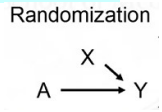
Module 3

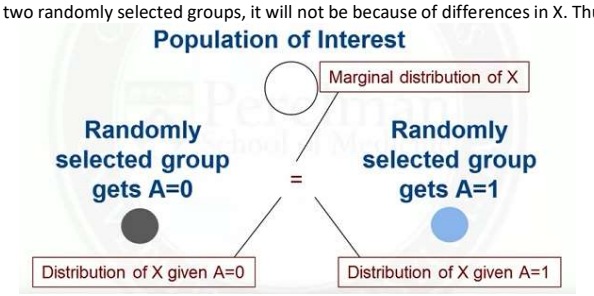
Monday, October 14, 2024 11:49 AM

Observational Studies

In a randomized trial, treatment assignment A would be determined by a coin toss, effectively erasing the arrow from X to A. So there is no backdoor path from A to Y via the covariates X:



Additionally, in a randomized trial, the distribution of covariates X will be the same in both treatment groups (covariate balance). Therefore, if the outcome distribution ends up differing in the two randomly selected groups, it will not be because of differences in X. Thus, X is dealt with during the design phase.



So why don't we always randomize trials?

- Randomized trials are expensive
- Sometimes randomizing treatment/exposure is unethical
- Some (even many) people refuse to participate in trials
- Randomized trials take time since you have to wait for the outcome data. In some cases, by the time you have the outcome data, the question may no longer be relevant.

So when you cannot do a randomized trial, instead try an observational study. Since the treatment isn't randomized in observational studies, the distribution of covariates X will differ between treatment groups.

There are two common methods of observational studies:

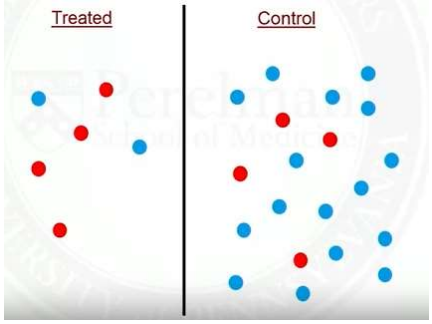
- Planned, prospective studies with active data collection. These are free of the potential unethical issues of randomized trials, though can still be slow to collect data and can still be expensive.
  - Like Trials: data collected on a common set of variables at planned times and outcomes are carefully measured using set protocols.
  - Unlike Trials: regulations are much weaker since you aren't intervening in any way (not manipulating treatment, randomizing the treatment). You're instead just observing what happens. A broader set of the population will be eligible for this type of study.
- Databases; retrospective studies with passive data collection. Since this data already exists, there are generally large sample sizes, and the studies are inexpensive. There is potential for rapid analysis. However, the data quality is typically lower since there is no uniform standard for data collection.

Matching is a method that attempts to make an observational study more like a randomized trial. The main idea is to match individuals in the treated group (A=1) to individuals in the control group (A=0) on the covariates X. The benefits of this include:

- Controlling for confounders is achieved at the design phase without looking at the outcome, so the difficult statistical work can be done completely blinded to the outcomes.
- Matching will reveal the lack of overlap in covariate distribution, so the positivity assumption will hold in the population that can be matched.
- Once data are matched, it is essentially treated as if it is from a randomized trial.
- Outcome analysis can be simple.

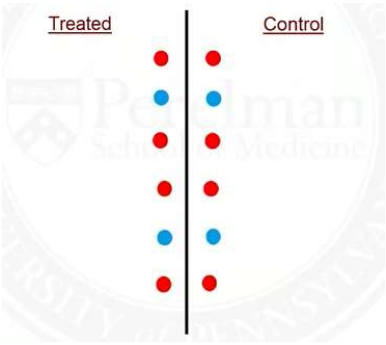
Matching Overview

Suppose we are dealing with a single covariate X of hypertension, where hypertensive patients (red) are more likely to be treated than patients without hypertension (blue).



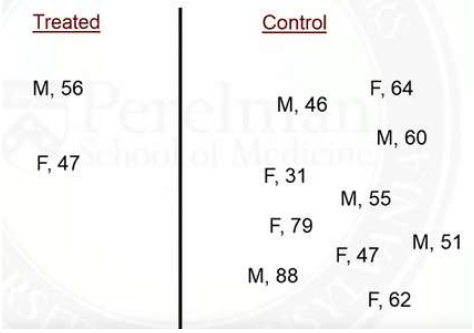
- 67% of treated subjects are hypertensive
- 20% of control subjects are hypertensive

This is handled by matching up each treated subject to a similar control subject, thus leveling out the covariate balance.

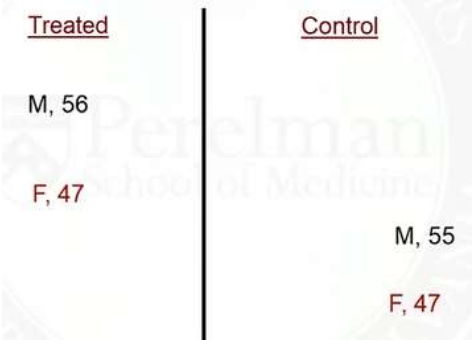


If there are many covariates, we will likely not be able to exactly match on the full set of covariates. In randomized trials, treated and control groups are not perfect matches either (the distribution of covariates is balanced between groups; this is stochastic balance). Matching closely on covariates can achieve stochastic balance; i.e., the matches will be close and the distribution of the covariates will be very similar between the treated and the control groups.

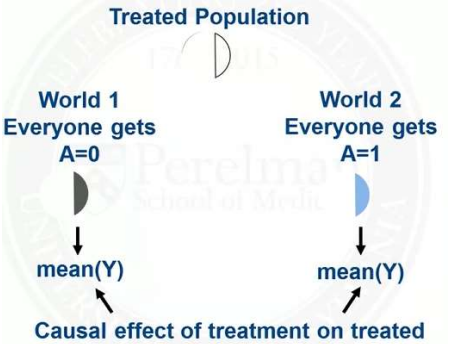
Suppose we have two covariates: sex and age.



We match as closely as possible on the two covariates.



Notice that through this process, we are making the distribution of covariates in the control group look like that of the treated group. This is the causal effect of treatment on the treated and we are making inferences about people who received treatment. We are comparing the outcome of the treated group to what their outcome would have been if they had not received treatment.



It is possible to match in such a way that you make the control group and treatment group look like each other, but that is a more involved and complex process.

Sometimes it is difficult to find great matches. We may be willing to accept some non-ideal matches if the control group and treated group have the same distribution of covariates. This is known as fine balance. In the below example, we haven't achieved stochastic balance, but we do have fine balance since the average age and gender distributions are the same in both groups, even though neither match is great.

- Match 1: Treated, Male, Age 40 | Control, Female, Age 45
- Match 2: Treated, Female, Age 45 | Control, Male, Age 40.

Another issue that occurs with matching is the number of matches.

- One to one (pair matching): match exactly one control to every treated subject
- Many to one: Match some fixed number K control subject to every treated subject (like 5 to 1 matching)
- Variable: sometimes match 1 or sometimes more than 1 control to treated subjects. If multiple good matches are available, use them. If not, do not.

Matching Directly on Confounders

Since we typically cannot match exactly, we need to choose some metrics for closeness.

Let X_j be the vector of covariates for subject j.

Then the Mahalanobis Distance between covariates for subject i and subject j is the square root of the sum of squared distances between each covariate vector ($vector^{transposed} * vector$), scaled by the covariance matrix $s = cov(X)$:

$$D(X_i, X_j) = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)}$$

Example with 3 covariates: age, COPD (1=yes, 0=no), Female (1=yes, 0=no)

Treated			Control			
Age	COPD	Female	Age	COPD	Female	**Distance**
78.17	0	1	70.25	1	0	4.23
			75.33	0	1	0.17
			86.08	1	1	3.72
			54.97	0	0	2.45
			43.63	0	0	2.89
			18.04	0	1	3.60

Next, the motivation behind the Robust Mahalanobis Distance is that outliers can create large distances between subjects even if the covariates are otherwise similar; therefore, ranks may be more relevant. Thus the steps for calculating Robust Mahalanobis Distance are:

1. Replace each covariate value with its rank.
    a. For each covariate, rank each subject 1,2,3, etc., based off how similar they are to the control subject.
2. there will be a constant diagonal on the covariance matrix
3. Calculate the usual Mahalanobis Distance on the ranks.

Practice Quiz

1. Balance refers to:
    a. The distribution of confounders being similar for treated and untreated subjects
2. In matching, distance is a measure of:
    a. How similar the values of confounders are for different people
3. A good match is one where:
    a. The distance is small

Greedy (nearest-neighbor) Matching

Suppose we have selected a set of pre-treatment covariates X that satisfy the ignorability assumption, we have calculated a distance d_{ij} between each treated subject and every control subject, and we have more control subjects than treated subjects. We are pair matching the subjects.

Greedy Matching is:

- Intuitive and easy to explain
- Computationally fast
    - Involves a series of simple algorithms
    - Fast even for large data sets
    - R package: MatchIt
- Not invariant to initial order of list; the matches may change depending on the order of the list
- Not globally optimal
    - Always taking the smallest distance match does not minimize total distance
    - Can lead to some bad matches

Then the steps to Greedy Matching (1 to 1) are:

1. Randomly order the list of treated subjects and control subjects
2. Start with the first treated subject and match to the control with the smallest distance (this is greedy)
3. Remove the matched control from the list of available matches
4. Move on to the next treated subject and match to the control with the smallest distance
5. Repeat steps 3-4 until all subjects have been matched

For Greedy Matching (Many to one):

1. After everyone has 1 match, go through the list again and find 2nd matches
2. Repeat until k matches

Tradeoffs:

- Pair matching has closer matches and computing times. Less bias because the matches are closer, but more variance.
- Many-to-one matching has a larger sample size. More bias, but smaller variance.

A bad match can be defined using a caliper; a maximum acceptable distance between subjects. We may prefer to exclude treated subjects for whom there does not exist a good match.

- Only match a treated subject if the best match has distance less than the caliper
- Recall the positivity assumption: that the probability of each treatment given X should be non-zero.
    - If there are no matches within caliper, it is a sign that the positivity assumption would be violated

- Excluding these subjects makes the assumption more realistic
- However, it makes the population harder to define ("the population is all treated subjects except those for which no match is available")

Optimal Matching

Optimal Matching refers to the set of matches that has the smallest total distance. Greedy Matching is not optimal because it doesn't guarantee the set of matches with the smallest total distance.

Whether or not it is feasible to perform Optimal Matching depends on the number of pairings to make:
- 100 treated subjects and 1000 controls results in 100,000 possible matches
- 1 million treatment-control pairings is feasible (but slow) on most computers
- 1 billion pairings (10,000 treated and 100,000 subjects) is not feasible

Sparse optimal matching refers to placing constraints to make optimal matching computationally feasible for larger data sets. For example, if you only match within the same disease category or the same hospitals. Mismatches can be tolerated if fine balance is still achieved.

Assessing Balance

After matching, we need to assess whether the matching worked.

One way to do this is to calculate the covariate balance by standardized differences to see if the means between the groups are similar. This should be done without looking at the outcome. The standardized mean difference is the difference in means between groups, divided by the (pooled) standard deviation.

$$smd = \bar{X}_{treatment} - \bar{X}_{control} \Big/ \sqrt{s^2_{treatment} - s^2_{control}/2}$$

- SMD doesn't depend on sample size
- Often |SMD| is reported
- Calculate SMD for each variable that is matched on and then the rule of thumb is as follows:
  - Values < 0.1 indicate adequate balance
  - Values (0.1, 0.2) are not too alarming
  - Values > 0.2 indicate serious imbalance

	Unmatched			Matched		
	No RHC	RHC	SMD	No RHC	RHC	SMD
n	3551	2184		2082	2082	
age (mean (sd))	61.8 (17.3)	60.8 (15.6)	0.06	61.6 (16.7)	61.0 (15.8)	0.039
sex = Male (%)	53.9	58.5	0.09	56.9	56.9	0.001
resp = Yes (%)	41.7	28.9	0.27	30.6	30.4	0.005
card = Yes (%)	28.4	42.3	0.30	39.3	39.5	0.004
neuro = Yes (%)	16.2	5.4	0.35	5.3	5.7	0.015

In the above example before matching, we see that covariates Age and Sex are adequately balanced, but Resp, Card, and Neuro indicate serious imbalance between groups. After (pair) matching, the SMD indicates balance in all covariates.

Another option is to assess balance with hypothesis tests; test for a difference in means between treated and controls for each covariate with two sample t-tests and then reporting the p-value for each test.
- This has the drawback of p-values being dependent on sample size, so small differences in means will have a small p-value if the sample size is large (though we probably don't care much if the mean differences are small and so hypothesis testing may not be the best approach).

Analyzing Data after Matching

After matching and assessing that the balance of matches is adequate, we need to identify statistical methods for analyzing the matched data. So we proceed with outcome analysis:
- Test for a treatment effect
- Estimate a treatment effect and confidence interval
- Methods should take matching into account

First, testing for the treatment effect involves Randomization Tests, otherwise known as Permutation Tests or Exact Tests. The general process is:
1. Compute the test statistic from observed data
2. Assume null hypothesis of no treatment effect is true
3. Randomly permute treatment assignment within pairs and re-compute the test statistic
4. Repeat many times to see how unusual the observed statistic is.

As an example, let's use the table of observed data below:

Matched pair	Treated	Control
1	0	0
2	1	0
3	1	0
4	0	0
5	1	1
6	0	0
7	0	1
8	0	1
9	0	0
10	1	0
11	0	0
12	1	0
13	1	0

- There are six rows in the treatment group where the outcome = 1; test statistic = 6
  - Would this number be unusual if the null hypothesis was true? i.e., if treatment had no effect, would this number be unusual?
- To check this, we need to do step 3:
  - First identify the discordant pairs:

Matched pair	Treated	Control
1	0	0
2	1	0
3	1	0
4	0	0
5	1	1
6	0	0
7	0	1
8	0	1
9	0	0
10	1	0
11	0	0
12	1	0
13	1	0

  - Next we randomly permute treatment assignment.
    - This can be thought of as this: for each matched discordant pair, I flip a coin. If heads, they stay the same. If tails, I flip the outcome of the treated and control group.

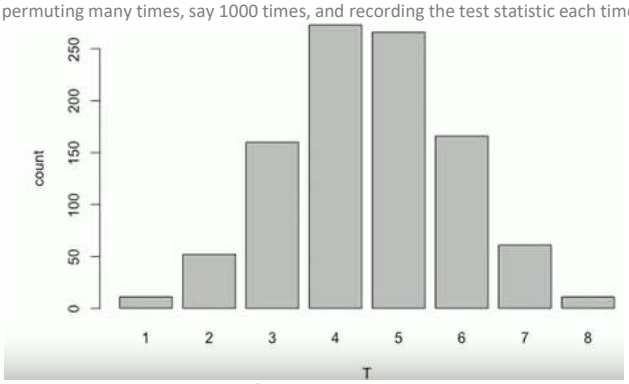
Matched pair	Treated	Control
1	0	0
2	1	0
3	0	1
4	0	0
5	1	1
6	0	0
7	1	0
8	0	1
9	0	0
10	0	1
11	0	0
12	1	0
13	1	0

  - Next we recalculate the test statistic
    - There are 5 rows in the treatment group where the outcome = 1: test statistic = 5
  - And repeat step 3 again with a new permutation

Matched pair	Treated	Control
1	0	0
2	1	0
3	1	0
4	0	0
5	1	1
6	0	0
7	1	0
8	1	0
9	0	0
10	1	0
11	0	0
12	1	0
13	0	1

        - Test statistic = 7
    - Repeat many times times to get more test statistics:
        - Test statistic = 4
        - Test statistic = 5
        - Test statistic = 6
    - Etc

After permuting many times, say 1000 times, and recording the test statistic each time, we get the following distribution:



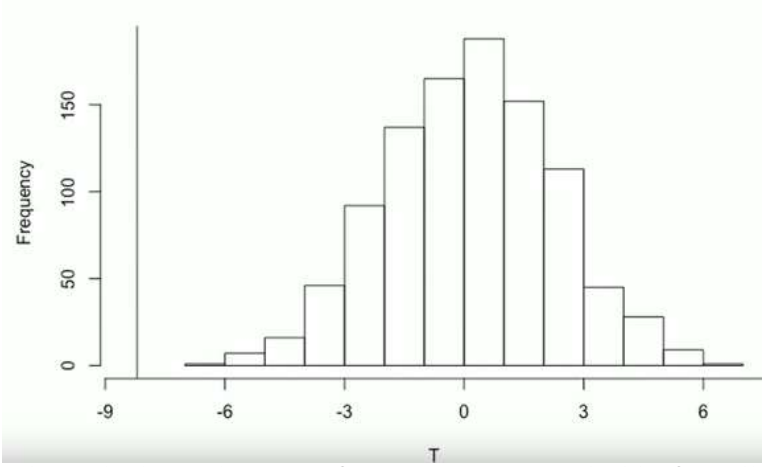
- We can see that test statistics of 1 and 8 are unusual, while 4 and 5 are very common.
- Our original observed data test statistic was 6, which is pretty consistent with what we would expect based off this distribution.
- Therefore what we observed is not inconsistent with the null hypothesis of no treatment effect and therefore we don't have any evidence that treatment is doing anything.

- In p-value terms, we would calculate the probability of seeing something as extreme or more extreme than our observed test statistic = 6. (probability of T = {6,7,8})
    - Since this is a symmetric distribution, we would need to include the probability of T = {1, 2, 3}
    - Thus P(T = {1,2,3,6,7,8})
- This test is equivalent to the McNemar test for paired data (can be done in R with package McNemar)

This basic approach also works for continuous data. Suppose we have systolic blood pressure outcome data from 20 matched pairs and the test statistic is the difference in sample means.

Pair ID	Control SBP	Treated SBP	Pair ID	Control SBP	Treated SBP
1	106	101	10	109	98
2	110	105	11	109	101
3	111	100	12	109	106
4	118	111	13	106	100
5	108	92	14	103	89
6	110	101	15	104	91
7	108	95	16	109	108
8	114	116	17	111	105
9	109	109	18	104	98
10	109	98	19	106	90
11	109	101	20	112	97

- $test\ statistic = \bar{x}_{control} - \bar{x}_{treatment} = -8.2\ mm\ Hg$
- Then we randomly permute data again to get estimated test statistics, then get the distribution of the estimated test statistics:



- And we see that the original test statistic of -8.2mm Hg seems very unusual and therefore it is likely that treatment is having an effect.
- The p-value is 0, since we never saw a single estimated statistic as extreme or more extreme than -8.2 mm Hg.
- Thus we reject the null hypothesis that there is no treatment effect
- So we have strong evidence that the treatment does have a negative-direction impact on systolic blood pressure.

- This can be calculated with a paired t-test (for continuous data).

Other outcome models you may want to use:

- McNemar Test: matched data
- T-Test: continuous data
- Conditional Logistic Regression: matched binary outcome data
- Stratified Cox Model: time-to-event (survival) outcome data. Baseline hazard stratified on matched sets
- Generalized Estimating Equations (GEE): Match ID variable used to specify clusters. For binary outcomes, can estimate a causal risk difference, causal risk ratio, or causal odds ratio (depending on the link function)

Practice Quiz

1. For Optimal Matching, what is optimized?
   a. Total distance is minimized
2. After matching, balance can be assessed by:
   a. Standardized differences
3. A large standardized difference for a covariate suggests:
   a. There is imbalance on this covariate
4. If there is a large treatment effect, then we expect the observed difference in mean of the outcome between matched pairs to be:
   a. Very difference from the difference in means if we randomly permute the treatment labels

Sensitivity Analysis

Randomized trials achieve balance on both observed and unobserved variables.

Matching aims to achieve balance on observed covariates. Overt bias could occur if there was an imbalance on observed covariates (i.e., if we did not fully control for the variables).

There is no guarantee that matching will result in balance on variables that we did not match on (including unobserved variables). If these unobserved variables are confounders, then we have hidden bias and the ignorability assumption was violated.

Sensitivity analysis addresses the hidden bias issue: if there is hidden bias, determine how severe it would have to be to change the conclusions; i.e., change from statistically significant to not significant or change direction of effect.

Consider the following:

- Let π_j be the probability that person j recieves treatment
- Let π_k be the probability that person k recieves treatment
- Suppose that person j and k are perfectly matched so that their observed covariates $\{X_j, X_k\}$ are the same

If $\pi_j = \pi_k$, then there is no hidden bias.

Consider this inequality:

$$\frac{1}{\Gamma} \leq \frac{\frac{\pi_j}{(1-\pi_j)}}{\frac{\pi_k}{(1-\pi_k)}} \leq \Gamma$$

- The numerator is the odds of treatment for person j
- The denominator is the odds of treatment for person k
- Therefore Γ is an odds ratio

Then:

- If $\Gamma = 1$ there is no overt bias
- If $\Gamma > 1$ there is hidden bias (person j is more likely to receive treatment than person k)
- The severity of the ignorability assumption's violation depends on how much bigger than one Γ is.

With the value of Γ, we can do a sensitivity analysis. Suppose we have evidence of treatment effect, under the assumption $\Gamma = 1$ and we assume no hidden bias.

- We can then increase Γ until evidence of treatment effect goes away (is no longer statistically significant).
  - If this happens quickly, like at $\Gamma = 1.1$, then we can say it is very sensitive to unmeasured confounding / hidden bias.
  - If this happens slowly, like at $\Gamma = 5$, then we can say it is not very sensitive to unmeasured confounding / hidden bias.
- R packages that can compute this: sensitivity2x2xk, sensitivityfull

Practice Quiz

1. Matching and calculation of standardized differences of the matched data can take place without using the outcome variable:
   a. True
2. Optimal Matching is less computationally demanding than greedy matching:
   a. False
3. A smaller value of the caliper would tend to lead to:
   a. Smaller standardized differences
4. Many-to-one matching, as opposed to pair matching, would tend to lead to estimates of causal effects that have:
   a. More bias, but less variability
5. A method for assessing the impact of violations of causal assumptions is:
   a. Sensitivity analysis
6. Standardized differences are very sensitive to sample size:
   a. False

Propensity Scores

Propensity Score: the probability of receiving treatment, rather than control, given covariates X. Let A = 1 for treatment and A=0 for control, then we will denote the propensity score for subject i by π_i. Thus the propensity score is a function of the covariates. Then:

$\pi_i = P(A = 1|X_i)$

For example, if age was the only covariate and older people were more likely to get treatment, then the propensity score would be larger for older people: $\pi_i > \pi_j$ if $age_i > age_j$. So if person i has a propensity value of 0.3, that means that given their particular covariate values, there is a 30% chance they will be treated.

Suppose two people have the same propensity score, but different covariate values X.

- This means that both subjects' covariates are just as likely to be found in the treatment group as the control group.
- If we restrict to a subpopulation of subjects who have the same value of the propensity score, then there should be balance between the two treatment groups.
- Thus the propensity score is a balancing score.
- So if we match on the propensity score, we should achieve balance. This makes sense considering we assumed ignorability (that treatment is randomized given X).
  - Conditioning on the propensity score is conditioning on an allocation probability.

Formally, this can be written as:

$P(X = x|\pi(X) = p \cap A = 1) = P(X = x|\pi(X) = p \cap A = 0)$

In a randomized trial, the propensity score is generally known, like $P(A = 1|X) = P(A = 1) = 0.5$

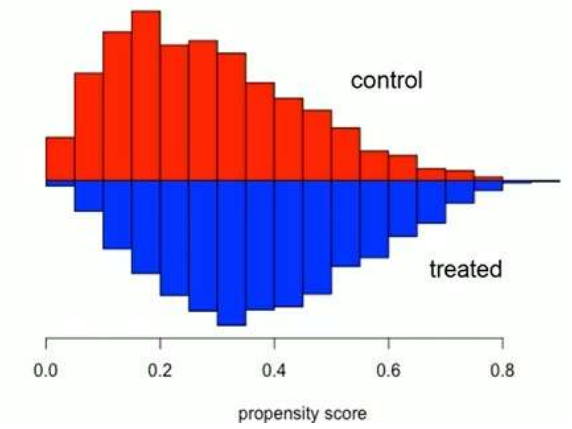
But in an observational study, the propensity score will be unknown. However, since the propensity score involves observed data of the treatment and covariates, we can estimate it. So when people typically talk about the propensity score, they are referring to the estimated score.

We need to estimate $P(A = 1|X)$, where the outcome here is A (a binary value). We therefore could estimate the propensity score using logistic regression (though other methods for estimating a propensity score could be used, including machine learning methods):

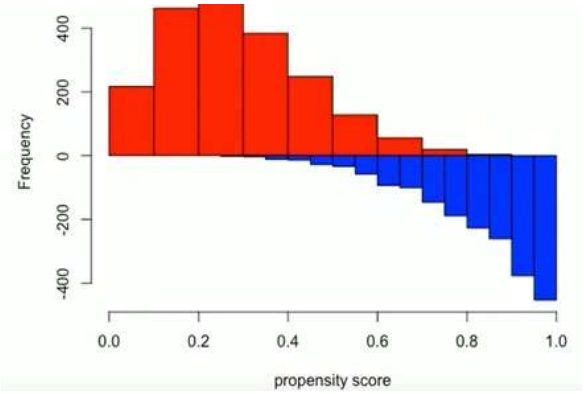
1. Fit a logistic regression model: Outcome A, Covariates X
2. From that model, get the predicted probability (fitted value) for each subject. This is the estimated propensity score.

Propensity Score Matching

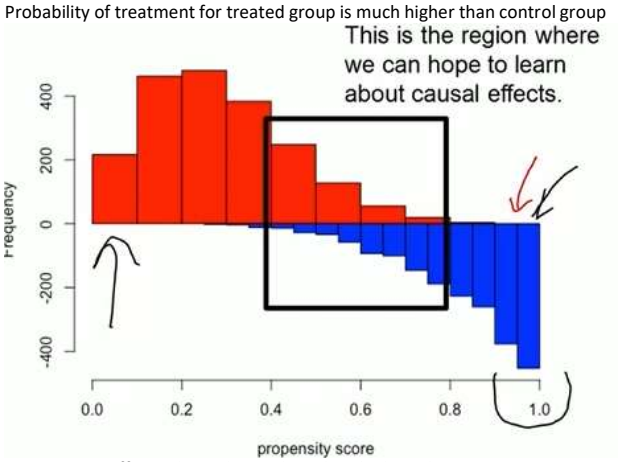
Recall that matching on the propensity score should achieve balance in the control and treatment groups. Once the propensity score is estimated, but before matching, it is useful to look for overlap (this is normally done with a plot). By overlap, we mean that there shouldn't be any cases where one group has a 0 people while the other more than zero. Even if the size of both groups at the same propensity score is different, one group shouldn't be zero.



- The treated group appears to have a higher probability of getting treatment (as expected)
- But there is overlap everywhere.
- Positivity assumption is reasonable



- Poor overlap
- Positivity assumption likely violated (since some people in treated group are guaranteed to get treatment).
- Probability of treatment for treated group is much higher than control group



- Treatment is effectively random only within the boxed-in range
  - The group smaller than the max control score and bigger than the min treated score.

If there is a lack of overlap, trimming the tails is an option; removing subjects who have extreme values of propensity score. Trimming the tails makes the positivity assumption more reasonable and prevents extrapolation.

Thus we can proceed with matching by computing a distance between the propensity score for each treated subject with every control. Then use optimal or greedy matching. In practice, logit (log-odds) of the propensity score is often used rather than the propensity score itself.

- The propensity score is bounded between 0 and 1, making many values seem similar
- Logit of the propensity score is unbounded - this transformation essentially stretches the distribution while preserving ranks.
- So match on $logit(\pi)$ rather than π.

Additionally, we can ensure that we do not accept any bad matches by using a caliper - the maximum distance we are willing to tolerate. In practice, a common choice for a caliper is $0.2 * SD(logit(\pi))$. Note that the smaller caliper means less bias, but more variance.

After matching, the outcome analysis methods can be used the same as would be done if matching on covariates; randomization tests, conditional logistic regression, GEE, stratified cox model, etc.

Practice Quiz

1. The propensity score is:
   a. The probability of treatment given covariates
2. Trimming the tails involves:
   a. Excluding subjects who have extreme values of the propensity score
3. If the propensity score is exactly equal to 0 or 1 for some subjects, what causal assumption is violated?
   a. Positivity assumption
4. Propensity score matching involves the following steps, in order:
   a. Estimate Propensity Score
   b. Check propensity Score Overlap
   c. Match on Propensity Score
   d. Check Covariate Balance
5. If we use a caliper on the propensity score of 0.1:
   a. Matches will never differ in the propensity score by more than 0.1