Abstract Submitted to Spring 2016 Conference of the Society for Research on Educational Effectiveness (SREE)

Abstract Title Page

Not included in page count.

Title:

Empirically Driven Variable Selection for the Estimation of Causal Effects with Observational Data

Authors and Affiliations:

Bryan Keller, Teachers College, Columbia University
Jianshen Chen, Educational Testing Service

Abstract Body

Background / Context:

Observational studies are common in educational research, where subjects self-select or are otherwise non-randomly assigned to different interventions (e.g., educational programs, grade retention, special education). Unbiased estimation of a causal effect with observational data depends crucially on the assumption of ignorability (Rosenbaum and Rubin, 1983), which specifies that potential outcomes under different treatment conditions are independent of treatment assignment, given the observed covariates. If an important confounder is not observed, and no suitable proxies are included in the method used to condition on the covariates, then bias may remain in the estimate of the causal effect even after adjustment for the observed covariates. In an effort to protect against this kind of omitted variable bias, statisticians have favored an inclusive approach to covariate selection for causal inference, advocating for inclusion so long as covariates were measured before any treatment was administered (Rubin, 2009; Gelman, Carlin, Stern, & Rubin, 2003, p. 228; Rosenbaum, 2002, p. 76; D'Agostino, 1998).

There are, however, three classes of variables which, when conditioned upon, can have an undesirable effect on the bias and/or variance of causal effect estimators: non-informative variables, instrumental variables, and collider variables. Non-informative variables (NVs) are those associated with neither the treatment Z nor the outcome Y; see Figure 1. Instrumental variables (IVs) are those associated with treatment Z and not with the error term of Y (Angrist, Imbens, & Rubin, 1996); see Figure 2. A variable C is called a collider if it lies on a directed path in which it is causally predicted by two adjacent variables A and B (Pearl, 2009a, p. 17); see Figure 3.

It is well known that conditioning on NVs decreases precision of estimation and, thereby, increases mean squared error; see Kuhn & Johnson (2013; p. 489) for simulation results with parametric and nonparametric regression methods. Decreased precision due to NVs extends directly to the estimation of causal effects, wherein conditioning on NVs in an attempt to satisfy ignorability only makes estimators less precise. As for IVs, conditioning on an IV will invariably increase the standard error of the causal effect; this has been clearly demonstrated in simulation studies (Myers et al, 2011; Austin, Grootendorst, & Anderson, 2007; Brookhart et al, 2006). Furthermore, conditioning on an IV, or a variable that is nearly an IV, because it is very weakly related to the outcome, has the potential to amplify bias due to unmeasured confounders (Steiner & Kim, 2014; Wooldridge, 2009). With respect to collider variables, if a variable C is a collider along the causal pathway from A to B, conditioning on C will induce a correlation between A and B, even if they are marginally independent. Thus, if interest centers on detecting the effect of treatment Z on an outcome Y, conditioning on a collider C can increase bias, because A and B are associated, given C. This type of bias due to conditioning on a collider is called M-Bias because of the shape of the causal diagram that describes it; see Figure 3.

There has been much debate centered around whether the strategy of adjusting for all pretreatment covariates is misguided given the potential bias-inducing impact of controlling for a collider (Pearl, 2009b; Rubin, 2009). However, true M structure is likely very rare in practice (Ding and Miratrix, 2015) and, as noted by Elwert & Winship (2014), the only way to decide whether to control for a potential collider C is to make an a priori decision based on theory about how the data were generated. Similarly, IVs and NVs are not identifiable in the presence of unobserved confounding because estimators for regression coefficients of observed covariates are inconsistent. However, IVs and NVs (and their near- counterparts) arguably are more likely to occur in practice, especially when covariates are chosen from a large database. The focus of

this paper is on the use of data-driven variable selection techniques to identify and drop potential non-informative and instrumental variables in the presence of unobserved heterogeneity in order to study the impact on the bias and efficiency of causal estimation.

Van der Weele and Shipster (2011) proposed a criterion for confounder selection that depends on the assumption that there exists a subset of the observed covariates, which, if conditioned upon, is sufficient to control for confounding. De Luna, Waernbaum, & Richardson (2011) proposed several covariate selection algorithms designed to identify minimal subsets of covariates sufficient for controlling for confounding, again, assuming a sufficient subset is actually observed. In practice, outside of tightly controlled laboratory settings, such as those created by Shadish, Clark, & Steiner (2008), the existence of a sufficient set of observed covariates for the complete control of confounding is often dubious in observational research. Thus, a practical approach for the identification and removal of variable types known to degrade the performance of causal effect estimators is called for. Although we focus our attention on empirical methods for identifying and removing potential non-informative and instrumental variables, we support the advice given by Sauer, Brookhart, Roy, and VanderWeele (2013) that an approach that combines prior knowledge about the relationships between variables with an empirical variable selection procedure is a practical compromise.

Purpose / Objective / Research Question / Focus of Study:

Our primary goals in this paper are to (a) propose and evaluate an empirically driven method for the identification and removal of potential instrumental and non-informative pretreatment variables based on lack of association with the outcome (though bias will exist in the presence of unmeasured confounders), and (b) to investigate, through simulation studies, the efficacy of three variable selection methods as measured by their success in identifying IVs and NVs and by improvement in bias and mean squared error relative to no variable selection at all.

Significance / Novelty of study:

The methods we use for variable selection are forward stepwise regression based on the AIC (Venables & Ripley, 2002), the lasso (Tibshirani, 1996), and recursive feature elimination with random forests (Kuhn & Johnson, 2013). Although stepwise regression has traditionally been the most commonly recommended method for empirically-based variable selection in causal inference (e.g., Brookhart et al, 2006), the lasso and machine learning methods such as random forests are being used more frequently (Sauer et al, 2014). Increased use notwithstanding, to our best knowledge, Ertefaie, Asgharian, & Stephens (2013) is the only published example including a simulation study in which the primary goal is to evaluate empirically driven variable selection methods with observational data. Our simulations differ from Ertefaie et al. (2013) in that (a) we use a very different data-generating process, (b) they focus on doubly robust estimation of the treatment effect, whereas we do not apply an outcome model after propensity score adjustment, and (c) we evaluate different methods for variable selection (although the use of the lasso is a point of overlap).

Research Design:

Two simulation studies are conducted. Data for the first simulation are generated with one continuous outcome (Y), one binary treatment (Z) and 24 continuous covariates including three IVs (IV_1 to IV_3), three near-IVs (IV_4 to IV_6), three NVs (NV_1 to NV_3), three near-NVs (NV_4 to NV_6), and 12 confounders (X_1 to X_{12}), two of which are chosen to be unobserved (X_1

and X_5); see Figure 4 for a graphical representation of the data-generating process. The data-generating procedure for the first simulation study was carried out as follows:
(1) Generate N = 1000 rows on 24 predictors (IV_1, …, IV_6, NV_1, …, NV_6, X_1, … X_{12}) from a standard multivariate normal density with off-diagonal pairwise correlations set to ρij = 0.3. Let X = (IV_1, …, IV_6, NV_1, …, NV_6, X_1, … X_{12}) be the resultant 1000 by 24 matrix of predictor variables.
(2) Obtain the true propensity scores by

$$PS\mathbf{X} = \exp(0 + \mathbf{X}\boldsymbol{\alpha}) \; 1 + \exp(0 + \mathbf{X}\boldsymbol{\alpha}) \;,$$

where $\boldsymbol{\alpha}$ represents a column vector of 24 coefficients.

(3)Generate the treatment variable Z by comparing the propensity score vector, $PS(\mathbf{X})$, to a vector of random draws Ui from Uniform(0, 1) and assign Zi = 1 if $PS\mathbf{X}\boldsymbol{i}$ > Ui and assign Zi = 0 otherwise.
(4) Generate the outcome Y by

$$\mathbf{Y} = 0 + \boldsymbol{\beta}\mathbf{X} + \tau\mathbf{Z} + \boldsymbol{\epsilon} \;,$$

where $\boldsymbol{\beta}$ is a column vector of 24 coefficients, τ=0 is the true treatment effect, and $\boldsymbol{\epsilon i} \sim N(0, 9)$. The error variance of 9 was chosen to yield a signal-to-noise ratio (Dicker, 2012) of approximately 6.65; this value is within the range of values studied by Tibshirani (1996), for evaluating the efficacy of the lasso for feature selection. For specific values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, see Tables 1 to 3.
(5) Data = ($\mathbf{X}$, $\mathbf{Z}$, $\mathbf{Y}$). Replicate 1000 times to produce 1000 data sets for simulation analysis.

For each replication of the simulation, three variable selection methods are run. For the lasso, the tuning parameter was selected via 10-fold cross-validation. Then the subset of variables retained is used in a logistic regression with one linear term for each variable in order to estimate propensity scores. Finally, the overall average treatment effect (ATE) is estimated via inverse probability of treatment weighting.

Note that in the first simulation study both the propensity score and outcome models include only linear (main effect) terms for each of the predictors. For the second simulation study, we add a quadratic term for X2 and use a smaller coefficient for the linear term for X_2 in order to investigate the performance of the variable selection algorithms in the presence of nonlinearity. Thus, steps (1), (3), and (5) and the post-subset selection analyses are identical. The procedure for steps (2) and (4) in the second simulation study are altered as follows:
The PS model coefficient for X_2, $\alpha 14$, is reduced from log(1.5) to log(1.05) and the propensity score model is augmented to include a quadratic term for X_2 with coefficient $\alpha 25$= log(1.2).
The outcome model coefficient for X_2, $\beta 14$, is reduced from 0.8 to 0.03 and the outcome model is augmented to include a quadratic term for X_2 with $\beta 25$= 0.8.

Findings / Results:

Simulation Study 1. Figure 5 shows the number of times each variable was dropped by each method across the 1000 replications of simulation study 1. All three methods dropped the first 12 variables (the IVs, near-IVs, NVs, and near-NVs), far more frequently than the confounders, though there were differences between the methods. For example, the lasso identified and dropped about 45% of true IVs, forward stepwise selection dropped about 68% of the true IVs, and recursive feature elimination with random forests dropped about 72% of the

true IVs. The frequency with which the weak confounders (X_4, X_8, and X_{12}; those with standardized regression coefficients of 0.30) were dropped varied as well. The lasso dropped about 1% of the weak confounders, stepwise dropped about 4% of the weak confounders, and recursive feature elimination with random forests dropped about 22% of the weak confounders.

Table 4 displays the bias, mean squared error (MSE), and simulation standard deviation for the average treatment effect for the different methods used for variable selection in Simulation Study 1. With no predictors included, the MSE of the prima facie estimator was estimated to be 36.43. On the other extreme, the MSE based on conditioning on all 22 observed predictors was 1.85. As expected, all three of the variable selection methods yielded lower MSEs; stepwise = 1.20, recursive feature elimination with random forests (RFE-RF) = 1.28, lasso = 1.42. While the MSEs for stepwise and RFE-RF were close, lasso trailed due to the decreased sensitivity for detection of IVs and NVs, displayed in Figure 5.

Simulation Study 2. The motivation for the second simulation study was to see how the three methods would handle a moderately sized nonlinear term in the outcome model. To that end, we decreased the magnitude of the coefficient for the linear term for X_2 and added a quadratic term with a moderately sized coefficient. Our hypothesis was that the two models based on linear regression (i.e., the lasso and forward stepwise selection) would drop X_2 more frequently because of the lack of a strong linear presence, but that RFE-RF would include X_2 because random forests algorithmically handle nonlinearities as they are based on regression trees, which can deal with a quadratic relationship, for example, by splitting twice on the same variable.

Figure 6 displays the number of times each variable was dropped by each method across the 1000 replications of simulation study 2. The general trends across methods are identical to simulation study 1 with one exception: the X_2 variable. X_2 was dropped from 62% and 39% of replications based on stepwise and lasso regression, whereas it was dropped from only 4% of replications based on RFE-RF. Table 5 presents the bias, mean squared error (MSE), and simulation standard deviation for the average treatment effect for the different methods used for variable selection in Simulation Study 2. Similar to Simulation Study 1, all three of the variable selection methods yielded lower MSEs than the MSEs based on no predictors or all predictors. However, in contrast with study 1, RFE-RF produced the smallest MSE (1.52), the lasso second (1.85) and the stepwise selection the third (1.90).

Conclusions:

In practice, when faced with a real data set, it is very unlikely that the IV/NV status of any variable will be known based on only prior substantive knowledge. Thus, empirically-based variable selection procedures have a role to play in identifying them, especially in cases where many variables with questionable relationship with the outcome are under consideration.

The conclusions of our simulations are twofold. First, we show that for the two data-generation processes used herein, preprocessing data to detect and remove potential instrumental and non-informative variables based on their relationships with the outcome improved the mean squared error of treatment effect estimation. Of course, this has to do with the bias/variance trade off: more sensitive variable selection methods do a better job detecting and removing non-informative variables (which decreases variance) while simultaneously dropping more weak confounders (which increases bias). Second, we find that recursive feature elimination with random forests is a promising method for predictor selection, as evidenced by strong performance in its naïve implementation across both simulation studies. Finally, we note that the ability of the methods to single out NVs and IVs depends in part upon the magnitudes and

directions of correlations between unmeasured confounders and other predictors. We are investigating these relationships in our ongoing work.

Appendices

Not included in page count.

Appendix A.

References

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association, 91,* 444-455.

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine, 26*, 734-753.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*, 1149–1156.

Ertefaie, A., Asgharian, M. & Stephens, D. A. (2014). Variable Selection in Causal Inference Using Penalization. *http://arxiv.org/pdf/1311.1283.pdf*

D'Agostino, R. B., Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*, 2265–2281.

De Luna, X., Waernbaum, I. & Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika, 98*, 861-875.

Dicker, L. H. (2012). Residual variance and the signal-to-noise ratio in high-dimensional linear models, arXiv preprint, arXiv:1209.0012.

Ding, P., & Miratrix, L. W. (2015). To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias. *Journal of Causal Inference, 3,* 41-57.

Elwert, F., & Winship, C. (2014). Endogenous selection bias: the problem of conditioning on a collider variable. *Annual Review of Sociology, 40*, 31-53.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian Data Analysis (Second Edition)*. Chapman & Hall/CRC Texts in Statistical Science.

Kuhn, M. & Johnson, K. (2013). *Applied Predictive Modeling*. New York: Springer.

Myers, J. A., Rassen, J. A., Gagne, J. J., et al. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology, 174*, 1213-1222.

Pearl, J. (2009a). *Causality: Models, Reasoning and Inference (2nd Edition)*. Cambridge University Press.

Pearl, J. (2009b). Letter to the editor: Remarks on the method of propensity score. *Statistics in Medicine, 28*, 1415–1416.

Rosenbaum, P. R. (2002). *Observational studies (2nd Edition)*. New York: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials, *Statistics in Medicine, 26*, 20–36.

Rubin, D. B. (2009). Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine, 28*, 1420–1423.

Sauer, B. C., Brookhart, M. A., Roy, J., & VanderWeele, T. (2013). A review of covariate selection for non-experimental comparative effectiveness research. *Pharmacoepidemiology and Drug Safety, 22*, 1139–1145.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 103*, 1334-1344.

Steiner, P. M., & Kim, Y. (2014). *On the Bias-Amplifying Effect of Near Instruments in Observational Studies*. Paper presented at the Spring Conference of the Society for Research on Educational Effectiveness (SREE), Washington D.C., 2014.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society (Series B), 58*, 267-288.

VanderWeele, T. J., & Shiptser, I. (2011). A new criterion for confounder selection. *Biometrics, 67*, 1406–1413.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S (4th Edition).* New York: Springer.

Walter, S. & Tiemeier, H. (2009). Variable selection: current practice in epidemiological studies. *European Journal of Epidemiology, 24*, 733-736.

Wooldridge, J. (2009). Should instrumental variables be used as matching variables? East Lansing, MI: Michigan State University.

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

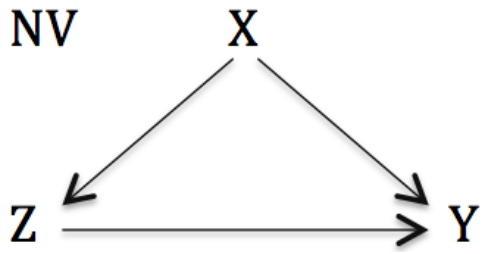

Figure 1. Causal diagram showing a confounding variable, X, and a non-informative variable, NV, for the treatment/outcome pair {Z, Y}.

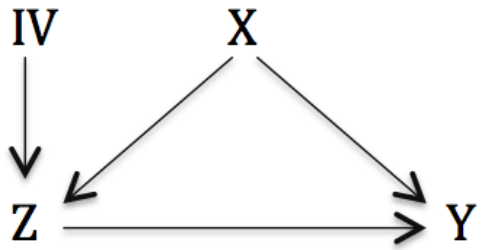

Figure 2. Causal diagram showing a confounding variable, X, and an instrumental variable, IV, for the treatment/outcome pair {Z, Y}.

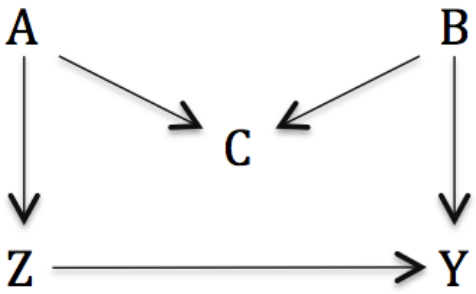

Figure 3. Causal diagram showing a collider variable, C, for the treatment/outcome pair {Z, Y}. Conditioning on C alone will result in *M*-Bias.

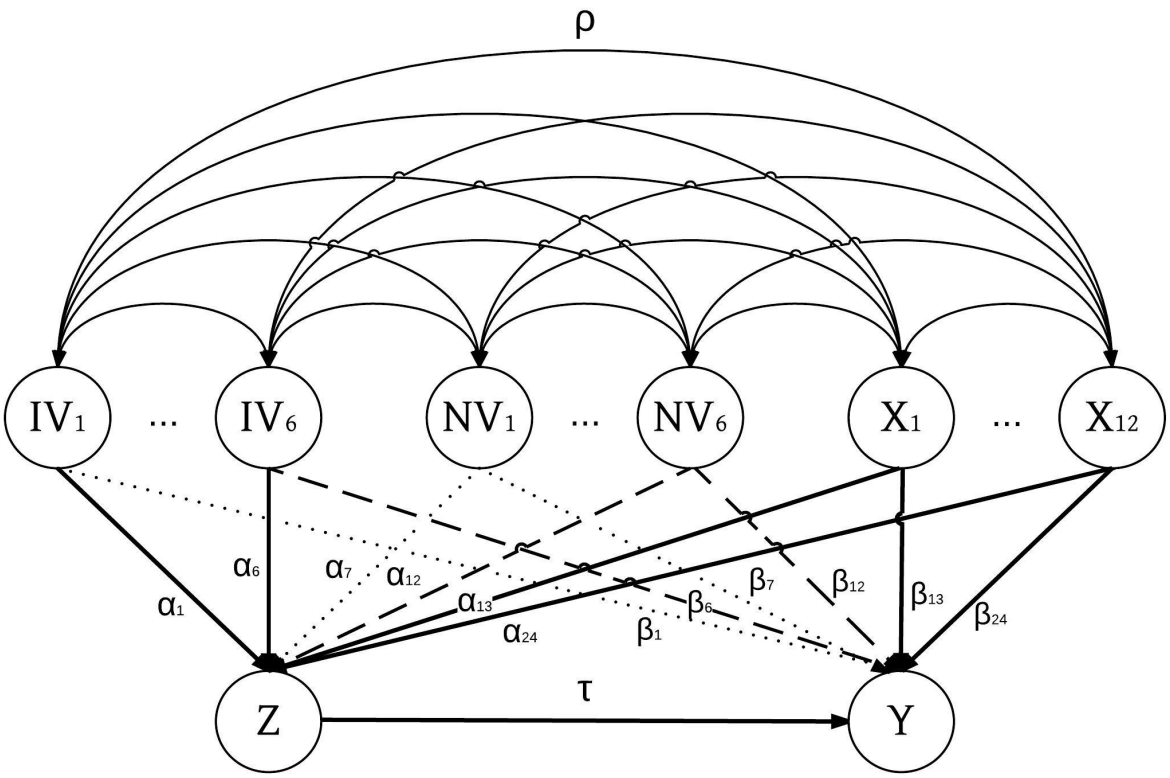

Figure 4. Causal diagram of the data-generating process. Solid arrows represent a causal effect; dotted arrows represent the absence of a causal effect; dashed arrows represent the *near* absence of a causal effect.

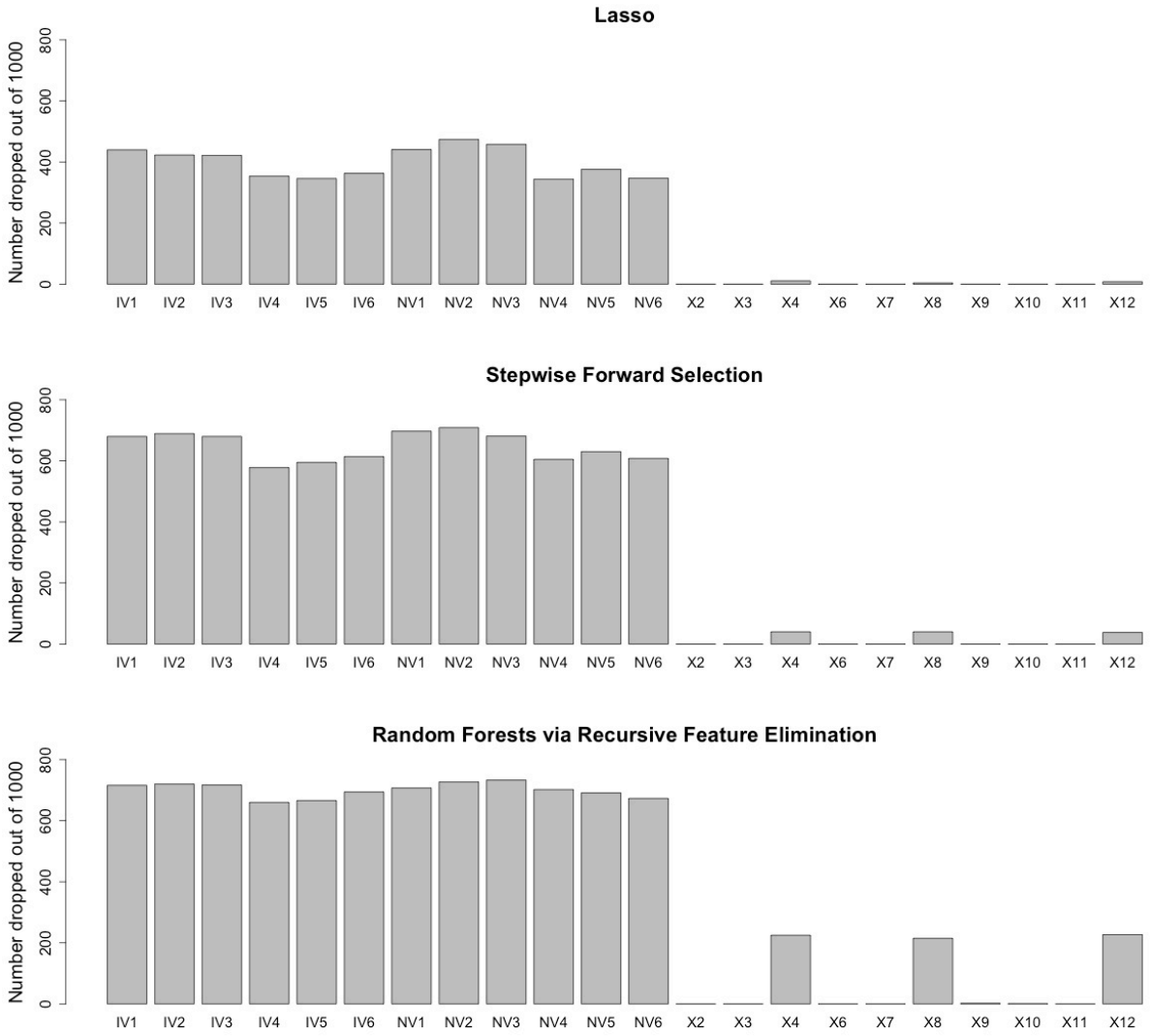

Figure 5. Simulation Study 1: Frequency plots of the number of times each variable was dropped out of 1000 simulation replications.

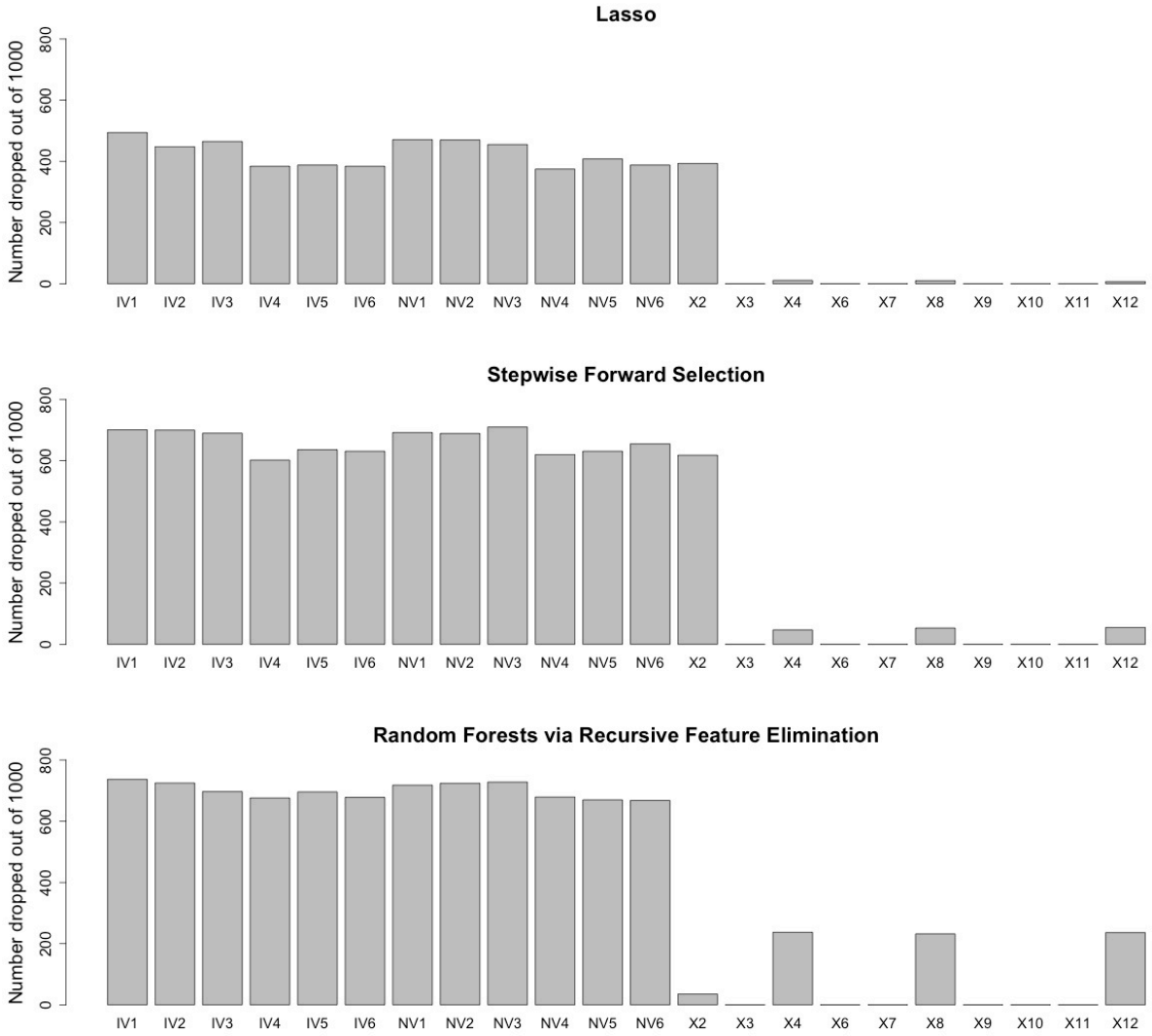

Figure 6. Simulation Study 2 (included a quadratic term for X_2): Frequency plots of the number of times each variable was dropped out of 1000 simulation replications. Note the differences with respect to X_2.

i	1	2	3	4	5	6
Name	IV_1	IV_2	IV_3	IV_4	IV_5	IV_6
α_i	log(1.20)	log(1.50)	log(2.00)	log(1.20)	log(1.50)	log(2.00)
β_i	0.00	0.00	0.00	0.03	0.03	0.03

Table 1. Data-generation coefficients for IVs and near-IVs.

i	7	8	9	10	11	12
Name	NV_1	NV_2	NV_3	NV_4	NV_5	NV_6
α_i	log(1.00)	log(1.00)	log(1.00)	log(1.05)	log(1.05)	log(1.05)
β_i	0.00	0.00	0.00	0.03	0.03	0.03

Table 2. Data-generation coefficients for NVs and near-NVs.

i	13	14	15	16	17	18	19	20	21	22	23	24
Name	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{10}	X_{12}
α_i	log(1.2)	log(1.5)	log(2)	log(1.2)	log(1.5)	log(2)	-log(1.2)	-log(1.5)	-log(2)	-log(1.2)	-log(1.5)	-log(2)
β_i	0.30	0.80	2.00	0.30	2.00	0.80	2.00	0.30	0.80	0.80	2.00	0.30

Table 3. Data-generation coefficients for confounding variables.

Method	Performance Measures		
	Bias	MSE	Simulation SD
No predictors	6.02	36.43	0.49
All observed	0.82	1.85	1.09
Stepwise	0.83	1.20	0.72
RF via RFE	0.87	1.28	0.72
Lasso	0.79	1.42	0.89

Table 4. Results from simulation study 1.

Method	Performance Measures		
	Bias	MSE	Simulation SD
No predictors	5.99	36.09	0.49
All observed	0.87	2.27	1.23
Stepwise	1.19	1.90	0.70
RF via RFE	1.03	1.52	0.68
Lasso	1.07	1.85	0.84

Table 5. Results from simulation study 2.