Optimal Data Analysis
Vol. 4 (January 12, 2015), 12-13

2155-0182/10/$3.00

Maximizing ESS of Regression Models in Applications with Dependent Measures with Domains Exceeding Ten Values

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

An inherent limitation of ordinary least-squares regression models is that dependent measure values near the mean for the sample are predicted well, while values greater or less than the mean are predicted poorly. A UniODA-based methodology for circumventing this limitation and explicitly maximizing classification accuracy (ESS) has been demonstrated for integer dependent variables having ten or fewer levels. This note describes three methods for extending this methodology for problems involving dependent measures having a larger domain of values.

It has been demonstrated that the well-known characteristic of ordinary least-squares regression models to predict dependent variable values near the mean, and to fail to predict values above or below the mean ("regression *toward* the mean") can be reversed ("regression *away* from the mean") using UniODA.[1-4] Prior demonstrations involved integer dependent variables with a maximum domain of ten—corresponding to the maximum number of class categories that can be analyzed by UniODA software.[5] This research note briefly describes three methods to accomplish this analysis for applications involving a dependent variable (the class variable in UniODA) having a domain of greater than ten values.

The first option is to simply divide the dependent measure into ten categories (e.g., deciles; domain/10), and replicate the analysis described in the prior papers.

If an exactly granular answer is instead desired (and the domain is at least 20), the second option involves employing weighted optimization.[5] Imagine a dependent measure having a maximum of 100 integer levels. First transform the dependent measure variable using integer [dependent measure/10]. Thus, values of 1-10 are transformed into a value of 1; values >10 to 20 are transformed into a value of 2; and values >90 to 100 are transformed into a value of 10. Within each of the ten "bins", a weight is used to define each of the 10 values: for a score of 1 the weight is 1.1; for a score of 2 the weight

2155-0182/10/$3.00

is 1.2; for a score of 9 the weight is 1.9; for a weight of 57 the weight is 1.7. The UniODA analysis uses integer [dependent measure/10] as the class variable, uses the dependent measure (X or function of X for an application having one, or two or more independent variables, respectively) as the attribute; and assigns the weight to complete the specification.[5] UniODA then maximizes the weighted accuracy. If desired—in order to predict outlying values with greatest accuracy, weights could be configured to reflect absolute distance of the bin from the mean (regression toward the *mean*, not toward the *median*, is being addressed). Alternatively, increasingly extreme weights can be assigned to outlying values (bins) until an *a priori* defined satisfactory level of accuracy is achieved.[6] It is of course essential to estimate the potential cross-generalizability of such models using leave-one-out, hold-out, or bootstrap methods.

The third possibility, if the sample is sufficiently large, is to use multiple optimized regression models: one to predict integer [dependent measure/10], and a second (or as many as required depending on the magnitude of the domain) optimized regression model within each bin to predict weighted values.

These methods still suffer from the likely problem of not-normally-distributed residuals for the initial regression models, in violation of a crucial assumption underlying the validity of estimated Type I error rate.[7]

References

[1]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.

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

[3]Yarnold PR (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.

[4]Yarnold PR (2013). Creating a data set with SAS™ and maximizing ESS of a multiple regression analysis model for a Likert-type dependent variable using UniODA™ and MegaODA™ software. *Optimal Data Analysis*, 2, 191-193.

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

[6]Harvey RL, Roth EJ, Yarnold PR, Durham JR, Green D (1996). Deep vein thrombosis in stroke: The use of plasma D-dimer level as a screening test in the rehabilitation setting. *Stroke*, *27*, 1516-1520. Abstracted in *American College of Physicians Journal Club*, 1997, *126*, 43.

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

Author Notes

E-mail: Journal@OptimalDataAnalysis.com.

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