Optimal Data Analysis
Vol. 9 (April 5, 2020), 79-83

2155-0182/10/$3.00

Implementing CTA from Within Stata: Modeling Time-to-Event (Survival) Data (*Invited*)

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

Health researchers frequently generate predictive models of time-to-event outcomes (e.g., death, onset of disease, hospital readmission) to assist clinicians to better understand the disease process and manage their patients. In this paper, I describe how the new Stata package for implementing CTA can be used to generate predictive models with time-to-event outcomes.

Prior papers[1-3] introduced the new Stata package called **cta**[4] for implementing CTA from within the Stata environment. This package is a wrapper for the CTA software[5], 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 generate predictive models with a time-to-event outcome such as death, onset of disease, or hospital readmission.[6,7] Time-to-event outcomes require specialized models designed to assess the influence of covariates on the outcome in the presence of *censoring*.[6] Survival times are called censored to indicate that the study terminated before the event occurred, or that the individual was lost to follow-up at some point during the study. Such models are an integral component of disease management.[8-12]

Generating a predictive model with a time-to-event outcome in **cta** is performed by specifying the outcome indicator (e.g., dead or alive at the end of follow-up) as the *class* variable, and all the covariates as *attributes*. To account for censoring, follow-up times are specified as a weight using the *wt*() option.

Methods

Data

I demonstrate the use of **cta** for survival analysis using a subset of data from the Framingham Heart Study, which has been collecting longitudinal data on residents of Framingham, Massachusetts since 1948, to gain insight into the epidemiology of coronary heart disease

2155-0182/10/$3.00

(CHD) and its risk factors. The data comprise 4,658 individuals free of CHD at their baseline exam and followed for up to 11,688 days (32 years). The variables include systolic and diastolic blood pressure (mmHg), age (years), serum cholesterol (mg/100 mL), body mass index (kg/m2), gender, follow-up time (days), and an indicator of whether the individual developed CHD or was otherwise censored. The original dataset had 4,699 observations, but for demonstrative purposes, only individuals with complete data were retained.

Analytic process

Splitting the sample

While it is not uncommon to see predictive models generated using the full available sample, in fact, the resultant models are not guaranteed to generalize to patients outside of that sample.[13] A well-accepted approach to test the generalizability of a predictive model is to first split the pooled data into two or more random sub-samples, generate a model using one sub-sample (called the "training" sample), and then test the accuracy of that model on the other sample[s] (called "testing" sample[s]). A generalizable model is one in which accuracy achieved in the testing sample is close to the accuracy achieved in the training sample.

For demonstrative purposes the pooled sample is split into two sub-samples using the Stata command **splitsample** using the following syntax:

```
splitsample, generate(sample) nsplit(2)
balance(chdfate)
```

The above syntax splits the data into two subsamples, generating a new variable called "sample" (with two values: 1 and 2), ensuring that the two sub-samples are balanced on the outcome "chdfate".

Generating a CTA model

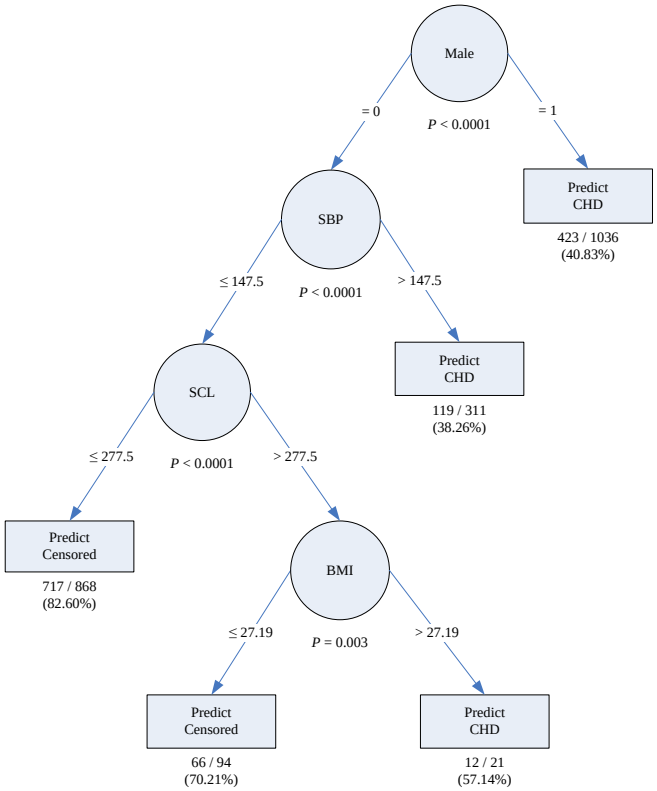
The following syntax is used to generate a predictive model using **cta** (see the help file for **cta** for a complete description of the syntax options):

```
cta chdfate male sbp dbp scl age bmi
if sample==1, pathcta("C:\CTA\")
store("C:\CTA\output") cat(male)
iter(10000) prune(0.05) enumerate
wt(followup)
```

The above syntax is explained as follows: The outcome variable is "chdfate" (dead or censored by day 11,688); the six variables listed until the comma are covariates specified as the *attributes*; the [if] statement limits the sample to the "training" sub-sample; 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 which attributes are categorical; the number of iterations (repetitions) for computing a permutation *P*-value is 10,000; the tree is pruned with a *P*-value of 0.05 used as the cutpoint for inclusion; an enumerated model (which enumerates the first three nodes) is conducted; and follow-up time is specified as the weight (Yarnold and Soltysik[5] 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 I include a diagram of the pruned model, which achieved overall weighted ESS of 24.54 (on the cusp of being a moderate effect)—which is slightly lower than achieved by the enumerated model (ESS=25.55), but is more parsimonious.

2155-0182/10/$3.00



Reviewing this diagram of the "training" sample, it is evident that patients predicted to develop CHD follow a different pathway than patients predicted to be either disease-free or censored at the end of follow-up. That is, the patients are predicted to develop CHD if: (1) they are male; (2) they are female with SBP > 147.5; or (3) they are female with SBP ≤ 147.5, SCL > 277.5, and BMI > 27.19. As seen, these pathways were all statistically significant with the largest *P* value < 0.003.

When applying the classification rules from this model to the "testing" sub-sample, the ESS = 22.98% indicating good generalizability to patients not included in the modeling process (the accuracy measures were computed using the package **classtabi**).[14]

Discussion

This paper demonstrates how to generate a predictive CTA model using the new Stata package **cta**. CTA provides accurate, parsimonious classification rules which 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 distinct ability to ascertain where optimal (maximum-accuracy) cutpoints exist on each variable, which in turn, facilitates the use of measures of predictive accuracy. Moreover, CTA can perform cross-validation using a variety of methodologies—in the present case using split-samples, which allows for assessing the cross-generalizability of the model to potentially new study participants or non-participants.[13]

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.[15-31]

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). CTA: Stata module for conducting Classification Tree Analysis. *Statistical Software Components S458729, Boston College Department of Economics.*

2155-0182/10/$3.00

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

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

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

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

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

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

[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 (2004). The generalizability of disease management program results: getting from here to there. *Managed Care Interface, 17*, 38-45.

[14]Linden A. (2020). CLASSTABI: Stata module for generating classification statistics and table for summarized data. *Statistical Software Components* S458127, *Boston College Department of Economics.*

[15]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).

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

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

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

[19]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.

[20]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.

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

[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]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

[24]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.

[25]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.

[26]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.

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

[28]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.

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

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

[31]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.