Optimal Data Analysis
Vol. 9 (April 25, 2020), 141-145

2155-0182/10/$3.00

Implementing ODA from Within Stata: Directional Hypothesis, Binary Class Variable, Ordinal Attribute

Paul R. Yarnold, Ph.D. and Ariel Linden, Dr.P.H.
Optimal Data Analysis, LLC Linden Consulting Group, LLC

This paper describes how a confirmatory (a priori, directional, one-tailed) hypothesis involving a binary (dichotomous) class variable and a five-level ordinal attribute is evaluated using MegaODA software via the new Stata package implementing ODA analysis.

Recent papers[1-15] introduce the new Stata package called **oda**[16] for implementing ODA from within the Stata environment. Because this package is a wrapper for the MegaODA software system[17-19], the MegaODA.exe file must be loaded on the computer for the **oda** package to work (MegaODA software is available at https://odajournal.com/resources/). To download the **oda** package, at the Stata command line type: "ssc install oda" (without the quotation marks). This paper demonstrates use of the **oda** package to evaluate a confirmatory hypothesis involving a binary class variable, and a five-level ordinal attribute.

Methods

Data

We consider data from Hyde and Plant (1995) comparing the relative strength of gender *vs*. other effects in psychology.[20] Arbitrary dummy-codes identified two types of *study*: gender=1, other=2 (use of this latter category can induce paradoxical confounding[21]). *Strength* of study outcome was rated using a five-category ordinal scale created by a statistically unmotivated[22] parse of Cohen's *d* statistic: 1=0-0.10; 2=0.11-0.35; 3=0.36-0.65; 4=0.66-1.0; and 5=over 1.0. Data for every study was entered in free format on a separate line as space-delimited text (ASCII) characters.[23]

Analytic Process

We repeat the ODA analysis previously conducted on these data (see example 5.8, *Optimal Data Analysis: A Guidebook with Software for Windows*[24]). The directional or "one-tailed" alternative hypothesis is that the binary class ("dependent") variable *study* can be discriminated on the basis of *strength* (ordinal attribute or "independent variable"): it is hypothesized that gender studies will have weaker effects. The null hypothesis is that this is not true. Weighting by prior odds (the default setting) identifies the model which maximizes ESS (i.e.,

2155-0182/10/$3.00

classification accuracy normed *vs*. chance), and a total of 25,000 Monte Carlo iterations are used to estimate Type I error (i.e., *p* value).[24]

For these data, **oda** is implemented using the following syntax to test the *a priori* hypothesis (see the **oda** help file for a complete description of syntax options):

```
oda study strength, pathoda("C:\ODA\")
store("C:\ ODA\output") iter(25000)
dir(< 1 2)
```

The above syntax is explained as follows: The variable "study" is the *class* variable; the variable "strength" is the *attribute*; the directory path where the MegaODA.exe file is located on the computer is "C:\ODA\"; the directory path where the output and other files generated during the analysis are stored is "C:\ODA\output"; 25,000 iterations (repetitions) are used to compute the permutation *p*-value; and it is hypothesized that class 1 studies should have lower strength scores than class 2 studies.

The **oda** package produces an extract of the total output produced by the ODA software (the complete output is stored in the specified directory with the extension ".out").

As seen in the **oda** output, the ODA model is interpreted as follows: "if *strength* ≤ 2.5 then predict *study* = 1; otherwise, predict *study* = 2." Studies of gender differences were accurately predicted to have lower Cohen's *d* scores than studies of other differences: the model correctly classified 60.23% of studies of gender differences, and 64.90% of other studies.

Effect strength for sensitivity (ESS) is labelled in the output as "Effect Strength PAC" (Percentage Accurate Classification). The ESS of 25.13% is marginally above the minimal criterion for an effect of moderate strength[24] and it has an associated permutation *p*<0.0001. In summary, ODA identified a model which discriminated between gender *vs*. other studies with moderate strength, and this finding was statistically significant.

```
ODA model:
----------
IF STRENGTH <= 2.5 THEN STUDY = 1
IF 2.5 < STRENGTH THEN STUDY = 2


Summary for Class STUDY  Attribute STRENGTH
------------------------------------------

Performance Index          Train
-----------------         ------
Overall Accuracy           63.21%
PAC STUDY=1                60.23%
PAC STUDY=2                64.90%
Effect Strength PAC        25.13%
PV STUDY=1                 49.28%
PV STUDY=2                 74.24%
Effect Strength PV         23.52%
Effect Strength Total      24.33%


Monte Carlo summary (Fisher randomization):
------------------------------------------
Iterations:  25000
Estimated p: 0.000000
```

Discussion

This paper shows how to use ODA to identify the model that maximally discriminates between any two categories of a class variable using a categorical ordinal attribute.

ODA should be considered the preferred approach over other methods because it avoids statistical assumptions required of conventional models, is 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.[24] Moreover, in contrast to other methods, ODA also has the unique ability to ascertain optimal (maximum-accuracy) assignments (categorical attributes) or cutpoints (ordered attributes) on the attribute, which facilitates the use of measures of predictive accuracy. Furthermore, ODA can perform cross-validation using LOO (and many other methods[24]) which allows for assessment of potential cross-generalizability of the model to independent random samples.

For these reasons we recommend that researchers employ ODA and CTA frameworks to evaluate the statistical hypotheses which are explored in their laboratory and field research endeavors.[25-39]

2155-0182/10/$3.00

References

[1]Linden A (2020). Implementing ODA from within Stata: An application to data from a randomized controlled trial (*Invited*). *Optimal Data Analysis*, *9*, 9-13.

[2]Linden A (2020). Implementing ODA from within Stata: Implementing ODA from within Stata: An application to estimating treatment effects using observational data (*Invited*). *Optimal Data Analysis*, *9*, 14-20.

[3]Linden A (2020). Implementing ODA from within Stata: An application to dose-response relationships (*Invited*). *Optimal Data Analysis*, *9*, 26-32.

[4]Linden A (2020). Implementing ODA from within Stata: assessing covariate balance in observational studies (*Invited*). *Optimal Data Analysis*, *9*, 33-38.

[5]Linden A (2020). Implementing ODA from within Stata: Evaluating treatment effects for survival (time-to-event) outcomes (*Invited*). *Optimal Data Analysis*, *9*, 39-44.

[6]Linden A (2020). Implementing ODA from within Stata: Evaluating treatment effects in multiple-group interrupted time series analysis (Invited). *Optimal Data Analysis*, *9*, 45-50.

[7]Linden A (2020). Implementing ODA from within Stata: identifying structural breaks in single-group interrupted time series designs (Invited). *Optimal Data Analysis*, *9*, 51-56.

[8]Linden A (2020). Implementing ODA from within Stata: Finding the optimal cut-point of a diagnostic test or index (Invited). *Optimal Data Analysis*, *9*, 74-78.

[9]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Exploratory hypothesis, binary class variable, and binary attribute. *Optimal Data Analysis*, *9*, 94-98.

[10]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Confirmatory hypothesis, binary class variable, and binary attribute. *Optimal Data Analysis*, *9*, 99-103.

[11]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Exploratory hypothesis, binary class variable, and binary attribute. *Optimal Data Analysis*, *9*, 104-108.

[12]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Exploratory hypothesis, binary class variable, and ordinal (rank) attribute. *Optimal Data Analysis*, *9*, 109-113.

[13]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Confirmatory hypothesis, binary class variable, and ordinal attribute. *Optimal Data Analysis*, *9*, 128-132.

[14]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Exploratory hypothesis, binary class variable, categorical ordinal attribute. *Optimal Data Analysis*, *9*, 133-136.

[15]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Nondirectional hypothesis, binary class variable, categorical ordinal attribute. *Optimal Data Analysis*, *9*, 137-140.

[16]Linden A (2020). ODA: Stata module for conducting Optimal Discriminant Analysis. *Statistical Software Components S458728, Boston College Department of Economics*.

[17]Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Separating the chaff. *Optimal Data Analysis*, *2*, 194-197.

[18]Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Harvesting the Wheat. *Optimal Data Analysis*, *2*, 202-205.

2155-0182/10/$3.00

[19]Yarnold PR, Soltysik RC (2013). MegaODA large sample and BIG DATA time trials: Maximum velocity analysis. *Optimal Data Analysis, 2*, 220-221.

[20]Hyde JS, Plant EA (1995). Magnitude of psychological gender differences: Another side to the story. *American Psychologist, 50,* 159-161.

[21]Yarnold PR (2018). Minimize usage of binary measurement scales in rigorous classical research. *Optimal Data Analysis, 7,* 3-9.

[22]Yarnold PR (2014). "Breaking-up" an ordinal variable can reduce model classification accuracy. *Optimal Data Analysis*, *3*, 19.

[23]Bryant FB, Harrison PR (2013). How to create an ASCII input data file for UniODA and CTA software (*Invited*). *Optimal Data Analysis, 2*, 2-6.

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

[25]Rhodes NJ (2020). Statistical power analysis in ODA, CTA and Novometrics (Invited). *Optimal Data Analysis*, *9*, 21-25.

[26]Yarnold PR, Linden A. (2016). Novometric analysis with ordered class variables: The optimal alternative to linear regression analysis, *Optimal Data Analysis*, *5*, 65-73.

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

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

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

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

[31]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.

[32]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.

[33]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.

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

[35]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.

[36]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.

2155-0182/10/$3.00

[37]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.

[38]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.

[39]Yarnold PR, Rhodes NJ, Linden A (2020). Selecting an appropriate weighting strategy in maximum-accuracy time-to-event (survival) analysis. *Optimal Data Analysis*, *9*, 3-6.

Author Notes

No conflicts of interest were reported.