Optimal Data Analysis
Vol. 10 (March 10, 2021), 3-9

2155-0182/10/$3.00

Implementing ODA from Within Stata: Exploratory Hypothesis, Three-Category Class Variable, Continuous Attribute

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

This paper describes how to test a non-directional (exploratory) hypothesis for a design relating a three-category class (“dependent”) variable and a continuous attribute vis-à-vis the Stata package for implementing ODA.

Recent papers[1-22] introduce the new Stata package called **oda**[23] for implementing ODA from within the Stata environment. This package is a wrapper for the MegaODA software system[24-26], so the MegaODA.exe file must be loaded on the computer for the **oda** package to work.[27] 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 two-tailed hypothesis for a design involving a three-category class variable and two continuous attributes.

Methods

Data

Melaragno, Smith, Kormann-Bortolotto and Neto presented data on two possible correlates of Alzheimer’s disease (AD) which are impaired by aging: an indicator of cellular response to DNA damage called sister chromatic exchange (SCE), and an indicator of cell reproduction rate called cell proliferation potential (CPP).[28] Both are ordered attributes. Data were collected for five patients with AD (class category 3), five older adults without AD matched for age with the AD patients (category 2), and five younger adults without AD (category 1).

Analytic Process

The non-directional (“exploratory”) alternative hypothesis is that the three class categories can be discriminated by SCE and CPP scores, and the null hypothesis is that this is not true. Due to the small sample and corresponding low statistical power, generalized p<0.05 is used to establish statistical significance for statistical hypotheses evaluated presently.

Analysis begins by evaluating SCE. 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 adstatus sce , pathoda("C:\ ODA\")
store("C:\ ODA") iter(25000) loo

This syntax is explained as follows: "adstatus" is the *class* variable and "sce" is the *attribute*; "C:\ODA\" is the directory path where the MegaODA.exe file exists on the computer, and where other files generated in analysis are stored; 25,000 iterations (repetitions) are used to obtain a permutation *p*-value; and LOO (leave-one-out validity) analysis is conducted.[29,30]

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 SCE <= 168.5 THEN ADSTATUS = 1
IF 168.5 < SCE <= 196.5 THEN ADSTATUS = 3
IF 196.5 < SCE THEN ADSTATUS = 2

Summary for Class ADSTATUS  Attribute SCE
-----------------------------------------

Performance Index           Train    LOO
-----------------          ------  ------
Overall Accuracy            73.33%  60.00%
PAC ADSTATUS=1              60.00%  60.00%
PAC ADSTATUS=2              80.00%  80.00%
PAC ADSTATUS=3              80.00%  40.00%
Effect Strength PAC         60.00%  40.00%
PV ADSTATUS=1              100.00%  60.00%
PV ADSTATUS=2               80.00%  80.00%
PV ADSTATUS=3               57.14%  40.00%
Effect Strength PV          68.57%  40.00%
Effect Strength Total       64.29%  40.00%

Monte Carlo summary (Fisher randomization):
-------------------------------------------
Iterations:  25000
Estimated p: 0.101440

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

Effect strength for sensitivity (ESS) is labelled in the output as "Effect Strength PAC" (Percentage Accurate Classification). For the exploratory hypothesis ESS is 60.0%, which exceeds the minimum criterion (ESS≥50) to be classified as a relatively strong effect.[29] Due to the tiny sample, this result is not statistically significant ($p<0.11$). LOO analysis shows the effect is unstable, with ESS=40.0% indicating a moderate effect[29] (ODA software only computes exact LOO *p* for two-category class variables).

The next analysis evaluates CPP. The **oda** code is the same as listed earlier, except that "cpp" is substituted for "sce".

As seen, for the exploratory hypothesis ESS is 70.0%, exceeding the minimum criterion (ESS≥75) for a strong effect.[29] Even with the tiny sample this result is statistically significant at the "per-comparison" criterion[29] ($p<0.026$). As seen, CPP values were greatest for young adults without AD, were lowest for older adults with AD, and were intermediate for older adults without AD. LOO analysis shows that the effect is unstable: ESS=60.0% indicates a relatively strong effect.[29]

```
ODA model:
----------
IF CPP <= 160.5 THEN ADSTATUS = 3
IF 160.5 < CPP <= 228.0 THEN ADSTATUS = 2
IF 228.0 < CPP THEN ADSTATUS = 1

Summary for Class ADSTATUS  Attribute CPP
-----------------------------------------

Performance Index           Train    LOO
-----------------          ------  ------
Overall Accuracy            80.00%  73.33%
PAC ADSTATUS=1             100.00%  80.00%
PAC ADSTATUS=2             100.00% 100.00%
PAC ADSTATUS=3              40.00%  40.00%
Effect Strength PAC         70.00%  60.00%
PV ADSTATUS=1               83.33%  80.00%
PV ADSTATUS=2               71.43%  62.50%
PV ADSTATUS=3              100.00% 100.00%
Effect Strength PV          77.38%  71.25%
Effect Strength Total       73.69%  65.62%

Monte Carlo summary (Fisher randomization):
-------------------------------------------
Iterations:  25000
Estimated p: 0.025280

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

.

All possible pairwise comparisons may be conducted to determine which aspects of the omnibus model are statistically reliable.

Groups 2 and 3 are compared using the following **oda** script.[29]

oda adstatus cpp if adstatus !=1,
pathoda("C:\Users\Ariel\Desktop\ODA\")
store("C:\Users\Ariel\Desktop\ODA\")
iter(25000) loo

2155-0182/10/$3.00

```
ODA model:
----------
IF CPP <= 160.5 THEN ADSTATUS = 3
IF 160.5 < CPP THEN ADSTATUS = 2

Summary for Class ADSTATUS  Attribute CPP
-----------------------------------------

Performance Index          Train    LOO
-----------------         ------  ------
Overall Accuracy           70.00%  50.00%
PAC ADSTATUS=2            100.00%  60.00%
PAC ADSTATUS=3             40.00%  40.00%
Effect Strength PAC        40.00%   0.00%
PV ADSTATUS=2              62.50%  50.00%
PV ADSTATUS=3             100.00%  50.00%
Effect Strength PV         62.50%   0.00%
Effect Strength Total      51.25%   0.00%

Monte Carlo summary (Fisher randomization):
-------------------------------------------
Iterations:  25000
Estimated p: 0.878480

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

Fisher's exact test (directional) classification table  p = .738095
```

.

The comparison between categories 2 and 3 was moderately strong in training analysis but was not statistically reliable (*p*<0.87). In LOO analysis a degenerate solution emerged classifying all observations as being category 2 (ESS=0).[29] The comparison between categories 1 and 3 was conducted via the following code.

```
oda adstatus cpp if adstatus !=2,
pathoda("C:\Users\Ariel\Desktop\ODA\")
store("C:\Users\Ariel\Desktop\ODA\")
iter(25000) loo
```

```
ODA model:
----------
IF CPP <= 230.5 THEN ADSTATUS = 3
IF 230.5 < CPP THEN ADSTATUS = 1

Summary for Class ADSTATUS  Attribute CPP
-----------------------------------------

Performance Index          Train    LOO
-----------------         ------  ------
Overall Accuracy          100.00%  80.00%
PAC ADSTATUS=1            100.00%  80.00%
PAC ADSTATUS=3            100.00%  80.00%
Effect Strength PAC       100.00%  60.00%
PV ADSTATUS=1             100.00%  80.00%
PV ADSTATUS=3             100.00%  80.00%
Effect Strength PV        100.00%  60.00%
Effect Strength Total     100.00%  60.00%

Monte Carlo summary (Fisher randomization):
-------------------------------------------
Iterations:  25000
Estimated p: 0.008600

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

Fisher's exact test (directional) classification table  p = .103175
```

The comparison between categories 1 and 3 was errorless in training analysis (ESS= 100), and statistically significant (*p*<0.0064). Classification degraded in LOO analysis (ESS= 60, a relatively strong effect). This finding indicated that older adults with AD have lower CPP values than younger adults without AD.

The final comparison between categories 1 and 2 was conducted via the following code.

```
oda adstatus cpp if adstatus !=3,
pathoda("C:\Users\Ariel\Desktop\ODA\")
store("C:\Users\Ariel\Desktop\ODA\")
iter(25000) loo
```

```
ODA model:
----------
IF CPP <= 228.0 THEN ADSTATUS = 2
IF 228.0 < CPP THEN ADSTATUS = 1

Summary for Class ADSTATUS  Attribute CPP
-----------------------------------------

Performance Index          Train    LOO
-----------------         ------  ------
Overall Accuracy          100.00%  90.00%
PAC ADSTATUS=1            100.00%  80.00%
PAC ADSTATUS=2            100.00% 100.00%
Effect Strength PAC       100.00%  80.00%
PV ADSTATUS=1             100.00% 100.00%
PV ADSTATUS=2             100.00%  83.33%
Effect Strength PV        100.00%  83.33%
Effect Strength Total     100.00%  81.67%

Monte Carlo summary (Fisher randomization):
-------------------------------------------
Iterations:  25000
Estimated p: 0.007520

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

Fisher's exact test (directional) classification table  p = .023810
```

.

The comparison between categories 1 and 2 was also errorless (p<0.0069) and remained strong effect in LOO analysis (ESS= 81.7). This finding indicates older adults without AD have lower CPP values than younger adults without AD.

Taken together the results of pairwise analysis reveal there is a relatively strong, statistically reliable effect whereby younger adults without AD have higher CPP values than older adults regardless of their AD status. However, CPP does not discriminate between adults who do *vs*. do not have AD.[29]

2155-0182/10/$3.00

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 ability to handle any variable metric including categorical, Likert-type integer, and real number measurement scales.[29] 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.[29]

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.[30-50]

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.

2155-0182/10/$3.00

[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]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]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Nondirectional, multicategorical class variable, multicategorical attribute. *Optimal Data Analysis*, *9*, 152-156.

[19]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Directional hypothesis, multicategorical class variable and attribute. *Optimal Data Analysis*, *9*, 157-161.

[20]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Directional hypothesis, multicategorical class variable, ordinal attribute. *Optimal Data Analysis*, *9*, 162-166.

[21]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: *A Priori* hypothesis, three-category class variable, four-level (integer) attribute. *Optimal Data Analysis*, *9*, 167-171.

[22]Linden A, Yarnold PR (2020). Implementing ODA from within Stata: A reanalysis of the National Supported Work Experiment. *Optimal Data Analysis*, *9*, 178-182.

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

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

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

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

[27]https://github.com/njrhodes/ODA

[28]Melaragno MI, Smith MAC, Kormann-Bortolotto MH, Neto JT (1991). Lymphocyte proliferation and sister chromatid exchange in Alzheimer's disease. *Gerontology*, *37*, 293-298.

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

[30]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.

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

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

2155-0182/10/$3.00

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

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

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

[36]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.

[37]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.

[38]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.

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

[40]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.

[41]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.

[42]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.

[43]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.

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

[45]Yarnold PR, Brofft GC (2013). ODA range test *vs*. one-way analysis of variance: Comparing strength of alternative line connections. *Optimal Data Analysis*, *2*, 198-201.

[46]Yarnold PR (2013). ODA range test *vs*. one-way analysis of variance: Patient race and lab results. *Optimal Data Analysis*, *2*, 206-210.

[47]Yarnold PR (2014). How to assess the inter-method (parallel-forms) reliability of ratings made on ordinal scales: Evaluating and comparing the Emergency Severity Index (Version 3) and Canadian Triage Acuity Scale. *Optimal Data Analysis*, *3*, 50-54.

[48]Yarnold PR (2016) Causality of adverse drug reactions: The upper-bound of arbitrated expert agreement for ratings obtained by WHO and Naranjo algorithms. *Optimal Data Analysis*, *5*, 37-40.

[49]Yarnold PR (2016). Novometric *vs*. ODA reliability analysis *vs*. polychoric correlation with relaxed distributional assumptions: Inter-rater reliability of independent ratings of plant health. *Optimal Data Analysis*, *5*, 179-183.

2155-0182/10/$3.00

[50]Rhodes NJ, Yarnold PR (2020). Generating novometric confidence intervals in R: Bootstrap analyses to compare model and chance ESS. *Optimal Data Analysis*, *9*, 172-177.

Author Notes

No conflicts of interest were reported.