PSC 504: Dynamic Causal Inference

Matthew Blackwell

4/18/2013

The problem

- Let's go back to a problem that we faced earlier, which is how to estimate causal effects with treatments that vary over time. We could have a panel data situation, where we repeated measurements of the outcome, or it could be a situation where we see a back and forth between the covariates and the treatment and observed only one outcome.
- An example of the latter is negative advertising in campaigns: a campaign decides to go negative based on past polling, but that negativity may have an impact on future polling. Of course, the outcome is the final vote share and that is only observed at the end of the campaign.
- What do we have to sort out here? One is what types of effects we are interested in. Earlier we saw that panel data regression models can only identify the contemporaneous effect of the treatment and cannot estimate the effect of treatment history. We'll see how to estimate these treatment history effects today.

Notation

- Let's let $\underline{A}_{it} = (A_{i1}, \ldots, A_{it})$ be the partial history of treatment up to time t. Define $\underline{X}_{it}$ similarly. Sometimes we'll need to reference a specific instance of these variables and we'll use $\underline{a}_t$ and $\underline{x}_t$ to do so. We drop the t subscript to refer to the entire sequence: $\underline{A}_i = (A_{i1}, \ldots, A_{iT})$.
- Of course, we'll need potential outcomes and these could be functions of the entire treatment history: $Y_(\underline{a})$. Note that this notation implies that the regimes are **static**, so that decision to go negative at time t is fixed by the treatment history and does not change based on covariates that are evolving over time.
- We can refer more generally to **treatment regimes** which are like strategies in game theory: they are rules that dictate what actions/treatments units should take given a certain covariate history. We define these as $g(\underline{x}_t)$, which takes values $\underline{a}_t$. Clearly the "static" treatment histories above are treatment regimes that are fixed across covariates.
- Obviously we might have different outcomes if we follow different treatment regimes, even if those regimes are observationally equivalent for some people. Thus, we want to write potential outcomes for the regimes: $Y_i(g)$.
- Treatment regimes are complicated and generally fairly hard to deal with, but the temptation is fairly obvious: this are the kinds of effects people want to know about.
- In order to connect the potential outcomes and the observed data:

$$Y_i = Y_i(g) \quad \text{if } \underline{A}_i = g(\underline{X}_i).$$

- This says that if a unit's observed history is equal to the perscription of the treatment regime, then the observed outcome equals the potential outcome under that regime.

g-computation

- We would like to estimate the effects of these regimes. Something like the follwing:

$$\tau(g, g') = E[Y(g) - Y(g')]$$

- In medical studies, the goal is often to estimate the "optimal" regime, which is the following:

$$\arg\max_g E[Y(g)],$$

- Here, we are trying to find the regime that maximizes the outcome (assuming the outcome is beneficial).
- Either of these estimands requires us to estimate the mean of the potential outcome under a given regime. How do we do that?
- In general, we are going to have to make an ignorability assumption, which we will call **sequential ignorability**:

$$Y(g) \perp\!\!\!\perp A_t | \underline{X}_t = \underline{x}_t, \underline{A}_{t-1} = g(\underline{x}_{t-1})$$

- This assumption says that the potential outcome under some regime is independent of the treatment at time t, conditional on the past values of the covariates and the treamtent regime. Again, this is simiar to running a sequential experiment, where the randomization can depend on the past.
- We also have to assume positivity, which says that there are no deterministic treatments. That is, if $\Pr[\underline{A}_{t-1} = \underline{a}_{t-1}, \underline{X}_t = \underline{x}_t] > 0$, then

$$\Pr[A_t = a_t | \underline{X}_t = \underline{X}_t, \underline{A}_{t-1} = \underline{a}_{t-1}] > 0$$

- But how do we calculate the marginal mean of the potential outcomes? The same way we have before, by marginalizing over the distribution of the covariates. Here it is more tricky because they vary over time.
- Jamie Robins came up with what he calls the **g-computational formula** for these types of marginal means. In general, it looks like this:

$$E[Y(g)] = \int_{x_t} \cdots \int_{x_0} E[Y|\underline{X} = \underline{x}, \underline{A} = g(\underline{x})] \prod_{j=0}^{T} \left\{ f(x_j | \underline{X}_{j-1} = \underline{x}_{j-1}, \underline{A}_{j-1} = g(\underline{x}_{j-1})) d\mu(x_j) \right\}$$

- Notice that the right hand side here only has observeable quantities. To get this, we had to invoke sequential ignorability. To see how this works, let's work with a simpler example: two time periods, with a binary covariate covariate between the two treatments, x.
- First note that, under consistency, we know that $E[Y(g)|\underline{A} = g(x)] = E[Y|\underline{A} = g(x)]$.

$$
\begin{aligned}
E[Y(g)] =& E[Y(g)|A_1 = g_1(1)] \\
=& E[Y(g)|X = 1, A_1 = g_1(1)] \Pr[X = 1|A_1 = g_1(1)] \\
&+ E[Y(g)|X = 0, A_1 = g_1(0)] \Pr[X = 0|A_1 = g_1(0)] \\
=& E[Y(g)|X = 1, A_2 = g_2(1), A_1 = g_1(1)] \Pr[X = 1|A_1 = g_1(1)] \\
&+ E[Y(g)|X = 0, A_2 = g_2(0), A_1 = g_1(0)] \Pr[X = 0|A_1 = g_1(0)] \\
=& E[Y|X = 1, A_2 = g_2(1), A_1 = g_1(1)] \Pr[X = 1|A_1 = g_1(1)] \\
&+ E[Y|X = 0, A_2 = g_2(0), A_1 = g_1(0)] \Pr[X = 0|A_1 = g_1(0)]
\end{aligned}
$$

- The first equal sign comes from sequential ignorability (the first period is randomized), the second from the law of iterated expectations, the third from sequential ignorability (conditional on first treatment and the covariate, the second treatment is random), and the last from consistency.
- Note that we can't just collapse the last line with the law of iterated expectations because the conditioning set for the outcome and the covariates are different. In the cross sectinal case this is true as well, but typically we average across the marginal distribution of X_i. This is relatively easy because we can usually just use the empirical distribution of the data. Here, though, we need the distribution of the X_{it} conditional on the past, which means we will almost certainly need a model for the relationship between the covariates and the past in addition to the model for the outcome. Ugh. Lots of modeling.
- Obviously, if the covariates are continuous, we are going to have to replace the sum over the distribution of the covariates with an integral, which is what we see in the g-computational formula.
- One approach is to write down models for the covariates and the outcome, then construct a likelihood and estimate the parameters of that likelihood and plug them into the g-formula. Robins has some work that shows that this is problematic for most situations because there may be no set of parameters for which if they are all zero, there is no effect. That is, it might be difficult to test the null hypothesis of no effect of the treatment regime. It might be possible to estimate effects in this situation, but this is relatively open question. There are ways to reparameterize the distribution of the data to make progress and these models are usually called "Structural Nested" models.

Marginal structural models

- With g-computation, we had to write down models for the covariates in addition to the outcomes, which is a large pain. Ideally, we would just be able to run a regression and read off the coefficients in a causal way. This is what marginal structural models are and we can use a weighting approach to estimate their parameters.
- A marginal structural model (MSM) is a model for the marginal mean of the potential outcome for a given treatment history (we're going to ignore treatment regimes for now). That is, it's a model of the form:

$$E[Y(\underline{a})] = h(\underline{a}; \beta)$$

- Here h is a link function and β are a set of parameters. There are a lot of modeling choices we might make here. With a binary treatment variable, there are 2^T possible treatment histories.

- Let's say we randomly assigned treatment histories. With the single-shot case, we could always non-parametrically estimate $E[Y(1)]$ and $E[Y(0)]$ using simple means. Here, though, we are unlikely to observe anyone following any particular history. Thus, simple means like these are not going to work. We are going to need a model for the marginal potential outcome means.

- How should we model this? It could be that the number of treated periods is all that matters:

$$E[Y(\underline{a})] = \beta_0 + \beta_1 \sum_t = 1^T a_t$$

- Or it could be that the effect varies over time:

$$E[Y(\underline{a})] = \beta_0 + \beta_1 \sum_t = 1^{T/2} a_t + \beta_2 \sum_{t=T/2+1}^{T} a_t$$

- In any case, we have to write a model down for the outcome here in terms of the treatment history.

IPTW

- But how do we actually estimate these models? Can we use regression or matching to estimate the parameters? Unfortunately not. Imagine we estimate the following model:

$$E[Y|\underline{A}, \underline{X}]$$

- This model conditions on the these variables that are changing over time (we call them time-varying confounders). Imagine we are in the two-period case and we estimate this:

$$E[Y|A_1, A_2, X] = \alpha_0 + \alpha_1 A_1 + \alpha_2 A_2 + \alpha_3 X$$

- Here we conditioned on the covariate to remove omitted variable bias, but we have actually introduced post-treatment bias. If the effect of negativity early in the race flows through polls and we condition on polls, this is going to underestimate the effect of earlier negativity. From this point of view maybe we omit polls and estimate this model:

$$E[Y|A_1, A_2] = \alpha_0 + \alpha_1 A_1 + \alpha_2 A_2$$

- Do these parameters equal the causal parameters β? No, because now there is confounding between polling and going negative later in the race. Thus, polling is simultaneously post- and pre-treatment. Controlling for it induces post-treatment bias and omitting induces confounding bias. What can we do?

- We can instead use a weighting approach. Remember that regression and matching looked within subsets of the covariates to find balance conditional on the covariates. Weighting instead reweighted the data to remove those imbalances.

- How do we weight? Well, we know that with a single-shot treatment, we can weight by the inverse of the propensity score:

$$W_i = \frac{A_i}{\Pr[A_i = 1|X_i]} + \frac{1 - A_i}{\Pr[A_i = 1|X_i]}$$

- Here, the propensity scores are more complicated because the treatment is more complicated. We have to weight by the probability of observing the entire history: $\Pr[\underline{A}|\underline{X}]$:

$$W_{it} = \frac{1}{\Pr(A_{it}|\underline{A}_{it-1}, \underline{X}_{it})}.$$

- Thus, in this case, we need to each unit by the probability of receiving the treatmeht history they did, conditional on the past. Let's look at this in the two-period example. Let's say we see a campaign that is positive in the first period, then they are trailing, then they negative later in the race. The weights we would calculate would be:

$$W_i = \frac{1}{\Pr(\text{pos}_1)} \cdot \frac{1}{\Pr(\text{neg}_2|\text{trail}, \text{pos}_1)}.$$

- Why does the weighting work? It balances the distribution of the data so that, in the reweighted data, any arrows pointing from $\underline{X}$ to the treatments are removed. Thus, in the reweighted data, there is no confounding, no backdoor paths, and no need to control for the covariates anymore.

- To see this, we can see how the weights affect the joint distribution of the observed data:

$$\begin{aligned} f_W(Y, A, X) &= \frac{W(A, X) \cdot f(Y, A, X)}{\omega} \\ &= f(Y|A, X) f(X) \frac{W(A, X) \cdot f(A|X)}{\omega} \\ &= f(Y|A) f(X) \cdot f_W(A) \end{aligned}$$

- Here, ω is a normalizing constant. Basically, the weights balance the distribution of the the treatment with respect to the covariates, but don't change the relationship between the treatment and the outcome. This is exactly what we want.