Mô hình với biến phụ thuộc bị giới hạn (Models with Limited Dependent Variables)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

21/04/2020

Các loại hình biến phụ thuộc bị giới hạn

- Đơn giản nhất là biến phụ thuộc là biến xác suất xảy ra một sự kiện, có hoặc không xảy ra.
  - Doanh nghiệp có bị phá sản hay không; có vay tiền ngân hàng không.
- Biến phụ thuộc thể hiện hành vi lựa chọn trong mô hình đa lựa chọn:
  - Lựa chọn smartphone thương hiệu gì trong số các mặt hàng bán trên thị trường: Apple, Samsung, LG, Xiaomi, Oppo...
- Biến phụ thuộc là biến xếp hạng/thứ tự:
  - Xếp hạng một bộ phim từ: rất kém, kém, trung bình, hay, rất hay.
- Biến phụ thuộc là số lần xảy ra một sự kiện:
  - Số lần một người vi phạm hành vi bạo lực gia đình, số lần đi khám bệnh một năm.
- Biến phụ thuộc có giá trị bị chặn dưới hoặc chặn trên:
  - Tiền lương từ các điều tra thu nhập bị chặn dưới ở 0 đồng; số giờ làm việc một tuần không vượt quá $7 * 24 = 168$ giờ.

Tại sao kiểm soát vấn đề biến phụ thuộc bị giới hạn rất quan trọng?

- Không thỏa các giả định của mô hình hồi quy tuyến tính cổ điển CLRM $\Rightarrow$ Ước lượng có thể gặp một hoặc nhiều các vấn đề sau:
  - Phương sai của sai số thay đổi
  - Ước lượng bị chệch
  - Ước lượng không nhất quán
  - Ước lượng không hiệu quả
- Để hiểu xảy ra vấn đề gì thì phải dựa vào hiểu biết của dữ liệu và lý thuyết để giải thích.
- Lựa chọn khi phải đối phó với biến phụ thuộc bị giới hạn:
  - Tiếp tục sử dụng OLS và chấp nhận các vấn đề có thể gặp phải.
  - Sử dụng phương pháp phù hợp với dữ liệu.

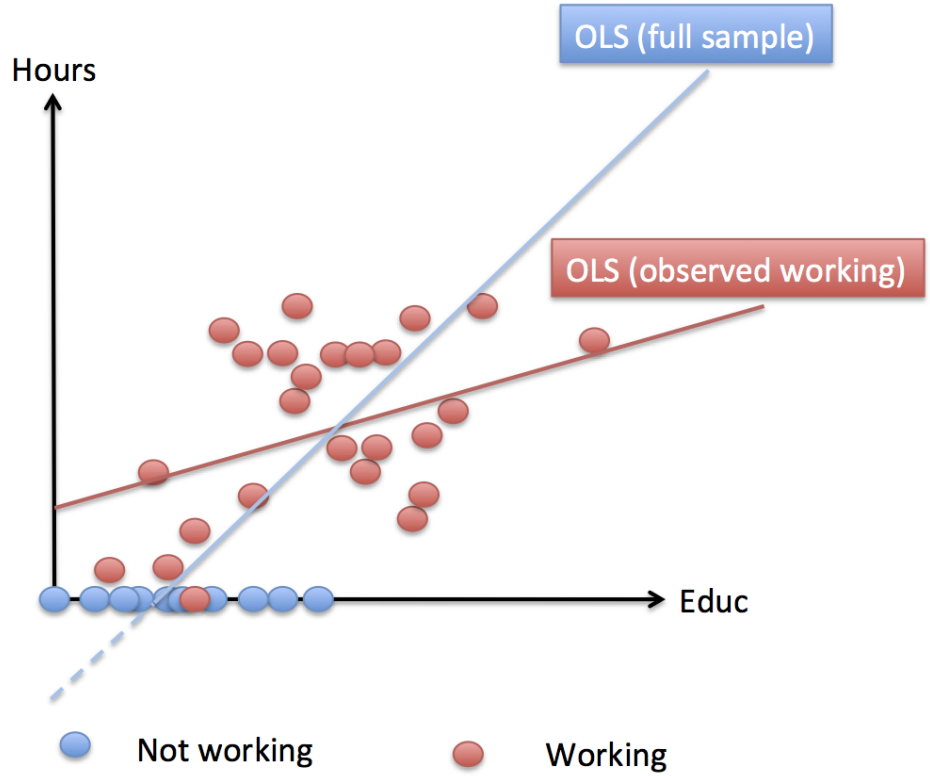
Các mô hình tương ứng với các loại biến phụ thuộc bị giới hạn

- Mô hình xác suất: **LPM, Logit, Probit**
- Mô hình đa lựa chọn: Multinomial logit/probit, conditional logit
- Mô hình biến xếp hạng: Ordered logit/probit
- Mô hình số lần xảy ra một sự kiện: Poisson count model
- Mô hình biến phụ thuộc bị chặn: **Tobit model for censored data**
- Mô hình với dữ liệu xảy ra hiện tượng lựa chọn mẫu: **Sample selection/Heckman correction model**

Khái niệm biến phụ thuộc bị chặn (censored data)

- Biến tiền lương bị chặn dưới bởi giá trị 0 đối với những người chưa đi làm, về hưu, hay đang thất nghiệp. Các giá trị quan sát được là dương.
- Rất nhiều biến số kinh tế bị chặn dưới bởi giá trị 0, ví dụ:
  - Số giờ lao động của phụ nữ đã có gia đình.
  - Số tiền làm từ thiện của một người trong một năm.
  - Số lít rượu bia một người uống trong một năm.
  - Chi tiêu cho hàng hoá xa xỉ của hộ gia đình trong dịp lễ tết.
  - Thời gian thất nghiệp của một người lao động.
- Dữ liệu có thể bị chặn trên hoặc chặn dưới do cách thức điều tra dữ liệu.
  - Tiền lương điều tra theo các mốc 0-10 triệu, 10-20,..., và trên 100 triệu/tháng. Những người thu nhập trên 100 triệu sẽ gom lại thành cùng một nhóm.

Hồi quy OLS của số giờ đi làm trong năm



- Biến phụ thuộc bị chặn dưới tại 0.
- Ước lượng bằng OLS với nhóm làm việc có thể bị thiên lệch giảm (downward bias) do bỏ qua nhóm không làm việc.
- Ước lượng OLS với toàn bộ dữ liệu gặp phải vấn đề số giờ làm việc âm tương tự như mô hình xác suất tuyến tính LPM.

Các cách xử lý biến phụ thuộc bị chặn

- Cách 1: ước lượng mô hình Logit/Probit với biến phụ thuộc là có làm việc hay không. Tuy nhiên cách làm này chỉ ước lượng được xác suất có làm việc hay không (biến định tính rời rạc), nhưng không ước lượng được tác động của biến giải thích lên số giờ làm việc của những người đi làm như thế nào (biến định lượng liên tục).

- Cách 2: mô hình Tobit xử lý được cả hai vấn đề trên.

Mô hình Tobit với biến phụ thuộc bị chặn

Bản chất của mô hình Tobit là hồi quy hai bước theo tuần tự:

- Bước 1: Ước lượng xác suất quan sát được một người có tham gia lao động hay không bằng hồi quy xác suất MLE.

- Bước 2: Ước lượng các nhân tố ảnh hưởng đến biến phụ thuộc (ví dụ số giờ lao động) bằng OLS, và điều chỉnh hệ số ước lượng có tính đến xác suất có đi làm hay không đã thực hiện ở bước 1.

Xây dựng mô hình Tobit

Thông thường hành vi làm việc của một người được diễn giải bởi hàm ẩn:

$$y^* = X * \beta + u, \qquad u \sim N(0, \sigma^2)$$

trong đó y^* là biến phụ thuộc ẩn (latent variable), không quan sát được. Chúng ta quan sát được biến y là số giờ làm việc trong năm:

$$y = max(0, y^*)$$

- Chúng ta quan sát được $y > 0$ đối với những người đi làm.
- Với những người không đi làm, $y = 0$.

Xây dựng mô hình Tobit

Chúng ta có thể tìm được phương trình ước lượng của biến phụ thuộc là trung bình có quyền số của xác xuất đi làm và số giờ đi làm:

$$E[y|x] = \underbrace{P(y=0|x) * E[y=0|x]}_{=0} + \underbrace{P(y>0|x) * E[y|y>0,x]}_{>0}$$

trong đó:

$$P(y>0|x) = P(X*\beta + u > 0) = P(\frac{u}{\sigma} > -\frac{X*\beta}{\sigma}) = \Phi(\frac{X*\beta}{\sigma})$$

với $\Phi(.)$ là hàm tích lũy phân phối chuẩn, được tính tại giá trị $\frac{X*\beta}{\sigma}$

Xây dựng mô hình Tobit

Ngoài ra, chúng ta có biểu thức sau (học viên tự chứng minh):

$$E[y|y > 0, x] = X * \beta + \sigma\lambda(\frac{X * \beta}{\sigma})$$

với $\lambda(c) = \frac{\phi(c)}{\Phi(c)}$, còn được gọi là tỷ số Mills nghịch đảo (inverse Mills ratio - IMR), là tỷ lệ giữa hàm mật độ và hàm tích lũy của phân phối chuẩn được tính tại giá trị c.

Xây dựng mô hình Tobit

Từ các công thức trên, chúng ta có phương trình hàm hồi quy Tobit như sau:

$$E[y|x] = \Phi(\frac{X * \beta}{\sigma}) * X * \beta + \sigma\phi(\frac{X * \beta}{\sigma})$$

So sánh với hồi quy OLS:

$$E[y|x] = X * \beta$$

- Hồi quy Tobit là hàm phi tuyến của các tham số và biến giải thích thông qua hàm tích lũy và phân phối xác suất.
- Có thể chứng minh (!) là giá trị dự báo của biến phụ thuộc của hàm Tobit là dương với mọi giá trị của X, khác so với hồi quy OLS có thể nhận giá trị dự báo âm.

Ước lượng mô hình Tobit và diễn giải ý nghĩa

- Mô hình Tobit được ước lượng bằng phương pháp MLE thay vì OLS.
- Diễn giải sự khác biệt của các hệ số ước lượng:
  - Với OLS thì β là tác động biên của các biến giải thích lên biến phụ thuộc và không đổi.

$$\frac{\partial E[y|x]}{\partial x_j} = \beta_j$$

  - Với Tobit thì chúng ta phải tính tác động biên từ phương trình hàm hồi quy bằng đạo hàm bậc nhất của biến phụ thuộc theo biến giải thích.

$$\frac{\partial E[y|x]}{\partial x_j} = \frac{\partial \left[\Phi(\frac{X*\beta}{\sigma}) * X * \beta + \sigma\phi(\frac{X*\beta}{\sigma}) \right]}{\partial x_j} = ?$$

Tác động biên trong mô hình Tobit

- Nếu biến giải thích là biến liên tục, chứng minh công thức sau bằng quy tắc đạo hàm chuỗi:

$$\frac{\partial E[y|x]}{\partial x_j} = \Phi(\frac{X * \beta}{\sigma}) * \beta_j$$

- Nếu biến giải thích là biến rời rạc x_0, x_1:

$$\Delta y = E[y|x_1] - E[y|x_0]$$

- Tác động biên của mô hình Tobit sẽ phụ thuộc vào giá trị tham chiếu thông qua xác suất quan sát được một cá nhân có tham gia lao động hay không $\Phi(\frac{X*\beta}{\sigma})$.
- Tương tự như hồi quy Logit/Probit, $\Phi(\frac{X*\beta}{\sigma})$ được tính tại các giá trị đặc trưng như trung bình, các tứ phân vị... của các biến giải thích.
- Tác động biên cũng phụ loại vào phân loại biến (liên tục hay rời rạc).

Ví dụ 1: Sử dụng bộ dữ liệu Labor.dta và ước lượng hàm cung lao động của phụ nữ đã có gia đình

Giả sử chúng ta muốn ước lượng mô hình hàm cung số giờ lao động như sau:

$$\begin{aligned} hours =& \beta_0 + \beta_1 netincome + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 age \\ &+ \beta_6 KIDS6 + \beta_7 KIDS7 + u \end{aligned}$$

với $KIDS6$ và $KIDS7$ là số con dưới 6 tuổi và từ 6-18 tuổi. Trong mẫu có 325/753 quan sát có số giờ làm việc bằng 0.

So sánh ước lượng OLS và Tobit thế nào?

	OLS b/se	Tobit b/se
main		
netincome	-3.4466	-8.8142*
	(2.5440)	(4.4591)
educ	28.7611*	80.6456***
	(12.9546)	(21.5832)
exper	65.6725***	131.5643***
	(9.9630)	(17.2794)
expersq	-0.7005*	-1.8642***
	(0.3246)	(0.5377)
age	-30.5116***	-54.4050***
	(4.3639)	(7.4185)
KIDS6	-442.0899***	-894.0217***
	(58.8466)	(111.8779)
KIDS7	-32.7792	-16.2180
	(23.1762)	(38.6414)
Constant	1330.4824***	965.3053*
	(270.7846)	(446.4358)
sigma		
Constant		1122.0217***
		(41.5790)
N	753	753

* p<0.05, ** p<0.01, *** p<0.001

Ước tính tác động biên

- Tác động biên của việc học thêm một năm lên số giờ lao động của phụ nữ, tại giá trị trung bình của các biến giải thích, là 80.65*.645 = 52 giờ. Ước lượng OLS là 28.76 giờ.
- Tác động biên lên số giờ lao động của phụ nữ chưa có con nhỏ dưới 6 tuổi so với có một con dưới 6 tuổi, tại giá trị trung bình của các biến giải thích khác, là 503.5 giờ.
- Chỉ giới hạn vào 428 phụ nữ đang tham gia lao động, ước lượng OLS và Tobit cho kết quả giống nhau.

Tổng kết mô hình Tobit

- Khi dữ liệu quan sát được bị chặn tại một ngưỡng giá trị nào đó thì ước lượng OLS có thể bị chệch hoặc gặp phải vấn đề dự báo không chính xác.
- Sử dụng mô hình Tobit và phương pháp MLE có thể sửa được lỗi của mô hình OLS.
- Diễn giải ý nghĩa của các tham số của mô hình Tobit phức tạp hơn mô hình OLS do giá trị dự báo là hàm phi tuyến của các biến giải thích và tham số ước lượng – tương tự như hàm hồi quy xác suất Logit hoặc Probit.

Học viên cần phân biệt hai tình huống và hai cách thức xử lý khác nhau đối với mỗi tình huống:

- Chỉ áp dụng hồi quy Tobit với dữ liệu bị chặn (có nghĩa là dữ liệu tồn tại, nhưng do quá trình thu thập hay tạo dữ liệu khiến dữ liệu thu thập được bị chặn tại một ngưỡng quan sát nào đó).

- Khi dữ liệu gặp phải vấn đề tự lựa chọn mẫu (ví dụ không quan sát được một số cá nhân có các thuộc tính nhất định) thì cần sử dụng hàm hồi quy điều chỉnh mẫu (Heckman sample selection model – phần sau) .

- Mô hình Tobit giải quyết vấn đề dữ liệu tồn tại nhưng bị thiếu thông tin. Mô hình Heckman sample selection giải quyết vấn đề không có hoặc không quan sát được dữ liệu. Do đó, nhà nghiên cứu phải thực sự hiểu dữ liệu và sử dụng giả định hợp lý khi đề xuất mô hình.

Mô hình với dữ liệu không ngẫu nhiên (Models with non-random sample/ sample selection)

Khái niệm dữ liệu không ngẫu nhiên/Vấn đề tự lựa chọn mẫu

- Do cách thiết kế mẫu khiến dữ liệu bị mất hoặc thiếu một cách hệ thống.
- Do dữ liệu bị thiếu một số thông tin nhất định.
- Do cách thiết kế chính sách dẫn đến chỉ quan sát được những nhóm đối tượng nhất định.

Hiệu lực nội tại khi xảy ra vấn đề lựa chọn mẫu

Giả sử chúng ta có mô hình hồi quy của thu nhập y theo các biến giải thích x:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$$

thỏa các điều kiện của mô hình CLRM và $E[u|x_1, ..., x_k] = 0$

- Nếu chúng ta quan sát được toàn bộ mẫu dữ liệu $\Rightarrow$ Ước lượng OLS không chệch và nhất quán.
- Khi dữ liệu bị thiếu:
  - Dữ liệu bị thiếu ngẫu nhiên?
  - Dữ liệu bị thiếu không ngẫu nhiên?

- Thiếu ngẫu nhiên: Ước lượng OLS đảm bảo hiệu lực nội tại, nhưng độ tin cậy của ước lượng sẽ bị giảm.

- Thiếu không ngẫu nhiên: Ước lượng bằng OLS **có thể** bị chệch và không có hiệu lực nội tại. Cần hiểu rõ bản chất của dữ liệu!!

Dữ liệu không ngẫu nhiên do quá trình chọn mẫu dựa trên biến giải thích

Xảy ra trong quá trình thiết kế hay điều tra mẫu, ví dụ chỉ điều tra những người làm việc ở HCM, hay có bằng cấp cao nhất không quá phổ thông trung học.

- Không ảnh hưởng đến hiệu lực nội tại, nhưng có thể ảnh hưởng đến hiệu lực ngoại vi.

- Ví dụ: Mô hình dựa trên điều tra thu nhập và tình trạng học vấn của nhóm cá nhân học không quá 12 năm sẽ không thể áp dụng cho nhóm học đại học hoặc cao hơn.

Dữ liệu không ngẫu nhiên do quá trình chọn mẫu xảy ra trên biến phụ thuộc

Xảy ra do không thể quan sát được hay quan sát không đủ dữ liệu.

- Ảnh hưởng đến hiệu lực nội tại.

- Ví dụ: Ước lượng hàm tiền lương của người trong độ tuổi lao động. Những người không đi làm (do đó tiền lương bằng không hoặc không được ghi nhận) có thể do nhiều lý do (tiền lương thấp hơn kỳ vọng, hoặc có lựa chọn khác). Nếu không xử lý vấn đề chọn mẫu thì ước lượng sẽ bị sai lệch.

Xử lý khi dữ liệu không ngẫu nhiên

Cần hiểu rõ bản chất của dữ liệu và nguồn gốc của vấn đề lựa chọn mẫu thì mới có thể đề xuất cách thức xử lý phù hợp!

- Nếu giả định những người không đi làm nhận mức lương bằng 0 $\Rightarrow$ Mô hình **Tobit** với biến phụ thuộc bị chặn dưới.

- Nếu giả định những người không đi làm là do có những lựa chọn khác tốt hơn (ví dụ làm tư, do đó không báo cáo thu nhập trong bảng câu hỏi tiền lương). Mặc dù những người này không được ghi nhận có thu nhập nhưng trên thực tế họ vẫn có thu nhập $\Rightarrow$ Dùng mô hình hồi quy điều chỉnh vấn đề lựa chọn mẫu **Heckman selection model/Heckit method**.

Ví dụ vấn đề chọn mẫu khi ước lượng hàm tỷ suất thu nhập của việc đi học

Chúng ta có thông tin của những người đi làm công ăn lương và có báo cáo thu nhập. Nhưng toàn bộ dữ liệu điều tra bao gồm cả những người trong độ tuổi lao động nhưng không báo cáo thu nhập do làm tư, kinh doanh tiểu thương.

- Nếu chỉ giới hạn ở mẫu dữ liệu những người đang đi làm và có thu nhập dương $\Rightarrow$ OLS có thể chệch và không nhất quán bởi nó bỏ qua những nhóm đối tượng có thu nhập nhưng không báo cáo.

- Nếu chúng ta đưa toàn bộ dữ liệu (gồm cả những người không báo cáo thu nhập) vào mô hình thu nhập $\Rightarrow$ Xử lý thế nào với những người không báo cáo thu nhập?

$\Rightarrow$ Chúng ta cần điều chỉnh hàm hồi quy để phản ánh vấn đề lựa chọn vào tham gia lực lượng lao động chính thức và có báo cáo thu nhập.

Xây dựng mô hình điều chỉnh vấn đề lựa chọn mẫu

Mô hình lựa chọn mẫu được viết dưới dạng **hệ phương trình cấu trúc**, bao gồm một phương trình diễn giải hành vi và một phương trình diễn giải vấn đề lựa chọn mẫu:

$$\begin{cases} y &= X\beta + u \\ s &= 1[Z\gamma + v \geq 0] \end{cases}$$

trong đó $E[u|X] = 0$, X là các biến giải thích của phương trình hành vi y, Z là các biến giải thích trong phương trình lựa chọn mẫu s.

Ý nghĩa của phương trình lựa chọn mẫu s

Phương trình lựa chọn được biểu diễn dưới dạng **hàm chỉ số (index function)** của các biến giải thích Z, mục đích để giải thích tại sao một số quan sát nằm trong mẫu nghiên cứu (ví dụ có thu nhập) còn những người khác nằm ngoài mẫu (không có thu nhập).

$$s = \begin{cases} 1 & \textit{if} \quad Z\gamma + v \geq 0 \\ \\ 0 & \textit{otherwise} \end{cases}$$

- Nếu $Z_i\gamma + v \geq 0 \Rightarrow s_i = 1$, có nghĩa là chúng ta quan sát được cá nhân i trong phương trình hành vi (cá nhân i có thu nhập).
- Nếu $s_i = 0$ có nghĩa là chúng ta không có cá nhân i trong phương trình hành vi (cá nhân i không có thu nhập).

Ý nghĩa của phương trình hành vi y

Với điều kiện quan sát được cá nhân có thu nhập thì phương trình hành vi ước lượng tác động của các nhân tố X ảnh hưởng như thế nào đến thu nhập y.

$$y = \quad X\beta + u$$

Phương trình hành vi chỉ áp dụng với các cá nhân lựa chọn vào trong mẫu (tức là các cá nhân đi làm chính thức và báo cáo thu nhập dương) chứ không áp dụng cho toàn bộ quần thể những người trong độ tuổi lao động.

Các bước xây dựng và ước lượng mô hình hồi quy điều chỉnh vấn đề lựa chọn mẫu

Bắt đầu bằng hệ phương trình cấu trúc:

$$\begin{cases} y &= X\beta + u \\ s &= 1[Z\gamma + v \geq 0] \end{cases}$$

Bỏ qua các bước biến đổi trung gian (học viên tự chứng minh), chúng ta có công thức của phương trình hành vi y với điều kiện quan sát được các cá nhân nằm trong mẫu là:

$$E[y|Z, s = 1] = X\beta + \rho\lambda(Z\gamma)$$

trong đó:

- $\lambda(.)$ là tỷ số Mills nghịch đảo (Mills Inverse Ratio-IMR), được tính tại giá trị $Z\gamma$.

$$\lambda(Z\gamma) = \frac{\phi(Z\gamma)}{\Phi(Z\gamma)}$$

  $\phi(.)$ và $\Phi(.)$ là hàm mật độ và hàm tích lũy phân phối chuẩn.
- $\lambda(Z\gamma)$ được coi như một biến giải thích phụ đưa vào để điều chỉnh vấn đề chọn mẫu.
- X là các biến giải thích trong mô hình cấu trúc.
- β và ρ là tham số cần ước lượng của phương trình hành vi có điều kiện.
- γ là tham số cần ước lượng của phương trình lựa chọn mẫu.

Tóm lại, chúng ta cần ước lượng phương trình hành vi có điều kiện (conditional expectation function):

$$E[y|Z, s = 1] = X\beta + \rho\lambda(Z\gamma)$$

với các đặc tính sau:

- Các tham số của mô hình hành vi có điều kiện là β và ρ.
- Các biến giải thích là X và tỷ số $\lambda(Z\gamma)$.

Do $\lambda(Z\gamma)$ phụ thuộc vào các tham số γ nên chúng ta phải ước lượng phương trình lựa chọn mẫu trước để tìm γ.

Heckman sample selection model

Bản chất của phương pháp điều chỉnh mẫu (các tên khác: hồi quy khi xảy ra vấn đề lựa chọn mẫu, phương pháp Heckman sample correction, phương pháp Heckit) là ước lượng phương trình hành vi có điều kiện bằng hồi quy hai bước:

1. Ước lượng phương trình tự lựa chọn mẫu s để tính $\lambda(Z\gamma)$.

2. Đưa $\lambda(Z\gamma)$ vào trong phương trình hành vi có điều kiện $E[y|Z, s = 1]$ như một biến giải thích nhằm điều chỉnh vấn đề lựa chọn mẫu. Ước lượng tham số cấu trúc từ bước 2 sẽ có hiệu lực nội tại.

Ước lượng hồi quy điều chỉnh mẫu bằng hồi quy hai giai đoạn

1. Ước lượng mô hình lựa chọn mẫu (cá nhân có thu nhập hay không) bằng hồi quy Probit để ước lượng các tham số γ, và sử dụng toàn bộ bộ dữ liệu của những người trong độ tuổi lao động:

$$P(s = 1|Z) = \Phi(Z\gamma + v)$$

Tính giá trị $\widehat{\lambda(Z\gamma)}$ bằng công thức:

$$\widehat{\lambda(Z\gamma)} = \frac{\phi(Z\hat{\gamma})}{\Phi(Z\hat{\gamma})}$$

Tương tự như phương pháp 2SLS/IV, **phải có ít nhất một biến ngoại sinh trong Z nhưng không thuộc X** (biến chỉ ảnh hưởng đến việc cá nhân có đi làm và có thu nhập chính thức chứ không ảnh hưởng đến thu nhập là bao nhiêu).

Ước lượng hồi quy điều chỉnh mẫu bằng hồi quy hai giai đoạn

2. Ước lượng mô hình hành vi có điều kiện bằng OLS, với dữ liệu trong mẫu (chỉ những cá nhân có thu nhập chính thức), với các biến giải thích X và $\widehat{\lambda(Z\gamma)}$ được tính ở bước 1:

$$y = X\beta + \rho\widehat{\lambda(Z\gamma)} + u$$

Ước lượng hồi quy điều chỉnh mẫu bằng hồi quy hai giai đoạn

- Bản chất của phương pháp Heckit là chúng ta đưa thêm một biến giải thích là tỷ số IMR được tính từ phương trình chọn mẫu vào hồi quy OLS của phương trình hành vi có điều kiện.
- X tác động lên biến phụ thuộc thu nhập trong phương trình hành vi, trong khi Z tác động lên xác suất tham gia lao động chính thức trong phương trình chọn mẫu.
- Điều kiện loại trừ của phương trình lựa chọn: Z phải có ít nhất một biến ngoại sinh không có trong X, tương tự như phương pháp 2SLS/IV.

Ví dụ 2: ước lượng tác động của thủy lợi đến năng suất lúa và ngô bằng phương pháp hàm sản xuất

Sử dụng bộ dữ liệu irrigation.dta.

- Chúng ta quan sát được sản lượng lúa và ngô trên từng mảnh đất, các đặc tính đất đai thổ nhưỡng của các khoảnh ruộng, biến nhân khẩu học... Biến chính sách là tình trạng tưới tiêu (đất có được tưới tiêu bằng thủy lợi hay không).
- Mảnh đất được tưới tiêu được kỳ vọng có sản lượng cao hơn. Chênh lệch sản lượng giữa các mảnh đất có và không có tưới tiêu sẽ cho phép ước lượng giá trị của thủy lợi.
- Biết được giá trị của thủy lợi sẽ giúp ước tính mức phí thủy lợi mà nông dân phải trả khi sử dụng nước.

Giả sử hàm sản xuất dạng logarithm như sau:

$$log(Q_i) = \alpha_0 + \alpha_1 \times D_{IRRI_i} + \sum_j INPUT^j{}_i \times \alpha_j + \sum_k LAND^k{}_i \times \alpha_k + \sum_n DEMO^n{}_i \times \alpha_n + u_i$$

trong đó:

- Q là tổng sản lượng trên một công đất (kg/1000m2) một năm.
- D_{IRRI} là biến mảnh ruộng có được tưới tiêu hay không.
- $INPUT$, $LAND$, $DEMO$ là các biến nhân tố đầu vào, đặc tính đất đai, và nhân khẩu học của hộ gia đình.

Nhận diện vấn đề lựa chọn mẫu trong bài toán thủy lợi

- Chính sách nông nghiệp ở Việt Nam yêu cầu một số loại đất chỉ được trồng lúa. Ngoài ra, việc trồng cây gì cũng phụ thuộc vào các đặc tính đất đai thổ nhưỡng của từng khoảng ruộng. Thường thì các mảnh đất tốt nhất được dành để trồng lúa, còn đất xấu hơn thì trồng màu hay câu lâu năm $\Rightarrow$ Dữ liệu quan sát được bị ảnh hưởng bởi vấn đề chọn mẫu.

- Nếu chỉ đất tốt nhất, có thủy lợi, hạ tầng... dành cho trồng lúa $\Rightarrow$ khả năng ước lượng giá trị của thủy lợi bằng hồi quy OLS sẽ bị chệch lên.

Mô hình 1: Ước lượng hàm sản xuất bằng OLS

$$log(Q_i) = \alpha_0 + \alpha_1 \times D_{IRRI_i} + \sum_j INPUT^j{}_i \times \alpha_j + \sum_k LAND^k{}_i \times \alpha_k + \sum_n DEMO^n{}_i \times \alpha_n + u_i$$

Mô hình 2: Hàm hồi quy có điều chỉnh vấn đề chọn mẫu bằng phương pháp Heckit. Ví dụ với đất lúa:

$$\begin{cases} log(Q_i^{rice}) = \alpha_0 + \alpha_1 \times D_{IRRI_i} + ... + \rho\lambda(Z_i\gamma) + u_i \\ P(s_i = 1|Z_i) = \Phi(Z_i\gamma + v_i) \end{cases}$$

trong đó Z là các đặc tính đất đai và chính sách có thể ảnh hưởng đến việc chọn loại cây trồng, bao gồm quy định mảnh đất đó chỉ được trồng lúa hay có thể trồng cây khác.

So sánh và kiểm định mô hình lựa chọn mẫu

- So sánh kết quả giữa mô hình OLS và Heckit: Đúng là có hiện tượng ước lượng thiên lệch quá với mô hình lúa, và thiên lệch giảm với mô hình ngô.
- Kiểm định có vấn đề tự lựa chọn mẫu: $H_o : \rho = 0$. Nếu bác bỏ H_o thì cần sử dụng mô hình lựa chọn mẫu.

RICE MODEL

	OLS b/se	Heckman Co~d b/se
main		
plotIrriga~n	**0.4885*** **	**0.4036*** **
	(0.0812)	**(0.0212)**
plotArea	**-0.0744*** **	**-0.0702*** **
	(0.0128)	**(0.0027)**

MAIZE MODEL

	OLS b/se	Heckman Co~d b/se
main		
plotIrriga~n	**0.0961***	**0.1234** **
	(0.0412)	**(0.0387)**
plotArea	**-0.0444** **	**-0.0509*** **
	(0.0108)	**(0.0067)**