Matching Methods for Causal Inference

Daniel E. Ho[1] Kosuke Imai[2] Gary King[3]
Elizabeth A. Stuart[4]

September 8, 2006 (11:33)

[1]J.D. candidate, Yale Law School, Ph.D. candidate, Department of Government, Harvard University. (Center for Basic Research in the Social Sciences, 34 Kirkland, Cambridge MA 02138, USA; http://www.people.fas.harvard.edu/~deho, Deho@Fas.Harvard.Edu).

[2]Assistant Professor, Department of Politics, Princeton University (Corwin Hall 041, Department of Politics, Princeton University, Princeton NJ 08544, USA; http://www.princeton.edu/~kimai, KImai@Princeton.Edu).

[3]David Florence Professor of Government, Harvard University (Center for Basic Research in the Social Sciences, 34 Kirkland Street, Harvard University, Cambridge MA 02138; http://GKing.Harvard.Edu, King@Harvard.Edu, (617) 495-2027).

[4]Ph.D. Candidate, Department of Statistics, Harvard University. (Science Center 702, One Oxford Street, Cambridge, MA 02138, USA; http://www.people.fas.harvard.edu/~estuart, Stuart@Stat.Harvard.Edu).

Abstract

Randomized experiments have long been considered the gold standard for the estimation of causal effects. Yet due to ethical and practical reasons, it is often infeasible to do randomized experiments in political science. We discuss the ideas of using matching to replicate randomized experiments using observational data to draw causal inferences. The main goal is to obtain treated and control groups similar to each other on all of the observed background covariates. The framework for these methods is the Rubin-Holland model for causal inference, which has penetrated the biomedical sciences and to a more limited extent the social sciences. Yet with a few unpublished exceptions the framework has not been applied to causal questions in political science, leaving many causal "answers" in the discipline to rest on unfounded functional form assumptions. We provide a summary and guidelines to apply matching to causal questions in political science, with applications in international relations, American, and comparative politics, as well as user-friendly "MatchIt" software for R / Splus that implements these matching methods.

Contents

under control. $T_i = 1$ indicates that unit i was assigned treatment, and $T_i = 0$ that unit i was assigned control. $Y_i(1)$ and $Y_i(0)$ are jointly unobservable, such that we only observe $Y_i^{obs} = T_i(Y_i(1)) + (1 - T_i)(Y_i(0))$. This is often termed as the "fundamental problem of causal inference" (Holland 1986, p. 947, **?**, p. 79).

The treatment effect for unit i is defined as $\alpha_i = Y_i(1) - Y_i(0)$. Two quantities of interest are the average treatment effect (ATE),

$$\overline{\alpha} = E(Y_i(1)) - E(Y_i(0)), \tag{1}$$

or the average treatment effect on the treated (ATT),

$$\overline{\alpha}_T = E(Y_i(1) - Y_i(0)|T_i = 1). \tag{2}$$

Variance of the Treatment Effect To calculate the variance of the ATT requires determining first whether the quantity of interest substantively is a sample or population parameter (Imbens 2003, pp. 28-29.)[1]

If we are interested in the sample ATT, estimating the variance of the treatment effect is straightforward, involving simply the variance of the estimated treatment effect for each treated unit.

To understand the logic of why the matching process does not affect the variance estimation, first consider pairwise exact matching on all covariates. Since we're fully conditioning on X in that case, the matching process involves no estimation uncertainty and the variance of sample ATE may simply be estimated by:

$$V(\overline{\alpha}_T) = V\big\{(Y_i^{obs}(1)|T = 1) - (\hat{Y}_i(0)|T = 1)\big\}, \tag{3}$$

where the missing potential outcome $\hat{Y}_i(0)|T = 1$ is estimated by the matched control unit to unit i. In other words, the variance of the sample ATT is simply the variance of the differences in the matched pairs.[2]

[1] This section discusses the estimation of the variance of the ATT (versus ATE) for simplicity of notation. With the exception of the note on the efficiency gain due to knowledge of the propensity score for ATE variance estimation, the calculations are analogous.

[2] An estimate due to Neyman (1923), which does not use the covariance between matched pairs and was initially applied to randomized experiments is sometimes used instead: $\hat{V}(\overline{\alpha}) = \frac{V(Y_i(1)^{obs}|T=1)}{M} + \frac{V(\hat{Y}_i(0)|T=1)}{N-M}$, where M represents the number of treated units (Imbens and Rubin 2002, Chapter 6). Imbens and Rubin (2002) note that this variance estimate may be understood as (a) an unbiased estimator under the assumption of constant additive treatment effects, (b) an upwardly biased estimator relaxing that assumption, or (c) an unbiased estimator of the superpopulation ATE.

Similar to exact matching, propensity score matching is simply a means to obtain a balanced dataset, and the variance of the sample ATT is the same as in Equation 3. The only difference is the fact that the missing potential outcome $\hat{Y}_i(0)|T = 1$ is now estimated by the matched control unit *via the propensity score*. Since the propensity score does not estimate any population parameter and serves its role purely as a "balancing score", it does not introduce any added uncertainty to the variance of the treatment effect.[3] (Note, however, that knowledge of the propensity score might decrease the variance bound for the ATT, but not the ATE (Frölich 2002; Hahn 1998; Hirano, Imbens and Ridder 2002).) In fact, this variance estimate might even be conservative for the sample ATT, providing over-coverage compared to the variance conditional on the covariates (**?**).

If on the other hand we are interested in the population ATT, the variance may be estimated by the bootstrap, or even the Neyman estimate in Note 2 (Imbens 2003, Imbens and Rubin 2002, Chapter 6).

Quantity of Interest. Average treatment effect vs. average treatment effect on the treated; limiting inferences to what's scientifically estimable.

Defining units, treatment and outcome. In assessing a causal effect, the definition of units is crucial. We divide units [variables???] into two classes: pre-exposure – those whose values are determined prior to exposure to the cause; post-exposure – those whose values are determined after exposure to the cause.[4] One particularly prevalent issue in comparative politics and international relations consists of cases where treatment occurs at some time t and persists indefinitely. In Simmons's data, for example, commitment to Article VIII occurs, but there is no possibility of "un-committing." In the comparative electoral systems data, with few exceptions, a country has the same electoral system for long periods of time. One approach taken by researchers has been to define units as unit-years, such as country-years or dyad-years. This, however, violates the stability assumptions of matching. [In more traditional econometric terms, the treatment indicator may

[3]There is one slight exception to this, namely that ties in the propensity score are resolved by a random draw. This is a scenario that is likely to happen when the explanatory covariates are largely categorical. Even in this instance, reporting estimates from one draw of matched pairs is correct (akin to drawing a random sample from a population), but to avoid potentially "spurious" findings, the user might well be advised to impute missing potential outcomes several times and combine estimates across these datasets via standard multiple imputation rules (see Rubin 1987).

[4]Holland, p. 946

thereby suffer of serial correlation[5] inducing an artificial causal effect.] We suggest redefining the units to be countries (rather than country/years as done by Simmons and others), which properly reflects the hypothesized assignment mechanism. In these examples where treatment occurs at time t and persists indefinitely, "treatment assignment" is not done every year for each state, as implied by the usual treatment of each country and year as a different unit. Rather, treatment is applied to each state only once (at time t) and the units modeled should reflect that. Thus, we recommend [?????] the use of modeling at the country level, where the variables before time t are treated as covariates and the variables after time t are treated as outcomes. This idea is explored by Bertrand et al. (2002), and they find that it has good properties. We will discuss this method more below.

Algorithms

A Sample Matching Algorithm: Optimal Nearest Neighbor One-to-one Matching with Replacement

1. Generate propensity score.
2. Check balance of pre-treatment covariates. If an imbalance exists, reestimate the assignment model Step 1 by adding higher order terms and interactions. Continue until a obtaining a balance in pre-treatment covariates.
3. Optional: Discard control units with propensity scores below the minimum score of treated units (note, other forms of discarding may be suitable).
4. Optional: Re-estimate propensity score.
5. To match units:
   (a) Calculate the absolute difference between propensity score of the *non-matched* treated unit with the highest propensity score (unit i) and all control units.
   (b) Match units i and j, where:

$$|(e_i|T=1)-(e_j|T=0)| < |(e_i|T=1)-(e_k|T=0)|, \forall k \neq j. \quad (4)$$

   In the instance multiple control units are within the minimum distance, draw randomly to match one of those control units.

[5] Duflo, 2003

6. Continue Steps 5a-5b for all non-matched treated units until all treated units are matched.

7. Calculate the average treatment effect on the treated for the sample ATE_s, where:

$$ATE_s = \frac{1}{N_T} \sum_{i=1}^{N_T} \big\{ (Y_i|T=1) - (Y_m|T=0) \big\} \tag{5}$$

   where N_T represents the number of treated units, Y_i refers to treated units, and Y_m refers to the control unit matched to i.

8. To estimate the point estimate and variance of ATE bootstrap the donor pool B times and repeat (Steps 5-6) for each bootstrapped sample, storing the estimated average treatment effect ATE_s from each bootstrap.

9. Lastly, calculate quantities of interest from the distribution of bootstrapped ATE_s, where the point estimate for the average treatment effect on the treated is:

$$ATE = \frac{1}{B} \sum_{s=1}^{B} ATE_s \tag{6}$$

   and the variance is:

$$Var(ATE) = \frac{1}{B-1} \sum_{s=1}^{B} (ATE - ATE_s)^2 \tag{7}$$

The Subclassification/Regression Adjustment Algorithm for a Binary Outcome Variable

1. Generate propensity score.

2. Construct 10 blocks defined by the deciles of the propensity score for the treated group (where to obtain balance in the pre-treatment covariates, (a) split imbalanced blocks further, and/or (b) respecify the asssignment model to take into account interactions and squares of imbalanced covariates).

3. Discard control units with propensity scores below the minimum score of treated units.

4. Re-estimate propensity score and regenerate blocks without the discarded units (optional)

5. For units within each substratum:

   (a) Estimate the logistic model, regressing the binary outcome on the propensity scores and the treatment indicator. (If sample sizes permit, this logistic model may also be estimated using all covariates.)

   (b) Generate a posterior distribution of the parameters from this model, drawing 1000 simulated values of β, where sim represents each simulated set of parameters.

   (c) For each substratum and one draw of the parameters β from the posterior distribution, impute the missing potential outcomes $\hat{Y}_i(1)|T_i = 0$ and $\hat{Y}_i(0)|T_i = 1$ with a logistic regression, where:

$$(\hat{Y}_i(1)|T_i = 0) = \frac{1}{1 + exp(-X_i^t\beta)} \tag{8}$$

   and

$$(\hat{Y}_i(0)|T_i = 1) = \frac{1}{1 + exp(-X_i^c\beta)} \tag{9}$$

   where the superscripts c and t represents the counterfactual treatment indicator of control and treatment, respectively.

   (d) Calculate the treatment effect for all units in the substratum

$$TE_d = \begin{pmatrix} (Y(1)|T = 1) \\ (\hat{Y}(1)|T = 0) \end{pmatrix} - \begin{pmatrix} (\hat{Y}(0)|T = 1) \\ (Y(0)|T = 0) \end{pmatrix} \tag{10}$$

   where d represents the subscript for the decile and TE_d is a vector of treatment effects for all i units in the substratum, $(Y(1)|T = 1)$ and $(Y(0)|T = 0)$ represent vectors of observed potential outcomes, and $(\hat{Y}(1)|T = 0)$ and $(\hat{Y}(0)|T = 1)$ represent vectors of imputed missing potential outcomes.

   (e) Calculate the average treatment effect for the substratum conditional on the set of simulated parameters $ATE_{d,sim}$:

$$ATE_{d,sim} = \frac{1}{n_d}\sum_{i=1}^{n_d} TE_{i,d} \tag{11}$$

   where n_d represents the number of units in the substratum, and $TE_{i,d}$ refers to the treatment effect for each unit i in substratum d.

(f) Repeat Steps 5a-5e for all draws of β.

(g) Calculate the average treatment effect for the substratum across all simulated parameters ATE_d:

$$ATE_d = \frac{1}{1000} \sum_{sim=1}^{1000} ATE_{d,sim} \tag{12}$$

(h) Calculate the variance of ATE_d for that substratum:

$$Var(ATE_d) = Var(ATE_{d,sim}) \tag{13}$$

(i) Finally repeat Steps 5a-5h to generate ATE_d and $Var(ATE_d)$ for all substrata.

6. Having completed each of the above steps for each substratum, calculate the overall average treatment effect ATE:

$$ATE = \sum_{d=1}^{10} \Big\{ \big(\frac{n_{d,t}}{N_t}\big) \big(ATE_d\big) \Big\} \tag{14}$$

where $n_{d,t}$ represents the number of treated units in each substratum d and N_t represents the total number of treated units across all substrata.

7. Finally, calculate the overall variance of the ATE:

$$Var(ATE) = \sum_{d=1}^{10} \Big\{ \big(\frac{n_{d,t}}{N_t}\big)^2 \big(Var(ATE_d)\big) \Big\}. \tag{15}$$

References

Frölich, Markus. 2002. "What is the Value of Knowing the Propensity Score for Estimating Average Treatment Effects?" IZA Discussion Paper 548, University of St. Gallen.

Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66:315–31.

Hirano, Keisuke, Guido W. Imbens and Geert Ridder. 2002. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." Manuscript.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.

Imbens, Guido W. 2003. "Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review." Manuscript, UC Berkeley.

Imbens, Guido W. and Donald B. Rubin. 2002. "Causal Inference." Book Manuscript.

Neyman, J. 1923. "On the application of probability theory to agricultural experiments. Essay on Principles. Section 9"." *Statistical Science* 5:465–480. Translated in 1990, with discussion.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley.