PSC 504: Weighting and post-treatment bias

Matthew Blackwell

2/21/2013

Post-treatment bias

- You may have noticed from most of the readings that conditioning on post-treatment variables has been *verboten* when attempting to identify treatment effects. There are intuitive reasons why this is the case—part of the effect of A on Y might "flow through" a post-treatment variable Z—but these intuitions can be very misleading.

- One intuition that is correct is that conditioning on a post-treatment variable implicitly attempts to estimate a treatment effect among a group who post-treatment variable is unaffected by the treatment. Obviously, this group might be very different than the overall population and we cannot in general average this problem away.

- To get a more precise view on the post-treatment bias problem, we'll follow the approach of Rosenbaum (1984), which is one of the canonical treatments (!) of post-treatment bias.

Decomposition of post-treatment bias

- Let $(Z_i(1), Z_i(0))$ be the potential outcomes for the post-treatment variable, with $Z_i = A_iZ_i(1) + (1 - A_i)Z_i(0)$. If we were interested in the effect of a campaign strategy on the campaign outcome, this might be a poll taken during the middle of the campaign. These potential outcomes are the outcomes that would occur if the candidate went negative ($A_i = 1$) or stayed positive ($A_i = 0$).

- Again, let's assume ignorability conditional on the covariates: $(Y(1), Y(0)) \perp\!\!\!\perp |X$. Thus, in this case the usual estimator for the conditional average treatment effect is unbiased (N_x is the number of units with $X_i = x$):

$$\hat{\tau}(x) = \frac{1}{N_x} \sum_{\{i:X_i=x\}} E[Y_i|A_i = 1, X_i = x] - E[Y_i|A_i = 0, X_i = x]$$

- And we can combine these estimates to get an estimate of the average treatment effect: $\hat{\tau} = E[\hat{\tau}(X)]$.

- Instead of using those estimators, let's see what happens when we control for the post-treatment variable:

$$\begin{aligned}\Delta(x,z) &= E[Y|A=1,Z=z,X=x]-E[Y|A=0,Z=z,X=x] \\ &= E[Y(1)|A=1,Z=z,X=x]-E[Y(0)|A=0,Z=z,X=x] && \text{(Consistency for } Y\text{)} \\ &= E[Y(1)|A=1,Z(1)=z,X=x]-E[Y(0)|A=0,Z(0)=z,X=x] && \text{(Consistency for } Z\text{)}\end{aligned}$$

- Then, we can average these over the distribution of (X, Z): $\Delta = E[\Delta(X, Z)]$. Now, we want to compare this estimator Δ to the average treatment effect τ. One way to investigate the bias of the estimator is to look at a different quantity, what Rosenbaum (1984) calls the net treatment difference $\nu(x, z)$:

$$\nu(x,z) = E[Y(1)|Z(1)=z,X=x]-E[Y(0)|Z(0)=z,X=x]$$

- Again, we'll take the average over (X, Z): $\nu = E[\nu(X, Z)]$. What is this quantity? If $\nu(x, z) = 0$ and $\tau > 0$, then the effect of A on Y flows entirely through Z. This is because there is no effect of A when we fix the value of Z to z under $A = 1$ or $A = 0$. This is simiar to estimating the effect of A when we remove the arrow from A to Z:

$$\begin{tikzcd} Z \arrow[dr] & \\ A \arrow[r] & Y \end{tikzcd}$$

- In the causal mechanisms literature, this is called the **controlled direct effect**, controlled because we fix the post-treatment variable at a given value, rather than let it take the value it would take under treatment or control.

- The bias of Δ can be written: $\Delta - \tau = (\Delta - \nu) + (\nu - \tau)$. You can think of this as highlighting the parts of the unbiased proof that are incomplete. In order to prove unbiasedness we need to show that $\Delta = \nu$ and then that $\nu = \tau$, but our current set of assumptions don't allow us to do that. Why?

- $(\Delta - \nu)$ measures our inability to estimate the controlled direct effect. It might be the case that $\Delta(x, z) \neq \nu(x, z)$ because Z is a collider. If we condition on Z, it opens a backdoor path between A and Y:

$$\begin{tikzcd} U \arrow[r] \arrow[d] & Z \\ A \arrow[r] \arrow[ur] & Y \arrow[u] \end{tikzcd}$$

- In this case, conditioning on Z opens the backdoor path from $A \leftarrow U \rightarrow Z \leftarrow Y$. Thus, $(\Delta - \nu)$ represents the bias due to unmeasured confounding between A and Z.

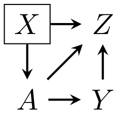
- The second term, $(\nu - \tau)$, represents the difference between the controlled direct effect and the average treatment effect. Somewhat obviously, this will be non-zero if A affects Z and Z affects Y. That is, τ is overall effect of A on Y, which includes the path from $A \rightarrow Z \rightarrow Y$. If we just look at the direct effect, this will ignore the indirect effect and will only equal the overall effect if that indirect effect is 0.

- Note about indirect/direct effects: while we discussing the controlled direct effect somewhat loosely here, you should be **very**, **very** cautious when attempting to make inferences about direct and indirect effects. Things become complicated quickly. We'll talk more about this when we get to causal mechanisms.

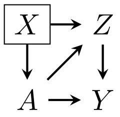
Conditions that eliminate post-treatment bias

- We know that the bias comes from $(\Delta - \nu)$ and $(\nu - \tau)$, so we can characterize assumptions that would make these two zero and give us unbiased estimates of τ.

- **Strong ingorability for post-treatment variable**: $(Y(0), Z(0), Y(1), Z(0)) \perp\!\!\!\perp A|X$. This extends ignorability to the post-treatment variable and gives us $\Delta = \nu$. Why? Because it eliminates the possibility that Z's role as a collider opens the backdoor path between A and Y. This is because $(Y(0), Z(0), Y(1), Z(0)) \perp\!\!\!\perp A|X$ implies $Y(0) \perp\!\!\!\perp A|X, Z(0)$, so that $E[Y(0)|A = 0, Z(0) = z, X = x] = E[Y(0)|Z(0) = z, X = x]$.



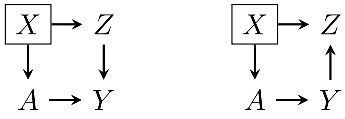
- This DAG shows the relationships when this condition holds. Conditional on X, there are no backdoor paths between A and Y or Z. This means that once we condition on X, it doesn't matter whether or not we condition on Z. The path from $A \leftarrow X \rightarrow Z \rightarrow Y$ is blocked by controlling for X, whether or not we condition on Z. But this isn't enough to prove $\Delta = \tau$ because we could have a situation which reverses the arrow between Z and Y:



- **No effect of treatment on the post-treatment variable**: $Z(1) = Z(0) = Z$ for all units. Under this condition, we have $\nu = \tau$ because the effect of A cannot go through Z since it doesn't affect Z. Technically,

$$\nu(x, z) = E[Y(1)|Z(1) = z, X = x] - E[Y(0)|Z(0) = z, X = x] = E[Y(1) - Y(0)|Z = z, X = x].$$

- So that when we take the average over (X, Z), we get $\nu = \tau$. In this case the above DAGs would be:



- You can see that once we apply these two assumptions, the decision to condition or not condition on Z no longer matters. In essence, we are assuming Z back to a pre-treatment variable.

- But, what does this buy us? Nothing, really, because $(Y(0), Z(0), Y(1), Z(0)) \perp\!\!\!\perp A|X$ implies $(Y(0), Y(1)) \perp\!\!\!\perp A|X$, so that we can estimate τ using the standard estimator $\hat{\tau}$. Controlling for Z does nothing.

Propensity score weighting

- Remember again that there were three approaches to estimating causal effects under ignorability: matching, weighting, and regression. This section covers the weighting approach which is closely linked to subclassification and matching on the propensity score.

- The intuition behind the weighting approach comes from a sampling mindset: each of the treated and control samples are unrepresentative of the overall population, which leads to imbalance in the covariates and the confounding. What do we usually do with unrepresentative samples? Reweight them to be more representative.

- How should we reweight the data from an observational study, though? It's not immediately clear, but we can get some insight from the distribution of the data:

$$p(Y, A, X) = p(Y|A, X)p(A|X)p(X)$$

- We know that a simple difference in means will not work here because there is dependence between A and X, captured in the term $p(A|X)$. If only A and X were independent, then we could just take a simple difference in means.

- The above factorization hints at a weighting scheme: if we were to reweight the data by $W_i = 1/p(A_i|X_i)$, then we would break the relationship between A and X. Note that this is the inverse of the probability of receiving the observed treatment conditional on the observed covariates. Thus, the denominator is simply the probability of seeing a unit like this, conditional on its covariates.

- Let's look at a situation with one binary covariate. There we would have four weights: $W_{ax} = 1/\Pr[A = a|X = x]$ for each of the four combinations of treatment and covariate. Thus, for the group observed to have $(A, X) = (1, 1)$.

$$\begin{aligned}
\Pr_W[A = 1|X = 1] &= \frac{W_{11} \cdot \Pr[A = 1|X = 1]}{\omega} \\
&= \frac{\frac{1}{\Pr[A=1|X=1]} \cdot \Pr[A = 1|X = 1]}{\frac{1}{\Pr[A=1|X=1]} \cdot \Pr[A = 1|X = 1] + \frac{1}{\Pr[A=0|X=1]} \cdot \Pr[A = 0|X = 1]} \\
&= \frac{1}{2}.
\end{aligned}$$

- The first equality simply relates the conditional distribution in the

weighted data relates to the conditional distribution in the original data, where W_{11} is the weight associated with the observations that follow $(A, X) = (1, 1)$ and ω is a normalizing factor to make the probabilities sum to one. The second equality simply fills in the definition of the weights: one over the probability of observing $A = 1$ conditional on $X = 1$. It also uses the fact that $\omega = E[W]$. If we do the same analysis for $X = 0$, we get end up with the same distribution:

$$\Pr_W[A=1|X=0] = \frac{W_{10} \cdot \Pr[A=1|X=0]}{\omega} \tag{1}$$

$$= \frac{\frac{1}{\Pr[A=1|X=0]} \cdot \Pr[A=1|X=0]}{\frac{1}{\Pr[A=1|X=0]} \cdot \Pr[A=1|X=0] + \frac{1}{\Pr[A=0|X=0]} \cdot \Pr[A=0|X=0]} \tag{2}$$

$$= \frac{1}{2}. \tag{3}$$

- Thus, we have shown that in the reweighted data, $p(A|X=1) = p(A|X=0) = p(A)$. That is, in the reweighted data, the treatment and the covariate are independent.

- We can show that the weighted mean of the treated units is the same as the mean of the potential outcome under treatment:

$$\begin{aligned}
E\left[\frac{AY}{e(X)}\right] &= E\left[\frac{AY(1)}{e(X)}\right] && \text{(Consistency)} \\
&= E\left[E\left[\frac{AY(1)}{e(X)}\Big|X\right]\right] && \text{(Law of Iterated Expectations)} \\
&= E\left[\frac{E[A|X]E[Y(1)|X]}{e(X)}\right] && \text{(Conditional Ignorability)} \\
&= E\left[\frac{e(X)E[Y(1)|X]}{e(X)}\right] && \text{(Propensity Score Definition)} \\
&= E[Y(1)] && \text{(Law of iterated Expectations)}
\end{aligned}$$

- The same logic would give us the mean potential outcomes under control:

$$E\left[\frac{(1-A)Y}{1-e(X)}\right] = E[Y(0)]$$

- These two facts provide an estimator for the average treatment effect:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{A_iY_i}{e(X_i)} - \frac{(1-A_i)Y_i}{1-e(X_i)}\right)$$

- The above two results give us that this esimator is unbiased.

Estimation of the propensity score

- Obviously, in order to estimate $\hat{\tau}$ we need to either know or estimate the propensity score, $e(X_i)$. How do we do that? Well, we already have done that to either subclassify on the propensity score or match on the propensity score. In some sense, it's old hat. We can simply perform a logistic regression of A on X. If X is discrete with a small number of categories, then we can even estimate the within-strata propensity scores (N_{xt}/N_x) and this will be a good, non-parametric estimate of the propensity score in each stratum of the data.

- But what happens when X has many categories? Then, for some x, there will only be treatment or control units and $\hat{e}(x) = 0$ or $\hat{e}(x) = 1$. What does this mean for our weights? Well, that means that $\frac{1}{\hat{e}(x)} = \frac{1}{0} = \infty$. When there is some value of x with no overlap between treated and control, we have a violation of positivity (which, remember, is $0 < p(A = 1|X) < 1$). There are two types of positivity violations: structural and random. Structural positivity violations are due to logical requirements: if we wanted to know the effect of getting tenure on your lifetime happiness, it's clear that people not in academia would be unable to receive the treatment because they cannot get tenure. No matter how many observations we have, we'll never see a non-academic with tenure. We can't get around this: we can only make inferences on susbets of the data where structural positivity holds.

- Random positivity violations, on the other hand, are due to the fact that we have a finite sample size and so we might, by chance, get a sample that has no treated or no control in some stratum of the data. In these cases, our non-parametric estimates of the propensity score give us infinite weights, as we saw. But we can also use models to borrow information from "nearby" strata to estimate the propensity score. This essentially entails using a model to estimate the propensity score, perhaps a logistic regression or a more complicated method, such as boosted regression, kernel estimation, or a series logit estimator.

- Entropy balancing: Hainmueller (2012) develops an approach that produces weights that balance the data directly. He observes that the covariates should be balanced in the reweighted data. Instead of modeling the propensity score, then, he chooses to find the weights that maximize the balance of the covariates.

Stablized weights

- Even if positivity isn't violated propensity scores that are close to 0 or 1 lead to very large weights, which can lead to larger-than-necessary standard errors. There are a couple of approaches to handle this problem in data analysis.

- One is **trimming the weights**. That is, pick some value ε (say, 0.1) and for any units with estimates $\hat{e}(X_i) < \varepsilon$ or $1 - \hat{e}(X_i) < \varepsilon$, we set their weight to $W_i = \frac{1}{\varepsilon}$. This is obviously equivalent to setting a maximum value of the weights. We can also set a minimum value.

- Another way to solve the extreme weight problem is to use **stabilized weights** instead of the usual weights. Note that in our proofs above, our weights were $1/\Pr[A|X]$. It turns out that we can replace the 1 with a function of A and still get the balancing property. This gives us the ability to reduce the variance of our estimates by using stablized weights instead of the usual weights: $SW_{ax} = \Pr[A = a]/\Pr[A = a|X = x]$.

Boostrapping to get the SEs

- In order to incorporate the uncertainty to due the weighting model, we need to bootstrap the whole estimation procedure. That is, repeat the following steps some large (roughly 1,000) times:

  1. Draw a sample of the data with replacement, call this, S_b.
  2. Estimate the propensity scores in this sample, $\hat{e}_b$ and create weights, W_b.
  3. Use the weights to get an estimate of the average treatment effect, τ_b in the sample S_b.
  4. Repeat.

- The distribution of the estimates, $\hat{\tau}_b$, will give us the bootstrapped standard errors and confidence intervals.