Fundamentals of causal inference: part 3

Justin Sheen

August 2, 2024

1 Introduction

The following is adapted from module three of Prof. Jason Roy's online Coursera course on causal inference https://www.coursera.org/learn/crash-course-in-causality.[1] I wanted to review the basics of causal inference for myself. This is part three of five.

2 Observational studies

Observational studies differ from randomized trials. They are subject to confounders, since treatment has not been randomized. Imagine $Z \rightarrow Y$ and $Z \leftarrow X \rightarrow Y$. In this scenario, what randomization effectively does is eliminate $Z \leftarrow X$ so that there are no backdoor paths from Z to Y so that ignorability is satisfied. More specifically, the distribution of X is now randomly distributed to both treatment groups, i.e., the marginal distribution of X is the same as the distribution of X given $Z = 1$ or $Z = 0$. This is known as *covariate balance*, and is dealt in the *design phase*.

So why not randomize everything? Because randomized trials are very expensive and lengthy, or could be unethical (e.g., randomizing people to smoke), or people might not want to have the chance to receive a placebo thus shrinking the study population. Planned, prospective observational studies could be used instead, although of course there are drawbacks such as low data quality, and they can still at times be expensive.

2.1 Matching

2.1.1 Description

Matching is a way that observational studies try to address confounding. Imagine X is age, and older people are more likely to have $Z = 1$ and younger people are more likely to have $Z = 0$. In matching, we try to find, for each treated person $Z = 1$ of a certain age, we try to find an untreated person $Z = 0$ with the same age to perform a comparison. In a randomized trial for any age we would expect to find about the same number whether $Z = 1$ or $Z = 0$. Matching tries to achieve this.

Matching is nice because we can do it before looking at outcomes, as it is still done in the design phase. It can also reveal whether, for example, some ages had no chance of receiving treatment, and we would try to exclude them from the study since we don't want to make inference on people who had no chance of either receiving treatment or not receiving treatment as this would violate the positivity assumption.

What occurs more specifically in matching is that we match a person with $Z = 1$ and a given X to a person with $Z = 0$ with the same X. Individuals in the smaller group will try to find matches in the bigger group; usually the treatment is the smaller group and control is the bigger group. Then we get rid of all the others that didn't have a match.

We will not be able to exactly match on the full set of covariates, for example maybe we couldn't find an exact match of age 55 in the control, but we could find an age 56, which may be close enough. But even in randomized trials we may not have exact matches, we have *stochastic balance.* Matching will also try to achieve stochastic balance.

It's subtle, but matching will only provide us with the causal effect of treatment on the treated, since we are finding matches in the control group that match individuals in the treatment group. That is, if we took individuals from the treated population, not the whole population, we ask the causal question of what would happen to them if $Z = 0$ or $Z = 1$. So we're not trying to match the distribution of X to the marginal distribution of X across treatment groups.

Sometimes it is difficult to find great matches. For example, if a treated male of age 40 is matches to a control female of age 45 and a treated female of age 45 is matched to a control male of age 40, this would not have stochastic balance. However, this does have *fine balance* since the distribution of X is the same among treated and control.

There's also a question of how many matches to make. For example, although so far we've done one to one matching, we can also do five to one matching where for every treated individual we can match them to five control individuals.

2.1.2 How do we create good matches?

We can create good matches by creating distance matrices between individuals based on their X covariates. There is the Mahalanobis distance (M distance) and robust Mahalanobis distance (robust M distance). Denote X_j as a vector of covariates for subject j. The Mahalanobis distance between subject i and j is:

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

Where S is the symmetric covariance matrix of X, which is done to scale the variables. So in the first matrix multiplication its looking at the difference in the first variable of X_j and then scaling by the variance of the first variable with itself plus the difference in the second variable with the second variable of X_j and then scaling by the covariance between the first variable and second variable, etc. for all other differences of the other variables and covariance with the first variable. Essentially we square the differences of each variable, scale by the covariance, and take the square root.

The robust Mahalanobis distance tries to account for the fact that outliers can create large distances, so we might be better off trying to match the ranks. One subtlety is that the diagonal of the covariance matrix will be a constant.

Greedy (nearest-neighbor) matching First, imagine the pool of control subjects is much larger than for treated subjects, which is often the case. The algorithm is:

1. Randomly order list of treated subjects and control subjects
2. Start with first treated subject. Match to control with the smallest distance (this is greedy)
3. Remove the matched control from the list of available matches
4. Move on to next treated subject and repeat for all treated subjects

This is intuitive and computationally fast, but it depends on the initial order of the list, and may not be globally optimal.

This is the algorithm for pair-matching, but for many to one matching the algorithm would slightly change to first find matches for all treated subjects, then go back and find second matches for all treated subjects. Many to one matching may be preferred since it uses more control subjects, but it may also lead to worse matches. There is essentially a bias-variance tradeoff where there may be closer matches in pair matching but larger variance, and vice-versa in many to one matching.

A *caliper* is a maximum acceptable distance to avoid very bad matches. This gets back to the positivity assumption, since probability of each treatment should be non-zero.

Optimal matching Finding the global optimal match is computationally expensive, but R packages have been developed to do this. But it also depends on the size of the problem. For example even if there are a million possible pairings (e.g. 1000 possible treated and 1000 possible control) this should still be runnable. We can constrain the number of pairings by blocking globally optimal matches within certain subsets, subpopulations, of the data (also known as sparse matching).

To assess the balance we can look for example at the means of the covariates and see whether they are similar between the two groups before and after matching. We can also take a look at the standardized difference between the two groups:

$$smd = \left| \frac{\overline{X}_{treatment} - \overline{X}_{control}}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}} \right|$$

Where values < 0.1 indicate adequate balance, values between 0.1 and 0.2 are O.K. and values greater than 0.2 indicate imbalance.

3 Analyzing matched data

Imagine we've successfully matched the data. We can analyze the data with *permutation tests*. The main idea is to first create a test statistic from the observed data, such as a difference in sample means between the two groups. Then assume the null hypothesis of no treatment effect is true. Randomly permute the treatment and control labels for each pair and calculate the test statistic, and do this thousands of time to create a null distribution.

Then check whether the observed test statistic is unusual compared to the null distribution. A McNemar test can also be used for paired data or a paired t-test for continuous data that tests against the null of no difference between the treatment conditions of each pair. There are also more complicated models that can be used such as a conditional logistic regression, stratified Cox model, and generalilzed estimating equations (GEE).

Sensitivity analyses should also be done for analysis. In observational studies especially, there may be hidden biases where the ignorability assumption is violated. For observational studies we can determine if there is hidden bias, how severe would it have to be for us to change conclusions? Let π_j and π_k be the probabilities that person j and k receive treatment. Suppose person j and k are perfectly matched so that the observed covariates X_j and X_k are the same. If $\pi_j = \pi_k$ then there is no hidden bias. Consider:

$$\Gamma = \frac{\frac{\pi_j}{1-\pi_j}}{\frac{\pi_k}{1-\pi_k}}$$

The numerator is the odds of treatment for person j and the denominator is the odds of treatment for person k. Γ is an odds ratio. If $\Gamma = 1$ then there is no overt bias and $\Gamma > 1$ or $\Gamma < 1$ implies hidden bias. So we can increase Γ until the conclusion changes due to, for example, a wider range of p-values at some level of Γ that cross the $\alpha = 5\%$ significance level. For an example of how the computation for this is done for the Wilcoxon-signed rank test, which essentially gives a range for the null-distribution depending on Γ see Rosenbaum (2005).[2] The point at which the conclusion changes due to a bias in the odds of treatment assignment gives us some sense of how sensitive our results are to hidden biases: for example if $\Gamma = 1.1$ for the conclusion to change, then this is very sensitive to confounding, compared to if conclusions change when $\Gamma = 5$, since the latter is saying that there must have been a confounder that assigned treatment 5 times more often to treatment than control to change the result.

4 Propensity scores

A propensity score is the probability of receiving treatment rather than control given covariates X. That is, for subject i:

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

To make it more concrete, if older people were more likely to get treatment, then $\pi_i > \pi_j$ if $age_i > age_j$.

Two individuals, i and j where $X_i \neq X_j$ may still have $\pi_i = \pi_j$, so they are just as likely to be found in the treatment group. If we only look at a subpopulation of subjects with the same π_i, there should be balance in the two treatment groups. In this case the propensity score is a *balancing score.* More formally:

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

If we match on the propensity score, rather than the raw covariate X as before, we should achieve balance, since the individuals of the treatment and the control should have had equal probability of being included in either the treatment or control group. In other words the probability of treatment is random given X if we match on the propensity scores. Note though that the propensity scores will not be randomly allocated to both groups since higher propensity scores should be seen more often in treatment since they are have a higher probability of being in the treatment group. Furthermore in theory the above states that the distribution of X given $\pi(X) = p$ should be the same for both groups if we match on the propensity scores. Conditioning on the propensity score is conditioning on an *allocation probability*.

In a randomized trial the propensity score is known, but in an observational study it is unknown, but we can estimate it. Treat the treatment as an outcome: $P(Z = 1|X)$. We can fit a logistic regression model of Z vs. X. From the model we can get a predicted probability for each subject, which is the estimated propensity score. This is very efficient to match on since we just have to match on a single variable for balance rather than balancing all covariates.

Before we match we need to look for *overlap*, where in the distribution of the propensity score among the two groups, there is always positivity no matter the propensity score. We could also check to see that there ended up being a higher peak among treated individuals, since they were more likely to have a higher probability of receiving treatment. Sometimes there is poor overlap at the tails, but causal effects can be estimated where there is overlap and trimming off these tails, and satisfying the positivity assumption. In practice we can also transform the propensity score to the log-odds to "stretch" the distribution, so we would match on $logit(\pi)$ rather than π, as well as use a caliper that is 0.2 times the standard deviation of the logit of the propensity score, although this is somewhat arbitrary. A smaller caliper has less bias, but more variance.

Then like before a randomization test for analysis can be used after matching on the propensity scores.

References

[1] J. Roy. A crash course in causality: Inferring causal effects from observational data. `https://www.coursera.org/learn/crash-course-in-causality`.

[2] Paul R Rosenbaum. Sensitivity analysis in observational studies. *Encyclopedia of statistics in behavioral science*, 4:1809–1814, 2005.