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

Follow this and additional works at: https://ecommons.luc.edu/psychology_facpubs

Part of the Psychology Commons

Recommended Citation
Yarnold, Paul R. and Bryant, Fred B.. The Role of Residuals in Optimal and Suboptimal Statistical Modeling. Optimal Data Analysis, 4, : 9-11, 2015. Retrieved from Loyola eCommons, Psychology: Faculty Publications and Other Works,

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 License. © Optimal Data Analysis LLC, 2015.
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 suboptimal3 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
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 UniODA12-26 and CTA27,28 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


Yarnold PR, Soltysik RC (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

eJournal: http://ODAJournal.com