Optimal Data Analysis
Vol. 7 (August 22, 2018), 54-58

2155-0182/10/$3.00

Reanalysis of the National Supported Work Experiment Using ODA

Ariel Linden, Dr.P.H. and Paul R. Yarnold, Ph.D.
Linden Consulting Group, LLC Optimal Data Analysis 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. In the present study we reanalyze the NSW data using ODA and compare results with those estimated using *t*-tests. Statistical results were largely consistent between methods, however ODA found 22.2% (2 of 9) pre-intervention characteristics to be imbalanced. Given that ODA avoids assumptions required of parametric methods, and is insensitive to skewed data and outliers, ODA should be considered the preferred approach when evaluating data from randomized experiments.

In a comprehensive series of papers, ODA and CTA frameworks were applied to observational data to draw causal inferences about treatment effects[1-17] as well as to reanalyzing data from a randomized controlled trial[18] for both count[19] and survival[20] outcomes. In this paper we reanalyze data from the National Supported Work (NSW) experiment which was originally discussed by LaLonde[21] in the context of economic evaluation approaches, but has since then been utilized frequently to demonstrate the implementation of various non-experimental techniques, such as propensity scoring methods, for assessing causal inference. Presently, we apply ODA to these data to assess whether results are consistent with those derived via the usually-employed parametric *t*-test.

Methods

Data

The NSW was a U.S. 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. We use the same subset of NSW data used by Dehejia and Wahba[22]—samples of 185 male treated and 260 control observations. Data were retrieved from: http://users.nber.org/~rdehejia/nswdata2.html. Nine preintervention (pre-treatment) variables were available to assess covariate balance, and the outcome variable was post-treatment earnings in 1978.

2155-0182/10/$3.00

Analyses

ODA models[23,24] were generated separately for each of the nine baseline covariates and for the outcome variable. In each model the binary indicator for treatment was specified as the class variable, and each covariate (and outcome) was set as the attribute. A total of 25,000 Monte Carlo simulations were used to compute *P* values, and leave-one-out (LOO) analysis was conducted to assess cross-generalizability. LOO analysis is inherently one-tailed (directional), and is only conducted for models having $P \leq 0.05$ in training (total sample) analysis.[23,24]

Results obtained by ODA were compared to those obtained using *t*-tests (*t*-tests on the equality of proportions were used for binary variables and *t*-tests on the equality of means were used for continuous/ordered variables). For each *t*-test analysis, treatment *vs*. control was configured as a between-groups factor, and each covariate (and outcome) as a dependent variable. Comparisons were made between ODA and *t*-tests on the full sample only, given that statistical software does not inherently provide LOO analysis for *t*-tests.

Results

Table 1 presents the results of all analyses conducted herein. When using *t*-tests, all preintervention characteristics appear to be balanced according to conventional criteria ($P > 0.05$), and the outcome is statistically significant ($P = 0.005$), indicating that the program was effective in increasing income levels for study participants versus controls.

All nine of the ODA models used to assess covariate balance had a relatively weak effect strength for sensitivity (ESS), indicating the existence of a modest degree of between-group imbalance (i.e., heterogeneity).[23,24]

In contrast to *t*-test findings of no differences between the treatment and control groups, two of nine (22.2%) ODA models had *P* – values ≤ 0.05, meeting the generalized (per-comparison) criterion for statistical significance.[24]

Table 1: Preintervention characteristics and outcome of NWS program participants and non-participants. Values represent cut-points on the covariate (and outcome), and values in parentheses represent sensitivity (for participants) and specificity (for non-participants).

	Treatment (*N*=185)	Controls (*N*=260)	Effect Strength Sensitivity	*P*-value *t*-test	*P*-value (Train)	*P*-value (LOO)
Preintervention Characteristics						
Age	> 20.5 (74.59)	≤ 20.5 (31.92)	6.52%	0.265	0.526	
Education	> 11.5 (29.19)	≤ 11.5 (83.46)	12.65%	0.135	0.015	0.001
Black	= 1 (84.32)	= 0 (17.31)	1.63%	0.649	0.701	
Hispanic	= 0 (94.05)	= 1 (10.77)	4.82%	0.076	0.087	
Married	= 1 (18.92)	= 0 (84.62)	3.53%	0.326	0.371	
Unemployed 1974	= 0 (29.19)	= 1 (75.00)	4.19%	0.325	0.327	
Unemployed 1975	= 0 (40.00)	= 1 (68.46)	8.46%	0.065	0.069	
Real earnings 1974	> 1671.57 (24.32)	≤ 1671.57 (80.38)	4.71%	0.982	0.589	
Real earnings 1975	> 782.60 (34.59)	≤ 782.60 (76.15)	10.75%	0.382	0.050	0.015
Outcome						
Real earnings 1978	> 465.53 (75.14)	≤ 465.53 (38.08)	13.21%	0.005	0.038	0.003

2155-0182/10/$3.00

For education ODA found a statistically significant cut-point ($P = 0.015$) at 12th grade. The ODA model was: if education > 11.5 years, then predict the observation was assigned to the intervention group, otherwise predict the observation was assigned to the control group: for this model ESS = 12.65 corresponding to a relatively weak effect.[23,24] ESS was stable in LOO validity analysis ($P = 0.001$).

For real earnings in 1975 ODA found a statistically significant cut-point ($P = 0.05$) at $782.60. The ODA model was: if real earnings in 1975 > $782.60, then predict the observation was assigned to the intervention group, otherwise predict the observation was assigned to the control group: for this model ESS = 10.75 corresponding to a relatively weak effect.[23,24] ESS was stable in LOO validity analysis ($P = 0.015$).

Finally, the ODA outcome model was: if earnings in 1978 > $465.53, then predict the observation was assigned to the intervention group, otherwise predict the observation was assigned to the control group: for this model ESS = 13.21 ($P = 0.038$) corresponding to a relatively weak effect.[23,24] ESS fell to 12.67 in LOO validity analysis ($P = 0.003$).

Discussion

In this paper we have once again demonstrated the value of using ODA for evaluating randomized experiments. ODA should be considered the preferred approach over commonly-used parametric models because ODA avoids the assumptions required of parametric models (e.g. linearity, sufficient sample size, independence, etc.), while by being insensitive to skewed data or outliers, and in its ability to handle any variable metric including categorical, Likert-type integer, and real number measurement scales.[23,24] Moreover, in contrast to regression models, ODA also has the distinct ability to ascertain where the optimal (maximum-accuracy) cut-points are on a variable of interest, which in turn facilitates the use of measures of predictive accuracy.

Furthermore, ODA has the capability to use cross-validation methods such as LOO which was employed presently, in addition to hold-out, multiple-sample, test-retest, and bootstrap[24,25] cross-validation methods to assess the generalizability of the model to other individuals (outside of the study) with similar characteristics.[26] Again, there is no equivalent in the parametric model-based framework, failing to provide insight into the likelihood that any observed intervention effect would generalize.

Another advantage of ODA lies in the analysis of designs with an attribute reflecting more than two qualitatively distinct categories. Using ODA a single multicategorical variable is created, on which each observation's actual category level is indicated by a dummy code—for example, 1, 2, and 3 for treatment A group, treatment B group, and control group, respectively. ODA will identify the model which maximizes ESS. The most accurate model might find that all three category levels differ: for example, A < B < C. Or, the most accurate model might find that one category level differs from two levels which do not differ from each other: for example, A > (B = C). Finally, the most accurate model might find that two category levels differ, with a third level intermediate between them: for example, A = B, B = C, and A > C. This method of dummy-coding cannot be used with multivariable linear models, for which one of the category levels is selected as a comparison reference.[27]

Finally, it should be noted that the objective of this paper was to demonstrate the use of ODA to draw causal inferences about baseline covariate balance and treatment effects, and compare results to findings obtained using *t*-tests. In contrast to the *t*-test, ODA identified imbalances on two of the attributes (covariates) for this sub-sample of the original RCT. The existence of statistically reliable baseline imbalances indicates that weighting by propensity scores may be appropriate with these data, for making causal inferences regarding the

2155-0182/10/$3.00

effect of the intervention. This methodology is demonstrated elsewhere, and falls beyond the scope of the present paper.[7, 10-12]

References

1. Linden A, Yarnold PR. Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice* 2016;22:839-847.

2. Linden A, Yarnold PR. Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice* 2016;22:848-854.

3. Linden A, Yarnold PR. Using machine learning to identify structural breaks in single-group interrupted time series designs. *Journal of Evaluation in Clinical Practice* 2016;22:855-859.

4. Linden A, Yarnold PR, Nallomothu BK. Using machine learning to model dose-response relationships. *Journal of Evaluation in Clinical Practice* 2016;22:860-867.

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

6. Linden A, Yarnold PR. Combining machine learning and matching techniques to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice* 2016;22:868-874.

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

8. Yarnold PR, Linden A. Novometric analysis with ordered class variables: The optimal alternative to linear regression analysis. *Optimal Data Analysis* 2016;22:65-73.

9. Yarnold PR, Linden A. Theoretical aspects of the D statistic. *Optimal Data Analysis* 2016;22:171-174.

10. Linden A, Yarnold PR. Using classification tree analysis to generate propensity score weights. *Journal of Evaluation in Clinical Practice* 2017;23:703-712.

11. Yarnold PR, Linden A. Computing propensity score weights for CTA models involving perfectly predicted endpoints. *Optimal Data Analysis* 2017;6:43-46.

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

13. Linden A, Yarnold PR. Minimizing imbalances on patient characteristics between treatment groups in randomized trials using classification tree analysis. *Journal of Evaluation in Clinical Practice* 2017;23:1309-1315.

14. Linden A, Yarnold PR. The Australian gun buy-back program and the rate of suicide by firearm. *Optimal Data Analysis* 2018;7:28-35.

15. Linden A, Yarnold PR. Identifying causal mechanisms in health care interventions using classification tree analysis. *Journal of Evaluation in Clinical Practice* 2018;24:353-361.

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

2155-0182/10/$3.00

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

18. Linden A, Butterworth SW. A comprehensive hospital-based Intervention to reduce readmissions for chronically ill patients: A randomized controlled trial. *American Journal of Managed Care* 2014;20(10):783-792.

19. Linden A, Yarnold PR. Using ODA in the evaluation of randomized controlled trials. *Optimal Data Analysis* 2018;7:46-49.

20. Linden A, Yarnold PR. Using ODA in the evaluation of randomized controlled trials: application to survival outcomes. *Optimal Data Analysis* 2018;7:50-53.

21. LaLonde R. Evaluating the econometric evaluations of training programs. *American Economic Review* 1986;76:604–620.

22. Dehejia RH, Wahba S. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics* 2002;84:151-61.

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

24. Yarnold PR, Soltysik RC. *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books, 2016. DOI: 10.13140/RG.2.1.1368.3286

25. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: An introduction to the bootstrap technique. *Disease Management and Health Outcomes* 2005;13:159-167.

26. Linden A, Adams J, Roberts N. The generalizability of disease management program results: getting from here to there. *Managed Care Interface* 2004;17:38-45.

27. Kleinbaum DG, Kupper LL, Muller KE. Applied Regression Analysis and Other Multivariable Methods. Boston, MA: PWS-Kent, 1988.

Author Notes

No conflict of interest was reported by either author.