Optimal Data Analysis
Vol. 9 (May 3, 2020), 152-156

2155-0182/10/$3.00

Implementing ODA from Within Stata: Nondirectional, Multicategorical Class Variable, Multicategorical Attribute

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

This paper describes how to evaluate an exploratory (nondirectional) hypothesis for a design involving a multicategorical class ("dependent") variable and a multicategorical attribute ("independent variable") using the new Stata package for implementing ODA.

Recent papers[1-17] introduce the new Stata package called **oda**[18] for implementing ODA from within the Stata environment. Because this package is a wrapper for the MegaODA software system[19-21], 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 nondirectional hypothesis for a square design involving a three-category class variable and attribute.

Methods

Data

Table 1 is the cross-classification of the class variable *vote* having three mutually-exclusive and exhaustive *response* options (yes, abstain, nay), and the attribute *region* also having three mutually-exclusive and exhaustive response options (north, border, and south).[22]

		Vote	
Region	*Yea*	*Abstain*	*Nay*
North	61	12	60
Border	17	6	1
South	39	22	7

Analytic Process

The nondirectional ("two-sided") *a priori* hypothesis is voting on the 1836 Pinckney Gag rule is related to region of residence. Exact *p* is estimated by a 25,000-iteration permutation test, and cross-generalizability of findings expected using the ODA model to classify independent random samples is estimated via leave-one-out (LOO) jackknife analysis. For the entire sample, **oda** is implemented with the following syntax (see the help file for **oda** for a complete description of syntax options):

2155-0182/10/$3.00

```
oda vote region, pathoda("C:\ODA\")
store("C:\ODA) iter(25000) loo cat
```

This syntax is explained as follows: "vote" is the *class* variable and "region" is the *attribute*; "C:\ODA\" is the directory path where the MegaODA.exe file exists on the computer, and where all other files generated in analysis are stored; the number of iterations (repetitions) that are used to compute a permutation *p*-value is 25,000; LOO analysis is conducted; and the attribute (region) is categorical. Data for each observation was entered in free format on a separate line using space-delimited text (ASCII) characters.[23,24]

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").

```
ODA model:
----------
IF REGION = 1 THEN VOTE = 3
IF REGION = 2 THEN VOTE = 1
IF REGION = 3 THEN VOTE = 2


Summary for Class VOTE  Attribute REGION
----------------------------------------

Performance Index           Train    LOO
-----------------          ------  ------
Overall Accuracy            44.00%  44.00%
PAC VOTE=1                  14.53%  14.53%
PAC VOTE=2                  55.00%  55.00%
PAC VOTE=3                  88.24%  88.24%
Effect Strength PAC         28.88%  28.88%
PV VOTE=1                   70.83%  70.83%
PV VOTE=2                   32.35%  32.35%
PV VOTE=3                   45.11%  45.11%
Effect Strength PV          24.15%  24.15%
Effect Strength Total       26.52%  26.52%


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


Results of leave-one-out analysis
---------------------------------
225 observations
(P-values are computed for binary class variables only)
```

As seen in the **oda** output, the ODA model is interpreted as follows: "if region is north then predict vote=nay; if region is border then predict vote=yea; and if region is south then predict vote=abstain. The effect strength for sensitivity (ESS) is labelled in the output as the "Effect Strength PAC" (Percentage Accurate Classification). In training and LOO analysis, ESS=28.88% (a moderate effect).[23] Permutation *p*-values for training and stable LOO analyses were <0.0001.

ODA software gives Type I error rates for LOO analyses involving 2 x 2 tables. For applications using multicategorical variables a directional ODA analysis must be conducted. Presently, for the entire sample, **oda** is implemented with the following syntax (here the dir command specifies the order of the three class categories as listed in the ODA model given in the first **oda** output):

```
oda vote region, pathoda("C:\ODA\")
store("C:\ODA) iter(25000) cat dir(< 3 1 2)
```

```
ODA model:
----------
IF REGION = 1 THEN VOTE = 3
IF REGION = 2 THEN VOTE = 1
IF REGION = 3 THEN VOTE = 2


Summary for Class VOTE  Attribute REGION
----------------------------------------

Performance Index           Train
-----------------          ------
Overall Accuracy            44.00%
PAC VOTE=3                  88.24%
PAC VOTE=1                  14.53%
PAC VOTE=2                  55.00%
Effect Strength PAC         28.88%
PV VOTE=3                   45.11%
PV VOTE=1                   70.83%
PV VOTE=2                   32.35%
Effect Strength PV          24.15%
Effect Strength Total       26.52%


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

In summary, ODA was able to find a statistically significant model which discriminated moderately well between regions associated with voting behavior, and was stable in LOO jackknife analysis.

We believe ODA should be considered the preferred statistical approach over other methods because it avoids statistical assumptions required of conventional models, is insensitive to skewed data or outliers, and has the

2155-0182/10/$3.00

ability to handle any variable metric including categorical, Likert-type integer, and real number measurement scales.[23] In contrast to alternative methods, only ODA can identify the optimal (maximum-accuracy) assignments (categorical attributes) or cutpoints (ordered attributes) that exist for the attribute, which in turn facilitates the use of measures of predictive accuracy.

Furthermore, ODA can evaluate model reproducibility by multiple methods, allowing assessment of potential cross-generalizability of the model applied to classify an independent random sample.[23]

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.[25-44]

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.

2155-0182/10/$3.00

[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]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Directional hypothesis, binary class variable, ordinal attribute. *Optimal Data Analysis*, *9*, 141-145.

[17]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Confirmatory hypothesis, binary class variable, continuous attribute. *Optimal Data Analysis*, *9*, 146-151.

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

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

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

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

[22]Bishop YMM, Feinberg SE, Holland PW (1975). *Discrete multivariate analysis*. Cambridge, England: Cambridge University Press.

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

[24]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.

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

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

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

[30]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.

[31]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.

[32]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.

2155-0182/10/$3.00

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

[34]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.

[35]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.

[36]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.

[37]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.

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

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

[40]Yarnold PR, Bryant FB (2013). Analysis involving categorical attributes having many categories. *Optimal Data Analysis*, *2*, 69-70.

[41]Yarnold PR (2013). Analyzing categorical attributes having many response categories. *Optimal Data Analysis*, *2*, 172-176.

[42]Yarnold PR (2013). Univariate and multivariate analysis of categorical attributes with many response categories. *Optimal Data Analysis*, *2*, 177-190.

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

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

Author Notes

No conflicts of interest were reported.