Fundamentals of causal inference: part 4

Justin Sheen

August 21, 2024

1 Introduction

The following is adapted from module four 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 four of five.

2 Inverse probability of treatment weighting (IPTW)

Imagine that the propensity scores for some covariate X is $P(Z = 1|X = 0) = 0.1$. Then it is likely that most of them will be assigned to control and if $X = 1$ it is likely that most of them will be assigned to treatment. Say there are 9 individuals with $X = 0$ in control and 1 individual with $X = 0$ in treatment. Then the single individual in treatment needs to matched to one in control, but we would discard 8 individuals in the control. Rather than discard this data, we can weight them by *the inverse of the probability of treatment (IPTW)*. So for example, for $X = 0$, for control subjects, we would weight by $P(Z = 0|X = 0)$ and for treatment subjects we would weight by $P(Z = 1|X = 0)$. In this example the individual in the treatment would have weight 10 and the individual in the control would have weight 10/9. If $P(Z = 1|X = 0)$ equalled some probability greater than 0.5 then treatment individuals would be down-weighted and control individuals would be up-weighted. In this way we don't have to discard data.

2.1 Further motivation

In survey sampling it is common to oversample some groups relative to the population. For example, maybe there is a subpopulation we want to make sure we have in our sample, so we consciously try to oversample that population. We need some way to account for this oversampling when we estimate the population mean, which we can do with the *Horvitz-Thompson estimator*.

Similarly, in observational studies, due to confounders we'll end up oversampling certain groups relative to in a randomized trial i.e., when $P(Z|X) \neq 0.5$ if there is a single treatment and control group and two possible values of Z and X. IPTW essentially creates a *pseudo-population* where there is no oversampling. For example in the above example, the

pseudopoulation would have 10 individuals in treatment and 10 in control once we up-weight the individual in treatment and down-weight each individual in control.

What does this look like in estimation? Say we want to know $E[Y(Z=1)]$. Under the assumption of exchangeability (i.e. ignorability) and positivity, we can estimate this as:

$$\frac{\sum_{i=1}^{n} I(Z_i=1)\frac{Y_i}{\pi_i}}{\sum_{i=1}^{n} \frac{I(Z_i=1)}{\pi_i}}$$

Where $\pi_i = P(Z=1|X_i)$ is the propensity score. So this is saying that we want to count, among all i individuals that were assigned treatment, in the numerator we want to count the sum of Y_i in the treated pseudopopulation, and in the denominator we just get the number of individuals in the treated pseudopopulation.

3 Marginal structural models

A marginal structural model is a model for the mean of potential outcomes. This is a model that is not conditional on confounders, and is made for potential outcomes rather than observed outcomes. To make things more concrete let's look at a linear model:

$$E[Y(Z)] = \psi_0 + \psi_1 z; z = 0, 1$$

$$E[Y(Z=0)] = \psi_0$$

$$E[Y(Z=1)] = \psi_0 + \psi_1$$

Therefore, ψ_1 is the *average causal effect* $E[Y(Z=1)] - E[Y(Z=0)]$, which is the expected difference in potential outcomes. Although this appears similar to a linear regression, it is not exactly because we are modeling potential outcomes rather than observed outcomes.

We can also make a logistic marginal structural model:

$$logit(E[Y(Z)]) = \psi_0 + \psi_1 z; z = 0, 1$$

Here, if we exponentiate ψ_1 this is a *causal odds ratio*:

$$\frac{\frac{P(Y(Z=1)=1)}{1-P(Y(Z=1)=1)}}{\frac{P(Y(Z=0)=1)}{1-P(Y(Z=0)=1)}}$$

In other words, in the numerator we have the odds that $Y=1$ if everyone was given treatment, and in the denominator we have the odds that $Y=1$ if everyone was given the control.

These are simple marginal structural models, but we can also these models with effect modifiers:

$$E[Y(Z)|V] = \psi_0 + \psi_1 z + \psi_3 V + \psi_4 zV$$

In this case,

$$E[Y(Z=1)|V] - E[Y(Z=0)|V] = \psi_1 + \psi_4 V$$

Where V is a variable that modifies the effect of z.

We can also have a general marginal structural model, much like a generalized linear model in statistics:

$$g(E[Y(Z)|V]) = h(z, V; \psi)$$

Where $g()$ is a link function and $h()$ is a function specifying the parametric form of z and V (typically additive and linear). Note again, we are modeling potential outcomes rather than observed outcomes.

4 IPTW estimation

First, recall how we estimate the coefficient, β in a regression model:

$$Y = X\beta + \epsilon$$

We estimate β by solving for the value of β that minimizes the sum of squared deviations between $X\beta$ and Y.

Marginal structural models look a lot like generalized linear models:

$$E[Y(Z)_i] = g^{-1}(\psi_0 + \psi_1 z)$$

This model is not equivalent to the regression model

$$E[Y_i|Z_i] = g^{-1}(\psi_0 + \psi_1 Z_i)$$

because of confounding. In the latter model, we are conditioning or restricting to the subpopulation of Z_i whereas in the former model we are setting Z to be whatever we want. However, again we can use a pseudo-population that's free from confounding (assuming ignorability and positivity) even without a randomized trial.

Here are our steps of how to do this:

1. Estimate propensity score
2. Create weights
   (a) 1 divided by propensity score for treated subjects
   (b) 1 divided by 1 minus propensity score for control subjects
3. Specify MSM of interest
4. Use software to fit a weighted generalized linear model
5. Use asymptotic (sandwich) variance estimator (or bootstrapping), which accounts for fact that pseudo-population could be larger than the sample size

To assess balance of the covariates after weighting, we can calculate the standardized mean differences (smd) defined earlier for each covariate. If things are still not balanced, we can still fiddle with our propensity score model since we still haven't looked at the outcomes, as this is still the design stage of analysis.

5 Weights

Let's consider weights a bit more, since they are the basis that we create our pseudopopulations.

Let's note some problems with large weights. First, we note that larger weights will lead to noisier estimate of causal effects. For example, if 1 person has a weight of 10,000, then they effectively represent 10,000 people. Whether or not the outcome event occurred for them will create larger differences in our estimate of some parameter. Thus, the standard error would be large. Second, large weights nearly violate the positivity assumption, since if someone has a large weight, it was highly unlikely for them to receive either the treatment or control.

What can we do if some individuals have large weights? First, we should check why they have large weights. Perhaps certain combinations of the covariates are uncommon, or some covariate variables are very extreme, i.e. an outlier (which could be a data error). It could also be that the weight isn't wrong.

We can also try trimming the tails of the propensity score distribution. A common trimming strategy is to remove treated individuals whose propensity scores (probability of being treated) are above the 98th percentile from the distribution of controls, and remove the control individuals whose propensity scores are below the 2nd percentile from the distribution of treated subjects. Both types of individuals were very unlikely to receive the other condition. However, we should note that trimming will change our population.

We can also just truncate the weight i.e., set a maximum weight. For example if someone has a weight of 10,000 we can just set their weight as 100. There is a bias-variance tradeoff here, as if we truncate, we introduce bias, but smaller variance, and vice-versa: if we don't truncate we don't have bias, but we have larger variance.

6 Doubly robust estimators

Recall estimation of $E[Y(Z = 1)]$ using IPTW:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{\pi_i(X_i)}$$

Where π_i is the propensity score and Z_i is an indicator variable of if individual i receives treatment. If the propensity score is correctly specified, the estimator is unbiased.

Alternatively, we could estimate $E[Y(Z = 1)]$ by specifying an outcome model $m_1(X) = E[Y|Z = 1, X]$ and then averaging over the distribution of X:

$$\frac{1}{n}\sum_{i=1}^{n} Z_i Y_i + (1 - Z_i)m_1(X_i)$$

Which is essentially averaging over all individuals: if $Z = 1$ for an individual, use their observed Y and for other subjects, use the *predicted* Y given X_i if Z had been 1.

A *doubly robust estimator* is an estimator that is unbiased if either the propensiity score model *or* the outcome regression model are correctly specified. Here is an example:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{\pi_i(X_i)} - \frac{Z_i - \pi_i(X_i)}{\pi_i(X_i)}m_1(X_i)$$

The first term resembles the IPTW estimator while the second term is referred to as an augmentation. The idea behind doubly robust estimators is that because we are using models, there's a chance we could get it wrong, but it would be nice to have an estimator where even if we get one of them wrong, we would still be alright.

Let's say that the propensity score is correctly specified but the outcome model is wrong. The second term has expectation 0 because the numerator of the second term will equal 0 in expectation since the expected value of Z_i, $E[Z_i]$ should equal the propensity score $\pi_i(X_i)$. What if it's the other way around? Well if we rearrange the terms:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i(Y_i - m_1(X_i))}{\pi_i(X_i)} + m_1(X_i)$$

We see that the numerator of the first term will go to 0 in expectation, leaving only the second term which is our model outcome.

References

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