The Effect of Varibale Selection Approaches on MSE of IPW ATE Estimator

Prelude

This document presents the final project completed collaboratively with fellow students as part of PUBH 7485: Methods in Causal Inference at the University of Minnesota School of Public Health in the Fall of 2023, a requirement in my MS Biostatistics Curriculum.

All relevant work files, including the `.qmd` file for document generation, are accessible in my GitHub repository. The repository contains : - code for making simulated data set - code for performing a causal inference analysis on one draw of said data set - code for storing all data that we need for the study of simulation results - scripts to interface with the computing cluster - final report reproducible document

This study is our first collective experience in research of statistical methods. We all were intrigued by the idea of finding a 'go-to' method for selecting variables to conduct a causal inference analysis in our careers beyond graduate school. We concluded that outcome adaptive LASSO variable selection is the best way to select variable for regression adjustment in the context of causal inference.

This document serves as a practical example of the final reports I can produce in my role as a data scientist or statistician. For more samples and a comprehensive view of my work, please explore my portfolio, showcasing various reports, studies, dashboards, and other analytical files.

Introduction

The average treatment effect (ATE) serves as a crucial measure for evaluating the causal impact of a specific treatment or intervention on an outcome variable. However, randomized experiments are typically necessary to establish a control group closely resembling the intervention group, ensuring accurate ATE estimation. Despite the widespread availability of data

today and the relatively lower costs compared to randomized trials, there is a growing interest in leveraging observational (or non-randomized) studies for estimating treatment effects, especially in social sciences, epidemiology, and certain clinical studies.

Inverse probability weighting, a propensity score-based technique, proves valuable for addressing imbalance in study groups within observational studies. Achieving an unbiased estimator, under the assumption of "no unmeasured confoundings," becomes a challenge in constructing a propensity score model. Real-world observational studies often contains substantial sample sizes and a high dimension of potential covariates, exemplified by studies such as Terzic et al. (2021) and Butler et al. (2023), which involve nearly 5,000 samples with hundreds of variables. To enhance the relevance of the study to real-world research, we will initiate the variation of our sample size, starting from 1500 with 50 covariates.

In the context of high-dimensional datasets, variable selection using machine-learning approaches has become an intriguing topic. In articles such as Tang et al. (2023) and Lu et al. (2018), the authors delve into variable selection for causal inference under ultra-high dimensionality, employing random forest approaches to build the model. This project aims to evaluate the performance of each variable selection method across diverse simulated scenarios. Taking advantage of knowing the truth, we can compare variable selection methods, contributing to a better understanding of these techniques in real-world applications.

Methods

Simulation Design

Covariates

We will simulate potential m covariates $\mathbf{X} = (X_1, ..., X_m)$ from a multivariate normal distribution $N(\mu, \Sigma)$, where $\mu = 0$. The correlation matrix Σ will be generated using the *rcorrmatrix* function from the *clusterGeneration* package. To enhance computational efficiency, the covariates will be simulated in units of 50 columns each. In other words, when considering a scenario with 150 covariates, we will first generate three independent subsets $\mathbf{Z}_1, \mathbf{Z}_2$, and $\mathbf{Z}_3$, where each subset $\mathbf{Z}$ will consist of 50 correlated covariates. Subsequently, we will construct $\mathbf{X}$ as $(\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3)$.

Treatment Assignment and Outcome Models

Selecting True Covariates

To determine the covariates that will be included in the model, we will first establish the number of true covariates denoted as s that we wish to incorporate. A vector $\mathbf{V} = (V_1, ..., V_m)$ will be generated, where we randomly select s columns to serve as the underlying predictors

for the model. Each element in the vector will function as an indicator, determining whether the column from $\mathbf{X}$ will be utilized in the model or not.

Generating Coefficients for True Covariates

For the coefficients of the treatment assignment model, we will generate another vector $\mathbf{U} = (U_1, ..., U_m)$ where $U_i \sim$ Unif(-0.5, 0.5). We then define the vector of coefficients $\beta_A = (V_1U_1, ..., V_mU_m)^T$. The final treatment assignment A is determined by a Bernoulii[expit($\mathbf{X}\beta_A$)].

For the outcome model, we will also generate a vector $\mathbf{R} = (R_1, ..., R_m)$ where $R_i \sim$ Unif(-1, 1). Again, we then define a vector of coefficients $\beta_Y = (V_1R_1, ..., V_mR_m)^T$. Potential outcomes Y^0 and Y^1 and the observed outcome Y will be:

Potential outcomes:

$$Y^0 = \mathbf{X}\beta_Y$$

$$Y^1 = \alpha + \mathbf{X}\beta_Y$$

Obsevered outcome:

$$Y = A \times Y^1 + (1 - A) \times Y^0 + \varepsilon$$

where α is the average treatment effect and $\varepsilon \sim N(0, \delta)$

The value of α will be determined by solving $\frac{\alpha}{SD(Y^0)} = 0.5$. δ will be chosen in a way that $R^2 \approx 0.5$ when fitting linear regression with our outcome on true predictors. Across 100 iterations in each simulating scenario, the same model will be used and only covariates $\mathbf{X} = (X_1, ..., X_m)$ will be regenerate.

Simulation Schemes

1. Given m and s, generate an indicator vector $\mathbf{V} = (V_1, ..., V_m)$ deciding ture covariates.
2. Simulate coefficient vector $\mathbf{U} = (U_1, ..., U_m)$ and $\mathbf{R} = (R_1, ..., R_m)$ to get β_Y.
   (i) Given n and s, simulate covariate matrix $\mathbf{X}$.
   (ii) Derive potential outcome $Y^0 = \mathbf{X}\beta_Y$.
   (iii) Repeat (i) and (ii) to get 100 replications.
3. Solve ATE (α) and residual (ε) based on the formulas above with all Y^0 from all the 100 replications.
4. Derive $Y^1 = \alpha + \mathbf{X}\beta_Y$ and $Y = A \times Y^1 + (1 - A) \times Y^0 + \varepsilon$

Factors and Simulation Scenarios

1. Sample Size (n): To evaluate the impact of sample size on the bias and MSE of $\hat{ATE}$, considering that observational studies often involve very large sample sizes, we will vary

the sample size. Specifically, we will investigate sizes of 1500, 3000, 4500, and 6000, assessing the effect on MSE across different sample sizes.

2. Number of Potential Covariates (m): A crucial aspect of our investigation involves comparing the performance of variable selection methods with the manual selection of covariates based on expertise and experience, simulating real-world research settings. Our goal is to determine if any selection method consistently outperforms others. Given that observational studies often deal with extensive datasets, we will explore three scenarios with varying numbers of covariates: 50, 100, and 150. Despite simulating 50 covariates in a unit for computational efficiency, this approach aligns with the complexities of real-world situations where covariates are inherently intricate, featuring a mix of correlated and independent variables.

3. True Covariates (s): We will vary the number of true covariates from 10, 20, to 30. The objective is to examine the impact on MSE as the true model becomes more complex and to assess each variable selection method's efficacy in correctly identifying relevant covariates.

In total, the study contains 36 distinct simulation scenarios, each replicated 100 times. These scenarios include variations in sample size, the number of potential covariates, and the number of true covariates. This comprehensive approach allows us to evaluate the performance of variable selection methods under diverse conditions, providing insights into their robustness and effectiveness in practical research settings.

Variable Selection Methods

The objective of this project is to assess the impact on MSE of the IPW ATE estimator when employing different variable selection approaches in modeling the propensity score. The selected methods for modeling the propensity score include Forward Selection, Lasso, Adaptive Lasso, and Experience-based selection. To establish benchmarks for both the best-case and worst-case scenarios, the Oracle method (constructing the propensity score model based on the true covariates) and the t-test for estimating ATE will also be incorporated.

Subset Selection - Forward Stepwise Selection

One advantage of forward selection is that it starts with smaller models. Also, this procedure is less susceptible to collinearity, as discussed by Chowdhury and Turin (2020).
1. Let M_0 denote the null model, which contains no predictors.
2. For $k = 0, ..., p-1$:
   (a) consider all $p-k$ models that add just one new variable to M_k
   (b) choose the best (smallest deviance) among these $p-k$ models and call it M_{k+1}
3. Select a single best model from $M_0, M_1, ..., M_p$ using AIC.

Shrinkage - Lasso

LASSO regression, recognized as L1 regularization, is a popular technique used in statistical modeling and machine learning for variable selection and modeling outcome. The LASSO proceeds by adding a penalty term to the coefficients and minimizing a regularized version of least squares:

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=i}^{m}\beta_j X_{ij})^2 + \lambda \sum_{j=1}^{m}|\beta_j|$$

where $\lambda > 0$ is a tuning parameter that will be separately determined which will minimizing 10-fold CV MSE.

Shrinkage - Adaptvie Lasso

In the article Zou (2006), Hui Zou demonstrates that the Lasso sometimes exhibits inconsistent variable selection, including noise variables. He illustrates that incorporating weights on the penalty term for each variable, known as the adaptive Lasso, can yield a more stable model compared to the standard Lasso method.
The adaptive Lasso estimates β by minimizing

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=i}^{m}\beta_j X_{ij})^2 + \lambda_n \sum_{j=1}^{m}\frac{1}{w_j}|\beta_j|$$

where $w_j = |\hat{\beta}_{OLS}|$ and λ_n is also determined by minimizing 10-fold CV MSE.

Experience-based - Selecting 5 Correct and 5 Incorrect Covariates

In practice, some researchers often select confounders relying on their experience, opting for variables that are more interpretable. However, this approach does not guarantee the inclusion of all true confounders. To assess the MSE and potential bias associated with the selection of incorrect covariates, we will randomly choose 5 true covariates and 5 incorrect covariates for use in the propensity model across all simulation scenarios.

T-test - Naive ATE estiamtor

In practice, some preliminary analyses will take the difference in average of obsrbed outcomes between the two groups without adjusting for confounding. It is a known established fact that this leads to a biased treatment effect estimation. We can also consider this method a 'null' variable selection, i.e. a method that performs no vairbale selection, and therefore does not perform any adjustment for existing confounders.

Together with the 'Oracle' and 'Experience Based' methods, these three form a set of benchmarking methods. We will compare the performance of data driven methods with benchmarking methods in this study.

Table 1: Variable Selection Methods Considered for Analysis.

Methods	Short Description
Oracle	Model with perfect information, only all true covariates.
No Variable Selection (T-test ATE Estimation)	Testing covariates on treatment outcome to see which have a significan association
Experience Based	Selecting 5 correct and 5 randomly chosen predictors from the set of unrelated covariates
Adaptive Lasso	A regularization method of LASSO by avoiding overfitting with penalizing large coefficients
Lasso	Adds a penalty term to the coefficients and minimizing a penalized version of least squares where lambda > 0
Forward Selection	Beginning with a null model, adding covariates that have a significant association treatment outcome one at a time.

Results

In this section, we analyze the outcomes of simulated data replications through a comprehensive examination utilizing data summaries and regression methodologies. We focus on IPW estimator that employs a propensity score model and weighted sample mean differences. Initially, we employ graphical tools to assess the results and compare variable selection techniques with more naive methods for estimating Average Treatment Effects (ATE). Subsequently, our investigation delves into the evaluation of Mean Squared Error (MSE) across varying simulation parameters and diverse variable selection methods. Our objective is to quantify the marginal effects of different conditions on MSE, identifying a variable selection approach that consistently achieves the lowest MSE. We adopt a Gaussian General Linear Regression Model with an identity link function, incorporating a natural logarithm transformation of Squared Errors to derive main and interaction effects in terms of percentage changes.

Simulation Results

Figure 1 illustrates and Mean Squared Error (MSE) for the various variable selection methods across varying model complexity scenarios. With growing number of true predictors,even simple linear additive models become complex. We expect that the task of constructing a regression model in such cases is a difficult task, which leads to potential of hugher bias, higher variance, or a steep trade off between the two. We introduced the two-sample t-test estimator as a benchmark for unadjusted Average Treatment Effect (ATE) estimation. It is acknowledged that neglecting confounders can lead to biased estimation, a fact visually confirmed in Figure 1.

Similarly, the 'Experience-Based' variable selection method yields estimates with high MSE, which is likely due to bias. Recall that the 'Experience-Based' method consistently selects five true and five random confounders, which leads to violation of 'No Unmeasured Confounders' assumption.

Conversely, the MSE associated with data-driven variable selection algorithms is relatively small. We verified that on average the average range of bias associated with data-driven methods is acceptable. Therefore, MSE is mostly comprised of variance of the estimator. This observation holds across different data generation scenarios.

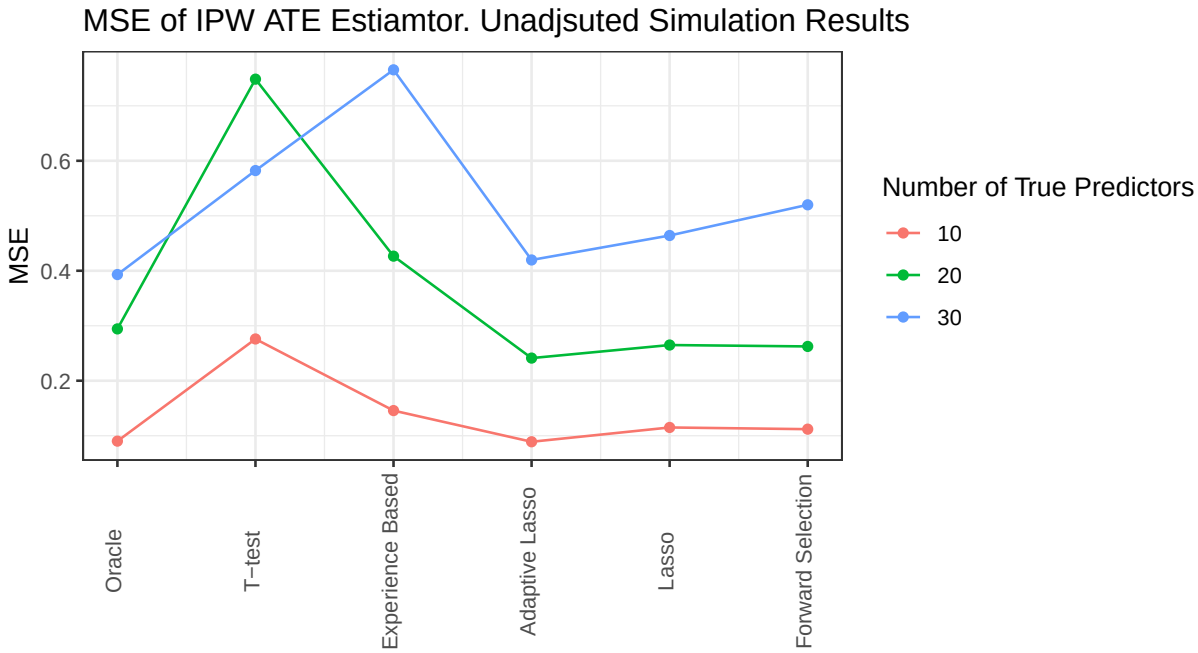

Figure 1: Mean Squared Error (MSE) directly from simulation results. Presented averages unadjusted for other potential factors explaining variations in MSE

Figure 1 shows that all data-driven methods exhibit a noticeable increase in MSE as the number of true confounders that need to be accounted for rises. We speculate that the rise in MSE and variance is a consequence of model misspecification.

Our speculation stems from the understanding that, under random sampling variability, it becomes increasingly challenging to select all true confounders as their number increases. Notably, for the Inverse Probability Weighting (IPW) class of Average Treatment Effect (ATE) estimators, model misspecification leads to higher variation in the estimator. The impact of this phenomenon is illustrated in the right graph in Figure 1.

In unadjusted comparisons, outcome adaptive lasso regression demonstrates behavior that, on average, closely approximates the performance of the 'Oracle' method in situations with a limited number of true covariates. Furthermore, both lasso and forward variable selection

methods exhibit nearly identical performance, and their effectiveness diminishes as the number of available variables for selection increases. This pattern persists even as the number of true confounders grows. However, as the complexity of the scenarios increases, all methods gradually converge to more similar results in terms of the attained Mean Squared Error (MSE).

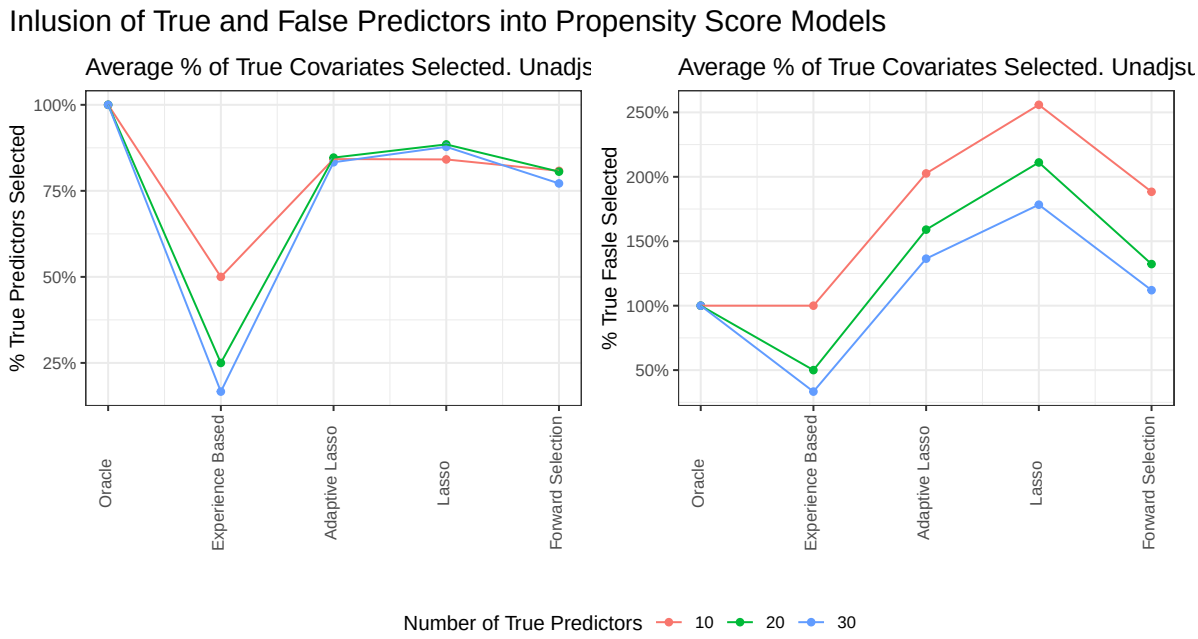

Figure 2: % True and False predictors calculated based of the number of true predictors. Values over 100% imply that the number of false predicotrs selected exceeds the number of true covariates

To assess the ability of each variable selection method to accurately specify the model, we examine the average proportion of true confounders selected and the average proportion of all predictors identified by the method in relation to the total number of true confounders. Figure 2 illustrates that, on average, all data-driven variable selection methods select approximately 75% to 85% of true confounders under varying data generation and sample size conditions. This higher percentage of correctly identified confounders is likely the primary factor contributing to low bias performance across all data generating scenarios. Notably, a steady deviation from this overall pattern is observed, indicating a consistent decrease in the number of true confounders as the number of options increases.

Conversely, Figure 2 also demonstrates that all data driven methods on average pick up big number of false positive covariates, likely due to sampling randomness. For example, when the number of true predictors is equal to 10, Lasso tends to pick up about 25 false positive predictors (i.e. 25 false positive / 10 true predictors = 250%). Additionally, on average, as the size of the true confounder space expands, the ratio of all selected covariates to the number of true confounders decreases for every method. Presumably, when the number of true

confounders is small, each additional irrelevant predictor added to the model has a more substantial impact on this proportion. The likelihood of selecting a false positive predictor in the model-building process is expected to increase when the number of true confounders is small, driven by sampling variability.

Main Effects

We employ a Gaussian General Linear Model with an identity link and a natural logarithm transformation of each squared error to derive marginal effects of the factors described in the previous section. Table 2 presents the effects of simulation parameters, variable selection results, and variable selection methods as the percent change in MSE.

Main effects of variable selection methods have limited significance in the context of our regression study. We have uncovered strong and suggestive evidence that the impact of variable selection methods on the percentage change in MSE varies with the number of true confounders. Therefore, we present the main effects of variable selection methods for scenarios where the number of true confounders is equal to zero, which is inherently meaningless. Instead, we focus on the effect of variable selection methods on MSE when compared with the best-case scenario (Oracle method) under varying numbers of true confounders. Detailed results are presented in Figure 3. Additionally, we provide a more nuanced exploration of variable selection method performance under different conditions in Figure 4 and Figure 5. These figures offer deeper insights into the effects based on varying simulation parameters and different numbers of true confounders.

We assessed the marginal effect of an increase in the number of true covariates, considering the size of the true confounder space as a factor measuring the difficulty of the model-building and variable selection task. Our estimation indicates that, with each additional true confounder, the average Mean Squared Error (MSE) is expected to increase by 4.6% (95% CI: -0.32% to 9.73%), after adjusting for other variables. While this result shows a strong suggestive trend ($P = 0.067$), it falls short of significance at the considered level of $\alpha = 0.05$. Additionally, we acknowledge a potential limitation; as the number of true predictors extends beyond 30, these findings may not be generalizable. A more in-depth discussion on these nuances is provided in the subsequent discussion section.

Figure 3 shows the effect of varaible selection method on MSE, and contrasted with the Oracle method. It is expected that the Oracle method will correctly specity the model at all times and therefore variance of an IPW esimator will only depend on the sampling variablity. As expected, this estimator has the lowest MSE through lowest bias and variance because model misspecificiation does not occur.

Table 2 provides expected percentage increase in MSE for each varaible selection mehtod under different data generating schemes, after adjsting for other variables, while Figure 3 compares and contrast expected MSE under our regression model for every variable selection method and true confounder space size.

Table 2: Gaussian GLM with log-transformed resposnse effect estiamtes. Coefficients are exponentiated and present the effect as % change

Predictor	Estimate	95% CI	P-value	P<0.05
Variable Selection Methods				
Adaptive Lasso	-1.3%	(-29.44 % , 38.09 %)	0.9395	
Lasso	72.3%	(19.7 % , 148.02 %)	0.0034	*
Forward Variable Selection	52.6%	(7.36 % , 116.96 %)	0.0185	*
Experience Based Selection	8.6%	(-39.87 % , 96.04 %)	0.7850	
Other Main Effects				
Total Covaraites Available	-1%	(-1.07 % , -0.89 %)	0.0000	*
Sample Size	-2.1%	(-4.21 % , -0.01 %)	0.0485	*
Number of True Confounders	4.6%	(-0.32 % , 9.73 %)	0.0672	
% True Confounders Selected	-34.9%	(-71.91 % , 51.03 %)	0.3177	
% Total Covariates Selected	-5.1%	(-13.59 % , 4.28 %)	0.2772	
Intercation Terms				
Number of True Confounders * Adaptive Lasso	0%	(-1.53 % , 1.5 %)	0.9697	
Number of True Confounders * Lasso	-1.8%	(-3.27 % , -0.35 %)	0.0152	*
Number of True Confounders * Forward Variable Selection	-1.2%	(-2.85 % , 0.5 %)	0.1655	
Number of True Confounders * Experience Based Selection	3.4%	(-0.43 % , 7.42 %)	0.0822	
Number of True Confounders * % True Confounders Selected	3.3%	(-1.44 % , 8.31 %)	0.1751	

[a] Regression model explains 10.26% of variation in Squared Errors of IPW estimator

[a] Variable selection methods are compared with the refernce 'Oracle' level

Overall, Figure 3 shows that in situations with a low number of true predictors, Oracle-based and Outcome Adaptive Lasso produce IPW estimators with similar MSE, which is also the lowest possible under our data-generating specifications. Lasso regression and Forward Variable Selection methods yield IPW estimators with higher MSE, as evidenced by mostly non-overlapping confidence intervals. In a situation with a low number of true covariates, the 'Experience-Based' method of variable selection results in the highest amount of MSE, although not significantly different from Lasso and Forward Selection approaches.

As we increase the number of true covariates and make the task of model building via variable selection more challenging, we observe that all data-driven methods produce estimators that converge to the same value of MSE, while the 'Experience-Based' approach tends to perform worse as the number of true covariates grows. While this finding was not expected, a possible explanation could be considered in the context of the bias-variance trade-off, which is discussed in detail in the subsequent discussion section.

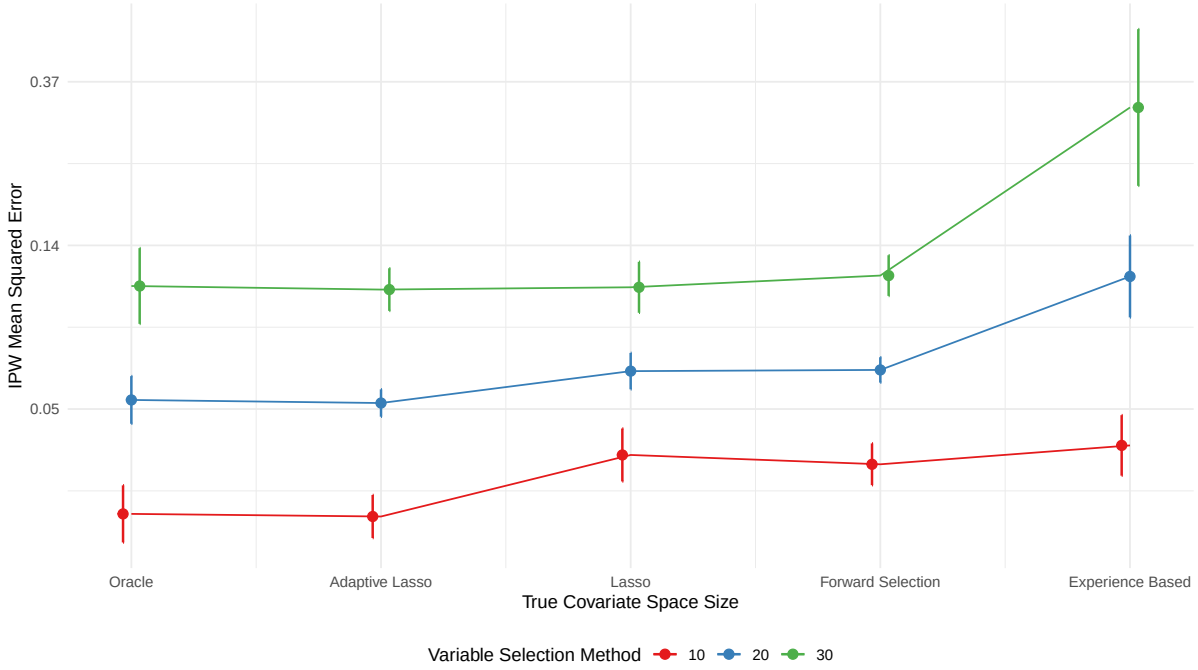

Figure 3: Marginal effect of variable selection method on MSE

The last phase of our analysis assesses the influence of the proportion of correctly identified confounders and false-positive covariates on Mean Squared Error (MSE) for each employed method. The consideration of statistically significant interactions prompts a detailed presentation of results for varying sizes of the true covariate space. Figure 4 incorporates key insights from Figure 3. In scenarios where the number of true confounders to capture is low, all methods yield estimators with comparable MSE. Furthermore, with an increase in the capture of true covariates, a marginal reduction in MSE is observed, though not statistically significant. However, as the size of the true confounder space expands, capturing an additional

percentage of covariates results in higher MSE. We posit that the complexity of data generating mechanisms necessitates the specification of intricate models, inevitably introducing bias, which contributes to increased MSE. Detailed discussions on these findings are provided in the subsequent section.

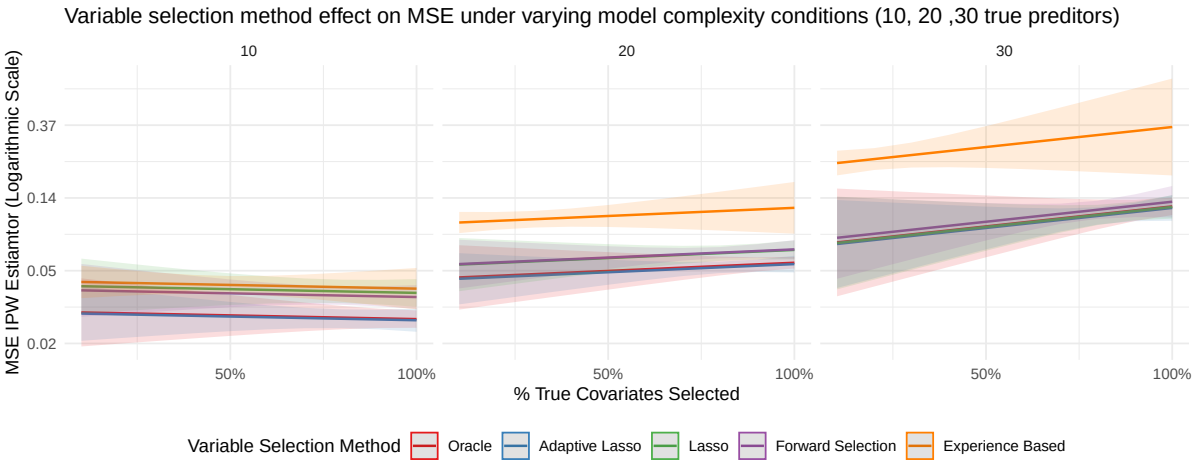

Figure 4: Effects of variable selection models are given for scenarios with 10, 20, and 30 true confounders

Figure 3 also includes broad confidence bands for estimated effect lines, particularly noteworthy when the true number of confounders is 30. An examination of the data summary reveals few instances where the true number of predictors was below 50%. The preceding Figure 2 indicates that, on average, 75% to 85% of true covariates were captured. Estimating and interpreting effects in scenarios with sparse data presentation lead to inherently uncertain results. Please note that some estiamted effects might not have actual reflection in the data. Oravle variable selction method always identifies 100% of corerct covariates. We use our regression model to estiamte expected MSE for the Oracle variable selection method at values lesser than 100% to enable comparison of this benchmark method with other variable seelction methods.

Furthermore, we explore the impact of the proportion of all selected covariates on the number of true confounders. Detailed results are presented in Figure 5 in the appendix.

Discussion

Our comprehensive investigation into the variable selection method for the Inverse Probability Weighting (IPW) propensity score model yielded both anticipated and unforeseen results. The primary objective of this study was to identify a method that yields a model with minimal degrees of misspecification. Our hypothesis posited that the least degree of misspecification would result in reduced variance and more precise estimators. The results section, however,

demonstrates that the answer to this question is contingent upon underlying parameters, including sample size and the complexity of the true model.

In our exploration, we revealed that, for a category of simpler additive logistic regression models, the adaptive lasso regression model consistently produces estimators with the lowest Mean Squared Error (MSE) across various data-generating and data-collection scenarios. It is crucial to note, however, that the true model is inherently unknown to researchers and analysts.

Consequently, based on the insights derived from this study, we advocate for the adoption of an adaptive lasso regression model for the IPW estimator by practicing statisticians. While the comparison with more flexible modeling techniques remains an open question, particularly in situations where interpretable models are preferred, the adaptive lasso consistently outperforms other methods across all unidentified data-generating scenarios considered in our study.

In instances where the underlying model is complex (with 30 true predictors), all data-driven methods exhibit similar results, and the choice of the variable selection model has minimal impact. However, when the number of true predictors is limited, and the inclusion of each additional false positive incurs a high cost, the selection of the variable selection method becomes pivotal.

It is imperative to reiterate that the true number of predictors in the underlying data-generating model remains unknown. Therefore, our strong recommendation is to employ an adaptive lasso model for variable selection, ensuring optimal performance of the IPW estimator in terms of minimizing MSE when compared to Lasso and Forward selection models.

Our investigation also delved into the influence of both true and false predictors on Mean Squared Error (MSE), visualizing these effects through graphical tools in Figure 4 and Figure 5, considering interactions and complex relationships. Surprisingly, as the percentage of true covariates selected by each method increased, we observed a concurrent increase in MSE. Our speculation is that augmenting model complexity leads to higher variance in the estimator associated with this model.

In contrast, Figure 5 in the appendix demonstrates that the marginal impact of including false positive covariates results in no discernible change in the MSE of the estimator. These findings may offer reassurance to practicing statisticians, indicating that the 'penalty' for the inclusion of false positive predictors appears to be absent. The emphasis lies solely on the presence of true positive predictors in the model.

Given that MSE encompasses both the variance and bias of the estimator, it is plausible that a model with a high number of predictors, achieved by maintaining the percentage of false positive predictors at around 150%, may exhibit low average bias due to the incorporation of numerous potential confounders. However, such a model is likely to possess a high degree of variance in the estimates it produces.

While our current analysis does not empower us to conduct a causal analysis, future studies could leverage these preliminary results to perform a more in-depth examination. Subsequent simulation studies may experiment with varying proportions of true positive and false positive predictors while keeping other data-generating parameters constant. This could help evaluate whether the rate of change in MSE worsens compared to the results we presented.

It appears that the inevitable increase in MSE as the number of predictors grows can be mitigated, and our best approach may be to minimize the rate at which MSE increases with the inclusion of more predictors. Such a finding could empower practicing analysts, providing them with a deeper understanding of how their modeling choices impact subsequent inferences.

Conclusion

In this study, we assessed the impact of variable selection methods on the construction of Inverse Probability Weighting (IPW) Average Treatment Effect Estimators within the propensity score model. Our evaluation encompassed diverse data-generating scenarios designed to emulate real-world datasets commonly used by researchers and practitioners. The focal points of interest included the number of true predictors in the data-generating mechanism and the variable selection method.

On average, our results revealed that all data-driven methods tend to capture approximately 75% of true confounders while incorporating around 150% of false positive confounders. Utilizing these findings, data-generating conditions, and variable selection methods, we constructed a Gaussian General Linear Model to investigate the impact of these factors on Squared Errors on the logarithmic scale. Our analysis indicated that MSE increases as models become more complex, prompting us to recommend further studies to evaluate whether the rate of MSE growth can be controlled using different methods or if it is an inevitable effect that can only be minimized.

A key discovery from our study was that the Adaptive Lasso regression model, which selects variables based on shrinkage parameters, outperforms all other data-driven variable selection methods across both known and unknown data-generating schemes. However, in scenarios with numerous confounders, all data-driven methods tend to exhibit similar performance, converging to the best-case scenario on average. Our advocacy rests on the suggestion that, in situations where interpretable regression models are preferred, practitioners should opt for Adaptive Lasso regression for robust model building.

References

Butler, J Lauren, Penny Gordon-Larsen, Lyn M Steffen, James M Shikany, David R Jr Jacobs, Barry M Popkin, and Jennifer M Poti. 2023. "Associations of 5-Year Changes in Alcoholic Beverage Intake with 5-Year Changes in Waist Circumference and BMI in the Coronary Artery Risk Development in Young Adults (CARDIA) Study." *PLoS One* 18 (3): e0281722. https://doi.org/10.1371/journal.pone.0281722.

Chowdhury, Mohammad Ziaul Islam, and Tanvir C Turin. 2020. "Variable Selection Strategies and Its Importance in Clinical Prediction Modelling." *Family Medicine and Community Health* 8 (1): e000262. https://doi.org/10.1136/fmch-2019-000262.

Lu, Min, Saad Sadiq, Daniel J Feaster, and Hemant Ishwaran. 2018. "Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods." *Journal of Computational and Graphical Statistics* 27 (1): 209–19. https://doi.org/10.1080/10618600.2017.1356325.

Tang, Dingke, Dehan Kong, Wenliang Pan, and Linbo Wang. 2023. "Ultra-high dimensional variable selection for doubly robust causal inference." *Biometrics* 79 (2): 903–14. https://doi.org/10.1111/biom.13625.

Terzic, Milan, Gulzhanat Aimagambetova, Sanja Terzic, Milena Radunovic, Gauri Bapayeva, and Antonio Simone Laganà. 2021. "Periodontal Pathogens and Preterm Birth: Current Knowledge and Further Interventions." *Pathogens* 10 (6): 730. https://doi.org/10.3390/pathogens10060730.

Zou, Hui. 2006. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101 (476): 1418–29. https://doi.org/10.1198/016214506000000735.

Appendix

Supplemental Figures

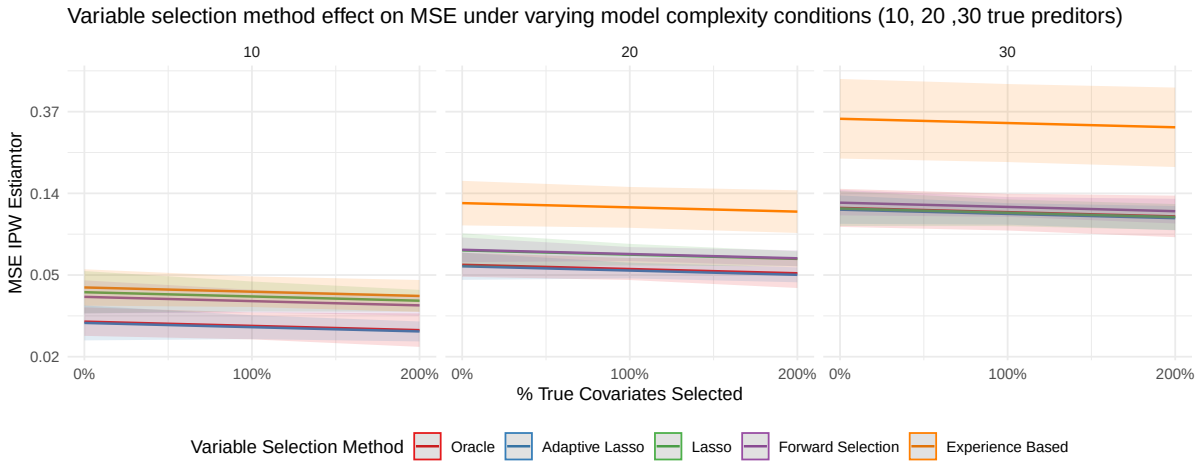

Figure 5: Effects of variable selection models are given for scenarios with 10, 20, and 30 true confounders

Table 3: Mean Squared Error of Adjusted Treatment Effect for Select Propensity Score Variable Selection Methods

	Propensity Score Stratification					Inverse Probability Weighting				
				95% CI					95% CI	
Method	Bias	MSE	SE	Lower	Upper	Bias	MSE	SE	Lower	Upper
Oracle	0.0037	0.0000841	0.116	0.974	1.430	-0.0078	0.0000921	0.131	0.933	1.447
Adaptive Lasso	0.0042	0.0000845	0.129	0.942	1.448	-0.0014	0.0000857	0.107	0.980	1.399
Lasso	0.0107	0.0001039	0.134	0.981	1.506	0.0019	0.0001062	0.125	0.990	1.479
Forward Selection	0.0130	0.0001044	0.135	0.981	1.510	0.0052	0.0001131	0.154	0.935	1.540
Experience Based	0.0724	0.0001516	0.115	1.038	1.489	0.0642	0.0001525	0.115	1.029	1.481
T-test	0.0711	0.0002287	0.130	1.049	1.558	0.0711	0.0002287	-	-	-

[a] Average value of 500 simulated combinations of each combination of sample size (1500, 3000, 4500, 6000), potential covariates (50, 100, 150), and true covariates (10, 20, 30)

Table 4: Mean Squared Error of Adjusted Treatment Effect for Select Propensity Score Variable Selection Methods

					95% CI					95% CI	
		Propensity Score Stratification					Inverse Probability Weighting				
Method	n	Bias	MSE	SE	Lower	Upper	Bias	MSE	SE	Lower	Upper
Oracle	1500	0.0104	0.0001684	0.174	0.863	1.544	0.0024	0.0001913	0.194	0.815	1.576
Oracle	3000	0.0098	0.0000734	0.111	0.942	1.377	-0.0139	0.0000776	0.123	0.895	1.377
Oracle	4500	0.0178	0.0000509	0.095	1.047	1.421	-0.0070	0.0000541	0.110	0.994	1.424
Oracle	6000	-0.0232	0.0000438	0.085	1.043	1.377	-0.0128	0.0000456	0.097	1.030	1.411
Adaptive Lasso	1500	0.0227	0.0001544	0.189	0.839	1.579	0.0277	0.0001563	0.153	0.915	1.514
Adaptive Lasso	3000	0.0004	0.0000957	0.135	0.953	1.483	-0.0073	0.0000971	0.112	0.990	1.430
Adaptive Lasso	4500	0.0034	0.0000398	0.101	0.959	1.353	-0.0111	0.0000417	0.082	0.981	1.303
Adaptive Lasso	6000	-0.0099	0.0000479	0.092	1.016	1.378	-0.0147	0.0000477	0.080	1.036	1.348
Lasso	1500	0.0431	0.0001930	0.201	0.872	1.659	-0.0063	0.0002021	0.174	0.876	1.557
Lasso	3000	-0.0009	0.0000950	0.128	0.955	1.455	0.0253	0.0000931	0.121	0.994	1.468
Lasso	4500	-0.0146	0.0000528	0.102	1.037	1.435	-0.0185	0.0000531	0.099	1.038	1.427
Lasso	6000	0.0177	0.0000602	0.091	1.100	1.457	0.0093	0.0000614	0.094	1.085	1.455
Forward Selection	1500	0.0319	0.0001948	0.200	0.862	1.647	0.0461	0.0002144	0.232	0.814	1.723
Forward Selection	3000	0.0108	0.0000953	0.130	0.963	1.471	-0.0249	0.0001013	0.146	0.894	1.468
Forward Selection	4500	-0.0093	0.0000527	0.103	1.040	1.443	-0.0086	0.0000561	0.117	1.013	1.472
Forward Selection	6000	0.0214	0.0000600	0.092	1.102	1.463	0.0096	0.0000642	0.106	1.063	1.478
Experience Based	1500	0.1115	0.0002844	0.162	0.980	1.616	0.1030	0.0002876	0.162	0.971	1.607
Experience Based	3000	0.0750	0.0001635	0.122	1.054	1.531	0.0667	0.0001626	0.122	1.045	1.523
Experience Based	4500	0.1517	0.0000624	0.090	1.128	1.481	0.1403	0.0000636	0.090	1.116	1.470
Experience Based	6000	-0.0485	0.0000963	0.086	0.990	1.327	-0.0530	0.0000964	0.086	0.984	1.323
T-test	1500	-0.0985	0.0005785	0.182	0.768	1.480	-0.0985	0.0005785	-	-	-
T-test	3000	0.2929	0.0001287	0.128	1.249	1.749	0.2929	0.0001287	-	-	-
T-test	4500	0.0038	0.0000815	0.105	1.049	1.459	0.0038	0.0000815	-	-	-
T-test	6000	0.0940	0.0000747	0.094	1.171	1.538	0.0940	0.0000747	-	-	-

[a] Average value of 500 simulated combinations of each combination of potential covariates (50, 100, 150), and true covariates (10, 20, 30)