Optimal Data Analysis
Vol. 9 (January 15, 2020), 9-13

2155-0182/10/$3.00

Implementing ODA From Within Stata: An Application to Data From a Randomized Controlled Trial (*Invited*)

Ariel Linden, Dr.P.H.
Linden Consulting Group, LLC

In this paper, the new Stata package for implementing ODA is introduced by reanalyzing data from a study by Linden and Butterworth (2014) that investigated the effect of a comprehensive hospital-based intervention in reducing readmissions for chronically ill patients. In the original analysis, negative binomial regression was used to evaluate readmission rates and emergency department visit rates at 30 and 90 days, and no treatment effects were found. However, ODA is a superior analytic approach because of its insensitivity to skewed data, model-free permutation tests to derive *P* values, identification of the threshold value which best discriminates intervention and control groups, use of a chance- and complexity-corrected indexes of classification accuracy, and cross-validation to assess generalizability of the findings.

A new Stata package called **oda**[1] is now available for implementing ODA from within the Stata environment. This package is a wrapper for the MegaODA software[2], and therefore the megaODA.exe file must be loaded on the computer for the **oda** package to work (ODA 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).

In this paper, I demonstrate the ease in using the **oda** package by way of reanalysis of data from a randomized controlled trial (RCT) that investigated the effect of a comprehensive hospital-based intervention in reducing readmissions for chronically ill patients.[3]

This particular study is highlighted for a couple of reasons. First, by virtue of the randomization process, in sufficiently large RCTs the study groups are expected to be equivalent on all characteristics, rendering the analysis a rather straightforward affair – that is, all that is needed in the evaluation is the outcome variable and the treatment variable. Second (and related to the first), although an analytic process may be uncomplicated (such as in an RCT), the correct model must still be chosen to fit the data's distribution. In the study by Linden and Butterworth[3] the outcome (readmissions) was a

2155-0182/10/$3.00

count variable (i.e., discrete positive values which are generally skewed due to many zeros). Choosing the correct count model is not a trivial exercise, requiring the investigator to understand the differences between the various count models and assess goodness of fit metrics to choose the most appropriate model for the data at hand.[4] Unfortunately, many researchers simply apply models based on what is most commonly used in their discipline, rather than whether the statistical assumptions are met for their data. Conversely, ODA requires no statistical assumptions, and the best fitting model will be identified algorithmically. As such, ODA overcomes many of the statistical limitations common in evaluation work.

Taken together, these data allow for a "gentle" introduction to ODA while demonstrating the superiority of ODA over traditional parametric methods. I leave more complex analyses for a later paper.

Methods

For brevity analyses are limited to the primary outcome – readmissions within 30 days – first analyzed on the entire sample, then separately for the congestive heart failure (CHF) and the chronic obstructive pulmonary disease (COPD) subgroups.

The **oda** syntax for analyzing the total sample readmissions is as follows (see the help file for **oda** for a complete description of the syntax options):

```
oda treat read30,  pathoda("C:\ODA\")
store("C:\ODA\output") iter(25000) loo cat
```

The above syntax is explained as follows: The variable "treat" is the *class* variable; the outcome variable "read30" is the *attribute*; the directory path where the megaODA.exe file is located on my computer is "C:\ODA\"; the directory path where the output and other files generated during the analysis should be stored is "C:\ODA\output"; the number of iterations (repetitions) for computing a permutation *P*-value is 25,000; leave-one-out (LOO) analysis should be performed, and the attribute (in this case "read30") should be treated as a categorical variable.

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 READ30 = 0 THEN TREAT = 0
IF READ30 = 1 THEN TREAT = 1
IF READ30 = 2 THEN TREAT = 1
IF READ30 = 4 THEN TREAT = 1


Summary for Class TREAT  Attribute READ30
-----------------------------------------

Performance Index           Train    LOO
-----------------          ------  ------
Overall Accuracy            51.56%  51.37%
PAC TREAT=0                 83.40%  83.40%
PAC TREAT=1                 18.97%  18.58%
Effect Strength PAC          2.37%   1.97%
PV TREAT=0                  51.31%  51.18%
PV TREAT=1                  52.75%  52.22%
Effect Strength PV           4.05%   3.41%
Effect Strength Total        3.21%   2.69%


Monte Carlo summary (Fisher randomization):
-------------------------------------------
Iterations:  25000
Estimated p: 0.658680



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

Fisher's exact test (directional) classification table  p = .318892
```

As shown in the **oda** output, the ODA model can be interpreted as follows: "if 30 readmissions = 0, then predict that the treatment group is 0 (controls). If 30 readmissions = 1, 2, or 4 then predict that the treatment group is 1 (treatment)."

The effect strength for sensitivity (ESS) is labelled in the output as "Effect Strength PAC" ("Percent Accuracy in Classification"). In the training data the ESS is 2.37% and in the LOO analysis it is 1.97% (a very weak effect).[5] The permutation *P*-value for the training sample was 0.659 and for the LOO analysis was 0.319. In summary, ODA could not find a model that sufficiently discriminated between treatment groups to elicit a statistically significant treatment effect.

2155-0182/10/$3.00

The **oda** syntax for analyzing readmissions in the CHF subgroup is as follows:

```
oda treat read30 if condition==1,
pathoda("C:\ODA\") store("C:\ODA\output")
iter(25000) loo cat
```

and the **oda** syntax for analyzing readmissions in the COPD subgroup is as follows:

```
oda treat read30 if condition==2,
pathoda("C:\ODA\") store("C:\ODA\output")
iter(25000) loo cat
```

As can be seen, the only change in the syntax is the addition of the qualifier "if condition == 1" (for CHF) and "if condition == 2" (for COPD). One would rightly surmise that "*condition*" is a variable in the dataset that denotes the subgroup.

The CHF output is as follows:

```
ODA model:
----------
IF READ30 = 0 THEN TREAT = 0
IF READ30 = 1 THEN TREAT = 1
IF READ30 = 2 THEN TREAT = 0
IF READ30 = 4 THEN TREAT = 1


Summary for Class TREAT  Attribute READ30
-----------------------------------------

Performance Index           Train    LOO
-----------------          ------  ------
Overall Accuracy            52.53%  52.14%
PAC TREAT=0                 87.50%  87.50%
PAC TREAT=1                 17.83%  17.05%
Effect Strength PAC          5.33%   4.55%
PV TREAT=0                  51.38%  51.14%
PV TREAT=1                  58.97%  57.89%
Effect Strength PV          10.35%   9.04%
Effect Strength Total        7.84%   6.80%


Monte Carlo summary (Fisher randomization):
-------------------------------------------
Iterations:  25000
Estimated p: 0.377000



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

Fisher's exact test (directional) classification table  p = .197051
```

And the COPD output is as follows:

```
ODA model:
----------
IF READ30 = 0 THEN TREAT = 0
IF READ30 = 1 THEN TREAT = 0
IF READ30 = 2 THEN TREAT = 1


Summary for Class TREAT  Attribute READ30
-----------------------------------------

Performance Index           Train    LOO
-----------------          ------  ------
Overall Accuracy            52.16%  52.16%
PAC TREAT=0                 97.71%  97.71%
PAC TREAT=1                  4.03%   4.03%
Effect Strength PAC          1.74%   1.74%
PV TREAT=0                  51.82%  51.82%
PV TREAT=1                  62.50%  62.50%
Effect Strength PV          14.32%  14.32%
Effect Strength Total        8.03%   8.03%


Monte Carlo summary (Fisher randomization):
-------------------------------------------
Iterations:  25000
Estimated p: 0.860520



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

Fisher's exact test (directional) classification table  p = .331156
```

As was the case for the total sample, ODA could not find a model for either subgroup that sufficiently discriminated between treatment groups to elicit a statistically significant treatment effect.

Discussion

In this paper, I demonstrated how the new Stata package **oda** can be used to evaluate an RCT. In this example data, the outcome under study was a count variable, which using parametric statistics would require a fair amount of expertise to determine the appropriate model.

ODA should be considered the preferred approach *vs*. commonly-used parametric models because ODA avoids the assumptions required of parametric 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.[2] Moreover, in contrast to regression models, ODA also has the distinct ability to ascertain where optimal (maximum-accuracy) cutpoints are located on the outcome variable, which in turn, facilitates the use of measures of predictive accuracy. Moreover, ODA can per-

2155-0182/10/$3.00

form cross-validation using LOO which allows for assessing the cross-generalizability of the model to potentially new study participants or non-participants.

Finally, the findings continue to support recommendations to employ the ODA and CTA frameworks to draw causal inferences regarding treatment effects in observational data, and in data from randomized controlled trials.[6-22] A large, ever-increasing body of evidence supports the use of ODA and CTA to evaluate the efficacy of health-improvement interventions and policy initiatives.[23-25]

References

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

[2]Yarnold PR, Soltysik RC (2016). *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[3]Linden A, Butterworth SW (2014). A comprehensive hospital-based Intervention to reduce readmissions for chronically ill patients: A randomized controlled trial. *American Journal of Managed Care*, *20*, 783-792.

[4]Long JS, Freese J (2006). *Regression Models for Categorical Dependent Variables Using Stata*. 2nd ed. College Station, TX: Stata Press; 2006.

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

[6]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.

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

[8]Yarnold PR, Linden A (2017). Computing propensity score weights for CTA models involving perfectly predicted endpoints. *Optimal Data Analysis*, *6*, 43-46.

[9]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.

[10]Linden A, Yarnold PR (2016). Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice*, *22*, 839-847.

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

[12]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.

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

[14]Yarnold PR, Linden A (2016). Using machine learning to model dose-response relationships via ODA: Eliminating response variable baseline variation by ipsative standardization. *Optimal Data Analysis*, *5*, 41-52.

2155-0182/10/$3.00

[15]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.

[16]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.

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

[18]Linden A, Yarnold PR (2018). The Australian gun buy-back program and the rate of suicide by firearm. *Optimal Data Analysis*, *7*, 28-35.

[19]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.

[20]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.

[21]Linden A, Yarnold PR (2018). Using ODA in the evaluation of randomized controlled trials. *Optimal Data Analysis*, *7*, 46-49.

[22]Linden A, Yarnold PR (2018). Using ODA in the evaluation of randomized controlled trials: Application to survival outcomes. *Optimal Data Analysis*, *7*, 50-53.

[23]Linden A, Adams J, Roberts N (October, 2003). *Evaluation methods in disease management: determining program effectiveness*. Position Paper for the Disease Management Association of America (DMAA).

[24]Linden A, Roberts N (2005). A Users guide to the disease management literature: recommendations for reporting and assessing program outcomes. *American Journal of Managed Care*, 11, 81-90.

[25]Linden A, Adler-Milstein J (2008). Medicare disease management in policy context. *Health Care Finance Review*, 29, 1-11.

Author Notes

No conflict of interest was reported.