5-16-2019

Logistic Discriminant Analysis and Structural Equation Modeling Both Identify Effects in Random Data

Ariel Linden
Linden Consulting Group

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

Paul R. Yarnold
Optimal Data Analysis

Optimal Data Analysis
Vol. 8 (May 16, 2019), 97-102

2155-0182/10/$3.00

Logistic Discriminant Analysis and Structural Equation Modeling Both Identify Effects in Random Data

Ariel Linden, Dr.P.H., Fred B. Bryant, Ph.D. and Paul R. Yarnold, Ph.D.
Linden Consulting Group, LLC Loyola University Chicago Optimal Data Analysis, LLC

Recent research compared the ability of various classification algorithms [logistic regression (LR), random forests (RF), support vector machines (SVM), boosted regression (BR), multi-layer perceptron neural net model (MLP), and classification tree analysis (CTA)] to correctly *fail to identify* a relationship between a binary class (dependent) variable and ten *randomly generated* attributes (covariates): only CTA failed to find a model. We use the same ten-variable N=1,000 dataset to assess training classification accuracy of models developed by logistic discriminant analysis (LDA), generalized structural equation modelling (GSEM), and robust diagonally-weighted least-squares (DWLS) SEM for binary outcomes. Except for CTA, all machine-learning algorithms assessed thus far have identified training effects in random data.

Recent research compared predictive accuracy obtained by CTA *vs*. by LR, RF, SVM, BR and MLP algorithms.[1-3] Prior research used artificial data involving 500 "group 1" and 500 "group 2" observations. Observations were independently assigned a random continuous value for each of ten covariates (attributes)—that by design have no association with the dichotomous dependent (class) variable. Among all of these algorithms *only CTA* correctly *failed* to discriminate the two groups (no CTA model emerged)—all other methods found a viable model in random data.

Using the same data, this study assesses if a consistent finding occurs for models which are identified by logistic discriminant analysis (LDA), or by generalized structural equation modeling (GSEM) for binary outcomes.

LDA

Rather than making assumptions regarding the distribution of the data and the residual scores within each group, LDA assumes the likelihood ratios of the groups have an exponential form. Multinomial logistic regression is the analytic methodology used to obtain the LDA model.[4]

As done previously a receiver operating characteristics (ROC) analysis[5] was conducted treating actual class status as the reference variable, and predicted probabilities from the model as the classification variable.[1-3] A model which perfectly discriminates the two groups has an AUC=1.0 (and effect strength for sensitivity or ESS=100); a model providing chance-level discrimination between groups has AUC=0.50

2155-0182/10/$3.00

(and ESS=0); and a model which misclassifies every observation in the sample has AUC=0 (and ESS=-100).[6-8]

In training analysis the ten-attribute LDA model obtained AUC=0.5665 (95% CI= 0.5310-0.6019). This corresponds to ESS=13.3, indicating a relatively weak effect.[6] Accuracy fell in cross-generalizability (hold-out) analysis, and the model 95% CI overlapped chance.

Failure of the LDA model to replicate in cross-validation reconfirms the necessity of conducting reproducibility analysis and supports the cautionary recommendation to only retain attributes having stable effects in training and LOO analysis within CTA models.[8-10]

Maximum Likelihood GSEM

GSEM is a more flexible modeling approach than SEM, as generalized linear model (GLM) is a more flexible alternative to ordinary least-squares regression. GSEM employs maximum likelihood (ML) estimation and allows the user to choose the particular distribution family and link to best fit the data at hand. In the current data, a GSEM model was fit using the Bernoulli distribution with a logit link. The results were identical to those obtained using LDA because in Stata (*Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC), GSEM derives its estimation using logistic regression, and LDA obtains estimates by using multinomial regression—which is a generalization of the logistic function.[11]

Robust Diagonally-Weighted Least-Squares (DWLS) SEM

Special-purpose SEM estimation methods are used for analysis involving binary and ordinal data.[12] For designs with a mixture of different measurement metrics, using DWLS estimation the input correlation matrix is a mixture of different correlation coefficients: *Pearson* if the variables are continuous measures; *polychoric* if the variables are ordinal measures; or *polyserial* if the variables are a continuous and an ordinal measure (it is assumed the binary measure reflects an unobserved, normally-distributed continuous variable aggregated into a binary measure). Presently, DWLS estimation in SEM was used to estimate a regression model consisting of a single, binary dependent variable predicted by ten continuous, independent variables which are allowed to correlate with one another.

A matrix of correlations among the ten continuous independent variables and the single binary outcome variable was created[13] involving 45 Pearson correlations among ten continuous variables, and ten polyserial correlations of the continuous variables and the binary outcome measure. The asymptotic covariance matrix for the 11 measured variables was employed to conduct robust estimation and correct the goodness-of-fit chi-square value and *SE*s of parameter estimates for nonnormality distortion.

SEM[14] was used to analyze these data and obtain robust DWLS estimates of unstandardized regression coefficients for the continuous independent variables, by regressing the dichotomous dependent variable on the set of ten continuous variables. Given that (a) the number of estimated parameters in the SEM is 66 [45 correlations among the independent variables] +[10 variances of the independent variables]+ [10 regression coefficients]+[1 residual variance term for the dependent variable], that (b) equals the number of elements in the covariance matrix of 11 measured variables ([11x12]/2=66), this regression analysis yields an exactly identified model with *df*=0 that, by definition, produces perfect, overall model fit (i.e., χ^2=0).

This DWLS SEM model explained 2.02% of the variance in the TREAT outcome variable, which is statistically significant: $F(10, 989)=2.0390$, $p<0.0269$. Robust DWLS parameter estimates for the regression model using the continuous variables to predict the binary outcome variable emerged for X3 (gamma=0.055, *SE*=0.0257, *Z*=2.1462, $p<0.0319$), X4 (gamma= -0.072, *SE*=0.0259, *Z*=2.7880, $p<0.0053$), and

2155-0182/10/$3.00

X10 (gamma=-0.084, *SE*=0.0261, *Z*=3.2058, *p*<0.0013)—which were statistically significant predictors of the binary dependent variable when holding constant at their mean the effects of all other predictors in the model. Standardized regression coefficients for statistically significant predictors were less than 0.10 in absolute value (considered a small effect in multiple regression analysis[15]) for X3 (β= 0.0551), X4 (β=-0.0723), and X10 (β=-0.0837).

Comments

The objective of the present paper, and of this line of research[1-3], is to focus awareness of and attention on the fact that most models—whether of classic theory or machine learning origin—are likely to find relationships in the data *that are not real*. Investigators should understand this crucial point when evaluating and placing confidence in their analytic results.

Findings obtained herein are consistent with prior research identifying an important limitation of machine-learning algorithms used for predicting binary class variables (outcomes) and to obtain propensity scores.[1-3] That is, the present study reveals that the LDA, GSEM, and DWLS SEM models are likely to find relationships in training analysis which in reality *don't exist* between variables.

Examination of model performance which is obtained in reproducibility analysis helps to inhibit such overfitting, but for some widely-used statistical analysis methods there is no standard methodology for assessing cross-generalizability. For example, SEM does not routinely use reproducibility analyses to assess the cross-sample generalizability of obtained model estimates. If the sample is very large researchers sometimes randomly split the sample in half and then fit the model to both halves to assess if identical results emerged.[16-18] Some studies with two or more independent data sets use one sample to create a training model, and use the other sample(s) to cross-validate the training model.[19,20]

Based on present results, developers of statistical software should in future program updates *for all statistical modeling approaches* add procedures which enable users to systematically assess reproducibility of obtained results, and thereby provide crucial safeguards against falling prey to chance. This is *not* an issue for ODA[6] and CTA[21] methods, for which a host of reproducibility analyses (e.g., jackknife, bootstrap, split-half, K-fold, holdout, and test-retest) *by axiom* are used in evaluating the alternative hypothesis.[8]

These findings should be replicated in independent laboratories, and the limits of this phenomenon should be identified. For example, research should assess the effect of the number of random attributes available to the algorithms, of significant digits used for measures (index of measurement precision), and of class category levels in the application, with regard to training and validity AUC. Research should also study designs with randomized *categorical* attributes having differing numbers of levels.

Finally, the present findings also bolster our recommendation to use the ODA and CTA frameworks to draw causal inferences regarding treatment effects in observational data, and in data from randomized controlled trials.[22-41] A large and rapidly-increasing mass of evidence supports the use of ODA and CTA to assess the efficacy of health-improvement interventions and policy initiatives.[42,43]

References

[1]Linden A, Yarnold PR (2019). Some machine learning algorithms find relationships between variables when none exist -- CTA doesn't. *Optimal Data Analysis, 8*, 64-67.

[2]Linden A, Yarnold PR (2019). Effect of sample size on discovery of relationships in random data by classification algorithms. *Optimal Data Analysis*, *8*, 76-80.

[3]Linden A, Yarnold PR (2019). Multi-layer perceptron neural net model identifies effect in random data. *Optimal Data Analysis*, *8*, 94-96.

[4]StataCorp (2017). *Stata 15 Multivariate Statistics Reference Manual*. College Station, TX: Stata Press.

[5]Linden A (2006). Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *Journal of Evaluation in Clinical Practice*, *12*, 132-139.

[6]Yarnold PR, Soltysik RC (2005). *Optimal Data Analysis: Guidebook with Software for Windows*. Washington, D.C.: APA Books.

[7]Yarnold PR (2017). What is optimal data analysis? *Optimal Data Analysis*, *6*, 26-42.

[8]Yarnold PR, Soltysik RC (2016). *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[9]Yarnold PR (2016). Using UniODA to determine the ESS of a CTA model in LOO analysis. *Optimal Data Analysis*, *5*, 3-10.

[10]Yarnold PR (2016). Determining jackknife ESS for a CTA model with chaotic instability. *Optimal Data Analysis*, *5*, 11-14.

[11]StataCorp (2017). *Stata 15 Structural Equation Modeling Reference Manual*. College Station, TX: Stata Press.

[12]Finney SJ, DiStefano C (2006). Nonnormal and categorical data in structural equation models. In GR Hancock & RO Mueller *(Eds.), A second course in structural equation modeling* (pp. 269-314). Greenwich, CT: Information Age.

[13]Jöreskog KG, Sörbom D (2002). *PRELIS 2 user's reference guide.* Scientific Software International: Lincolnwood, IL.

[14]Jöreskog KG, Sörbom D (2002). *LISREL 8 user's reference guide.* Scientific Software International: Lincolnwood, IL.

[15]Cohen J (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

[16]Brockway JH, Carlson KA, Jones SK, Bryant, FB (2002). Development and validation of a scale for measuring cynical attitudes toward college. *Journal of Educational Psychology, 94,* 210-224.

[17]Streitweiser B, Bryant FB, Drane D, Light, G (2019). Assessing student conceptions of international experience: Developing a validated survey instrument. *International Journal of Intercultural Relations, 68,* 26-43.

[18]Travers L, Randall E, Bryant FB, Conley C, Bohnert A (2015). The cost of perfection with apparent ease: Theoretical foundations and development of the Effortless Perfectionism Scale. *Psychological Assessment, 27,* 1147-1159.

[19]Bryant FB, Smith BD (2001). Refining the architecture of aggression: A measurement model for the Buss-Perry Aggression Questionnaire. *Journal of Research in Personality, 35,* 138-167.

[20]Bryant FB, Yarnold PR (1995). Comparing five alternative factor-models of the Student Jenkins Activity Survey: Separating the wheat from the chaff. *Journal of Personality Assessment, 64,* 145-158.

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

2155-0182/10/$3.00

[22]Yarnold PR (1996). Discriminating geriatric and non-geriatric patients using functional status information: An example of classification tree analysis via UniODA. *Educational and Psychological Measurement*, *56*, 656-667.

[23]Linden A, Yarnold PR (2016). Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice*, *22*, 839-847.

[24]Linden A, Yarnold PR (2016). Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*, *22*, 848-854.

[25]Linden A, Yarnold PR (2016). Using machine learning to identify structural breaks in single-group interrupted time series designs. *Journal of Evaluation in Clinical Practice*, *22*, 855-859.

[26]Linden A, Yarnold PR, Nallomothu BK (2016). Using machine learning to model dose-response relationships. *Journal of Evaluation in Clinical Practice*, *22*, 860-867.

[27]Yarnold PR, Linden A (2016). Using machine learning to model dose-response relationships via ODA: Eliminating response variable baseline variation by ipsative standardization. *Optimal Data Analysis*, *5*, 41-52.

[28]Linden A, Yarnold PR (2016). Combining machine learning and matching techniques to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*, *22*, 868-874.

[29]Linden A, Yarnold PR (2016). Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. *Journal of Evaluation in Clinical Practice*, *22*, 875-885.

[30]Yarnold PR, Linden A (2016). Novometric analysis with ordered class variables: The optimal alternative to linear regression analysis, *Optimal Data Analysis*, *5*, 65-73.

[31]Yarnold PR, Linden A (2016). Theoretical aspects of the D statistic. *Optimal Data Analysis*, *5*, 171-174.

[32]Yarnold PR, Linden A (2017). Computing propensity score weights for CTA models involving perfectly predicted endpoints. *Optimal Data Analysis*, *6*, 43-46.

[33]Linden A, Yarnold PR (2017). Minimizing imbalances on patient characteristics between treatment groups in randomized trials using classification tree analysis. *Journal of Evaluation in Clinical Practice*, *23*, 1309-1315.

[34]Linden A, Yarnold PR (2018). The Australian gun buy-back program and the rate of suicide by firearm. *Optimal Data Analysis*, *7*, 28-35.

[35]Linden A, Yarnold PR (2018). Identifying causal mechanisms in health care interventions using classification tree analysis. *Journal of Evaluation in Clinical Practice*, *24*, 353-361.

[36]Linden A, Yarnold PR (2018). Estimating causal effects for survival (time-to-event) outcomes by combining classification tree analysis and propensity score weighting. *Journal of Evaluation in Clinical Practice*, *24*, 380-387.

[37]Linden A, Yarnold PR (2018). Using machine learning to evaluate treatment effects in multiple-group interrupted time series analysis. *Journal of Evaluation in Clinical Practice*, *24*, 740-744.

[38]Linden A, Yarnold PR (2018). Using ODA in the evaluation of randomized controlled trials. *Optimal Data Analysis*, *7*, 46-49.

2155-0182/10/$3.00

[39]Linden A, Yarnold PR (2018). Using ODA in the evaluation of randomized controlled trials: Application to survival outcomes. *Optimal Data Analysis*, *7*, 50-53.

[40]Linden A, Yarnold PR (2017). Using classification tree analysis to generate propensity score weights. *Journal of Evaluation in Clinical Practice*, *23*, 703-712.

[41]Linden A, Yarnold PR (2017). Modeling time-to-event (survival) data using classification tree analysis. *Journal of Evaluation in Clinical Practice*, *23*, 1299-1308.40

[42]Linden A, Roberts N (2005). A Users guide to the disease management literature: recommendations for reporting and assessing program outcomes. *American Journal of Managed Care*, 11, 81-90.

[43]Linden A, Adler-Milstein J (2008). Medicare disease management in policy context. *Health Care Finance Review*, 29, 1-11.

Author Notes

No conflict of interest was reported.