Routledge
Taylor & Francis Group

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: http://www.tandfonline.com/loi/hmbr20

Abstract: Data Mining Alternatives to Logistic Regression for Propensity Score Estimation: Neural Networks and Support Vector Machines

Bryan S. B. Keller , Jee-Seon Kim & Peter M. Steiner

To cite this article: Bryan S. B. Keller , Jee-Seon Kim & Peter M. Steiner (2013) Abstract: Data Mining Alternatives to Logistic Regression for Propensity Score Estimation: Neural Networks and Support Vector Machines, Multivariate Behavioral Research, 48:1, 164-164, DOI: 10.1080/00273171.2013.752263

To link to this article: http://dx.doi.org/10.1080/00273171.2013.752263

Multivariate Behavioral Research, 48:164, 2013

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2013.752263

Routledge
Taylor & Francis Group

Abstract: Data Mining Alternatives to Logistic Regression for Propensity Score Estimation: Neural Networks and Support Vector Machines

Bryan S. B. Keller, Jee-Seon Kim, and Peter M. Steiner
Department of Educational Psychology, University of Wisconsin-Madison

Logistic regression (LR) has traditionally been the most frequently used method for modeling selection in propensity score (PS) analysis. The dominance of a single method is not for a lack of alternatives; rather, there is a perception among practitioners and methodologists that the extant research on alternatives to LR has not yet made a strong enough case for considering an alternative method.

There are, however, circumstances under which logistic regression may not perform well. If the response surface is not a hyperplane, the LR selection model requires polynomial and interaction terms. When there are many covariates, the number of terms to consider can be overwhelmingly large. In addition, when the ratio of the number of covariates to the sample size is high, the estimates produced by LR might be unstable. Data mining methods such as the neural networks (NN) and support vector machines (SVM) are designed to deal with high-dimensional data and algorithmically handle nonlinearities in the selection surface, thus avoiding the need for iterative model respecifications.

The performance of the data mining methods relative to LR is examined via simulation. Design factors in the study include the PS estimation method (LR with main effects only, NN with 8 hidden nodes, SVM with radial basis), the true selection model (linear and additive logistic vs. not), and the true outcome model (linear and additive vs. not). Performance outcomes include absolute bias; mean squared error; and standard error of the treatment effect estimate and covariate balance on linear, quadratic, and 2-way interaction terms.

Our results show that NN uniformly outperforms SVM on all outcomes under all scenarios. NN and LR perform similarly when the LR model is correctly specified. When the LR is misspecified, it is outperformed by NN. In the most extreme case, with quadratic and interaction terms in both the true PS and outcome models, the absolute bias averaged over 1,000 replications was 145% for LR and less than 6% for NN. The results underscore the importance of checking balance on higher order terms. We recommend estimating PSs via NN in conjunction with LR. If NN achieves better balance, either respecify the LR model or use the estimates from the NN.

Bryan S. B. Keller is grateful to his SMEP sponsor, David Kaplan. Correspondence concerning this abstract should be addressed to Bryan S. B. Keller, Department of Educational Psychology, University of Wisconsin-Madison, 859 Education Sciences, 1025 W. Johnson St., Madison, WI 53706. E-mail: bskeller@wisc.edu