Lecture 8: Balancing acts: Non-parametric and semi-parametric matching
DRAFT LECTURE NOTES
Meredith Fowlie

NB: Usual caveats apply. These notes are a work-in-progress. Errors, suggestions, clarifying questions welcome

1 Introduction

This lecture continues our discussion of causal inference absent experimental data, and conditioning strategies more specifically. We will maintain our focus on estimating the average effect of a program or treatment on an outcome, allowing for heterogeneous effects in the population.

A jumping off point for todays lecture is the "conditional unconfoundedness" assumption or "selection on observables". It implies that, within a population that is homogeneous in observable covariates X, potential outcomes are independent of D_i :

$$D_i \perp (Y_i(0), Y_i(1))|X_i$$

If we can assume conditional unconfoundedness (CU), our challenge is then to ensure that the covariates in X are balanced across treated and untreated units. Recall that if X is distributed differently across treatment and control groups, this can generate bias. Once differences in observed covariates across treated and control units have been removed, any differences in outcomes can be causally attributed to the intervention (given CU).

Even if conditional unconfoundedness is out of reach, it can still be very useful to condition carefully on observables. For example, if an unconditional means comparison suggests a large effect, but this effect disappears once you condition on observables, this is informative.

In what follows, we will work through alternative approaches to conditioning on observables and discuss some specific issues that arise when you actually get down to implementing these approaches.

1.1 Standard regression-based conditioning strategies

The first suite of strategies we will consider are based on parametric models of the relationship between the covariates and the potential outcomes. The basic idea: specify a parametric model

that can be used to impute the missing potential outcomes and use this model to predict what would have happened to a specific unit had this unit been exposed to the counterfactual treatment state.

First consider the simplest case where the "true" relationship between potential outcomes and the covariates in X is assumed to be linear and symmetric:

$$\begin{aligned} E[Y_i(0)|X_i &= x] = \alpha_C + \beta' x \equiv \mu_C(x) && (1)\\ E[Y_i(1)|X_i &= x] = \alpha_T + \beta' x \equiv \mu_T(x) && (2)\\ E[Y_i|X_i &= x, D_i = D] = \alpha_C + (\alpha_T - \alpha_C)D + \beta' x && (3)\\ &= \alpha_0 + \tau D + \beta' x && (4) \end{aligned}$$

We can also specify a slightly less restrictive model that accommodates separate regression regimes for $\mu_C(x)$ and $\mu_T(x)$. Consider the simple case of two linear relationships:

$$\begin{aligned} E[Y_i(0)|X_i &= x] = \alpha_C + \beta_C' x \equiv \mu_C(x) && (5)\\ E[Y_i(1)|X_i &= x] = \alpha_T + \beta_T' x \equiv \mu_T(x) && (6)\\ E[Y_i|X_i &= x, D_i = D] = \alpha_C + (\alpha_T - \alpha_C)D + \beta_C' x + (\beta_T - \beta_C)Dx + \varepsilon && (7)\\ &= \alpha_C + \tau D + \beta_C' x + \gamma' x \cdot D + \varepsilon && (8) \end{aligned}$$

Now suppose we estimate the following regression equation using observational data:

$$Y_i = \alpha_C + \tau D_i + \beta_C' X_i + \gamma' X_i \cdot D_i + \varepsilon_i \quad (9)$$

Question. Do we need to invoke more than conditional unconfoundedness, SUTVA, and the overlap condition in order to interpret τ as an unbiased estimate of the average treatment effect among individuals characterized by average values of X_i?

Yes!

In order to obtain an unbiased estimate, we must add another assumption/condition. If covariates are not identically distributed across treatment/control groups, we must assume that our model of the relationship between the potential outcomes and covariates is properly specified. Our desire to eschew this additional assumption will push us towards the non-parametric and semi-parametric alternatives discussed below. But first, let us understand why misspecification can bias our estimates.

Note that we can rewrite our average treatment effect estimator as a weighted average of our estimates of $\mu_C(x)$ and $\mu_T(x)$.Let's first think about our ATT estimate:

$$\begin{aligned}
\widehat{\tau}_T &= \frac{1}{PN}\sum_{i:D_i=1}(Y_i - (\widehat{\alpha}_C + X_i'\widehat{\beta}_C)) && (10)\\
&= \overline{Y}_T^{OBS} - \widehat{\alpha}_c - \overline{X}_T\widehat{\beta}_C && (11)\\
&= \overline{Y}_T^{OBS} - \widehat{\alpha}_c - \overline{X}_T\widehat{\beta}_C + \widehat{\alpha}_c + \overline{X}_c\widehat{\beta}_C - \overline{Y}_C^{OBS} && (12)\\
&= \overline{Y}_T^{OBS} - \overline{Y}_C^{OBS} - (\overline{X}_T - \overline{X}_C)'\widehat{\beta}_C && (13)
\end{aligned}$$

Similarly, constructing counterfactual outcomes for the control units....

$$\begin{aligned}
\widehat{\tau}_C &= \frac{1}{(1-P)N}\sum_{i:D_i=0}(\widehat{\alpha}_C + \widehat{\tau} + \widehat{\beta}_C'X_i + \widehat{\gamma}'X_i - Y_i) && (14)\\
&= \widehat{\alpha}_T + \overline{X}_c\widehat{\beta}_T - \overline{Y}_C^{OBS} && (15)\\
&= \widehat{\alpha}_T + \overline{X}_c\widehat{\beta}_T - \overline{Y}_C^{OBS} + \overline{Y}_T^{OBS} - \widehat{\alpha}_T - \overline{X}_T\widehat{\beta}_T && (16)\\
&= \overline{Y}_T^{OBS} - \overline{Y}_C^{OBS} - (\overline{X}_T - \overline{X}_C)'\widehat{\beta}_T && (17)
\end{aligned}$$

Thus, our OLS estimate of the average treatment effect:

$$\widehat{ATE} = P\left(\overline{Y}_T - \overline{Y}_C - (\overline{X}_C - \overline{X}_T)'\widehat{\beta}_C\right) - (1-P)\left(\overline{Y}_T - \overline{Y}_C - (\overline{X}_T - \overline{X}_C)'\widehat{\beta}_T\right) \quad (18)$$

Notice where our assumptions regarding the functional form of the relationships between the outcome variable and the covariates in X start to bite. The estimate of $\widehat{\beta}_C$ is estimated using data from the control sample where the average is $\overline{X}_C$. It provides a good approximation to the conditional mean function in the neighborhood around $\overline{X}_C$. But if this linear approximation is not globally accurate, regression adjustments may lead to bias when the linear function $X'\beta_C$ is used to predict counterfactual outcomes in the treated population. Similar arguments pertain to the $\widehat{\beta}_T$

Punchline: If our covariates are unbalanced (such that the averages of the covariates in the two treatment arms are very different), we are leaning on our parametric assumptions to construct our counterfactual. Without knowing how our assumed functional form differs from the true relationship, it is hard to sign the bias that may arise from unbalanced covariates.

Rather than lose sleep over misspecification, it behooves us to reduce the extent to which we are relying on these parametric assumptions by improving the extent to which covariates are balanced across treatment and control groups.

1.2 Covariate "matching": the very basics

Recall that the bias described in the previous section is not an issue when we were working with experimental data, provided we have a sufficiently large sample. In expectation, the covariates will be similarly distributed across the treatment and control groups. In practice, the difference $(\overline{X}_C - \overline{X}_T)$ should be small in large samples.

When working with non-experimental data, covariate "matching" tries to mimic these ideal conditions by aligning the distribution of the observed characteristics such that X is similarly distributed among the treated and control groups. The non-parametric and semi-parametric matching methods we'll be considering have a long history in non-experimental program evaluation. Most often, they are applied in settings where:

- The estimand of primary interest is the ATT
- There is a large reservoir of potential controls
- You have data on many (ideally all!) variables that determine both participation and outcomes.

Within a matching framework, the emphasis is more often placed on the ATT, so this is the estimand we will be focusing on. Prior to estimating the ATT, controls are selected from among the units that are untreated such that the selected are as similar as possible to the treated. How "similar" varies by matching method.

A general representation of the ATT matching estimator :

$$\widehat{\tau}_{ATT} = \frac{1}{PN} \sum_{i \in I_T} \{Y_i - \sum_{j \in I_C} \omega(i,j) Y_j\} \tag{19}$$

I_C and I_T denote a set of indices for treated and untreated observations, respectively. i indexes treated cases and j is the index over untreated.

The $\omega(i,j)$ is a positively valued weight that is defined such that, for each i:

$$\sum_{j \in I_C} w_T(i,j) = 1. \tag{20}$$

The weights are, in general, a decreasing function of the "distance" between the treated unit i and the control unit j. Different matching estimators take different approaches to constructing

these weights. Much of the matching literature focuses on approaches that eschew parametric restrictions on the relationship between X and the potential outcomes.

The matching estimator in [19] assumes cross-sectional matching. That is, observationally equivalent treatment and control units are matched, and observed outcomes are compared across these matches.

Of course, there may be systematic differences between treated and control outcomes even after conditioning on all observables. For example, suppose cross-sectional matching accounts for some- but not all- of the confounding factors that determine treatment assignment and potential outcomes. But suppose the remaining (unobserved) confounders are constant over time. If you have panel data, you can remove the effect of these time-invariant confounders by constructing a DID matching estimator.

Heckman, Ichimura and Todd (1997) and Heckman, Ichimura, Smith and Todd (1998) introduce a DID matching strategy that allows for temporally invariant differences across participants and non-participants.

Let t^0 denote a pre-treatment period and $t^{'}$ denote a post-treatment period. The DID matching estimator requires that:

$$E[Y_{t'}(1) - Y_{t^0}(0)|X, D = 1] = E[Y_{t'}(1) - Y_{t^0}(0)|X, D = 0]. \tag{21}$$

The estimator also requires that the covariate overlap condition is satisfied in both pre- and post-treatment periods. Note that this is a different, and generally weaker, assumption as compared to cross-sectional matching.

The generalized DID matching estimator is given by:

$$\tau_{DID} = \frac{1}{PN}\sum_{i\epsilon I_1} Y_{it'}(1) - Y_{it^0}(0)) - \sum_{j\epsilon I_0} \omega(i,j)(Y_{jt'}(0) - Y_{jt^0}(0))\}.$$

Although I made reference to "matching estimators", above, I think this language is somewhat misleading. The matching process occurs before any estimation happens. Matching is more accurately a *pre*-estimation data processing step that involves pruning the data based on exogenous, pre-determined covariates. Once this pruning is complete, the estimation can take a variety of forms. It may be a simple comparison of means. Alternatively, mutivariate regression can be used to address remaining imbalances.

1.3 Matching versus regression

Before we move onto the mechanics of non-parametric and semi-parametric matching, it is worth highlighting some of the similarities and differences, strengths and weaknesses, vis a vis parametric, regression based estimators.

Important similarities between covariate matching strategies and the parametric, regression-based estimator:

- Both compare outcomes at treated units and units assigned to the control state in the construction of estimates of the unobserved counterfactual.
- The validity of either approach to estimating average causal effects depends critically on the conditional independence assumption. Neither approach yields a solution to the causal inference problem if there is selection on unobservables that cannot be controlled for using fixed effects or differencing strategies.
- Matching estimators can be interpreted/implemented as a weighted regression. The essential difference lies in how the weights used to pool estimates at different values of the covariates are defined.

This last point is worth elaborating upon because it helps us understand why matching and regression-based strategies may yield very different results even when we condition on the very same covariates. It also provides another way to illustrate how/why the parametric assumptions imposed by linear regression can yield misleading causal effects estimates in the presence of treatment effect heterogeneity and unbalanced covariates.

1.3.1 Matching as weighted regression

Suppose that treated units in a population are not evenly distributed across sub-strata (where the sub-strata are defined in terms of observable covariates X).

If we take a matching approach, we will estimate the ATT using a weighted average of differences in outcomes across treatment and control subgroups, where the sub-groups are defined in terms of observable covariates X. The weights will be proportional to the probability of treatment at each value of the covariates.
A linear regression uses not only the number of units in the strata to weight the results, but also the conditional variance of the treatment variable at each value of the covariates. Recall that, if

the treatment is binary, the variance is $P(1-P)$ where P is the probability that a unit will receive treatment. The variance is highest when P=0.5.

These two approaches will only take us to the same place if the probability of receiving treatment is the same in each strata AND/OR the stratum-specific effect is the same for each strata (i.e. no heterogeneous treatment effects).

For a classic - and insightful- example of how a failure to meet one or both of the aforementioned conditions can generate biased OLS estimates, we take a quick detour into the labor literature.

Back in 1998, Josh Angrist was interested in estimating the effect of voluntary military service on civilian earnings. This was a policy relevant question. Between 1989 and 1992, the size of the military declined sharply because of increasing enlistment standards. Policymakers wanted to know whether people who would have served under the old rules but were unable to enlist under the new rules were hurt by the lost opportunity for service.

The potential outcomes are $Y_i(0)$ (earnings conditional on not serving) and $Y_i(1)$ (earnings conditional on serving). The estimand of interest is the average impact of service on earnings among those individuals who voluntarily served in the military:= $E[Y(1) - Y(0)|D = 1]$. The covariates include year of birth, military test score group, year of application to the military, and educational attainment at the time of application.
The following table (adapted from MHE, p. 73) shows that a simple comparison of average earnings by veteran status (conditioning on application to serve) suggest that those who were accepted to serve ended up earning more than their non-veteran counterparts. This effect appears larger for non-whites.

	Average earnings 1988-1991	Difference in means by veteran status
Whites	14,537	1,233
		(60)
Non-white	11,664	2,449
		(47)

Can we stop here? Why might we be worried about bias?

These treatment effect estimates assume that treatment status D_i is independent of the potential outcomes $Y_i(0)$ within these sub-populations. The military only accepts those with high school diplomas and high test scores. Given selection into voluntary mlitary service was non-random, one can imagine that unconditional independence will not hold. It seems more plausible, therefore, to

invoke an independance assumption that conditions on observables such as year of birth, education, test scores.:

$$E[Y_i(0)|X_i, D_i = 1] = E[Y_i(0)|X_i, D_i = 0].$$

Given conditional unconfoundedness, the conditional effect of treatment on the treated can be constructed as:

$$\begin{aligned}
\tau_x &= E[Y_i(1) - Y_i(0)|D_i = 1] \\
&= E\{E[Y_i(1)|X_i, D_i = 1] - E[Y_i(0)|X_i, D_i = 1]|D_i = 1\} \\
&= E\{E[Y_i(1)|X_i, D_i = 1] - E[Y_i(0)|X_i, D_i = 0]|D_i = 1\} \\
&= E[Y_i(1)|X_i, D_i = 1] - E[Y_i(0)|X_i, D_i = 0]
\end{aligned}$$

Here τ_x is a random variable that represents the differences in mean earnings by veteran status corresponding w covariates X.

Matching estimates are constructed from covariate-value-specific differences in earnings. The covariates in this case were the age, schooling, and test-score variables used to select soldiers from the pool of applicants. The overall average treatment effect is a weighted average of the conditional treatment effects:

$$\widehat{\tau} = \sum_x \widehat{\tau}_x P(X_i = X|D_i = 1)$$

Where $P(X_i = X|D_i = 1)$ is the probability mass function for X_i given $D_i = 1$ and the summation is over the values of X_i. Intuitively, the weights are proportional to the probability of veteran status conditional on the value of the covariates. In other words, the men that are most likely to serve get the most weight in estimates of the effect of treatment on the treated.

	Difference in means by veteran status	Matching estimate
Whites	1233	-197
	(60)	(71)
Non-white	2449	840
	(47)	(63)

Comparing these two columns, we see that simple (unweighted) comparisons seem to overstate the effect of military service on earnings. Why? The military selects only the most competent individuals. So the earnings among non-veterans are going to under-estimate what those who

served would have earned had they not served. Once we condition on measures of competence (i.e. test scores), we find the estimated ATT to be much lower than a naive comparison of means. Among whites, this point estimate is actually negative!

Now consider the following linear regression:

$$Y_i = \sum_x S_{ix}\beta_x + \tau D_i + \varepsilon_i.$$

This is a highly saturated model, meaning that there is an S_x indicator variable for every possible combination (or "stratum") of covariates. It is not fully saturated, however, because there are no $D_i \cdot X_i$ interactions.

The following table reports the results:

	Difference in means by veteran status	Matching estimate	Regression estimate
Whites	1233	-197	-89
	(60)	(71)	(29)
Non-white	2449	840	1074
	(47)	(63)	(50)

Despite the fact that this parametric regression controls for precisely the same covariates, regression estimates are larger than the matching estimates for non-white and less negative for whites. The reason for the difference is that the regression estimates weight observations differently from the matching estimator.

Intuitively, OLS minimizes the variance of the parameter of interest. As a result, it will weight more heavily the stratum-specific effects with the lowest expected variance. Recall that the variance is minimized when the probability of treatment = 0.5. So OLS will weight more heavily those strata where P is close to 0.5. Recall that, if treatment effects are not heterogeneous, this weighting is not an issue. And if the propensity score does not vary across strata, this weighting is a non-issue.

In this application, neither of these conditions are met. The share of men who serve varies significantly across strata. And the men who were most likely to be accepted to serve in the military (i.e. educated guys who score well on tests) appear to benefit least (in terms of income) from their service. Matching, which uses the distribution of covariates among those who served to weight covariate-specific effects, weights these cells relatively more heavily. In contrast, the regression weights relatively more heavily those cells iin which the treatment fraction is close to 0.5. In this case, matching yields an estimate of the effect of military service that is smaller than

regression estimates which condition on the very same controls. But this bias could easily have worked the other way.

1.3.2 Advantages of matching:

- Matching does not impose any functional form restrictions on the outcome equations. This eliminates the potential for bias due to misspecification of functional form. More precisely, if conditional independence holds, but linearity does not, matching estimators are consistent, but linear regression estimators are not. Smith and Todd (2005) directly compare the results of matching and regression estimates and show that avoiding functional form restrictions can reduce bias substantively. On the other hand, if the functional form is known, imposing it increases the efficiency of the estimator.
- For non-technical audiences, matching may have expositional advantages. It offers an intuitive, conceptually straightforward approach to conditioning on covariates (as compared to regression adjustments).
- Matching algorithms force the analyst to examine the alternative distribution of covariates across treatment groups. This helps you recognize which observations have no partner in the opposite treatment group, and makes you realize when you are relying upon parametric assumptions as you project into areas of the data where you might have no business being in(because you have not data).
- Although matching and regression-based estimates can be implemented as weighted regressions, the weights differ. Matching weights these covariate-specific effects according to the proportion of treated at each point in X. This ensures that covariates are uncorrelated with the treatment indicator. In contrast, a simple regression estimator weights marginal effects proportional to the variance of the treatment status at each point in X.
- Whereas non-parametric matching leaves individual causal effects unrestricted and accomodates heterogeneity in treatment effects, regression imposes structural assumptions on the relationship between covariates and treatment effects.

Note that matching and OLS estimates will be similar if:

- there is common support / extensive overlap in the distributions of X.
- there is little heterogeneity in treatment effects with respect to X
- the outcome regression equations are not seriously misspecified.

Finally, it is also worth noting that matching can be viewed as a compliment, versus a substitute, for parametric models. That is, matching can be used to "preprocess" the data, and then a parametric model can be estimated.

2 Apples to apples?

So you have collected your data, you have an identification strategy in mind, you are ready to get started. You have yet to decide how you are going to adjust for differences in covariate distributions: linear regression, semi-parametric matching, etc. A good first step is to get a sense of how off-balance you are to begin with.

2.1 Step1: Assess overlap

A first step is to check the overlap of the covariate distributions of the treatment and control groups. It will often be the case that overlap is not perfect. Political scientists have introduced the notion of "FSATT". The feasible sample treatment effect on the treated is the causal effect among those treated individuals who find a close match in the control group.

2.2 Step 2: Assess balance in covariates

Imbens and Rubin (2010) suggest some informative summary statistics that can be useful for assessing the differences between covariate distributions in the treatment (T) and control (C) groups, respectively. One measure that I find particularly useful measures what fraction of the treated (control) units have covariate values that are near the center of the distribution of the covariate values for the control (treated) group. In a randomized experiment, in expectation, 95% of observations should lie within the 0.975 and 0.025 quantiles of the covariate distribution of the opposite group. We can assess how close our observational data come to this ideal.

To implement this approach, we define the following empirical distribution functions:

$$\widehat{F}_C(x) = \frac{1}{(1-P)N} \sum_{i:D_i=0} 1(x_i \leq x) \tag{22}$$

$$\widehat{F}_T(x) = \frac{1}{PN} \sum_{i:D_i=1} 1(x_i \leq x) \tag{23}$$

We can then define the inverse:

$$\widehat{F}_C^{-1}(z) = \min\{X_i : \widehat{F}_C(x) \geq z\} \quad (24)$$
$$\widehat{F}_T^{-1}(z) = \min\{X_i : \widehat{F}_T(x) \geq z\} \quad (25)$$

Let $\pi_T^{0.95}$ identify the proportion of treated units with covariate values within 0.025 and 0.975 quantiles of the empirical distribution of the covariate values in the control group:

$$\pi_T^{0.95} = \widehat{F}_T(\widehat{F}_C^{-1}(0.975)) - \widehat{F}_T(\widehat{F}_C^{-1}(0.025)). \quad (26)$$

We can define an analagous $\pi_C^{0.95}$ proportion for the control group.

3 Non-parametric covariate matching

This discussion of matching techniques will focus on the approaches that are most commonly used in the literature. Our discussion will touch on some of the key issues that arise when implementing these estimators. We will not delve too deeply into any of the implementation issues we consider. My goal is to flag the key design choices you will encounter, and talk briefly about possible implications.

3.1 Non-parametric matching estimators

If we want to eschew parametric functional forms entirely (a noble goal!), then we will need to match directly on the covariates. The simplest method of matching compares observations with exactly the same values of the observed variables X. For obvious reasons, this method is typically impossible to implement in practice.

Here, we will consider two approaches that involve inexact matching on covariates. Inexact matching procedures reduce the dimension of the problem by defining a distance metric on X and then matching using the distance rather than the X. Asymptotically, all inexact matching schemes are in some sense equivalent since they all tend towards exact matches as the sample gets larger and everyone finds the perfect match. However, these can yield very different answers in finite samples that tend to fall short of the perfect match ideal.

We are going to claim that these methods allow us to remain agnostic with respect to the functional form of the underlying potential outcome equations. Let us be clear about what we mean. We can write the two potential outcome equations as:

$$Y_i(1) = f_T(X_i) + \varepsilon_{iT} \quad (27)$$

$$Y_i(0) = f_C(X_i) + \varepsilon_{iC} \quad (28)$$

where ε_{i0} and ε_{i1} are assumed to be *i.i.d.* with zero conditional means (conditioning on X_i).

A non-parametric matching estimator does impose *some* structure on these relationships. The basic idea behind covariate matching is that:

$$X_i = X_j \Rightarrow f_T(X_i) = f_T(X_j),\ f_C(X_i) = f_C(K_j) \quad (29)$$

and:

$$d(X_i, X_j) < \varepsilon \Rightarrow d'(f_T(X_i), f_T(K_j)) < \delta, d'(f_C(X_i), f_C(K_j)) < \delta, \quad (30)$$

where d and d'represent the metric you are using to measure the distance between observations.

So, to be more precise, although we can avoid having to specify a particular functional form for the outcome equations, we do need to assume these are a continuous functions over the relevant range of X_i. Why? The continuity of the outcome equations justify neighborhood matching when exact matching is impossible.

"Nearest neighbor" matching (NN) The NN matching estimator offers the simplest approach to non-parametric covariate matching. It is commonly used due to its ease of implementation and conceptual clarity. The idea is to construct the counterfactual for each treated unit using the m "nearest" controls.

Notation introduced by Imbens (Abadie et al, 2001), helps make this discussion more precise. Given a sample $\{(Y_i, D_i, X_i)\}_{i=1}^N$, let $d_M(i)$ be the distance from the covariates for unit i, X_i, to the mth nearest match with the opposite treatment. Allowing for the possibility of ties, this is the distance such that strictly fewer than m units are closer to unit i than $d_m(i)$. More formally, $d_m(i) > 0$ is the distance satisfying the following inequality:

$$\sum_{j|D_j \neq D_i} 1\{||X_j - X_i|| \leq d_m(i)\} < m, \quad (31)$$

$$D_j \neq D_i, \quad (32)$$

where $||X_j - X_i||$ measures the "distance" between X_j and X_i. We will discuss the choice of distance metric below.

Let $\vartheta_m(i)$ denote the set of indices for the matches for unit i that are as least as close as $d_m(i)$. If there are no ties, the number of elements in $\vartheta_m(i)$ is m. In constructing a counterfactual estimate for observation i, the m nearest neighbors are weighted $\omega(i,j) = 1/m$. For all other control cases $\omega(i,j) = 0$.

Consider the example of setting $m = 1$. Observation i in the treated sample is matched with observation j in the control sample if $X_i = X_j = x$:

$$\begin{aligned}
\widehat{\tau}_i &= Y_i(1) - Y_i(0) && (33)\\
&= f_T(X_i) + \varepsilon_{iT} - (f_C(X_j) + \varepsilon_{jC}) && (34)\\
&= f_T(X_i) + \varepsilon_{iT} - (f_C(X_i) + \varepsilon_{iC}) + \{\varepsilon_{iC} - \varepsilon_{jC}\} && (35)\\
&= \tau + \{\varepsilon_{iC} - \varepsilon_{jC}\} && (36)
\end{aligned}$$

The treatment effect on the treated can thus be estimated by:

$$\begin{aligned}
\widehat{\tau}_{TT} &= \frac{1}{PN}\sum \widehat{\tau}_i && (37)\\
&= \tau_{TT} + \frac{1}{PN}\sum_i \varepsilon_{iC} + \frac{1}{PN}\sum_j \varepsilon_{jC} && (38)
\end{aligned}$$

This is an unbiased and consistent estimate of the PATT.

What about the more likely scenario in which matches are inexact and we use more than one "neighbor":

$$\begin{aligned}
\widehat{\tau}_i &= Y_i(1) - Y_i(0) && (39)\\
&= f_T(X_i) + \varepsilon_{iT} - \frac{1}{M}_{j\in\vartheta_m(i)}(f_C(X_j) + \varepsilon_{jC}) && (40)\\
&= f_T(X_i) + \varepsilon_{iT} - (f_C(X_i) + \varepsilon_{iC}) + \frac{1}{M}_{j\in\vartheta_m(i)}\{\varepsilon_{iC} - \varepsilon_{jC}\} + \frac{1}{M}_{j\in\vartheta_m(i)} f_C(X_i - X_j) && (41)\\
&= \tau + \varepsilon_{iC} - \frac{1}{M}_{j\in\vartheta_m(i)}\varepsilon_{jC} + \frac{1}{M}_{j\in\vartheta_m(i)} f_C(X_i - X_j) && (42)
\end{aligned}$$

Using these to build our estimate of PATT:

$$\widehat{\tau}_{TT} = \frac{1}{PN}\sum \widehat{\tau}_i \tag{43}$$

$$= \tau_{TT} + \frac{1}{PN}\sum_i \varepsilon_{iC} + \frac{1}{PN}\sum_i \frac{1}{M}_{j\in\vartheta_m(i)} \varepsilon_{jC} + \frac{1}{PN}\sum_i \frac{1}{M}_{j\in\vartheta_m(i)} f_C(X_i - X_j) \tag{44}$$

In general, we would expect $\frac{1}{M}_{j\in\vartheta_m(i)} f_C(X_i - X_j)$ to be increasing in M as the average match quality gets poorer. Clearly, if m approaches $(1-P)N$, $\frac{1}{M}_{j\in\vartheta_m(i)}(f_C(X_j) + \varepsilon_{jC})$ is systematically biased toward the mean covariate value in the control group, regardless of evidence of local deviation from the overall mean in the neighborhood of X_i. At the other extreme, if $m = 1$, under reasonable conditions, bias approaches zero as the number of controls approaches infinity (and are crowded densely into the covariate space). However, when you do not have an infinite pool of controls, setting $m = 1$ could make your estimate very sensitive to idiosyncracies of a particular neighbor; estimates will be more variable.

3.1.1 Implementation details

How should I measure "distance" Distance metrics have the general form:

$$d_{jk} = (X_j - X_k)'\Sigma_x(X_j - X_k) \tag{45}$$

where Σ_x is a weighting matrix.

The simplest distance measure is the standard Euclidean metric:

$$d^E_{jk} = (X_j - X_k)'(X_j - X_k). \tag{46}$$

This metric will weight each component equally when computing the scalar measure of difference. This may not be sensible. For example, if your covariates are measured in different scales. Instead, it is far more standard to use the Mahalanobis distance which uses the inverse of the covariance matrix of X to weight the differences as the weighting matrix:

$$d^M_{jk} = (X_j - X_k)'\Sigma_X^{-1}(X_j - X_k). \tag{47}$$

For more formal discussions of distance metrics in matching, see Rubin and Thomas (1992) and Zhao (2004).

How many neighbors? NN matching estimators have the attractive feature that, once you have chosen your distance metric, you need only choose how many neighbors to match on (and whether to match with replacement). presents some interesting trade offs when you actually get down to implementing it.

The first issue: Should you match with replacement?

This choice boils down to a trade off between bias and variance. If we match without replacement, our match quality may suffer if the number of comparison ($D = 0$) observations comparable to the treated observations is small. Matching with replacement will improve the average quality of matches (i.e. reduce the distance between matched observations). This will reduce the bias. But it will also reduce the number of observations that are used as controls, and this can increase the variance.

Also note that matching without replacement is order dependent, though there are variants, called "optimal matching" in the applied statistics literature, that try to find the best (by some distance criterion) set of matches without replacement. See Hansen (2004) Journal of the American Statistical Association.

In sum, the decision to match with replacement or not depends in part on the data. Is there a unique close match for each treated observation?

Second issue: How many near neighbors?

The choice of how many neighbors also boils down to a trade off between bias and variance. Matching only on the closest neighbor maximizes match quality. This minimizes the bias that can be an issue if match quality is poor.

Matching on additional neighbors reduces the average match quality. However, the variance will likely decrease because we are averaging across neighbors to construct a counterfactual for each unit, thus reducing noise. The choice generally depends on the data. If you have lots of close matches, more neighbors is probably better.

Sensitivity analysis is always good.

In the literature, researchers sometimes inform their choices using "leave-one-out cross-validation". As the name implies, this "leave-one-out" approach drops the jth control observation in the comparison group and uses the remaining control observations in the comparison group to construct an estimate of $Y_j(0)$. Let the forecast outcome for unit j be:

$$\widehat{E}_{-j}[Y_j(0)|X_j, m], \tag{48}$$

where m denotes the number of neighbors used in the matching. We compute the forecast error for this j. We repeat the process for the remaining $(1-P)N-1$ control observations. After we have finished leaving one out, we compute the mean squared error or root mean squared error associated with the different matching estimators:

$$MSE(m) = \sum_{j \in D_i=0} \frac{1}{(1-P)N} \left([Y_j(0) - \widehat{E}_{-j}[Y_j(0)|X_j, m] \right)^2 \tag{49}$$

.A comparison of the mean squared error or root mean squared error guides the choice of how many neighbors to use. We choose the m associated with the smallest MSE.

Kernel matching Kernel matching, a natural extension of NN matching, is a non-parametric matching estimator that constructs a match for each of the N_T program participants using a weighted average over all N_C members of the comparison group (Heckman, Ichimura and Todd, 1997; Heckman, Ichimura, and Todd, 1998; Heckman, Ichimura, Smith, and Todd, 1998). The contributions of relatively more distant observations are weighted relatively less.

Kernel matching can thus be seen as a weighted regression of counterfactual outcomes on an intercept with weights determined by kernel weights. The weights are given by

$$w_{ij} = \frac{G(\frac{d_{ij}}{a_{N_C}})}{\sum_k G(\frac{d_{ij}}{a_{N_C}})}.$$

Here, $G(.)$ is a kernel function (i.e. a symmetric function that integrates to one) that weights observations based on the distance between X_i and X_j. Commonly used kernels include the normal (where G is the normal pdf) or the "triangle" kernel (where G is the Epanechinikov kernel).
d_{ij} is the distance between treated facility i and untreated facility j;

a_{N_C} is a bandwidth parameter that scales the distances between the treated and untreated based on the comparison group size.

The numerator is therefore a transformed distance between eacth control case and the target treatment case. The denominator insures that the weights sum to one.
One advantage of kernel matching over NN matching is that the variance of the estimate tends to be lower because information on all control cases is used to construct the counterfactual estimate for each treated individual. A second advantage has to do with variance estimation. Whereas Abadie and Imbens (2006) demonstrate that bootstrapping methods are invalid for NN matching, bootstrapping standard errors is valid for drawing inference in kernel matching (HIT, 1998).

Bandwidth selection When implementing a kernel matching estimator, you need to choose the kernel function and the bandwidth. The choice of kernel function appears to be relatively unimportant in practice.
Estimates can be sensitive to the choice of bandwidth. The choice of bandwidth is analagous to our choice of neighbors, as it involves trading off variance and bias. A narrow bandwidth increases the weight on close matches, thus reducing bias. A wider bandwidth gives more weight to more units, thus reducing variance, but increasing bias as more dissimilar units receive more weight in the counterfactual construction.

There is an entire literature on optimal bandwidth selection(!). In practice, it seems the approach to choosing the bandwidth parameter is very similar to the approach to choosing the optimal number of neighbors. Sensitivity analysis and leave-one-out cross validation.

4 Semi-parametric covariate matching

In covariate matching, the goal is to set the bias elements $B1 = B2 = 0$. The problem is, as the number of covariates to match gets large, covariate matching is unlikely to be exact.

Abadie and Imbens (2002) show that, for the NN estimator, the bias term corresponding to the matching discrepancies will be of the order $N^{-1/k}$. Asymptotically, this bias can dominate the large sample variance if you have three or more continuous covariates.

There are a couple of ways around this problem. First, if you are blessed with a deep pool of potential controls and you are only interested in estimating the ATT, you can appeal to an asymptotic sequence where the numper of potential controls increases faster than the number of treated units. Asymptotically, the bias disappears as the match quality approaches perfection.

Another approach involves making regression-based adjustments for poor match quality. A number of corrections have been proposed. We will consider two such corrections.

Both adjustments involve estimating μ_C using a parametric regression equation:

$$Y_i(0) = \alpha_C + \beta_C' X_i + \varepsilon_i \tag{50}$$

to capture how the outcome under the control state, $f_0(.)$.varies with observables. This assumes we are interested in estimating the ATT. This function is estimated by least squares using data on the untreated observations.

Abadie and Imbens (2002) suggest a very similar approach. The difference is that they use only those controls that are used as matches for the treated units with weights that correspond to the number of times the control unit is used as a match. Compared to the Rubin method, this approach can be less efficient if you are discarding lots of control observations. However, the controls you discard are more likely to be outliers. By using the AI approach, the weighted empirical distribution is closer to the distribution of covariates that we are ultimately interested in.

The bias-corrected estimate for unit i:

$$\widehat{\tau_i} = Y_i - (\frac{1}{M} \underset{j \in \vartheta_m(i)}{} (Y_j + \beta_1 X_i - \beta_1 X_j))$$

One question. We were so excited about getting away from imposing functional form assumptions on the outcome equation. Why re-introduce parametric assumptions in this "back room" bias adjustment?

Keep in mind that the model-based adjustments are performed post-match on matched observations only. This means there is much less extrapolation going on (as compared to the regression-based estimators). Regression-based methods are very well suited for approximating smooth functions over short intervals. This is what we are doing with this bias adjustment. Regression-based methods can potentially get you intro trouble if you extrapolate over larger ranges of sparsely populated (with observations) covariate space.

Abadie and Imbens (2002, 2004) evaluate the performance of this semi-parametric approach. They demonstrate how this bias-adjusted matching estimator can outperform conventional alternatives (namely non-parametric and fully parametric approaches).

5 Propensity-score matching

When you have many covariates in X, it can get increasingly difficult to non-parametrically adjust for multi-dimensional differences in the distribution of the covariates. One approach, described above, involves mapping multiple covariates into a distance metric. Alternatively, you can focus on adjusting for differences in the propensity score (i.e. the probability of receiving the treatment conditional on observables).

"Propensity score matching" refers to a class of multivariate matching methods that can be used to match treated and control units that have similar distributions for many covariates. Matching

on the scalar propensity score can be very effective in reducing the differences in sample means of many variables across treatment and control sub-groups (and thus mitigating bias components $B1$ and $B2$).

Rosenbaum and Rubin (1983) lay the foundation for most of the work that has been done in this area. They introduce the concept of a "balancing score" $b(X)$, a function of observed covariates X such that the conditional distribution of x given $b(x)$ is the same for the treated and comparison groups:

$$X \perp D \mid b(X). \tag{51}$$

Conditioning on covariates and conditioning on a balancing score will result in identical distribution of X across treated and control groups.

The theory underlying propensity score matching is quite different from covariate matching. The propensity score is the conditional probability that a unit with $X = x$ will receive the treatment:

$$p(x) = \Pr(D = 1|X = x). \tag{52}$$

Their first important result is that, the propensity score *is a balancing score.*

$$\begin{aligned} \Pr(X, D|p) &= \Pr(X|p) \Pr(D|p) && (53) \\ &\Longrightarrow X \perp D|p && (54) \end{aligned}$$

Why is this useful? If treatment assignment is strongly ignorable given X (i.e., if our assumption of conditional unconfoundedness holds), then treatment assignment is strongly ignorable conditional on a balancing score. Put differently, if Y is independent of D given X, then it should also be independent of D conditional on the scalar $P(X)$ which summarizes the information in X that is relevant to D.

The second important claim is that if exact matching is impossible, then approximate matching on the propensity score will yield approximately the same distribution of X in the treated and control sample:

$$d(p_i, p_j) < \varepsilon \Rightarrow d'(\Pr(X_i|p_i), \Pr(X_j|p_j)) < \delta \tag{55}$$

The intuition is that two groups with the same probability of participation will show up in the treated and control samples in equal proportion. Thus, they can be combined for purposes of comparison.

To see the detailed exposition of these key results, please see Rosenbaum and Rubin (1983). This seminal paper generated lots of excitement about matching methods that avoid adjusting directly for all covariates, and instead adjust for differences in the propensity score. Although propensity score matching does seem to offer a solution to the curse of dimensionality, there are strings attached.

Note that propensity score matching is "semi-parametric" because it combines a parametric model of selection into the treatment group with a non-parametric comparison of outcomes (typically). Of course, it is not possible to estimate a fully non-parametric selection model, because if the curse of dimensionality befalls the outcome equation, it will also befall the selection equation. Estimation of the propensity score is typically accomplished parametrically (using a probit or logit model).

What does this mean? Recall that we abandoned regression-based methods in favor of covariate matching because we wanted to eschew parametric assumptions. But propensity score methods have us back to making parametric assumptions, although these assumptions apply to a different aspect of the data generating process.

Let's think about what this means....

First, if the researcher knows the propensity score (as in a randomized experiment where the researcher is in control of the assignment mechansim), then propensity score methods can be very effective in removing bias. Moreover, in the case of many covariates, propensity score estimation allows you to avoid reliance on many dimensional nonparametric regression. And this implies superior performance in finite samples.

However, it is more often the case that the researcher does not know the true selection process. By choosing a probit or logit functional form, you restrict the predicted probabilities to lie between 0 and 1 (a very reasonable restriction!) but you also impose other restrictions on the form of this selection equation. We cannot assume a priori that the standard logit or probit is the correct specification for the underlying selection process. Put differently, although propensity score matching techniques can allow you to remain agnostic about the form of the relationship between the covariates and the outcomes, a parametric model with many covariates still lurks in the background. We are trading one set of parametric assumptions for another when we choose propensity score matching over a parametric regression approach. You have just moved the parametric assumptions from the outcome equation to the selection equation.

When might propensity score methods likely to be the best choice? If your intuition about the correct form of the selection equation is better than your intuition about the outcome equation,

propensity score methods may be preferred. For example, if you have more information about the selection process (in particular, is it more smooth than the outcome equation?).

Another thing to keep in mind: a precise estimate of the "true" propensity score is not required for unbiased estimation. What we really care about is the ability of this single scalar measure to act as a balancing function. So long as we have achieved that, we really do not care whether the functional form you choose looks anything like the true relationship between pre-treatment variables and the treatment assignment. But if we do not achieve this balance, then we are back to relying upon parametric assumptions about the underlying data generating process to establish a basis for causal inference.

One last observation. In models with only discrete covariates, matching on the estimated propensity score and non-parametric covariate matching will be equivalent (because the propensity score is nothing other than the within stratum probability of receiving the treatment).

5.1 Now what?

Once you have your propensity scores, what do you do with them? You have several options:

5.1.1 Matching!

Rosenbaum and Rubin's result implies that it is sufficient to adjust solely for differences in the propensity score between treated and control units. So the NN matching or kernel matching methods described above can be readily implemented (matching on the estimated scores).

5.1.2 Inverse probability weighting

Hirano, Imbens, and Ridder (2003) advocate weighting by the inverse of the probability score in an effort to balance covariates across treated and control units.

Why?

Intuitively, this is akin to "undoing" a stratified sampling design, except that the inverse probabilities are estimated. Michael Anderson provides a more detailed derivation in the ARE 213 notes. But the basic idea is as follows.

To estimate the ATT, we would construct our IPW estimate as follows:

$$\tau_{TT} = \frac{1}{N} \sum_i \widehat{p}(X_i) \left(\frac{Y_i D_i}{\widehat{p}(X_i)} - \frac{Y_i(1 - D_i)}{1 - \widehat{p}(X_i)} \right) \tag{56}$$

The first term is just the average of the outcomes among the treated because the probability terms cancel.

The second term weights the outcomes of the control units by the probability of not being treated. This weighting scheme ensures that an observation with covariate value x is equally represented (in expectation) in the treated and control groups when we make our comparison.

An advantage of this estimator is that avoids the bandwidth selection problem. A disadvantage is that this assumes that $\widehat{p}(X_i)$ is a true balancing score. This is not always a safe assumption!

5.1.3 Blocking

In their original propensity score paper, Rosenbaum and Rubin suggest using the propensity score to divide the sample into M blocks based on $p(x)$. Blocking is akin to a crude form of non-parametric regression. The $[0, 1]$ interval is partitioned into b intervals. Within each block, the average treatment effect is estimated as if random assignment had occured within the block:

$$\hat{\tau}_b = \frac{1}{N_{Tb}} \sum_{i=1}^{N_{Tb}} D_i Y_i - \frac{1}{N_{Cb}} \sum_{i=1}^{N_{Cb}} (1 - D_i) Y_i. \tag{57}$$

The overall treatment effect is calculated as a weighted sum. If we are interested in average effect on the treated, the weights correspond to the number of treated units within each block.

5.1.4 Blocking with regression:

Subsequent researchers have augmented this blocking procedure by conditioning on covariates within the blocks. That is, you estimate the propensity score, block on the propensity score. and then within each block, estimate a linear regression with the full set of covariates (including the treatment indicator). This leaves you with b estimates of τ_b. These are then averaged in the same way as above.

5.1.5 The "doubly robust" estimator

Recall the regression-based parametric adjustment we started out with:

$$Y_i = \alpha_0 + \tau D_i + \beta' X_i + \gamma' X_i \cdot D_i + \varepsilon_i \tag{58}$$

In order to interpret τ as an unbiased estimate of the average treatment effect, it must be the case that the postulated linear outcome equation is properly specified OR covariates are perfectly balanced across the treatment and control groups OR we have unconditional unconfoundedness. As the latter two conditions are rarely satisfied, we end up relying on our parametric assumptions when imputing counterfactual outcomes.

The estimator based on inverse propensity score weighting requires that the specification of the selection equation be correct, or at least sufficient to yield a true balancing score.

If we combine both approaches, we have two chances to get lucky. That is, if we estimate our regression equation, but weight observations using the IPW, we have a doubly robust estimator. If the postulated propensity score model yields a true balancing score, but the regression model is misspecified, the misspecification of the regression model is of no consequence. If the covariates are identically distributed across the treatment and control groups (thanks to the inverse probability weighting), we do not need to make regression-based adjustments. Converseley, if the postulated regression model is correct, but the postulated propensity score model is not correct, we are also alright. This is because the regression corrections will compensate for imbalances in the covariate distribution.

In sum, this combination of methods offers protection against misspecification of one form or another..but not both!

Abadie, A., and G. Imbens, (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects,"Econometrica, 74(1), 235-267.

Abadie, A., and G. Imbens, (2007), "On the Failure of the Bootstrap for Matching Estimators,"unpublished manuscript, department of Economics, UC Berkeley.

Abadie, A., D. Drukker, H. Herr, and G. Imbens, (2003), "Implementing Matching Estimators for Average Treatment Effects in STATA,"The Stata Journal, 4(3), 290-311.

Caliendo, M., (2006), Microeconometric Evaluation of Labour Market Policies, Springer Verlag,Berlin.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd (1996): "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evdience on the Effectiveness of Matching as a Program Evaluation Method," Proceedings of the National Academy of Sciences, 93(23), 13416-13420.

Heckman, James, Hidehiko Ichimura and Petra Todd (1997): "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," Review of Economic Studies, 64(4), 605-654.

Heckman, James, Hidehiko Ichimura and Petra Todd (1998), "Matching As An Econometric Evaluation Estimator," Review of Economic Studies, 65(2), 261-294.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd (1998): "Characterizing Selection Bias Using Experimental Data," Econometrica , 66(5), 1017-1098.

Imbens, G., (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,"Review of Economics and Statistics, 86(1): 1-29.

Rosenbaum, P., Rubin, D., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Rosenbaum, P., Rubin, D., 1984. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 79, 516–524.

Rosenbaum, P., Rubin, D., 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity. American Statistician 39, 33–38.

Rubin, D., 1973. Matching to remove bias in observational studies. Biometrics 29, 159–183.

Rubin, D., 1979. Using multivariate matched sampling and regression adjustment to control bias in observation studies. Journal of the American Statistical Association 74, 318–328.