Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010), 46-49

2155-0182/10/$3.00

Aggregated *vs*. Referenced Categorical Attributes in UniODA and CTA

Paul R. Yarnold, Ph.D. and Robert C. Soltysik, M.S.
Optimal Data Analysis, LLC

Multivariable linear methods such as logistic regression analysis, discriminant analysis, or multiple regression analysis, for example, directly incorporate binary categorical attributes into their solution. However, for categorical attributes having more than two levels, each level must first be individually dummy-coded, then one level must be selected for use as a reference category and omitted from analysis. Selection of one or another level as the reference category can mask effects which otherwise would have materialized, if a different level had been chosen. Neither UniODA nor CTA require reference categories in analysis using multicategorical attributes.

Using a categorical attribute with three or more levels in a linear multivariable analysis requires separately dummy-coding each level, selecting one level as a reference category, and omitting it from analysis.[1] For example, imagine that a study assessed three ethnic categories: Navajo, Sumatran, and Inuit. Preparing this attribute for linear analysis first requires creating three new binary attributes: [a] Navajo (1) *vs*. others (0); [b] Sumatran (1) *vs*. others (0); and [c] Inuit (1) *vs*. others (0). Only two of the dummy-variables can be used as attributes in analysis, and one's choice can mask an effect depending on which class is selected as reference category. As an increasing number of polychotomous attributes are used, the associated design matrix becomes massive rapidly, increasing the likelihood of sparse or empty cells, imbalanced marginal distributions and nonnormality, toxic properties for linear methods. In addition to possibly masking effects, inducing numerical instability, undermining assumptions underlying the validity of *p*, and contributing to overdetermined models, the use of reference categories is also antithetical to the axiom of parsimony. Finally, in computer-intensive methods such as CTA, a larger number of attributes increases both memory and time resources needed to obtain an optimal solution.

In contrast, UniODA[2] and CTA[3] use aggregated multicategory attributes. Using the current example one "ethnicity" attribute having three levels (rather than three ethnicity attributes each having two levels) requires coding: Navajo (1), Sumatran (2), or Inuit (3).

This paper illustrates some advantages of using aggregated attributes in both bivariate (UniODA) and multivariable (CTA) analyses, using an application involving predicting use of mechanical ventilation for hospitalized patients with *Pneumocystis cariini* pneumonia (PCP).[4]

2155-0182/10/$3.00

UniODA

The analysis selected for exposition contrasts intubation rate for a total sample of 1,211 patients hospitalized for PCP in Chicago, Los Angeles, Miami, New York, and Seattle. The first analysis used the aggregated attribute, arbitrarily using dummy-codes of 1-5 for cities, respectively. The resulting UniODA model was: if city=Los Angeles or Chicago then predict a higher ventilation rate; otherwise predict a lower ventilation rate. This model correctly classified 54.9% of 1,418 non-ventilated, and 61.9% of 147 ventilated patients, yielding a relatively weak ESS=16.8 (p<0.0006), which was stable in jackknife validity analysis.

Using the aggregated city attribute and therefore one test of a statistical hypothesis, UniODA determined three cities have a lower ventilation rate than two other cities, and even though the effect is statistically significant and likely to cross-generalize for an independent random sample, the effect is weak, reflecting only 16.8% of the gain in accuracy theoretically possible to achieve beyond chance.

UniODA was next used to assess the ability of all five binary city attributes to predict ventilation: the test for Los Angeles (p<0.0006) alone achieved the criterion[2] for statistical significance with a weak effect of ESS=12.6. This result indicates that Los Angeles had a higher ventilation rate than the other four cities. Five tests of statistical hypotheses were conducted in reaching this conclusion, and must be accounted for in assessing the statistical significance of all hypothesis tests conducted within the study.

CTA

In the original research from which the example was drawn, ventilation was modeled by logistic regression analysis.[4] Predictive factors which emerged included a PCP severity score developed previously via CTA[6], location (Los Angeles), ethnicity (African-American), and a cytological confirmation of PCP diagnosis. For clarity in exposition, the same attributes selected by logistic regression were modeled presently. Algorithmic CTA[3] was run via ODA automated CTA software, using a minimum endpoint denominator of N=25 to ensure adequate statistical power.[7]

The first analysis used aggregated race and city attributes. The "aggregated attributes" model selected three attributes, and correctly classified 66.4% of intubated and 68.1% of non-intubated patients, yielding a moderate effect: ESS=34.5 (Figure 1).

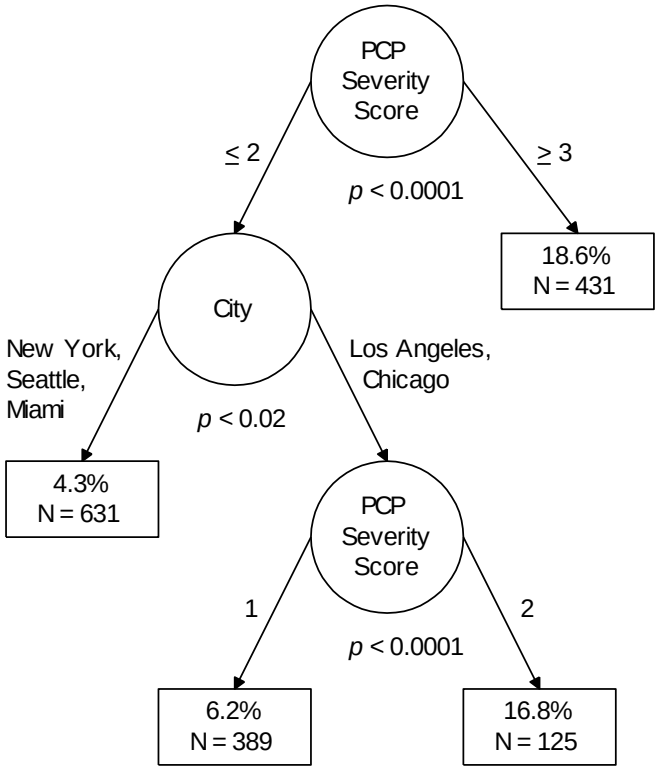

Figure 1: CTA Intubation Model using Aggregated Race and City Attributes

The second analysis used individually dummy-coded race and city attributes, although unlike linear models which require omission of a reference attribute from analysis, with CTA all of the binary attributes compete for admission to the model. The "separately coded attributes"

2155-0182/10/$3.00

model selected five attributes and correctly classified 78.0% of intubated and 57.0% of non-intubated patients, yielding a moderate effect: ESS=35.0 (Figure 2).

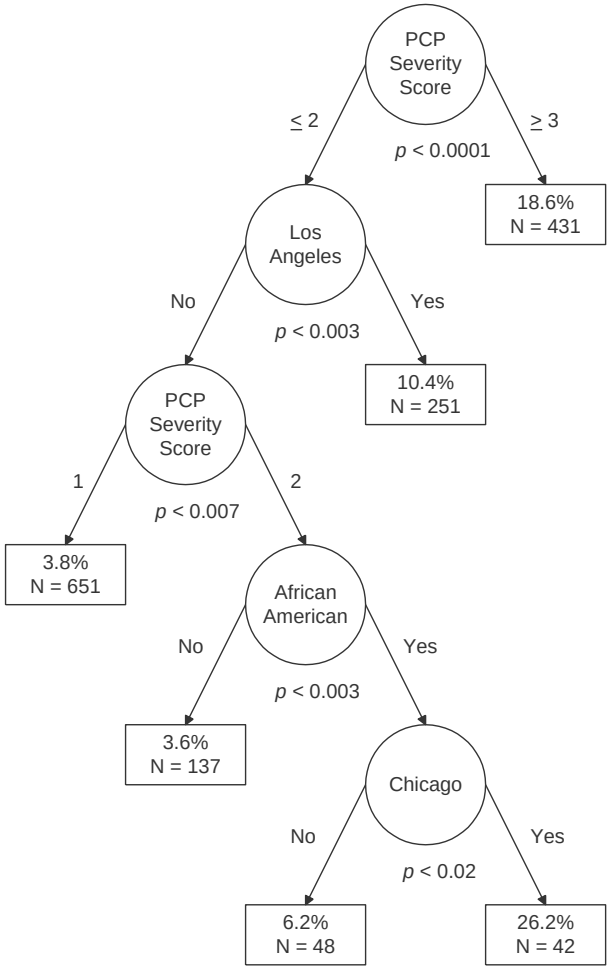

Figure 2: CTA Intubation Model using Separately Coded Race and City Attributes

The models selected the same attributes except for one separately-coded race attribute. The aggregated attributes model employed one attribute to model city, and achieved an overall model efficiency=34.5/3 or 11.5 ESS units-per-attribute. In contrast, the separately-coded attributes model used two attributes to model city, and achieved an overall model efficiency= 35.5/5 or 7.1 ESS units-per-attribute). Thus, the aggregated attributes model is 62% more efficient than the separately-coded attributes model.

Note that the final "Chicago" attribute in the separately-coded attributes CTA model was retained on the basis of model-wise Bonferroni criterion.[2] However, had one additional test of a statistical hypothesis been conducted (e.g., as in any random typical published study), then the Chicago attribute would have been pruned from the model.

Yet another advantage of parsimonious CTA models is that by having fewer endpoints into which observations are partitioned, the minimum endpoint denominators may be larger. Presently, the minimum endpoint denominator for the aggregated attributes model (N=125) is nearly three times larger than for the separately-coded attributes model (N=42). Estimates for the aggregated attributes model are thus more robust over sampling anomalies and likely to cross-generalize, especially for smaller samples.

Using a 3 GHz Intel Pentium D microcomputer, the separately-coded attributes model required 78 CPU seconds to solve, 34.5% more than the 58 CPU seconds required to solve the aggregated attributes model. These problems were relatively simple for automated CTA software to solve, so computing efficiency gained by using aggregated categorical attributes was relatively modest compared to gains obtained in complex analyses. Presently, for example, enumerated CTA models (not shown) involving aggregated (1,394 CPU seconds) or separately-coded (4,054 CPU seconds) attributes revealed a 190.8% gain in computing efficiency.

References

[1]Kleinbaum DG, Kupper LL, Nizam A, Muller KE. *Applied regression analysis and other multivariable methods* (4th Ed.). Thomson Higher Education, Belmont, Ca, 2008.

[2]Yarnold PR, Soltysik RC. *Optimal data analysis: a guidebook with software for Windows*. APA Books, Washington DC, 2005.

2155-0182/10/$3.00

[3]Soltysik RC, Yarnold PR. Introduction to automated CTA software. *Optimal Data Analysis*, In press.

[4]Curtis JR, Yarnold PR, Schwartz DN, Weinstein RA, Bennett CL. Improvements in outcomes of acute respiratory failure for patients with human immunodeficiency virus-related *Pneumocystis carinii* pneumonia. *American Journal of Respiratory and Critical Care Medicine* 2000, 162: 393-398.

[5]Yarnold PR. Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement* 1996, 56:430-442.

[6]Arozullah AM, Yarnold PR, Weinstein RA, Nwadiaro N, McIlraith TB, *et al*. A new pre-admission staging system for predicting in-patient mortality from HIV-associated *Pneumocystis carinii* pneumonia in the early-HAART era. *American Journal of Respiratory and Critical Care Medicine* 2000, 161:1081-1086.

[7]Yarnold PR, Soltysik RC. Statistical power analysis for UniODA and CTA. *Optimal Data Analysis*, In press.

Author Notes

Address correspondence to authors at: Optimal Data Analysis 1220 Rosecrans St., Suite 330, San Diego, CA 92106. Send E-mail to: Journal@OptimalDataAnalysis.com.