Optimal Data Analysis
Vol. 7 (March 17, 2018), 19-22

2155-0182/10/$3.00

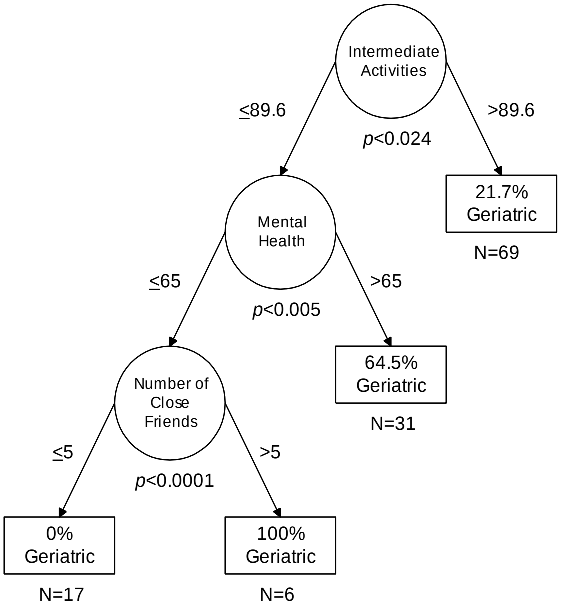
CTA Models and Staging Tables

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

This note reviews creation and use of staging tables for CTA models.

The initial CTA model discriminated geriatric (65 years of age or older) *vs*. non-geriatric adult ambulatory medical patients on the basis of self-reported well-being (Figure 1). Forty geriatric and 85 non-geriatric patients completed a survey which assessed five aspects of functional status [Basic Activities; Intermediate Activities; Mental Health (absence of depression); Social Activity; Quality of Social Interaction], and included five single-item measures assessing health satisfaction, physical limitations, and quantity of social interaction.[1]

Figure 1: CTA Model Discriminating Geriatric *vs*. Non-Geriatric Ambulatory Medical Patients



CTA Models

CTA models initiate with a *root node* from which at least two *branches* emerge. Branches are pathways through the tree, and ultimately terminate in model *endpoints* which represent sample *strata*. With respect to the attributes that CTA selects, sample observations are homogeneous within and heterogeneous between strata. CTA finds the model predicting the class variable (e.g., geriatric class) with maximum accuracy, assessed as effect strength for sensitivity or ESS. In every analysis ESS=0 is the level of accuracy which is expected by chance, and ESS=100 is perfect accuracy.[2-4]

In the schematic diagram the circles are nodes (attributes), arrows are branches (pathways), and rectangles are endpoints (unique patient strata). Numbers (words, for categorical attributes) adjacent to branches are the value (or categories) constituting the *optimal threshold* for the node. The number beneath a node is the associated generalized (per-comparison) permutation *p* value (all *p* in the model must satisfy a Bonferroni criterion[2] for experimentwise *p*<0.05). The number of observations classified into each endpoint (strata *N*) is given beneath each endpoint, and the percentage of class category=1 (here, geriatric) observations is indicated within each endpoint.

Using a CTA model to classify individual observations is rudimentary. For example, consider a hypothetical observation having an Intermediate Activities score = 88, a Mental Health score = 63, and 6 close friends. Starting with the first node, since the person's Intermediate Activities score is $\leq$ 89.6, the left branch is appropriate. At the second node the left branch is again appropriate because the person's Mental Health score is $\leq$ 65. Finally, at the third node the right branch is appropriate since the person has more than 5 close friends. The person is classified into the corresponding model endpoint: as seen, all six observations classified into this model endpoint were geriatric. The empirical probability of being geriatric in this endpoint is $p_{geriatric}$ = 1, but for prognostic purposes the probability of being geriatric in this endpoint is $p_{geriatric} \geq$ 6/7. In contrast, if the patient had reported $\leq$5 close friends, then the left-hand endpoint would be used, with $p_{geriatric}$ = 0 and $\leq$1/18, respectively.

Many intuitive aspects of CTA models are conceptually appealing.[5] For example model "coefficients" (optimal thresholds) consist of numerical values or category descriptions expressed in their natural measurement units. Using CTA models sample stratification unfolds in a "flow" process which is easily visualized across attributes in the schematic model. The manner in which CTA handles observations with missing data is also intuitive: while linear models drop observations missing data on any attributes used in the model, CTA only drops observations missing data on attributes actually used in their classification. Imagine for example an observation had an Intermediate Activities score > 89.6, and had missing data on Number of Close Friends and/or Mental Health: using a linear model this observation would be dropped, using CTA this observation would be classified. Within model endpoints, *percent* of strata that are members of class=1 is an intuitive, standardized metric. And, strata *N* (beneath model endpoints), and exact permutation *p* value (beneath model nodes) likewise are expressed using intuitive, standardized metrics.[3,5]

Staging Tables

An alternative representation of CTA models, *staging tables* are used to determine numerical "severity scores" for observations (Table 1). The *rows* of the staging table are CTA *model endpoints* arranged in increasing order of percent of class 1 membership. The *columns* of the staging table are the CTA *model nodes* starting with the root node in column two, followed by other attributes according to their relative depth in the tree (i.e., the attributes found deep in the tree appear in columns found on the right-hand side of the staging table).[2,3]

Table 1: Staging Table for Predicting Geriatric Likelihood

Stage	Intermediate Activities	Mental Health	Number of Close Friends	N	Odds	$p_{Geriatric}$
1	$\leq$ 89.6	$\leq$ 65	$\leq$ 5	17	$\leq$ 1:17	0.056*
2	> 89.6	---	---	69	2:7	0.217
3	$\leq$ 89.6	> 65	---	31	7:4	0.645
4	$\leq$ 89.6	$\leq$ 65	> 5	6	$\geq$ 6:1	0.857*

Note: Increasing scores on Intermediate Activities indicate *increasing* adaptability, and on Mental Health indicate *decreasing* depression. An asterisk indicates the endpoint had perfect classification so $p_{Geriatric}$ is based on minimum odds—otherwise *p* is empirically determined.

2155-0182/10/$3.00

With respect to the rows in Table 1, the first row (Stage 1) corresponds to the left-most endpoint in Figure 1, for which 0% of N=17 observations were geriatric; the second row (Stage 2) corresponds to the right-most endpoint, for which 21.7% of N=69 observations were geriatric; the third row (Stage 3) corresponds to the second-from-the-right endpoint, for which 64.5% of N=31 observations were geriatric; and the fourth row (Stage 4) corresponds to the second-from-the-left endpoint, for which 100% of N=6 observations were geriatric. Note that stage is an ordinal index of increasing likelihood of class=1 (here, geriatric) membership. Of course, had the class variable been mortality (instead of geriatric) status, then stage would have been an ordinal index of increasing likelihood of mortality. Were the class variable bankruptcy status, stage would have been an ordinal index of increasing likelihood of bankruptcy, etcetera.

And, with respect to the columns in Table 1, note that the root node of the CTA model (Intermediate Activities) is represented in column two; the middle node of the CTA model (Mental Health) is represented in column three; and the bottom-most node of the CTA model (Number of Close Friends) is represented in column four of the staging table.

After all of the model attributes have been included in the staging table, the following column gives N for each Stage (i.e., endpoint).

The second-to-last column of the staging table gives the odds of observations in a given Stage being geriatric (class=1).

If endpoint (Stage) classification was imperfect—less than 100% accurate, then the odds of class status=1 for the Stage is determined empirically. For example, Stage 2 in Table 1 corresponds to the right-most endpoint in Figure 1, for which the empirical probability of class=1 membership is 0.217 for N=69 observations. The tabled approximate odds, 2:7, corresponds to estimated probability of class=1 membership of 2/(2+7)=0.222 (an overestimate of 0.005). And, Stage 3 in Table 1 corresponds to the second-from-the-right endpoint in Figure 2, for which the empirical probability of class=1 membership is 0.645 for N=31 observations. The tabled approximate odds, 7:4, corresponds to estimated probability of class=1 membership of 7/(7+4)= 0.636 (an underestimate of 0.009).

When predictions made within a given endpoint are perfect—100% accurate, then computing the odds is done differently.[6] For example, Stage 1 in Table 1 had 17 observations all correctly predicted to be non-geriatric. Thus, the *highest* that the probability of being geriatric could be in this Stage is 1 of 18 observations—if the 18th observation in the sample who exactly satisfied the Stage 1 profile was incorrectly predicted to be non-geriatric. And, Stage 4 in Table 1 had 6 observations all correctly predicted to be geriatric: the *lowest* that the probability of being geriatric could be in this Stage is 6 of 7 observations (i.e., if the 7th observation in the sample was incorrectly predicted to be geriatric).

Obviously $p_{geriatric}$ is more precise than Stage. For example, when compared to Stage 1, $p_{geriatric}$ is approximately 0.217/0.056=3.875-times higher in Stage 2, 11.518-times higher in Stage 3, and 15.304-times higher in Stage 4. And, Stages 1 and 4 are identical except for Number of Close Friends: having six or more *vs*. five or fewer close friends corresponds to the 15.304-times higher likelihood of class=1 membership in Stage 4. Of course, if working with small samples the findings based on numerically small denominators may change dramatically with the addition of a few observations.

Using the staging table to estimate the likelihood (either Stage or $p_{geriatric}$) of a given observation being geriatric is straightforward: simply evaluate the fit between the data for the observation and each stage descriptor. Begin at Stage 1 and work sequentially through stages until identifying the descriptor that is *exactly true* for the observation undergoing staging. Consider the hypothetical person discussed earlier. Stage 1 does not fit because the person has >5 close friends. Stage 2 does not fit since

2155-0182/10/$3.00

the person’s Intermediate Activities score is ≤89.6. Stage 3 does not fit because the person’s Mental Health score is ≤65. However, since the hypothetical person has an Intermediate Activities score ≤89.6, Mental Health score ≤65, and >5 close friends, Stage 4 exactly fits the data of the hypothetical person.

References

[1]Yarnold PR (1996). Discriminating geriatric and non-geriatric patients using functional status information: An example of classification tree analysis via UniODA. *Educational and Psychological Measurement, 56*, 656-667.

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

[3]Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[4]Yarnold PR (2017). What is optimal data analysis? *Optimal Data Analysis*, *6*, 26-42.

[5]Linden A, Yarnold PR (2016). Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice*, *6*, 839-847.

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

Author Notes

This methodological article uses publically-available data and is exempt from IRB review. No conflict of interest was reported.