Optimal Data Analysis
Vol. 9 (June 22, 2020), 183-189

2155-0182/10/$3.00

Implementing CTA from Within Stata: Reassessing the Propensity Score Estimation Approach Used in the National Supported Work Experiment (*Invited*)

Ariel Linden, Dr.P.H.
Linden Consulting Group, LLC

Data from the National Supported Work (NSW) randomized experiment have been used frequently over the past 30 years to demonstrate the implementation of various non-experimental methods for drawing causal inferences about treatment effects. The present paper reassesses the approach used by Dehejia and Wahba (2002) for estimating propensity scores and compares it to a propensity score CTA model which is generated by using the new Stata package for implementing CTA.

Studies in which participants are randomized to treatment are considered the gold standard for assessing causal inference because randomization putatively ensures that the study groups do not differ systematically in their characteristics, and consequently, treatment effects are assumed to be unbiased.[1] If randomization is infeasible, investigators rely on statistical techniques which model treatment assignment in order to control for threats to validity[2,3] which may compromise causal interpretation of the results.[4-8]

Herein I reanalyze data taken from the National Supported Work (NSW) experiment, originally discussed by LaLonde[9] in the context of economic evaluation, and also widely used to demonstrate the implementation of a variety of non-experimental techniques, such as propensity scoring methods, in assessing causal inference. Specifically, I reassess Dehejia and Wahba's[10] approach to estimating the propensity score used to compare a subset of participants in the NSW experiment to a pool of potential control participants in the Current Population Survey (CPS). I use the new Stata package called **cta**[11] which implements CTA within the Stata environment to generate a CTA propensity score model, and I assess if the resulting model is consistent with Dehejia and Wahba's[10] propensity score estimation model. As **cta** is a wrapper for CTA software[12], the CTA64.exe file (which is available

2155-0182/10/$3.00

at https://odajournal.com/resources/) must be on the computer for **cta** to work. To download the **cta** package, at the Stata command line type: "ssc install cta" without the quotation marks.

Methods

Data

The NSW was a US federally- and privately-funded program that aimed to provide work experience for individuals who had faced economic and social problems prior to enrollment in the program. Candidates for the experiment were selected on the basis of eligibility criteria, and then were either randomly assigned to, or excluded from, the training program. I use the same subset of NSW data used by Dehejia and Wahba[10], joining the 185 treated units from the NSW experiment to comparison units from the 15,992 individuals in the Current Population Survey (CPS). Data were retrieved from: http://users.nber.org/~rdehejia/nswdata2.html.

Variables (attributes) available for individuals across both sets of data (i.e. NSW and CPS) were age, education, black, Hispanic, no degree, married, real earnings in 1974, 1975 and 1978 (adjusted to 1982 US dollars), and indicators for 1974 and 1975 unemployed status. The outcome (primary model attribute) was real earnings in 1978, and the treatment (class) variable indicates whether individuals participated in the NSW intervention, or were untreated from the CPS data.

Analysis

Dehejia and Wahba[10] estimated a propensity score in which the binary treatment indicator was regressed on age, age^2, age^3, education, education^2, married, no degree, black, Hispanic, earnings in 1974 and 1975, unemployed in 1974 and 1975, and an interaction of education and earnings in 1974. The logic they used to choose the right-hand side variables in the propensity score model was as follows[10]:

- Start with a parsimonious logit function to estimate the score.
- Sort data according to estimated propensity score (ranking from lowest to highest).
- Stratify all observations such that estimated propensity scores within a stratum for the treated and control units are close (i.e., no significant difference); e.g., start by dividing observations in blocks of equal score range (0-0.2, ..., 0.8-1).
- Statistical test: for all covariates, the differences-in-means across treated and control units within each block are not significantly different from zero.
    1. If covariates are balanced between treated and control observations for all blocks, stop.
    2. If covariate *i* is not balanced for some blocks, divide block into finer blocks and re-evaluate.
    3. If covariate *i* is not balanced for all blocks, modify the logit by adding interaction terms and/or higher-order terms of the covariate *i* and re-evaluate.

It is apparent that this approach is both labor intensive and will miss relationships that may exist between the covariates, which are not explored in step 3. In contrast, I utilize CTA to predict treatment assignment from the set of covariates (attributes), thereby ensuring that all statistically significant covariates as well as all statistically significant interactions between covariates are identified. Clearly, a manual approach could not accomplish such an analysis in any reasonable amount of time.[13]

The following syntax generated the CTA model (see the help file for **cta** for a complete description of the syntax options):

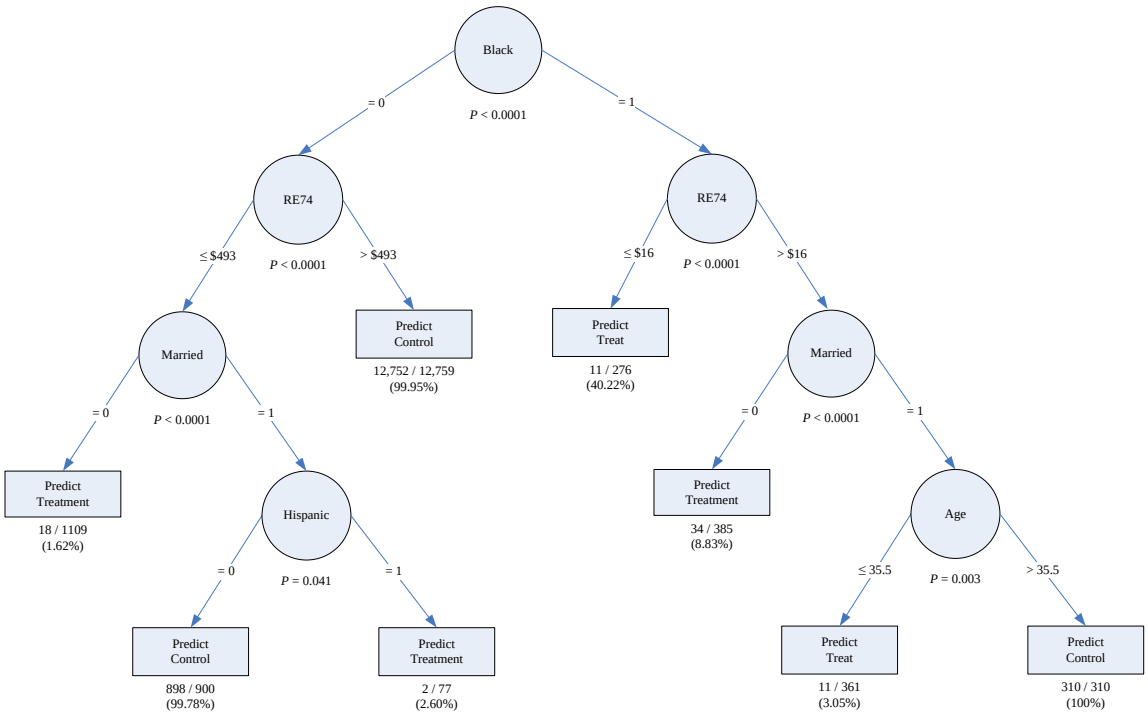
2155-0182/10/$3.00

```
cta treat age educ black hispan married re74
re75, pathcta("C:\ CTA\")
store("C:\CTA\output") cat( black hispan
married) loo(stable) inter(10000) prune(0.05)
enumerate name(NWS)
```

The above syntax is explained as follows: The outcome variable "treat" is the *class* variable; the seven variables listed before the comma are covariates specified as *attributes*; the directory path for the CTA64.exe file on my computer is "C:\CTA\"; the directory path where output and other files generated during the analysis are stored is "C:\CTA\output"; the *cat*() option indicates categorical attributes; the number of iterations (repetitions) for computing a permutation *P*-value is 10,000; leave-one-out (LOO) cross-generalizability analysis is used to identify and retain attributes having the same classification performance in LOO and training (total sample) analysis; the tree is pruned with experimentwise $P < 0.05$ used as the cutpoint for inclusion; and a model enumerating the first three nodes was used (Yarnold and Soltysik[12] describe the CTA modeling process as well as the interpretation of CTA results). **cta** produces an extract of the total output produced by CTA software: the complete output is stored in the specified directory with the extension ".out".

Below we present a diagram of the pruned model, which achieved overall ESS of 82.43 (a strong effect)—slightly less than the enumerated model (ESS=83.95), but much more parsimonious (4 levels and 15 nodes, *vs*. 6 levels and 61 nodes).

In reviewing this diagram, it is evident that individuals from NSW were very different than those surveyed in the CPS. That is, an individual is predicted to have participated in the intervention if: (1) they were not Black, had real income in 1974 ≤ $493, and not married; (2) they were Hispanic, had real income in 1974 ≤ $493, and were married; (3) they were Black and had income in 1974 ≤ $16; (4) they were Black, had income in 1974 > $16, and were not married; and (5) they were Black, married, and ≤ 35.5 years of age. All these pathways were statistically less than $P < 0.05$.



2155-0182/10/$3.00

Comparing this model to that estimated by Dehejia and Wahba[10] we see many notable differences. (1) CTA found that age was a factor for only a small subset of all individuals in the sample, whereas Dehejia and Wahba[10] applied the linear, squared and cubed version of age across all individuals in the sample. (2) CTA found that education did not predict treatment assignment, whereas Dehejia and Wahba[10] used education and a derivative variable called "no degree" as covariates in the model. (3) CTA did not find real earnings in 1975 to be predictive of treatment assignment, whereas Dehejia and Wahba[10] included this variable, and a derivative covariate representing being unemployed in 1975, in their model.

Next, I generated a CTA model using the variables specified by Dehejia and Wahba[10] in their propensity score model, to determine if CTA would find those covariates predictive of treatment assignment (i.e., higher order terms of age and interactions between education and real income in 1974). The CTA model was specified using the following **oda** syntax:

```
cta treat age age2 age3 educ educ2
educXre74 black hispan married nodegree
re74 re75 u74 u75 , pathcta("C:\CTA\")
store("C:\CTA\output") cat( black hispan
married ) loo(stable) iter(10000) prune(0.05)
enumerate
```

The CTA model produced from this specification was identical to that of the first CTA model. This supports the contention that, in addition to being *inefficient* and *incomplete*, manually choosing covariates and generating higher order and interaction terms is likely to *miss the true predictive relationships between covariates and the outcome variable*.

As a final step, I compare the estimated treatment effects using propensity score models of Dehejia and Wahba[10] *vs*. the propensity score model derived by CTA (see references[14,15] for a description of propensity score weighting using CTA). I evaluate these data using **oda** with the following syntax (see the help file for **oda** for a complete description of syntax options):

```
oda treat re78, pathoda("C:\ ODA\")
store("C:\ODA\output") iter(10000) loo
seed(1234) wt(ctawt)
```

This syntax is explained as follows: The variable "treat" is the *class* variable; the outcome variable "re78" (earnings in 1987) is the *attribute*; the directory path where the megaODA.exe file is located on my computer is "C:\ODA\"; the directory path where the output and other files generated during the analysis should be stored is "C:\ODA\output"; the number of iterations (repetitions) for computing a permutation *P*-value is 10,000; LOO analysis is performed; the seed is set to 1234 to ensure replication of the permutation results; and the CTA weights are specified in the wt() option.

The **oda** package produces an extract of the total output produced by the ODA software (the complete output is stored in the specified directory with the extension ".out").

```
ODA model:
----------
IF RE78 <= 12595.355 THEN TREAT = 1
IF 12595.355 < RE78 THEN TREAT = 0


Summary for Class TREAT  Attribute RE78
----------------------------------------

Performance Index           Train    LOO
-----------------           ------  ------
Overall Accuracy            60.88%  60.87%
Overall Wtd Accuracy        60.54%  60.41%
PAC TREAT=0                 60.59%  60.59%
PAC TREAT=1                 85.95%  85.41%
Effect Strength PAC         46.53%  45.99%
Wtd PAC TREAT=0             60.13%  60.13%
Wtd PAC TREAT=1             96.91%  85.42%
Effect Strength Wtd PAC     57.04%  45.55%
PV TREAT=0                  99.73%  99.72%
PV TREAT=1                   2.46%   2.45%
Effect Strength PV           2.19%   2.17%
Wtd PV TREAT=0              99.94%  99.73%
Wtd PV TREAT=1               2.68%   2.37%
Effect Strength Wtd PV       2.62%   2.09%
Effect Strength Total       24.36%  24.08%
Effect Strength Wtd Total   29.83%  23.82%


Monte Carlo summary (Fisher randomization):
-------------------------------------------
Iterations:  10000
Estimated p: 0.000700


Results of leave-one-out analysis
---------------------------------
16177 observations

Fisher's exact test (directional) classification table  p = .102E-0036
```

2155-0182/10/$3.00

As shown in the **oda** output above, the ODA model is interpreted as follows: "if real earnings in 1978 <= $12595.335, then predict that the treatment group is 1 (treatment). If the earnings are > $12595.335, then predict that the treatment group is 0 (controls)."

The effect strength for sensitivity (ESS) is labelled in the output as "Effect Strength Wtd PAC". In the training data the ESS is 57.04% (a relatively strong effect) and in the LOO analysis is 45.55% (moderate effect).[16] The permutation *P*-value for the training sample was 0.0007 and for the LOO analysis was < 0.0001.

In contrast, the ODA model used to assess treatment effect based on the propensity score and matching used by Dehejia and Wahba[10] predicted the control group had real earnings in 1978 <= $1237.291, and treatment group had real earnings > $1237.291.[17] Given that these are observational data, and that the CPS pool of "control" individuals was vastly different from the treated sample from the NSW experiment, it is impossible to know the true effect. Discrepancy of results based on different propensity scoring techniques further supports the reliance on randomized trials to provide unequivocal estimates of treatment effects.

Discussion

This paper demonstrates how the new Stata package **cta** can be used to generate a propensity score model that captures all the covariates and interactions between covariates that predict treatment assignment. In utilizing CTA for this procedure, the modeling process is automated and the resulting model is maximally accurate. CTA should therefore be considered the preferred approach over commonly-used parametric models because CTA avoids the assumptions required of parametric models, is insensitive to skewed data or outliers, and can use combinations of variable metrics including categorical, Likert-type integer, and real number measurement scales. Moreover, in contrast to regression models, CTA has the unique ability to ascertain the precise location of optimal (maximum-accuracy) cutpoints on the outcome variable (in this case, treatment assignment), which in turn, facilitates the use of measures of predictive accuracy.

Finally, the findings continue to support the recommendation to use the ODA and CTA frameworks to evaluate the efficacy of health-improvement interventions and policy initiatives.[18-33]

References

[1]Linden A, Yarnold PR (2017). Minimizing imbalances on patient characteristics between treatment groups in randomized trials using classification tree analysis. *Journal of Evaluation in Clinical Practice, 23*, 1309-1315. DOI: 10.1111/jep.12792

[2]Linden A (2007). Estimating the effect of regression to the mean in health management programs. *Disease Management and Health Outcomes, 15*, 7-12.

[3]Linden A, Yarnold PR (2016). Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice, 22*, 839-847.

[4]Linden A, Adams J (2006). Evaluating disease management program effectiveness: an introduction to instrumental variables. *Journal of Evaluation in Clinical Practice, 12*, 148-154.

[5]Linden A, Adams JL (2010). Using propensity score-based weighting in the evaluation of health management programme effectiveness. *Journal of Evaluation in Clinical Practice, 16*, 175-179.

[6]Linden A, Adams JL (2010). Evaluating health management programmes over time. Application of propensity score-based weighting to longitudinal data. *Journal of Evaluation in Clinical Practice, 16*, 180-185.

2155-0182/10/$3.00

[7]Linden A (2014). Combining propensity score based stratification and weighting to improve causal inference in the evaluation of health care interventions. *Journal of Evaluation in Clinical Practice, 20,* 1065-1071.

[8]Linden A, Uysal SD, Ryan A, Adams JL (2016). Estimating causal effects for multivalued treatments: A comparison of approaches. *Statistics in Medicine, 35*, 534-552.

[9]LaLonde R (1986). Evaluating the econometric evaluations of training programs. *American Economic Review, 76*, 604–620.

[10]Dehejia RH, Wahba S (2002). Propensity score matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, *84*, 151-61.

[11]Linden A. (2020). CTA: Stata module for conducting Classification Tree Analysis. *Statistical Software Components S458729, Boston College Department of Economics.*

[12]Yarnold PR, Soltysik RC (2016). *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[13]Yarnold PR (2013). Univariate and multivariate analysis of categorical attributes with many response categories. *Optimal Data Analysis*, *2*, 177-190.

[14]Linden A, Yarnold PR (2017). Using classification tree analysis to generate propensity score weights. *Journal of Evaluation in Clinical Practice, 23*, 703-712.

[15]Linden A, Yarnold PR (2018). Estimating causal effects for survival (time-to-event) outcomes by combining classification tree analysis and propensity score weighting. *Journal of Evaluation in Clinical Practice, 24*, 380-387.

[16]Yarnold PR, Soltysik RC. *Optimal data analysis: Guidebook with software for Windows*. Washington, D.C.: APA Books, 2005.

[17]Linden A, Yarnold PR (2020). Implementing ODA from Within Stata: A Reanalysis of the National Supported Work Experiment, *Optimal Data Analysis, 9*, 178-182.

[18]Linden A, Adams J, Roberts N (October, 2003). *Evaluation methods in disease management: determining program effectiveness*. Position Paper for the Disease Management Association of America (DMAA).

[19]Linden A, Roberts N (2005). A Users guide to the disease management literature: recommendations for reporting and assessing program outcomes. *American Journal of Managed Care, 11*, 81-90.

[20]Linden A, Adler-Milstein J (2008). Medicare disease management in policy context. *Health Care Finance Review, 29*, 1-11.

[21]Linden A, Yarnold PR (2018). Identifying causal mechanisms in health care interventions using classification tree analysis. *Journal of Evaluation in Clinical Practice, 24*, 353-361.

[22]Linden A, Yarnold PR (2016). Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. *Journal of Evaluation in Clinical Practice, 22*, 875-885.

[23]Yarnold PR, Linden A (2017). Computing propensity score weights for CTA models involving perfectly predicted endpoints. *Optimal Data Analysis, 6*, 43-46.

[24]Linden A, Yarnold PR (2016). Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice, 22*, 848-854.

[25]Linden A, Yarnold PR (2016). Using machine learning to identify structural breaks in single-group interrupted time series designs. *Journal of Evaluation in Clinical Practice, 22*, 855-859.

2155-0182/10/$3.00

[26]Linden A, Yarnold PR, Nallomothu BK (2016). Using machine learning to model dose-response relationships. *Journal of Evaluation in Clinical Practice*, *22*, 860-867.

[27]Yarnold PR, Linden A (2016). Using machine learning to model dose-response relationships via ODA: Eliminating response variable baseline variation by ipsative standardization. *Optimal Data Analysis*, *5*, 41-52.

[28]Linden A, Yarnold PR (2016). Combining machine learning and matching techniques to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*, *22*, 868-874.

[29]Linden A, Yarnold PR (2017). Modeling time-to-event (survival) data using classification tree analysis. *Journal of Evaluation in Clinical Practice*, *23*, 1299-1308.

[30]Linden A, Yarnold PR (2018). The Australian gun buy-back program and the rate of suicide by firearm. *Optimal Data Analysis*, *7*, 28-35.

[31]Linden A, Yarnold PR (2018). Using machine learning to evaluate treatment effects in multiple-group interrupted time series analysis. *Journal of Evaluation in Clinical Practice*, *24*, 740-744.

[32]Linden A, Yarnold PR (2018). Using ODA in the evaluation of randomized controlled trials. *Optimal Data Analysis*, *7*, 46-49.

[33]Linden A, Yarnold PR (2018). Using ODA in the evaluation of randomized controlled trials: Application to survival outcomes. *Optimal Data Analysis*, *7*, 50-53.

Author Notes

No conflict of interest was reported.