Optimal Data Analysis
Vol. 5 (July 9, 2016), 62-64

2155-0182/10/$3.00

How Many EO-CTA Models Exist in My Sample and Which is the Best Model?

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

As concerns the existence of statistically reliable enumerated-optimal classification tree analysis (EO-CTA) model(s) for a given application, possible alternative analytic outcomes are: no EO-CTA model exists; one model exists; or a descendant family (DF) that consists of two or more models exists. Models in a DF maximize ESS for unique partitions of the sample, and the model with the lowest observed *D* statistic is the globally-optimal CTA (GO-CTA) model for the application. The brute-force method of identifying a DF involves obtaining an initial EO-CTA model without specifying minimum endpoint sample size, then applying the minimum denominator selection algorithm (MDSA) to the initial model. A more efficient methodology for obtaining the GO-CTA model involves including only the attribute subset identified using structural decomposition analysis (SDA). The DF for the SDA attribute subset differs from the DF identified for the entire attribute set because the DF is data-specific. These methods are illustrated for an application using rated aspects of nursing and physician care to discriminate 1,045 very satisfied *vs.* 671 satisfied Emergency Department (ED) patients.

The modus operandi when using legacy analytic methods in applied research involves reporting a featured statistical model that meets methodological criteria and substantive aspects of the research.[1,2] However, novometric theory asserts that a DF consisting of multiple optimal models may exist in a given application.[3] Models in the DF vary in relative parsimony (number of strata) and accuracy (ESS), that together define the *D* statistic that indicates the number of equivalent effects that are needed to obtain a theoretically ideal statistical model.[3] Models constituting the DF are identified by the MDSA, and selection of the featured model is based on analytic (*D*) as well as substantive (qualitative) model features.

Identifying GO-CTA Model by Brute Force

Data for exposition of these methods were obtained from patients receiving care in the ED of a private Midwestern hospital, who were mailed and returned a completed satisfaction survey. Items were answered using a five-point Likert-type scale: 1=very poor, 2=poor, 3=fair, 4=good, 5=very good. The class variable (sat) was overall satisfaction with care received in the ED: data from 671 patients rating overall

2155-0182/10/$3.00

satisfaction as good were compared with data from 1,045 patients rating overall satisfaction as very good. Attributes were patient ratings of six aspects of nurse (N1=courtesy; N2=took the patient's problem seriously; N3=attention paid to patient; N4=concern to keep patient informed; N5=concern for patient privacy; N6=technical skill) and physician (P1=waiting time in treatment area; P2=courtesy; P3=took patient's problem seriously; P4=concern for patient comfort: P5=explanation of test/treatment; P6=explanation of illness/injury) patient-care behavior.[4]

The first step of MDSA is accomplished using CTA software[3] command syntax below (although the default setting for the software explicitly maximizes ESS, this may also be manually specified by setting prior odds on[3]):

```
OUTPUT satis.out;
OPEN satis.dat;
VARS sat n1 TO n6 p1 TO p6;
CLASS sat;
ATTR n1 TO n6 p1 TO p6;
MISSING all (-9);
PRIORS ON;
MC ITER 10000 CUTOFF .05 STOP 99.9;
PRUNE .05;
ENUMERATE;
GO;
```

CTA software output lists the EO-CTA model after pruning by a sequentially-rejective Sidak Bonferroni-type multiple comparisons procedure used to ensure experimentwise $p < 0.05$, as well as by a pruning algorithm that ensures explicitly maximized ESS.[3] All attributes in this study were stable in leave-one-out validity analysis.[3,5,6] In the first step of MDSA no minimum strata size ("denominator") is specified. Presently the smallest N in any of the initial ("unrestricted") EO-CTA model endpoints was 6 observations (Table 1), so the second model in the DF is identified by forcing the CTA algorithm to increase the minimum endpoint N by at least one observation:

```
MINDENOM 7;
GO;
```

The MDSA continues until the minimum denominator is so large that no EO-CTA model can be identified. The DF identified presently is summarized in Table 1.

Table 1: MDSA Discriminating Very- *vs*. Moderately-Satisfied Patients: 12 Attributes

Step	Strata	MinD	ESS	Eff	PAC	*D*	Secs
1	12	5	69.6	5.80	85.1	5.25	234
2	10	15	69.2	6.92	84.8	4.44	234
3	10	20	68.7	6.87	84.5	4.55	232
4	9	22	67.9	7.54	84.4	4.26	228
5	9	27	67.4	7.49	83.4	4.35	228
6	8	56	67.2	8.40	84.2	3.90	226
7	8	62	66.2	8.33	83.8	4.01	215
8	6	64	66.1	11.0	83.0	3.08	215
9	5	83	65.2	13.1	81.5	2.67	214
10	5	104	64.6	12.9	81.4	2.75	208
11	5	112	64.1	12.8	81.3	2.80	202
12	4	115	63.0	15.8	82.4	2.35	196
13	4	150	61.7	15.4	80.6	2.48	195
14	4	261	61.1	15.3	80.4	2.54	188
15	4	265	60.1	15.0	80.0	2.66	161
16	3	360	59.9	20.0	77.9	2.01	159
17	3	370	59.6	19.9	77.8	2.04	144
18	3	375	58.6	19.5	77.2	2.12	141
19	2	769	56.4	28.2	77.3	1.55	133
20	0	---	---	---	---	---	7

Note: There were 20 steps in this brute-force MDSA. Strata is the number of endpoints identified by the EO-CTA model. MinD is the smallest number of observations (patients) in any of the strata (i.e., the smallest model endpoint N). ESS is a normed index of classification accuracy on which 0 represents the level of accuracy expected by chance and 100 represents perfect (errorless) classification: by rule-of-thumb: ESS≤25 is a relatively weak effect; ESS≤50 is a moderate effect; ESS≤75 is a relatively strong effect; and ESS>75 is a strong effect.[3] Efficiency, an index of parsimony, is ESS/number of strata. PAC (percent accurate classification) is the overall proportion of the sample correctly predicted by the model. *D* is the number of additional equivalent effects needed to achieve a theoretically ideal model for this application.[3] Secs is number of CPU seconds required to identify the indicated EO-CTA model using CTA software running on a 3 GHz Intel Pentium D microcomputer.

For this application the DF consists of 19 EO-CTA models. If an arbitrary minimum endpoint sample size is specified then analysis returns the first EO-CTA model in the DF with a minimum endpoint sample size that is equal to or larger than the arbitrary specified value. For

2155-0182/10/$3.00

example, if MINDENOM is set at 34 then the EO-CTA model in Step 6 will be obtained. For MINDEMON set at 178 the EO-CTA model in Step 14 will be obtained. For any MINDENOM greater than 769 no EO-CTA model exists. A total of 3,760 CPU seconds were used in identifying the DF using brute force.

The two-strata model identified in step 19 of the MDSA (Table 1) had the smallest *D* statistic, and thus is the GO-CTA model for this application.

Identifying GO-CTA Model by SDA

Described elsewhere, SDA is the maximum-accuracy conceptual analogue of principal components analysis (PCA): whereas PCA identifies the attributes defining factors that maximize explained *variation* (indexed by the eigenvalue), SDA identifies the attributes defining GO-CTA models that maximize classification *accuracy* (indexed by ESS).[1,2] When performed with the present sample, SDA selected three attributes: n3, p1, p3. According to the second axiom of novometric theory, the GO-CTA model for the original set of 12 attributes (six each for nurses and physicians) is obtained by applying the MSDA to the attributes identified by SDA.[1] The DF identified for the subset of three attributes is summarized in Table 2.

Table 2: MDSA Discriminating Very- *vs*. Moderately-Satisfied Patients: 3 Attributes

Step	Strata	MinD	ESS	Eff	PAC	*D*	Secs
1	8	26	62.3	7.79	82.1	4.84	16
2	6	29	61.0	10.2	79.8	3.83	12
3	4	120	60.6	15.2	81.3	2.60	13
4	4	261	59.8	14.9	79.7	2.69	12
5	3	370	59.6	19.9	77.8	2.04	11
6	2	769	56.4	28.2	77.3	1.55	10
7	0	---	---	---	---	---	2

Note: There were 7 steps in this MDSA using attributes identified by SDA. The EO-CTA model in step 5 here is the same as the model in step 17 in Table 1, and the GO-CTA model in step 6 here is the same as the GO-CTA model in step 19 in Table 1. See notes to Table 1.

The GO-CTA model identified in step 19 of the MDSA without SDA (Table 1) is identical to the GO-CTA model identified in step 6 of the MDSA with SDA (Table 2). A total of 76 CPU seconds were used in identifying the DF using SDA prior to MDSA.

References

[1]Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago: ODA Books.

[2]Grimm LG, Yarnold PR (Eds.). *Reading and Understanding Multivariate Statistics*. Washington, DC: APA Books, 1995.

[3]Grimm LG, Yarnold PR (Eds.). *Reading and Understanding More Multivariate Statistics*. Washington, DC: APA Books, 2000.

[4]Yarnold PR (2014). What most satisfies Emergency Department patients? *Optimal Data Analysis, 3*, 98-101.

[5]Yarnold PR (2016). Using UniODA to determine the ESS of a CTA model in LOO analysis. *Optimal Data Analysis, 5*, 3-10.

[6]Yarnold PR (2016). Determining jackknife ESS for a CTA model with chaotic instability. *Optimal Data Analysis, 5*, 11-14.

Author Notes

The study analyzed de-individuated data and was exempt from Institutional Review Board review. No conflict of interest was reported.

Mail: Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., #163
Chicago, IL 60646
USA