

Phân tích đặc trưng mô hình

Trong bài giảng này chúng ta tiếp tục thảo luận đã bắt đầu từ lần trước và tập trung vào những dạng hàm số nào có thể là phù hợp với biến phụ thuộc và các biến hồi qui.

Cho tới nay trong khoá học này, chúng ta đã sử dụng từ *tuyến tính* trong hai tình huống quan trọng. Tình huống thứ nhất xảy ra khi chúng ta nói tới mô hình của chúng ta như là mô hình hồi qui *tuyến tính*. Tình huống thứ hai xảy ra khi chúng ta ghi nhận rằng các hàm tuyến tính của các biến ngẫu nhiên có phân phối xác suất chuẩn cũng có phân phối chuẩn.

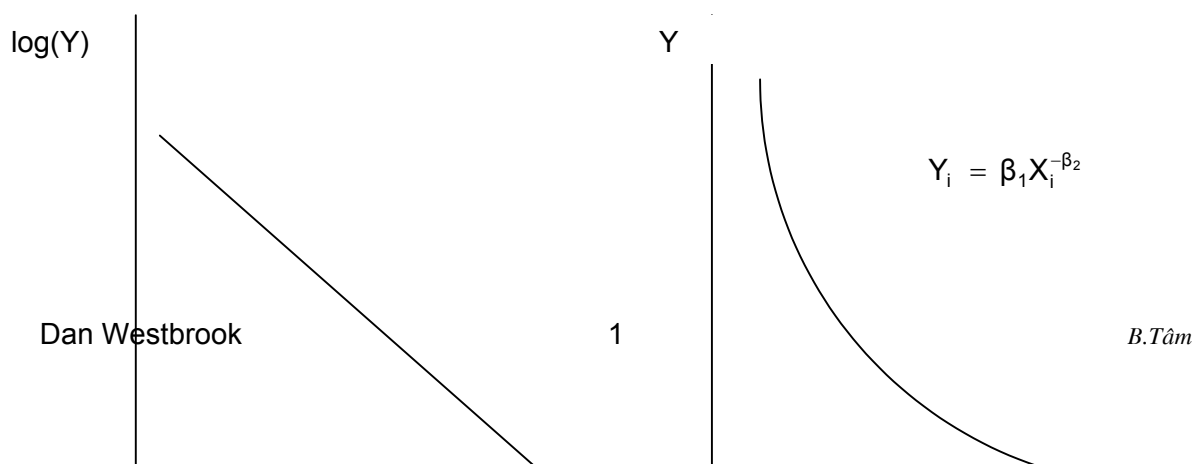
Khía cạnh quan trọng trong đó mô hình hồi qui *tuyến tính* là tuyến tính là nó là *tuyến tính trong các tham số*. Nó không cần là tuyến tính trong các biến. Các mô hình là tuyến tính trong các tham số để ước lượng bởi các phương pháp truyền thống (OLS), nhưng phần mềm của máy vi tính hiện đại cũng tạo điều kiện cho chúng ta ước lượng các mô hình phi tuyến trong các tham số.

Chúng ta sẽ xét nhiều trường hợp trong đó các đặc trưng mô hình phi tuyến đối với các biến trở nên rất hữu ích. Sau này, chúng ta cũng sẽ thu được kinh nghiệm nào đó với những mô hình nhất định là phi tuyến trong các tham số.

Mô hình Logarit kép

Mô hình logarit kép là một mô hình trong đó biến phụ thuộc và các biến độc lập ở dạng logarit. Mô hình này có nhiều công dụng khác nhau trong kinh tế học: các mô hình nhu cầu có độ co giãn không đổi, hay các hàm sản xuất Cobb-Douglas hay trans-log.

Trong khi chúng ta thường dùng các đường thẳng để thể hiện các đường cầu khi giảng dạy các nguyên tắc mô, thì chúng có thể không đại diện tốt cho dữ liệu thực tế. Thường là, mối quan hệ giữa giá và lượng cầu được mô tả một cách tuyệt vời bởi một mô hình logarit kép. Hai đồ thị dưới đây chỉ ra liên hệ giữa mối quan hệ tuyến tính dưới dạng logarit của các biến và mối quan hệ tương ứng giữa chính những biến này:



log(X)

X

Mối tương quan mô tả đường cong không thể được ước lượng bởi phương pháp OLS. Tuy nhiên, nếu chúng ta lấy logarit cả hai vế, thì kết quả này là một mối tương quan mà chúng ta có thể ước lượng bằng OLS.

Một nét đặc trưng hữu ích của mô hình logarit kép là độ co giãn của biến phụ thuộc theo một biến giải thích được cho trực tiếp bởi hệ số độ dốc.

Hãy nhớ lại định nghĩa của độ co giãn điểm : $\eta = \frac{\partial Y}{\partial X} \frac{X}{Y}$

Nếu chúng ta ước lượng một phép hồi qui tuyến tính, thì chúng ta có một hàm ước lượng cho độ dốc của Y theo X. Tuy nhiên, nếu chúng ta ước lượng một mô hình logarit kép, thì chúng ta có kết quả sau:

$$\begin{aligned}\partial \log(Y) &= \partial(\beta \log(X)) \\ \frac{1}{Y} \partial Y &= \frac{\beta}{X} \partial X \\ \Rightarrow \\ \beta &= \frac{\partial Y}{\partial X} \frac{X}{Y}\end{aligned}$$

Một ứng dụng thường gặp nhất của mô hình logarit kép là để ước lượng các hàm sản xuất. Hàm sản xuất Cobb-Douglas đã được phát hiện là cung cấp một đại diện tốt cho sản xuất trong nhiều tình huống. Đặc trưng mô hình này là :

$$Y = \beta_1 K^{\beta_2} L^{\beta_3} e^{\varepsilon}$$

Đây là một đặc trưng mô hình khác mà không thể ước lượng được bởi OLS. Tuy nhiên, nếu chúng ta lấy logarit cả hai vế, chúng ta tìm ra

$$\log(y) = \log(\beta_1) + \beta_2 \log(K) + \beta_3 \log(L) + \varepsilon$$

Dạng này của mô hình là tuyến tính trong các tham số, nên nó có thể được ước lượng bởi OLS. Một đặc trưng của hàm sản xuất Cobb-Douglas là rằng lợi thế kinh tế theo qui mô (RTS) là tổng các hệ số trên $\log(K)$ và $\log(L)$, nên để ước lượng RTS.

Hàm sản xuất trans-log được xác định như là một hàm số có chứa logarit của từng nhập lượng, logarit bình phương của từng nhập lượng, và các tích của từng đôi logarit nhập lượng. Hàm này là một xấp xỉ bậc hai cho bất cứ hàm sản xuất thực nhưng chưa biết nào. Đối với trường hợp p mà trong đó các nhập lượng là K và L , đặc trưng mô hình trans-log là:

$$\log(Y) = \log(\beta_1) + \beta_2 \log(K) + \beta_3 \log(L) + \beta_4 \log(K)^2 + \beta_5 \log(L)^2 + \beta_6 \log(K) * \log(L) + \varepsilon$$

Chúng ta đã tìm ra đặc trưng mô hình này trong Bài tập 12.

Mô hình Log tuyến tính

Các mô hình tăng trưởng thường có đặc trưng mô hình sau:

$$Y_t = (1+g)Y_{t-1}$$

trong đó g là tỉ lệ tăng trưởng, có thể được tính xấp xỉ bởi: $\frac{\Delta Y}{Y}$.

Thay thế tiếp cho ta mức của biến phụ thuộc tại thời điểm t như một hàm của giá trị biến phụ thuộc trong giai đoạn đầu tiên:

$$Y_t = Y_0(1+g)^t$$

Ở đây, số các giai đoạn đã qua kể từ giai đoạn đầu là biến giải thích, và các tham số chưa biết là giá trị ban đầu và tỉ lệ tăng trưởng. Chúng ta không thể ước lượng tỉ lệ tăng trưởng và giá trị ban đầu bởi OLS.

Lấy logarit cả hai vế:

$$\log(Y_t) = \log(Y_0) + \log(1+g)*t$$

Nếu chúng ta làm các thay thế sau, chúng ta có một mô hình hồi qui đơn biến với t là biến giải thích :

$$\beta_1 = \log(Y_0) \quad \text{và} \quad \beta_2 = \log(1+g) \Rightarrow$$

$$\log(Y_t) = \beta_1 + \beta_2 * t$$

Để dàng chứng minh rằng $\beta_2 = \frac{\partial \log(Y_t)}{\partial t} = \frac{\partial Y_t / Y_t}{\partial t}$ là thay đổi theo tỉ lệ của Y trên mỗi thay đổi một đơn vị trong thời gian.

Đặc trưng mô hình của chúng ta được hoàn thiện bằng cách bổ sung thêm một thành phần nhiễu ngẫu nhiên có các tính chất cổ điển :

$$\log(Y_t) = \beta_1 + \beta_2 * t + \varepsilon$$

Vì phương trình này là tuyến tính , nên chúng ta có thể ước lượng nó bằng OLS.

Giải để tìm hàm ước lượng của tỉ lệ tăng trưởng , chúng ta có:

$$\hat{g} = e^{\hat{\beta}_2} - 1$$

Trong khi hàm ước lượng của β_2 là không chệch, chúng ta không thể đưa ra khẳng định đó đối với hàm ước lượng của g .

Lịch hướng : Định lý không biến thiên (Invariance)

Xét một tham số chưa biết là một hàm của tham số chưa biết khác :

$$\theta = f(\varphi)$$

Nếu $\hat{\varphi}$ là một hàm ước lượng nhất quán của φ , thì $f(\hat{\varphi})$ là một hàm ước lượng nhất quán của θ .

Ứng dụng của định lý này bảo đảm với chúng ta rằng hàm ước lượng của g mà chúng ta đã rút ra ở trên là nhất quán.

Nếu chúng ta muốn dự báo các giá trị tương lai của biến phụ thuộc, thì chúng ta phải đặc biệt cẩn thận. Phương trình của chúng ta là phương trình này:

$$\log(Y_t) = \beta_1 + \beta_2 * t + \varepsilon_t$$

Nếu chúng ta lấy hàm số mũ của cả hai vế, ta có:

$$Y_t = e^{\beta_1 + \beta_2 * t + \varepsilon} = e^{\beta_1 + \beta_2 * t} e^{\varepsilon_t}$$

Có thể chỉ ra rằng $E[e^{\varepsilon_t}] = e^{\sigma^2/2} \neq 1$, vì vậy nếu chúng ta muốn thử dự báo Y_t bằng cách sử dụng

$$\hat{Y}_t = e^{\hat{\beta}_1 + \hat{\beta}_2 * t}$$

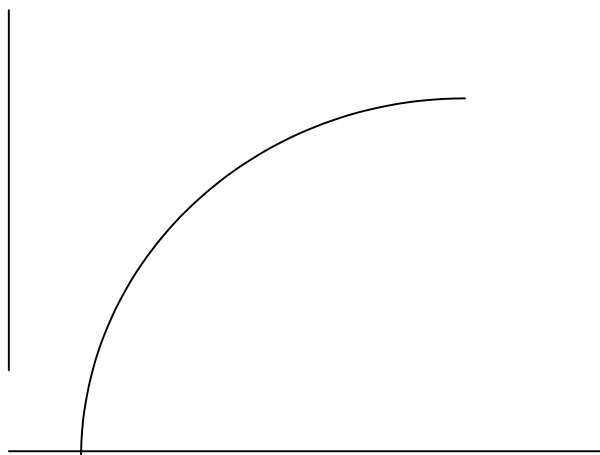
thì dự báo là chệch. Phương trình dự báo chính xác là phương trình này:

$$Y_t = e^{\beta_1 + \beta_2 * t + \sigma^2/2}$$

EViews tự động tạo ra các dự báo phù hợp (xem Bài tập 13). Hãy nhớ rằng các mô hình chuỗi thời gian thường yêu cầu các kỹ thuật đặc biệt ngoài phạm vi của khóa học này.

Các mô hình tuyến tính - Logarit

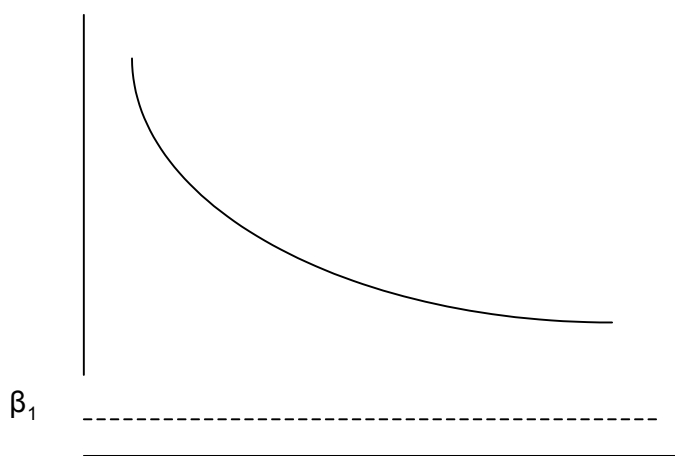
Mối quan hệ $Y = \beta_1 + \beta_2 \log(X)$ tạo nên một đồ thị trong X-Y trông giống như thế này:



Đồ thị này có thể là phù hợp p với đồ thị thu nhập-tiêu dùng của một hàng hoá thông thường.

Các mô hình nghịch đảo

Tương quan $Y = \beta_1 + \beta_2 \left(\frac{1}{X} \right)$ tạo ra một đồ thị trong X-Y trông giống như thế này:



Đồ thị này có thể là phù hợp p với một đường cầu hay một đường chi phí.

Lựa chọn dạng hàm số

Tồn tại nhiều khả năng. Hướng dẫn tốt nhất là kiểm tra các đồ thị, các tác động biên, và các độ co giãn được tạo ra bởi các đặc trưng mô hình thay thế và chọn ra một đặc trưng thích hợp theo lý thuyết kinh tế có liên quan.

So sánh R^2 giữa các mô hình

Để so sánh R^2 giữa các mô hình, ESS và RSS của hai mô hình này phải được đo theo cùng đơn vị. Vì thế, so sánh trực tiếp các R^2 của hai mô hình này là không phù hợp:

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

$$\log(Y_i) = \beta_1 + \beta_2 X_i + \varepsilon_i$$

Một kỹ thuật để tạo ra một R^2 có thể so sánh dành cho mô hình thứ hai là làm việc sau:

$$\hat{Y}_i = \exp \left[\log(Y_i) + \frac{s^2}{2} \right]$$

Hãy lưu ý sự điều chỉnh cho thiên lệch này!

Cuối cùng, hãy tính hệ số tương quan mẫu bình phương giữa Y_i và \hat{Y}_i : Việc này cho ta một R^2 tương hợp giá trị được tạo ra cho mô hình hồi quy thứ nhất.

Tính đa cộng tuyến

Hoàn toàn không có tính đa cộng tuyến

Trên thực tế, hiếm có các tập hợp biến hồi quy mà giữa chúng không có tương quan tuyến tính; chúng ta sẽ không xét tiếp trường hợp này.

Đa cộng tuyến hoàn hảo

Khi một tập hợp các biến hồi quy có tương quan tuyến tính hoàn hảo, thì thường là do một sai lầm của nhà kinh tế lượng: có thể là nhà kinh tế lượng này đã xây dựng nên một biến từ các biến khác trong dữ liệu theo cách dẫn tới tính đa cộng tuyến hoàn hảo

Tính đa cộng tuyến cao

Tính đa cộng tuyến cao là tình thế ta quan tâm mà đôi khi xuất hiện trên thực tế. Hãy trở lại với biểu thức mà chúng ta đã sử dụng để thể hiện tính đa cộng tuyến hoàn hảo trong Bài giảng 11:

$$\lambda_1 + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \dots + \lambda_K X_{Ki} = 0$$

Một cách tương đương để viết biểu thức này là một biến hồi qui cụ thể là một hàm tuyến tính của các biến khác. Không làm mất tính tổng quát, chúng ta có thể giả định rằng biến hồi qui cuối cùng X_K là một hàm tuyến tính của các biến khác:

$$X_{Ki} = \gamma_1 + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \dots + \gamma_{K-1} X_{(K-1)i}$$

Như chúng ta đã nêu trong ngày Thứ Hai, sự tồn tại của tính đa cộng tuyến hoàn hảo làm cho không thể tìm được các lời giải trong phép bình phương tối thiểu. Tuy nhiên, nếu tính đa cộng tuyến không hoàn hảo, thì vẫn có thể tìm được các lời giải.

Sự tồn tại của tính đa cộng tuyến cao làm tăng các phương sai (và các sai số chuẩn) của các hàm ước lượng bình phương tối thiểu. Do đó, một chỉ số cho sự hiện diện của tính đa cộng tuyến cao là sự kết hợp của R^2 cao với trị thống kê t thấp. Trong các trường hợp p thái cực, những hệ số này tỏ ra rất nhạy trước việc bỏ một quan sát hay trước các thay đổi nhỏ trong đặc trưng mô hình. Điều này là do máy vi tính không thể thực hiện được các tính toán một cách chuẩn xác và những kết quả bị lấn át bởi sai số làm tròn.

Tính đa cộng tuyến không hoàn hảo có thể được trình bày theo cách y như tính đa cộng tuyến hoàn hảo, nhưng với một thành phần sai số được bổ sung thêm vào phương trình:

$$X_{Ki} = \gamma_1 + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \dots + \gamma_{K-1} X_{(K-1)i} + v_i$$

Điều này gợi ý một cách để phát hiện tính đa cộng tuyến sẽ là lần lượt chạy một phép hồi qui phụ trên mỗi biến hồi qui. Nếu một hoặc nhiều hơn các phép hồi qui tạo ra một R^2 rất cao thì chúng ta có bằng chứng về tính đa cộng tuyến cao.

Các phép hồi qui phụ:

$$X_{Ki} = \gamma_1 + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \dots + \gamma_{K-1} X_{(K-1)i} + v_{Ki} \Rightarrow R_K^2$$

$$X_{(K-1)i} = \gamma_1 + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \dots + \gamma_{(K-2)} X_{(K-2)i} + \gamma_K X_{Ki} + v_i \Rightarrow R_{K-1}^2$$

⋮

$$X_{2i} = \gamma_1 + \gamma_3 X_{3i} + \dots + \gamma_K X_{Ki} + v_i \Rightarrow R_2^2$$

Chúng ta có thể làm gì với tính đa cộng tuyến?

Đôi khi các sinh viên thử tìm biến hồi qui với R^2 phụ cao nhất và bỏ nó đi. Đây là một qui trình nguy hiểm, vì biến này có thể là phù hợp p và việc bỏ nó đi có thể gây ra thiên lệch do bỏ biến.

Có thể có các giải pháp khác: người ta có thể chỉnh lại đặc trưng mô hình theo cách nào đó. Ví dụ, nếu người ta đang sử dụng dữ liệu ở dạng các mức chung, thì việc thay đổi sang các giá trị theo đầu người có thể giúp ích.

Tuy nhiên, nói chung, tính đa cộng tuyến đơn giản là một đặc trưng của dữ liệu. Nếu nó là một nét đặc trưng mạnh, thì nó có thể làm cho một số biến phù hợp p tỏ ra không có ý nghĩa thống kê. Lời khuyên của tôi trong những trường hợp như vậy là giữ lại tất cả những biến hồi qui này trong mô hình đang xét, nhưng cũng báo cáo các kết quả của phép hồi qui phụ mạnh nhất.

Các biến giả

Giả sử rằng Anh/Chị phải ước lượng mối tương quan giữa tiền lương của các giáo sư và số năm công tác của họ.

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

Y_i = tiền lương hàng năm của giáo sư i .

X_i = số năm công tác của giáo sư i .

Bây giờ giả sử rằng Anh/Chị muốn điều tra xem liệu có phải các giáo sư nữ bị phân biệt đối xử trong tiền lương của họ không.

Một cách để tiến hành sẽ là ước lượng hai phép hồi qui riêng biệt: một cho các giáo sư nam trong mẫu của Anh/Chị và một cho các giáo sư nữ trong mẫu của Anh/Chị. Có hai khó khăn trong cách giải quyết này. Một là, mỗi phép hồi qui là kém hiệu quả hơn so với phép hồi qui tổng hoà sử dụng toàn bộ dữ liệu. Hai là, tương đối không thuận tiện khi kiểm định liệu có phải hai phép hồi qui này khác nhau không.

Giải pháp cho thách thức ước lượng này là xác định một biến mới ghi nhận sự có mặt hay vắng mặt của thuộc tính "nữ".

$D_i = 0$ nếu quan sát i thuộc về một giáo sư nam

$D_i = 1$ nếu quan sát i thuộc về một giáo sư nữ

Bây giờ xác định mô hình sau :

$$Y_i = \beta_1 + \beta_2 X_i + \delta D_i + \varepsilon_i$$

Hãy ghi nhận rằng các bậc tự do của mô hình này là $n_M + n_F - 3$. Sự có mặt của biến giả này làm không thể thể hiện hai phép hồi qui trong một mô hình. Xét các giá trị kỳ vọng có điều kiện:

$$E[Y_i | D_i = 0] = \beta_1 + \beta_2 X_i \quad \text{nam}$$

$$E[Y_i | D_i = 1] = (\beta_1 + \delta) + \beta_2 X_i \quad \text{nữ}$$

Chúng ta thấy rằng nếu δ là dương thì các giáo sư nữ có một hàm hồi qui tổng thể có tung độ gốc cao hơn so với các giáo sư nam.

Hệ số δ được gọi là *chênh lệch tung độ gốc* và nó cho thấy chênh lệch của các tung độ gốc đối với hai mẫu phụ. Dễ dàng kiểm định xem liệu chênh lệch này có ý nghĩa thống kê không: đơn giản là đánh giá trị thống kê t kèm với δ .

Tất nhiên, lương khởi điểm có thể không phải là một nguồn gốc duy nhất của tình trạng phân biệt đối xử. Có thể là độ dốc của các phép hồi qui cũ khác nhau. Chúng ta có thể đưa khả năng này vào bằng cách xác định một *biến tương tác* như sau :

$$DX_i = D_i \times X_i$$

Bổ sung biến này vào phép hồi qui của chúng ta và ta có:

$$Y_i = \beta_1 + \beta_2 X_i + \delta_1 D_i + \delta_2 DX_i + \varepsilon_i$$

Các kỳ vọng có điều kiện giờ đây trở thành

$$E[Y_i | D_i = 0] = \beta_1 + \beta_2 X_i \quad \text{nam}$$

$$E[Y_i | D_i = 1] = (\beta_1 + \delta_1) + (\beta_2 + \delta_2) X_i \quad \text{nữ}$$

Hệ số δ_2 được gọi là *chênh lệch độ dốc* vì nó là chênh lệch giữa các độ dốc trong những phép hồi qui đối với hai mẫu phụ của chúng ta. Dễ dàng kiểm định rằng liệu chênh lệch này có ý nghĩa thống kê hay không. Chúng ta chỉ cần đánh giá mức độ ý nghĩa của trị thống kê t kèm với ước lượng của δ_2 .

Kiểm định xét xem liệu giới tính có bất cứ tác động nào không có thể được thực thi với kiểm định Wald cho giả thuyết sau :

$$H_0 : \delta_1 = \delta_2 = 0$$

H_1 : không phải cả hai cùng là zero

Các biến giả đối với các chủng loại đa biến

Giả sử rằng Anh/Chị quyết định mở rộng nghiên cứu của mình và ước lượng xem biến thiên bao nhiêu trong tiền lương là do biến thiên trong trình độ học vấn và rằng mẫu của Anh/Chị có các cá nhân có các bằng đại học, bằng thạc sĩ, và bằng tiến sĩ.

Anh/Chị có thể mã hoá những bằng này ra sao với các biến giả?

Các sinh viên đôi khi gợi ý điều sau:

D_i	=	0	đối với bằng đại học
D_i	=	1	đối với bằng thạc sĩ
D_i	=	2	đối với bằng tiến sĩ

Khó khăn với đặc trưng mô hình này là chênh lệch giữa bằng đại học và bằng tiến sĩ lớn đúng gấp đôi chênh lệch giữa bằng đại học và bằng thạc sĩ; còn chênh lệch giữa bằng thạc sĩ và bằng tiến sĩ đúng bằng chênh lệch giữa bằng đại học và bằng thạc sĩ. Để thấy được điều này, Anh/Chị cần tìm các kỳ vọng có điều kiện như chúng ta đã làm trước đây. Giới hạn này có thể không tương hợp với dữ liệu của chúng ta và nó là không cần thiết.

Xét con đường thay thế sau: hãy chỉ ra một cặp biến giả. Chúng ta thấy rằng chúng có ba cấu hình xác định một cách duy nhất ba kết quả học vấn này.

D_{1i}	D_{2i}	
0	0	Đại học

1	0	Thạc sĩ
0	1	Tiến sĩ

Với ba chủng loại, các sinh viên đôi khi thử sử dụng ba biến giả như thế này :

D_{1i}	D_{2i}	D_{3i}	
1	0	0	Đại học
0	1	0	Thạc sĩ
0	0	1	Tiến sĩ

Khó khăn với chiến lược này là nó tạo ra tính đa cộng tuyến hoàn hảo giữa ba biến giả và biến $X_1 = 1$ đại diện cho hằng số :

$$D_{1i} + D_{2i} + D_{3i} - X_{1i} = 0 \text{ đối với mọi } i.$$

Điều này được gọi là "bẫy biến giả" và nó cung cấp một ví dụ cho điểm được nêu trước đây rằng sự đa cộng tuyến hoàn hảo thường được tạo ra một cách ngẫu nhiên bởi nhà kinh tế lượng.

