Optimal Data Analysis
Vol. 4 (December 16, 2015), 173-174

Copyright 2015 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Generalized Linear Interactive Modeling: Four Wrongs Don't Make a Right

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

Several examples used to illustrate generalized linear interactive modeling (GLIM) violate crucial assumptions underlying chi-square, advocate arbitrary parsing of attributes, and conduct statistically unmotivated agglomeration of class categories. Violating assumptions call the validity of the estimated effect and associated Type I error rate into question, and arbitrary parsing and agglomerating procedures can reduce model classification accuracy at best, and mask effects that exist or identify paradoxical effects at worst.

Why are so many articles published that contain obvious statistical errors? For example, why is chi-square often used for analysis of ordered attributes?[1] It should be common knowledge among researchers that chi-square is appropriate for analysis of attributes measured using what is known as a nominal, qualitative, or categorical measurement scale. Where are students and professional scientists and educators learning these incorrect statistical methods? Why do journal reviewers and editors publish papers containing obvious statistical errors? Could part of the explanation be that many (most that I've examined) texts used to teach statistical methods espouse and include examples of incorrect statistical methods? An example, by no means unique, is a book discussing GLIM analysis.[2]

For example, near the beginning of this book a table is given that shows the distribution of data in a 5 x 5 contingency table in which 18 of the 25 cells have an expected value of less than five—a clear violation of the minimum expectation assumption (Table 1.2, p. 9).[3]

Later, in discussion about preparing data for analysis, attention turned to using respondent age as an attribute: "The first step is to classify the (age) data into ten-year age groups" (p. 25). Should all studies including age (no matter what focus of investigation) be transformed in this manner—what is the statistical motivation? What is the criterion for deciding which of the ordered attributes should be transformed in this manner: is age the only attribute that should be transformed in this manner, or does this procedure equally apply to all ordered attributes? It is easy to demonstrate that arbitrary parsing of attributes can reduce the accuracy of classification methods[4] and may induce paradoxical confounding (e.g., see pp. 41-42 and 68-69 of the book[2]).[1,5]

The book continues: "(This) computes a new variable … (that transforms) the Age variable … into these ten year categories (p. 25)." The intention is to use statistical methods (e.g., log-linear model) that treat the attribute as being categorical, when in reality the transformed

Copyright 2015 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

attribute remains ordered. Confusing ordinal scales (e.g., Likert-type scores) for nominal scales is a commonly repeated error in the literature.[2,6]

Finally, throughout the book examples are given of marginal subtables examined in a search for statistically significant effects: a form of analytic fishing. Arbitrary inclusion or exclusion of specific attributes in an application can induce paradoxical confounding, and it is best to use algorithms capable of identifying non-confounded models that explicitly maximize (weighted) classification accuracy as well as parsimony.[7,8]

References

[1]Yarnold PR, Soltysik RC (In Review). *Maximizing predictive accuracy*. Chicago, IL: ODA Books.

[2]Gilbert N (1993). Analyzing tabular data: Loglinear and logistic models for social researchers. London, England: UCL Press.

[3]Yarnold JK (1970). The minimum expectation in χ^2 goodness of fit tests and the accuracy of approximations for the null distribution. *Journal of the American Statistical Association*, *65*, 864-886. URL: http://www.jstor.org/stable/2284594

[4]Yarnold PR (2014). “Breaking-up” an ordinal variable can reduce model classification accuracy. *Optimal Data Analysis*, *3*, 19. URL: http://optimalprediction.com/files/pdf/V3A7.pdf

[5]Yarnold PR (1996). Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement*, *56*, 430-442. DOI: 10.1177/0013164496056003005

[6]Yarnold PR, Soltysik RC (2010). UniODA vs. chi-square: Ordinal data sometimes feign categorical. *Optimal Data Analysis*, *1*, 62-66. URL: http://optimalprediction.com/files/pdf/V1A12.pdf

[7]Yarnold PR (2015). Optimal statistical analysis involving a confounding variable. *Optimal Data Analysis*, *4*, 87-103. URL: http://optimalprediction.com/files/pdf/V4A16.pdf

[8]Yarnold PR (2015). Optimal statistical analysis involving multiple confounding variables. *Optimal Data Analysis*, *4*, 107-112. URL: http://optimalprediction.com/files/pdf/V4A18.pdf

Author Notes

The study analyzed de-individuated data and was exempt from Institutional Review Board review. No conflict of interest was reported.

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