, một sinh viên ít có khả năng vỡ nợ hơn một người không phải là sinh viên.
- Thông tin về `income` không thực sự ảnh hưởng tới khả năng vỡ nợ tín dụng.

GLM cho dữ liệu nhị phân - Ví dụ

Dự đoán

Sử dụng kết quả ước lượng mô hình, ta có thể dự đoán xác suất vỡ nợ:

- người dùng là sinh viên, $\texttt{balance} = 1,500\$$ và $\texttt{income} = 40,000\$$:

$$\begin{aligned}\widehat{\Pr}(\texttt{default} = 1) &= \frac{\exp\left(-10.8691 + 0.0057 \times 1500 - 0.6467 \times 1 + 0.0303 \times 10^{-4} \times 40000\right)}{1 + \exp\left(-10.8691 + 0.0057 \times 1500 - 0.6467 \times 1 + 0.0303 \times 10^{-4} \times 40000\right)} \\ &= 0.0579\end{aligned}$$

- người dùng là phải sinh viên, $\texttt{balance} = 1,500\$$ và $\texttt{income} = 40,000\$$:

$$\begin{aligned}\widehat{\Pr}(\texttt{default} = 1) &= \frac{\exp\left(-10.8691 + 0.0057 \times 1500 - 0.6467 \times 0 + 0.0303 \times 10^{-4} \times 40000\right)}{1 + \exp\left(-10.8691 + 0.0057 \times 1500 - 0.6467 \times 0 + 0.0303 \times 10^{-4} \times 40000\right)} \\ &= 0.1050\end{aligned}$$

GLM cho dữ liệu nhị phân - Áp dụng

GLM cho dữ liệu nhị phân nói chung hay mô hình logistic nói riêng thường được áp dụng trong:

- dự đoán khả năng xảy ra 1 sự kiện, ví dụ: vỡ nợ, giao dịch thành công, có việc/thất nghiệp, sử dụng mạng xã hội (có/không);
- xử lý dữ liệu cho nhiệm vụ xây dựng mô hình credit scoring trong lĩnh vực tài chính;
- xử lý dữ liệu cho nhiệm vụ phân loại hai danh mục đối tượng;
- xử lý dữ liệu số lần sự kiện "thành công" trên tổng số m lần quan sát, ví dụ: số côn trùng chết trong tổng số côn trùng bị phun thuốc diệt côn trùng (dữ liệu dose-response).

GLM cho dữ liệu nhị phân - Vấn đề overdispersion/underdispersion

Đối với mô hình Logistic, tham số phân tán, ϕ, là bằng 1 theo lý thuyết.

$\Rightarrow \mathbb{V}\mathrm{ar}(Y) = \mu(1-\mu)$.

Tuy nhiên, với dữ liệu thực tế, ước lượng $\widehat{\phi}$ của ϕ được xác định bởi

$$\widehat{\phi} = \frac{1}{n-p-1}\sum_{i=1}^{n}\frac{(Y_i-\widehat{\mu}_i)^2}{\widehat{\mu}_i(1-\widehat{\mu}_i)}$$

có thể khác 1. Cụ thể,

- $\widehat{\phi} > 1$, khi đó ta nói dữ liệu quá phân tán - **overdispersion**, so với mô hình Logistic;
- $\widehat{\phi} < 1$, khi đó ta nói dữ liệu phân tán thấp - **underdispersion**, so với mô hình Logistic.

Ví dụ, với mô hình Logistic cho dữ liệu `default`.

Ta tính được $\widehat{\phi} = 0.7007$, nhỏ hơn 1 đáng kể.

$\Rightarrow$ mô hình đang bị vấn đề underdispersion.

GLM cho dữ liệu nhị phân - Vấn đề overdispersion/underdispersion

Overdispersion hoặc underdispersion làm giảm độ chính xác trong ước lượng sai số chuẩn của hệ số, $\mathrm{SE}(\widehat{\beta}_j)$.

Cụ thể

- overdispersion $\Rightarrow \mathrm{SE}(\widehat{\beta})$ bị ước lượng thấp (underestimated);
- underdispersion $\Rightarrow \mathrm{SE}(\widehat{\beta})$ bị ước lượng cao (overestimated).

Điều này sẽ ảnh hưởng tới kết quả của kiểm định giả thuyết $H_0 : \beta_j = 0$.

Phương án xử lý

Để xử lý vấn đề overdispersion/underdispersion, ta có thể dùng 1 trong các cách sau:

- sử dụng ước lượng $\widehat{\phi}$ trong tính toán $\mathrm{SE}(\widehat{\beta})$, thay vì dùng 1 (giá trị lý thuyết);
- sử dụng mô hình quasi-binomial.

GLM cho dữ liệu đếm

Biến phản hồi dạng số đếm

Xét biến định lượng Y miêu tả số lần một sự kiện A xuất hiện ngẫu nhiên trong

- một khung thời gian;
- một địa điểm/khu vực;
- một nhóm quần thể đối tượng;

với một tỷ lệ cụ thể.

Khi đó, Y cho kết quả là số đếm: 0, 1, 2, 3, ...

Ví dụ:

- Y là lượng thuê xe đạp trong 24h (1 ngày) tại một thành phố;
- Y là số lượng loài cây trong một hòn đảo;
- Y là số lượng khách hàng mua sản phẩm tại một cửa hàng trong một ngày;
- Y là số lượng người bị chết vì một căn bệnh tại một thành phố trong 1 năm.

GLM cho dữ liệu đếm

Giả sử ta có bộ dữ liệu ngẫu nhiên $(Y_i, X_{1i}, X_{2i}, \ldots, X_{qi})$, với $i = 1, \ldots, n$:

- các quan sát thứ i là độc lập;
- Y_i là biến phản hồi dạng số đếm, nhận giá trị $0, 1, 2, \ldots$;
- $X_{1i}, \ldots, X_{qi}$ là các biến giải thích có thể liên quan tới Y_i.

Do Y_i có dạng số đếm $\Rightarrow$ ta có thể dùng phân phối Poisson để mô hình hóa.

Khi đó, $Y_i \sim \mathcal{P}(\lambda_i), \lambda_i > 0$, và $\mathbb{E}(Y_i) = \mathbb{V}\mathrm{ar}(Y_i) = \lambda_i$.

Mô hình tiên đoán trung bình của Y_i khi biết $X_{1i}, X_{2i}, \ldots, X_{qi}$:

$$\mathbb{E}(Y_i | X_{1i}, X_{2i}, \ldots, X_{qi}) = \exp(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_q X_{qi}),$$

được gọi là mô hình **Poisson log-linear**. Mô hình này tương đương với,

$$\log(\mu_i) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_q X_{qi},$$

với $\mu_i = \mathbb{E}(Y_i | X_{1i}, X_{2i}, \ldots, X_{qi})$.

GLM cho dữ liệu đếm

Trong một số trường hợp, giá trị đếm Y_i là tỷ lệ với một chỉ số t_i, với

- t_i là một lượng thời gian;
- t_i là kích cỡ quần thể;
- t_i là diện tích môt khu vực không gian.

Khi đó, tỷ lệ mẫu Y_i/t_i tương ứng với kỳ vọng λ_i/t_i. Điều này dẫn tới mô hình

$$\log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_q X_{qi},$$

hay tương đương,

$$\log(\mu_i) = \log(t_i) + \beta_0 + \beta_1 X_{1i} + \ldots + \beta_q X_{qi},$$

thành phần $\log(t_i)$ được gọi là **offset**.

Mean-variance relationship

Theo tính chất của phân phối Poisson,

$$\mu_i = \mathbb{E}(Y_i|X_{1i}, X_{2i}, \ldots, X_{qi}) = \mathbb{V}\mathrm{ar}(Y_i|X_{1i}, X_{2i}, \ldots, X_{qi})$$

$\Rightarrow$ mô hình Poisson log-linear có khả năng xử lý vấn đề trung bình và phương sai có liên hệ nhau (1 trường hợp của phương sai không đồng nhất).

GLM cho dữ liệu đếm - Diễn giải mô hình

Theo thiết lập mô hình Poisson log-linear, ta có

$$\log(\mu) = \log(t) + \beta_0 + \beta_1 X_1 + \ldots + \beta_q X_q.$$

Từ dạng biểu diễn tuyến tính này, ta có quy tắc sau:

- nếu $\beta_j > 0$, thì khi giá trị X_j tăng một đơn vị thì $\log(\mu)$ tăng β_j đơn vị;
- nếu $\beta_j < 0$, thì khi giá trị X_j tăng một đơn vị thì $\log(\mu)$ giảm $|\beta_j|$ đơn vị,

khi các biến khác, $X_k, (k \neq j)$ được giữ là hằng số.

Quy tắc này cũng tương đương với

- nếu $\beta_j > 0$, thì khi giá trị X_j tăng một đơn vị thì μ tăng $\exp(\beta_j)$ lần;
- nếu $\beta_j < 0$, thì khi giá trị X_j tăng một đơn vị thì μ giảm $\exp(\beta_j)$ lần,

khi các biến khác, $X_k, (k \neq j)$ được giữ là hằng số.

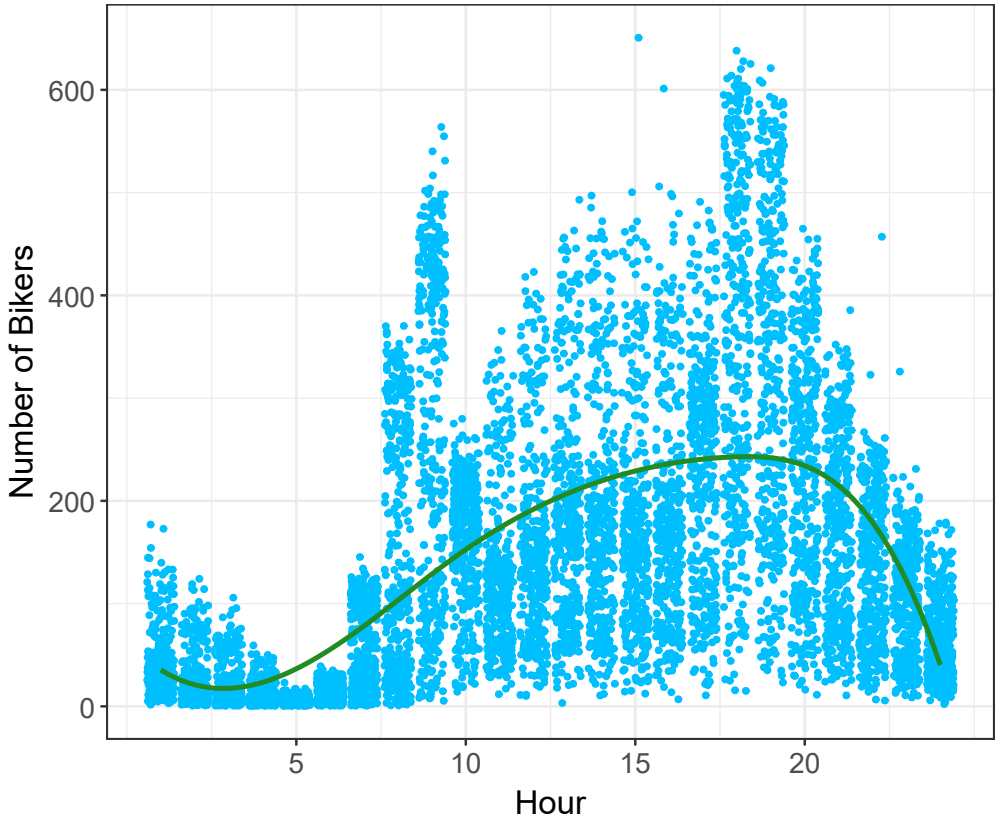
GLM cho dữ liệu đếm - Ví dụ

Ta xét dữ liệu `Bikeshare` cung cấp số lượng xe đạp cho thuê theo giờ và theo ngày trong giai đoạn 2011 và 2012 trong hệ thống Capital Bikeshare (Mỹ) cùng thông tin thời tiết và theo mùa tương ứng.

	bikers	season	mnth	day	hr	holiday	weekday
1	16	1	Jan	1	0	0	6
2	40	1	Jan	1	1	0	6
3	32	1	Jan	1	2	0	6
4	13	1	Jan	1	3	0	6
5	1	1	Jan	1	4	0	6
6	1	1	Jan	1	5	0	6
7	2	1	Jan	1	6	0	6
8	3	1	Jan	1	7	0	6
9	8	1	Jan	1	8	0	6
10	14	1	Jan	1	9	0	6

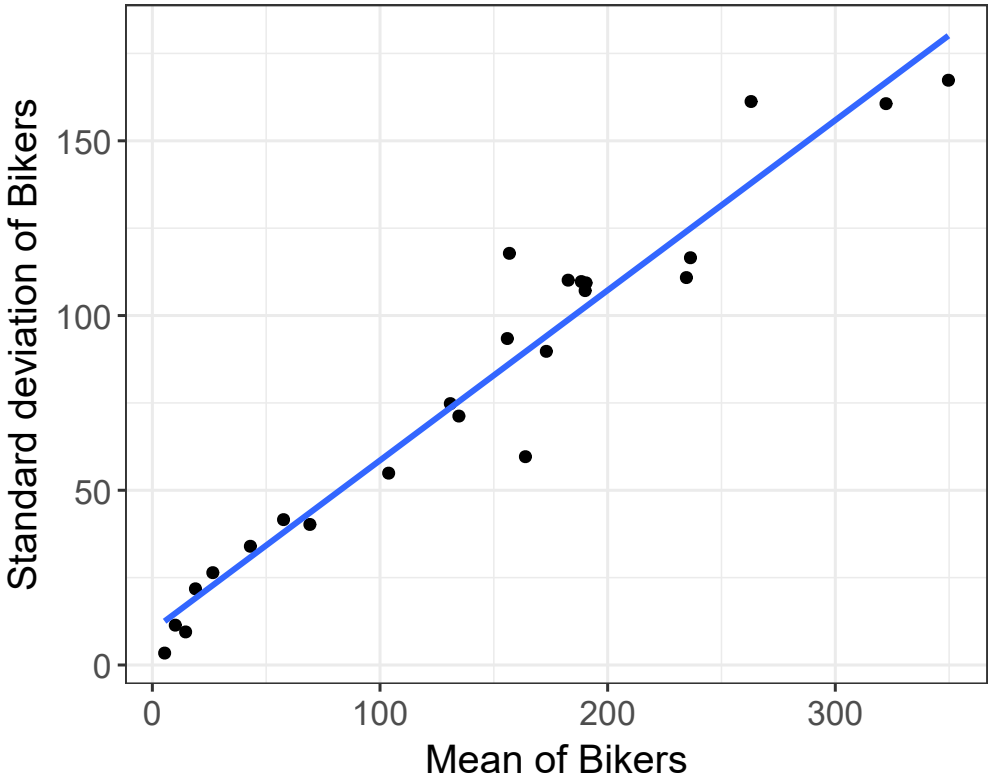
`bikers` là số xe đạp được cho thuê trong 1 ngày tương ứng theo giờ.

GLM cho dữ liệu đếm - Ví dụ



⇒ số lượng xe thuê thay đổi theo giờ và là có giá trị: 0, 1, 2, ...

GLM cho dữ liệu đếm - Ví dụ



Trung bình và Độ lệch chuẩn của `bikers` theo giờ có liên hệ với nhau.

$\Rightarrow$ mean-variance relationship.

GLM cho dữ liệu đếm - Ví dụ

Ta muốn xây dựng mô hình ước lượng số lượng xe đạp được thuê (`bikers`) dựa vào thông tin:

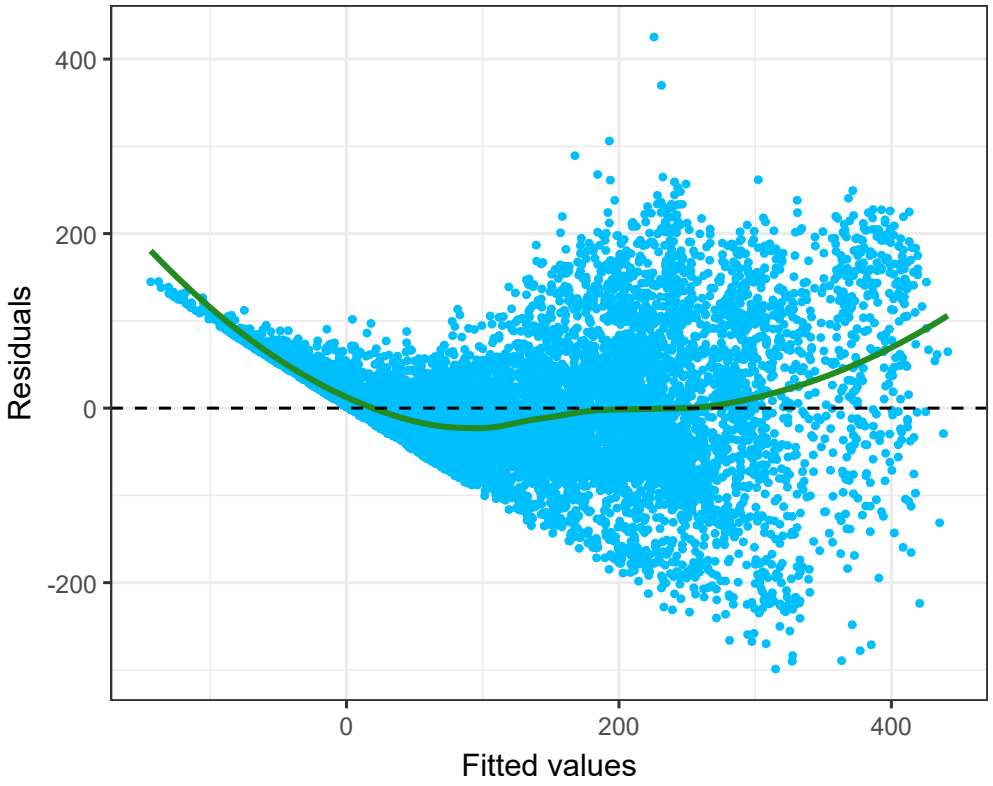
- `mnth`: 12 tháng;
- `hr`: 24 giờ;
- `workingday`: ngày làm việc/ngày nghỉ;
- `temp`: nhiệt độ;
- `weathersit`: thời tiết trong ngày.

Phương án 1: Mô hình hồi quy tuyền tính

$$\texttt{bikers}_i = \beta_0 + \beta_1 \texttt{mnth}_i + \beta_2 \texttt{hr}_i + \beta_3 \texttt{workingday}_i + \beta_4 \texttt{temp}_i + \beta_5 \texttt{weathersit}_i + \varepsilon_i$$

GLM cho dữ liệu đếm - Ví dụ

Kết quả phân tích thặng dư

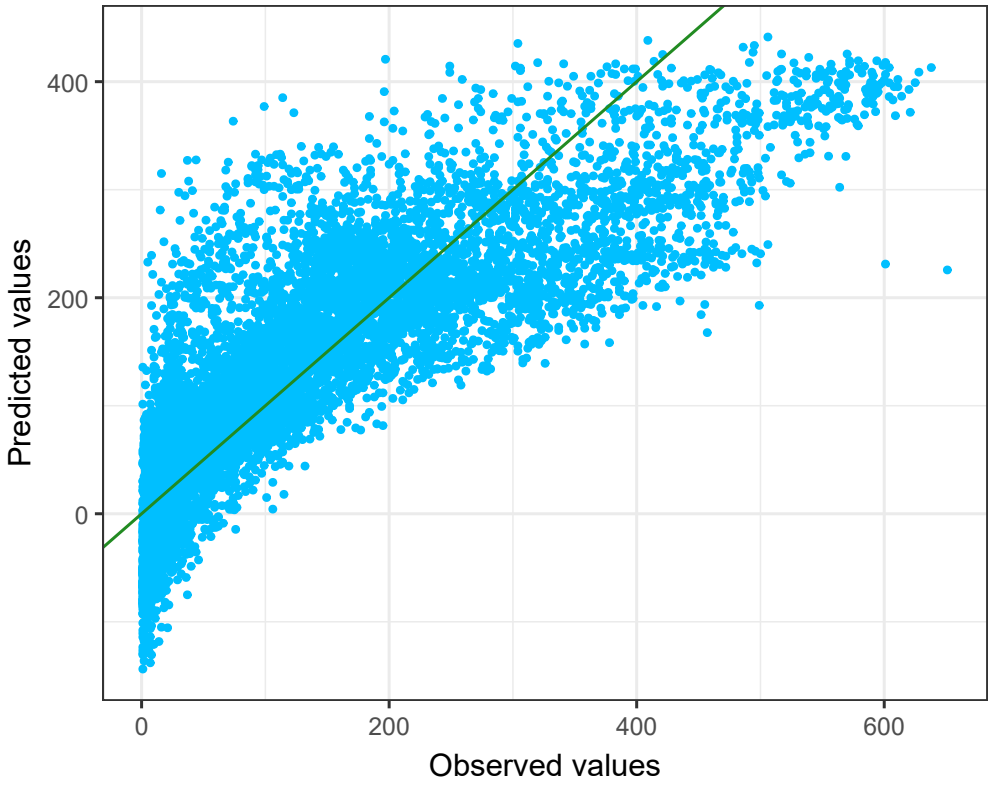


cho thấy

- cấu trúc tuyến tính là không hợp lý;
- phương sai không đồng nhất.

GLM cho dữ liệu đếm - Ví dụ

Mặt khác,



kết quả tiên đoán cho thấy:

- sai lệch nhiều với dữ liệu `bikers`;
- giá trị tiên đoán bị âm, khoảng 9.63%.

GLM cho dữ liệu đếm - Ví dụ

Phương án 2: Mô hình hồi quy Poisson log-linear

- $\texttt{bikers}_i \sim \mathcal{P}(\mu_i)$;
- mô hình

$$\log(\mu_i) = \beta_0 + \beta_1\texttt{mnth}_i + \beta_2\texttt{hr}_i + \beta_3\texttt{workingday}_i + \beta_4\texttt{temp}_i + \beta_5\texttt{weathersit}_i.$$

Bảng tổng hợp kết quả

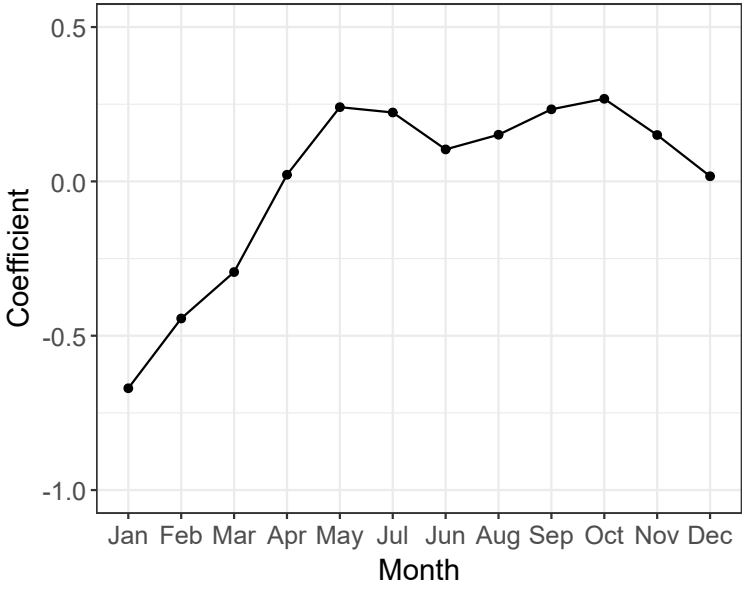
	Coefficient	Std. error	z-value	p-value
Intercept	4.1182	0.0305	135.183	<0.0001
workingday	0.0147	0.0099	1.483	0.138
temp	0.7853	0.0581	13.526	<0.0001
weathersit[cloudy/misty]	−0.0752	0.0110	−6.824	<0.0001
weathersitlight[light rain/snow]	−0.5758	0.0205	−28.047	<0.0001
weathersitheavy[heavy rain/snow]	−0.9263	0.8438	−1.098	0.272

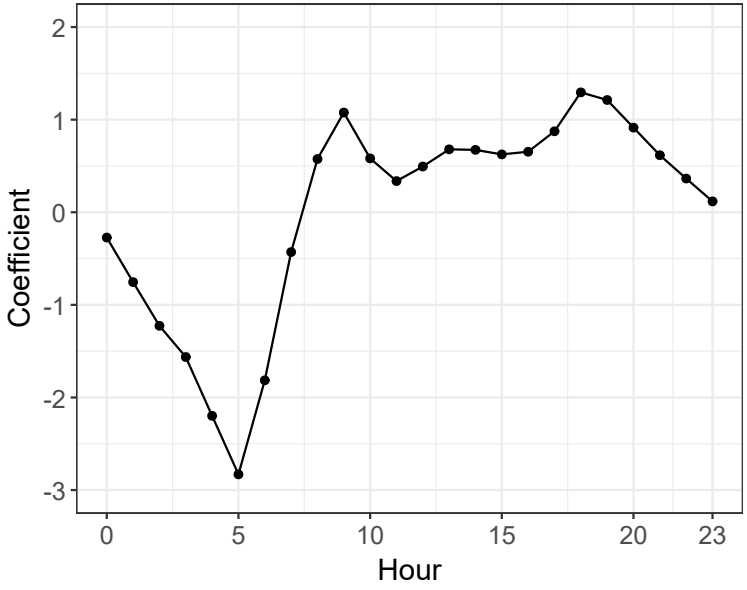
Kết quả ước lượng cho hệ số của các biến nhân tố:

- mnth
- hr

được miêu tả bởi hình sau.

GLM cho dữ liệu đếm - Ví dụ





Ta có một số nhận xét:

- giá trị âm ở các tháng 1, 2, 3, 4 phản ánh lượng thuê xe thấp ở mùa đông;
- lượng thuê xe tăng cao từ mùa xuân tời mùa thu;
- lượng thuê xe giảm vào các khung giờ khuya, và tăng lại từ 7h sáng.

GLM cho dữ liệu đếm - Ví dụ

Một số diễn giải cho kết quả của biến thời tiết, `weathersit`, trong mô hình như sau:

- dưới điều kiện trời nhiều mây, `weathersit[cloudy/misty]`, lượng thuê xe đạp thay đổi bởi $\exp(-0.0752) \approx 0.9276$

  $\Rightarrow$ chỉ 92.76% số người sử dụng xe đạp khi trời nhiều mây so với khi trời quang;

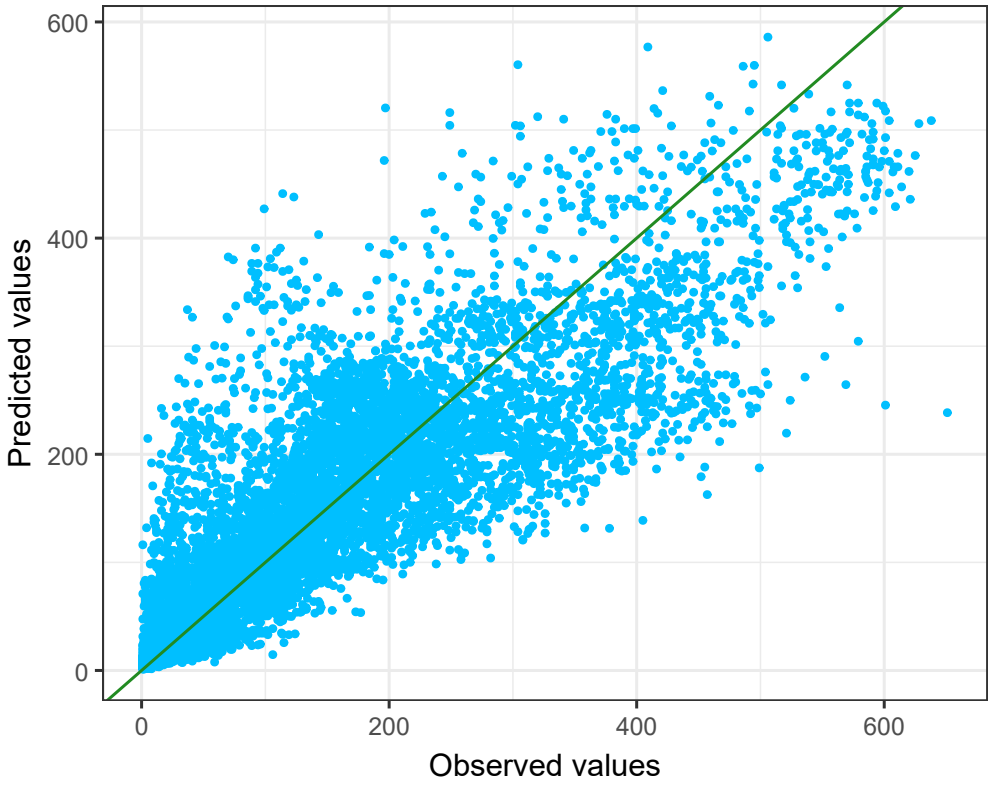
- dưới điều kiện trời mưa nhẹ (hoặc tuyết nhẹ), `weathersit[light rain/snow]`, lượng thuê xe đạp thay đổi bởi $\exp(-0.5758) \approx 0.5622$

  $\Rightarrow$ chỉ 56.22% số người sử dụng xe đạp khi trời mưa nhẹ (hoặc tuyết nhẹ) so với khi trời quang;

- dưới điều kiện trời mưa lớn (hoặc tuyết rơi dày), `weathersit[heavy rain/snow]`, lượng thuê xe đạp thay đổi bởi $\exp(-0.9263) \approx 0.3960$

  $\Rightarrow$ chỉ 39.60% số người sử dụng xe đạp khi trời mưa lớn (hoặc tuyết rơi dày) so với khi trời quang;

GLM cho dữ liệu đếm - Ví dụ



Kết quả tiên đoán cho thấy:

- so sánh được với dữ liệu `bikers`;
- giá trị tiên đoán không bị âm.

GLM cho dữ liệu đếm - Vấn đề overdispersion/underdispersion

Đối với mô hình Poisson, tham số phân tán, ϕ, là bằng 1 theo lý thuyết.

$\Rightarrow \mathbb{E}(Y) = \mathbb{V}\mathrm{ar}(Y) = \mu$.

Tuy nhiên, với dữ liệu thực tế, ước lượng $\widehat{\phi}$ của ϕ được xác định bởi

$$\widehat{\phi} = \frac{1}{n-p-1} \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}$$

có thể khác 1. Cụ thể,

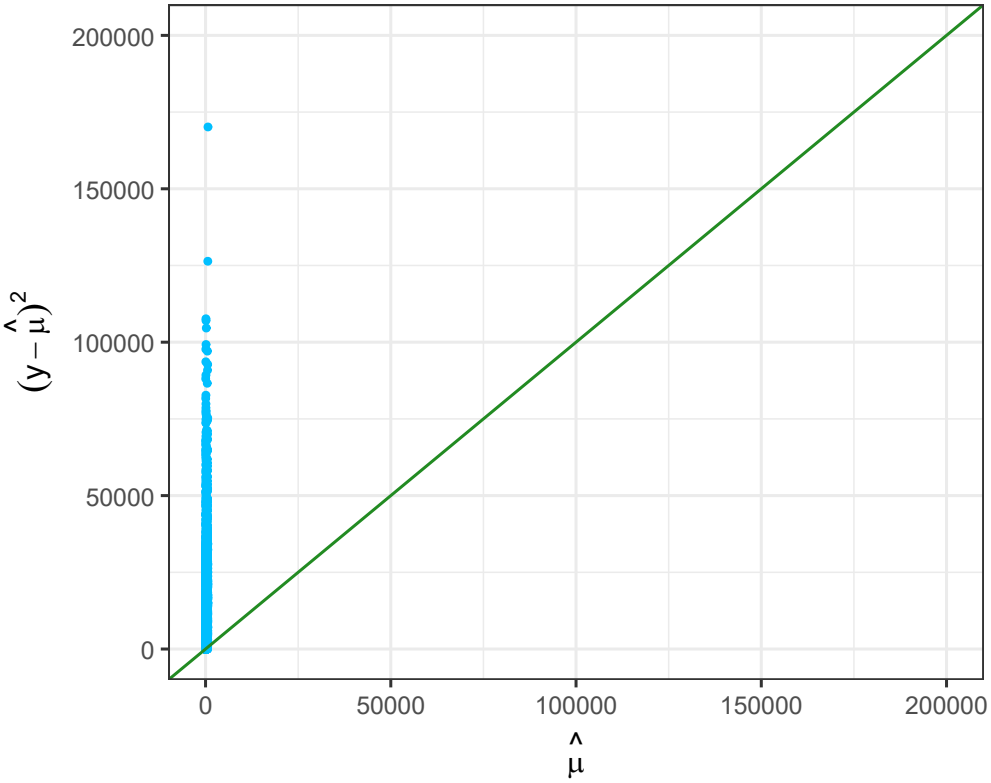
- $\widehat{\phi} > 1 \Rightarrow \mathbb{E}(Y) < \mathbb{V}\mathrm{ar}(Y)$, khi đó ta nói dữ liệu quá phân tán - **overdispersion**, so với mô hình Poisson;
- $\widehat{\phi} < 1 \Rightarrow \mathbb{E}(Y) > \mathbb{V}\mathrm{ar}(Y)$, khi đó ta nói dữ liệu phân tán thấp - **underdispersion**, so với mô hình Poisson.

GLM cho dữ liệu đếm - Vấn đề overdispersion/underdispersion

Chẳng hạn với mô hình Poisson cho dữ liệu `bikers`.

Ta tính được $\widehat{\phi} = 25.59892$, lớn hơn 1 rất nhiều.

Biểu đồ so sánh giữa $\widehat{\mu}_i$ và $(Y_i - \widehat{\mu}_i)^2$ dưới đây,



cho thấy vấn đề overdispersion là rất trầm trọng.

GLM cho dữ liệu đếm - Vấn đề overdispersion/underdispersion

Overdispersion hoặc underdispersion làm giảm độ chính xác trong ước lượng sai số chuẩn của hệ số, $\mathrm{SE}(\widehat{\beta}_j)$.

Cụ thể

- overdispersion $\Rightarrow \mathrm{SE}(\widehat{\boldsymbol{\beta}})$ bị ước lượng thấp (underestimated);
- underdispersion $\Rightarrow \mathrm{SE}(\widehat{\boldsymbol{\beta}})$ bị ước lượng cao (overestimated).

Điều này sẽ ảnh hưởng tới kết quả của kiểm định giả thuyết $H_0 : \beta_j = 0$.

Phương án xử lý

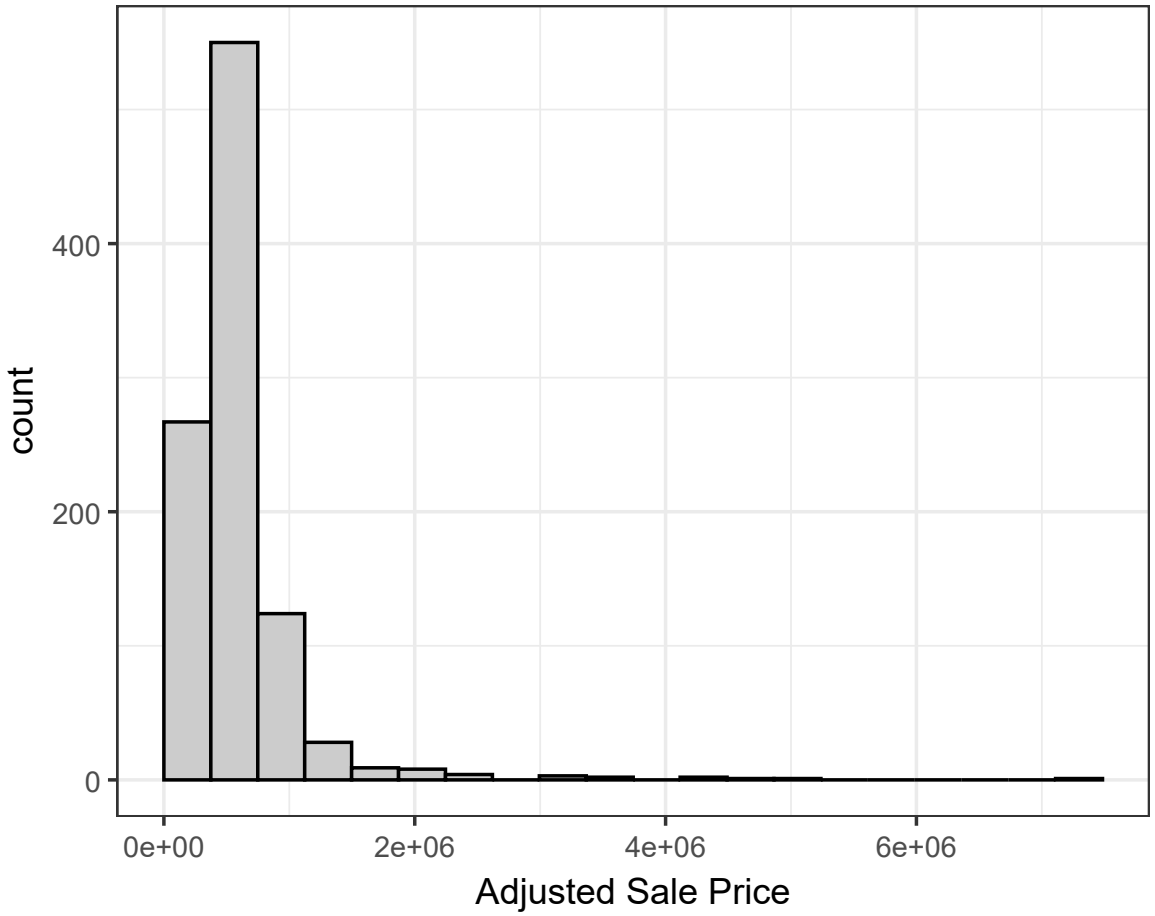
Để xử lý vấn đề overdispersion/underdispersion, ta có thể dùng 1 trong các cách sau:

- sử dụng ước lượng $\widehat{\phi}$ trong tính toán $\mathrm{SE}(\widehat{\boldsymbol{\beta}})$, thay vì dùng 1 (giá trị lý thuyết);
- sử dụng mô hình Quasi-Poisson.

GLM cho dữ liệu lệch phải

Trong một số dữ liệu, biến phản hồi Y có thiên hướng bị lệch phải.

Ví dụ giá bán nhà



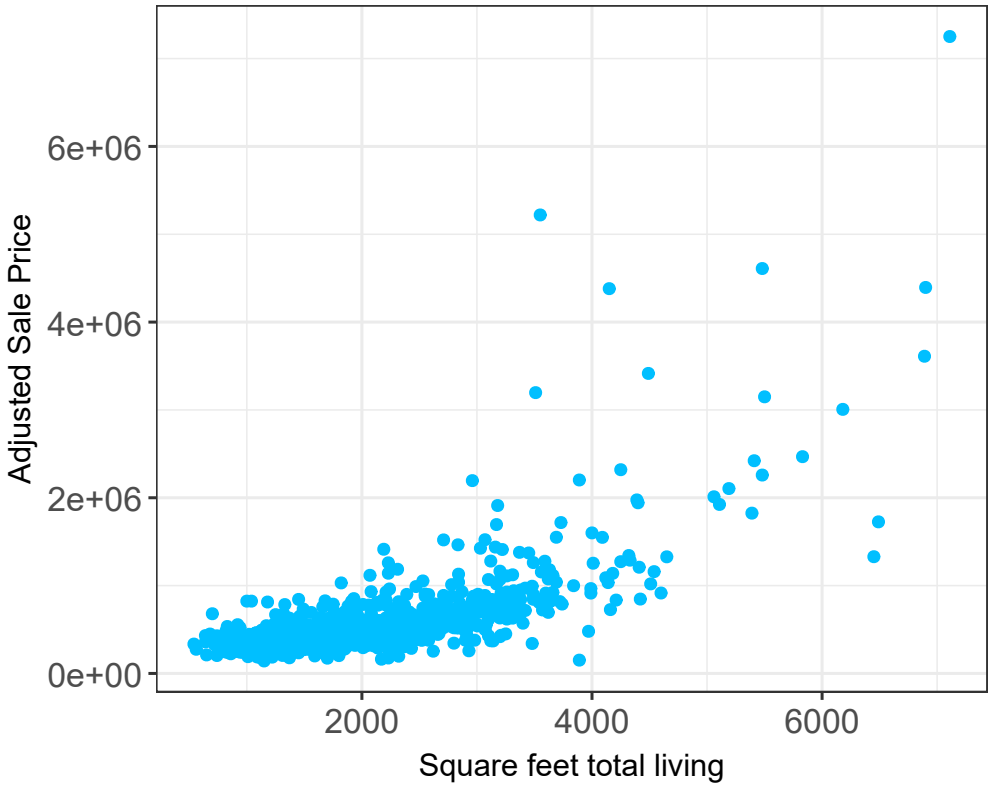
GLM cho dữ liệu lệch phải

Trong trường hợp này, nếu ta sử dụng mô hình hồi quy tuyến tính, có thể sẽ có các vấn đề:

- phương sai không đồng nhất;
- giá trị ngoại lai ảnh hưởng tới mô hình.

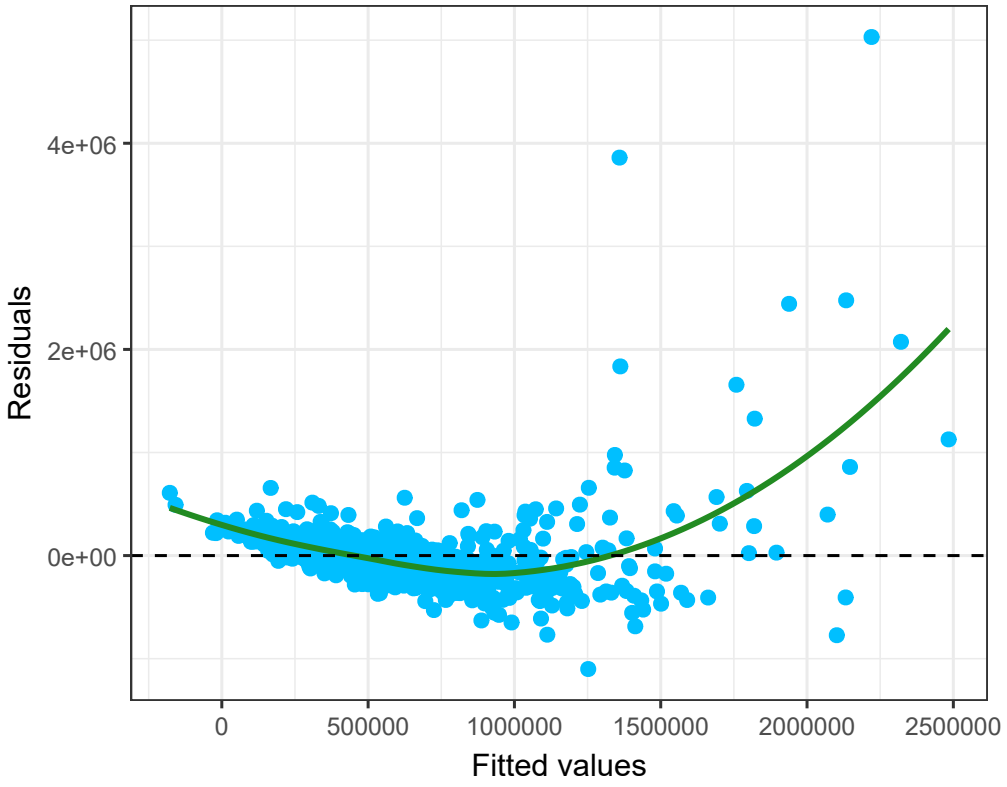
Ví dụ: xét mô hình hồi quy tuyến tính dự đoán giá bán nhà

$$\texttt{adj_sale_price}_i = \beta_0 + \beta_1 \texttt{sq_ft_tot_living}_i + \beta_2 \texttt{bldg_grade}_i + \beta_3 \texttt{yr_built}_i + \varepsilon_i,$$



GLM cho dữ liệu lệch phải

Kết quả phân tích thặng dư

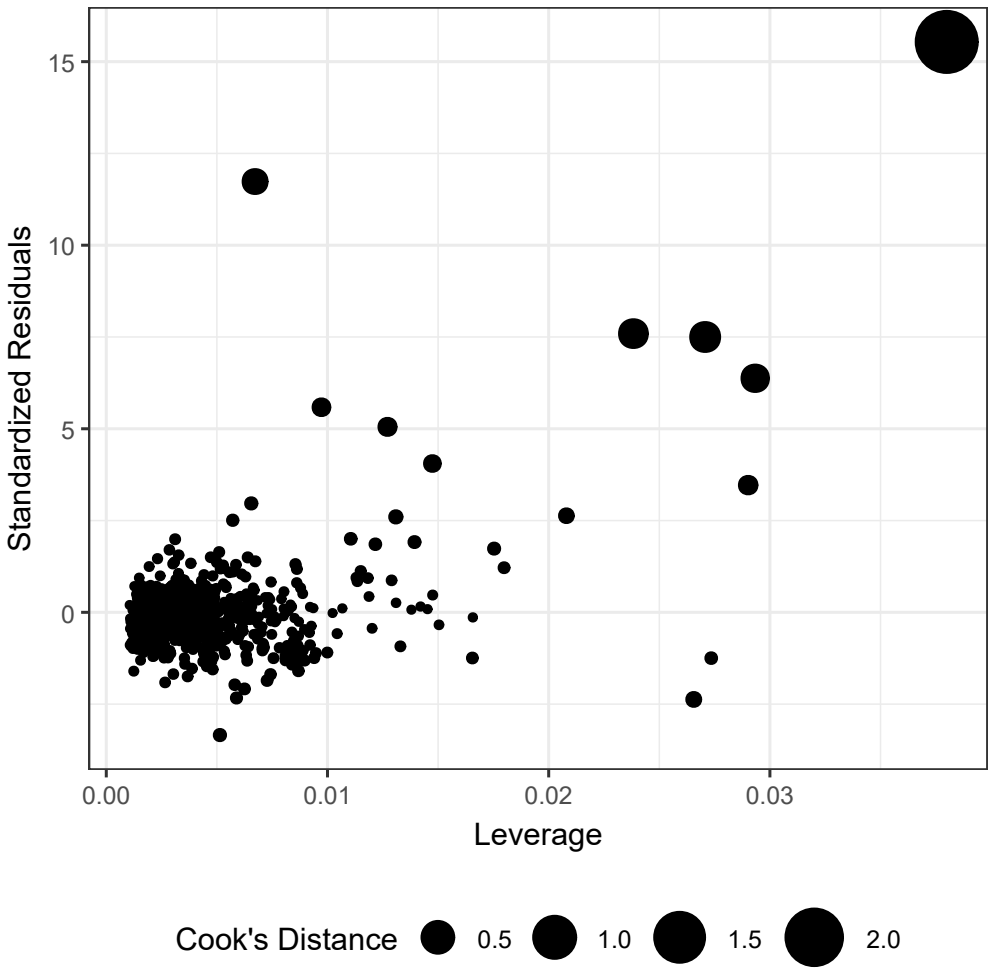


cho thấy

- cấu trúc tuyến tính là không hợp lý;
- phương sai không đồng nhất.

GLM cho dữ liệu lệch phải

Kết quả Cook's distance và Leverage



cho thấy có nhiều điểm có Cook's distance lớn ⇒ các điểm có ảnh hưởng.

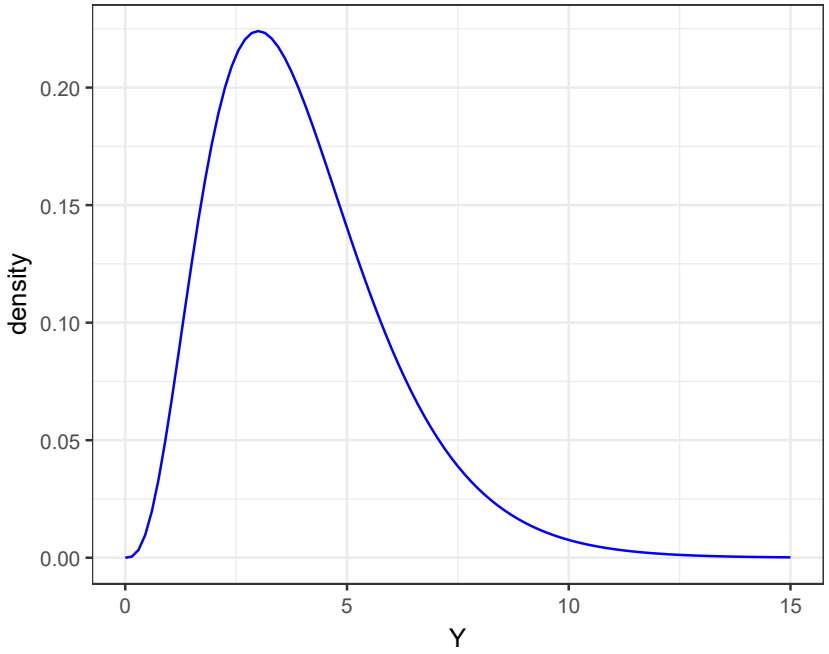
GLM cho dữ liệu lệch phải

Phân phối Gamma và dữ liệu lệch phải

Biến $Y \sim \mathcal{G}(\alpha, \beta)$ với $\alpha, \beta > 0$ được gọi là shape và rate parameter:

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

với $y > 0$ và $\Gamma(\alpha)$ là hàm Gamma. Ví dụ, với $\alpha = 4, \beta = 1$, ta có hàm mật độ



$\Rightarrow$ hàm mật độ có xu hướng lệch phải.

GLM cho dữ liệu lệch phải

Đối với phân phối Gamma, ta có tính chất:

- hàm mật độ có xu hướng lệch phải;

  $\Rightarrow$ các điểm giá trị lớn có thể xuất hiện và được coi là giá trị tất yếu;
- $\mathbb{E}(Y) = \mu = \dfrac{\alpha}{\beta}$;
- $\mathbb{V}\mathrm{ar}(Y) = \sigma^2 = \dfrac{\alpha}{\beta^2} = \dfrac{\mu^2}{\alpha}$

  $\Rightarrow$ phương sai σ^2 thay đổi theo μ.

Do đó, sử dụng phân phối Gamma cho biến phản hồi Y giải quyết được hai vấn đề:

- giá trị lớn trong dữ liệu của biến phản hồi;
- phương sai không đồng nhất.

GLM cho dữ liệu lệch phải

Giả sử ta có bộ dữ liệu ngẫu nhiên $(Y_i, X_{1i}, X_{2i}, \ldots, X_{qi})$, với $i = 1, \ldots, n$:

- các quan sát thứ i là độc lập;
- Y_i là biến phản hồi có phân phối dạng lệch phải, nhận giá trị dương;
- $X_{1i}, \ldots, X_{qi}$ là các biến giải thích có thể liên quan tới Y_i.

Do Y_i có phân phối dạng lệch phải $\Rightarrow$ có thể dùng phân phối Gamma để mô hình hóa. Khi đó, $Y_i \sim \mathcal{G}(\alpha_i, \beta)$ với $\alpha_i > 0, \beta > 0$.

Mô hình tiên đoán trung bình của Y_i khi biết $X_{1i}, X_{2i}, \ldots, X_{qi}$:

$$\mathbb{E}(Y_i | X_{1i}, X_{2i}, \ldots, X_{qi}) = g(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_q X_{qi}),$$

được gọi là mô hình **Gamma**. Mô hình này tương đương với,

$$g^{-1}(\mu_i) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_q X_{qi},$$

với $\mu_i = \mathbb{E}(Y_i | X_{1i}, X_{2i}, \ldots, X_{qi})$. Hàm $g(\cdot)$ được xác định:

- $g(u) = u^{-1}$ (inverse link) – không đảm bảo $\mu > 0$;
- $g(u) = \exp(u)$ (log link);
- $g(u) = u$ (identity link).

GLM cho dữ liệu lệch phải - Ví dụ

Xét lại dữ liệu về giá bán nhà.

Do `adj_sale_price` có phân phối lệch phải (histogram slide 105).

$\Rightarrow$ ta xét mô hình Gamma với log link:

$$\mu_i = \exp\left(\beta_0 + \beta_1 \texttt{sq_ft_tot_living}_i + \beta_2 \texttt{bldg_grade}_i + \beta_3 \texttt{yr_built}_i\right),$$

Bảng tổng hợp kết quả

	Coefficient	Std. error	t-value	p-value
`Intercept`	19.9764	0.6345	31.48	<0.0001
`sq_ft_tot_living`	0.0002	1.635×10^{-5}	13.81	<0.0001
`bldg_grade`	0.2284	0.2284	17.31	<0.0001
`yr_built`	-0.0046	0.0132	-13.58	<0.0001

GLM cho dữ liệu lệch phải - Ví dụ

Đối với mô hình Gamma, tham số phân tán, ϕ, được ước lượng bởi

$$\widehat{\phi} = \frac{1}{n-p-1} \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i^{\,2}}.$$

Với giá trị $\widehat{\phi}$, ta tính được phương sai điều kiện của Y dưới mô hình Gamma là $\widehat{\sigma}^2 = \widehat{\mu}^2\,\widehat{\phi} \Rightarrow \widehat{\sigma} = \widehat{\mu}\sqrt{\widehat{\phi}}$.

Ví dụ, với mô hình trên, ta tính được $\widehat{\phi} = 0.0822$

$\Rightarrow \widehat{\sigma} = \widehat{\mu}\sqrt{0.0822} = 0.2867\widehat{\mu}$. Chẳng hạn:

- $\widehat{\mu} = 500000$ thì $\widehat{\sigma} = 0.2867 \times 500000 = 143376.2$;
- $\widehat{\mu} = 350000$ thì $\widehat{\sigma} = 0.2867 \times 350000 = 100363.3$;
- $\widehat{\mu} = 150000$ thì $\widehat{\sigma} = 0.2867 \times 150000 = 43012.85$.

$\Rightarrow$ trung bình số tiền bán nhà càng tăng thì độ lệch chuẩn càng tăng.

GLM cho dữ liệu trong (0, 1)

Trong một số trường hợp, dữ liệu Y có thể nhận giá trị trong (0, 1), chẳng hạn:

- tỷ lệ;
- điểm đánh giá được rescale (ví dụ, điểm chỉ số sức khỏe).

Khi đó, ta có thể dùng phân phối Beta để mô hình hóa Y.

Phân phối Beta, $\mathcal{B}e(\alpha, \beta)$

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

với $y \in [0, 1]$, và $\alpha, \beta > 0$. Ta xác định được

- trung bình $\mathbb{E}(Y) = \mu = \dfrac{\alpha}{\alpha+\beta} \in (0, 1)$;
- tham số phân tán $\phi = \alpha + \beta > 0$;
- phương sai $\mathbb{V}\mathrm{ar}(Y) = \sigma^2 = \dfrac{\mu(1-\mu)}{1+\phi}$.

Khi đó, ta có hàm mật độ theo cách đổi tham số μ và ϕ, $\mathcal{B}e(\mu, \phi)$:

$$f_Y(y|\mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},$$

GLM cho dữ liệu trong (0, 1)

Giả sử ta có bộ dữ liệu ngẫu nhiên $(Y_i, X_{1i}, X_{2i}, \ldots, X_{qi})$, với $i = 1, \ldots, n$:

- các quan sát thứ i là độc lập;
- Y_i là biến phản hồi có giá trị trong (0, 1);
- $X_{1i}, \ldots, X_{qi}$ là các biến giải thích có thể liên quan tới Y_i.

Do Y_i có giá trị trong (0, 1) $\Rightarrow$ có thể dùng phân phối Beta để mô hình hóa. Khi đó, $Y_i \sim \mathcal{B}e(\mu_i, \phi)$ với $0 < \alpha_i < 1, \phi > 0$.

Mô hình tiên đoán trung bình của Y_i khi biết $X_{1i}, X_{2i}, \ldots, X_{qi}$:

$$\mathbb{E}(Y_i | X_{1i}, X_{2i}, \ldots, X_{qi}) = g(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_q X_{qi}),$$

được gọi là mô hình hồi quy **Beta**. Mô hình này tương đương với,

$$g^{-1}(\mu_i) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_q X_{qi},$$

với $\mu_i = \mathbb{E}(Y_i | X_{1i}, X_{2i}, \ldots, X_{qi})$. Hàm $g(\cdot)$ được xác định:

- $g(u) = \dfrac{\exp(u)}{1 + \exp(u)}$ (logistic link);
- $g(u) = \Phi(u)$ (probit link).