Optimal Data Analysis
Vol. 9 (May 3, 2020), 157-161
Copyright 2020 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Implementing ODA from Within Stata: Directional Hypothesis, Multicategorical Class Variable and Attribute

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

This paper demonstrates how to evaluate a confirmatory (directional) 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[1-18] introduce the new Stata package called **oda**[19] for implementing ODA from within the Stata environment. Because this package is a wrapper for the MegaODA software system[20-22], 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

As an example of a directional hypothesis involving a multicategorical class variable and a multicategorical attribute, consider Reynold's data on political affiliation of 1,852 high school students and their parents.[23]

Table 1 is the cross-classification of the class variable *student* having seven mutually-exclusive exhaustive *response* options (strong Democrat=1; Democrat=2; Independent-Democrat=3; Independent=4; strong Republican =5; Republican=6; Independent Republican=7), and the attribute *region* which consists of the identically-coded set of seven mutually-exclusive exhaustive response options. For example, the intersection of column 3 (Ind-Dem) and row 4 (Ind) indicates 32 Independent students have parents who are Independent-Democrats.

Note that the *left-hand* side of the coding scheme begins with Strong Democrat, code=1. To preserve the symmetric (unidimensional) ordinal structure of measurement codes across the entire scale, the *right-hand* side of the coding scheme *should* begin with Strong Republican, having code=7. Similarly, to preserve the measurement symmetry, the second element of

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

Table 1: Political Affiliation Status of High School *Students* and Their *Parents*[23]: Dem=Democrat; Ind=Independent; Rep=Republican (Tabled are *N*)

Student's Political Affiliation	Parent's Political Affiliation: Strong Dem	Dem	Ind-Dem	Ind	Strong Rep	Rep	Ind-Rep
Strong Dem	180	108	30	20	2	5	3
Dem	147	167	39	30	10	38	17
Ind-Dem	63	78	38	30	14	30	14
Ind	33	49	32	50	17	42	14
Strong Rep	9	13	14	23	17	35	45
Rep	16	29	14	23	17	92	61
Ind-Rep	9	13	4	10	9	35	64

the right-hand coding scheme should be Republican (code=6), and the third element should be Independent-Republican (code=5).

Analytic Process

The directional ("one-sided") *a priori* hypothesis is family members have the same political affiliation, and the null hypothesis is that family members *do not* have the same political affiliation. Exact *p* is estimated by a 25,000-iteration permutation test. For the entire sample, **oda** is implemented with the following syntax (see the help file for **oda** for a complete description of syntax options):

oda student parents, pathoda("C:\ODA\")
store("C:\ODA) iter(25000) cat
dir(< 1 2 3 4 5 6 7)

This syntax is explained as follows: "student" is the *class* variable and "parents" 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) used to compute a permutation *p*-value is 25,000; the attribute (parents) is categorical; and the directional hypothesis is that code assignments made by the students and the parents agree. Data for each observation was entered in free format on a separate line using space-delimited text (ASCII) characters.[24,25]

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 PARENTS = 1 THEN STUDENT = 1
IF PARENTS = 2 THEN STUDENT = 2
IF PARENTS = 3 THEN STUDENT = 3
IF PARENTS = 4 THEN STUDENT = 4
IF PARENTS = 5 THEN STUDENT = 5
IF PARENTS = 6 THEN STUDENT = 6
IF PARENTS = 7 THEN STUDENT = 7


Summary for Class STUDENT  Attribute PARENTS
--------------------------------------------

Performance Index           Train
-----------------           ------
Overall Accuracy             32.83%
PAC STUDENT=1                51.72%
PAC STUDENT=2                37.28%
PAC STUDENT=3                14.23%
PAC STUDENT=4                21.10%
PAC STUDENT=5                10.90%
PAC STUDENT=6                36.51%
PAC STUDENT=7                44.44%
Effect Strength PAC          19.36%
PV STUDENT=1                 39.39%
PV STUDENT=2                 36.54%
PV STUDENT=3                 22.22%
PV STUDENT=4                 26.88%
PV STUDENT=5                 19.77%
PV STUDENT=6                 33.21%
PV STUDENT=7                 29.36%
Effect Strength PV           17.90%
Effect Strength Total        18.63%


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

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

Seen in the **oda** output, the ODA model is interpreted as follows: "student's and parents' political affiliation codes are the same". The effect strength for sensitivity (ESS) is labelled in the output as the "Effect Strength PAC" (i.e., Percentage Accurate Classification). Presently, ESS=19.36% (a relatively weak effect).[26] The permutation *p*-value was <0.0001.

In summary, ODA was able verify that political affiliations of students and their parents exhibit a statistically significant, but relatively weak tendency to have the identical political affiliation .

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.[24] 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.[24]

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.[27-46]

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.

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

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

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

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

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

[23]Reynolds HT (1977). *The analysis of cross-classifications*. New York: Free Press.

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

[25]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.

[26]Yarnold PR (2017). What is optimal data analysis? *Optimal Data Analysis*, *6*, 26-42.

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

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

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

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

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

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

[32]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.

[33]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.

[34]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.

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

[36]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.

[37]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.

[38]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.

[39]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.

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

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

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

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

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

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

[46]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.