2015

The Role of Residuals in Optimal and Suboptimal Statistical Modeling

Paul R. Yarnold
Optimal Data Analysis LLC

Fred B. Bryant
Loyola University Chicago, fbryant@luc.edu

Optimal Data Analysis
Vol. 4 (December 27, 2014), 9-11

2155-0182/10/$3.00

The Role of Residuals in Optimal and Suboptimal Statistical Modeling

Paul R. Yarnold, Ph.D., and Fred B. Bryant, Ph.D.
Optimal Data Analysis, LLC Loyola University Chicago

This note contrasts the importance of the analysis of model residual values in assessing the *invalidity* of estimated Type I error rates for parametric methods, versus in determining ways of improving the *validity* of maximum-accuracy methods.

Analysis of residuals—the difference between the predicted and actual values of *observations* with respect to the dependent variable—is important in assessing the validity of parametric statistical methods.[1,2] In particular, a crucial assumption is that the residuals are normally distributed: failing this assumption threatens the validity of Type I error estimates. This is an important limitation of suboptimal[3] methods, as residuals are always greatest for absolutely extreme values of the dependent measure for general linear model-based methods (e.g., ordinary least-squares regression), and for the smallest class (category) of the dependent measure for maximum-likelihood-based methods (e.g., logistic regression analysis).[4,5] Another limitation of suboptimal methods is there is no established algorithmic procedure for assessing how different independent variables or their interactions are specifically related to the value of residuals for individual observations.

Residual values are also an integral part of structural equation modeling (SEM), which uses a "fitting function" to obtain parameter estimates that minimize the size of the residuals between the elements of the *observed* covariance matrix based on the set of measured variables being analyzed (S) and the elements of the *predicted* covariance matrix implied by the parameter estimates in the model (Σ). The most commonly used method of estimation in SEM is maximum-likelihood, which finds parameter estimates that maximize the likelihood that the fitted residuals (S – Σ) are due to chance.[6,7] In SEM, the overall size of residuals is used to assess a structural model's goodness-of-fit to the data (e.g., via a chi-square value testing the statistical significance of the size of fitted residuals, or descriptive fit indices reflecting the average size of residuals); individual elements in the matrix of fitted residuals can be inspected to identify specific relationships between measured variables that the model explains poorly; and the model can be modified to include additional estimated parameters to improve its fit to the data. Note that this statistical method does NOT address the residuals associated with individual *observations*.

In contrast, in the optimal (maximum-accuracy) data analysis (ODA) paradigm no distributional assumptions underlie theoretical distributions of optima, so the validity of the Type I error rate is never in doubt.[8-11] However, in the ODA paradigm the analysis of residuals is

2155-0182/10/$3.00

arguably the most important aspect of an analysis—in terms of assessing ways in which prediction of observations' actual class categories can be improved. Residuals tell one what remains to be explained. The ultimate objective is to eliminate all such errors—that is, to correctly classify all of the observations in the sample.

Compared to suboptimal methods, the ability of residuals to indicate ways to improve statistical models is a major benefit of both the UniODA[12-26] and CTA[27,28] algorithms. Model endpoints that are homogeneous are well explained, and there is little room for further improvement; and model endpoints that are heterogeneous are poorly explained, and leave much room for improvement.[29,30] When an endpoint has a large N, and is heterogeneous, it is the most appropriate area in which to work to improve overall model performance—and thus understanding of the phenomenon.[31] It also is clear that none of the measured attributes used to find the model will help in this regard—or they would be included in the model. Clues to the characteristic nature of the observations in the targeted strata are garnered by content analysis of attributes (and their cut-point values) defining the endpoint. This not only paves the way toward fastest improvement in performance (knowledge), but it indicates what the subject inclusion criteria for future research should be (observations classified into the targeted endpoint), thereby providing "bread crumbs" pointing the way to new attributes to study. In a word, residuals lie at the heart of the matter.

References

[1]Grimm LG, Yarnold PR (Eds.). *Reading and Understanding Multivariate Statistics*. Washington, DC: APA Books, 1995.

[2]Grimm LG, Yarnold PR (Eds.). *Reading and Understanding More Multivariate Statistics*. Washington, DC: APA Books, 2000.

[3]Yarnold PR (2014). "A statistical guide for the ethically perplexed" (Chapter 4, Panter & Sterba, *Handbook of Ethics in Quantitative Methodology*, Routledge, 2011): Clarifying disorientation regarding the etiology and meaning of the term *Optimal* as used in the Optimal Data Analysis (ODA) paradigm. *Optimal Data Analysis*, *3*, 30-31.

[4]Yarnold PR, Bryant FB, Soltysik RC (2013). Maximizing the accuracy of multiple regression models via UniODA: Regression *away from* the mean. *Optimal Data Analysis*, *2*, 19-25.

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

[6]Bollen KA (1989). Structural equations with latent variables. New York: Wiley.

[7]Kline, RB. (2011). *Principles and practice of structural equation modeling* (3rd ed). New York: Guilford Press.

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

[9]Soltysik RC, Yarnold PR (1994). Univariable optimal discriminant analysis: One-tailed hypotheses. *Educational and Psychological Measurement*, *54*, 646-653.

[10]Carmony L, Yarnold PR, Naeymi-Rad F (1998). One-tailed Type I error rates for balanced two-category UniODA with a random ordered attribute. *Annals of Operations Research*, *74*, 223-238.

[11]Yarnold PR, Soltysik RC (2010). Optimal data analysis: A general statistical analysis paradigm. *Optimal Data Analysis*, *1*, 10-22.

2155-0182/10/$3.00

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

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

[14]Yarnold PR (2010). GenUniODA *vs*. log-linear model: Modeling discrimination in organizations. *Optimal Data Analysis, 1*, 59-61.

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

[16]Yarnold PR (2014). UniODA *vs*. t-Test: Comparing two migraine treatments. *Optimal Data Analysis*, *3*, 6-8.

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

[18]Yarnold PR (2014). UniODA *vs*. weighted kappa: Evaluating concordance of clinician and patient ratings of the patient's physical and mental health functioning. *Optimal Data Analysis*, *3*, 12-13.

[19]Yarnold PR (2014). UniODA *vs*. kappa: Evaluating the long-term (27-year) test-retest reliability of the Type A Behavior Pattern. *Optimal Data Analysis*, *3*, 14-16.

[20]Yarnold PR (2014). UniODA *vs*. logistic regression analysis: Serum cholesterol and coronary heart disease and mortality among middle aged diabetic men. *Optimal Data Analysis*, *3*, 17-18.

[21]Yarnold PR (2014). UniODA *vs*. polychoric correlation: Number of lambs born over two years. *Optimal Data Analysis*, *3*, 113-114.

[22]Yarnold PR (2014). UniODA *vs*. Bray-Curtis dissimilarity index for count data. *Optimal Data Analysis*, *3*, 115-116.

[23]Yarnold PR (2014). UniODA vs. ROC analysis: Computing the "optimal" cut-point. *Optimal Data Analysis*, *3*, 117-120.

[24]Yarnold PR (2014). UniODA vs. Kendall's Coefficient of Concordance (*W*): Multiple rankings of multiple movies. *Optimal Data Analysis*, *3*, 121-123.

[25]Yarnold PR (2014). UniODA vs. Mann-Whitney *U* test: Sunlight and petal width. *Optimal Data Analysis*, *4*, 3-5.

[26]Yarnold PR (2014). UniODA vs. Mann-Whitney *U* test: Comparative effectiveness of laxatives. *Optimal Data Analysis*, *4*, 6-8.

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

[28]Soltysik RC, Yarnold PR (2010). Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis, 1*, 144-160.

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

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

[31]Yarnold PR (2014). Triage algorithm for chest radiography for community-acquired pneumonia of Emergency Department patients: Missing data cripples research. *Optimal Data Analysis*, *3*, 102-106.

Author Notes

eMail: Journal@OptimalDataAnalysis.com.

eJournal: http://ODAJournal.com