Optimal Data Analysis
Vol. 9 (April 25, 2020), 128-132
Copyright 2020 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Implementing ODA from Within Stata: Confirmatory Hypothesis, Binary Class Variable, and 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 an ordinal (quintiles) attribute is evaluated using MegaODA software using the new Stata package implementing ODA analysis.

Recent papers[1-12] introduce the new Stata package called **oda**[13] for implementing ODA from within the Stata environment. Because this package is a wrapper for the MegaODA software system[14-16], 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 involving a single binary class variable, and a single ordinal (quintile) attribute.

Methods

Data

We consider Berry and Mielke's data on socio-economic status (SES) and political affiliation, called *politics* in the Stata program.[17] Arbitrary dummy-codes were used to identify *politics*: Democrat=11, and Republican=34. SES was assessed by two attributes—education (*educate*) and occupational prestige (*prestige*), both of which were measured using quintiles. For both attributes the lowest-scoring fifth of the sample (lowest 20% of scores) was coded as quintile=1, the next-lowest 20% of scores were coded as quintile=2, and the highest 20% of scores were coded as quintile=5. Data for every subject was entered in free format on a separate line as space-delimited text (ASCII) characters.[18]

Analytic Process

We repeat the ODA analysis previously conducted on these data (see example 5.5, *Optimal Data Analysis: A Guidebook with Software for Windows*[19]). The directional or "one-tailed" alternative hypothesis is that the binary class ("dependent") variable *politics* can be discriminated on the basis of SES assessed by *educate* and *prestige* (ordinal attributes or "independent variables"). The null hypothesis is that this is not true. Weighting by prior odds (the default

Copyright 2020 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

setting) is used to obtain a model which maximizes ESS (i.e., classification accuracy normed *vs*. chance).[19] A total of 25,000 Monte Carlo iterations are used to estimate Type I error (i.e., *p* value), and because *two* tests of statistical hypotheses are being evaluated, the Sidak per-comparison criterion of $p<0.02532$ is used to ensure experimentwise $p<0.05$.[19]

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

```
oda politics educate, pathoda("C:\ODA\")
store("C:\ ODA\output") iter(25000)
dir(< 34 11) Sidak(2)
```

The above syntax is explained as follows: The variable "politics" is the *class* variable; the variable "educate" 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"; the 25,000 iterations (repetitions) are used for computing the permutation *p*-value; the directional (one-tailed) hypothesis is that Republicans will score at lower levels than Democrats on *educate*; and *p*-values will be evaluated using the Sidak criterion for two tests of statistical hypotheses.

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 *educate* $\leq$ 3.5 then predict *politics* = 34; otherwise, predict *politics* = 11." As hypothesized, Republicans scored in the lowest three education quintiles. As seen, this model correctly classified 100% of the Republican subjects, but only 50% of the Democratic subjects.

Effect strength for sensitivity (ESS) is labelled in the output as "Effect Strength PAC" (Percentage Accurate Classification). The ESS is 50% (the minimum criterion for a relatively strong effect[19]) and the permutation $p<0.027$. This Type I error rate is considered statistically significant by the *per-comparison* criterion[19], but not by the *experimentwise* criterion: here the Sidak adjusted Type I error rate is $p<0.052$. In summary, ODA found a model which discriminated the political parties relatively strongly, but due to the small sample and corresponding weak statistical power[20], the effect was marginally significant.

```
ODA model:
----------
IF EDUCATE <= 3.5 THEN POLITICS = 34
IF 3.5 < EDUCATE THEN POLITICS = 11


Summary for Class POLITICS  Attribute EDUCATE
-------------------------------------------

Performance Index           Train
-----------------          ------
Overall Accuracy            80.00%
PAC POLITICS=34            100.00%
PAC POLITICS=11             50.00%
Effect Strength PAC         50.00%
PV POLITICS=34              75.00%
PV POLITICS=11             100.00%
Effect Strength PV          75.00%
Effect Strength Total       62.50%


Monte Carlo summary (Fisher randomization):
-------------------------------------------
Iterations:  25000
Estimated p: 0.026280
Sidak Adjusted (2) p: .05186936
```

Next, **oda** is implemented using the following syntax to test the *a priori* hypothesis for the attribute *prestige*:

```
oda politics prestige, pathoda("C:\ODA\")
store("C:\ ODA\output") iter(25000)
dir(< 34 11) Sidak(2)
```

As seen in the **oda** output, the ODA model is interpreted as follows: "if *prestige* $\leq$ 2.5 then predict *politics* = 34; otherwise, predict *politics* = 11." As hypothesized, Republicans scored in the lowest two prestige quintiles. This

Copyright 2020 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

model correctly classified 91.67% of Republicans, and 87.50% of Democratics.

```
ODA model:
----------
IF PRESTIGE <= 2.5 THEN POLITICS = 34
IF 2.5 < PRESTIGE THEN POLITICS = 11


Summary for Class POLITICS  Attribute PRESTIGE
-----------------------------------------------

Performance Index           Train
-----------------          ------
Overall Accuracy            90.00%
PAC POLITICS=34             91.67%
PAC POLITICS=11             87.50%
Effect Strength PAC         79.17%
PV POLITICS=34              91.67%
PV POLITICS=11              87.50%
Effect Strength PV          79.17%
Effect Strength Total       79.17%


Monte Carlo summary (Fisher randomization):
-------------------------------------------
Iterations:  25000
Estimated p: 0.000480
Sidak Adjusted (2) p: .00095977
```

The ESS of 79.17% meets the criterion for a strong effect.[19] The Sidak adjusted Type I error rate is $p<0.00096$. Thus, ODA identified a model which discriminated the political parties strongly, and met the experimentwise criterion for statistical significance.

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 single ordinal (quintile) 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.[19] 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[19]) 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.[21-34]

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.

Copyright 2020 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

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

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

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

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

[17]Berry JR, Mielke PW (1985). Goodman and Kruskal's TAU-B statistic: A nonasymptotic test of significance. *Sociological Methods and Research*, *13*, 543-550.

[18]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.

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

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

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

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

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

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

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

[26]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.

Copyright 2020 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

[27]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.

[28]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.

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

[30]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.

[31]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.

[32]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.

[33]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.

[34]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.