ESTIMATION OF TREATMENT EFFECTS VIA MATCHING

AAEC 5126
INSTRUCTOR: KLAUS MOELTNER

Textbooks: Wooldridge (2010), Ch.21; Greene (2012), Ch.19;
Angrist and Pischke (2010), Ch. 3
R scripts: `mod5s3`

General Approach

The main difference between matching and regression adjustment is that in matching only the missing counterfactuals are estimated ("imputed"). The actually observed outcomes are used directly. The second big difference is that the imputed counterfactuals are based on just a handful of "close matches", and not the entire sub-sample of observations from the opposite treatment group.

In other words, if we have y_{1i}, the outcome for a treated observation, we only need to estimate y_{0i}, the unobserved counterfactual, and vice versa.

Finding nearest neighbors. How do we select the closest matches, aka "nearest neighbors" amongst the opposite treatment group? This is generally based on similarity in $\mathbf{x}$. Specifically, let $\mathbf{x}_i$ be the regressor vector for some individual i. Now consider $\mathbf{x}_l$, which belongs to individual l from the opposite treatment group. Then a measure of closeness or "distance" can be computed as

$$d_{i,l} = \sqrt{(\mathbf{x}_i - \mathbf{x}_l)' \, \mathbf{V}^{-1} \, (\mathbf{x}_i - \mathbf{x}_l)}, \tag{1}$$

where $\mathbf{V}$ is the sample variance-covariance matrix for $\mathbf{x}$ (for the *full* sample), or, alternatively, its diagonal. In the first case, the distance measure in (1) is referred to as *Mahalanobis* metric. We will follow Abadie et al. (2004) and use the diagonal version.

Our goal is now to find M nearest neighbors for observation i, all of which must come from the opposite treatment group. How many do we need? Abadie and Imbens (2004) suggest a small number, such as three or four. This is because the quality of the matches (as measured by $d_{i,l}$ will typically deteriorate as M increases. By convention, M is interpreted as the *minimum number of matched observations*. If there are ties, the actual set of matched observations might be larger than M.

For simplicity, let's refer to i as a treated observation, and l as a control observation, though the following naturally holds for the reverse case. Let's refer to the set of M nearest neighbors for i as

$\mathcal{J}_{i,M}$. That is, the elements l in $\mathcal{J}_{i,M}$ have a value of $d_{i,l}$ that is no larger than the 4^{th} closest distance from i. Thus, if we let the number of elements in $\mathcal{J}_{i,M}$ be denoted as $L_{i,M}$, we have $L_{i,M} \geq M$.

There is another important construct that we need to define upfront. It is the number of times each l was used as a match for some i, weighted, in each case, by the total number of matches for i, i.e. $L_{i,M}$. Denoting this quantity as $\mathcal{K}_l$, we have

$$\mathcal{K}_l = \sum_{i=1}^{n_1} \left(\frac{I\left(l \in \mathcal{J}_{i,M}\right)}{L_{i,M}} \right), \tag{2}$$

where n_1 is the sample size for the treated, and I is an indicator function, which takes the value of 1 if the condition it represents holds, and a value of 0 otherwise.

Enforcing exact matches. One of the main advantages of this "nearest neighbors" approach to matching (as opposed to, for example, matching on propensity scores) is that we can "force" the elements in $\mathcal{J}_{i,M}$ to match *exactly* on one or more variables in $\mathbf{x}$. For example, in matching single family homes, we may force nearest neighbors to have the exact same number of bedrooms and bathrooms as the target property. In our labor application, we might want to only consider matches with the exact same level (number of years) of education.

Such restrictions can be easily incorporated in this framework by multiplying the corresponding diagonal elements in $\mathbf{V}^{-1}$ by a large number, say 1000. See script `mod5s3` for an example of this "distance penalty".

In a spatial context, this penalty could be used to force all matches to come from the same geographic area (say an urban neighborhood, school district, or school attendance zone), as implemented in Abbott and Klaiber (forthcoming). This could be used to (indirectly) control for unobserved neighborhood effects.

Checking for covariate balance (overlap). As discussed in Ho et al. (2007), it is very important to check for overlap in covariates (also referred to as "covariate balance") between the treated observations and the control observations that were actually chosen as matches. We can use again normalized distances or some other measure of "distribution similarity" for the explanatory variables across both groups. If this does not produce satisfactory results, the matching process may have to be repeated, using different matching criteria (e.g. a different distance metric) and/or a different minimum number of matches per treated observation. Diamond and Sekhon (2013) propose an adaptive, computer-directed approach to find the matching process that maximizes overlap.

Deriving the estimators. The counterfactual outcome for observation i can than be computed as the average over all outcomes for the matched controls, i.e.

$$\hat{y_{0i}} = \frac{1}{L_{i,M}} \sum_{l \in \mathcal{J}_{i,M}} y_l, \tag{3}$$

if, as we continue to assume, i is a treated observation. The analogous expression holds if i is a member of the control group (replacing $\hat{y_{0i}}$ with $\hat{y_{1i}}$). Thus, we can generally write

$$\begin{aligned} \hat{y_{0i}} &= \begin{cases} y_{0i} & \text{if} \quad w_i = 0 \\ \frac{1}{L_{i,M}} \sum_{l \in \mathcal{J}_{i,M}} y_l & \text{if} \quad w_i = 1 \end{cases} \\ &\text{and} \\ \hat{y_{1i}} &= \begin{cases} y_{1i} & \text{if} \quad w_i = 1 \\ \frac{1}{L_{i,M}} \sum_{l \in \mathcal{J}_{i,M}} y_l & \text{if} \quad w_i = 0 \end{cases} \end{aligned} \tag{4}$$

The ATE can then be computed as

$$\hat{\tau_{ate}} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{1i} - \hat{y}_{0i}) \tag{5}$$

Similarly, the ATT can then be derived as

$$\hat{\tau_{att}} = \frac{1}{n_1} \sum_{i=1}^{n_1} (y_{1i} - \hat{y}_{0i}) \tag{6}$$

Consistent standard errors for the basic-matching ATE and ATT can be derived using the analytical expressions given in Abadie et al. (2004) and employed in script `mod5_s3`. They are the same expressions as for the regression-corrected estimator, which are given below.

Matching with regression correction

As discussed in Abadie and Imbens (2006) and Abadie and Imbens (2011) the generic matching estimators outlined above exhibit a bias that does not vanish even under large samples due to (highly likely) inexact matching on continuous covariates. As would be expected, the bias increases with the number of (imprecisely matched) continuous elements in $\mathbf{x}$.

Abadie and Imbens (2011) propose a *bias-corrected* version of the nearest-neighbors matching estimator that is based on a regression adjustment. The correction proceeds in two steps. In step one, we estimate a linear regression model, using only the observations from the opposite treatment group *that were actually used as matches* by the "nearest neighbor" criterion. In addition, we need to weight each such observation by $\mathcal{K}_l$, i.e. the (weighted) number of times it was used as a match.

Thus, we regress $\tilde{\mathbf{y}}_l$ on $\tilde{\mathbf{X}}_l$, where, for a single observation,

$$\begin{aligned} \tilde{y_l} &= \sqrt{\mathcal{K}_l} * y_l \quad \text{and} \\ \tilde{\mathbf{x}}_l &= \sqrt{\mathcal{K}_l} * \mathbf{x}_l \end{aligned} \tag{7}$$

Let the estimated regression coefficients from this model be denoted as $\hat{\boldsymbol{\beta}}$. Consider a specific observation i. We can generate a regression-based prediction for i as $\hat{\mu}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$. Similarly, we can generate regression-based predictions for all matched observations that go with i, i.e the elements in $\mathcal{J}_{i,M}$, as $\hat{\mu}_l = \mathbf{x}_l\hat{\boldsymbol{\beta}}$.

In step 2, the regression-corrected counterfactuals can be derived as:

$$\hat{y_{0i}} = \begin{cases} y_{0i} & \text{if} \quad w_i = 0 \\ \frac{1}{L_{i,M}} \sum_{l \in \mathcal{J}_{i,M}} (y_l + \hat{\mu}_i - \hat{\mu}_l) & \text{if} \quad w_i = 1 \end{cases}$$
$$\text{and} \tag{8}$$
$$\hat{y_{1i}} = \begin{cases} y_{1i} & \text{if} \quad w_i = 1 \\ \frac{1}{L_{i,M}} \sum_{l \in \mathcal{J}_{i,M}} (y_l + \hat{\mu}_i - \hat{\mu}_l) & \text{if} \quad w_i = 0 \end{cases}$$

These counterfactuals are then used in the expressions for ATE and ATT as given in equations (5) and (6).

Consistent standard errors

As discussed in Abadie and Imbens (2011), the bootstrapping approach to obtain standard errors for the matched ATE and ATT estimates is generally inappropriate. Instead, they derive analytical solutions for the variance (and thus standard errors) for these estimates.

A central component in this derivation is the term $\hat{\sigma^2}$. For the ATE, it can be written as

$$\hat{\sigma}^2_{ate} = \frac{1}{2n} \sum_{i=1}^{n} \left[\frac{1}{L_{i,M}} \sum_{l \in \mathcal{J}_{i,M}} \{ w_i (y_i - y_l - \hat{\tau}_{ate}) + (1 - w_i)(y_l - y_i - \hat{\tau}_{ate}) \}^2 \right] \tag{9}$$

For the ATT, it is given as

$$\hat{\sigma}^2_{att} = \frac{1}{2n_1} \sum_{i:w_i=1} \left[\frac{1}{L_{i,M}} \sum_{l \in \mathcal{J}_{i,M}} (y_i - y_l - \hat{\tau}_{att})^2 \right] \tag{10}$$

These estimated $\hat{\sigma^2}$ is then inserted in the general variance formula. For the ATE:

$$\hat{V}_{ate} = \frac{1}{n^2} \sum_{l=1}^{n} \left[(1 + \mathcal{K}_l)^2 \hat{\sigma}^2_{ate} \right] \tag{11}$$

where, as defined in (2), $\mathcal{K}_l$ is the (weighted) number of times observation l was used as a match for the opposite treatment group. For the ATT we have

$$\hat{V}_{att} = \tfrac{1}{n_1^2} \sum_{l=1}^{n} \left[\{w_l - (1 - w_l)\, \mathcal{K}_l\}^2\, \hat{\sigma}_{att}^2 \right] \tag{12}$$

Note that as for the ATE, the summation is over *all n observations*. If an observation l is in the control group (with $w_l = 0$), its contribution to the variance is weighted by $(\mathcal{K}_l)^2$. If it is in the treated group, it will not have been used as a match by definition, i.e. $\mathcal{K}_l = 0$, and its contribution to the overall variance is simply $\hat{\sigma}_{att}^2$.

The standard errors for either estimator can then be computed as the square root of the estimated variance.[1]

Script `mod5s3` illustrates these variance derivations for the case of the ATT, using the Lalonde (1986) labor application.

References

Abadie, A., Drucker, D., Herr, J. and Imbens, G. (2004). Implementing matching estimators for average treatment effects in Stata, *The Stata Journal* **4**: 290–311.

Abadie, A. and Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review, *Review of Economics and Statistics* **86**: 4–29.

Abadie, A. and Imbens, G. (2006). Large-sample properties of matching estimators for average treatment effects, *Econometrica* **74**: 235–267.

Abadie, A. and Imbens, G. (2011). Bias-corrected matching estimators for average treatment effects, *Journal of Business & Economic Statistics* **29**: 1–11.

Abbott, J. and Klaiber, H. (forthcoming). The value of water as an urban club good: A matching approach to community-provided lakes, *Journal of Environmental Economics and Management* .

Angrist, J. D. and Pischke, J. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics, *Journal of Economic Perspectives* **24**: 3–30.

Diamond, A. and Sekhon, J. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies, *The Review of Economics and Statistics* **95**: 932–945.

Greene, W. (2012). *Econometric Analysis*, 7th edn, Pearson / Prentice Hall.

Ho, D., Kosuke, I., King, G. and Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference, *Policitcal Analysis* **15**: 199–236.

Lalonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data, *American Economic Review* **76**: 604–620.

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*, MIT Press.

[1] Abadie and Imbens (2011) also illustrate the derivation of a heteroskedasticity-robust version of this variance estimators. For details see also Abadie et al. (2004).