Optimal Data Analysis
Vol. 9 (April 28, 2020), 146-151

2155-0182/10/$3.00

Implementing ODA from Within Stata: Confirmatory Hypothesis, Binary Class Variable, Continuous 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 continuous (interval or ratio) attribute is evaluated via MegaODA software using the new Stata package implementing ODA analysis.

Recent papers[1-16] introduce the new Stata package called **oda**[17] for implementing ODA from within the Stata environment. Because this package is a wrapper for the MegaODA software system[18-20], 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 directional hypothesis for a binary class variable and a continuous (interval or ratio) attribute.

Methods

Data

We consider data from Martin et al[21] on heart rate variability (*Singer*) and susceptibility to sudden cardiac death (*SCD*). Arbitrary dummy-codes were used to identify SCD susceptibility: susceptible=1, not susceptible=0. Increasing values of Singer scores indicate lower levels of heart rate variability. Data for every subject was entered in free format on a separate line as space-delimited text (ASCII) characters.[22]

Analytic Process

We repeat the ODA analysis previously conducted on these data (see example 5.9, *Optimal Data Analysis: A Guidebook with Software for Windows*[23]). The directional or "one-tailed" alternative hypothesis is that the binary class ("dependent") variable SCD can be discriminated on the basis of Singer scores (continuous attribute or "independent variable"): susceptible people are hypothesize to have higher Singer scores than non-susceptible people. The null hypothesis is that this is not true. Weighting by prior odds (the default setting) is used to obtain a model that maximizes ESS (i.e., 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).[23]

2155-0182/10/$3.00

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 scd singer, pathoda("C:\ODA\")
store("C:\ ODA\output") iter(25000)
dir(> 1 0)

The above syntax is explained as follows: The variable "scd" is the *class* variable; the variable "singer" is the *attribute*; the directory path where the MegaODA.exe file is located on the computer is "C:\ODA\"; the directory path where output and other files generated during analysis are stored is "C:\ODA\output"; 25,000 iterations (repetitions) are used to compute the permutation *p*-value; and the directional hypothesis is that people who are susceptible to SCD have greater singer scores than the non-susceptible people.

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 *singer* $\leq$ 0.0796 then predict *SCD* = 0; otherwise, predict *SCD* = 1." People with lower Singer scores are predicted to be less susceptible to SCD than are people with higher Singer scores. As seen, this model correctly classified 86.36% of the non-susceptible people, and 63.64% of susceptible people. This level of classification accuracy is statistically significant: *p*<0.0036. The effect strength for sensitivity (ESS) is labelled in the output as "Effect Strength PAC" (Percentage Accurate Classification). ESS is 50.00% which exactly meets the minimal criterion for a relatively strong effect.[23]

In summary, consistent with the *a priori* hypothesis, ODA identified a model which discriminated relatively strongly between people who are susceptible *vs*. not susceptible to SCD, and this effect was statistically significant.

```
ODA model:
----------
IF SINGER <= 7.959999900000000E-002 THEN SCD = 0
IF 7.959999900000000E-002 < SINGER THEN SCD = 1


Summary for Class SCD  Attribute SINGER
---------------------------------------

Performance Index           Train
-----------------          ------
Overall Accuracy            75.00%
PAC SCD=0                   86.36%
PAC SCD=1                   63.64%
Effect Strength PAC         50.00%
PV SCD=0                    70.37%
PV SCD=1                    82.35%
Effect Strength PV          52.72%
Effect Strength Total       51.36%


Monte Carlo summary (Fisher randomization):
-------------------------------------------
Iterations:  25000
Estimated p: 0.003600
```

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 continuous 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.[23] 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 evaluate model reproducibility using multiple methods, which allows for assessment of potential cross-generalizability of the model to independent random samples.[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.[24-53]

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

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

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

2155-0182/10/$3.00

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

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

[21]Martin GJ, Magid NM, Myers G, Barnett PS, Schaad JW, Weiss JS, et al. (1987). Heart rate variability and sudden cardiac death secondary to coronary artery disease during ambulatory electrocardiographic monitoring. *American Journal of Cardiology*, *60*, 86-89.

[22]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.

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

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

[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.

[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.

2155-0182/10/$3.00

[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]DeArmon JS, Lacher AR (1997). Aggregate flow directives as a ground delay strategy: Concept analysis using discrete-event simulation. *Air Traffic Control Quarterly, 4,* 307-323.

[39]Cohen R, Wiley S, Oswald DP, Eakin KB, Best AM (1999). Applying utilization management principles to a comprehensive service system for children with emotional and behavioral disorders and their families: A feasibility study. *Journal of Child and Family Studies*, *8*, 463-476.

[40]Axelrod BN, Fichtenberg NL, Millis SR, Wertheimer JC (2006). Detecting incomplete effort with digit span from the Wechsler Adult Intelligence Scale—Third Edition. *The Clinical Neuropsychologist*, *20*, 513-523.

[41]Belmares J, Gerding DN, Parada JP, Miskevics S, Weaver F, Johnson S (2007). Outcome of metronidazole therapy for *Clostridium difficile* disease and correlation with a scoring system. *Journal of Infection, 55*, 495-501.

[42]Diesfeldt HFA (2007). Discrepanties tussen de Informantvragenlijst (IQCODE) en cognitieve tests bij deelnemers aan psychogeriatrische dagbehandeling.*Tijdschrift voor Gerontologie en Geriatrie, 38*, 199-209.

[43]Stalans LJ, Seng M (2006). Identifying subgroups at high risk of dropping out of domestic batterer treatment: The buffering effects of a high school education. *International Journal of Offender Therapy and Comparative Criminology*, *10*, 1-19.

[44]Brocklehurst PR, Baker SR, Speight PM (2010). Factors which determine the referral of potentially malignant disorders by primary care dentists. *Journal of Dentistry, 38*, 569-578.

[45]Brocklehurst PR, Baker SR, Speight PM (2010). Primary care clinicians and the detection and referral of potentially malignant disorders of the mouth: A summary of the current evidence. *Primary Dental Care, 17*, 65-71.

[46]Stalans LJ, Hacker R, Talbot ME (2010). Comparing nonviolent, other-violent, and domestic batterer sex offenders: Predictive accuracy of risk assessments on sexual recidivism. *Criminal Justice and Behavior, 37*, 613-628.

[47]Albuquerque K, Giangreco D, Morrison C, Siddiqui M, Sinacore J, Potkul R, Roeske J (2011). Radiation-related predictors of hematologic toxicity after concurrent chemoradiation for cervical cancer and implications for bone marrow-sparing pelvic IMRT. *International Journal of Radiation Oncology * Biology * Physics*, *79*, 1043-1047.

[48]Jones A, Ingram MV (2011). A comparison of selected MMPI-2 and MMPI-2-RF validity scales in assessing effort on cognitive tests in a military sample. *The Clinical Neurologist*, *7*, 1207-1227.

2155-0182/10/$3.00

[49]Yung RC, Zeng MY, Stoddard GJ, Garff M, Callahan K (2012). Transcutaneous computed bioconductance measurement in lung cancer: A treatment enabling technology useful for adjunctive risk stratification in the evaluation of suspicious pulmonary lesions. *Journal of Thoracic Oncology*, *7*, 681-689.

[50]Herrold AA, Pape TLB, Guernon A, Mallinson T, Collins E, Jordan N (2014). Prescribing multiple neurostimulants during rehabilitation for severe brain injury. *The Scientific World Journal*, Article ID 964578, 1-7. DOI: 10.1155/2014/964578

[51]Clark-Raymond A, Meresh E, Hoppensteadt D, Fareed J, Sinacore J, Halaris A (2014). Vascular endothelial growth factor: A potential diagnostic biomarker for major depression. *Journal of Psychiatric Research*, August 19. DOI: http://dx.doi.org/10.1016/j.jpsychires.2014.08.005

[52]Oguoma VM, Nwose EU, Ulasi II, Akintunde AA, Chukwukelu EE, Araoye MA, Ijoma CK, Onyia IC, Ogbu II, Onyeanusi JC, Digban KA, Onodugo OD, Adediran O, Richards RS, Skinner TC (2016). Maximum accuracy obesity indices for screening metabolic syndrome in Nigeria: A consolidated analysis of four cross-sectional studies. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *10*, 121-127.

[53]Martin J, Benjamin E, Craver C, Kroch E, Nelson E, Bankowitz R (2016). Measuring adverse events in hospitalized patients: An administrative method for measuring harm. *Journal of Patient Safety*, *12*, 125-131.

Author Notes

No conflicts of interest were reported.