Optimal Data Analysis
Vol. 5 (November 17, 2016), 218-222

2155-0182/10/$3.00

Novometric *vs.* Log-Linear Model: Intergenerational Occupational Mobility of White American Men

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

Prior research[1] modeling intergenerational occupational mobility (professional and managerial=1; clerical and sales=2; craftsman=3; operatives and laborers=4; farmers=5) using several log-linear model approaches failed to identify a model representing a satisfactory fit of the data (p. 66). For these data an exploratory novometric analysis[2-28] (constrained by the investigator *a priori* to obtain stable accuracy in LOO analysis) predicting son's occupation (class variable) as a function of father's occupation (multicategorical attribute) identified a parsimonious, relatively strong model.

SAS™ code used to construct the data analyzed herein[1] is given in the Appendix. Novometric analysis identified a single two-strata model that had stable classification performance in LOO analysis: if father=farmer, predict son=farmer; otherwise predict son=not a farmer. Table 1 is the confusion matrix for this model (relatively strong ESS=60.03, D=1.33, $p<0.001$). As seen, the model accurately classified 3 in 4 of sons who were not farmers (50% sensitivity is expected by chance for each class category in two-category designs without analytic weights[2,30-35]), and 7 in 8 sons who were farmers.

Table 1: Novometric Model Confusion Matrix

		Son's Predicted Occupation		
		Non-Farmer	Farmer	
Son's Actual	Non-Farmer	2312	822	73.8%
Occupation	Farmer	36	226	86.3%

The novometric model indicates farming is a particularly stable occupation considered relative to the other occupations represented in the sample. The failure of the novometric model to identify a more granular separation of sons with non-farming occupations suggests that the other occupational categories are chaotic—that is, data are either widely dispersed off-diagonal and/or model performance is unstable in cross-generalizability analysis, and/or that the category-coding scheme integrates occupations that induce Simpson's paradox (e.g., masking effects that in reality exist).[2,36-37]

Whereas it is not possible to evaluate the latter possibility via the present data, ODA may be used to shed light on the former possibilities. Using ODA to predict sons' occupation (5-category class variable) as a linear function (data are hypothesized to fall in the major diagonal of the table, indicating perfect stability[2,30,38-41]) returns

2155-0182/10/$3.00

a relatively weak effect: ESS=23.42, D=16.35, p<0.001. Setting the correctly classified cells to zero and conducting an exploratory ODA to model off-diagonal cells returned a relatively weak but stable (in LOO analysis) effect: ESS= 17.34, D=23.84, p< 0.001. Iterating this procedure twice more[42] extracts a total of four ODA models, and integrating their performance yields very strong ESS=90.00, D=100/(90.00/20)-20= 2.22. Compared to the two-strata novometric model (D=1.33), this complex ODA model is 2.22/1.33 or 66.9% further from its corresponding theoretically ideal model.

The novometric model—which is closest to theoretically ideal among all possible models in this application—clearly indicates the need for increased granularity in coding occupational groupings other than (and including, since the model sensitivity was imperfect) the farming category. However, the ODA model identifies reliable, cross-generalizable patterns of change in the off-diagonal data. For applied purposes, such information may be mission-critical: in such cases the theoretical shortcoming of the complex model as indicated by a large D statistic is secondary to the clinical or substantive strength of the complex model as indicated by strong predictive values.[2,30,43-44]

References

[1]Knoke D, Burke PJ (1980). *Log-linear models*. Beverly Hills, CA: Sage (pp. 47-48).

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

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

[4]Yarnold PR, Bennett CL (2016). Novometrics *vs*. correlation: Age and clinical measures of PCP survivors, *Optimal Data Analysis, 5*, 74-78.

[5]Yarnold PR, Bennett CL (2016). Novometrics *vs*. multiple regression analysis: Age and clinical measures of PCP survivors, *Optimal Data Analysis, 5*, 79-82.

[6]Yarnold PR (2016). Novometrics *vs*. regression analysis: Literacy, and age and income, of ambulatory geriatric patients. *Optimal Data Analysis, 5*, 83-85.

[7]Yarnold PR (2016). Novometrics *vs*. regression analysis: Modeling patient satisfaction in the Emergency Room. *Optimal Data Analysis, 5*, 86-93.

[8]Yarnold PR (2016). Matrix display of pairwise novometric associations for ordered variables. *Optimal Data Analysis, 5*, 94-101.

[9]Yarnold PR, Batra M (2016). Matrix display of pairwise novometric associations for mixed-metric variables. *Optimal Data Analysis, 5*, 104-107.

[10]Yarnold PR (2016). Novometrics *vs*. ODA *vs*. One-Way ANOVA: Evaluating comparative effectiveness of sales training programs, and the importance of conducting LOO with small samples. *Optimal Data Analysis, 5*, 131-132.

[11]Yarnold PR (2016). Parental smoking behavior, ethnicity, gender, and the cigarette smoking behavior of high school students. *Optimal Data Analysis, 5*, 136-140.

[12]Yarnold PR (2016). Using gender of an imaginary rated smoker, and subject's gender, ethnicity, and smoking behavior to identify perceived differences in peer-group smoking standards of American high school students. *Optimal Data Analysis, 5*, 141-143.

[13]Yarnold PR (2016). Novometric models of smoking habits of male and female friends of American college undergraduates: Gender, smoking, and ethnicity. *Optimal Data Analysis, 5*, 146-150.

2155-0182/10/$3.00

[14]Yarnold PR (2016). Predicting daily television viewing of senior citizens using education, age and marital status. *Optimal Data Analysis*, *5*, 151-152.

[15]Yarnold PR (2016). Novometric statistical analysis and the Pearson-Yule debate. *Optimal Data Analysis*, *5*, 162-165.

[16]Yarnold PR (2016). Comparing WAIS-R qualitative information for people 75 years and older, with *vs*. without brain damage. *Optimal Data Analysis*, *5*, 166-170.

[17]Yarnold PR (2016). Using novometrics to disentangle complete sets of sign-test-based multiple-comparison findings. *Optimal Data Analysis*, *5*, 175-176.

[18]Yarnold PR (2016). Novometric analysis *vs*. MANOVA: MMPI codetype, gender, setting, and the MacAndrew Alcoholism scale. *Optimal Data Analysis*, *5*, 177-178.

[19]Yarnold PR (2016). Novometric *vs*. ODA reliability analysis *vs*. polychoric correlation with relaxed distributional assumptions: Inter-rater reliability of independent ratings of plant health. *Optimal Data Analysis*, *5*, 179-183.

[20]Yarnold PR (2016). Novometrics *vs*. polychoric correlation: Number of lambs born over two years. *Optimal Data Analysis*, *5*, 184-185.

[21]Yarnold PR (2016). Comparing MMPI-2 *F-K* Index normative data among male and female psychiatric and head-injured patients, individuals seeking disability benefits, police and priest job applicants, and substance abusers. *Optimal Data Analysis*, *5*, 186-193.

[22]Yarnold PR, Linden A (2016). Theoretical aspects of the D statistic. *Optimal Data Analysis*, *5*, 171-174.

[23]Yarnold PR (2016). Novometric analysis predicting voter turnout: Race, education, and organizational membership status. *Optimal Data Analysis*, *5*, 194-197.

[24]Yarnold PR (2016). Novometrics *vs*. Yule's Q: Voter turnout and organizational membership. *Optimal Data Analysis*, *5*, 198-199.

[25]Yarnold PR (2016). Novometric *vs*. recursive causal analysis: The effect of age, education, and region on support of civil liberties. *Optimal Data Analysis*, *5*, 200-203.

[26]Yarnold PR (2016). Novometric analysis *vs*. GenODA *vs*. log-linear model: Temporal stability of the association of presidential vote choice and party identification. *Optimal Data Analysis*, *5*, 204-207.

[27]Yarnold PR (2016). Novometric analysis *vs*. ODA *vs*. log-linear model in analysis of a two-wave panel design: Assessing temporal stability of Catholic party identification in the 1956-1960 SRC panels. *Optimal Data Analysis*, *5*, 208-212.

[28]Yarnold PR (2016). GenODA structural decomposition *vs*. log-linear model of one-step Markov transition data: Stability and change in male geographic mobility in 1944-1951 and 1951-1953. *Optimal Data Analysis*, *5*, 213-215.

[29]Bryant FB, Harrison PR (2013). How to create an ASCII input data file for UniODA and CTA software. *Optimal Data Analysis*, *2*, 2-6.

[30]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington, DC, APA Books.

[31]Linden A, Yarnold PR (In Press). Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. *Journal of Evaluation in Clinical Practice*.

2155-0182/10/$3.00

[32]Linden A, Yarnold PR (In Press). Combining machine learning and matching techniques to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*.

[33]Linden A, Yarnold PR (In Press). Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*. DOI: 10.1111/jep.12538

[34]Linden A, Yarnold PR (In Press). Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice*.

[35]Yarnold PR, Soltysik RC (1991). Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences, 22*, 739-752.

[36]Yarnold PR (1996). Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement, 56*, 430-442.

[37]Soltysik RC, Yarnold PR (2010). The use of unconfounded climatic data improves atmospheric prediction. *Optimal Data Analysis, 1*, 67-100.

[38]Yarnold PR (2014). How to assess inter-observer reliability of ratings made on ordinal scales: Evaluating and comparing the Emergency Severity Index (Version 3) and Canadian Triage Acuity Scale. *Optimal Data Analysis, 3*, 42-49.

[39]Yarnold PR (2014). How to assess the inter-method (parallel-forms) reliability of ratings made on ordinal scales: Evaluating and comparing the Emergency Severity Index (Version 3) and Canadian Triage Acuity Scale. *Optimal Data Analysis, 3*, 50-54.

[40]Yarnold PR (2015). Estimating inter-rater reliability using pooled data induces paradoxical confounding: An example involving Emergency Severity Index triage ratings. *Optimal Data Analysis, 4*, 21-23.

[41]Yarnold PR (2016). Novometric *vs*. ODA reliability analysis *vs*. polychoric correlation with relaxed distributional assumptions: Inter-rater reliability of independent ratings of plant health. *Optimal Data Analysis, 5*, 179-183.

[42]Yarnold PR (2015). UniODA-based structural decomposition vs. legacy linear models: Statics and dynamics of intergenerational occupational mobility. *Optimal Data Analysis, 4*, 194-196.

[43]Stalans LJ, Yarnold PR, Seng M, Olson DE, Repp M (2004). Identifying three types of violent offenders and predicting violent recidivism while on probation: A classification tree analysis. *Law & Human Behavior, 28*, 53-271.

[44]Yarnold PR (2013). Standards for reporting UniODA findings expanded to include ESP and all possible aggregated confusion tables. *Optimal Data Analysis, 2*, 106-119.

Author Notes

This study analyzed publically available data. No conflict of interest was reported.

Mail: Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., #163
Chicago, IL 60646
USA

2155-0182/10/$3.00

Appendix

SAS™ Code used to Construct (Reproduce[1]) the Data File for Analysis by ODA Software[2,29]

```
data real;
infile datalines;
input row column;
cards;
1 1
;
Data example;
Do n=1 to 152;
put '1 1';
end;
Do n=1 to 66;
put '1 2';
end;
Do n=1 to 33;
put '1 3';
end;
Do n=1 to 39;
put '1 4';
end;
Do n=1 to 4;
put '1 5';
end;
Do n=1 to 201;
put '2 1';
end;
Do n=1 to 159;
put '2 2';
end;
Do n=1 to 72;
put '2 3';
end;
Do n=1 to 80;
put '2 4';
end;
Do n=1 to 8;
put '2 5';
end;
Do n=1 to 138;
put '3 1';
end;
Do n=1 to 125;
put '3 2';
end;
Do n=1 to 184;
put '3 3';
end;
Do n=1 to 172;
put '3 4';
end;
Do n=1 to 7;
put '3 5';
end;
Do n=1 to 143;
put '4 1';
end;
Do n=1 to 161;
put '4 2';
end;
Do n=1 to 209;
put '4 3';
end;
Do n=1 to 378;
put '4 4';
end;
Do n=1 to 17;
put '4 5';
end;
Do n=1 to 98;
put '5 1';
end;
Do n=1 to 146;
put '5 2';
end;
Do n=1 to 207;
put '5 3';
end;
Do n=1 to 371;
put '5 4';
end;
Do n=1 to 226;
put '5 5';
end;
Output;
Run;
```