PSC 504 - Matching

Matthew Blackwell

2/14/2013

Why match?

- Let's say that we know that ignorability holds conditional on X_i. We know that we have to "control for" X_i in some way, but what is the best way to do this? There are three broad approaches that overlap in parts and have much in common, but also have fundamental differences. They are matching, weighting, and regression. We'll talk about each of these in the coming weeks.
- Matching has a number of nice properties that have made it appealing the last few years. The most important is that, under ignorability, if we are able to find a matching solution with good balance on the covariates, then no further modeling of the covariates is necessary. We get to side-step the rather strong assumptions of linear relationships between X_i and Y_i that are required by regression.
- Remember that matching doesn't justify a causal effect, ignorability does. Matching doesn't make ignorability more plausible, it simply represents a non-parametric way of estimating causal effects under ignorability. As Sekhon says:

  > Without an experiment, natural experiment, a discontinuity, or some other strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive.

- Matching's appeal is twofold: first, it can greatly simplify the estimation of certain causal parameters and second, it can reduce the dependence of such estimates on parametric models.

Causal Estimates

- Ignorability + Balance = Ignorability conditional on the match.
- As we did last week, we are going to always assume that ignorability holds: $Y_i(a) \perp\!\!\!\perp A_i | X_i$ for all values a. Again, this is just the selection on the observables assumption. We'll also assume overlap: $0 < \Pr[A_i = 1|X_i] < 1$.
- ATT is identified using exact matching without making assumptions about the relationship between X_i and Y_i. After matching $E[Y_i|A_i = 0] = E[E[Y_i|A_i = 0, X_i = x]]$ which means we can just use the difference in means.

- Let's say that for each treated unit we can find an **exact match**: a control unit with the same values of X_i and suppose we drop any control units that are not matched. What does this imply? Well, for one, we know that the distriubtion of X_i will be the same across the treated and control groups $\Pr(X_i = x|A_i = 1) = \Pr(X_i = x|A_i = 0)$ for all values of x. This is because in the matched data, for every treated unit, there is one (and, in this case, only one) control unit with the same exact value of X_i. The two groups must have the same distribution in X_i. Let's show that the ATT is identified if the data is exactly matched:

$$
\begin{aligned}
\tau_{\text{ATT}} &= E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 1] \\
&= E[Y_i|A_i = 1] - \sum_{x \in \mathcal{X}} E[Y_i(0)|X_i = x, A_i = 1] \Pr(X_i|A_i = 1) && \text{(Consistency \& Interated Expectations)} \\
&= E[Y_i|A_i = 1] - \sum_{x \in \mathcal{X}} E[Y_i(0)|X_i = x, A_i = 0] \Pr(X_i|A_i = 1) && \text{(Ignorability)} \\
&= E[Y_i|A_i = 1] - \sum_{x \in \mathcal{X}} E[Y_i|X_i = x, A_i = 0] \Pr(X_i|A_i = 1) && \text{(Consistency)} \\
&= E[Y_i|A_i = 1] - \sum_{x \in \mathcal{X}} E[Y_i|X_i = x, A_i = 0] \Pr(X_i|A_i = 0) && \text{(Exactly Matched Data)} \\
&= E[Y_i|A_i = 1] - E[Y_i|A_i = 0] && \text{(Iterated Expectations)}
\end{aligned}
$$

- As you can see, with ignorability, consisitency, and exact matches, we can identify the ATT. In fact, we don't even need full ignorability here, but rather a weaker condition: $E[Y_i(0)|X_i, A_i = 1] = E[Y_i(0)|X_i, A_i = 0]$. There are two features of this weaker assumption: one is that we only have to make assumptions about the potential outcome under control, not the potential outcome under treatment. Second, we only have to assume condition mean independence (sometimes called CMI), not full independence (which would include higher moments).

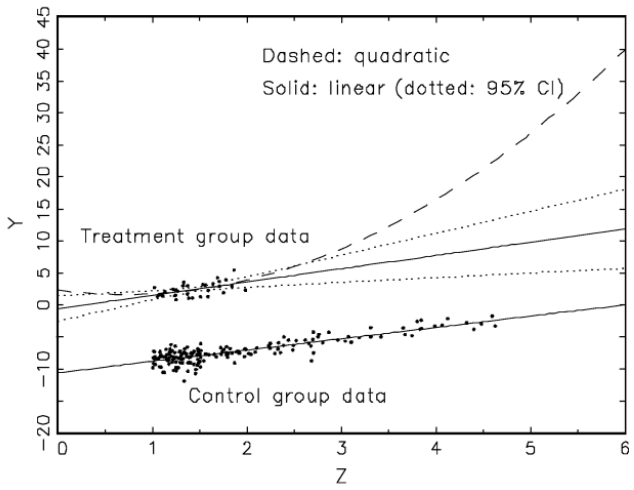
- Obviously the nice part about this analysis is that in the matched dataset, all we need is a simple difference in means. That is, we can ignore X_i. We just take the mean of Y_i among the control units. If there are different numbers of matches for each treated unit, then we need to take the weighted mean since $\Pr[X_i|A_i = 1] = \frac{1}{M} \Pr[X_i|A_i = 0]$, where M is the number of matched controls.

- One way to think of this approach is that we are "imputing" the missing values $Y_i(0)$ for the treated units, using control units with very similar values of X_i. In this sense, matching is very similar to some approaches to missing data (namely, hot-deck imputation).

- Here we used exact matching, but that is not necessary. To justify matching, all we need is balance, conditional on the matching solution. A **matching solution** is a subset of the data produced by the matching procedure. Let's call that $\mathcal{S}$. Note that $\mathcal{S}$ is function of the covariates X_i, so that ignorability on the covariates implies ignorability on the covariates **and** the matching soluation: $Y_i(a) \perp\!\!\!\perp A_i|X_i, \mathcal{S}$. Now, if we can achieve balance through that matching solution, then we should have the distribution of X_i and A_i be independent, conditional on that solution: $A_i \perp\!\!\!\perp X_i|\mathcal{S}$. This is obviously a checkable condition: we can assess balance under any matched data set. These two properties, combined with Lemmas 4.2 and 4.3 of Dawid (1979) imply that there is ignorability conditional on the just the matching solution: $Y_i(a) \perp\!\!\!\perp A_i|\mathcal{S}$. Thus, if we can ensure balance in the matched dataset, then we will be able to identify causal parameters.

Model dependence

- The exact model we use will be less relevant because (a) we won't be extrapolating to regions of the data where there is no overlap and (b) the lack of assumptions above make the effects identified without any modeling assumptions.



- Without matching, the model we choose for the relationship between X_i and Y_i will affect our estimates of the the relationship between A_i and Y_i. If we match, though, we have a dataset with good balance so that A_i and X_i are approximately independent. Therefore, in regressions or other models, the coefficient on A_i will be less affected by the inclusion or exclusion of X_i or functions of X_i.

The matching procedure

1. Choose a number of matches (1 control:1 treated, 2:1, k:1, etc), whether to match with replacement or not, and a distance metric (propensity scores, Mahalanobis distance).
2. Choose a set of pre-treatment covariates that satify ignorability.
3. Find matches (nearest neighbor, GenMatch, optimal matching), dropping control units that are not matched.
4. Check balance (difference-in-means, medians, eQQ, etc)
5. Repeat (1)-(4) until balance is acceptable, adding variables or functions of variables to improve balance.
6. Calculate the effect of the treatment on the outcome in the matched datasets.

Number of matches

- The number of matches for each control group must strike a balance because small numbers of matches means fewer observations, but more matches means that each match might be a worse match.

- If there are variable numbers of control matches for each treated unit, we need to weight the controls each matched stratum according the number of controls in that stratum.
- Matching with replacement is, in general, a good idea because it allows for better matches, but we may have to use weights to account for the units being in the data twice. In addition, if we match **all** units (match treated to control and control to treated), then we can estimate the ATE in addition to the ATT. Of course, the estimator will be slightly more complicated because we have to calculate the imputed potential outcomes for each unit, control and treated.

Distance metrics

- In order to choose a matching control unit for each treated unit, we need some way of measuring the distance between two units in terms of the covariates, X_i.
- Exact: only match units to other units that have the same exact values of X_i. This obviously works for a small number of discrete variables, but as we either add continuous variables or increase the dimensionality of X, exact matching won't be feasible.

$$D_{ij} = \begin{cases} 0 & \text{if } X_i = X_j \\ \infty & \text{if } X_i \neq X_j \end{cases}$$

- Propensity scores: When there are many covariates, we can match on the propensity score. The justification for this comes from last week when we showed that conditioning on the (true) propensity score is equivalent to conditioning on the entire set of covariates. Of course, we have to estimate the propensity score, so it is no longer necessarily true that matching on the estimated propensity score will increase balance. Rubin and his colleagues have shown that propensity score matching has good properties if the distribution of the covariates is ellipsodally symmetric (such as Normal or t), but if this isn't true then the properties can be quite bad.

$$D_{ij} = |e(X_i) - e(X_j)|$$

- Linear propensity scores: When estimating the propensity scores from a logistic regression it is often better to use the linear propensity score, which is just the linear predictor, $\text{logit}(e(X_i)) = X_i\beta$.

$$D_{ij} = |\text{logit}(e(X_i)) - \text{logit}(e(X_j))|$$

- Mahalanobis distance: the Mahalanobis distance is an alternative to Euclidean distance that takes into account the distribution of the data. This is useful for finding "nearby" control units in a multidimensional space with continuous covariates. The intuition here is that we want to normalize the distance between two points by the standard deviation of each variable. If two units are very far apart on the nominal scale, but the standard deviation is also high, then we might want to count this as "close" compared to two units that are close on a nominal scale with an extremely small SD. But we could achieve this with just Euclidean distance on standardized variables. The Mahalanobis distance takes into account covariances as well. We need a covariance matrix, Σ, to calculate the MD. For the ATT we'll use the covariance matrix of the treated data and for the ATE, we'll use the covariance matrix of the entire data.

$$D_{ij} = \sqrt{(X_i - X_j)'\Sigma^{-1}(X_i - X_j)}$$

- Calipers. Sometimes forcing a match between two units produces poor balance, so we would rather not match treated units to control units that are too far away on the propensity score. The maximum distance in terms of the (linear) propensity score that we would be willing to accept is called the caliper, c. Thus, we would be dropping treated and control units potentially. For those control units within the caliper, we might use Mahalanobis distance to find matches:

$$D_{ij} = \begin{cases} \sqrt{(X_i - X_j)'\Sigma^{-1}(X_i - X_j)} & \text{if } |\text{logit}(e(X_i)) - \text{logit}(e(X_j))| \leq c \\ \infty & \text{if } |\text{logit}(e(X_i)) - \text{logit}(e(X_j))| > c \end{cases}$$

Estimands

- In general, what we showed at the beginning is that the ATT is easy to calculate with 1:1 exact matches on the treated units: we just take a difference in means. We can identify the ATE with 1:1 exact matching as well, but we need full ignorability and we need to keep all control units. The calculation becomes more cumbersome if we have multiple matches for each unit. See Imbens (2004) for a description of ATE estimators.

- "Moving the goalposts": When we keep all of the treated units and only drop control units from the data, we can, in general identify the ATT, but if we begin to drop treated units, then it's unclear what we are estimating. It becomes the treatment effect among the matched units, which may or may not be an interesting group.

- Common support: related to calipers is the notion of common support. If there are areas for which positivity doesn't (empirically) because there are either no controls or no treated units, then we have to choose between extrapolating to that region or calculating the effect for the common support of the data (where there are both treated and control observations). We should definitely **not** extrapolate to portions of the covariate space where it is theoretically impossible for there to be treated or control units. We wouldn't want to compare the effect of voting for presidential candidates for those under age 18, since it is theoretically impossible for them to vote (legally).

Matching methods

- Nearest Neighbor: Using the chosen distance matric, D_{ij}, find a control unit that is closest to each treated unit. Obviously, the order of the matching matters in terms of which units get matched to which other units. This is sometimes called "greedy" matching.

- Optimal matching: Finds the matching solution that minimizes overall distance.

- GenMatch: The key insight of GenMatch is that using MD as a distance metric might fail in certain circumstances. Instead, they attempt to find the balance metric that induces the best balance in the data. They augment MD with a set of variable weights. Then, GenMatch uses a genetic algorithm to find the set of variable weights that produces a match that maximizes balance in the data. A genetic algorithm is needed because the optimization problem is irregular.

- CEM: An alternative method is akin to stratification from last week. Suppose we have a set of continuous covariates. Obviously, we cannot use exact matching, but if we can find a stratification/coarsening of the data that produces good balance, then ignorability will hold within those strata by the arugments

we used last week. Thus, we'll coarsen the data (say, splitting years of education into less than H.S., H.S. degree, some college, B.A./B.S., Advanced degree), then calculate the ATT within each stratum where there are control and treated units, dropping any stratum without both types of units. Thus, we might drop both treated and control units.

- One nice feature of CEM is that it allows you to control the amount of imbalance up front by setting the fineness of the coarsening. Coarser means more imbalance, finer means less imbalance but also fewer matched units.

Assessing balance

- Because all matching methods attempt to minimize balance, the choice of balance metric will determine which matching method performs better.
- Differences-in-means/medians: fairly straightfoward.
- Quantile-quantile plots/KS statistics: The difference in means doesn't tell us about the comparison of distributions between the treated and control groups. That is, we would like to compare the entire density a covariate under control and treatment. We can visually inspect the quantile-quantile plots for a given variable and/or attempt to summarize the difference between two histograms using the eQQ statistics, Kolmogorov-Smirnov tests, and Kullback-Leibler distances. These are useful ways to measure difference between covariate distributions.
- $\mathcal{L}_1$: Ideally, we would want to meaure imbalance by a multivariate histograms. This measure of imbalance tries to replicate that by coarsening the data into a multivariate histogram and calculating the difference in counts within each stratum of that stratification. Obviously, this is closely tied to CEM.