

CAUSALITY IN STATISTICAL INVESTIGATIONS

Royal Statistical Society

27 May 2009

PROPENSITY SCORE MATCHING AND CAUSAL INFERENCE

Barbara Sianesi

Institute for Fiscal Studies

barbara_s@ifs.org.uk

Largely based on:

Blundell, R., Dearden L., and Sianesi, B. (2005), "Evaluating the Impact of Education on Earnings in the UK: Models and Results from the NCDS", *Journal of the Royal Statistical Society Series A*, 168, 473-512.

Roadmap

1. THE EVALUATION PROBLEM

→ returns to educational qualifications in the UK

2. GENERAL FRAMEWORK

HETEROGENEITY

- individual characteristics → observed / unobserved
- individual impacts → homog./heterog. returns model
- treatments → single / multiple-treatment model

3. TYPES OF EVALUATION METHODS

I. excluded instrument

II. measure all relevant variables



Selection

- importance of ability & background information
- ‘testing’ matching

Mis-specification & comparability

- matching *versus* OLS and fully interacted OLS
- interpretation of results

Framework

$\{0, 1, \dots, J\}$ → set of $J + 1$ educational qualifications/levels

$S_{ji} = 1$ → attainment of qualification j by individual i

$Y_i^0, Y_i^1, \dots, Y_i^J$ → set of potential outcomes for individual i
associated to each of the $J + 1$ qualifications

$$Y_i = Y_i^k + \sum_{j=1}^J \underbrace{(Y_i^j - Y_i^k)}_{\beta_{jk,i}} S_{ji}$$

$\beta_{jk,i} \equiv$ causal effect on Y of schooling level j
relative to level k for individual i
i.e. individual return to j (relative to k)

Which parameter of interest?

- average in the population whether or not one takes level j

$$\beta_{ATE(j,k)} \equiv E[Y^j - Y^k]$$

- average among those observed to take level j

$$\beta_{ATT(j,k|j)} \equiv E[Y^j - Y^k \mid S_j = 1]$$

- average among those who have not taken level j

$$\beta_{ATNT(j,k|l)} \equiv E[Y^j - Y^k \mid S_l = 1] \quad l \neq j$$

The homogeneous returns model

$$\beta_{jk,i} = \beta_{jk} \text{ for all } i \quad j, k \in \{0, 1, \dots, J\}$$

For $j=0,1,\dots,J$:

$$Y_i^j = f_j(X_i, u_i^j) \quad \text{with } X_i^j = X_i$$

$$Y_i^j = m_j(X_i) + u_i^j \quad \text{with } E(Y_i^j | X_i) = m_j(X_i) \quad j = 0,1$$

$$\boxed{u_i^j = \alpha_i + b_{ji}} \quad \text{with } b_{0i}=0$$

then

$$\boxed{Y_i = m_0(X_i) + \sum_{j=1}^J \beta_{ji} S_{ji} + \alpha_i}$$

Return to educational level j (relative to level 0) for individual i with characteristics X_i :

$$\beta_{ji} \equiv \beta_{j0,i} \equiv Y_i^j - Y_i^0 = [m_j(X_i) - m_0(X_i)] + [u_i^j - u_i^0]$$

$$= \underbrace{b_j(X_i)}_{\Downarrow} + \underbrace{b_{ji}}_{\Downarrow}$$

common effect for all individuals with characteristics X_i idiosyncratic impact component for individual i

observable heterogeneity in impacts unobservable heterogeneity in impacts

Homogenous returns

$$Y_i = m_0(X_i) + \beta_1 S_{1i} + \beta_2 S_{2i} + \dots + \beta_J S_{Ji} + \alpha_i$$

Single Treatment Model $\rightarrow J=1$

$$Y_i = m_0(X_i) + b(X_i)S_i + \{b_i S_i + \alpha_i\}$$

Naive estimator for the ATT

$$\begin{aligned} E[Y | S=1] - E[Y | S=0] \\ = E[Y^1 - Y^0 | S=1] - \underbrace{\{ E[Y^0 | S=1] - E[Y^1 | S=0] \}}_{= B_1 + B_2 + B_3} \end{aligned}$$

Heckman, Ichimura, Smith and Todd (1998):

- B_1 Non-overlapping support of the observables
- B_2 Differences in the distribution of the observables between the two groups over the common support
- B_3 Selection on unobservables

LEAST SQUARES

ASSUMPTIONS

A1) Identifying assumption: selection on observables

$$E(\alpha | S, X) = E(\alpha | X) = 0$$

A2) Parametric assumptions to specify functional form of

- $E(Y^0 | X) \equiv m_0(X)$
- $E(Y^1 - Y^0 | X) \equiv b(X)$

Standard OLS specification

$$Y = \gamma' X + \delta S + \eta$$

- observed variables affect potential outcomes linearly
- common impact of the programme

VIOLATIONS OF ASSUMPTIONS

A1) selection on unobservables (B_3)

$$Y = m_0(X) + \beta_{ATE}(X)S + \underbrace{\{bS + \alpha\}}_e$$

$$Y = m_0(X) + \beta_{ATT}(X)S + \underbrace{\{(b - E[b | X, S=1])S + \alpha\}}_w$$

where

$$\beta_{ATE}(X) \equiv b(X)$$
$$\beta_{ATT}(X) \equiv b(X) + E[b | X, S=1]$$

OLS biased for either parameter if correlation between S and e or w (given X):

- (1) 'Ability' bias $E(\alpha | S, X) \neq 0$
- (2) 'Returns' bias $E(b | S=1, X) \neq 0$

For example:

- if only (1) and $E(\alpha | S=1) > E(\alpha | S=0)$
⇒ upward bias in OLS for β_{ATE} or β_{ATT}
- if only (2) and $E(b | S=1) > b_0$
⇒ upward bias in OLS for β_{ATE} , but
⇒ **OLS recovers β_{ATT}**

A2) mis-specification bias (B_2, B_1)

- If true model for $m_0(X)$ has higher-order terms of X
- If true model for $m_0(X)$ has interactions between the X 's
- If impact of the treatment $b(X)$ varies according to X

⇒ OLS in general biased for ATT, ATNT or ATE

Misspecification issue also linked to source of bias B_1 :

OLS approximation of the regression function over the non-overlapping region purely based on chosen functional form.

MATCHING

In recovering causal impacts, matching makes the same identifying assumption as OLS but avoids any additional ones.

A1) Identifying assumption for ATT: selection on observables

$$E(Y^0 | X, S=1) = E(Y^0 | X, S=0)$$

To give it empirical content: common support requirement

$$P(S=1 | X) < 1$$

Note: for ATNT or ATE (as with OLS), need to rule out selection on unobserved returns.

Bias decomposition

B_1 difference in the supports of X
Eliminated by performing matching only over CS
 \Rightarrow Matching might actually recover a different causal impact, $ATT(\text{Sup}_{10}) \neq ATT(\text{Sup}_1)$ – external validity

B_2 difference of the distribution of X over CS
Eliminated since matching reweights $S=0$ data to equate the distribution of X in the $S=1$ sample

B_3 difference in unobservables
Matching would be just as biased as OLS.

\Rightarrow Matching focuses on comparability in terms of observables, i.e. on constructing a suitable comparison group by carefully matching treated and non-treated on X / reweighting the non-treated to realign their X

BUT we don't need matching to make OLS less parametric...

- model can become almost arbitrary non-parametric
- functional form and homogeneity assumptions can be tested

FULLY INTERACTED OLS

$$Y = m_0(X_1, X_2) + \delta S + \delta_1(X_1 S) + \delta_2(X_2 S) + \delta_{12}(X_1 X_2 S) + e$$

Average over relevant group

$$\beta_{ATT} = \delta + \delta_1 \bar{X}_{1|S=1} + \delta_2 \bar{X}_{2|S=1} + \delta_{12} (\overline{X_1 X_2})_{|S=1}$$

$$\beta_{ATNT} = \delta + \delta_1 \bar{X}_{1|S=0} + \delta_2 \bar{X}_{2|S=0} + \delta_{12} (\overline{X_1 X_2})_{|S=0}$$

$$\beta_{ATE} = \delta + \delta_1 \bar{X}_1 + \delta_2 \bar{X}_2 + \delta_{12} (\overline{X_1 X_2})$$

- can F-test for presence of heterogeneous effects

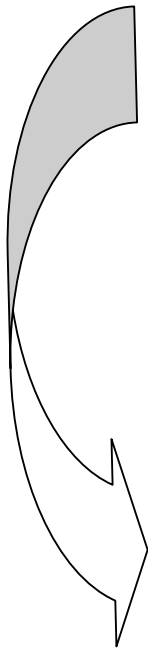
(Note: if correctly specified, OLS is BLUE)

STILL, matching (\neq OLS) highlights comparability of groups

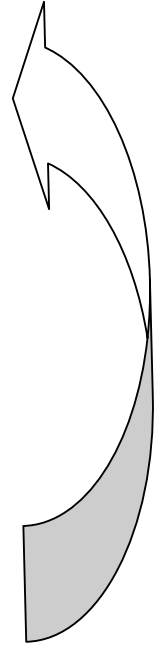
Check matching quality

Check (and possibly improve on) balancing of observables:

$$S \perp X \mid \hat{p}(X)$$



- Propensity score
 - more ‘structural’ model
 - more flexible specification
 - probit/logit
 - probability/index/odds ratio
- Matching
 - metric: X , $\hat{p}(X)$ or $\{X, \hat{p}(X)\}$
 - type of matching
 - smoothing parameters
 - common support
- Assessment of matching quality



CAN we get the two groups balanced?

Application

The data: 1958 NCDS birth cohort – males

The outcome: individual wages at age 33 in 1991

The treatments:

- no qualifications
 - O levels or vocational equivalent
 - A levels or vocational equivalent
 - some type of higher education (HE)
-
- X = (gender and age), ethnicity, math and reading ability at 7 and 11, family background (mother's and father's education and age, father's social class, mother's employment, number of siblings), school type, region.

Single-Treatment Model

Higher Education *versus* anything less

————— HOMOGENEOUS RETURNS —————

ATT \equiv ATE \equiv ATNT

Basic OLS specification age, gender, ethnicity, region	39.8 (37.1; 42.5)
Full OLS specification	28.7 (25.7; 31.8)

————— HETEROGENEOUS RETURNS —————

	ATT	ATE	ATNT
FULLY INTERACTED OLS	26.5 (23.0; 30.1)	30.8 (27.6; 34.1)	31.8 (28.9; 36.2)
MATCHING	26.8 (23.5; 31.1)	31.3 (28.7; 34.9)	33.1 (30.0; 36.7)

- omitted ability & family background bias of ~48%
- $b(X)$
 - significant ($F=1.80, p=0.002$)
 - matching \approx fully interacted OLS
 - ATNT > ATT

BUT *caveats*

- ATNT: more restrictive assumption
- no-HE includes no-quals group
- more demanding on data

Treated	N_1	Compar.	N_0	ps-R ²	ps-R ²	$P > \chi^2$	Med bias	Med bias	% lost
				<i>bef</i>	<i>aft</i>	<i>aft</i>	<i>bef</i>	<i>aft</i>	<i>aft</i>
HE	1030	no-HE	2609	.209	.006	.9963	9.1	1.9	0.1
no-HE	2609	HE	1030	.209	.037	.0000	9.1	4.3	0.8

Matching (in contrast to even fully interacted OLS):

highlights comparability of groups hence reliability of results

What about selection on unobservables?

- unobserved characteristics → Matching biased
- unobserved returns → Matching OK for ATT

‘Testing’ Matching with Control Function

Assume:

- joint N distribution for the unobservables in Y and S equations
- an exclusion restriction Z
(parental interest/birth order/bad financial shock)



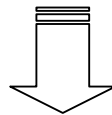
$$Y = m_0(X) + \beta_{ATE}(X)S + \rho_{\alpha v} \{(1-S)\lambda_0 + S\lambda_1\} + \rho_{bv}S\lambda_1 + \omega$$

with $E[\omega | X, S, \{(1-S)\lambda_0 + S\lambda_1\}, S\lambda_1] = 0$

- Recovers ATE directly, even if heterogeneous returns
(Exploiting the structure, ATT and ATNT)
- $\hat{\rho}$'s are informative on selection process

Most general CF model allowing for:

1. selection on unobserved ability (one CF term)
2. selection on unobserved returns (additional CF term)
3. observably heterogeneous returns $b(X)$ (interactions $X \cdot S$)
(Z interacted with S in 1st-stage Probit)



- ATT = 29.4 (9.8; 47.9)
- no evidence of residual selection on unobserved ‘ability’ ($\rho_{\alpha v}$)
- interactions $X \cdot S$ matter
- no evidence of selection on unobserved returns (ρ_{bv})
- but without $X \cdot S$: evidence of selection on ‘unobserved’ returns

Multiple-Treatment Model

- the treatments S_j ($j=0,1,2,3$)
 - dropped out of school with no qualifications
 - stopped with O levels or vocational equivalent
 - stopped with A levels or vocational equivalent
 - completed higher education
- assume NCDS information is enough to control directly for selection
- estimate the incremental return to each of the three qualifications by actual qualification
- CS imposed also by only including individuals who are matched for every possible transition

Incremental treatment effects: Matching and OLS estimates

	O-level	A-level		HE			<i>N</i>
	<i>vs</i> None	<i>vs</i> O-level	<i>vs</i> None	<i>vs</i> A-level	<i>vs</i> O-level	<i>vs</i> None	
None	13.2 (9.1; 17.3)	5.5 (0.1; 10.1)	18.7 (13.6; 23.2)	24.8 (17.7; 31.6)	30.3 (23.2; 36.3)	43.5 (36.8; 49.7)	624
O-level	17.8 (12.9; 22.1)	5.9 (2.3; 9.9)	23.7 (19.1; 29.4)	24.6 (20.5; 29.3)	30.5 (26.6; 34.4)	48.2 (43.4; 53.3)	963
A-level	18.1 (13.2; 22.6)	5.7 (2.0; 9.8)	23.8 (18.5; 28.7)	25.6 (21.7; 30.2)	31.3 (27.5; 35.5)	49.4 (43.5; 54.0)	911
HE	21.6 (14.1; 29.6)	8.0 (3.9; 12.6)	29.6 (22.0; 37.5)	21.7 (17.4; 25.6)	29.7 (25.8; 33.7)	51.3 (43.8; 58.7)	871
any: ATE	18.0 (13.3; 22.4)	6.3 (2.9; 10.1)	24.2 (19.7; 28.7)	24.2 (20.6; 28.2)	30.5 (27.1; 34.2)	48.4 (43.2; 52.7)	3,251
OLS	14.8 (11.2; 18.4)	6.4 (3.1; 9.7)	21.2 (17.3; 25.1)	23.5 (20.0; 27.1)	29.9 (26.5; 33.4)	44.7 (40.1; 48.9)	3,639
basic	21.1 (17.4; 24.7)	9.0 (5.6; 12.4)	30.0 (26.2; 33.8)	28.9 (25.6; 32.3)	37.9 (34.7; 41.1)	59.0 (55.3; 62.6)	3,639

- set of conditioning X matters
- no large heterogeneity in impacts for O and A levels groups
- no-quals group
 - lowest returns
 - 13% – cf. Harmon and Walker
- HE *versus* no-quals
 - loss to CS
 - close matches (on observables)?
 - outlier-dependent?

Treated	N_1	Compar.	N_0	ps R^2		$P > \chi^2$	Med bias		% lost
				<i>bef</i>	<i>aft</i>		<i>bef</i>	<i>aft</i>	
none	651	O lev	993	.150	.005	1.0000	1.3	1.6	0.6
		A lev	965	.248	.012	0.9985	1.8	2.3	2.9
		HE	1030	.512	.091	0.0000	5.8	7.7	5.5
O lev	993	none	651	.150	.016	0.9491	2.7	3.2	1.4
		A lev	965	.045	.002	1.0000	0.6	0.7	1.3
		HE	1030	.227	.019	0.4570	2.8	3.3	0.3
A lev	965	none	651	.248	.041	0.0005	4.3	4.6	7.2
		O lev	993	.045	.002	1.0000	0.7	0.9	1.1
		HE	1030	.127	.008	0.9999	1.5	2.1	0.5
HE	1030	none	651	.512	.162	0.0000	9.7	10.3	20.2
		O lev	993	.227	.022	0.1906	2.8	3.1	5.9
		A lev	965	.127	.005	1.0000	1.3	1.7	0.5

- the larger the educational gap, the harder to balance their X s
- more difficult when potential comparison group smaller
- matching highlights comparability hence reliability
- lack of comparability → is the causal impact of relevance?

Wrapping Up..

The 3 sources of bias in our application

SELECTION ON UNOBSERVABLES

- Set of conditioning X **matters**
 - results argue for careful choice of the matching variables
 - importance of ability and family background info

⇒ **better data help a lot!**
- NCDS: there seem to be enough variables to control **directly** for selection on unobservables
 - Matching not ‘rejected’ by Control Function

SELECTION ON OBSERVABLES:

NON-OVERLAPPING SUPPORT & DIFFERENT DISTRIBUTION OVER CS

- avoid use of potentially misleading functional forms in constructing counterfactual
 - ⇒ **(matching \approx fully interacted OLS) $>$ simple OLS**

Matching *versus* simple OLS:

no mis-specification bias; ATT *versus* ATNT

- compare comparable people
 - ⇒ **matching $>$ fully interacted OLS**

Matching *versus* fully interacted OLS:

highlights actual comparability of groups, hence reliability – and relevance – of estimates