Econometrics 2

Matching Methods and Propensity Scores

Sergey Lychagin

Central European University

Winter 2020

Review

- What is the potential outcome model for a binary treatment D and outcome Y?
- How is the treatment effect defined in the potential outcome model?
- Can we estimate an individual treatment effect?
- If we have control variables X, what is the conditional independence assumption (CIA)?

Estimation of Treatment Effects

Suppose we have a binary treatment model with treatment variable D_i and potential outcomes $\{Y_{1i}, Y_{0i}\}$. In addition, we have a number of control variables in X_i. Suppose further that the CIA holds

$$\{Y_{1i}, Y_{0i}\} \perp D_i | X_i$$

How can we estimate the treatment effect $E\left[Y_i|X_i, D_i = 1\right] - E\left[Y_i|X_i, D_i = 0\right] = E\left[Y_{1i} - Y_{0i}|X_i\right]$?

- linear regression approximates $E\left[Y_i|X_i, D_i\right]$
- TODAY: non-linear alternatives

What are we estimating?

Average Treatment Effect

$$\begin{aligned}\delta_{ATE} &= E\left[Y_{1i} - Y_{0i}\right] \\ &= E\{E\left[Y_{1i} - Y_{0i}|X_i\right]\}\end{aligned}$$

Average Treatment Effect on the Treated

$$\begin{aligned}\delta_{TOT} &= E\left[Y_{1i} - Y_{0i}|D_i = 1\right] \\ &= E\{E\left[Y_{1i} - Y_{0i}|X_i\right]|D_i = 1\}\end{aligned}$$

Matching Estimators

Define δ_X as

$$\delta_X := E\left[Y_i | X_i, D_i = 1\right] - E\left[Y_i | X_i, D_i = 0\right]$$

Matching estimators for ATE and TOT are given by

$$\begin{aligned} \delta_{TOT} &= \sum_x \delta_X P\left(X_i = x | D_i = 1\right) \\ \delta_{ATE} &= \sum_x \delta_X P\left(X_i = x\right) \end{aligned}$$

(to make things simple, assume X is discrete).
$P\left(X_i = x\right)$ and $P\left(X_i = x | D_i = 1\right)$ can be seen as different weights for combining the effects δ_X.

Example

We are interested in the effect of smoking on mortality. Evaluation strategy: compare mortality rates of smokers and non-smokers. Available covariate: Age groups. Selection problem: mean age of smokers < mean age of non-smokers

Age	Smokers	Non-Smokers	# obs	# Smokers
<30	$\bar{Y}_{S,30}$	$\bar{Y}_{N,30}$	n_1	ns_1
30-40	$\bar{Y}_{S,40}$	$\bar{Y}_{N,40}$	n_2	ns_2
40-50	$\bar{Y}_{S,50}$	$\bar{Y}_{N,50}$	n_3	ns_3
>50	$\bar{Y}_{S,60}$	$\bar{Y}_{N,60}$	n_4	ns_4

$$\begin{aligned}
\delta_{TOT} &= (\bar{Y}_{S,30} - \bar{Y}_{N,30})\frac{ns_1}{NS} + (\bar{Y}_{S,40} - \bar{Y}_{N,40})\frac{ns_2}{NS} + ... \\
\delta_{ATE} &= (\bar{Y}_{S,30} - \bar{Y}_{N,30})\frac{n_1}{N} + (\bar{Y}_{S,40} - \bar{Y}_{N,40})\frac{n_2}{N} + ...
\end{aligned}$$

This allows for **arbitrary**, nonlinear relationship between δ_X and X.

1. We **matched** observations based on X.
2. We estimated δ_X for each X.
3. We put the estimates together to form δ_{ATE} and δ_{TOT}.

Questions:

1. Why bother? Why is this better than linear regression + OLS?
2. What if X is so multidimensional that we have too few obs. in each cell (or some cells) to estimate δ_X?

Matching and Regression

Define as $d_{ix} = 1\left[X_i = x\right]$ a dummy variable indicating $X_i = x$ and consider

$$Y_i = \sum_x \alpha_x d_{ix} + \sum_x \beta_x D_i d_{ix} + \varepsilon_i$$

This is a **saturated** model.

Note that $\delta_x = \beta_x$ (why?) and $\widehat{\beta}_x^{OLS}$ is consistent! Therefore, OLS estimates δ_X.

Matching and Regression

Now, consider this model

$$Y_i = \sum_x d_{ix}\alpha_x + \delta_R D_i + u_i$$

This is called a model **saturated-in-X**.
OLS applied to this model differs from the matching estimators in the weights used to combine the δ_X:

- δ_{TOT} puts most weight on X-cells where it is most likely to be treated
- δ_R puts most weight on X-cells where the variance of treatment is highest

Proof

Recall the regression anatomy formula:

$$\delta_R = \frac{cov(Y_i, \tilde{D}_i)}{V[\tilde{D}]}, \text{ where } \tilde{D}_i = D_i - E[D|X = X_i]$$

Let's express δ_R via δ_X:

$$\begin{aligned}
\delta_R &= \frac{\sum_x E(Y_i \tilde{D}_i | X_i = x) P(X_i = x)}{V[\tilde{D}]} \\
&= \frac{\sum_x E((E[Y_i | X_i = x, D_i = 0] + \delta_x D_i)\tilde{D}_i | X_i = x) P(X_i = x)}{V[\tilde{D}]} \\
&= \frac{\sum_x \delta_x E(D_i \tilde{D}_i | X_i = x) P(X_i = x)}{V[\tilde{D}]} \\
&= \frac{\sum_x \delta_x E(\tilde{D}_i^2 | X_i = x) P(X_i = x)}{V[\tilde{D}]} \\
&= \sum_x \delta_x \frac{V(D_i | X_i = x)}{V[\tilde{D}]} P(X_i = x)
\end{aligned}$$

Propensity Score

The propensity score is defined as

$$p(X_i) := E\left[D_i | X_i\right] = P\left[D_i = 1 | X_i\right]$$

Theorem (Propensity Score Theorem)

Suppose the CIA holds $\{Y_{1i}, Y_{0i}\} \perp D_i | X_i$, *then*

$$\{Y_{1i}, Y_{0i}\} \perp D_i | p(X_i)$$

Proof: show that $P\left[D_i = 1 | Y_{ji}, p(X_i) = p\right]$ does not depend on Y_{ji} with $j = 0, 1$.

Propensity Score Theorem - Proof

Proof.

$$\begin{aligned}
P\left[D_i = 1 | Y_{ji}, p(X_i) = p\right] &= E\left[D_i | Y_{ji}, p(X_i) = p\right] \\
&= E\{E\left[D_i | Y_{ji}, p(X_i), X_i\right] | Y_{ji}, p(X_i) = p\} \\
&= E\{E\left[D_i | Y_{ji}, X_i\right] | Y_{ji}, p(X_i) = p\} \\
&= E\{E\left[D_i | X_i\right] | Y_{ji}, p(X_i) = p\} \\
&= E\{p(X_i) | Y_{ji}, p(X_i) = p\} = p
\end{aligned}$$

□

Propensity Score Theorem - Practical Implications

The propensity score theorem reduces the dimension of the matching problem to a single variable $p(X_i)$. This motivates a 2-step estimation procedure

1. Estimate the propensity score
2. Generate a matching estimator based on the propensity score

Application

Rosenbaum + Rubin JASA (1984) "Reducing Bias in Observational Studies Using Sub-classification on the Propensity Score"

Medical vs. surgical treatment of coronary artery disease

- $D_i = 1$ patient receives bypass surgery $N_T = 590$
- $D_i = 0$ patient receives medical therapy $N_C = 925$

Outcome variables: survival rates, health improvements
Selection: patients with worse health are more likely treated with surgery

Estimating the Propensity Score

In total 74 covariates available, this means many options to specify a propensity score estimator!
What is a good estimator for the propensity score?
The propensity score theorem

$$\{Y_{1i}, Y_{0i}\} \perp D_i | X_i \Rightarrow \{Y_{1i}, Y_{0i}\} \perp D_i | p(X_i)$$

implies that observations with "similar" values of the propensity score should also have "similar" X-characteristics.

Estimating the Propensity Score

Subclassification Algorithm

- Estimate a parsimonious logit for $P(D_i = 1|X_i)$
- Stratify the data by quintile blocks of $\hat{p}(X_i)$
- Compare $\bar{X}_T - \bar{X}_C$ in each block. Use a t-test or F-test of significant differences in means
  1. if X_i are balanced in each block STOP
  2. if not balanced, divide block in 2 parts and re-evaluate
  3. if X_i not balanced in all blocks re-specify the logit: add interaction terms and polynomials of variables with high F-/t-stats.
- Rosenbaum + Rubin "5 subclasses constructed from the propensity score are sufficient to remove 90% of the bias"

Tests of Balance

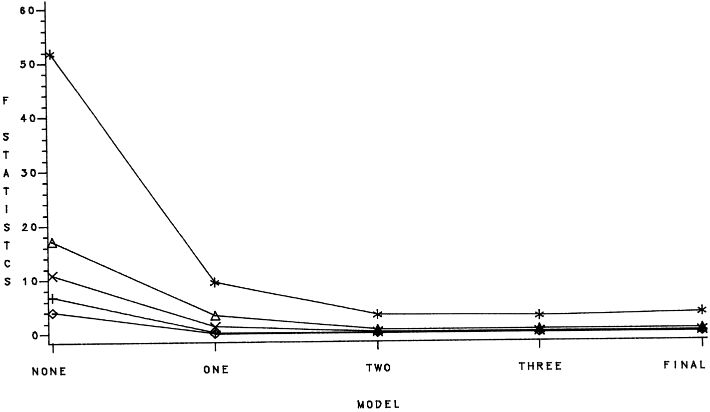


Figure 1. F Tests of Balance Before and After Subclassifications: Main Effects (5-point summary). (Minimum ◇; lower quartile +; median ×; upper quartile △; maximum ∗.)

Balance for single variables

Good:

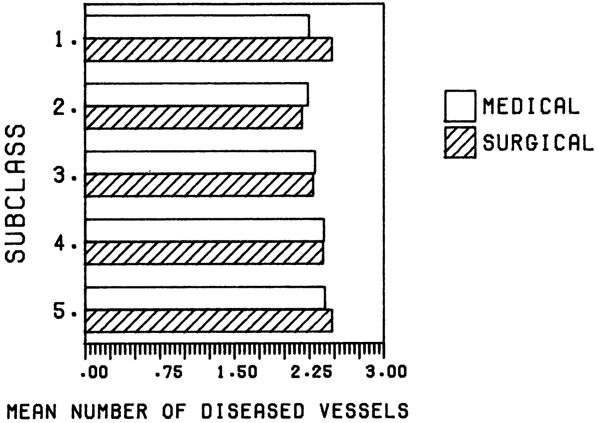
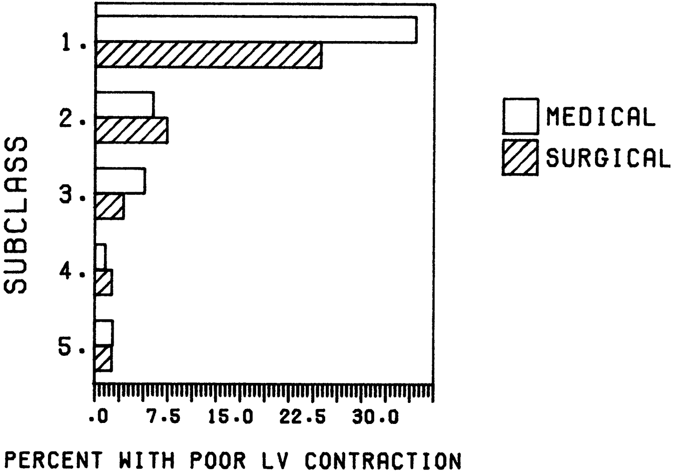


Figure 3. Balance Within Subclasses: Number of Diseased Vessels.

Balance for single variables

Not so good:



Propensity Score as diagnostic tool

Compare box-plots or histograms of the distribution of $\hat{p}(X_i)$ for observations with $D_i = 0$ and $D_i = 1$.
Is there sufficient overlap in the distributions?
If observable variables X_i are very different among treated and controls, it is also more likely that unobservables differ a lot.

Perfect prediction of D_i is a **bad sign**! E.g., imagine $D_i =$ surgery for all patients in subclass 1. Cannot compare their outcomes to similar patients undergoing therapy.

Box Plots

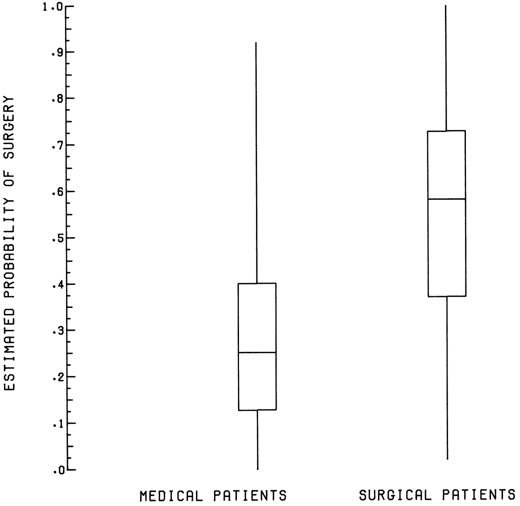


Figure 6. Boxplots of the Estimated Propensity Score.

Matching Estimates for ATE

Many ways to implement the matching . The easiest — follow the age-smoking example:

- Find treatment effect δ_c for each subclass c.
- Combine δ_c to form δ_{ATE} (or δ_{TOT}, δ_{TOU}, etc).

Matching Estimates for ATE

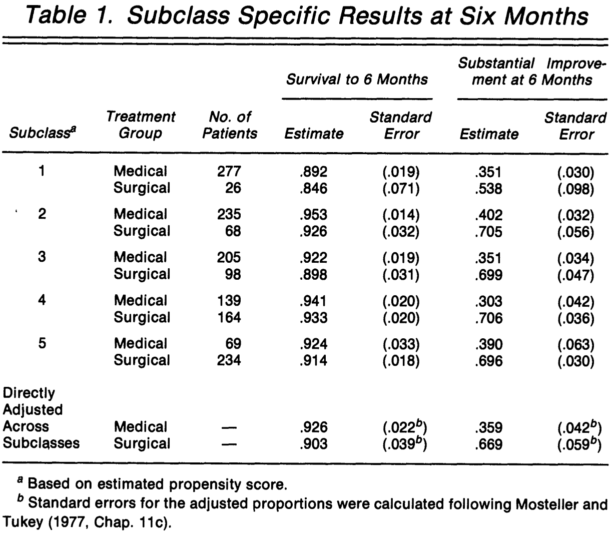


Table 1. Subclass Specific Results at Six Months

			Survival to 6 Months		Substantial Improvement at 6 Months	
Subclass[a]	Treatment Group	No. of Patients	Estimate	Standard Error	Estimate	Standard Error
1	Medical	277	.892	(.019)	.351	(.030)
	Surgical	26	.846	(.071)	.538	(.098)
2	Medical	235	.953	(.014)	.402	(.032)
	Surgical	68	.926	(.032)	.705	(.056)
3	Medical	205	.922	(.019)	.351	(.034)
	Surgical	98	.898	(.031)	.699	(.047)
4	Medical	139	.941	(.020)	.303	(.042)
	Surgical	164	.933	(.020)	.706	(.036)
5	Medical	69	.924	(.033)	.390	(.063)
	Surgical	234	.914	(.018)	.696	(.030)
Directly Adjusted Across Subclasses	Medical	—	.926	(.022[b])	.359	(.042[b])
	Surgical	—	.903	(.039[b])	.669	(.059[b])

[a] Based on estimated propensity score.

[b] Standard errors for the adjusted proportions were calculated following Mosteller and Tukey (1977, Chap. 11c).

Matching Estimates

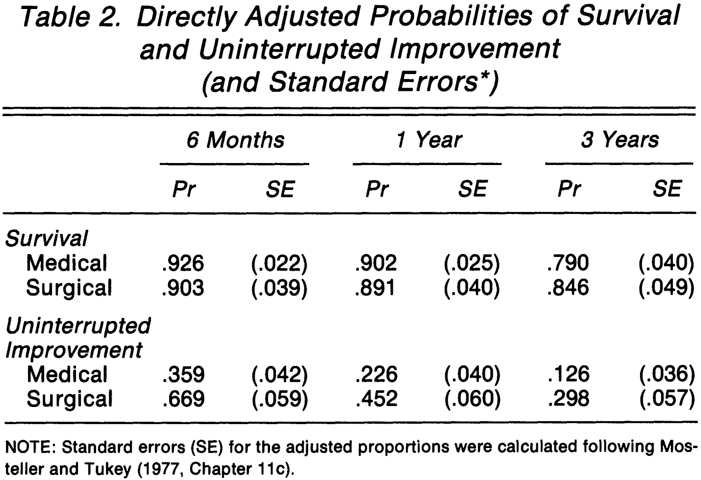


Table 2. Directly Adjusted Probabilities of Survival and Uninterrupted Improvement (and Standard Errors)*

	6 Months		*1 Year*		*3 Years*	
	Pr	*SE*	*Pr*	*SE*	*Pr*	*SE*
Survival						
Medical	.926	(.022)	.902	(.025)	.790	(.040)
Surgical	.903	(.039)	.891	(.040)	.846	(.049)
Uninterrupted Improvement						
Medical	.359	(.042)	.226	(.040)	.126	(.036)
Surgical	.669	(.059)	.452	(.060)	.298	(.057)

NOTE: Standard errors (SE) for the adjusted proportions were calculated following Mosteller and Tukey (1977, Chapter 11c).

Matching Estimates

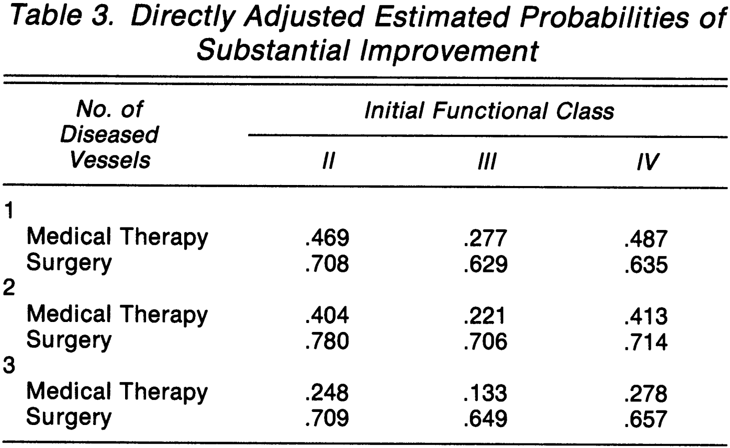


Table 3. Directly Adjusted Estimated Probabilities of Substantial Improvement

No. of Diseased Vessels	*Initial Functional Class*		
	II	*III*	*IV*
1			
Medical Therapy	.469	.277	.487
Surgery	.708	.629	.635
2			
Medical Therapy	.404	.221	.413
Surgery	.780	.706	.714
3			
Medical Therapy	.248	.133	.278
Surgery	.709	.649	.657

Another way to implement matching: K-nearest neighbors

The above example used **stratification** matching (by subclass).

Popular alternative — **k nearest neighbors**. Suppose we somehow estimated the propensity score $\widehat{p}(X_i)$. Let's find δ_{TOT}:

1. For every treated i, find K untreated observations closest to i in terms of $\widehat{p}$; call them C_i.
2. Construct a counterfactual for i: $\widehat{Y}_{0i} = \frac{1}{K} \sum\limits_{j \in C_i} Y_j$
3. Estimated treatment effect for i: $Y_i - \widehat{Y}_{0i}$. Treatment effect on the treated:

$$\widehat{\delta}_{TOT} = \frac{1}{N_{treated}} \sum_{i:D_i=1} \left(Y_i - \widehat{Y}_{0i} \right)$$

K nearest neighbors, k=2

i	D_i	$\widehat{p}_i(X_i)$	Y_i
1	0	0.01	0.4
2	1	0.05	$\boxed{2.1}$
3	1	0.12	$\boxed{1.8}$
4	0	0.12	–0.1
5	1	0.23	$\boxed{0.9}$
6	0	0.31	1.3
7	0	0.33	0.2
8	1	0.52	–0.2
9	0	0.61	1.7
10	1	0.83	1.1

i	D_i	$\widehat{p}_i(X_i)$	Y_i
1	0	0.01	0.4
2	1	0.05	2.1
3	1	0.12	1.8
4	0	0.12	–0.1
5	1	0.23	0.9
6	0	0.31	1.3
7	0	0.33	0.2
8	1	0.52	–0.2
9	0	0.61	1.7
10	1	0.83	1.1

$$\widehat{Y}_{02} = \widehat{Y}_{03} = \frac{0.4 - 0.1}{2}, \quad \widehat{Y}_{05} = \frac{1.3 + 0.2}{2}, \quad \text{etc..}$$

Warning: the gap between $\widehat{p}_{10}$ and $\widehat{p}_7$ is huge. It's safer to estimate ATE only for the subpopulation with $p >= 0.61$.

Yet another way: linear regression

Use OLS to estimate

$$Y_i = \alpha_0 + \alpha_1 \widehat{p}_i + (\beta_0 + \beta_1 \widehat{p}_i) D_i + \varepsilon_i$$

treatment effect: $\widehat{\delta}_p = \widehat{\beta}_0 + \widehat{\beta}_1 p$.
Or, use a more flexible approximation:

$$Y_i = \alpha_0 + \alpha_1 \widehat{p}_i + \alpha_2 \widehat{p}_i^2 + (\beta_0 + \beta_1 \widehat{p}_i + \beta_2 \widehat{p}_i^2) D_i + \varepsilon_i$$

Recall the saturated regression in the “effect of smoking by age” example — same logic here.

..And another one — propensity score weighting

Let's say we are looking for ATE:

$$\delta_{ATE} = E[Y_{1i} - Y_{0i}]$$

Note that

$$\begin{aligned} E\left[\frac{Y_i D_i}{p(X_i)}\right] &= E\left[E\left[\left.\frac{Y_i D_i}{p(X_i)}\right| X_i\right]\right] \\ &= E\left[p(X_i) E\left[\left.\frac{Y_i D_i}{p(X_i)}\right| X_i, D_i = 1\right] + (1 - p(X_i)) \cdot 0\right] \\ &= E\left[E\left[\left. Y_{1i} \right| X_i, D_i = 1\right]\right] = E\left[E\left[\left. Y_{1i}\right| X_i\right]\right] = E\left[Y_{1i}\right] \end{aligned}$$

Similarly, $E[Y_{0i}] = E\left[\frac{Y_i(1-D_i)}{1-p(X_i)}\right]$. Thus, $\delta_{ATE} = E\left[\frac{(D_i - p(X_i))Y_i}{p(X_i)(1-p(X_i))}\right]$
In finite samples,

$$\widehat{\delta}_{ATE} = \frac{1}{N}\sum_{i=1}^{N} \frac{(D_i - \widehat{p}(X_i))Y_i}{\widehat{p}(X_i)(1 - \widehat{p}(X_i))}$$

Concluding remarks

- There is a myriad of ways to implement matching. Some of them come with fancy names, lofty promises.
- But remember, there is no magic here. Matching is based on same exogeneity assumptions as OLS.
- Most importantly: **matching won't help if treatment depends on unobservables**.