Tóm tắt một số điểm chính của phương pháp hồi quy với dữ liệu bảng

Lê Việt Phú

1. Mục đích của sử dụng dữ liệu bảng là để khắc phục hạn chế với nhân tố không quan sát được không thay đổi theo thời gian (unobserved heterogeneity). Nhân tố không quan sát được, chẳng hạn như tố chất cá nhân trong mô hình tỷ suất thu nhập của đi học, tương quan với số năm đi học, và cũng có tác động lớn đến thu nhập. Do đó, nhân tố này dẫn đến hệ quả biến giải thích (số năm đi học) tương quan với phần dư (thu nhập), vi phạm điều kiện 4.2 của mô hình CLRM.

   Nếu có dữ liệu bảng thì dùng sai phân bậc nhất (first differencing data - dữ liệu năm sau trừ đi năm trước) sẽ loại bỏ được tất cả các nhân tố không thay đổi theo thời gian.

   Mô hình gốc:

$$Y_{it} = \beta_0 + \beta_k X_{it} + Ability_i + u_{it} \tag{1}$$

   Mô hình với sai phân bậc nhất:

$$\Delta Y_i = \beta_k \Delta X_i + \Delta u_{it} \tag{2}$$

   Do đó, chúng ta sẽ loại được vấn đề tố chất cá nhân không quan sát được làm mất hiệu lực của mô hình OLS với dữ liệu chéo. Hồi quy bằng OLS với dữ liệu sai phân có hiệu lực nội tại. Ước lượng bằng dữ liệu sai phân bậc nhất gọi là ước lượng sai phân bậc nhất – **first differencing estimator**, hoặc ước lượng khác biệt trong khác biệt (**diffirence-in-difference estimator** - dùng khác biệt của biến giải thích để ước lượng mô hình của khác biệt của biến phụ thuộc).

2. Mở rộng, khi có dữ liệu bảng thì chúng ta có thể thiết kế nghiên cứu nhằm phục vụ mục đích đánh giá tác động chính sách. Học viên cần nắm rõ khái niệm đánh giá tác động chính sách – program evaluation hay impact evaluation – là ước lượng tác động của một can thiệp lên kết quả với nhóm hưởng lợi so với tình huống nhóm đó không được hưởng lợi. Ở đây chúng ta nhấn mạnh mối quan hệ nhân quả giữa một can thiệp với kết quả xảy ra, thay vì chỉ dừng lại ở mối quan hệ tương quan.

Tác động chính sách là sự khác biệt về mặt kết quả thực tế xảy ra so với kết quả đáng lẽ đã xảy ra. Kết quả đáng lẽ đã xảy ra được gọi là phản chứng (counterfactual). Đương nhiên, chúng ta chỉ quan sát được một nhóm đối tượng tại một thời điểm, do đó không bao giờ quan sát được phản chứng. Phản chứng phải được ước lượng thông qua các thiết kế nghiên cứu. Do yêu cầu phải ước lượng được phản chứng để thiết lập quan hệ nhân quả nên các nghiên cứu về tác động chính sách yêu cầu cao hơn rất nhiều so với các mô hình định lượng thuần túy.

Với dữ liệu bảng, chúng ta có thể xây dựng được phản chứng bằng giả định song song. Theo đó, xu hướng thay đổi kết quả giữa hai nhóm hưởng lợi và nhóm kiểm soát được giả định là tương đồng nếu như không xảy ra can thiệp.

Chúng ta có thể ước lượng tác động can thiệp trung bình (ATE) bằng hồi quy dữ liệu gộp với biến tương tác (**pooled regression with interaction terms**). Tác động của chính sách là tham số của biến tương tác.

$$Y_{it} = \beta_0 + \beta_1 T_i + \beta_2 Year_t + \beta_3 * (T_i \times Year_t) + \beta_k X_{it} + u_{it} \tag{3}$$

Cách thức ước lượng này là đơn giản nhất (chỉ sử dụng OLS), và cho phép sử dụng dữ liệu bảng không cân bằng, không bỏ sót bất cứ quan sát nào. Tuy nhiên, lưu ý về cấu trúc dữ liệu phải đúng trước khi ước lượng mô hình.

Phương pháp này có nhược điểm là nó không tận dụng cấu trúc dữ liệu lặp để xử lý vấn đề thiếu biến quan trọng không quan sát được trong mô hình tương tự như các ước lượng OLS với dữ liệu chéo. Đồng thời, nhà nghiên cứu cũng phải cảnh giác nếu dữ liệu thiếu không ngẫu nhiên có thể dẫn đến sai lệch kết quả.

3. Để tận dụng lợi thế lặp của dữ liệu bảng thì chúng ta có các phương pháp hồi quy dữ liệu bảng, đặc biệt là ước lượng với tác động cố định (**fixed effects estimator**). Tác động cố định là tất cả các nhân tố không quan sát được nhưng không thay đổi theo thời gian (time invariant unobserved heterogeneity). Ví dụ đối với bài toán tỷ suất thu nhập của đi học thì tác động cố định có thể là tố chất cá nhân. Còn với bài toán tín dụng vi mô thì tác động cố định có thể là yếu tố vốn xã hội (social capital) như các mối quan hệ hay khiếu kinh doanh (entrepreneurship). Có nhiều hình thức để ước lượng với tác động cố định:

   - Cách đơn giản nhất để loại trừ nhân tố không quan sát được không thay đổi theo thời gian là dùng sai phân bậc nhất (dữ liệu năm sau trừ đi năm trước). Khi này các nhân tố không thay đổi sẽ bị loại trừ, và chúng ta có thể dùng hồi quy OLS với sai phân. Cách này sẽ làm mất tất cả những dữ liệu không thay đổi theo thời gian như giới tính, dân tộc, một số đặc tính nhân khẩu học...

- Cách thứ hai là hồi quy với biến giả (**least squares dummy variables - LSDV**). Dùng mỗi biến giả đại diện cho một hộ, và đưa (N-1) biến giả vào trong mô hình. Ước lượng của tham số của biến giả sẽ là các thuộc tính cố định. Thông thường thì các tham số này không có ý nghĩa hay hàm ý chính sách gì, do đó chúng ta không cần quan tâm đến chúng. Có một điểm lưu ý với hồi quy biến giả là việc đưa vào một số lượng lớn biến giả dẫn đến R2 tăng rất cao. Trong khi đó R2 với phương pháp sai phân thường thấp hơn rất nhiều. Việc R2 tăng cao hay thấp không phải do mô hình sai mà do cách sử dụng dữ liệu khác nhau. Đồng thời, việc đưa nhiều biến vào làm giảm số bậc tự do và giảm sức mạnh của kiểm định thống kê.

$$Y_{it} = \beta_0 + \beta_1 T_{it} + \beta_2 Year_t + \beta_k X_{it} + \sum_{j=1}^{N-1} \sigma_j D_j + u_{it} \tag{4}$$

Cả hai phương pháp trên bản chất đều là ước lượng OLS.

- Cách thứ ba là dùng hồi quy dữ liệu bảng với tác động cố định bằng ước lượng bên trong (**within estimator**). Đây là phương pháp tối ưu nhất với dữ liệu lặp. Phương pháp này ước lượng bằng cách chuyển đổi dữ liệu loại trừ giá trị trung bình (**within transformation/time-demeaned transformation**).

$$Y_{it} = \beta_0 + \beta_1 T_{it} + \beta_2 Year_t + \beta_k X_{it} + a_i + u_{it} \tag{5}$$

Ước lượng này yêu cầu dữ liệu lặp, do đó cũng gặp vấn đề tương tự như ước lượng bằng sai phân. Lưu ý tổ chức dữ liệu với hồi quy dữ liệu bảng khác với hồi quy dữ liệu gộp với biến tương tác.

4. Lựa chọn phương pháp ước nào và giải thích kết quả như thế nào phụ thuộc vào dữ liệu, lý thuyết và bối cảnh nghiên cứu.

   - Hồi quy dữ liệu gộp đơn giản dễ thực hiện, nhưng không xử lý được vấn đề thiếu biến quan trọng không quan sát được tương quan với biến chính sách dẫn đến ước lượng không có hiệu lực nội tại.

   - Hồi quy dữ liệu bảng với tác động cố định xử lý được vấn đề trên và có hiệu lực nội tại tốt nhất, nhưng yêu cầu dữ liệu lặp.

   - Nếu dữ liệu thiếu không ngẫu nhiên thì kết quả có thể bị sai lệch.

   - Để có dữ liệu lặp yêu cầu cao về công tác thu thập.

   - Hồi quy dữ liệu bảng với tác động cố định không thể áp dụng để ước lượng tác động của

nhân tố không hoặc ít thay đổi theo thời gian.

5. Lưu ý khi giải thích kết quả:

   - R2 khác biệt lớn giữa các mô hình, mặc dù dữ liệu gốc, mô hình được xây dựng và cách giải thích các tham số giống nhau. Sự khác biệt này là do cách thức dữ liệu được sử dụng khi ước lượng mô hình.

   - Thứ nhất, LSDV có R2 cao do đưa vào một số lượng lớn biến giả nên R2 tăng.
   - Thứ hai, pooled OLS và LSDV sử dụng nguyên dữ liệu gốc, trong khi ước lượng tác động cố định (within hoặc first differencing estimator) sử dụng dữ liệu đã chuyển đổi (time-demeaned (TD) hoặc first-differenced (FD) data). Khi chuyển đổi dữ liệu thì sẽ làm mất dữ liệu, theo 2 hướng:

     * Số quan sát (và bậc tự do) giảm. Giả sử nếu chỉ có 2 kỳ quan sát thì sau khi chuyển đổi TD hay FD số quan sát chỉ còn một nửa.

     * Mất dao động của dữ liệu (lose data variations). Từ học kỳ thu chúng ta đã biết cần dữ liệu có thay đổi từ quan sát này sang quan sát khác để ước lượng được mô hình. Chuyển đổi TD hay FD chỉ giữ lại các dữ liệu thay đổi theo thời gian. Ngay cả với dữ liệu có thay đổi nhưng nếu thay đổi ít thì kết quả ước lượng cũng kém chính xác (do đó R2 thấp).

     Với hai lý do trên thì ước lượng bằng within hoặc first differencing estimator làm cho R2 thấp. Tuy nhiên, R2 không phải là tiêu chí chính để lựa chọn mô hình.

   - Mở rộng, hồi quy đưa nhiều biến giả vào có thể giúp tăng R2 tùy ý. Do đó, cần tránh lạm dụng việc đưa biến giả tùy tiện nhằm khắc phục mô hình yếu. Ngoài ra, việc tăng R2 có thể dẫn đến hậu quả là mô hình bị ước lượng quá khớp (overfitting). Khi đưa nhiều biến giả hay tăng độ phức tạp của mô hình thì có thể tăng mức độ dự đoán với dữ liệu trong mẫu (in-sample prediction) nhưng khả năng dự báo ngoài mẫu giảm (out-of-sample prediction). Công cụ kiểm chứng chéo (cross-validation) có thể được sử dụng để lựa chọn mô hình tối ưu (dùng dữ liệu trong mẫu để xây dựng mô hình tối đa hóa khả năng dự báo ngoài mẫu).

6. Kinh nghiệm thực tế:

   - Các bộ dữ liệu điều tra đại đa số là dữ liệu chéo gộp (pooled cross-sectional data), có nghĩa là các bộ dữ liệu được điều tra nhiều vòng (ví dụ VHLSS) nhưng số hộ gia đình được điều tra lặp rất ít, dẫn đến khả năng dùng thiết kế nghiên cứu sử dụng dữ liệu bảng khá hạn chế.

   - Một số bộ điều tra được thiết kế ngay từ đầu để đánh giá quá trình tiến hóa (longitudinal analysis) của mẫu như Young Lives, VARHS, SME (CIEM). Mặc dù vẫn có hiện tượng mẫu bị

rơi rụng và bổ sung mẫu (do đó cần cảnh giác khi sử dụng) nhưng kết quả có hiệu lực nội tại tốt.

Mô hình dữ liệu bảng nâng cao: Fixed Effects (FE) & Random Effect (RE) Estimators

$$Y_{it} = \beta_k * X_{it} + \underbrace{a_i + u_{it}}_{v_{it}} \tag{6}$$

Mục đích chính của việc sử dụng dữ liệu bảng với tác động cố định là để loại trừ nhân tố a_i *(1) không quan sát được (2) không thay đổi theo thời gian và (3) có tương quan với biến chính sách* X. Trong bài toán tỷ suất thu nhập của đi học, nếu không xử lý vấn đề tố chất cá nhân tương quan với biến chính sách và thu nhập sẽ dẫn đến ước lượng của biến chính sách bị lệch và không nhất quán.

Dùng sai phân bậc nhất (first differencing) hoặc chuyển đổi loại trừ giá trị trung bình (time demeaned transformation) sẽ loại được a_i khỏi mô hình. Chúng ta cũng có thể dùng LSDV để ước lượng các tác động cố định a_i bằng việc đưa $(N-1)$ biến giả vào mô hình, mặc dù chúng không có hàm ý gì về mặt chính sách.

Tuy nhiên, điều gì xảy ra nếu *tác động cố định* a_i *không tương quan với biến chính sách* X*?*

Chúng ta vẫn có thể sử dụng within hay first-differencing estimator, nhưng đây *không phải là cách làm tối ưu* bởi bản chất của FE estimator là chuyển đổi dữ liệu để loại trừ a_i, do đó làm mất dữ liệu, trong khi *sự tồn tại của* a_i *không làm mô hình hồi quy bằng OLS mất hiệu lực nội tại (do giả định 4.2 vẫn thỏa). Ước lượng OLS với mô hình (6) vẫn không chệch và nhất quán, nhưng không hiệu quả (inefficient).* Tại sao hồi quy OLS không hiệu quả khi có nhân tố a_i không quan sát được không thay đổi theo thời gian và không tương quan với biến giải thích X?

Khi học về dữ liệu chuỗi thời gian chúng ta sẽ cần chứng minh rằng phần dư gộp v_{it} (composite error term – bao gồm phần dư tức thì (instantaneous errors) u_{it} và tác động cố định không quan sát được a_i) sẽ tương quan chuỗi với nhau theo thời gian (serial correlation) thông qua nhân tố a_i.

$$cov(v_{it}, v_{is}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2} \neq 0 \tag{7}$$

Khi xảy ra tương quan chuỗi thì một trong các giả định của mô hình CLRM bị vi phạm (giả định *iid* – independent, identically distributed error terms – phần dư độc lập và phân phối đồng nhất). Khi này, ước lượng bằng OLS vẫn không chệch (unbiased) và nhất quán nhưng không phải là hiệu

quả nhất (không phải là best hay most efficient estimator – thể hiện ở phương sai của ước lượng tăng dẫn đến ước lượng kém chính xác).

Tóm lại, khi có nhân tố a_i *(1) không quan sát được (2) không thay đổi theo thời gian và (3) không tương quan với biến chính sách* X thì sử dụng FE estimator hay OLS đều không phải là tối ưu. Khi này, chúng ta có thể sử dụng mô hình dữ liệu bảng với tác động ngẫu nhiên - Random Effects Model và ước lượng RE sẽ *hiệu quả* hơn ước lượng FE. Bản chất của RE estimator là ước lượng lai giữa pooled OLS với FE estimator. RE được thực hiện bằng cách chuyển đổi dữ liệu theo tham số θ nhằm phản ánh mức độ quan trọng tương đối giữa nhân tố cố định không quan sát được a_i so với phần dư tức thì u_{it} trong phần dư gộp v_{it}.

$$\theta = 1 - \sqrt{\frac{\sigma_u^2}{(\sigma_u^2 + T\sigma_a^2)}} \tag{8}$$

và chuyển đổi dữ liệu theo công thức: $Y_{it} - \theta\overline{Y_i}$, $X_{it} - \theta\overline{X_i}$, và $v_{it} - \theta\overline{v_i}$. Sau đó ước lượng mô hình với dữ liệu đã chuyển đổi:

$$Y_{it} - \theta\overline{Y_i} = \beta_k * (X_{it} - \theta\overline{X_i}) + (v_{it} - \theta\overline{v_i}) \tag{9}$$

Nếu nhân tố cố định rất quan trọng (thể hiện qua phương sai của a lớn hơn nhiều so với phương sai của u) thì θ sẽ tiệm cận giá trị 1 và cách chuyển đổi dữ liệu như trên sẽ rất giống với chuyển đổi dữ liệu loại trừ giá trị trung bình (**within transformation/time-demeaned transformation**), và ước lượng RE sẽ tiệm cận ước lượng FE. Nếu nhân tố cố định không quan trọng trong mô hình (phương sai của a rất nhỏ so với phương sai của u) thì θ sẽ tiệm cận giá trị 0 và ước lượng RE sẽ giống như ước lượng pooled OLS.

Chúng ta chỉ nên nhìn nhận RE estimator như một mô hình mở rộng nâng cao so với mô hình FE. Có dữ liệu bảng thì FE estimator luôn là lựa chọn đầu tiên nhằm giải quyết a_i (1) không quan sát được (2) không thay đổi theo thời gian và (3) tương quan với biến chính sách X. Nới lỏng điều kiện (3) cho phép ước lượng mô hình RE thì kết quả sẽ hiệu quả hơn mô hình FE. Sau đó, dùng lý thuyết và kiểm định Hausman để lựa chọn FE hay RE. Việc đưa thêm nội dung RE vào nghiên cứu làm cho kết quả đầy đủ và thuyết phục hơn, nhưng nó chỉ nên coi là nội dung bổ sung làm vững kết quả. Làm đủ nội dung FE đã đảm bảo kết quả đúng (nhưng có thể không phải là ước lượng tối ưu nhất). Bổ sung thêm RE có thể cải thiện kết quả của ước lượng FE. Nhưng nếu áp dụng RE sai thì hậu quả còn nguy hiểm hơn là không làm bởi áp dụng RE sai làm cho ước lượng không nhất quán (inconsistent). Ước lượng không nhất quán không có hiệu lực nội tại và không thể chỉnh sửa được ngay cả khi có mẫu quan sát lớn.