Abstract Title Page
Not included in page count.

Title: Propensity Score Estimation with Data Mining Techniques: Alternatives to Logistic Regression

Authors and Affiliations:

Bryan S. B. Keller
University of Wisconsin-Madison

Jee-Seon Kim
University of Wisconsin-Madison

Peter M. Steiner
University of Wisconsin-Madison

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

Propensity score analysis (PSA) is a methodological technique which may correct for selection bias in a quasi-experiment by modeling the selection process using observed covariates. Because logistic regression is well-understood by researchers in a variety of fields and easy to implement in a number of popular software packages it has traditionally been the most frequently used method for modeling selection in PSA. The dependence on a single method is not for a lack of alternatives; any method that relates a binary outcome to multiple predictors is appropriate for modeling selection. Rather, there is a perception among practitioners and methodologists that the extant research on alternatives to logistic regression has not yet made a strong enough case for considering a different method (Stuart, 2010; Steiner & Cook, in press).

There are, however, circumstances under which logistic regression may not perform well. If the response surface is not a hyperplane, the logistic regression selection model will require more than just linear terms in order to capture nonlinear relationships. Although polynomial and interaction terms may be included in the logistic model in order to better approximate a nonlinear selection process, 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 logistic regression will be unstable. Data mining methods such as the neural network (NN; Ripley, 1996) and the support vector machine (SVM; Cortes and Vapnik, 1995) are potentially useful in such situations because they are designed to deal with high-dimensional data and they automatically detect and model nonlinearities in the selection surface, thus avoiding the need for iterative model respecification.

Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook (2008) compared the performance of neural networks and main-effects only logistic regression in a simulation study that included ten covariates. They found that neural networks outperformed logistic regression in terms of percent bias reduction in some scenarios, including those in which the selection model was most nonlinear and nonadditive. To our best knowledge, the performance of neural networks for the estimation of propensity scores has not been compared with logistic regression in any other empirical investigation. In a review of potential alternatives to logistic regression for PS estimation, Westreich, Lessler, & Funk (2010) noted that the SVM is promising because it is well-suited to classification problems with high-dimensional data and does not require specification of a parametric model. Ratkovic (2012) adapted the SVM classifier to carry out case matching directly to estimate the average treatment effect on the treated in a nonequivalent comparison group setting. However, the performance of the SVM has not been examined (either in a simulation or case study) in the context of propensity score estimation.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

Simulation Study 1: A simulation study explores the effect of PS estimation method on (a) the mean square error and standard error of the treatment effect estimates and (b) covariate balance after conditioning on the estimated propensity scores via optimal full matching. Design factors in the study include the PS estimation method (logistic regression, the neural network, and the SVM), the data-generating selection model (linear and additive vs. nonlinear and non-additive), the data-generating outcome model (linear and additive vs. nonlinear and non-additive), and the number of covariates. The purpose of the simulation study is to examine the performance of the data mining methods relative to logistic regression under the different data generation scenarios and for varying numbers of covariates.

Simulation Study 2: Both the NN and the SVM require the specification of tuning parameters in order to be implemented. When models are used for prediction, the optimal tuning parameters should be selected by running an extensive grid search and selecting the parameters which minimize the cross-validated prediction error. In the context of using data mining techniques to estimate propensity scores, however, prediction is not the ultimate goal. McCaffrey, Ridgeway, & Morral (2004) used generalized boosted modeling for PS estimation and recommended maximizing the balance instead of minimizing the prediction error. We develop a cross-validation procedure in R which maximizes the balance as measured by the average absolute standardized mean difference on first and second order terms and evaluate its performance relative to minimizing prediction error via simulation.

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.

Since Rosenbaum and Rubin (1983), logistic regression has been the traditional choice for PS estimation. The most important disadvantage of a propensity score estimation approach that uses logistic regression is the need for iterative specification of the model, which can be rather time intensive and comes with no guarantee of success, in particular with many covariates. A careful review of the burgeoning PS estimation literature has shown that the neural network and the SVM are promising alternatives to logistic regression which avoid the need for respecification because they automatically model nonlinearities in the selection response surface, and are well suited for high-dimensional data. These two methods, although promising, are heretofore largely or completely empirically untested in this context.

Through simulation, we examine the conditions under which logistic regression is relatively robust to model misspecification and the conditions under which the neural network or the support vector machine will provide a less biased estimate of the effect of a treatment. We also evaluate through simulation and make available a program written in R which carries out a cross-validated grid search for the optimal tuning parameters for the data mining methods based on maximizing the balance as opposed to minimizing the prediction error.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

Here we describe the data generation models for simulation study 1 with 12 covariates of size *n* = 2000. Let $\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_{12}})$ represent the 2000×12 matrix of covariates. The $\boldsymbol{X_i}$ are independently generated from a standard normal distribution and correlations are introduced by Cholesky decomposition such that $\rho(\boldsymbol{X_i}\boldsymbol{X_j}) = 0.3$ for all $i \neq j$. Let $\boldsymbol{Z}$ represent the dichotomous 2000×1 treatment assignment vector and $\boldsymbol{Y}$ the 2000×1 continuous outcome vector. Then $e(\boldsymbol{X}) = P(\boldsymbol{Z} = \boldsymbol{1}|\boldsymbol{X})$ is the propensity score vector. The data-generating ***propensity score*** model are as follows.

Scenario A_{PS} – propensity score model is linear and additive (all main effects only):

$$e(\boldsymbol{X}) = (1 + \exp\{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{12} X_{12})\})\text{^} - 1$$

Scenario B_{PS} – propensity score model is nonlinear and nonadditive (all main effects, five two-way interactions, and four quadratic terms):

$$e(\boldsymbol{X}) = (1 + \exp\left\{-\left(\begin{array}{c}\beta_0 + \beta_1 \boldsymbol{X_1} + \beta_2 \boldsymbol{X_2} + \cdots + \beta_{12} \boldsymbol{X_{12}} + \\ \beta_{13} \boldsymbol{X_1} \boldsymbol{X_{12}} + \beta_{14} \boldsymbol{X_2} \boldsymbol{X_{12}} + \beta_{15} \boldsymbol{X_2} \boldsymbol{X_{10}} + \beta_{16} \boldsymbol{X_4} \boldsymbol{X_{12}} + \beta_{17} \boldsymbol{X_1} \boldsymbol{X_8} + \\ \beta_{18} \boldsymbol{X_2^2} + \beta_{19} \boldsymbol{X_5^2} + \beta_{20} \boldsymbol{X_8^2} + \beta_{21} \boldsymbol{X_{11}^2} + \epsilon\end{array}\right)\right\})\text{^} - 1$$

The data-generating ***outcome*** model are as follows.

Scenario A_{OC} – outcome model is linear and additive (all main effects only):

$$\boldsymbol{Y} = \alpha_0 + \alpha_1 \boldsymbol{X_1} + \alpha_2 \boldsymbol{X_2} + \cdots + \alpha_{12} \boldsymbol{X_{12}} + \epsilon$$

Scenario B_{OC} – outcome model is nonlinear and nonadditive (all main effects, five two-way interactions, and four quadratic terms):

$$\boldsymbol{Y} = \left(\begin{array}{c}\alpha_0 + \alpha_1 \boldsymbol{X_1} + \alpha_2 \boldsymbol{X_2} + \cdots + \alpha_{12} \boldsymbol{X_{12}} + \\ \alpha_{13} \boldsymbol{X_1} \boldsymbol{X_{12}} + \alpha_{14} \boldsymbol{X_2} \boldsymbol{X_{12}} + \alpha_{15} \boldsymbol{X_2} \boldsymbol{X_{10}} + \alpha_{16} \boldsymbol{X_4} \boldsymbol{X_{12}} + \alpha_{17} \boldsymbol{X_1} \boldsymbol{X_8} + \\ \alpha_{18} \boldsymbol{X_2^2} + \alpha_{19} \boldsymbol{X_5^2} + \alpha_{20} \boldsymbol{X_8^2} + \alpha \boldsymbol{X_{11}^2} + \epsilon\end{array}\right)$$

Here we give a brief overview of the data mining methods used to estimate propensity scores in the simulation studies. The term *neural network* refers to a class of models inspired by theories about how the human brain uses neurons to send messages. The back-propagation neural network is made up of an *input layer*, which consists of the observed covariates, an *output layer*, which consists of one unit for a dichotomous classification problem, and one *hidden layer* of unobserved variables. The NN may be thought of as a nonlinear extension of logistic regression. In particular, the NN with no hidden layer and one dichotomous output is equivalent to logistic regression. The addition of a hidden layer involves a weighted transformation of the data via an activation function, usually chosen to be the logistic function, $f(x) = (1 + \exp(-x))^{-1}$; the hidden layer with non-linear activation function is what affords the NN its added flexibility. Weight matrices $\boldsymbol{Z}$ and $\boldsymbol{W}$ contain the connection weights between the covariates and the hidden layer and the hidden layer and the output, respectively, and are iteratively estimated through forward and backward passes through the network until a stopping criteria is met.

The support vector *classifier* is a technique for classifying training data by constructing a maximally separating hyperplane in the covariate space. If the data are quite noisy, it may not be possible to construct a hyperplane that perfectly separates the points. This problem is dealt with

by defining slack variables which are proportional to how far from the boundary the point is on the wrong side. The margin is maximized subject to the constraint that the sum of the slack variables be less than or equal to a constant. Thus, points near the boundary play a bigger role in shaping it and points which are correctly classified play no role in shaping the boundary. The support vector *machine* is a nonlinear extension of the support vector classifier which maps the covariates into a high dimensional space via a basis function (often the radial basis function, $K(X_1, X_2) = \exp\left(-\gamma\left||X_1 - X_2|\right|^2\right)$). The support vector classifier is then run in the transformed space. Linear boundaries in the transformed space translate to non-linear boundaries in the original space.

Usefulness / Applicability of Method:
Demonstration of the usefulness of the proposed methods using hypothetical or real data.

The results of the simulation study clearly demonstrate that the misspecification of the PS model via logistic regression leads to the potential for gross bias in the estimate of the treatment effect when there are nonlinear or nonadditive confounders. This can be seen in Table 1 in the BB condition. The absolute percent bias for the misspecified logistic regression model for that condition is 145%, compared with 6% for the NN and 28% for the SVM. An examination of the average standardized absolute mean difference reveals that although logistic regression was able to attain better balance on the linear confounders, it failed to balance the higher order terms. The NN and the SVM are fully automated algorithmic approaches which were able to achieve better overall balance which resulted in substantially better estimates of the treatment effect in this condition as measured by both the absolute bias and the mean-square error.

Conclusions:
Description of conclusions, recommendations, and limitations based on findings.

Results of the simulation study demonstrate that when there are higher order confounders (i.e., nonlinear terms present in both the selection and the outcome model), misspecification of the logistic PS estimation model can result in a very biased estimate of the treatment effect. The data mining techniques were less biased and had smaller mean square error in that case. The simulation study further explores the effect of the number of covariates and the number and strength of higher order confounders on the performance of the PS estimation methods. We develop and assess the use of a cross-validation procedure to choose tuning parameters based on maximizing balance as opposed to minimizing prediction error. Finally, we use the results of the simulations to inform the reanalysis of several educational data sets which used (a) PSA with PSs estimated by logistic regression or (b) a multiple regression approach. In particular, we check for robustness of the treatment effect estimate across the different PS estimation methods and use balance (or lack thereof) on higher order terms as an indicator of the appropriateness of PS estimation method.

Given the widespread use of propensity score methods in education and across a variety of other substantive areas, the potential impact of improved estimates of treatment effect based on appropriate selection of PS estimation techniques is enormous. It is our hope that the recommendations based on the simulation study results will help to guide researchers to make informed decisions about which propensity score estimation technique to use for their given situation in order to maximize the accuracy and efficiency of research.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning, 20,* 273-297.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York, New York: Springer.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9,* 403-425.

Ratkovic, M. (2012). *Achieving optimal covariate balance under general treatment regimes.* Unpublished manuscript. Retrieved from http://www.princeton.edu/politics/about/file-repository/public/RATKOVIC-REVISED-MatchingRD3.pdf

Ripley, B. D. (1996). *Pattern recognition and neural networks*. New York, NY: Cambridge University Press.

Steiner, P. M. & Cook, T. D. (in press). Matching and propensity scores. In T. Little (Ed.), *Oxford handbood of quantitative methods*. Oxford: Oxford University Press.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*, 1-21.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety, 17*, 546-555.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: Neural networks, SVMs, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology, 63*, 826-833.

Appendix B. Tables and Figures
Not included in page count.

Table 1. Simulation results averaged over 1000 replications for the case with $p = 12$ covariates and $N = 2000$ subjects. The average treatment effect was estimated by optimal full matching in each case.

		Scenario†			
Metric	Method*	$A_{PS}A_{OC}$	$A_{PS}B_{OC}$	$B_{PS}A_{OC}$	$B_{PS}B_{OC}$
Absolute bias (per cent)	LR	0.11	1.46	12.83	145.30
	NN	2.43	3.81	3.07	5.57
	SVM	6.61	7.44	6.82	28.16
Mean-square error × 10^2	LR	0.63	1.26	0.63	34.57
	NN	0.77	1.27	0.75	1.16
	SVM	0.89	1.52	0.98	2.53
Standard error × 10^2	LR	0.25	0.35	0.19	0.28
	NN	0.28	0.35	0.27	0.33
	SVM	0.29	0.38	0.30	0.36
ASAMD‡: main effects	LR	0.040	0.039	0.030	0.030
	NN	0.047	0.046	0.048	0.047
	SVM	0.050	0.060	0.062	0.061
ASAMD: squared terms	LR	0.047	0.047	0.125	0.125
	NN	0.049	0.050	0.045	0.046
	SVM	0.047	0.052	0.059	0.060
ASAMD: two-way interactions	LR	0.065	0.066	0.135	0.134
	NN	0.062	0.063	0.060	0.060
	SVM	0.064	0.074	0.071	0.071

*LR: logistic regression, NN: neural network, SVM: support vector machine.
† A_{PS} = data generating PS model is linear and additive; B_{PS} = data generating PS model is nonlinear and nonadditive; A_{OC} = data-generating outcome model is linear and additive; B_{OC} = data-generating outcome model is nonlinear and nonadditive;
‡ASAMD: average standardized absolute mean difference of the covariates after PS matching.