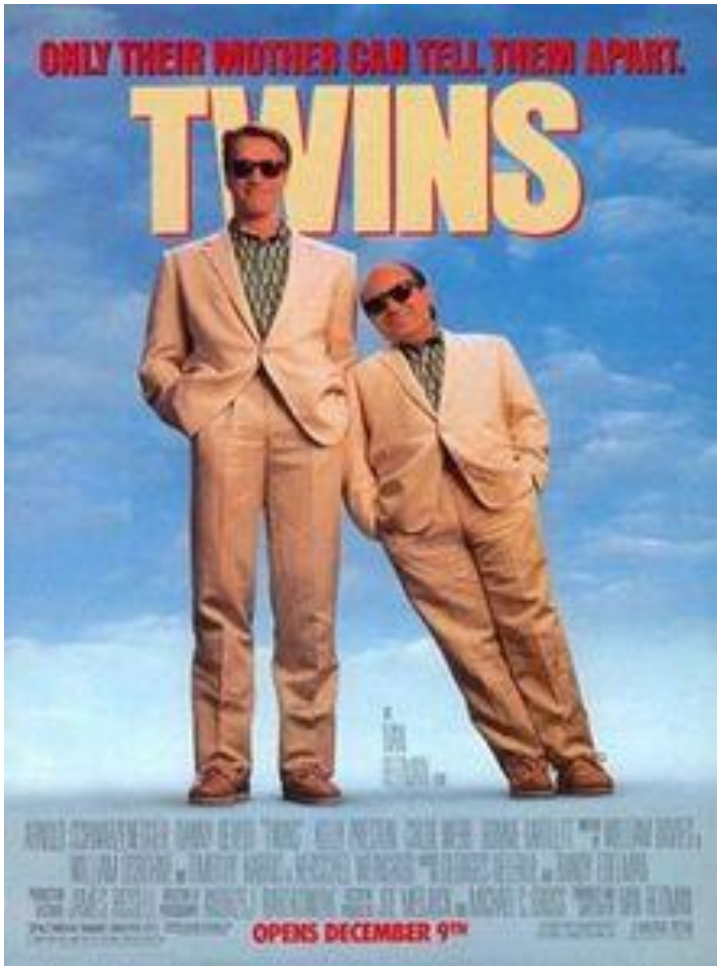
Bài giảng 9:
Phương pháp đánh giá ghép cặp dựa trên điểm xu hướng (Propensity Score Matching)

Edmund Malesky, Ph.D.

July 13, 2018

Duke University

Chiến lược ghép cặp



Thế nào là ghép cặp

- Công cụ để ước lượng nhân quả dựa trên ước lượng phản thực
- Xây dựng nhóm so sánh nhân tạo bằng các công cụ thống kê:
  - Tìm cách ghép một hoặc nhiều hộ gia đình/cá nhân không tham gia với mỗi hộ gia đình/cá nhân tham gia.
  - Các cặp ghép được với nhau dựa trên các đặc tính quan sát được giống nhau.
- Các cá nhân hoặc hộ không tham gia được sử dụng làm nhóm đối chứng cho nhóm hưởng lợi
- Cần giả định mạnh: việc lựa chọn tham gia chương trình chỉ dựa trên các đặc tính quan sát được
  - Giả định này khắt khe hơn nhiều so với phương pháp Diff-in-Diff
  - Không thể kiểm chứng được, nhưng có thể đánh giá mức độ hợp lý
  - Là hạn chế lớn nhất của phương pháp ghép cặp
- Thông thường thì kém vững hơn phương pháp DD/RDD/thử nghiệm ngẫu nhiên
  - Sử dụng để thay thế khi các phương pháp khác không thể dùng được

Động lực

DEATH RATES PER 1,000 PERSON-YEARS

Smoking group	Study		
	Canadian	British	U. S.
Non-smokers	20.2	11.3	13.5
Cigarettes only	20.5	14.1	13.5
Cigars, pipes	35.5	20.7	17.4

MEAN AGES, YEARS

Smoking group	Study		
	Canadian	British	U. S.
Non-smokers	54.9	49.1	57.0
Cigarettes only	50.5	49.8	53.2
Cigars and/or pipe	65.9	55.7	59.7

Source: Cochran, 1968.

Lời nguyền về thông tin đa chiều (Curse of Multidimensionality)

Treated units			
Age	Gender	Months unemployed	Secondary diploma
19	1	3	0
35	1	12	1
41	0	17	1
23	1	6	0
55	0	21	1
27	0	4	1
24	1	8	1
46	0	3	0
33	0	12	1
40	1	2	0

Untreated units			
Age	Gender	Months unemployed	Secondary diploma
24	1	8	1
38	0	2	0
58	1	7	1
21	0	2	1
34	1	20	0
41	0	17	1
46	0	9	0
41	0	11	1
19	1	3	0
27	0	4	0

- Khi có rất nhiều tiêu chí để so sánh, tiêu chí gì là quan trọng nhất?
- Có thể so sánh nhóm hưởng lợi/đối chiếu có cùng các đặc điểm quan sát được
- Nhưng với rất nhiều biến thì rất khó có thể đảm bảo tương đồng
- Thường thì khó có thể tìm được hai hộ gia đình giống hệt nhau, chỉ khác về tình trạng hưởng lợi
- Phương pháp ghép cặp bằng điểm xu hướng có thể xử lý vấn đề này

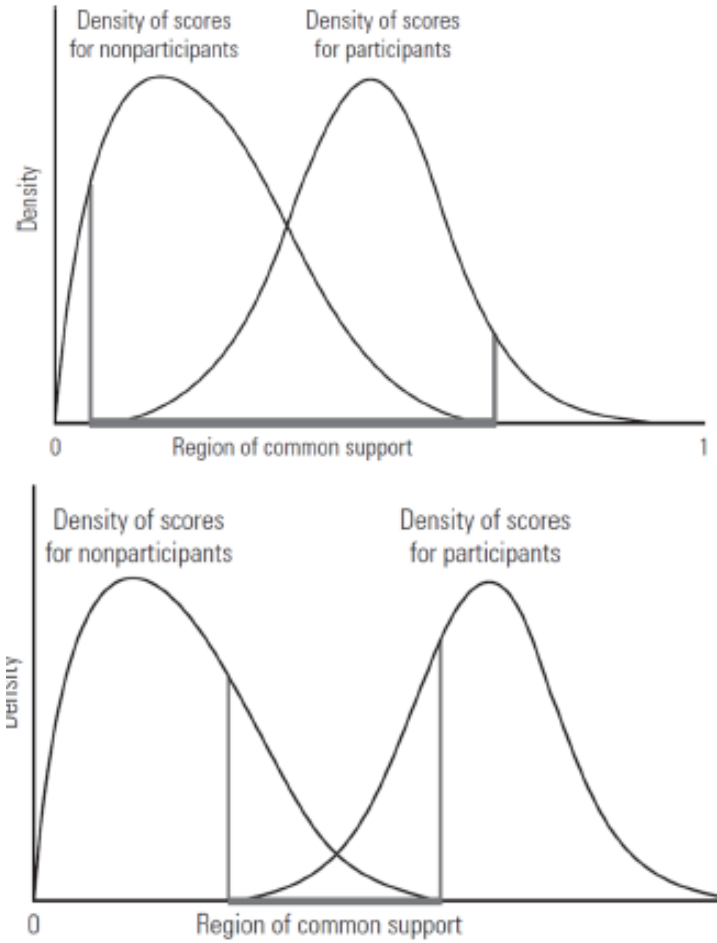
Phương pháp đánh giá ghép cặp dựa trên điểm xu hướng (Propensity Score Matching-PSM)

- Ghép cặp dựa vào xác suất tham gia chương trình được ước lượng dựa trên các <u>đặc tính quan sát được</u>
- *Điểm xu hướng, P(X):* là xác suất mà một quan sát sẽ tham gia chương trình dựa trên các đặc tính quan sát được
  - Là một chỉ số tổng hợp tất cả các đặc tính quan sát được có ảnh hưởng đến trạng thái tham gia
- Phương pháp PSM ghép các quan sát tham gia với đối chứng khi giá trị *P(X)* là gần nhau nhất
- Hiệu lực của PSM phụ thuộc vào 2 giả định:
  1. **<u>Độc lập có điều kiện:</u>** $(Y_i^T, Y_i^C) \perp T_i | X_i$
  2. **<u>Có vùng hỗ trợ chung:</u>** $0 < P(T_i = 1|X_i)$

1. **<u>Độc lập có điều kiện:</u>** Sau k[illegible] tất cả các khác biệt liên quan đến các biến quan sát được X, tình trạng tham gia chương trình hoàn toàn độc lập với kết quả tham gia [given set of observable covariates X that are not affected by treatment, potential outcomes Y are independent of (orthogonal to) treatment assignment T]
2. **<u>Vùng hỗ trợ chung</u>:** việc tham gia chương trình chỉ phụ thuộc vào các đặc tính quan sát được

PSM & Vùng hỗ trợ chung (common support)

- Vùng hỗ trợ chung đảm bảo tìm được nhóm đối chứng cho nhóm tham gia do có giá trị P(X) gần giống nhau
- Vùng đuôi của phân phối nằm ngoài vùng hỗ trợ chung
- Có số mẫu lớn sẽ giúp tìm được nhóm đối chứng cho nhóm tham gia
- Vùng hỗ trợ chung kém có thể dẫn đến ước lượng bị chệch
  - Ví dụ loại bỏ các quan sát nằm ngoài vùng hỗ trợ chung có thể là làm mất dữ liệu một cách không ngẫu nhiên



Các bước để thực hiện PSM

1. Sử dụng các điều tra thống nhất của cả nhóm tham gia và nhóm đối chứng
2. Gộp các dữ liệu và ước lượng xác suất tham gia chương trình dựa trên các đặc tính quan sát được – gọi là điểm xu hướng hay P(X)
   - Cụ thể là chúng ta sử dụng một mô hình hồi quy sau:
     i. Biến phụ thuộc là tình trạng tham gia, =1 nếu tham gia, và =0 nếu không tham gia.
     ii. Sử dụng hồi quy *logit* hoặc *probit* để ước lượng xác suất tham gia, với các biến giải thích là các đặc tính quan sát được.
3. Hạn chế mẫu phân tích vào khu vực có vùng hỗ trợ chung
4. Xếp dữ liệu theo điểm xu hướng – P(X).

   - Đối với nhóm tham gia, tìm các quan sát không tham gia nhưng có điểm xu hướng gần giống
5. So sánh kết quả của nhóm tham gia với nhóm không tham gia.
6. Khác biệt về kết quả trung bình = tác động của chương trình lên nhóm tham gia
7. Trung bình của các khác biệt = Tác động can thiệp trung bình

Các phương pháp tính tác động khác nhau PSM

Có nhiều phương pháp ghép nhóm tham gia và nhóm đối chứng

1. Ghép quan sát gần nhất
2. Ghép theo khoảng giá trị
3. Ghép theo tầng
4. Ghép bằng quyền số dựa trên phân phối kernel & hồi quy nội tại
5. Ghép bằng quyền số dựa trên thuật toán genetic.

Các phương pháp trên nói chung đều cho ra kết quả giống nhau, mặc dù có độ chính xác khác nhau.

Sử dụng PSM khi nào

- Sử dụng PSM chỉ khi các biến quan sát được có ảnh hưởng đến trạng thái tham gia chương trình
  - Tùy thuộc vào định hướng chương trình và các nhân tố ảnh hưởng đến việc tự lựa chọn tham gia (self-selection)
  - Không thể chứng minh một cách chắc chắn được
  - Yêu cầu phải hiểu bối cảnh của việc thực hiện chương trình, và sử dụng điều tra để đánh giá
- Chỉ phù hợp khi thông tin cung cấp là phù hợp
  - Càng nhiều dữ liệu càng tốt, đặc biệt là một số biến trọng yếu
- Cảnh giác với việc ghép cặp sau khi thực hiện chương trình
  - Ghép cặp phải sử dụng dữ liệu tham chiếu (trước khi thực hiện chương trình)
  - Rủi ro với điều tra sau khi thực hiện chương trình: Việc thực hiện ảnh hưởng đến các biến quan sát được
- Có thể kếp hợp phương pháp ghép cặp với các phương pháp khác như Diff-in-Diff
- Có thể sử lý được vấn đề chệch lựa chọn (selection bias) do các nhân tố không quan sát được nhưng không thay đổi theo thời gian

Ví dụ HISP

Table 7.1 Estimating the Propensity Score Based on Observed Characteristics

Dependent Variable: *Enrolled* = 1	
Explanatory variables / characteristics	**Coefficient**
Head of household's age (years)	−0.022**
Spouse's age (years)	−0.017**
Head of household's education (years)	−0.059**
Spouse's education (years)	−0.030**
Head of household is female = 1	−0.067
Indigenous = 1	0.345**
Number of household members	0.216**
Dirt floor = 1	0.676**
Bathroom = 1	−0.197**
Hectares of land	−0.042**
Distance to hospital (km)	0.001*
Constant	0.664**

Source: Authors.

Note: Probit regression. The dependent variable is 1 if the household enrolled in HISP, and 0 otherwise. The coefficients represent the contribution of each listed explanatory variable / characteristic to the probability that a household enrolled in HISP.

* Significant at the 5 percent level; ** Significant at the 1 percent level.

Source: Gertler et al., 2011.

Ví dụ về trợ cấp bảo hiểm y tế

Table 7.2 Case 7—HISP Impact Using Matching (Comparison of Means)

	Enrolled	Matched comparison	Difference	*t*-stat
Household health expenditures	7.8	16.1	−8.3	−13.1

Table 7.3 Case 7—HISP Impact Using Matching (Regression Analysis)

	Multivariate linear regression
Estimated impact on household health expenditures	−8.3** (0.63)

Source: Authors.

Note: Standard errors are in parentheses.

** Significant at the 1 percent level.

Source: Gertler et al., 2011.

Tác động của việc tư nhân hóa cấp nước đến tỷ lệ tử vong của trẻ em

	Full Sample			Using Observations on Common Support			Kernel Matching on Common Support*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Private water services (=1)	−.334	−.320	−.283	−.540	−.541	−.525	−.604
	(.169)**	(.170)*	(.170)*	(.177)***	(.178)***	(.178)***	(.168)***
	[.157]**	[.163]**	[.162]*	[.191]***	[.198]***	[.195]***	
	[.195]*	[.203]	[.194]	[.261]**	[.274]**	[.266]**	
%Δ in mortality rate	−5.3	−5.1	−4.5	−8.6	−8.6	−8.4	−9.7
Other covariates:							
Real GDP per capita		.007	.009		.005	.006	
		(.005)	(.006)		(.006)	(.006)	
		[.006]	[.006]		[.007]	[.007]	
		[.007]	[.007]		[.007]	[.008]	
Unemployment rate		−.555	−.686		−.778	−.896	
Income inequality		5.171	5.085		2.932	3.052	
		(2.868)*	(2.880)*		(2.907)	(2.926)	
		[3.468]	[3.445]		[3.314]	[3.289]	
		[3.696]	[3.691]		[3.833]	[3.838]	
Public spending per capita		−.028	−.035		−.068	−.070	
		(.038)	(.038)		(.039)*	(.039)*	
		[.055]	[.055]		[.059]	[.059]	
		[.054]	[.055]		[.049]	[.050]	
Local government by Radical party (=1)			.482			.166	
			(.267)*			(.284)	
			[.281]*			[.301]	
			[.288]*			[.365]	
Local government by Peronist party (=1)			−.202			−.168	
			(.191)			(.193)	
			[.202]			[.230]	
			[.254]			[.309]	
R^2	.1227	.1256	.1272	.1390	.1415	.1420	
Observations	4,732	4,597	4,597	3,970	3,870	3,870	3,970

Source: Galiani et al, 1995.

Jalan và Ravillion (2003)

- Mỗi năm có 4 triệu trẻ em chết vì bệnh tiêu chảy
  - Nguyên nhân chính: nước uống không an toàn
- Bài nghiên cứu này đánh giá tác động của chương trình cấp nước máy ở Ấn độ
  - 1.5 triệu trẻ em chết hàng năm do bệnh tật liên quan đến chất lượng nước
  - Cao nhất thế giới
- Nhận thấy khu vực có nước máy có tỷ lệ nhiễm bệnh và thời gian mắc tiêu chảy thấp hơn
- Nhưng tác động này biến mất ở nhóm hộ nghèo hoặc có bà mẹ có tình trạng học vấn thấp
- Cần thêm các dữ liệu khác, chẳng hạn như có biết đun sôi nước và bảo quản tốt hơn không

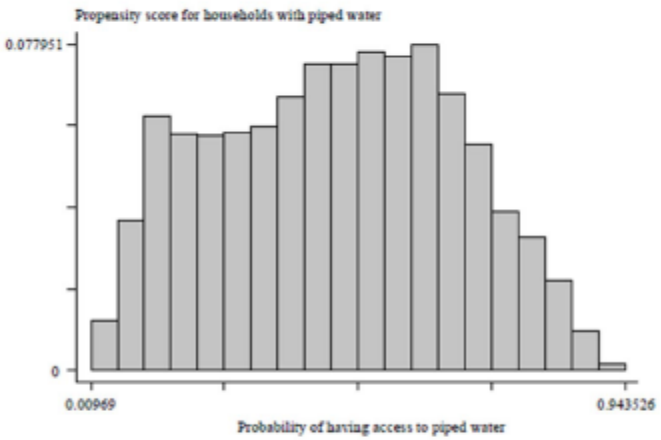
Ước lượng điểm xu hướng được tiếp cận nước sạch

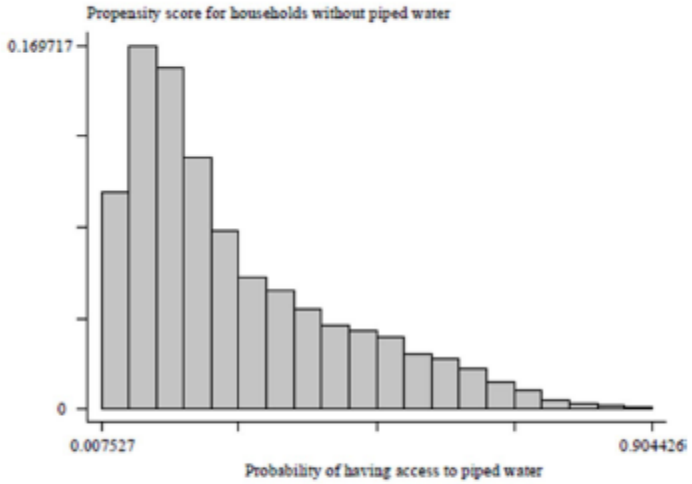
Logit regression for piped water

	Coefficient	*t*-statistic
Village variables		
Village size (log)	0.08212	4.269
Proportion of gross cropped area which is irrigated: > 0.75	−0.04824	−1.185
Proportion of gross cropped area which is irrigated: 0.5–0.75	0.19399	4.178
Whether village has a day care center	−0.07249	−2.225
Whether village has a primary school	−0.08136	−1.434
Whether village has a middle school	−0.09019	−2.578
Whether village has a high school	0.26460	7.405
Female to male students in the village	0.10637	3.010
Female to male students for minority groups	−0.07661	−2.111
Main approachable road to village: pucca road	0.19441	3.637
jeepable/kuchha road	−0.00163	−0.033
Whether bus-stoop is within the village	0.11423	2.951
Whether railway station is within the village	0.00920	0.179
Whether there is a post-office within the village	0.02193	0.550
Whether the village has a telephone facility	0.33059	9.655
Whether there is a community TV center in the village	0.09859	2.661
Whether there is a library in the village	−0.04153	−1.116
Whether there is a bank in the village	0.19084	4.655
Whether there is a market in the village	0.31690	6.092
Student teacher ratio in the village	0.00242	5.295
Household variables		
Whether household belongs to the Scheduled Tribe	−0.21288	−4.203
Whether household belongs to the Scheduled Caste	−0.01045	−0.288
Whether it is a Hindu household	−0.24195	−1.709
Whether it is a Muslim household	−0.21631	−1.427
Whether it is a Christian household	0.40367	2.426
Whether it is a Sikh household	−0.86645	−4.531
Household size	0.00337	0.571
Utilization of landholdings: used for cultivation?	0.17109	1.914
Whether the house belongs to the household	−0.18988	−2.854
Whether the household owns other property	0.00181	0.044
Whether the household has a bicycle	−0.26514	−8.243
Whether the household has a sewing machine	0.01183	0.252
Whether the household owns a thresher	−0.05790	−0.577
Whether the household owns a winnower	0.21842	1.820
Whether the household owns a bullock-cart	−0.25900	−5.430
Whether the household owns a radio	0.01036	0.251
Whether the household owns a TV	0.08095	1.335
Whether the household owns a fan	0.01336	0.321
Whether the household owns any livestock	−0.07780	−2.339

Source: Jalan & Ravallion, 2003.

Giả định có vùng hỗ trợ chung





Source: Jalan & Ravallion, 2003.

Kết quả của việc được tiếp cận nước sạch

Impacts of piped water on diarrhea prevalence and duration for children under five

	Prevalence of diarrhea		Duration of illness	
	Mean for those with piped water (st. dev.)	Impact of piped water (st. error)	Mean for those with piped water (st. dev.)	Impact of piped water (st. error)
Full sample	0.0108	−0.0023*	0.3254	−0.0957*
	(0.046)	(0.001)	(1.650)	(0.021)
Stratified by household income per capita (quintiles)				
1 (poorest)	0.0155	0.0032*	0.4805	0.0713
	(0.055)	(0.001)	(2.030)	(0.053)
2	0.0136	0.0007	0.4170	0.0312
	(0.051)	(0.001)	(1.805)	(0.051)
3	0.0083	−0.0039*	0.2636	−0.1258*
	(0.038)	(0.001)	(1.418)	(0.042)
4	0.0100	−0.0036*	0.3195	−0.1392*
	(0.044)	(0.001)	(1.703)	(0.048)
5	0.0076	−0.0068*	0.1848	−0.2682*
	(0.042)	(0.001)	(1.254)	(0.036)
Stratified by highest education level of a female member				
Illiterate	0.0131	−0.0000	0.3588	−0.0904*
	(0.053)	(0.001)	(1.710)	(0.036)
At most primary school educated	0.0112	−0.0015	0.3502	−0.0465
	(0.045)	(0.001)	1.739)	(0.036)
At most matriculation educated	0.0074	−0.0065*	0.2573	−0.1708*
	(0.038)	(0.001)	(1.476)	(0.039)
Higher secondary or more	0.0050	−0.0080*	0.1880	−0.2077*
	(0.027)	(0.002)	(1.158)	(0.076)

*Indicates significance at the 5% level or lower.

Source: Jalan & Ravallion, 2003.

Tác động của nước máy lên xác suất mắc bệnh tiêu chảy

Child-health impacts of piped water by income and education

	Illiterate		At most primary		At most matriculation		Higher secondary or more	
	Prevalence of diarrhea	Duration of illness	Prevalence of diarrhea	Duration of illness	Prevalence of diarrhea	Duration of illness	Prevalence of diarrhea	Duration of illness
1 (poorest quintile)	0.0100*	0.1028	0.0010	0.0548	−0.0118*	−0.1091	Small Sample	
	(0.002)	(0.089)	(0.002)	(0.094)	(0.003)	(0.132)		
2	0.0057*	0.0777	0.0013	0.1061	−0.0121*	−0.2580*	Small Sample	
	(0.003)	(0.083)	(0.002)	(0.083)	(0.002)	(0.087)		
3	−0.0038*	−0.1503*	−0.0008	0.0056	−0.0069*	−0.1659*	Small Sample	
	(0.002)	(0.069)	(0.002)	(0.081)	(0.002)	(0.059)		
4	−0.0062*	−0.2224*	−0.0041*	−0.1691	0.0008	−0.0186	Small Sample	
	(0.002)	(0.097)	(0.002)	(0.070)	(0.003)	(0.091)		
5	−0.0075*	−0.2932*	−0.0051*	−0.2435*	−0.0063*	−0.2578*	−0.010*	−0.2637*
	(0.000)	(0.045)	(0.002)	(0.075)	(0.002)	(0.008)	(0.003)	(0.085)

Note: Figures in parentheses are the respective standard errors.
*Indicates significance at 5% or lower.

Source: Jalan & Ravallion, 2003.