Optimal Data Analysis
Vol. 9 (April 15, 2020), 94-98

2155-0182/10/$3.00

Implementing ODA from Within Stata: Exploratory Hypothesis, Binary Class Variable, and Binary Attribute

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

This paper describes how an exploratory (i.e., post hoc, nondirectional, or two-tailed) hypothesis involving a binary (i.e., dichotomous) class variable and a binary attribute can be evaluated using MegaODA software vis-à-vis the new Stata package for implementing ODA analysis.

Recent papers[1-8] introduced the new Stata package called **oda**[9] for implementing ODA from within the Stata environment. Because this package is a wrapper for the MegaODA software system[10-12], 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).

In this paper, we demonstrate how to use the **oda** package to evaluate a nondirectional hypothesis involving a single binary class variable, and a single binary attribute.

Methods

Data

Using Newmark’s random samples of wasps and honeybees, we evaluate the exploratory hypothesis that the proportion of males and females differs between species.[13] Newmark’s data are summarized in the following table.

Gender	Wasp	Honeybee
Female	40	103
Male	1,600	1,748

Analytic Process

We repeat the ODA analysis previously performed on these data (see example 5.1, *Optimal Data Analysis: A Guidebook with Software for Windows*[14]). The nondirectional or “two-tailed” alternative hypothesis is that the binary class (or “dependent”) variable *gender* can be discriminated of the basis of the binary attribute (or “independent variable”) *species*, and the null hypothesis is that this is not true. Arbitrary dummy-codes were used to identify categories of gender (female=0, male=1) and of species (wasp=1, honeybee=2), and the data for each observation was listed on a separate line using space-delimited text (ASCII) characters[15].

2155-0182/10/$3.00

LOO analysis was conducted to assess potential cross-generalizability of the ODA model when used to classify observations other than those in the original study sample.[14]

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

oda sex insect, pathoda("C:\ ODA\")
store("C:\ODA\output") iter(25000) loo cat

The above syntax is explained as follows: The variable "sex" is the *class* variable; the variable "insect" is the *attribute*; the directory path where the megaODA.exe file is located on my computer is "C:\ODA\"; the directory path where the output and other files generated during the analysis should be stored is "C:\ODA\output"; the number of iterations (repetitions) for computing the permutation *p*-value is 25,000; leave-one-out (LOO) analysis should be performed; and the attribute is categorical.

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 insect= 1 (wasp), then predict sex=1 (male). If insect=0, then predict sex=1(female)."

The effect strength for sensitivity (ESS, the classification accuracy normed *vs*. chance) is labelled in the output as "Effect Strength PAC" (Percentage Accurate Classification). In training as well as LOO analysis the ESS is 19.82% (a relatively weak effect).[14] The permutation *p*-value in both the training and LOO analysis was < 0.0001. In summary, ODA was able to find a model discriminating relatively weakly between wasps (which have a larger proportion of males) and honeybees (which have a larger proportion of females), which is likely to cross-generalize to an independent random sample, and that was statistically significant.

```
ODA model:
----------
IF INSECT = 1 THEN SEX = 1
IF INSECT = 2 THEN SEX = 0

Summary for Class SEX  Attribute INSECT
--------------------------------------

Performance Index           Train    LOO
-----------------          ------  ------
Overall Accuracy           48.78%  48.78%
PAC SEX=0                  72.03%  72.03%
PAC SEX=1                  47.79%  47.79%
Effect Strength PAC        19.82%  19.82%
PV SEX=0                    5.56%   5.56%
PV SEX=1                   97.56%  97.56%
Effect Strength PV          3.13%   3.13%
Effect Strength Total      11.47%  11.47%

Monte Carlo summary (Fisher randomization):
------------------------------------------
Iterations:  25000
Estimated p: 0.000040

Results of leave-one-out analysis
---------------------------------
3491 observations

Fisher's exact test (directional) classification table  p = .000002
```

Discussion

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

ODA should be considered the preferred approach over other methods because it avoids many of the 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.[14] Moreover, in contrast to other methods, ODA also has the distinct ability to ascertain the optimal (i.e., 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 other methods[14]) which allows for the 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.[16-52]

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

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

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

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

[13]Newmark J (1983). *Statistics and probability in modern life* (3rd ed.). Philadelphia: Saunders.

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

[15]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.

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

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

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

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

2155-0182/10/$3.00

[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 (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.

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

[23]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.

[24]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.

[25]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.

[26]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.

[27]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.

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

[29]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.

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

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

[32]Linden A, Yarnold PR (2018). Using ODA in the evaluation of randomized controlled trials: Application to survival outcomes. *Optimal Data Analysis*, *7*, 50-53.

[33]Yarnold PR (2016). CTA *vs*. *not* chi-square: Fear and specific recommendations *do* synergistically affect behavior. *Optimal Data Analysis*, *5*, 108-111.

[34]Yarnold PR (2016). CTA *vs*. *not* chi-square: Differentiating statistical and ecological significance. *Optimal Data Analysis*, *5*, 112-115.

[35]Yarnold PR (2016). CTA *vs*. chi-square: Differentiating statistical and ecological significance. *Optimal Data Analysis*, *5*, 116-117.

[36]Yarnold PR (2016). CTA *vs*. disintegrated chi-square: Integrated *vs*. piecemeal analysis. *Optimal Data Analysis*, *5*, 118-120.

[37]Yarnold PR (2016). CTA *vs*. chi-square: Comparing voter sentiment in political wards. *Optimal Data Analysis*, *5*, 129-130.

2155-0182/10/$3.00

[38]Yarnold PR (2016). CTA vs. non-disentangled omnibus chi-square: Comparing samples (not) selected for study participation. *Optimal Data Analysis, 5*, 154-157.

[39]Yarnold PR (2016). Using EO-CTA to disentangle sets of sign-test-based multiple-comparisons. *Optimal Data Analysis, 5*, 158-159.

[40]Yarnold PR (2010). UniODA *vs*. chi-square: Ordinal data sometimes feign categorical. *Optimal Data Analysis, 1*, 62-65.

[41]Yarnold PR (2014). UniODA *vs*. chi-square: Audience effect on smile production in infants. *Optimal Data Analysis*, *3*, 3-5.

[42]Yarnold PR (2014). UniODA *vs*. chi-square: Discriminating inhibited and uninhibited infant profiles. *Optimal Data Analysis, 3*, 9-11.

[43]Yarnold PR (2015). UniODA *vs*. chi-square: Measures of effect size. *Optimal Data Analysis, 4*, 137-138.

[44]Yarnold PR (2016). CTA *vs*. *not* chi-square: Differentiating statistical and ecological significance. *Optimal Data Analysis, 5*, 112-115.

[45]Yarnold PR (2016). CTA *vs*. chi-square: Differentiating statistical and ecological significance. *Optimal Data Analysis, 5*, 116-117.

[46]Yarnold, PR (2015). UniODA *vs*. chi-square: Deciphering *R* x *C* contingency tables. *Optimal Data Analysis*, *4*, 156-158.

[47]Yarnold, PR (2015). UniODA *vs*. *not* chi-square: Vaccine administration and flu. *Optimal Data Analysis, 4*, 159-160.

[48]Yarnold, PR (2015). UniODA *vs*. chi-square: Voter sentiment and political ward. *Optimal Data Analysis, 4*, 161-162.

[49]Yarnold, PR (2015). UniODA *vs*. *not* chi-square: Work shift and raw material production quality. *Optimal Data Analysis, 4*, 168-170.

[50]Yarnold PR (2016). UniODA *vs*. chi-square: Describing baseline data from the National Pressure Ulcer Long-Term Care Study (NPULS). *Optimal Data Analysis, 5*, 24-28.

[51]Yarnold PR (2016). ODA *vs*. undocumented chi-square: Clarity *vs*. confusion. *Optimal Data Analysis*, *5*, 121-123.

[52]Yarnold PR (2019). Value-added by ODA *vs*. chi-square. *Optimal Data Analysis, 8*, 10-14.

Author Notes

No conflicts of interest were reported.