Copyright 2020 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Implementing CTA from Within Stata: Identifying Causal Mechanisms in Interventions (*Invited*)

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

Mediation analysis identifies causal pathways by testing the relationships between the treatment, the outcome, and an intermediate variable that mediates the relationship between the treatment and the outcome. In this paper, I describe how the new Stata package for implementing CTA can be used to assess mediation effects.

Prior papers[1-4] introduced the new Stata package called **cta**[5] for implementing CTA from within the Stata environment. This package is a wrapper for the CTA software[6], thus the CTA64.exe file must be loaded on the computer for the **cta** package to work (CTA software is available at https://odajournal.com/resources/). To download the **cta** package, at the Stata command line type: “ssc install cta” (without the quotation marks).

This paper demonstrates how the **cta** package can be used to assess mediation effects in interventions. A common criticism of only estimating an overall treatment effect is this fails to reveal insights about the causal mechanism by which the intervention is hypothesized to influence the outcome. Mediation analysis overcomes this limitation by identifying intermediate variables that lie on the causal pathway and consequently mediate the relationship between treatment and outcome. Additionally, mediation analysis can serve to optimize an intervention by identifying activities that most reliably lead to the desired outcome.[7]

The CTA mediation strategy overcomes several limitations of the existing alternatives for mediation analysis, by not requiring any assumptions about the distribution of the mediator or of the outcome, or regarding the functional form of the model. Moreover, CTA will systematically identify a treatment-by-mediator interaction if it exists, as well as any other interaction between variables. Finally, the correct temporal relationships between treatment, mediator(s), and the outcome can be forced in CTA-mediation models.[8]

Generating a mediation model in **cta** is performed by specifying the outcome indicator as the *class* variable and all the covariates as *attributes*. To ensure the correct temporal relationship between variables (treatment » mediator » outcome), the treatment variable is specified in the *forcenode*() option at node 1.

Methods

Data

Data are from the Job Search Intervention Study (JOBS II), a randomized field experiment investigating the efficacy of a job training workshop on unemployed workers. Study investigators hypothesized that individuals attending the workshop (treatment) would increase their job search self-efficacy (mediator), which in turn would lead to lower depressive symptoms and increase reemployment (outcomes) compared to controls.[9]

A subset of the original data is used that includes 899 individuals (600 treated and 299 controls) and the following main variables: a binary treatment variable, a continuous scale measuring the level of job-search self-efficacy (mediator), and a binary variable representing whether the respondent had become employed at follow-up (outcome). The data also include the following baseline covariates: education, income, race, marital status, age, sex, previous occupation, the level of economic hardship, and preintervention level of depressive symptoms.

Analytic process

This example focuses on the case in which the treatment is a binary variable, the mediator is continuous, and the outcome is binary. See Linden and Yarnold[8] for a comprehensive discussion of other combinations.

Generating a CTA model

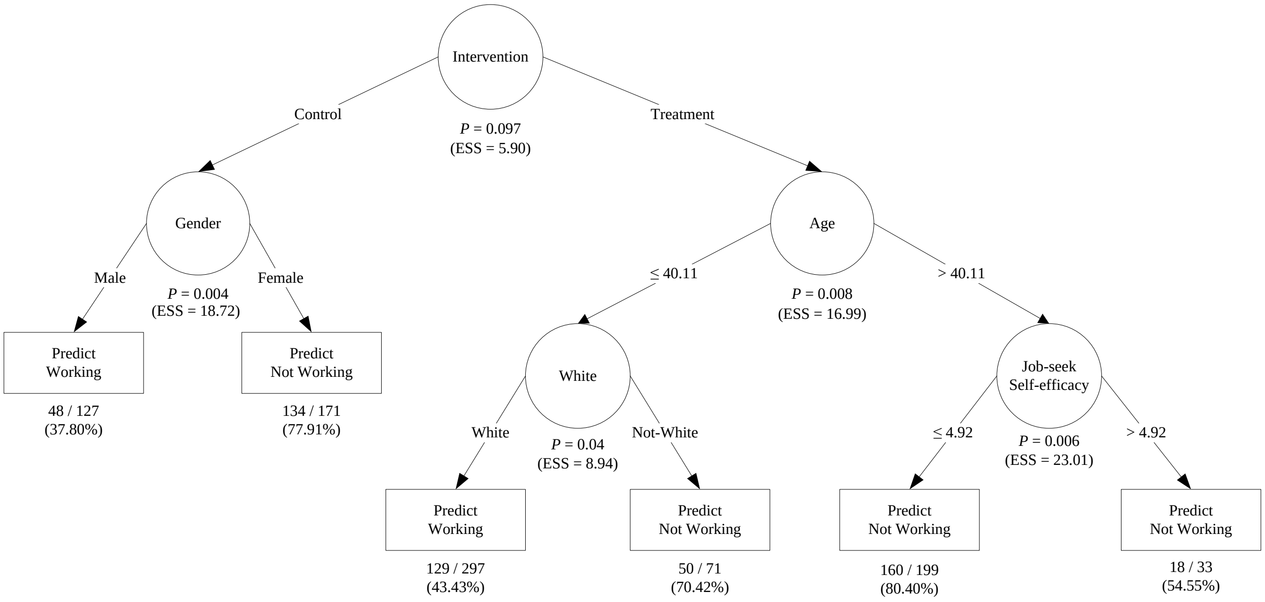
The following syntax is used to generate a mediation model using **cta** (see the help file for **cta** for a complete description of the syntax options):

```
cta work1 treat job_seek econ_hard
depress1 sex age occp marital nonwhite
educ income, pathcta("C:\ CTA\")
store("C:\ CTA\output") cat( treat sex
occp marital nonwhite) iter(10000)
stop(99) cutoff(0.10) prune(0.10)
forcenode(1 treat)
```

The above syntax is explained as follows: The outcome variable “work1” (reemployed) is specified as the *class* variable; the 11 variables listed until the comma are covariates specified as the *attributes*; the directory path where the CTA64.exe file is located on my computer is “C:\CTA\”; the directory path where the output and other files generated during the analysis should be stored is “C:\CTA\output”; the *cat*() option indicates the categorical attributes; the number of iterations (repetitions) for computing a permutation P-value is 10,000; the stopping rule is that the model stops when the 99% CI is reached or P-value of 0.10; the tree is pruned with a P-value of 0.10; and the treatment should be forced into the first node. Cutpoints are set at P=0.10 for an attribute to be included in the model to ensure the treatment variable will be included (from prior work[8] we know that the treatment effect is $P > 0.05$). Yarnold and Soltysik[6] provide a complete description of the CTA modeling process and interpretation of results.

The **cta** package produces an extract of the total output produced by CTA software (the complete output is stored in the specified directory with the extension “.out”). Here we include a diagram of the pruned model, which achieved overall weighted ESS of 23.32 (on the cusp of being considered a moderate effect).

In reviewing the diagram, it is evident that a complex model emerged in which increased job-seeking self-efficacy played a limited role in mediating the effect between treatment and reemployment. Specifically, individuals in the treatment group, older than 40.11 years, with perfect self-efficacy (i.e., > 4.92) had a 54.55% probability of being reemployed at follow-up.



As seen, the CTA model also identifies other pathways to the outcome in addition to the hypothesized mediated process. For example, individuals in the treatment condition who are ≤ 40.11 years of age and who are white have a 43.43% probability of reemployment at follow-up, compared to a 100% − 70.42% = 29.58% probability of reemployment for nonwhites. Males in the control condition had a 37.80% probability of reemployment at follow-up, versus 100% − 77.91% = 22.09% for females. These covariate interactions may be considered to be “moderators” of the treatment-outcome relationship. Interactions identified between the mediator and other covariates may be thought of as a moderated-mediator process.

Finally, in addition to maximizing the value of overall ESS, CTA models enable the investigator to ascertain the effect strength of every term in the model. For example, in the diagram, the root (initial) attribute in the model was the intervention assignment—which was not statistically significant (*P* = 0.097) and yielded very weak effect strength (ESS = 5.9). In contrast, for people in the control group (the left-hand branch from the intervention root variable), the statistically significant effect of gender (*P* < 0.004) had ESS = 18.72. While still relatively weak, this effect is nearly 3.2 times stronger than the effect of the intervention. The strongest such partial effect in the model, which also was relatively weak, occurred for job-seeking self-efficacy (ESS = 23.01).[8]

Discussion

This paper demonstrates how to generate mediation models in CTA model using the new Stata package **cta**. CTA provides accurate, parsimonious decision rules that are easy to visually display and interpret, while reporting *P* values derived via permutation tests at every node, in addition to corresponding partial ESS statistics. CTA is also insensitive to skewed data or outliers, and has the ability to handle any variable metric including categorical, Likert-type integer, and real number measurement scales. Moreover, CTA also has the unique ability to ascertain where optimal (maximum-accuracy) cutpoints are located on each variable, which in turn, facilitates the use of measures of predictive accuracy. And, CTA can perform cross-validation using LOO (among a host of reproducibility methods[6]) which allows for

Copyright 2020 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

assessing the cross-generalizability of the model to potentially new study participants or non-participants.[10]

Finally, the findings continue to support the recommendation to employ the ODA and CTA frameworks to evaluate the efficacy of health-improvement interventions and policy initiatives.[11-35]

References

[1]Linden A (2020). Implementing CTA from within Stata: Assessing the quality of the randomization process in randomized controlled trials (*Invited*). *Optimal Data Analysis, 9*, 57-62.

[2]Linden A (2020). Implementing CTA from within Stata: Characterizing participation in observational studies (*Invited*). *Optimal Data Analysis, 9*, 63-67.

[3]Linden A (2020). Implementing CTA from within Stata: Using CTA to generate propensity score weights (*Invited*). *Optimal Data Analysis, 9*, 68-73.

[4]Linden A (2020). Implementing CTA from within Stata: Modeling time-to-event data (*Invited*). *Optimal Data Analysis, 9*, 68-72.

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

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

[7]Linden A, Karlson KB (2013). Using mediation analysis to identify causal mechanisms in disease management interventions. *Health Services Outcomes Research Methodology, 13,* 86-108.

[8]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.

[9]Vinokur A, Schul Y (1997). Mastery and inoculation against setbacks as active ingredients in the jobs intervention for the unemployed. *Journal of Consulting and Clinical Psychology, 65,* 867-877.

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

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

[12]Linden A, Roberts N (2005). A User's guide to the disease management literature: recommendations for reporting and assessing program outcomes. *American Journal of Managed Care, 11,* 113-120.

[13]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).

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

[15]Yarnold PR, Linden A (2016). Theoretical aspects of the D statistic. *Optimal Data Analysis, 22,* 171-174.

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

[17]Linden A, Adams J, Roberts N (2004). Evaluating disease management program effectiveness: an introduction to survival analysis. *Disease Management*, *7*, 180-190.

[18]Linden A, Roberts N (2004). Disease management interventions: What's in the black box? *Disease Management*, *7,* 275-291.

[19]Linden A, Butterworth S, Roberts N (2006). Disease management interventions II: what else is in the black box? *Disease Management*, *9*, 73-85.

[20]Biuso TJ, Butterworth S, Linden A (2007). Targeting prediabetes with lifestyle, clinical and behavioral management interventions. *Disease Management*, *7*, 6-15.

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

[22]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.

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

[24]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.

[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.

[26]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.

[27]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.

[28]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.

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

[30]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.

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

[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.