

# The Effect of Variable 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 contain 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  $\mathbf{X} = (X_1, \dots, X_m)$  from a multivariate normal distribution  $N(\mu, \Sigma)$ , where  $\mu = 0$ . The correlation matrix  $\Sigma$  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  $\mathbf{V} = (V_1, \dots, V_m)$  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  $\mathbf{U} = (U_1, \dots, U_m)$  where  $U_i \sim \text{Unif}(-0.5, 0.5)$ . We then define the vector of coefficients  $\beta_A = (V_1 U_1, \dots, V_m U_m)^T$ . The final treatment assignment  $A$  is determined by a Bernoulli[expit( $\mathbf{X}\beta_A$ )].

For the outcome model, we will also generate a vector  $\mathbf{R} = (R_1, \dots, R_m)$  where  $R_i \sim \text{Unif}(-1, 1)$ . Again, we then define a vector of coefficients  $\beta_Y = (V_1 R_1, \dots, V_m R_m)^T$ . Potential outcomes  $Y^0$  and  $Y^1$  and the observed outcome  $Y$  will be:

Potential outcomes:

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

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

Observed outcome:

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

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

The value of  $\alpha$  will be determined by solving  $\frac{\alpha}{SD(Y^0)} = 0.5$ .  $\delta$  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  $\mathbf{X} = (X_1, \dots, X_m)$  will be regenerate.

### Simulation Schemes

1. Given  $m$  and  $s$ , generate an indicator vector  $\mathbf{V} = (V_1, \dots, V_m)$  deciding true covariates.
2. Simulate coefficient vector  $\mathbf{U} = (U_1, \dots, U_m)$  and  $\mathbf{R} = (R_1, \dots, R_m)$  to get  $\beta_Y$ .
  - (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 ( $\alpha$ ) and residual ( $\varepsilon$ ) 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  $k = 0, \dots, p - 1$ :
  - (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  $M_0, M_1, \dots, M_p$  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:

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

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

### Shrinkage - Adaptive 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  $\beta$  by minimizing

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

where  $w_j = |\hat{\beta}_{OLS}|$  and  $\lambda_n$  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 confoundings. 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.

Table 1: Methods and Descriptions

Methods	Short Description
Oracle	Model with perfect information, only all true covariates.
T-test	Testing covariates on treatment outcome to see which have a significant association
Experience Based	Selecting 5 correct and 5 incorrect 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
Forward Selection	Beginning with a null model, adding covariates that have a significant association treatment

## 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 average MSE. We adopt a Gaussian General Linear Regression Model with an identity link function, incorporating a natural logarithm transformation of MSE to derive main and interaction effects as percentage changes.

### Simulation Results

Figure 1 illustrates the Bias and Mean Squared Error (MSE) for the various variable selection methods across diverse data generating scenarios. We introduce 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 biased estimates, possibly attributable to a violation of the No Unmeasured Confounders assumption. Recall that the ‘Experience-Based’ method consistently selects five true and five random confounders.

Conversely, the bias associated with data-driven variable selection algorithms is centered around zero, suggesting acceptability. This observation holds across different data generation scenarios, considering variations in the number of true covariates and the scale of available data (sample size and total number of predictors). Notably, all variable selection methods result in an Inverse Probability Weighting (IPW) estimator with zero bias, on average, aligning with the anticipated behavior for this class of estimators.

Figure 1 also incorporates the variance of the estimator through Mean Squared Error (MSE). We utilize the number of true confounders as a proxy for the difficulty of the model-building task. It is evident that all data-driven methods exhibit a noticeable increase in MSE as the number of true confounders that need to be accounted for rises. This escalation is likely driven by an increase in variance, given the observation that the average bias was consistently distributed around zero with minor variations for the data-driven methods. 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)

Relative ATE Bias (Bias/True ATE) and MSE for IPW Estimator under varying conditions

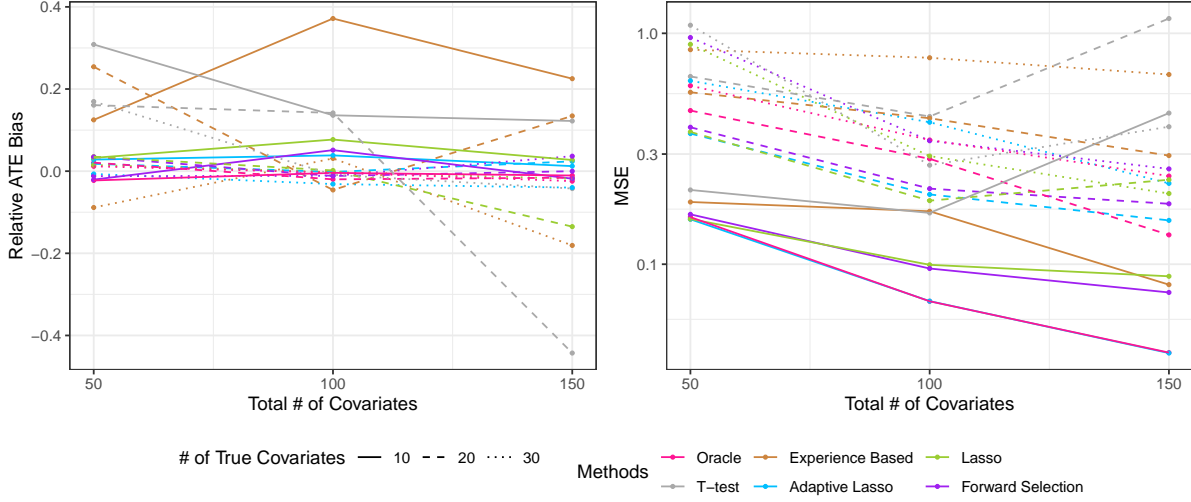


Figure 1: CAPTION FOR THIS PLOT REQUIRED

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).

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, as the number of potential predictors grows, every variable selection method tends to increase the ratio of all selected predictors to the number of true predictors. Additionally, on average, as the size of the true confounder space

## Covariate Selection Results Under varying condiitons

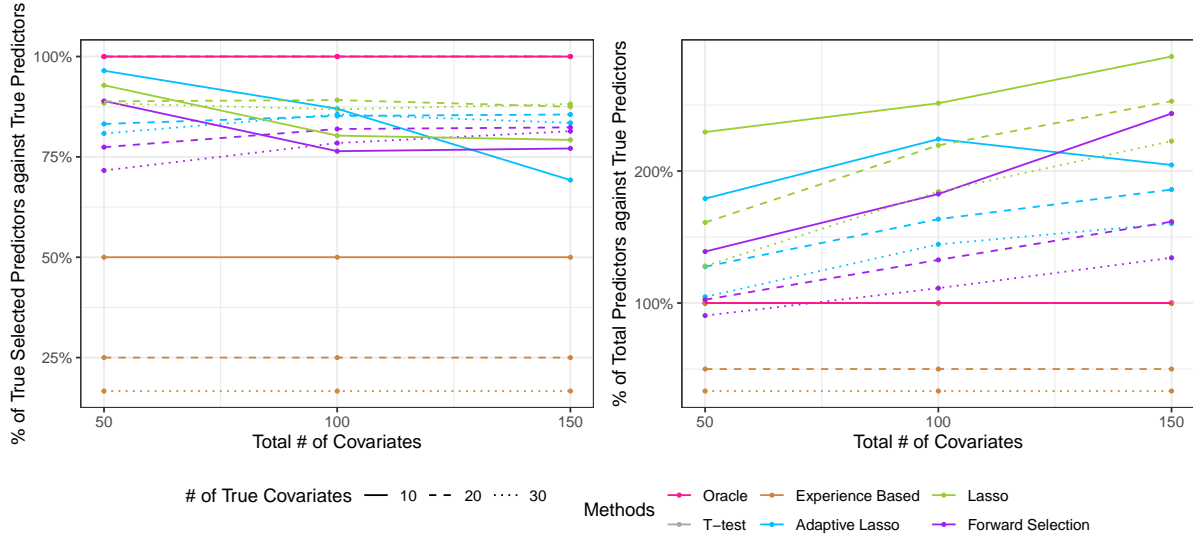


Figure 2: CAPTION FOR THIS PLOT REQUIRED

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 1 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



Table 1: Gaussian GLM with log-transformed response effect estimates. Coefficients are exponentiated and present the effect as % change

Predictor	Estimate	95% CI	P-value	P<0.05
<b>Variable Selection Methods</b>				
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	
<b>Other Main Effects</b>				
Total Covariates 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	
<b>Interaction Terms</b>				
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	

<sup>a</sup> Regression model explains 10.26% of variation in Squared Errors of IPW estimator

<sup>a</sup> Variable selection methods are compared with the reference 'Oracle' level

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 variable selection method on MSE, and contrasted with the Oracle method. It is expected that the Oracle method will correctly specify the model at all times and therefore variance of an IPW estimator will only depend on the sampling variability. As expected, this estimator has the lowest MSE through lowest bias and variance because model misspecification does not occur.

Table 1 provides expected percentage increase in MSE for each variable selection method under different data generating schemes, after adjusting 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.

In a scenario with a relatively low number of true confounders, adaptive lasso variable selection produces an Inverse Probability Weighting (IPW) estimator with Mean Squared Error (MSE) almost identical to the best-case scenario. We estimate that the average change in MSE for IPW based on adaptive lasso variable selection, compared with IPW based on all true covariates, was PLACEHOLDER () when the number of true covariates was 10, after adjusting for other predictors. For detailed estimates for all variable selection methods under different values of true covariate space size, please refer to the appendix.

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.

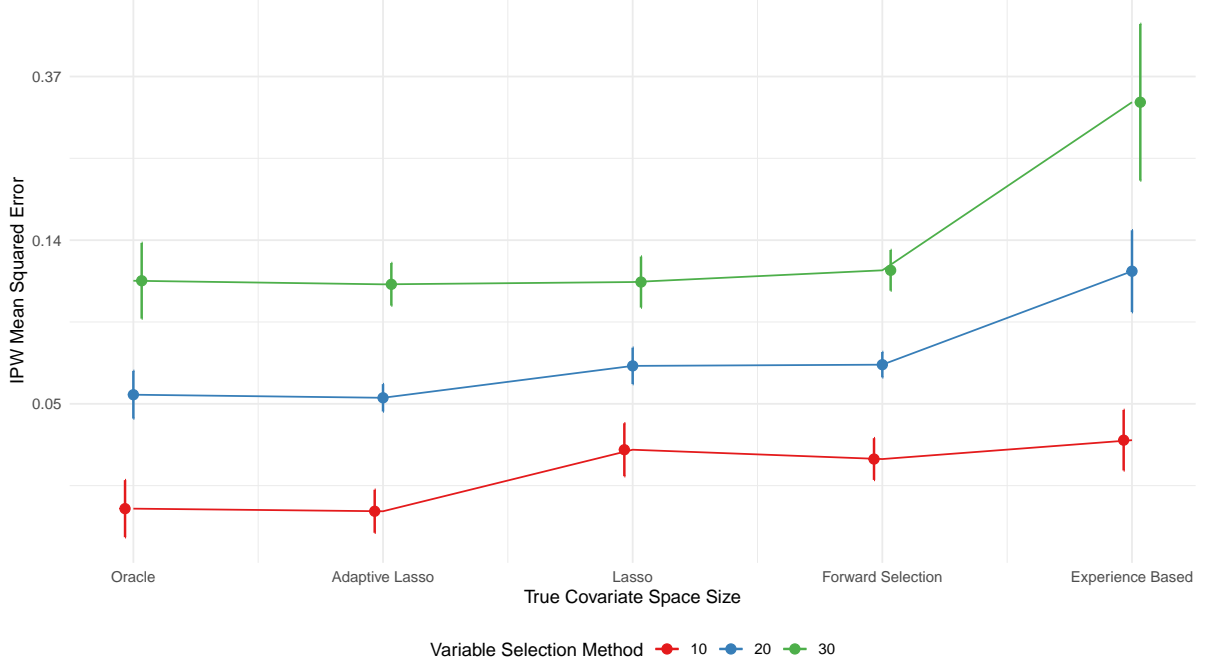


Figure 3: Marginal effect of variable selection method on MSE

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 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 estimated effects might not have actual reflection in the data. Oracle variable selection method always identifies 100% of correct covariates. We use our regression model to estimate expected MSE for the Oracle variable selection method at values lesser than 100% to enable comparison of this benchmark method with other variable selection methods.

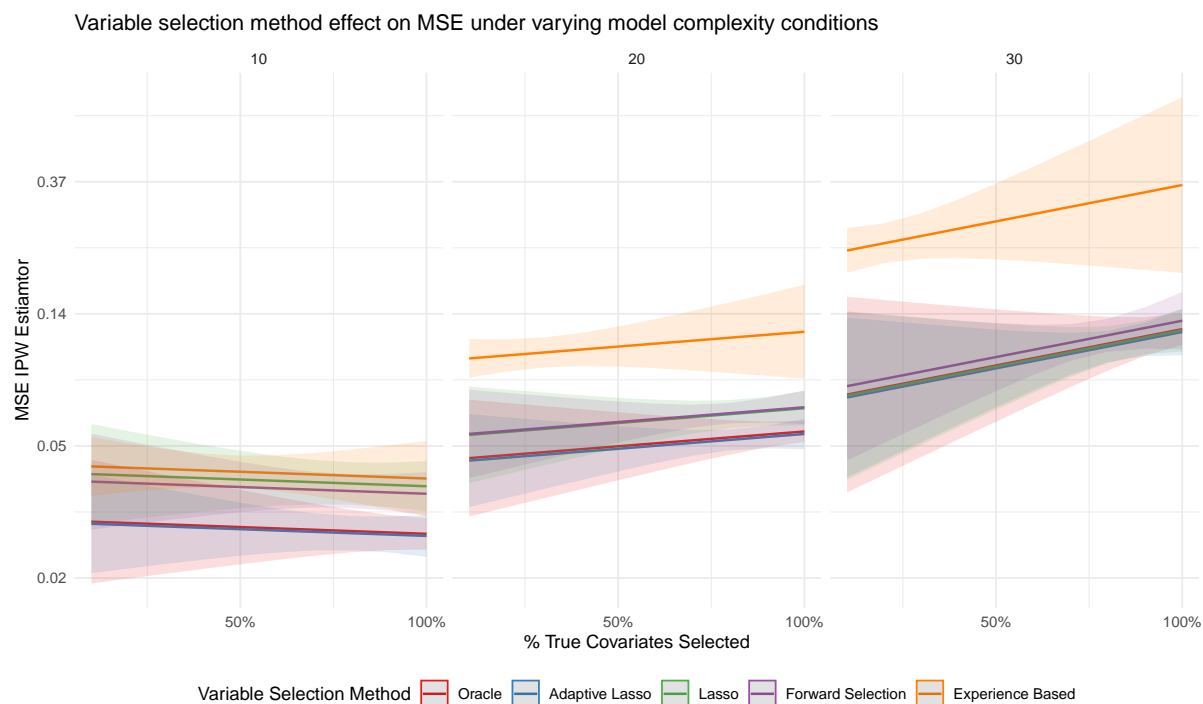


Figure 4: Effects of variable selection models are given for scenarios with 10, 20, and 30 true confounders

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**

Will change this section the most

## **Conclusion**

## 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/10.1198/0162145060000000735>.

## Appendix

### Tables

Mean Squared Error of Adjusted Treatment Effect for select Propensity Score Variable Selection Methods

Method	Propensity Score Stratification <sup>1</sup>				Inverse Probability Weighting <sup>1</sup>			
	Bias	MSE	SE	95% CI	Bias	MSE	SE	95% CI
Oracle	0.0037	8.410646e-05	0.116	0.974, 1.430	−0.0078	9.214985e-05	0.131	0.933, 1.4
Adaptive Lasso	0.0042	8.446108e-05	0.129	0.942, 1.448	−0.0014	8.569968e-05	0.107	0.980, 1.3
Lasso	0.0107	1.039089e-04	0.134	0.981, 1.506	0.0019	1.061700e-04	0.125	0.990, 1.4
Forward Selection	0.0130	1.044025e-04	0.135	0.981, 1.510	0.0052	1.130763e-04	0.154	0.935, 1.5
Experience Based	0.0724	1.516325e-04	0.115	1.038, 1.489	0.0642	1.525472e-04	0.115	1.029, 1.4
T-test	0.0711	2.286879e-04	0.130	1.049, 1.558	0.0711	2.286879e-04	–	–, –

<sup>1</sup>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)

Mean Squared Error of Adjusted Treatment Effect for select Propensity Score Variable Selection Methods

n	Propensity Score Stratification <sup>1</sup>				Inverse Probability Weighting <sup>1</sup>			
	Bias	MSE	SE	95% CI	Bias	MSE	SE	95% CI
Oracle								
1500	0.0104	1.683509e-04	0.174	0.863, 1.544	0.0024	1.913133e-04	0.194	0.815, 1.576
3000	0.0098	7.336256e-05	0.111	0.942, 1.377	−0.0139	7.758931e-05	0.123	0.895, 1.377
4500	0.0178	5.093292e-05	0.095	1.047, 1.421	−0.0070	5.412340e-05	0.110	0.994, 1.424
6000	−0.0232	4.377948e-05	0.085	1.043, 1.377	−0.0128	4.557334e-05	0.097	1.030, 1.411
Adaptive Lasso								
1500	0.0227	1.544392e-04	0.189	0.839, 1.579	0.0277	1.562887e-04	0.153	0.915, 1.514
3000	0.0004	9.572008e-05	0.135	0.953, 1.483	−0.0073	9.706605e-05	0.112	0.990, 1.430
4500	0.0034	3.983297e-05	0.101	0.959, 1.353	−0.0111	4.170648e-05	0.082	0.981, 1.303
6000	−0.0099	4.785207e-05	0.092	1.016, 1.378	−0.0147	4.773748e-05	0.080	1.036, 1.348
Lasso								
1500	0.0431	1.930140e-04	0.201	0.872, 1.659	−0.0063	2.021053e-04	0.174	0.876, 1.557
3000	−0.0009	9.500305e-05	0.128	0.955, 1.455	0.0253	9.312165e-05	0.121	0.994, 1.468
4500	−0.0146	5.284774e-05	0.102	1.037, 1.435	−0.0185	5.312026e-05	0.099	1.038, 1.427

6000	0.0177	6.020197e-05	0.091	1.100, 1.457	0.0093	6.141411e-05	0.094	1.085, 1.455
Forward Selection								
1500	0.0319	1.948440e-04	0.200	0.862, 1.647	0.0461	2.143709e-04	0.232	0.814, 1.723
3000	0.0108	9.528285e-05	0.130	0.963, 1.471	-0.0249	1.013384e-04	0.146	0.894, 1.468
4500	-0.0093	5.265025e-05	0.103	1.040, 1.443	-0.0086	5.607398e-05	0.117	1.013, 1.472
6000	0.0214	6.004828e-05	0.092	1.102, 1.463	0.0096	6.424452e-05	0.106	1.063, 1.478
Experience Based								
1500	0.1115	2.843693e-04	0.162	0.980, 1.616	0.1030	2.876405e-04	0.162	0.971, 1.607
3000	0.0750	1.634612e-04	0.122	1.054, 1.531	0.0667	1.625917e-04	0.122	1.045, 1.523
4500	0.1517	6.244657e-05	0.090	1.128, 1.481	0.1403	6.356623e-05	0.090	1.116, 1.470
6000	-0.0485	9.625289e-05	0.086	0.990, 1.327	-0.0530	9.639052e-05	0.086	0.984, 1.323
T-test								
1500	-0.0985	5.784727e-04	0.182	0.768, 1.480	-0.0985	5.784727e-04	-	-, -
3000	0.2929	1.287203e-04	0.128	1.249, 1.749	0.2929	1.287203e-04	-	-, -
4500	0.0038	8.154148e-05	0.105	1.049, 1.459	0.0038	8.154148e-05	-	-, -
6000	0.0940	7.468183e-05	0.094	1.171, 1.538	0.0940	7.468183e-05	-	-, -

<sup>1</sup>Average value for combination of 500 of each combination of potential covariates (50, 100, 150) and true covariates (10, 20, 30)

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8227634/#:~:text=In%20the%20Obstetrics%20and%20Perio>

[https://accpjournals.onlinelibrary.wiley.com/doi/abs/10.1592/phco.23.8.1037.32876?casa\\_token=-v6XJv5w0DkAAAAA:DJdg3MRA\\_a9xY093j773zpMOCKSILYQ2jeJSjeCjR8hc7\\_J-WIoEeOiwNEuCvI-oUhyVgf-9bc6aNA](https://accpjournals.onlinelibrary.wiley.com/doi/abs/10.1592/phco.23.8.1037.32876?casa_token=-v6XJv5w0DkAAAAA:DJdg3MRA_a9xY093j773zpMOCKSILYQ2jeJSjeCjR8hc7_J-WIoEeOiwNEuCvI-oUhyVgf-9bc6aNA)

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3065283/>

Use of random forest with a lot of variables <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5920646/>

- Super high dim variable selection and ANDI data

[https://onlinelibrary.wiley.com/doi/full/10.1111/biom.13625?casa\\_token=TPy\\_un3cN-EAAAAA%3Aasm\\_EoAthJYwoV489HGuY\\_CnDi0BV7JIWSOvWyTcynDH6dYDq0TOQJGe6pHMzhcFI4z42w](https://onlinelibrary.wiley.com/doi/full/10.1111/biom.13625?casa_token=TPy_un3cN-EAAAAA%3Aasm_EoAthJYwoV489HGuY_CnDi0BV7JIWSOvWyTcynDH6dYDq0TOQJGe6pHMzhcFI4z42w)

- Some study with bias and mse

[https://onlinelibrary.wiley.com/doi/full/10.1002/sim.4469?casa\\_token=N3XCrX6yrUsAAAAA%3AQ2uBYcmIikpXzZnL8\\_FPmqQhPGPP1No3OxG74JK0fw](https://onlinelibrary.wiley.com/doi/full/10.1002/sim.4469?casa_token=N3XCrX6yrUsAAAAA%3AQ2uBYcmIikpXzZnL8_FPmqQhPGPP1No3OxG74JK0fw)



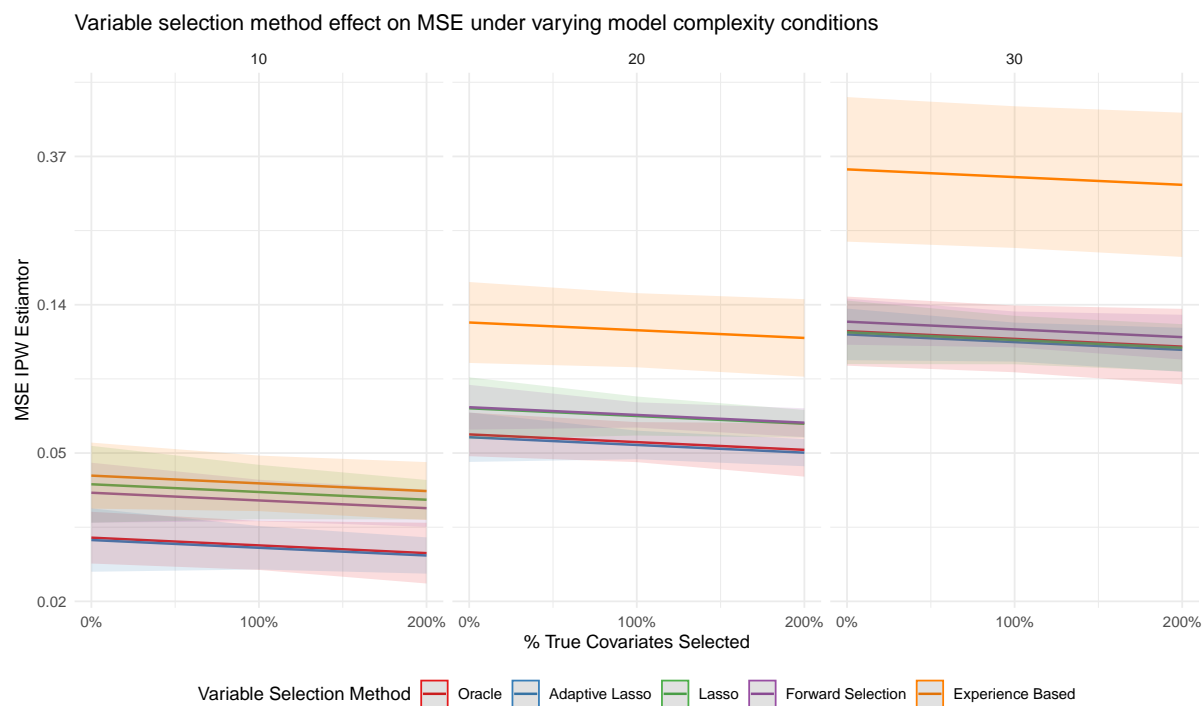


Figure 5: Effects of variable selection models are given for scenarios with 10, 20, and 30 true confounders