Lecture 06 - Propensity Score Matching

Samuel DeWitt

Overview of Lecture

1) Curse of Dimensionality Redux
2) Evaluating Covariate Balance
3) Conditional Independence Assumption
4) Selection on (Un)Observables
5) Applied Example - NLSY97

Propensity Score Matching - Important Terms

As we discussed in a prior lecture, we use propensity score matching to sidestep curse of dimensionality issues we typically encounter in an exact 1-to-1 match procedure using just a small handful of simple variables.

As a means of review, the curse of dimensionality issue refers to a problem where additional variables in an exact matching procedure creating progressively smaller and smaller subsamples to match within.

Eventually, the data become so sparse in those subsamples that we cannot find equivalent cases in the control group to match to treated cases (and vice versa).

Propensity Score Matching - Curse of Dimensionality Redux

The underlying issue with the curse of dimensionality problem is that causal inference using matching is dependent upon satisfying an additional assumption known as **common support**.

In plain terms, the assumption of **common support** means that our treatment and control cases have *overlapping* values on the matching variable.

Overlapping means that, although one group is treated and the other is not, cases among either group can have similar (or the very same) values on some other variable we use for matching.

Curse of Dimensionality - Example

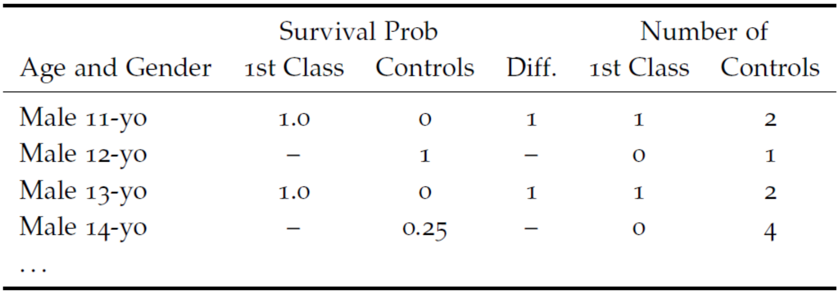
Let's say that we have a treatment (D) that we cannot assign randomly, and we know that age and gender have something to do with who gets treated and what their outcomes tend to be.

We want to then **match** individuals who were treated to those who were not on their exact age (in years) and gender.

To satisfy the **common support** assumption for our treatment effect, we need to have overlap across groups within each age-gender combination in the data. If we don't we have either adopt a different approach, or reduce the generalization of our treatment effect.

Curse of Dimensionality - Example

Here's an example of what that type of distribution may look like:



	Survival Prob			Number of	
Age and Gender	1st Class	Controls	Diff.	1st Class	Controls
Male 11-yo	1.0	0	1	1	2
Male 12-yo	–	1	–	0	1
Male 13-yo	1.0	0	1	1	2
Male 14-yo	–	0.25	–	0	4
...					

Can you locate the evidence in the table that the common support assumption is violated?

Sidestepping the Curse of Dimensionality

Propensity score matching avoids the **curse of dimensionality** by reducing the many variables we *want* to match on to a **single** value.

This is accomplished using a logistic regression to predict the probability that a case is **treated** (i.e., that $D = 1$).

Common support can then be evaluated by whether the groups overlap in the propensity to be treated - it is **not necessary** that each subcategory of the matching variables overlap.

Sidestepping the Curse of Dimensionality

But, we now have an issue - how are we to be sure that the matching procedure produced groups that are actually *similar* on the **matching variables**?

We now need to evaluate what is known as **covariate balance**.

This is accomplished by comparing matched groups on the **averages** of matching variables.

Evaluating Covariate Balance

To say that a characteristic is **balanced** across the **treated** and **untreated** groups we evaluate whether average values are *statistically equivalent* across the groups.

They **do not** have to be exactly equivalent, merely similar enough such that the **average** person in the treated group looks about the same as the **average** person in the untreated group.

Evaluating Covariate Balance

There is one primary way to evaluate covariate balance - estimating **mean comparison** tests (t-tests) across groups pre- and post-matching.

We *expect* there to be difference pre-matching and if our matching procedure works well, these differences should be much smaller across **matched** groups.

Evaluating Covariate Balance - Example

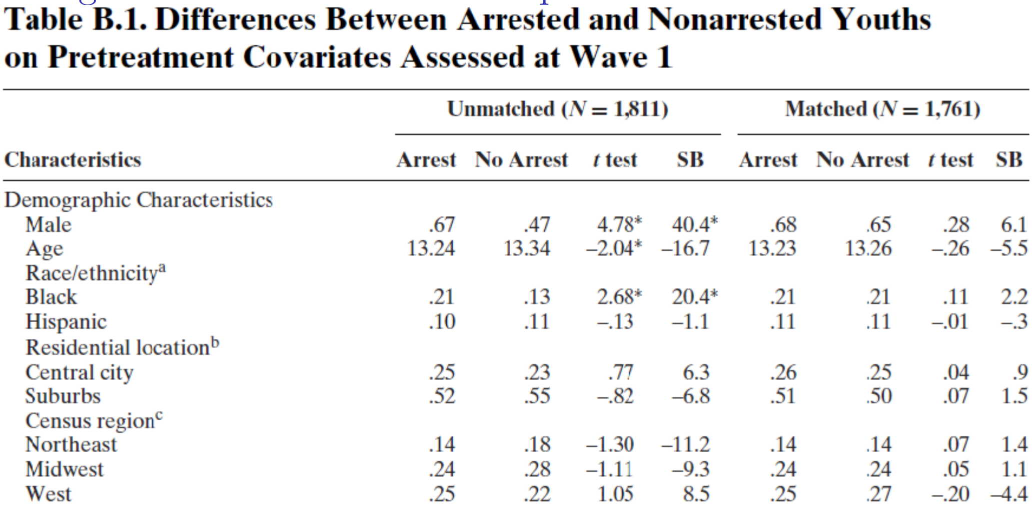


Table B.1. Differences Between Arrested and Nonarrested Youths on Pretreatment Covariates Assessed at Wave 1

	Unmatched ($N = 1{,}811$)				Matched ($N = 1{,}761$)			
Characteristics	**Arrest**	**No Arrest**	***t* test**	**SB**	**Arrest**	**No Arrest**	***t* test**	**SB**
Demographic Characteristics								
Male	.67	.47	4.78*	40.4*	.68	.65	.28	6.1
Age	13.24	13.34	–2.04*	–16.7	13.23	13.26	–.26	–5.5
Race/ethnicity[a]								
Black	.21	.13	2.68*	20.4*	.21	.21	.11	2.2
Hispanic	.10	.11	–.13	–1.1	.11	.11	–.01	–.3
Residential location[b]								
Central city	.25	.23	.77	6.3	.26	.25	.04	.9
Suburbs	.52	.55	–.82	–6.8	.51	.50	.07	1.5
Census region[c]								
Northeast	.14	.18	–1.30	–11.2	.14	.14	.07	1.4
Midwest	.24	.28	–1.11	–9.3	.24	.24	.05	1.1
West	.25	.22	1.05	8.5	.25	.27	–.20	–4.4

Evaluating Covariate Balance - Example

I'll draw your attention to two main differences - percent male and percent black.

In the **unmatched** sample these characteristics are quite different - the **arrested** (treated) group is snignificantly more likely to be male and black as compared to the **unarrested** (untreated group).

What happens after we **match** youth on their propensity scores to be arrested, though?

Evaluating Covariate Balance - Review

With propensity score matching, we sidestep the **curse of dimensionality** issue by collapsing all variables into a single probability to match with.

Ths does not, by itself, guarantee the groups are similar in the underlying characteristics we use in the logistic regression model.

We evaluate **covariate balance** by conducting **t-tests** on the matching variables pre- and post-matching.

Propensity Score Matching - Conditional Independence

The overarching procedure of estimating a propensity score and then matching on it is to accomplish a particular outcome - that of **conditional independence**.

What this means is that, *conditional* on the propensity score, treated versus untreated units are no different with respect to any variable that could have influenced both selection into being treated and the outcome of interest.

This assumption is generally referred to as CIA (conditional independence assumption) and is the topic of some controversy regarding the use of propensity score matching in the social sciences.

Propensity Score Matching - Conditional Independence (cont)

Why do think there's controversy surrounding this assumption?

Propensity Score Matching - Conditional Independence (cont)

Why do think there's controversy surrounding this assumption?

If you said there could be doubts about the logistic regression (selection) model, you are right.

In order for conditional independence to be true, we need to control for every relevant variable that has an impact on selection into treatment (and possibly the outcome of interest).

Propensity Score Matching - Conditional Independence (cont)

How plausible is it that we have met this assumption by controlling for everything that has an impact on selection?

Propensity Score Matching - Conditional Independence (cont)

How plausible is it that we have met this assumption by controlling for everything that has an impact on selection?

In many cases, not very plausible. This is because it would take a lot of effort (and time and $$$) to measure all the characteristics we *think* matter.

Further, even if we were to spend the time to measure everything we could think of that *could* matter, there are always characteristics that are either unmeasurable or insurmountably difficult to measure.

Propensity Score Matching - Selection on (Un)Observables

This leads us to important terms for propensity score matching: **selection on observables** and **selection on unobservables**.

Selection on observables implies that we can control for the influence of *observable* characteristics on selection into treatment.

By contrast, **selection on unobservables** means that we can control for the influence of *unobservable* characteristics on selection into treatment.

Propensity Score Matching - Selection on (Un)Observables

Propensity score matching **can** account for observable variables by **balancing** treated and untreated groups on observed and measurable characteristics we enter into the selection model.

It can only account for **unobserved** characteristics if the variables we enter into the selection model (i.e., the variables we **do** observe) have a perfect correlation with the **unobserved** variables we also want to control for.

By definition, we cannot ever be certain we have **balanced** both groups on measures we cannot observe, since our loose definition for **balance** is that group means on those variables are statistically similar.

Conditional Independence & Selection Redux

To review, the goal of propensity score matching is to achieve **conditional independence** across the treated and untreated groups. This is equivalent to saying that **treatment status** is as good as random assignment.

The extent to which we satisfy this assumption is contingent upon the variables we match on and whether they are good proxies for important variables we **cannot observe**.

However, we cannot ever formally test for **selection on unobservables** so this will remain an issue for all propensity score models.

Bringing it All Together - Sweeten & Apel (2007)

It's common knowledge that one of the purposes of prison is incapacitation.

That is, one of the reasons we sentence people to prison is to prevent them from committing more crimes in the near future.

Consistent with this purpose, it is often asked how many crimes we can *expect* to be avoided by sentencing someone to prison.

Bringing it All Together - Sweeten & Apel (2007)

Stated slightly differently, we want to know how many crimes someone sentenced to incarceration **would have** committed during their imprisonment had they instead not been incarcerated.

This maps pretty neatly onto the structure of **counterfactuals** we have already discussed and in terms of potential outcomes, we want to estimate the following effect where Y equals the number of crimes and D is a treatment indicator for incarcerated or not.

$$Y_{D=1} - Y_{D=0}$$

Since we know we **cannot** observe both outcomes for the same person, we need to find a different way to estimate this effect.

Bringing it All Together - Sweeten & Apel (2007)

However, estimating the effect of prison on future criminal behavior is difficult, at least in part, because we cannot randomly assign people to prison.

This means that it is reasonable to assume that people who go to prison are not a **random** sample, and we need to be very careful who we compare them to.

Therefore, we cannot simply compare an incarcerated person to a non-incarcerated person to figure out how many crimes we prevented by sending the former to prison.

Bringing it All Together - Sweeten & Apel (2007)

For the purposes of this example, we will focus on the 16 to 17-year old comparison in Sweeten & Apel (2007).

There were 113 youth in the NLSY97 who were incarcerated for the first time at 16 or 17-years old and 6,100 youth who were not incarcerated at that age.

Bringing it All Together - Sweeten & Apel (2007)

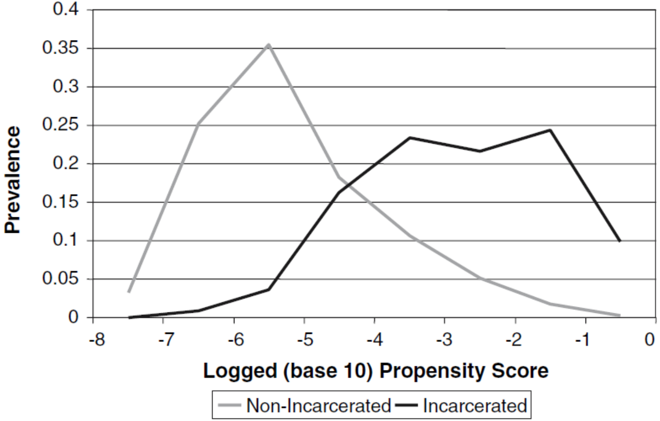
The authors first compute a direct comparison of means, then conduct simple matching which conditions the **untreated** subsample on being arrest or convicted at age 15.

The authors then conduct two types of propensity score matching: **nearest neighbor** and **kernel matching**, both with a *maximum distance* in the propensity score of .01.

In plain terms, **treated** and **untreated** cases are only matched together if their probability of being incarcerated is within .01 of each other.

Bringing it All Together - Sweeten & Apel (2007)

First, let's take a look at the common support assumption:



Bringing it All Together - Sweeten & Apel (2007)

This slide is just to have written down what I say in the prior slide:

1) Propensity score value distributions are **clearly** different for incarcerated and non-incarcerated youths. Non-incarcerated youths tend to have very low scores, while incarcerated youth have scores that are a bit higher.

2) That said, there is some overlap in the distributions - that's where we can find appropriate matches.

3) The ease of matching will be greatest where there is more overlap and vice versa.

Bringing it All Together - Sweeten & Apel (2007)

Now let's take a look at covariate balance:

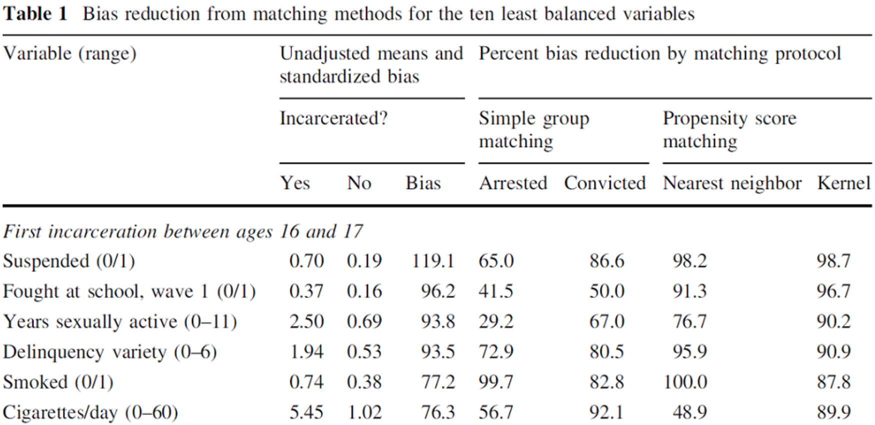


Table 1 Bias reduction from matching methods for the ten least balanced variables

Variable (range)	Unadjusted means and standardized bias			Percent bias reduction by matching protocol			
	Incarcerated?			Simple group matching		Propensity score matching	
	Yes	No	Bias	Arrested	Convicted	Nearest neighbor	Kernel
First incarceration between ages 16 and 17							
Suspended (0/1)	0.70	0.19	119.1	65.0	86.6	98.2	98.7
Fought at school, wave 1 (0/1)	0.37	0.16	96.2	41.5	50.0	91.3	96.7
Years sexually active (0–11)	2.50	0.69	93.8	29.2	67.0	76.7	90.2
Delinquency variety (0–6)	1.94	0.53	93.5	72.9	80.5	95.9	90.9
Smoked (0/1)	0.74	0.38	77.2	99.7	82.8	100.0	87.8
Cigarettes/day (0–60)	5.45	1.02	76.3	56.7	92.1	48.9	89.9

Bringing it All Together - Sweeten & Apel (2007)

Again, just putting what I am saying on the slide:

1) Bias is standardized by the shared standard deviation across both groups. A value of 119.1 indicates that incarcerated youth are 1.19 standard deviations above non-incarcerated youth.

2) Percent bias reduction is computed the same way as percent difference. We want to reduce bias to be below 20 in order to consider the covariate balanced in the matched sample.

3) It's clear that simple matching does not reduce bias enough for each variable, but either type of propensity score matching can reduce bias by about 90 to 99%.

Bringing it All Together - Sweeten & Apel (2007)

After matching, what happens to the group mean comparisons?

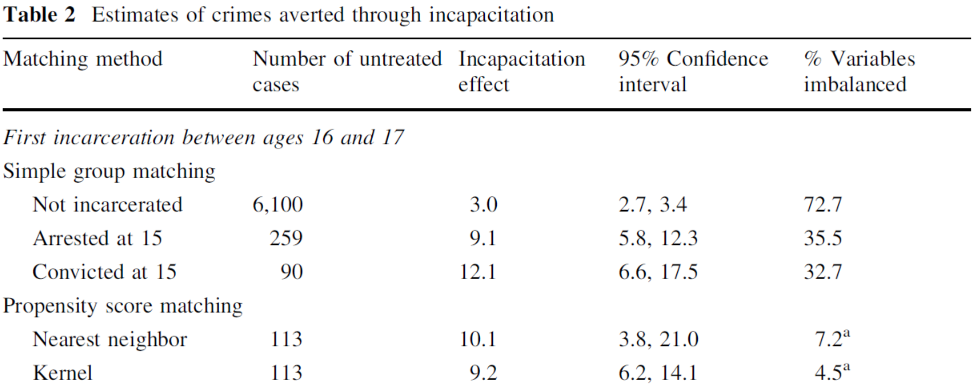


Table 2 Estimates of crimes averted through incapacitation

Matching method	Number of untreated cases	Incapacitation effect	95% Confidence interval	% Variables imbalanced
First incarceration between ages 16 and 17				
Simple group matching				
Not incarcerated	6,100	3.0	2.7, 3.4	72.7
Arrested at 15	259	9.1	5.8, 12.3	35.5
Convicted at 15	90	12.1	6.6, 17.5	32.7
Propensity score matching				
Nearest neighbor	113	10.1	3.8, 21.0	7.2[a]
Kernel	113	9.2	6.2, 14.1	4.5[a]

Bringing it All Together (Again) - Widdowson et al. (2016)

We tend to think that involvement in the criminal justice system affects educational milestones through preventing them altogether or delaying them.

Either effect is of concern, though the former is obviously much more serious.

Bringing it All Together (Again) - Widdowson et al. (2016)

But, again, we may not randomly assign youth to criminal justice interventions of interest - like arrest - so we have strong reasons to believe that arrested youth could already look pretty different than non-arrested youth in terms of their educational trajectories.

How, then, can we estimate the causal effect of an arrest on future educational achievement?

Bringing it All Together (Again) - Widdowson et al. (2016)

Again, we could use propensity score matching to ensure that, at least with respect to the characteristics we can observe about these youth, they are similarly likely to **have been** arrested, even if some youth do not go on to be arrested.

But, as detailed above, we need to examine the degree to which arrested and not arrested NLSY97 youth are similar along a host of covariates we **think** to be related to arrest and/or educational attainment.

Bringing it All Together (Again) - Widdowson et al. (2016)

Using the NLSY97 sample (again), Widdowson et al. (2016) match arrested and unarrested youth ($n = 1811$) across several dozen matching variables.

Let's briefly take a look at those (see pages 650-651).

Were the observed differences surprising, or unsurprising?

Bringing it All Together (Again) - Widdowson et al. (2016)

Can we evaluate common support here?

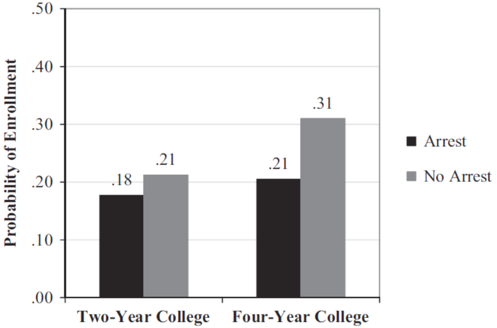
Slightly difficult, since they do not provide a lot of information about the range of propensity scores.

They do have some information, though - has anyone found it?

Bringing it All Together (Again) - Widdowson et al. (2016)

Now it’s time to see the results:

Figure 1. Predicted Probability of Postsecondary Enrollment Status Within 9 Months of High-School Graduation Date in the Matched Sample (*N* = 1,761)



Bringing it All Together (Again) - Widdowson et al. (2016)

And here are the regression model results:

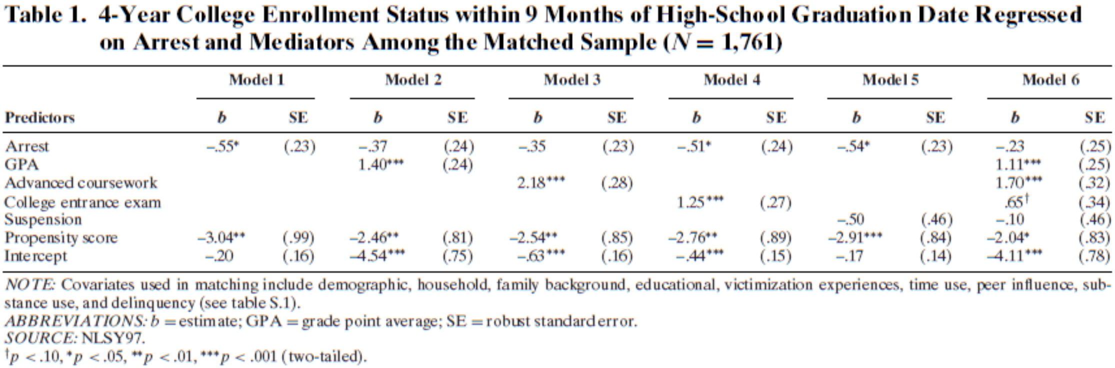


Table 1. 4-Year College Enrollment Status within 9 Months of High-School Graduation Date Regressed on Arrest and Mediators Among the Matched Sample ($N = 1{,}761$)

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
Predictors	***b***	**SE**	***b***	**SE**	***b***	**SE**	***b***	**SE**	***b***	**SE**	***b***	**SE**
Arrest	−.55*	(.23)	−.37	(.24)	−.35	(.23)	−.51*	(.24)	−.54*	(.23)	−.23	(.25)
GPA			1.40***	(.24)							1.11***	(.25)
Advanced coursework					2.18***	(.28)					1.70***	(.32)
College entrance exam							1.25***	(.27)			.65†	(.34)
Suspension									−.50	(.46)	−.10	(.46)
Propensity score	−3.04**	(.99)	−2.46**	(.81)	−2.54**	(.85)	−2.76**	(.89)	−2.91***	(.84)	−2.04*	(.83)
Intercept	−.20	(.16)	−4.54***	(.75)	−.63***	(.16)	−.44***	(.15)	−.17	(.14)	−4.11***	(.78)

NOTE: Covariates used in matching include demographic, household, family background, educational, victimization experiences, time use, peer influence, substance use, and delinquency (see table S.1).

ABBREVIATIONS: *b* = estimate; GPA = grade point average; SE = robust standard error.

SOURCE: NLSY97.

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$ (two-tailed).

Bringing it All Together (One Last Time)

To sum up, propensity score matching is useful in scenarios where we cannot randomly assign units to be treated, but have some background information on them that can help us to predict their likelihood of being treated.

The primary concerns of propensity score matching have to do with **conditional independence**, **common support**, and **covariate balance**.

Propensity score matching can satisfy **selection on observables** in most applications, but **selection on unobservables** always remains a concern.