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 introduce the new Stata package called oda for implementing ODA from within the Stata environment. Because this package is a wrapper for the MegaODA software system, 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. 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.

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). 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
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) and the permutation \( p < 0.027 \). This Type I error rate is considered statistically significant by the \( \text{per-comparison criterion} \), but not by the \( \text{experimentwise criterion} \): 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, the effect was marginally significant.

Next, ODA is implemented using the following syntax to test the \( \text{a priori hypothesis} \) for the attribute \texttt{prestige}:

\begin{verbatim}
oDA politics prestige, pathoda(“C:\ODA\”) store(“C:\ODA\output”) iter(25000) dir(< 34 11) Sidak(2)
\end{verbatim}

As seen in the ODA output, the ODA model is interpreted as follows: “if \texttt{prestige} \leq 2.5 then predict \texttt{politics} = 34; otherwise, predict \texttt{politics} = 11.” As hypothesized, Republicans scored in the lowest two prestige quintiles. This
model correctly classified 91.67% of Republicans, and 87.50% of Democrats.

**ODA model:**

\[
\text{IF PRESTIGE} \leq 2.5 \text{ THEN POLITICS} = 34 \\
\text{IF 2.5 < PRESTIGE THEN POLITICS} = 11
\]

**Summary for Class POLITICS Attribute PRESTIGE**

<table>
<thead>
<tr>
<th>Performance Index</th>
<th>Train</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>90.00%</td>
</tr>
<tr>
<td>PAC POLITICS=34</td>
<td>91.67%</td>
</tr>
<tr>
<td>PAC POLITICS=11</td>
<td>87.50%</td>
</tr>
<tr>
<td>Effect Strength PAC</td>
<td>79.17%</td>
</tr>
<tr>
<td>PV POLITICS=34</td>
<td>91.67%</td>
</tr>
<tr>
<td>PV POLITICS=11</td>
<td>87.50%</td>
</tr>
<tr>
<td>Effect Strength PV</td>
<td>79.17%</td>
</tr>
<tr>
<td>Effect Strength Total</td>
<td>79.17%</td>
</tr>
</tbody>
</table>

Monte Carlo summary (Fisher randomization):

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Iterations:</td>
<td>25000</td>
</tr>
<tr>
<td>Estimated p:</td>
<td>0.000480</td>
</tr>
<tr>
<td>Sidak Adjusted (2) p:</td>
<td>.00095977</td>
</tr>
</tbody>
</table>

The ESS of 79.17% meets the criterion for a strong effect.\textsuperscript{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.\textsuperscript{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\textsuperscript{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.\textsuperscript{21-34}

**References**

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.
8Linden 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.

9Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Exploratory hypothesis, binary class variable, and binary attribute. *Optimal Data Analysis*, 9, 74-78.

10Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Confirmatory hypothesis, binary class variable, and binary attribute. *Optimal Data Analysis*, 9, 94-98.


**Author Notes**

No conflicts of interest were reported.