PS C236A/ Stat C239A
Univariate Matching and Propensity Score

Erin Hartman

September 30, 2009

1 ATE, ATT, ATC

Based on Rocio Titiunik's notes: ATT, ATE and Potential Outcomes
`http://www-personal.umich.edu/~titiunik/fall2007_ps236/section_notes_4_ateatt.pdf`

We know, from the potential outcomes framework, that the unit level treatment effect is defined as:

$$\tau_i = Y_{i1} - Y_{i0}$$

However, this is unobservable, so we often have to rely on other moments to estimate the effect of our treatment. The difference $Y_{i1} - Y_{i0}$ is a random variable, so we must decide what aspect of its distribution we are interested in estimating. One common moment we are interested in is the average treatment effect (ATE), τ. We may also be interested in the average treatment effect on the treated (ATT). We define these quantities as follows:

$$ATE = \mathbb{E}[Y_{i1} - Y_{i0}]$$

$$ATT = \mathbb{E}[Y_{i1} - Y_{i0}|T_i = 1]$$

However, in general, we will have to make a selection on observable assumption, therefore we will be interested in the conditional versions of ATE and ATT:

$$ATE = \mathbb{E}[Y_{i1} - Y_{i0}|X]$$

$$ATT = \mathbb{E}[Y_{i1} - Y_{i0}|T_i = 1, X]$$

1.1 Random Assignment

When we have random assignment, then we know that $T \perp (Y_{i1}, Y_{i0})$, then we know that

$$ATT = \mathbb{E}[Y_{i1} - Y_{i1}|T = 1] = \mathbb{E}[Y_{i1} - Y_{i1}] = ATE$$

because conditioning on T is irrelevant due to independence. It can be shown that

$$ATE = ATT = \mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]$$

which can be estimated by

$$\widehat{\mathbb{E}[Y|T=1] - \mathbb{E}}[Y|T=0] = \frac{1}{N_1}\sum_{i:T_i=1}^{N_1} Y_i - \frac{1}{N_0}\sum_{i:T_i=0}^{N_0} Y_i$$

Note that this is just a simple difference in means between outcomes of the treatment group and outcomes of the control group. A randomized treatment assignment guarantees that the difference-in-means estimator is unbiased, consistent and asymptotically normal.

Note: This result also holds if we replace the assumption of independence by the assumption of mean independence, ie. $\mathbb{E}[Y_{i1}|T] = \mathbb{E}[Y_{i1}]$ and $\mathbb{E}[Y_{i0}|T] = \mathbb{E}[Y_{i0}]$

1.2 Non-random Assignment

Typically we don't have random assignment. Rather, there is usually self-selection of treatment, and often this decision is related to the benefits of treatment, i.e. to $Y_{i1} - Y_{i0}$. Since this will make treatment correlated with the potential outcomes, we need an additional assumption in order to identify the treatment effects we're interested in. The typical approach is to make an assumption based on X, a vector of pretreatment covariates. Rosenbaum and Rubin (1983) introduced the assumption of unconfoundedness (note: let $\mathbb{X}$ denote the support of X.

Assumption: *Unconfoundedness* For almost every $x \in \mathbb{X}$, T is independent of $(Y_{i1} - Y_{i0})$ conditional on $X = x$

The intuition behind the unconfoundedness assumption is as such. While $(Y_{i1} - Y_{i0})$ and treatment assignment are correlated, given enough information (from the X covariates) that determines treatment assignment, then $(Y_{i1} - Y_{i0})$ will be independent of T, conditional on those X. Thus the name *selection on observables*. The implications are:

$$\mathbb{E}[Y_{i1}|T, X] = \mathbb{E}[Y_{i1}|X]$$

$$\mathbb{E}[Y_{i0}|T, X] = \mathbb{E}[Y_{i0}|X]$$

From this assumption, we see that ATE conditional on X (defined here as $ATE(X)$) and the ATT conditional on X (defined here as $ATT(X)$) are equal:

$$\begin{aligned} ATT(X) &= \mathbb{E}[Y_{i1} - Y_{i0}|X, T=1] \\ &= \mathbb{E}[Y_{i1}|X, T=1] - \mathbb{E}[Y_{i0}|X, T=1] \\ &= \mathbb{E}[Y_{i1}|X] - \mathbb{E}[Y_{i0}|X] \\ &= \mathbb{E}[Y_{i1} - Y_{i0}|X] \\ &= ATE(X) \end{aligned}$$

Note here that by saying conditional on $T = 1$, we mean that we are conditional on the distribution of the X in the treated population, not conditioning on treatment being received.

See the appendix for a discussion of Rubin's notation on this issue and why it provides a little more intuition for this notion.

We know that at $X = x$, the average treatment effect is:

$$ATE(x) = \mathbb{E}[Y_{i1} - Y_{i0}|X = x]$$

However, this is still a missing data problem, so to see how to estimate this, notice that:

$$\begin{aligned}\mathbb{E}[Y_i|X, T = 1] &= \mathbb{E}[Y_{i0} + T(Y_{i1} - Y_{i0})|X, T = 1] \\ &= \mathbb{E}[Y_{i1}|X, T = 1] \\ &= \mathbb{E}[Y_{i1}|X]\end{aligned}$$

Where the last line comes from the unconfoundedness assumption. It can be shown that $\mathbb{E}[Y_i|X, T = 0] = \mathbb{E}[Y_{i0}|X]$. Therefore, we see that:

$$\begin{aligned}ATE(X) &= \mathbb{E}[Y_{i1} - Y_{i0}|X] \\ &= \mathbb{E}[Y_{i1}|X] - \mathbb{E}[Y_{i0}|X] \\ &= \mathbb{E}[Y_i|X, T = 1] - \mathbb{E}[Y_i|X, T = 0]\end{aligned}$$

Meaning we simply estimate $ATE(X)$ at $X = x$ by subtracting the sample mean of the outcomes for the control units whose value of $X = x$ from the sample mean of the outcomes for the treated units who value of $X = x$. Note that $ATE(X)$ is a function of X that tells us the average effect of the treatment for every value x of X. Thus, $ATE(x)$ tells us the average treatment effect for the subpopulation with $X = x$ and $ATT(x)$ tells us the average treatment effect for the subpopulation with $X = x$ and $T = 1$ (where $T = 1$ refers to the distribution of X in the treated population). See the appendix for how this discussion relates to Rubin's response surface discussion.

It is worth noting here that for ATC, parallel arguments can be set up as for ATT, except that we condition on the distribution of the observable covariates in the controls.

2 Univariate Matching

Based on Rosenbaum 2009 chapter 3

Working in the observational framework and allowing subjects to select treatment, we will define the following quantities. Let $\mathbf{x}_i$ refer to a set of observed covariates for person i, let u_i refer to an unobserved covariate for subject i, T_i refers to treatment assignment (1 for treatment, 0 for control), and Y_{i1} refers to potential outcome under treatment, and Y_{i0} refers to potential outcome under control. Assume that there are N subjects. This situation is different from an experimental design because T_i is not assigned by the equitable flip of a fair coin.

In the population before matching, we imagine that subject i received treatment with probability π_i, independently of other subjects, where π_i may vary from one person to the next and is not known. More precisely:

$$\pi_i = Pr(T_i = 1|Y_{i1}, Y_{i0}, \mathbf{x}_i, u_i)$$

$$Pr(T_1 = t_1, \ldots, T_N = t_n | Y_{11}, Y_{10}, \mathbf{x}_1, u_1, \ldots, Y_{n1}, Y_{n0}, \mathbf{x}_n, u_n) = \prod_{i=1}^{N} \pi_i^{t_i} (1 - \pi_i)^{1-t_i}$$

Note that the second statement is trivially true for $u_i = T_i$, so there is nothing assumed in the second statement. Assumptions arise when we impose restrictions on the behavior of u_i.

2.1 The Ideal Match

Suppose that we could find two subjects, say k and l, such that exactly one was treated, $T_k + T_l = 1$, but they had the same probability of treatment, $\pi_k = \pi_l$. We can pair these two subjects and call them a match pair. Note, though, that we are imposing an assumption because we now require that $0 < \pi_i < 1$, otherwise we wouldn't be able to find matches. It is difficult to create this matched pair because we don't observe u_k or u_l, and we either observe Y_{k1} or Y_{l1} (but not both) and either Y_{k0} or Y_{l0}. [Note: We could create this matched pair by matching on $\mathbf{x}$ and randomizing treatment with the flip of a coin, however experiments are not always feasible nor ethical]. We will get back to how to estimate π_i.

Supposing that we could create a matched pair with $\pi_k = \pi_l$ and $T_k + T_l = 1$, then what would this give us?

$$\begin{aligned}
&Pr(T_k = 1, T_l = 0 | Y_{k1}, Y_{k0}, \mathbf{x}_k, u_k, Y_{l1}, Y_{l0}, \mathbf{x}_l, u_l, T_k + T_l = 1) \\
&= \frac{Pr(T_k = 1, T_l = 0 | Y_{k1}, Y_{k0}, \mathbf{x}_k, u_k, Y_{l1}, Y_{l0}, \mathbf{x}_l, u_l)}{Pr(T_k + T_l = 1 | Y_{k1}, Y_{k0}, \mathbf{x}_k, u_k, Y_{l1}, Y_{l0}, \mathbf{x}_l, u_l)} \\
&= \frac{\pi_l^{1+0}(1 - \pi_l)^{(1-1)+(1-0)}}{\pi_l^{1+0}(1 - \pi_l)^{(1-1)+(1-0)} + \pi_l^{0+1}(1 - \pi_l)^{(1-0)+(1-1)}} \\
&= \frac{\pi_l(1 - \pi_l)}{\pi_l(1 - \pi_l) + \pi_l(1 - \pi_l)} = \frac{1}{2}
\end{aligned}$$

This implies that if we can find this matched pair, then it is as though we flipped a coin for our treatment assignment within this matched pair. This also means that if we could find, among the N subjects in the population, a total of $2I$ different subjects matched in I pairs in the above way, then we could reconstruct the distribution of treatment assignments $\mathbf{T}$ in a paired randomized experiment. Inference about causal effects would be straight forward. Note, though, that randomization guarantees our selection on observable assumptions and our justifications for our estimates, however, in this world we are basing our inference on "suppositions". We could hope for a natural experiment, however these are often rare . . . , so what are our options?

2.2 The Naive approach: Those who *look* comparable are comparable

This model posits that people who look comparable in terms of a set of measured covariates, $\mathbf{x}$, are actually comparable. The consequences of this are as follows: 1) the probability of being assigned to treatment no longer depends on potential outcomes or on the unobserved covariate, and 2) it assumes that every unit has some chance of being assigned to treatment.

$$\pi_i = Pr(T_i = 1 | Y_{i1}, Y_{i0}, \mathbf{x}_i, u_i) = Pr(T_i = 1 | \mathbf{x}_i)$$

$$0 < \pi_i < 1 \qquad i = 1, 2, \ldots, N$$

with

$$Pr(T_1 = t_1, \ldots, T_N = t_n | Y_{11}, Y_{10}, \mathbf{x}_1, u_1, \ldots, Y_{n1}, Y_{n0}, \mathbf{x}_n, u_n) = \prod_{i=1}^{N} \pi_i^{t_i} (1 - \pi_i)^{1-t_i}$$

It is important to note that this is always true if treatment is assigned by the fair flip of a fair coin, or even by independent flips of a group of biased coins where the same biased coin is used when subject i and subject j have the same observable characteristics (assuming no coins have a probability of 0 or 1).

This model gives us the feature of "strongly ignorable treatment assignment given $\mathbf{x}$", or "strong ignorability".

$$Ti \perp\!\!\!\perp Y_{i1}, Y_{i0}, u_i | \mathbf{x}_i$$

2.3 Under the Naive Model: How to construct matched Pairs: Exact Matching and The Propensity Score

2.3.1 Exact Matching

If the naive model is true, then it is clear that if we can exactly match on $\mathbf{x}$, then the model will follow and we can reconstruct the distribution of treatment assignments in a randomized paired experiment simply by matching based on the observed covariates. If there is only one covariate that determines how treatment is assigned, then this is straight forward: we just matched on that covariate. With a large enough sample, it might even be straight forward to exactly match on a couple of covariates, however it becomes very difficult to exactly match on many covariates, especially with finite samples.

2.3.2 The Propensity Score

The propensity score is a way of collapsing the multidimensional problem of matching on many covariates into a unidimensional problem of matching on one covariate, the propensity score. The propensity score is defined as the conditoinal probability of treatment, $T = 1$ given the observed covariates $\mathbf{x}$.

$$e(\mathbf{x}) = Pr(T = 1 | \mathbf{x})$$

Note that the propensity score is defined only in terms of observable characteristics, T and $\mathbf{x}$, regardless of whether or not our naive model is true. In a randomized experiment we know the true propensity score, often it is $\frac{1}{2}$, because the experimenter decides the probability of treatment conditional on observed covariates. However, in an observational study the propensity score is typically unknown, however since it is based on observable characteristics, it is straight forward enough to estimate it. If the naive model *were* true, then $\pi_i = e(\mathbf{x}_i)$.

2.3.3 The Balancing Property of the Propensity Score

The balancing property is always true, regardless of if the naive model holds or not. The balancing property states that treated and control units with the same propensity score have

the same distribution of the *observed* characteristics. This gives us that treatment and observed covariates are conditionally independent given the propensity score.

$$Pr\{\mathbf{x}|T=1, e(\mathbf{x})\} = Pr\{\mathbf{x}|T=0, e(\mathbf{x})\} \Leftrightarrow T \perp\!\!\!\perp \mathbf{x}|e(\mathbf{x})$$

It is important to see that within a given matched pair, it is not necessary that subject k and subject l have the same values of $\mathbf{x}$, only that they have the same propensity score, $e(\mathbf{x}_k) = e(\mathbf{x}_l)$. Within the matched pair, the balancing property implies that the specific values of the observed covariates will be independent of the treatment assignment. Over all of the matched pairs, the distribution of $\mathbf{x}$ will look about the same in the treatment and control groups. If there are 10 covariates that contribute to the propensity score, then propensity score matching will balance all 10 covariates despite the fact that it is only a univariate matching method.

We often estimate the propensity score, coming up with an estimate $\hat{e}(\mathbf{x})$ to produce balance on the observed covariates $\mathbf{x}$. Randomization, however, is a much more powerful tool for balancing. Randomization will guarantee that strong ignobility of treatment holds, while matching on the propensity score will only guarantee that there is balance on the observed covariates, and you must from there assume that selection on observables holds in order to get balance on the unobservables.

If the naive model *were* true, then from the propensity score we could get ignorable treatment assignment. We could produce the "ideal match" from the propensity score, since it just reduces our dimensionality of $\mathbf{x}$. If the naive model holds, then $\pi_i = e(\mathbf{x})$, so matching on the propensity score is matching on π_i. In the naive model:

$$T \perp\!\!\!\perp Y_{i1}, Y_{i0}, u_i|\mathbf{x} \Rightarrow T \perp\!\!\!\perp Y_{i1}, Y_{i0}, u_i|e(\mathbf{x})$$

3 Estimating the Propensity Score

Based on Rosenbaum 2009 chapter 3

3.1 An Example: Welders and DNA

Summary: "Welders get exposed to chromium and nickel, substances that can cause inappropriate links between DNA and proteins. Costa, Zhitkovich, and Toniolo measured DNA-protein cross-links in samples of white blood cells from 21 railroad arc welders exposed to chromium and nickel and from 26 unexposed controls. All 47 subjects were male. In their data ... there are three covariates, namely age, race and current smoking behavior. The response is a measure of DNA-protein cross-links."

Before matching, we get the following descriptive statistics for the means of the two groups:

```
          control      treat
age    42.6923077 38.2380952
black   0.1923077  0.0952381
smoker  0.3461538  0.5238095
```

It would appear from these means that these covariates could be used to predict treatment. The age of the treatment group is lower, and there are more smokers and more whites. This implies that the propensity score is not constant. By saying that welders tend to be younger, we could also say that the chance of being a welder is lower for an older person. Two people who have the same propensity scores do not have to have the same values of $\mathbf{x}$, so a younger nonsmoker and an older smoker might have the same propensity score (notice that these two covariates appear to work in opposite directions in the observed data). The important fact to notice is that despite the fact that two people may differ on these observable covariates, these differences will not be helpful in predicting treatment assignment.

However, just because we balance on observable covariates $\mathbf{x}$, there is no reason to believe that the unobservables u are balanced. In this example, imagine that having a father who was a welder is highly predictive of a son being a welder, however this is unmeasured in this dataset. Successful matching on a pscore based on age, race, and smoking will guarantee balance on those covariates, but not necessarily on whether or not the father was a welder.

How do we estimate the propensity score? Often we use a linear logit model (this bounds our propensity score between 0 and 1). The propensity is then estimated by:

$$log\big(\frac{e(\mathbf{x}_i)}{1-e(\mathbf{x}_i)}\big) = \zeta_0 + \zeta_1 age_i + \zeta_2 black_i + \zeta_3 smoker_i$$

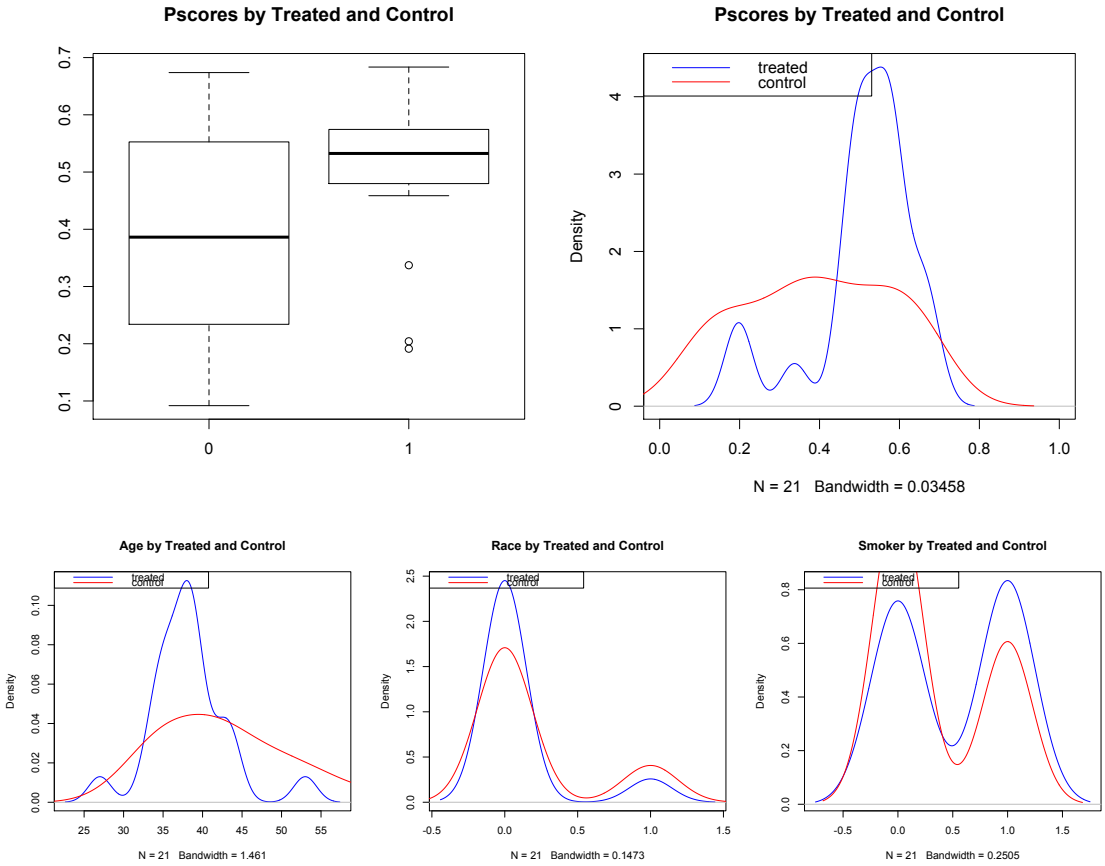
$\hat{e}(\mathbf{x}_i)$ are the fitted values from this model

So, we run the following code in R to find our propensity score:

```
pscore = glm(treat ~ age + black + smoker, family = binomial(link = logit),
 data = data)$fitted.values
```

We see that for observation number 5, a 35 year old white smoker (who received treatment), their estimated pscore $\hat{e}(\mathbf{x}_5) \approx 0.65$ which implies that there is about an estimated 65% chance of subject 5 being a welder. We see that observation 33, a 36 year old white smoker (who did not receive treatment) had about a 64% chance of receiving treatment, $\hat{e}(\mathbf{x}_{33}) \approx 0.64$, although we observed that he did not.

Overall, we see that the mean propensity score for those people who were treated is about 0.51, where it is only 0.39 for controls, so our observed covariates are imbalanced.



3.2 Distance Metrics

In order to figure out what the "closest" match is, we have to decide what our metric for the distance between observations k and l. Since we are only matching on one covariate, in this case the propensity score, we can use the squared distance between the two estimated propensity scores.

$$d = (\hat{e}(\mathbf{x}_k) - \hat{e}(\mathbf{x}_k))^2$$

This will punish large differences more than small distances. Alternatively, we could use the absolute value of the distance between the estimated propensity scores. Whatever our distance metric, "nearest-neighbor" matching matches the closest control unit to each treated unit (in the case of ATT) or the closest treated unit to each control unit (ATC). There are a few things that we have to consider when we match. For the rest of the example, we will focus on the ATT, meaning that we will find a matching control unit for every treated unit (meaning for every treated unit, we find the nearest-neighbor from the pool of control units).

- Do we match with replacement or without replacement?
- What do we do with ties?
- What do we consider a "good" match?
- How many neighbors do we match?

3.2.1 With or Without Replacement

If we match without replacement, then once we match a control unit, we take it out of the pool of potential matches for all remaining treated units. It is important to notice that if we do this, then depending on the order of the controls and the algorithm we use to sort through them, we may get different matches. If we match with replacement, then this means that after a control gets matched to a treated unit, it goes back into the pool of potential matches for the remaining treated units. This means that a control unit could be matched to multiple treated units. In general, we'd like to match with replacement to make sure that we get the "best" match every time.

3.2.2 What do we do with ties?

The case may arise that when we look for matches to a given treated unit i, there are two control units that are the same distance from i based on our distance metric d. We can either flip a coin and pick unit to make the matched pair, or we can allow ties. If we randomly pick a value, then this will decrease our variance estimate, so instead we should allow ties. Ties means that we match both control units to treated unit i, but we give each of these controls a weight of $\frac{1}{2}$ in our matched data set (in effect, we average the control units).

3.2.3 What do we consider a "good" match?

What if the closest control unit to treated unit i has a large distance, d. We may want to say that treated unit i cannot be matched because there is no control unit that is "close" to it. To do this, we would enforce a caliper, which says that if there is no "nearest neighbor" to treated unit i, defined as being within a certain distance of i, we say that we cannot match treated unit i.

$$|\hat{e}(\mathbf{x}_i) - \hat{e}(\mathbf{x}_k)| > w$$

Where, if the distance is greater than the caliper w, we set the distance to infinity. Notice that this allows us to exclude both outliers and inliers. In practical terms, we often say a caliper that is 20% of the standard deviation of the propensity score is an appropriate caliper, of course it depends on your data and your story. When we drop treated observations, we are changing what we are estimating, but it is no longer ATT ...

3.2.4 How many neighbors do we match?

We may want to use multiple neighbors, either always or only if they are within a certain distance of one another. Keep in mind here that we then must reweigh the observations, much like we did with ties. However, with ties we equally weighed the observations, in many-to-one matching, the weights may be more complex.

3.3 `Match()`

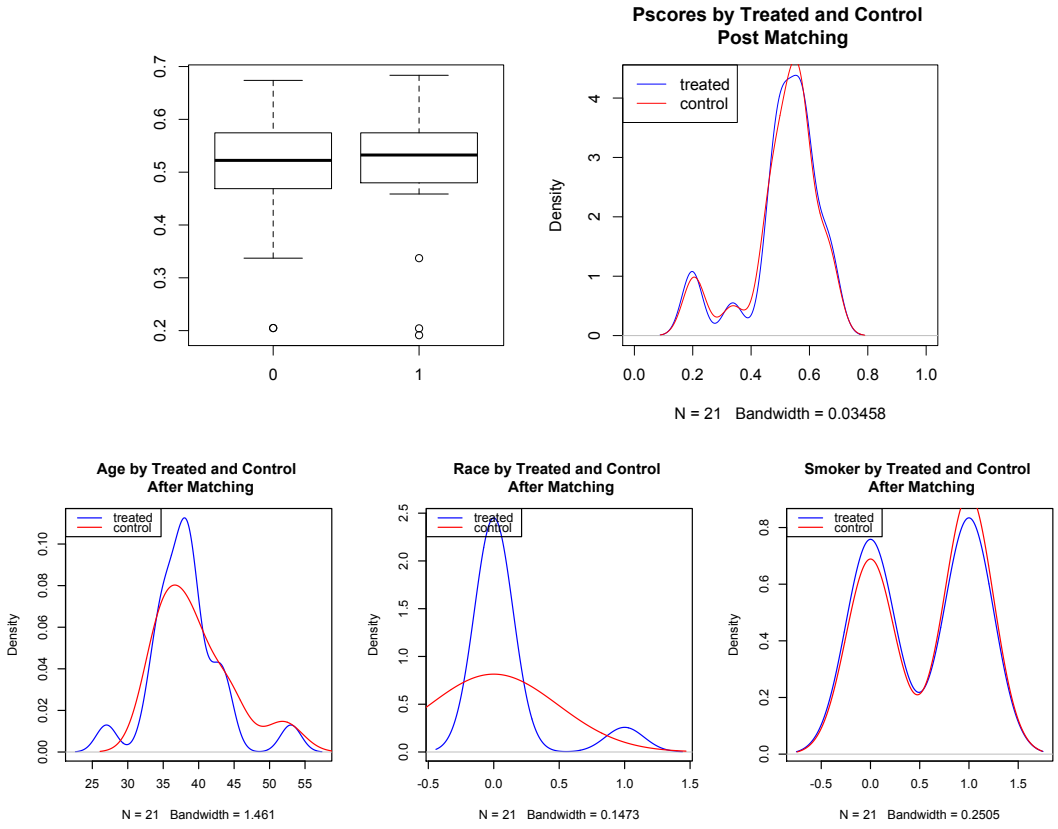
`Match` is a function in the `Matching` library that performs nearest neighbor matching.

```
library(Matching)
?Match
```

`Match` requires a treatment vector and an X vector on which you want to match, and an estimand (which defaults to ATT) so that it knows which units to match to which, and optionally you can pass in an outcome vector, in which case `Match` will also calculate the specified treatment effect. See the associated code for examples on how to use `Match`.

Here is an example from the R call:

```
m1 = Match(Tr = data$treat, X = data$pscore, estimand = "ATT")
```



3.4 A property of estimated propensity scores

Usually when we estimate parameters, the estimate doesn't perform as well as the true parameter. However, this is not true for propensity scores. Estimated propensity scores, $\hat{e}(\mathbf{x})$ tends to work slightly better than the true scores, $e(\mathbf{x})$ because estimated scores tend to slightly overfit the data. The effect of this is to produce slightly better than chance balance on the observed covariates in the data set that we estimated the propensity score from. Since our goal is to balance data, this is not a problem, and in fact helps us!

4 Beyond the Naive Model

The naive model helps us separate the two tasks in inference using matching. The first task is how do we create matched pairs. This is a fairly mechanical task. The second task is to decide whether or not those people that look comparable are comparable ... and this is not such a trivial task. We can try to get at the answer to this question using natural experiments, quasi-experiments, sensitivity analysis, etc. Ultimately, we are asking ourselves if our mechanical operations are sufficient for identification of our treatment effect. As Rosenbaum says, "The second task is not a mechanical but rather a scientific task, one that can be controversial and difficult to bring to a rapid and definitive closure; this task is, therefore, more challenging, and hence more interesting." This is where what we include in our propensity score model comes in ... if we condition on non-baseline covariates, we bias our estimate, but if we condition on too few covariates then we may not satisfy the strong ignorability of treatment assignment assumption.

5 Appendix

5.1 Law of Iterated Expectations

Based on Rocio's ATE and ATT notes

Theorem 1.1 *Law of Iterated Expectations (LIE). If X and Y are any two random variables, then*

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$$

Proof (Continuous Case) Let $f_{X,Y}(x,y)$ be the joint PDF of (X,Y), $f(x|y)$ be the conditional PDF of given Y = y and $f_Y(y)$ be the marginal PDF of Y. By definition of unconditional expectation we have

$$\begin{aligned}\mathbb{E}[X] &= \int x f_X(x) dx \\ &= \int\int x f_{X,Y}(x,y) dx dy\end{aligned}$$

Since $f_{X,Y}(x,y) = f(x|y)f_Y(y)$ we can replace $f_{X,Y}(x,y)$ to obtain:

$$\begin{aligned}\mathbb{E}[X] &= \int x f(x|y) f_Y(y) dx dy \\ &= \int \Big[\int x f(x|y) dx \Big] f_Y(y) dy\end{aligned}$$

where the expression in brackets is the conditional expectation $\mathbb{E}[X|y]$ and therefore:

$$\begin{aligned}\mathbb{E}[X] &= \int \mathbb{E}[X|y] f_Y(y) dx dy \\ &= \mathbb{E}[\mathbb{E}[X|Y]]\end{aligned}$$

Note that the law of iterated expectations does not assume independence, it holds for *any* two random variables.

5.2 Rubin's ATE and ATT notation

Rubin uses the response surface notation, where

$$R_1(x) \equiv \mathbb{E}[Y|X, T = 1]$$

$$R_2(x) \equiv \mathbb{E}[Y|X, T = 0]$$

So, the response surfaces are just the conditional expectation of Y given X in both the treated and the control populations. Rubin defined the effect of the treatment variable at $X = x$ as

$$R_1(x) - R_2(x)$$

which under the notation from earlier is just:

$$ATE(x) = ATT(x) = \mathbb{E}[Y|X, T = 1] - \mathbb{E}[Y|X, T = 0]$$

If $ATE(x) \equiv R_1(x) - R_2(x)$ does not depend on x (i.e. it is constant) then the response surfaces are parallel and the objective of the study will be to study this constant difference. If $ATE(x) \equiv R_1(x) - R_2(x)$ does depend on x then the response surfaces are not parallel and *there is no single parameter that completely summarizes the effect of the treatment variable.* When this happens, we will be interested in estimating some average effect of the treatment variable. In particular, we are usually interested in estimating the average difference between non-parallel response surfaces over P_1 (the treated population). This will be called the (unconditional) average treatment effect on the treated (ATT).

To recover $ATE(x)$ we average $ATE(x)$ over the distribution of X (we can only do this under some sort of identification assumption, which we assume holds). When we average $ATE(x)$ over the distribution of X we are actually using the law of iterated expectations. If we average over the entire distribution of X, we get the average treatment effect (ATE).

$$\begin{aligned}
\mathbb{E}[ATE(x)] &= \mathbb{E}\{\mathbb{E}[Y|X = x, T = 1] - \mathbb{E}[Y|X = x, T = 0]\} \\
&= \mathbb{E}\{\mathbb{E}[Y_1|X = x] - \mathbb{E}[Y_0|X = x]\} \\
&= \mathbb{E}\{\mathbb{E}[Y_1|X = x]\} - \mathbb{E}\{\mathbb{E}[Y_0|X = x]\} \\
&= \mathbb{E}[Y_1] - \mathbb{E}[Y_0] \\
&= \mathbb{E}[Y_1 - Y_0] \\
&\equiv ATE
\end{aligned}$$

where we used the law of iterated expectations to move from the third to the fourth line. Remembering that $ATE(x) = ATT(x)$, the average effect of treatment on the treated can be recovered by averaging $ATT(x) \equiv \mathbb{E}[Y|X = x, T = 1] - \mathbb{E}[Y|X = x, Y = 0]$ over the distribution of X conditional on $T = 1$.

$$\begin{aligned}
\mathbb{E}[ATT(x)|T=1] &= \mathbb{E}\{\mathbb{E}[Y|X=x,T=1] - \mathbb{E}[Y|X=x,T=0]|T=1\} \\
&= \mathbb{E}\{\mathbb{E}[Y_1|X=x] - \mathbb{E}[Y_0|X=x]|T=1\} \\
&= \mathbb{E}\{\mathbb{E}[Y_1|X=x]|T=1\} - \mathbb{E}\{\mathbb{E}[Y_0|X=x]|T=1\} \\
&= \mathbb{E}[Y_1|T=1] - \mathbb{E}[Y_0|T=1] \\
&= \mathbb{E}[Y_1 - Y_0|T=1] \\
&\equiv ATT
\end{aligned}$$

where the move between the fourth and fifth line is valid because in $\mathbb{E}[ATT(x)|T=1]$ the outer expectation is taken with respect to the distribution of X in the treatment group.

In Rubin's notation, the ATT:

$$ATT = \mathbb{E}_1[R_1(x) - R_2(x)]$$

There is one way in which Rubin's notation is actually more clear. For Rubin, $\mathbb{E}_1\{\cdot\}$ is the expectation over the distribution of X in the treated population (which he calls P_1). In the other notation, $\mathbb{E}\{\cdot|T=1\}$ means exactly the same thing, but it is less clear because when we condition on $T=1$ it is not obvious that we are actually conditioning on the distribution of X in the treated population. We are *not* conditioning on the fact that treatment was received, as it might appear.