Econometrics

Matching

Paul P. Momtaz

The Anderson School
UCLA

Outline

Motivation

- Matching or *selection on observables* can help if controlled randomization impossible and natural experiments unavailable
- Underlying conditional independence assumption often criticized
- But
  - observations selected without reference to the outcome
  - and might reduce selection bias

Conditional Independence Assumption

Conditional Independence Assumption

$$y(1), y(0) \perp D | x$$

$x :=$ pretreatment variables

Under the CIA

$$\mathbb{E}[y_i(0)|D_i = 0, x] = \mathbb{E}[y_i(0)|D_i = 1, x] = \mathbb{E}[y_i(0)|x]$$

$$\mathbb{E}[y_i(1)|D_i = 0, x] = \mathbb{E}[y_i(1)|D_i = 1, x] = \mathbb{E}[y_i(1)|x]$$

Conditional Independence Assumption

Causal Effects under CIA

$$
\begin{aligned}
\delta_x &= ATE \\
&= \mathbb{E}[\Delta_i | x] \\
&= \mathbb{E}[y_i | D_i = 1, x] - \mathbb{E}[y_i | D_i = 0, x]
\end{aligned}
$$

$$
\begin{aligned}
\tau &= ATT \\
&= \mathbb{E}[\Delta_i | D_i = 1] \\
&= \mathbb{E}[\mathbb{E}[\Delta_i | D_i = 1, x] | D_i = 1] \\
&= \mathbb{E}[\mathbb{E}[y_i | D_i = 1, x] - \mathbb{E}[y_i | D_i = 0, x] | D_i = 1] \\
&= \mathbb{E}[\delta_x | D_i = 1] \qquad \text{By LIE}
\end{aligned}
$$

Conditional Independence Assumption

Estimation Strategy under CIA

Estimation strategy for ATE and ATT under CIA:

- Stratify data by x-values.
- Compute ATE for each stratification.
- Average ATEs in population of treated units.

Essential difference between matching and regression are different weighting schemes.

Comparison of Matching and Regression

$$
\begin{aligned}
ATT = \Delta_M &= \mathbb{E}[y_i(1) - y_i(0)|D_i = 1] \\
&= \delta_0 Pr(x_i = 0|D_i = 1) + \delta_1 Pr(x_i = 1|D_i = 1) \\
&= \delta_0 \frac{Pr(D_i = 1|x_i = 0)Pr(x_i = 0)}{Pr(D_i = 1)} \\
&+ \delta_1 \frac{Pr(D_i = 1|x_i = 1)Pr(x_i = 1)}{Pr(D_i = 1)}
\end{aligned}
$$

Where weights are proportional to probability of treatment at each covariate value.

$$
\begin{aligned}
\delta_0 &= \mathbb{E}[y_i(1) - y_i(0)|D_i = 1, x_i = 0] \\
\delta_1 &= \mathbb{E}[y_i(1) - y_i(0)|D_i = 1, x_i = 1]
\end{aligned}
$$

Comparison of Matching and Regression

This compares to (fully saturated) regression

$$y_i = \alpha + \beta x_i + \Delta r D_i + \varepsilon_i$$

$$\begin{aligned}\Delta r &= \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|x_i])y_i]}{\mathbb{E}[(D_i - \mathbb{E}[D_i|x_i)D_i]} \\ &= \delta_0 \frac{Pr(D_i = 1|x_i = 0)[1 - Pr(D_i = 1|x_i = 0)]Pr(x_i = 0)}{\mathbb{E}[Pr(D_i = 1|x_i)[1 - Pr(D_i = 1|x_i)]]} \\ &+ \delta_1 \frac{Pr(D_i = 1|x_i = 1)[1 - Pr(D_i = 1|x_i = 1)]Pr(x_i = 1)}{\mathbb{E}[Pr(D_i = 1|x_i)[1 - Pr(D_i = 1|x_i)]]}\end{aligned}$$

Where weights are proportional to variance of D at each value of covariate

$\Rightarrow$ Highest variance of probability of $D = 0.5$

Comparison of Matching and Regression

Feasibility of matching and regression estimators?

- Regression: Curse of dimensionality $\Rightarrow$ non-saturated model not a solution
- Matching: Problem if in some stratification only treated or only controls

$\Rightarrow$ Propensity score mitigates dimensionality problem.

Outline

Propensity Score Matching (PSM)

The Two PSM Lemmas

$$\rho(x) = Pr(D = 1|x) = \mathbb{E}[D|x]$$

Lemma 1:

$$D \perp x|\rho(x)$$

Proof:

$$\begin{aligned} Pr(D = 1|x, \rho(x)) &= \mathbb{E}[D|x, \rho(x)] \\ &= \mathbb{E}[D|x] \\ &= Pr(D = 1|x) \\ &= \rho(x) \end{aligned}$$

Propensity Score Matching (PSM)

The Two PSM Lemmas

$$
\begin{aligned}
Pr(D = 1|\rho(x)) &= \mathbb{E}[D|\rho(x)] \\
&= \mathbb{E}[\mathbb{E}[D|x, \rho(x)]|\rho(x)] \\
&= \mathbb{E}[\rho(x)|\rho(x)] \\
&= \rho(x)
\end{aligned}
$$

$$\Rightarrow Pr(D = 1|x, \rho(x)) = Pr(D = 1|\rho(x)) \blacksquare$$

Propensity Score Matching (PSM)

The Two PSM Lemmas

Lemma 2:

$$y(1), y(0) \perp D|x \Rightarrow y(1), y(0) \perp D|\rho(x)$$

Proof:

$$\begin{aligned} Pr(D=1|y(1), y(0), \rho(x)) &= \mathbb{E}[D|y(1), y(0), \rho(x)] \\ &= \mathbb{E}[\mathbb{E}[D|x, y(1), y(0)]|y(1), y(0), \rho(x)] \\ &= \mathbb{E}[\mathbb{E}[D|x]|y(1), y(0), \rho(x)] \\ &= \mathbb{E}[\rho(x)|y(1), y(0), \rho(x)] = \rho(x) \end{aligned}$$

$$Pr(D=1|\rho(x)) = \rho(x) \qquad \text{From Lemma 1}$$

$$\Rightarrow Pr(D=1|y(1), y(0), \rho(x)) = Pr(D=1|\rho(x)) \blacksquare$$

Outline

Motivation

Conditional Independence Assumption

Comparison of Matching and Regression

Propensity Score Matching (PSM)

- The Two PSM Lemmas
- Identification
- Estimation Strategy
- PSM Estimation Algorithm
- Estimation of ATT by Stratification on the Score
- Nearest Neighbor, Kernel and Radius
- Weighting on the Score

Propensity Score Matching (PSM)

Identification

$$\mathbb{E}[y_i(0)|D_i = 0, \rho(x_i)] = \mathbb{E}[y_i(0)|D_i = 1, \rho(x_i)] = \mathbb{E}[y_i(0)|\rho(x_i)]$$
$$\mathbb{E}[y_i(1)|D_i = 0, \rho(x_i)] = \mathbb{E}[y_i(1)|D_i = 1, \rho(x_i)] = \mathbb{E}[y_i(1)|\rho(x_i)]$$

$$\begin{aligned}\delta_{\rho(x)} &= \mathbb{E}[\Delta_i|\rho(x_i)] \\ &= \mathbb{E}[y_i(1) - y_i(0)|\rho(x_i)] \\ &= \mathbb{E}[y_i(1)|D_i = 1, \rho(x_i)] - \mathbb{E}[y_i(0)|D_i = 0, \rho(x_i)] \\ &= \mathbb{E}[y_i|D_i = 1, \rho(x)_i] - \mathbb{E}[y_i|D_i = 0, \rho(x_i)]\end{aligned}$$

$$\begin{aligned}ATT = \tau &= \mathbb{E}[\Delta_i|D_i = 1] \\ &= \mathbb{E}[\mathbb{E}[\Delta_i|D_i = 1, \rho(x_i)]] \\ &= \mathbb{E}[\mathbb{E}[y_i(1)|D_i = 1, \rho(x_i)] - \mathbb{E}[y_i(0)|D_i = 0, \rho(x_i)]|D_i = 1] \\ &= \mathbb{E}[\delta_{\rho(x)}|D_i = 1]\end{aligned}$$

Outline

Propensity Score Matching (PSM)

Estimation Strategy

1. Estimate $\rho(x)$
2. Estimate ATE given $\rho(x)$; actual matching
   - Stratification on $\rho(x)$
   - Nearest neighbour
   - Radius matching
   - Kernel matching
   - weighting based on $\rho(x)$

$\Rightarrow$ No dimensionality problem, thanks to Lemma 1.

Outline

Propensity Score Matching (PSM)

PSM Estimation Algorithm

1. Start with parsimonious logit or probit.
2. Sort data from low to high.
3. Sorting all observations into blocks such that treated and controls not statistically different.
4. Test whether balancing property holds in all blocks for all covariates.
   a) Test whether means of all covariates in each block different from zero for treated and controls.
   b) If one covariate not balanced in ONE block, split the block.
   c) If one covariate not balanced in ALL blocks, add interactions and/or higher-order terms to logit.

Tips:

- $\rho(x)$-ranges for treated and controlled ideally the same.
- Frequency of treated and controls in all blocks the same.

Outline

Estimation of ATT by Stratification on the Score

$$\tau_f^s = \frac{\sum_{i \in I(f)} y_i^T}{N_f^T} - \frac{\sum_{j \in I(f)} y_j^C}{N_f^C} \qquad f \text{ is index for strate}$$

$$\tau^s = \sum_{f=1}^{Q} \tau_f^s \frac{\sum_{i \in I(f)} D_i}{\sum_{\forall i} D_i}$$

$$Var[\tau_f^s] = \frac{Var[y_f^T]}{N_f^T} + \frac{Var[y_f^C]}{N_f^C}$$

$$Var[\tau^s] = \frac{1}{N^T}\left[Var[y_i^T] + \sum_{f=1}^{Q} \frac{N_f^T}{N^T}\frac{N_f^T}{N_f^C} Var[y_j^C]\right]$$

$$\overset{\text{from}}{=} \sum_{f=1}^{Q} \left(\frac{N_f^T}{N^T}\right)^2 Var[\tau_f^s]$$

- Irrelevant controls for ATT: discard controls with $\rho(x) >$ max or $<$ min.
- Penalty for unequal number of treated and controls in a block by increasing variance.

Outline

Nearest Neighbor, Kernel and Radius

Nearest neighbor closest to ideal,
but what if nearest neighbor already used?

- Nearest neighbor or radius matching with replacement.
- Kernel matching: every treated is matched with weighted average of all controls with weights that are inversely proportional to the distance Q between the two.

Outline

Weighting on the Score

Lemma 1

Lemma 1: (ATE)

$$y(1), y(0) \perp D|x$$

$$\Rightarrow w = \mathbb{E}[y_i(1)] - \mathbb{E}[y_i(0)]$$
$$= \mathbb{E}[y_i D_i|\rho(x_i)] - \mathbb{E}[y_i(1-D_i)|(1-\rho(x_i))]$$

Proof:

$$\mathbb{E}[y_i D_i|\rho(x_i)] - \mathbb{E}[y_i(1-D_i)|(1-\rho(x_i))]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{y_i D_i}{\rho(x_i)}|x\right] - \mathbb{E}\left[\frac{y_i(1-D_i)}{1-\rho(x_i)}|x\right]\right]$$

Weighting on the Score

Lemma 1

Using definition of $\rho(x)$ and fact that CIA makes conditioning on treatment irrelevant in two internal expectations, we have:

$$\mathbb{E}[\mathbb{E}[y_i(1)|x] - \mathbb{E}[y_i(0)|x]] = \mathbb{E}[y_i(1)] - \mathbb{E}[y_i(0)] \blacksquare$$

Weighting on the Score

Lemma 2

Lemma 2: (ATT)

$$y(1), y(0) \perp D | x$$

$$\Rightarrow \tau = \mathbb{E}[y_i(1)|D_i = 1] - \mathbb{E}[y_i(0)|D_i = 1]$$
$$= \mathbb{E}[y_i D_i] - \mathbb{E}\left[y_i(1 - D_i)\frac{\rho(x_i)}{1 - \rho(x_i)}\right]$$

Weighting on the Score

Lemma 1

Proof:

$$\mathrm{E}[y_i D_i] - \mathbb{E}\left[y_i(1-D_i)\frac{\rho(x_i)}{1-\rho(x_i)}\right]$$

$$=\mathrm{E}\left[\mathbb{E}[y_i D_i|x] - \mathbb{E}\left[y_i(1-D_i)\frac{\rho(x_i)}{1-\rho(x_i)}|x\right]\right]$$

$$=\mathrm{E}[\mathbb{E}[y_i(1)|D_i=1,x]Pr(D_i=1|x)$$
$$-\mathrm{E}\left[y_i(0)\frac{\rho(x_i)}{1-\rho(x_i)}|D_i=1,x\right]Pr(D_i=0|x)$$

Using the definition of $\rho(x_i)$ and CIA makes conditioning irrelevant,

$$\mathbb{E}[\mathbb{E}[y_i(1)|D_i=1,x] - \mathbb{E}[y_i(0)|D_i=1,x]|D_i=1]$$
$$=\mathbb{E}[y_i(1)|D_i=1] - \mathbb{E}[y_i(0)|D_i=0] \blacksquare$$

Weighting on the Score

- Problem with Weighting:
  Sensitive to method $\rho(x)$ is estimated
- Advantage of Weighting:
  Does not rely on stratification or matching
- Automated covariate balancing *i* genetic matching and entropy balancing.

$\Rightarrow$ Overall: Matching generally preferable to OLS