Optimal Data Analysis
Vol. 3 (September 29, 2014), 107-109

2155-0182/10/$3.00

Increasing the Validity and Reproducibility of Scientific Findings

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

Globally optimal statistical methods eliminate the greatest challenges to the validity and reproducibility of findings in the literature.

In my view ubiquitous use of linear statistical models (LSMs) to aide in "understanding" of sample data ranks among the greatest challenges to the validity and reproducibility of empirical findings reported in the literature.

By far, the most widely-reported LSMs reflect parametric methods—which clearly are the most challenged of all LSM formulations. For parametric LSMs an omnipresent challenge is the ability of sample data to comply with all of the underlying assumptions. Violation of any crucial assumption(s) underlying any method is problematic because any and all such violation undermines both internal and external validity of findings obtained by the method. Extorting the virtue of "robustness over violations" begs the question of how "incorrect" can something be, and still be considered "correct".[1-3]

An inherent limitation of parametric LSMs is inaccuracy: most models are only capable of accurately predicting values close to the sample mean or mode (for variance- and maximum-likelihood function-based methods, respectively). For example, weak accuracy is obtained by correlation and multiple regression analysis-based LSMs[4-6] and by chi-square and logistic regression analysis-based LSMs.[7-11]

To eradicate both issues for problems involving a binary class variable and multiple attributes ("independent variables") an optimal (maximum-accuracy) LSM method called MultiODA was created.[12-13] An analogue to logistic regression analysis, MultiODA requires no assumptions and explicitly proves maximum-accuracy: it not only identifies more accurate LSMs than parametric methods, it sometimes identifies accurate LSMs in applications where parametric models find nothing.[14-15] Indeed, a MultiODA model involving the use of *unit-weight beta coefficients* is more accurate and parsimonious than parametric models.[16]

Multicategorical attributes—categorical variables with three or more possible response levels—are inherently difficult for all LSMs. Such attributes are rearranged using "reference groups".[17] Type I error changes as a function of reference group definition which determines the constitution of the design matrix: as the number of levels of a multicategorical variable increases the design matrix can rapidly be overwhelmed.[18] This is easily seen by reconstructing a problem originally developed for logistic regression analysis instead as a log-linear model.[1-2] Use of reference groups decreases parsimony of LSMs, and imperfectly-specified reference groups can reduce model accuracy.[19-20] Arbitrary parsing of ordered variables can influence both Type I error and model accuracy in all methods.[21]

2155-0182/10/$3.00

The most important inherent challenge for all LSMs is avoiding Simpson's Paradox, a phenomenon whereby pooling (combining) of different groups (e.g., ethnic categories) produces spurious confounding and useless findings.[22-24] Explicitly optimal (maximum accuracy) classification tree analysis (CTA) was created to eliminate all shortcomings of LSMs, including paradoxical confounding.[25-26] Initial research using CTA reported strongest models obtained in some areas of research.[27-28] Most recently, algorithms were discovered to identify globally optimal models for a given sample.[29-30] New statistical methods eliminate challenges to validity and reproducibility of findings, and offer promise of increasing the accuracy and efficiency of programmatic research.

References

[1]Grimm, L.G., & Yarnold, P.R. (Eds.). *Reading and Understanding Multivariate Statistics*. Washington, D.C.: APA Books, 1995.

[2]Grimm, L.G., & Yarnold, P.R. (Eds.). *Reading and Understanding More Multivariate Statistics*. Washington, D.C.: APA Books, 2000.

[3]Yarnold, P.R., & Soltysik, R.C. *Optimal data analysis: Guidebook with software for Windows*. Washington, D.C.: APA Books, 2005.

[4]Yarnold, P.R., Bryant, F.B., & Soltysik, R.C. (2013). Maximizing the accuracy of multiple regression models via UniODA: Regression *away from* the mean. *Optimal Data Analysis*, *2*, 19-25.

[5]Yarnold, P.R. (2013). Maximum-accuracy multiple regression analysis: Influence of registration on overall satisfaction ratings of emergency room patients. *Optimal Data Analysis*, *2*, 72-75.

[6]Yarnold, P.R. (2013). Assessing technician, nurse, and doctor ratings as predictors of overall satisfaction ratings of Emergency Room patients: A maximum-accuracy multiple regression analysis. *Optimal Data Analysis*, 2, 76-85.

[7]Yarnold, P.R., & Soltysik, R.C. (1991). Refining two-group multivariable classification models using univariate optimal discriminant analysis. *Decision Sciences*, *22*, 1158-1164.

[8]Yarnold, P.R., Hart, L.A., & Soltysik, R.C. (1994). Optimizing the classification performance of logistic regression and Fisher's discriminant analyses. *Educational and Psychological Measurement*, *54*, 73-85.

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

[10]Yarnold, P.R. (2014). UniODA *vs*. chi-square: Audience effect on smile production in infants. *Optimal Data Analysis*, *3*, 3-5.

[11]Yarnold, P.R. (2014). UniODA *vs*. chi-square: Discriminating inhibited and uninhibited infant profiles. *Optimal Data Analysis*, *3*, 9-11.

[12]Soltysik, R.C., & Yarnold, P.R. (2010). Two-group MultiODA: Mixed-integer linear programming solution with bounded M. *Optimal Data Analysis*, *1*, 31-37.

[13]Shinmura, S. (2014). Improvement of CPU time of linear discriminant functions based on MNM criterion by IP. *Statistics, Optimization and Information Computing*, *2*, 114-129.

[14]Yarnold, P.R., Soltysik, R.C., & Martin, G.J. (1994). Heart rate variability and susceptibility for sudden cardiac death: An example of multivariable optimal discriminant analysis. *Statistics in Medicine*, *13*, 1015-1021.

[15]Yarnold, P.R., Soltysik, R.C., McCormick, W.C., Burns, R., Lin, E.H.B., Bush, T., & Martin, G.J. (1995). Application of multivaria-

2155-0182/10/$3.00

ble optimal discriminant analysis in general internal medicine. *Journal of General Internal Medicine*, *10*, 601-606.

[16]Yarnold, P.R., Soltysik, R.C., Lefevre, F, & Martin, G.J. (1998). Predicting in-hospital mortality of patients receiving cardiopulmonary resuscitation: Unit-weighted MultiODA for binary data. Statistics in Medicine, 17, 2405-2414.

[17]Yarnold, P.R., & Bryant, F.B. (2013). Analysis involving categorical attributes having many categories. *Optimal Data Analysis*, *2*, 69-70.

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

[19]Yarnold, P.R. (2010). Aggregated *vs*. referenced categorical attributes in UniODA and CTA. *Optimal Data Analysis*, *1*, 46-49.

[20]Yarnold, P.R. (2010). Unconstrained covariates in CTA. *Optimal data Analysis*, *1*, 38-40.

[21]Yarnold, P. (2014). "Breaking-up" an ordinal variable can reduce model classification accuracy. *Optimal Data Analysis*, *3*, 19.

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

[23]Yarnold, P.R. (2013). Ascertaining an individual patient's *symptom dominance hierarchy*: Analysis of raw longitudinal data induces Simpson's Paradox. *Optimal Data Analysis*, *2*, 159-171.

[24]Soltysik, R.C., & Yarnold, P.R. (2010). The use of unconfounded climatic data improves atmospheric prediction. *Optimal Data Analysis*, *1*, 67-100.

[25]Yarnold, P.R., Soltysik, R.C., & Bennett, C.L. (1997). Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: An example of hierarchically optimal classification tree analysis. *Statistics in Medicine*, *16*, 1451-1463.

[26]Soltysik, R.C, & Yarnold, P.R. (2010). Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis*, *1*, 144-160.

[27]Yarnold, P.R. (2013). Initial use of hierarchically optimal classification tree analysis in medical research. *Optimal Data Analysis*, 2, 7-18.

[28]Bryant, F.B. (2010). The Loyola experience (1993-2009): Optimal data analysis in the Department of Psychology. *Optimal Data Analysis*, *1*, 4-9.

[29]Yarnold, P.R., & Soltysik, R.C. (2014). Globally optimal statistical classification models, I: Binary class variable, one ordered attribute. *Optimal Data Analysis*, *3*, 55-77.

[30]Yarnold, P.R., Soltysik, R.C. (2014). Globally optimal statistical classification models, II: Unrestricted class variable, two or more attributes. *Optimal Data Analysis*, *3*, 78-84.

Author Notes

eMail: Journal@OptimalDataAnalysis.com