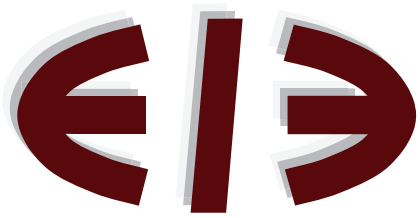
EERI
Economics and Econometrics Research Institute

A Comparison of Bayesian and Frequentist Variable Selection Methods for Estimating Average Treatment Effects in Logistic Regression

Alex H. Martinez, Brian Christensen, Elizabeth F. Sutton, Andrew G. Chapple

EERI Research Paper Series No 01/2025

ISSN: 2031-4892



EERI
Economics and Econometrics Research Institute
Avenue Louise
1050 Brussels
Belgium

Tel: +32 2271 9482
Fax: +32 2271 9480
www.eeri.eu

A Comparison of Bayesian and Frequentist Variable Selection Methods for Estimating Average Treatment Effects in Logistic Regression

Alex H. Martinez, Brian Christensen, Elizabeth F. Sutton, Andrew G. Chapple

Abstract

In many manuscripts, researchers use multivariable logistic regression to adjust for potential confounding variables when estimating a direct relationship of a treatment or exposure on a binary outcome. After choosing how variables are entered into that model, researchers can calculate an estimated average treatment effect (ATE), or the estimated change in the outcome probability with and without an exposure present. Which potential confounding variables should be included in that logistic regression model is often a concern, which is sometimes determined from variable selection methods. We explore how forward, backward, and stepwise confounding variable selection estimate the ATE compared to *spike-and-slab* Bayesian variable selection across 1,000 randomly generated scenarios and various sample sizes. Our large simulation study allow us to make pseudo-theoretical conclusions about which methods perform best for different sample sizes, rarities of coutcomes, and number of confounders. An R package is also described to implement variable selection on the confounding variables only and provide estimates of the ATE. Overall, results suggest that Bayesian variable selection is more appealing in smaller sample sizes than frequentist variable selection methods in terms of estimating the ATE. Differences are minimal in larger sample sizes.

1 Introduction and Methods

In many observational studies, authors are interested in estimating the direct effect of a binary exposure on a binary outcome. Often these observational studies are filled with additional confounding variables, which can mislead researchers about the direct effect of that exposure on outcome, and are most commonly accounted for via multivariable logistic regression. Naturally, choosing which to adjust for is a widely discussed topic with no consensus guidelines [1], [2] [3] [4] [5]. Some reccomendations (i.e. avoiding colliders, mediators, instrumental variables, etc) require the researchers to know the causal structure which is often unknown. Variable selection is an automated way to choose what variables should be included in the abscence of that causal knowledge.

For any regression model chosen, researchers can compute an Average Treatment Effect (ATE) which measures the adjusted change in the outcome probability when a typical observation goes from not exposed to exposed (or treated vs not treated). The ATE is a measure of the overall effect of a treatment or intervention from a confounding-adjusted estimated model averaged over the sample. Since this quantity depends on which covariates are adjusted for, variable selection plays a key role in estimating this quantity.

This research aims to investigate three common frequentist methods of variable selection for confounder selection (forwards, backwards, and stepwise), and how they perform compared to *spike-and-slab* Bayesian Variable Selection (BVS) method for estimating the ATE of an exposure/treatment using logistic regression models. Here we used a *spike-and-slab* prior first described by Geweke (1996), but with a dirac measure for the spike which directly samples 0 coefficients independently [6], [7]. We used independent normal priors on the covariate regression coefficients and did not investigate alternative slab formulations (like Cauchy or T-distributions) or spike formulations (like the horseshoe or slab-slab approaches). Our justification here is that the coefficient estimates from deviance-based variable selection methods (i.e. stepwise) are asymptotically efficient (i.e. normally distributed) so assuming a normal prior on our coefficients made this comparison fair.

A simulation study will be utilized to analyze thousands of different randomily generated scenarios to see which of these 4 approaches performs best in terms of estimating ATEs. We perform simulation across 1,000 randomly generated scenarios instead of cherry-picking a few to really answer the question of what works best, since in logistic regression there are not theoretical justifications of this. For each simulation scenario, we generated 100 random datasets of various sample sizes and compared the bias and coverage probabilities of each method for estimating the ATE.

There are many other methods that exist for estimating average treatment effects that we do not describe here. These include propensity score matching, using the propensity score as a covariate, the inverse probability treatment weighting estimator (IPTW), the augmented inverse probability treatment weighting estimator (aIPTW), the do-operator, and instrumental variable methods [8], [9], [10] [11]. We decided to only explore covariate adjustment in the variable selection setting since that is what is most commonly done in the literature. We also decided to only explore traditional confounding relationships, where covariates affect the outcome and the exposure directly. We avoided colliders, mediators, proxy variables, and exploration of under or overcontrol directly [12], [13], [14], [15]. Instead we are choosing to focus on automated variable selection methods in assessing ATEs which is common practice. We explore both normally distributed correlated covariates and binary correlated covariates across these thousands of randomly generated scenarios to determine which variable selection performs best.

The rest of this paper is structured as follows. In section 2 we will outline the assumed logistic regression model, calculation of the average treatment effect (ATE), deviance-based frequentist confounder selection, and Bayesian confounder selection. In section 3 we will discuss the design of the simulation study including generation of random scenario truths. In section 4 we will discuss the results from our study separated by sample size. Three applications are explored in section 5 and discussion is included in section 6. Additionally we have included supplemental material containing supplemental tables and a user-guide for the R package (`VARSELECTEXPOSURE`), which contains functions for implementing all the confounding variable selection methods described here while estimating the ATE.

2 Methods

This paper explores the estimation of a treatment effect of a binary exposure E_i on a binary outcome Y_i in the presence of potential confounding variables $\boldsymbol{X}_i$ that appear linearly in a true logistic regression relationship. Our goal is to determine whether *spike-and-slab* Bayesian variable selection or Deviance-based variable selection (Forwards, Backwards or Stepwise) performs better in terms of estimating the average treatment effect of E_i on Y_i. We assume the following logistic regression model both in truth and for our full variable selection models:

$$logit(P[Y_i = 1|E_i, \mathbf{X}_i, \boldsymbol{\beta}]) = \beta_0 + \beta_E E_i + \beta_1 X_{i1} + ... + \beta_p X_{ip}, \tag{1}$$

where $\beta_0, \beta_1, ..., \beta_p$ are real constants. Observational studies aim to find causal relationship between a binary treatment variable and an outcome, however at an individual level these causal effects are impossible to calculate because of the inability of subjects to have both the exposure and non-exposure simultaneously. This is sometimes referred to as the fundamental problem of causal inference [16], and it is because of this that researchers often utilize the Average Treatment Effect (ATE). The ATE is the average difference between the predicted probability that $Y_i = 1$ given the data and that patient i is in the treatment/exposure group, and the predicted probability that $Y_i = 1$ given the data and that patient i is not a member of the treatment/exposure group. Formally, the true ATE can be calculated as

$$ATE = \frac{1}{n}\sum_{i}^{n}\big[P(Y_i = 1|E_i = 1, \mathbf{X}_i, \boldsymbol{\beta}) - P(Y_i = 1|E_i = 0, \mathbf{X}_i, \boldsymbol{\beta})\big]. \tag{2}$$

It can therefore be interpreted as a contrast of means of counterfactual outcomes [17]. We can use an estimated adjusted statistical model $\hat{\boldsymbol{\beta}}$ with some (potentially different than reality) set of covariates to estimate ATE, obtaining $\widehat{ATE}$, by the formula

$$\widehat{ATE} = \frac{1}{n}\sum_{i}^{n}\big[P(Y_i = 1|E_i = 1, \mathbf{X}_i, \hat{\boldsymbol{\beta}}) - P(Y_i = 1|E_i = 0, \mathbf{X}_i, \hat{\boldsymbol{\beta}})\big]. \tag{3}$$

In the frequentist methods, the maximum likelihood estimates for the parameters are used to calculate these probabilities, while in the Bayesian method we use the posterior parameter estimates in place of $\hat{\boldsymbol{\beta}}$ and obtain a posterior sample of ATE. Here the bth posterior sample of β^b is used in (3) to estimate the posterior sample ATE^b. We then compute the Bayesian posterior mean estimate, $\widehat{ATE}$, as the average of the $\{ATE^b\}_{b=1}^{B}$ posterior samples.

2.1 Frequentist Variable Selection Methods

First we describe forward, backward, and stepwise confounding variable selection using deviance-based metrics in logistic regression. The general idea of variable selection is similar in principal between these methods, with distinctions in their computational approaches. The difference in deviances of a multivariable logistic regression model with and without covariate k tests whether the coefficient β_k is equal to 0, which follows a chi-squared distribution with 1 degree of freedom under the

null hypothesis. If the .05 significance level is used, one can add (delete) a confounding variable k if the deviance difference is greater (less) than 3.84.

The forward selection algorithm in the context of adjusting for potential confounding variables $X_1, ..., X_p$ in estimating an average treatment effect of E_i on Y_i is laid out as follows:

1. Begin with a model containing only binary exposure E_i. Calculate this model's deviance, denoted D_0.
2. Calculate $D_1, D_2, ..., D_k$ for $k = 1, 2, ..., p$, where each D_k represents the deviance resulting from the model adding the k^{th} potential confounding variable. The corresponding covariate $\text{argmin}_k\{D_1, D_2, ..., D_p\}$ is now the best candidate for addition, indicated by the smallest deviance, denoted $D_* = \min\{D_1, D_2, ..., D_k\}$.
3. Calculate the test statistic $T = |D_0 - D_*|$ and compare it to a chi-square distribution with 1 degree of freedom. If significant at $\alpha = 0.05$, i.e. if $T > 3.84$, we add the variable to our model, reassign $D_0 = D_*$ and remove the added variable from consideration for future additions. The size of the set of potential confounding variables is reduced by 1.
4. Repeat this process until none of the remaining variables significantly improve our model when added.

Using forward variable selection is advantageous because it provides smaller models on average, since it starts with only the exposure included making it less susceptible to problems caused by multicollinearity and overfitting. In forward selection once a variable is added to the candidate model it cannot be removed, which may hurt estimation of our direct exposure effect [18].

The backward variable selection algorithm can be seen as the "opposite" of forward selection. In the context of confounding variable selection, the algorithm works as follows:

1. Begin with the full model containing the binary exposure variable E_i and all covariates $X_1, ...X_p$. Calculate this model's deviance, called D_0.
2. Calculate $D_1, D_2, ..., D_k$ where $k = 1, 2, ..., p$, where each D_k represents the deviance resulting from the model removing the k^{th} covariate. Let $D_* = \min\{D_1, D_2, ..., D_k\}$, the corresponding covariate$\text{argmin}_k\{D_1, D_2, ..., D_p\}$ is now a candidate for deletion.
3. Calculate the test statistics $T = |D_0 - D_*|$ and compare it to a chi-square distribution with 1 degree of freedom. If insignificant at $\alpha = 0.05$, i.e. if $T < 3.84$, we delete the variable from our model and reassign $D_0 = D_*$.
4. Repeat this process (letting p denote the new number of covariates included) until none of the remaining variables significantly improve our model when removed.

It is worth noting that because these models will be used for effect estimation, our binary exposure variable E_i is never a candidate for deletion. One advantage of the backward selection method over forward selection is in its ability to assess the joint predictive ability of multiple variables which forward selection may not do as efficiently. Backward selection alleviates this by beginning with a model that includes all variables. Similar to forwards selection however, due to the fact that deleted variables cannot be considered there may be unseen problems that arise in the case that a deleted variable would become important in a later visited model post-deletion.

The stepwise variable selection algorithm can be thought of as a compromise between both the forward and backward variable selection techniques that addresses these issues. The algorithm in confounding variable selection is as follows:

1. Begin with empty model similar to in forward selection, and attempt to add a variable using the same process described for forward selection.
2. Once a variable has been added, attempt to delete a variable from the current model using the same process as in backward variable selection. Note, in the first iteration this step will not change the model.
3. Continue sequentially attempting to add and delete variables until there are two successive instances of no change, i.e. either no variables are added then no variables are deleted, or no variables are deleted then no variables are added.

The stepwise variable selection method is perhaps the most popular method of variable selection. Some reasons for this include that it is easy to apply in statistical software, it allows evaluation of different models that may have otherwise not been proposed, and it remains objective in that the same variables are generally chosen from the same data set no matter who is conducting the analysis, allowing researchers the ability to validate the model and reproduce results [19] [20]. It's also been said to prevent researchers from thinking about the problem itself [18].

2.2 Bayesian Variable Selection

Frequentist logistic regression provides one value of $\hat{\boldsymbol{\beta}}$ based on a given data set. This is in contrast to the Bayesian framework, which obtains a random sample (posterior distribution) of $\boldsymbol{\beta}$ values where each posterior sample is used in estimating ATE. Here we explore the *spike-and-slab* method described by many authors but initially introduced by George and Mcculloch. [7] [21] [22] [23]. We chose this approach because it mimicks the selection method seen in stepwise variable selection, with coefficients β_k iteratively being added (i.e. $\beta_k \neq 0$) or deleted (i.e. $\beta_k = 0$) from a regression model.

This idea is directly applicable to variable selection and creates a foundation for the central idea of Bayesian Variable Selection (BVS). In essence, the frequentist task of variable selection is transformed into a problem of parameter estimation in the Bayesian framework. As opposed to searching for one "optimal" model, this method explored the uncertainty on whether a covariate k should be included in the model by repeatedly proposing setting $\beta_k = 0$ (removal) or $\beta_k \neq 0$ (addition) in various Markov Chain Monte Carlo iterations [23]. Throughout these iterations, candidate confounding variables are constantly being added and deleted similar to stepwise variable selection, but with each iteration we also obtain an estimate of β_0, β_E and ATE. β_0 and β_E are not candidates for variable selection due to our interest in estimating ATE.

For variable selection of potential confounders $k = 1, ..., p$, we introduce the random binary indicator variable η_k. This variable represents whether or not a given β_k is non-zero (included in the adjusted model). We assume that each β_k is distributed *a priori* as

$$\beta_k|\eta_k \sim \eta_k N(0, \sigma^2) + (1 - \eta_k)\delta_0(\beta_k), \tag{1.11}$$

where $\delta_0(\cdot)$ is a point mass function that is always 0 and $N(0, \sigma^2)$ is our prior distribution on non-zero β_k entries with mean 0 and variance σ^2. For our study, we use a weakly informative $\sigma^2 = 1$, since we are applying this in logistic regression settings and aim for models that do not induce separation. Similarly, we assume that $\beta_E \sim N(0, \sigma^2)$ and that $\beta_0 \propto 1$. Here $P[\eta_k = 1] = \pi_\eta = 1 - P[\eta_k = 0]$ is a hyperparameter denoting the prior probability that any variable is included in the multivariable regression model. This is a parameter that we explore the effect of in our simulation study, with a higher (lower) value including more variables in visited regression models *a priori*. We specifically explore choices of $\pi_\eta = .1, .25, .5$ in our simulation study.

To obtain posterior samples from this multivariate posterior distribution for $(\beta_0, \beta_E, \beta_1, ..., \beta_p, \eta_1, ..., \eta_p)$, we perform Gibbs sampling with adaptive Metropolis-Hastings steps. A generic iteration of the Markov Chain Monte Carlo (MCMC) proceeds as follows:

- **Sampling β_0, β_E, and $\beta_k|\eta_k = 1$:** This is done sequentially via Gibbs steps and adaptive Metropolis-Hastings. Let θ denote any one of these parameters. We propose $\theta* \sim N(\theta, c_\theta)$ where c_θ is adaptvely adjusted every 100 iterations so that the acceptance rate for that parameter over those 100 iterations is between .2 and .6. If the acceptance rate is too small, we divide c_θ by 2, and if it's too large, we multiply by 2. The proposal value $\theta*$ is accepted over the previous value with probability equal to the prior ratio (which is 1 for β_0) times the likelihood ratio.

- **Sampling $(\boldsymbol{\beta}, \boldsymbol{\eta})$:** For each iteration of the MCMC, we randomly pick a value k from $1, ..., p$ with equal probability and do the following.

  - If $\eta_k = 1$, i.e. a covariate k is currently included in the logistic regression model, we propose removing it from the regression by setting $\beta_k^* = 0$ and $\eta_k^* = 0$. This move is accepted with probability equal to the resulting likelihood ratio times the prior ratio for β_k and $(1 - \pi_\eta)/\pi_\eta$ which is the prior ratio for η_k. You can see here how larger values affect the MCMC, which decrease this ratio and make the likelihood of accepting $\beta_k^* = 0$ over $\beta_k \neq 0$ smaller for any given iteration.
  - If $\eta_k = 0$, i.e. a covariate k is currently not included in the logistic regression model, we propose settign $\eta_k^* = 1$ and β_k equal to some non-zero value. This sampling step can be tricky, since making β_k non-zero greatly adjusts how other parameters β_m for $m \neq k$ affect the outcome model. These β_m coefficients have been sampled without covariate k and may settle around their maximum likelihood estimates with large sample sizes, making the choice of β_k^* challenging to accept. To get around this, we randomly generate β_k^* from $N(\hat{\beta}_k, 1)$. Here, $\hat{\beta}_k$ is the maximum likelihood value of the logistic regression with variable X_k added given the remaining $\boldsymbol{\beta}$ values are equal to the previous iteration's values. This move is accepted with probability equal to the resulting likelihood ratio times the prior ratio for β_k and $\pi_\eta/(1 - \pi_\eta)$ which is the prior ratio for η_k.

We continue this process for 2000 iterations, and discard or "burn-in" the first 1000. Adaptive adjustment of c_θ for each parameter is also done during the "burn-in" period but not afterwards. This is intended to protect against an unfavorable starting point, where regions that are actually low in probability may be over-sampled. After this burn-in period, the posterior sample of our model parameters is $\{\beta_0^b, \beta_E^b, \beta_k^b, \eta_k^b\}_{b=1}^B$, and our Bayesian estimate $\widehat{ATE}$ is the average of the $\{ATE^b\}_{b=1}^B$

posterior samples. We also compute the posterior probability of inclusion is the percentage of times each $\eta_k = 1$, i.e. the average of $\{\eta_k^b\}_{b=1}^B$, which can be used as an indicator for how "important" each variable is. In the context of this research, we are much more interested in the ATE estimates than we are in the posterior distribution of the coefficients. However, functions included in the R package `VARSELECTEXPOSURE` do output posterior distributions of all model quantities.

3 Data Generation & Randomization of Parameter Settings for Simulation Study

In this section we discuss the data generation model for $\{X_i, E_i, Y_i\}_{i=1}^n$ based on a true set of parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \Sigma, \boldsymbol{\mu})$ and how we randomly generated these parameters. We randomly generate 1,000 sets of parameters for $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \Sigma, \boldsymbol{\mu})$ in order to pseduo-theoretically test which of these discussed variable selection method performs the best and not cherry-pick several scenarios that may favor one method vs another.

3.1 Data Generation Model

We randomly generated outcome, exposure, and covariate data $\{X_i, E_i\}$ based on a true set of parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma$, and $\boldsymbol{\mu}$. For our binary outcome data we use a logit model to get the probability that $Y_i = 1$ for each patient given our randomly generated exposure variable, covariates, and outcome-specific parameters. These probabilities are used to generate each individual Y_i from a Bernoulli distribution. The true model can be written as

$$logit(P[Y_i = 1|E_i, \mathbf{X}_i, \boldsymbol{\beta}]) = \beta_0 + \beta_E E_i + \beta_1 X_{i1} + ... + \beta_p X_{ip} \tag{4}$$

where β_0, β_E and $\beta_1, ..., \beta_p$ are our outcome-specific parameters. We use a similar model in the same manner to generate our binary exposure values E_i, the model can be written

$$logit(P[E_i = 1|\mathbf{X}_i, \boldsymbol{\alpha}]) = \alpha_0 + \alpha_1 X_{i1} + ... + \alpha_p X_{ip} \tag{5}$$

where α_0 and $\alpha_1, ...\alpha_p$ are exposure-specific parameters. For our covariates, we generated both continuous and binary versions of $\mathbf{X}_i$ for analysis. The continuous p-dimensional confounding variable data $\boldsymbol{X}_i$ is sampled from the multivariate normal distribution with parameters $\boldsymbol{\mu}$ and Σ. When generating binary confounding variables $\boldsymbol{X}_i$, the same process is used to generate a latent vector of continuous random data $\mathbf{Z}_i \sim MVN(\boldsymbol{\mu}, \Sigma)$, then the Binary covariate matrix entries are defined as $X_{ik} = I[Z_{ik} > 0]$, where I is an indicator function on whether or not an entry of $\mathbf{Z}$ is positive. If the entry Z_{ik} is positive, the corresponding entry of $\mathbf{X}$ will be 1, it will be 0 otherwise. For each dataset generated, we compute the true value of the ATE using (2).

3.2 Random Parameter Generation

Here we explain how we randomly generated the parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \Sigma, \boldsymbol{\mu})$. We begin by randomly generating our intercept coefficients β_0, β_E and α_0 from a standard normal distribution.

We then randomly generate our coefficients $\beta_1, ..., \beta_p$ as well as $\alpha_1, ...\alpha_p$ for our exposure. These are generated from a normal distribution with mean 0 and standard deviation $\frac{1}{2}$, i.e. $\beta_k \sim N(0, \frac{1}{2})$ and $\alpha_k \sim N(0, \frac{1}{2})$. We chose these distributions because on the log-odds scale, small coefficients still have a large effect on the model. Next for both the generated $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ vectors, we randomly choose p^*, the number of parameters that will be set to 0, by randomly selecting a number between 1 and p with equal probability (i.e. randomly drawing from a discrete uniform distribution on the numbers 1 to p). After this number is chosen we randomly select p^* coefficients from the set $\{1, ..., p\}$ without replacement, and set those corresponding coefficients to 0. This is done in both $\beta_1, ..., \beta_p$ and $\alpha_1, ..., \alpha_p$. Here a confounder by definition is a covariate k with $\beta_k \neq 0$ and $\alpha_k \neq 0$, but it's also possible that X_{ik} that affect only E_i or Y_i are meaningful in method comparison.

To generate the covariate data $\boldsymbol{X}_i$ we need a vector of $\boldsymbol{\mu}$ values and a covariate matrix Σ. A p-length vector of μ values was generated from the standard normal distribution, and a $p \times p$ matrix was randomly generated for Σ using the Wishart distribution with p degrees of freedom and an identity matrix.

One problem that may arise in this process is separation, which refers to a situation where all of the observations can be perfectly classified based on a given set of predictor variables, which makes estimating different entries of $\hat{\boldsymbol{\beta}}$ difficult [24]. Separation is called complete when $X_{ik} > c$ for some constant c (or $E_i = 1$) perfectly predicts $Y_i = 1$ or $Y_i = 0$. It's quasi-complete when a linear combination of X_{ik} and E_i perfectly predict Y_i. There have been many possible remedies proposed, such as removing the variable causing separation or adding 0.5 to each of the observations [25]. The presence of separation leads to issues with estimating $\hat{\boldsymbol{\beta}}$ so we took steps to remove scenarios that caused complete separation.

Any scenario that resulted in separation for $n = 1000$ was removed from consideration as pathological, and replaced. To do this, we constructed a 2 by 2 contingency table of E_i and Y_i (and for binary covariates X_{ik} and Y_i. We checked if any count in this table (i.e. $E_i = m, Y_i = j$) happened 0 times. If it did, we removed this scenario from consideration. Without removing these scenarios, estimates of $\beta_E, \beta_1, ..., \beta_p$ could potentially have magnitudes over 1,000, which would create unreliable estimates of ATE and could drastically affect averaged simulation results for a given scenario. We do the same thing with continuous versions of $\boldsymbol{X}_i$ by checking if there exists some c where $X_{ik} > c$ causes complete separation.

We generate the parameter vector $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \Sigma, \boldsymbol{\mu})$ using the process described above until we obtain 1,000 randomly generated scenarios, recording the number of true confounders and other interesting scenario properties. We obtain average operating characteristics for each simulation scenario by simulating 100 randomly generated datasets from each scenario. We perform these simulations by sending out the 1,000 randomly generated scenarios to different cores and simulating 100 data sets each examine using resources provided by the Open Science Grid Cluster, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science [26] [27].

4 Simulation Study Results

To compare the performance of the 6 variable selection methods considered (forwards, backwards, stepwise, BVS with $\pi_\eta = .1, .25, .5$), we examined 4 different measures. First is the mean bias in $\widehat{ATE}$ estimating ATE, the average distance from $\widehat{ATE}$ to the true ATE across the 100 simulation replications. Formally this is

$$\frac{1}{100}\sum_{r=1}^{100} |\widehat{ATE}^r - ATE|, \tag{3.1}$$

where $\widehat{ATE}^r$ is the estimated ATE for the r^{th} generated dataset. The second measure is coverage probability, which tells us the percentage of the time the 95% confidence interval (for the frequentist methods) or the 95% quantile-based credible interval (for the Bayesian methods) contained the true value of the ATE. The third measure is the coverage length, which tells us the average length of the credible/confidence intervals calculated from each method. Lastly we discuss the coverage ratio, which is the coverage probability divided by the coverage length. The coverage ratio is a compromise of coverage probability, where larger is better (more accurate), and coverage length, where smaller is better (more precise). Collectively a larger coverage ratio is better. These results will be categorized by sample size and various scenario properties such as the true number of confounders or how rare Y_i and E_i are.

4.1 Overall

Figure 1 displays the average ATE bias (top left), the coverage probability (top right), the coverage legnth (bottom left), and the coverage ratio (bottom right) along with error bars for the 10 and 90 % quantiles across the 1,000 randomly generated scenarios. We chose to display the 10 and 90 % quantiles instead of the maximum and minimum or the standard deviation since these were nearly identical for the methods considered. These quantiles give a better measure of the central tendency than the outlier scenarios, where all methods performed especially well or poorly.

ATE bias seems to be decreasing as sample size increases, with all methods performing almost the same in the $n = 1000$ section. This is expected since it is well known that frequentist and Bayesian methods are consistent estimators. In the smaller sample sizes ($n = 100$ and $n = 200$), the Bayesian methods outperform the frequentist methods with a lower average ATE bias, but it's traded for a larger variability than the frequentist methods indicated by the quantile bars.

In coverage probability, the frequentist methods perform similarly across all sample sizes, outperforming the Bayesian methods in every sample size. This may be explained in the plot for coverage length, where in every sample size the frequentist methods have a higher average coverage length than the Bayesian methods. Because they have a larger coverage length on average, it makes sense that the coverage probabilities for the frequentist methods would be larger on average than the Bayesian methods.

Bayesian methods outperform the frequentist methods in average coverage ratio across all sample sizes, but not in variability. A larger ratio indicates a larger coverage probability and/or a smaller coverage length, so we will consider larger better. As the sample size increases, the difference in average coverage ratio and variability across methods decreases, but the frequentist methods never overtake the Bayesian methods.

Next we will look at pairwise comparisons of each methods superiority through heat maps. Each cell in these heat maps represents the proportion of simulation scenarios where the row method outperformed the column method. For example a more blue shade in the cell corresponding to the forward row and BVS 0.5 column would indicate a higher proportion of simulation scenarios where the forward method outperformed the BVS 0.5 method, while a more red shade in the same cell would indicate a lower proportion of simulation scenarios where the forward method outperformed the BVS 0.5 method. In

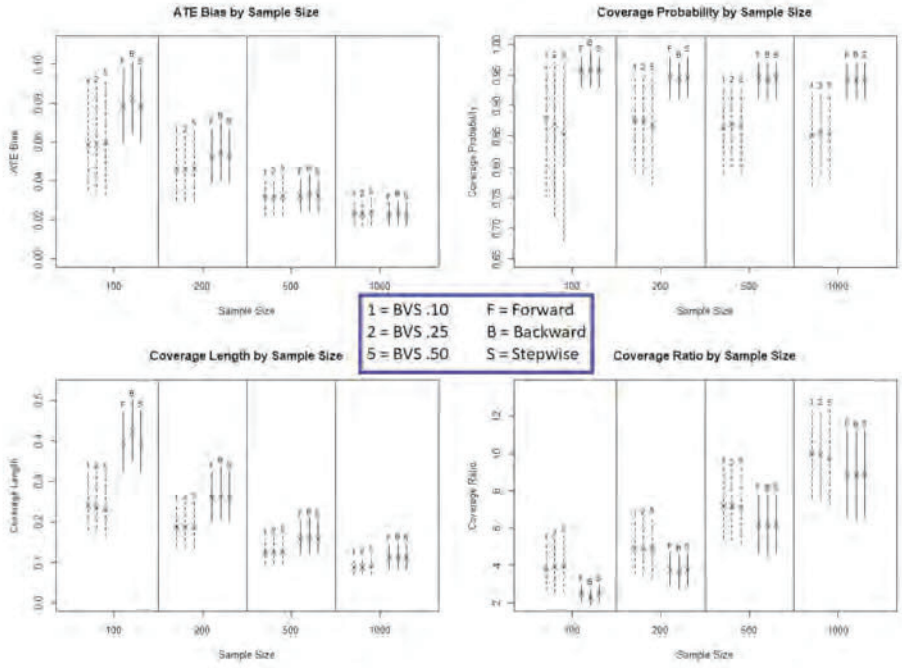


Figure 1: Continuous Simulation Results: Plots displaying average ATE bias (smaller=better), coverage probability (larger=better), coverage length (smaller=better), and coverage ratio (larger=better) for each method separated by sample size.

the context of ATE bias, coverage probability, coverage length, and coverage ratio, we prefer the values to be lower, higher, lower, and higher respectively - and the colors reflect superiority for each measure.

Figure 2 displays the comparative results for the ATE bias. They are separated by sample size, with $n = 1000$ in the top left, $n = 500$ in the top right, $n = 200$ in the bottom left, and $n = 100$ in the bottom right. Here the Bayesian methods outperformed the frequentist methods in a majority of scenarios for $n = 100$ and $n = 200$, indicated by the dark blue cells in the heatmap for the BVS rows and frequentist columns. Bayesian methods had a better performance in more scenarios in $n = 500$, but the light shade indicates that difference is not strong. Similarly, frequentist methods performed better than the Bayesian methods for $n = 1000$, but the shading indicates this difference was not strong. Within frequentist methods, forward and stepwise selection were markedly better than backwards selection for $n = 100$, but that difference dissapeared with sample size. There were not major differences for Bayesian variable selection with different choices of π_η, with all shades being light.

Supplemental figure 2 shows the pairwise comparison of the coverage probability via the same heatmap. In almost all simulation scenarios for each sample size the frequentist methods had a higher average coverage probability than the Bayesian methods. This means that in most simulations scenarios, a higher proportion of the confidence intervals calculated by the frequentist methods contained the true ATE value than the credible intervals calculated by the Bayesian methods. This trend is consistent in each sample size, but does get weaker as the sample size decreases.

Supplemental figure 3 displays the heatmap showing the pairwise comparative credible/confidence interval length between methods. For all sample sizes, the average length of the credible intervals calculated by the Bayesian methods was smaller than that of the confidence intervals calculated by the frequentist methods in almost every simulation. This result can help explain the trend in the previous set of heat maps, as a smaller coverage length will obviously result in a lower proportion of intervals containing the true ATE.

Figure 3 displays the pairwise comparative differences in the coverage ratio, which is a compromise between the coverage probability and coverage length. Here a more blue (red) cell indicates a higher (lower) proportion of simulations where the row method had a larger coverage ratio than the column method. We chose to display this in the main manuscript since this represents a compromise between coverage probability of the ATE and the length of the interval used for that probability.

In each sample size the Bayesian methods outperform the frequentist methods in most of the scenarios indicated by the dark blue shades for BVS rows and frequentist columns of Figure 3. In addition, the forward and stepwise methods outperformed the backward method very often in the $n = 100$ plot, but the comparative performance evens out as sample size increases.

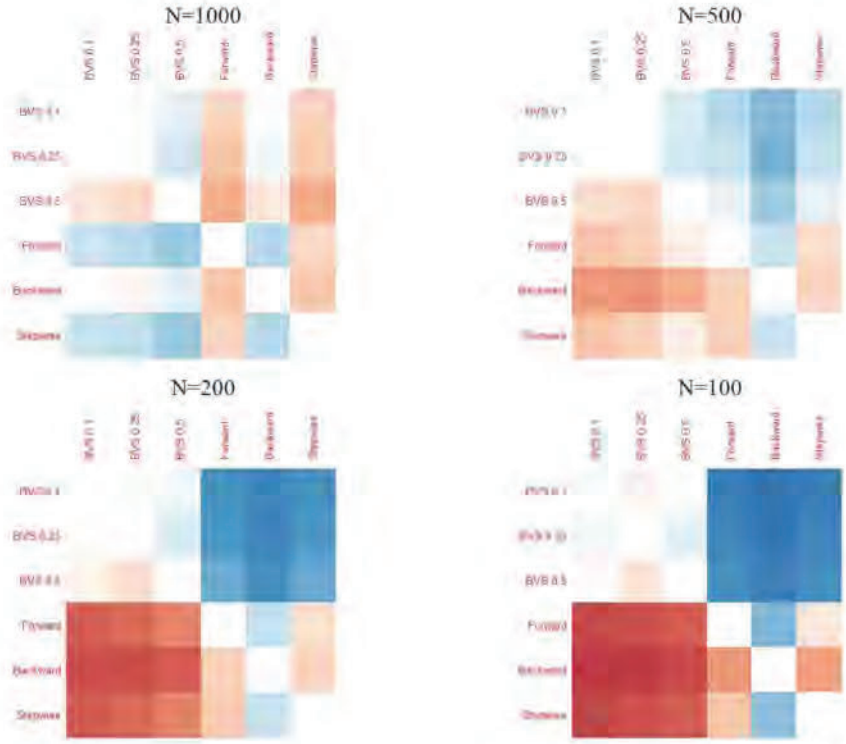


Figure 2: Continuous Simulation Results: Heat maps of method-wise bias comparisons separated by sample size, where a blue shade represents superiority of the row method in a higher proportion of simulations, and a red shade represents inferiority of the row method in a higher proportion of simulations than the column method.

4.2 Investigating ATE Bias Under Various Scenario Settings

The results for $n = 1000$ separated by various scenario types are shown in figure 4. The top left plot contains the average ATE bias and error bars for each method separated by rarity of Y. The rarity of Y is calculated as the smaller value between the mean of Y or the mean of 1-Y, averaged across all simulated datasets. The top right plot contains average ATE bias and error bars for each method separated by rarity of E, which is calculated in the same way as rarity of Y. In this sample size all simulations ran without the presence of separation.

The bottom left plot contains ATE bias and error bars for each method separated by number of non-zero Y coefficients. The number of non-zero Y coefficients is calculated as $\sum_{k=1}^{p} I[\beta_k \neq 0]$. The bottom right plot contains average ATE bias and error bars for each method separated by non-zero E coefficients, which is calculated as $\sum_{k=1}^{p} I[\alpha_k \neq 0]$.

In the rarity of Y plot, we can see the methods performed fairly similarly. When Y is very rare, the methods are all fairly close in terms of average ATE bias, all staying around 0.021, however there is slightly more variability in the Bayesian methods than in the frequentist methods. As Y becomes less and less rare, both the average ATE bias and variability for all methods become closer together, with the highest Bayesian method being BVS 0.5 at 0.0241, and the highest frequentist method being backward at 0.0235. In the rarity of E plot, we once again see high variability in the rare simulations. In this case though the frequentist methods have higher variability than the Bayesian methods, however the average ATE bias is still similar, with all methods around 0.038. Once again as E becomes less and less rare, the average bias and variability across methods comes closer together, converging to about 0.022.

In the non-zero Y coefficients plot, we see a less obvious trend over all the methods. In each section the average ATE bias between methods stays relatively close, always staying in between 0.02 and 0.03, however within the Bayesian methods we see that the average ATE bias tends to increase as the prior probability of inclusion increases. This trend may be due to more overfitting, as a higher π_η means that we include more variables a priori. In addition we see the stepwise and forward methods performing similarly, having lower ATE bias than the backward method. Once again this may be due to overfitting, since the backward method is more likely to include variables given it begins with a full model. These trends can be seen in all bins except when there are 4+ non-zero coefficients, where all methods perform more or less the same at around 0.021. In the non-zero E coefficients plot, the variability of the methods stays relatively consistent across all bins, but we see an overall slight increase in ATE bias as the number of non-zero E coefficients increases, from around 0.019 in the 0 bin, to an average of 0.023 in the 4+ bin.

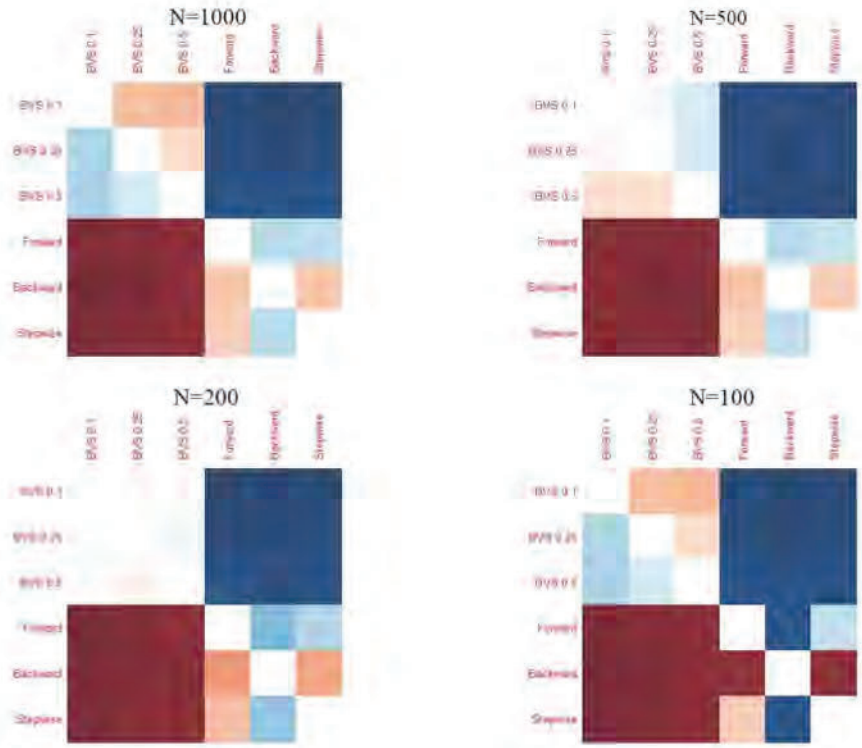


Figure 3: Continuous Simulation Results: Heat maps of method-wise average coverage ratio comparisons separated by sample size, where a blue shade represents superiority of the row method in a higher proportion of simulations, and a red shade represents inferiority of the row method in a higher proportion of simulations than the column method.

The results for $n = 500$ separated by various scenario types are shown in supplemental figure 4. Overall in the plots we see similar trends to the previous set of results, it should also be noted that the average ATE biases across all methods and bins were generally higher than in the results for $n = 1000$. In the Y rarity plot, we see when Y is more rare the variability and ATE bias in the Bayesian methods (average of 0.0268) is greater than in the frequentist methods (average of 0.0253), and as Y becomes less rare we see a slight increase in ATE bias across all methods, with all converging to around 0.033. In the E rarity plot we see when E is very rare both sets of methods tend to have higher ATE bias and variability, with the frequentist methods (average of 0.056) being noticeably higher than the Bayesian methods (average of 0.048) in bias. As E becomes less rare the bias and variability for all methods quickly decreases and the difference between methods becomes very small, eventually settling at around 0.03 for all methods. This trend when E is rare may be due to the fact that when there are less patients in the exposure group, there are less data points with which to assess the exposure's effectiveness, leading to more variability and higher bias.

In the non-zero Y coefficients plot we again see that in general the ATE bias tends to increase as prior probability of inclusion increases in the Bayesian methods, increasing from 0.034 in BVS 0.1 to 0.038 in BVS 0.5. We see this across bins except the 4+ bin where the results even out to an average of 0.030 in the Bayesian methods and an average of 0.031 in the frequentist methods. We also again see that the stepwise and forward methods perform similarly in all bins, while the backward method tends to have higher ATE bias. Once again in the bin with 4+ non-zero coefficients all methods even out. In the non-zero E coefficients plot, the variability between method groups tended to be slightly higher in the Bayesian methods, while the average ATE bias tended to be fairly close. Overall across methods we see an increase in average ATE bias as the number of non-zero coefficients increases, going from 0.028 to 0.033 in the Bayesian methods, and 0.028 to 0.034 in the frequentist methods.

The results for $n = 200$ separated by various scenario types are shown in supplemental figure 5, where 3 simulations were flagged for presence of separation and were removed from consideration. In the Y rarity plot we see a noticeable increase in variability from the Bayesian methods in the previous results when Y is more rare. In addition, as Y becomes less and less rare the ATE bias fluctuates slightly in the Bayesian methods and increases in the frequentist methods, leading to an overall higher ATE bias in the frequentist methods in the last bin. In the E rarity plot we see a different trend than previously with the Bayesian methods. We see a slight decrease in ATE bias in the Bayesian methods as E becomes less and less rare, rater than the sharp decrease from the very rare bin we saw previously. In the frequentist methods we see the same sharp decrease

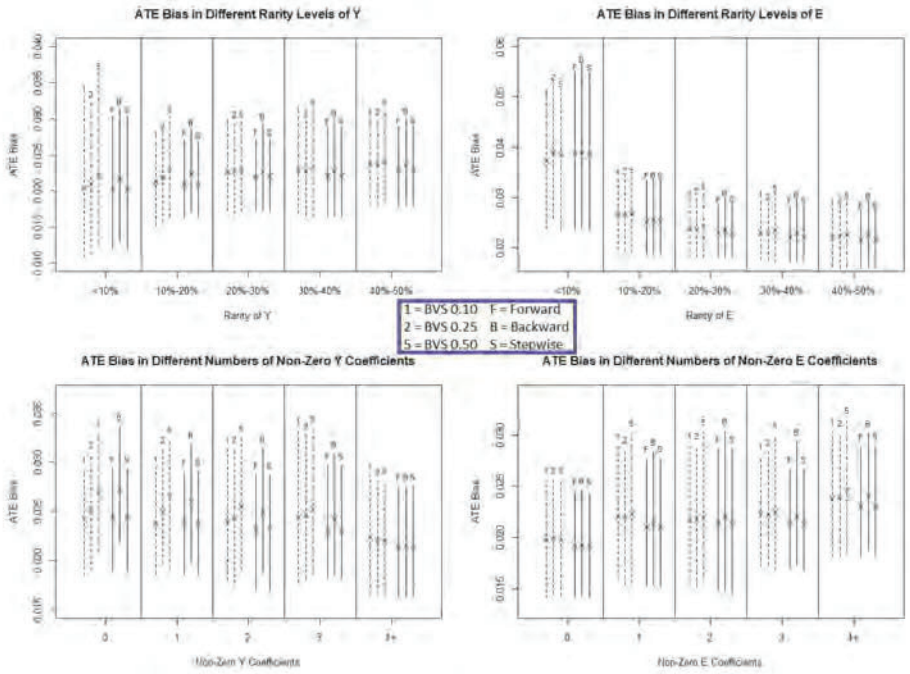


Figure 4: Continuous Simulation Results: Plots of ATE bias and $10^{th}/90^{th}$ quantile bars separated by rarity of Y, rarity of E, non-zero Y coefficients, and non-zero E coefficients for $n = 1000$. Here rarity is the minimum of $\bar{Y}, 1-\bar{Y}$ for Y rarity, and $\bar{E}, 1-\bar{E}$ for E rarity.

from the very rare bin, but overall the Bayesian methods have a lower ATE bias than the frequentist methods.

In the non-zero Y coefficients plot, we see a similar trend to previous results. With the exception of the 4+ bin the ATE bias increases as the prior probability of inclusion increases in the Bayesian methods, but it's generally still smaller than that of the frequentist methods. In addition, once again the backward method had higher ATE bias than both the forward and stepwise methods in all bins except 4+ where the methods performed about the same. In the non-zero E coefficients plot, we see fluctuations in the ATE bias of the Bayesian methods, and a steady increase in ATE bias of the frequentist methods, leading to the ATE bias being lower in the Bayesian methods in every bin.

The results for $n = 100$ separated by various scenario types are shown in supplemental figure 6, where 11 simulations were flagged for the presence of separation and removed. In the Y rarity plot we see another noticeable increase in variability in the Bayesian methods, most notably in the more rare bins. In the less rare bins, the ATE bias of the Bayesian methods decreases and eventually plateaus, while the ATE bias in the frequentist methods increases, never dipping back below the Bayesian methods as they did in the rarest bin. In the E rarity plot we once again only see slight fluctuation in the ATE bias of the Bayesian methods, while we see a steady decline in that of the frequentist methods. In addition, the Bayesian methods have a lower ATE bias than the frequentist methods in every rarity bin.

In the non-zero Y coefficients plot we see a steady decrease in the ATE bias of the Bayesian methods. The ATE bias in the frequentist methods also very slightly decrease in ATE bias as the number of non-zero coefficients increases, however overall they are higher than the Bayesian methods. Once again we see that the backwards method has higher ATE bias than the forward and stepwise methods across all bins. In the non-zero E coefficients plot we again see a slight fluctuation in the ATE bias of the Bayesian methods, staying around 0.06 in all bins. In the frequentist methods we see a steady increase, leading to higher ATE bias than the Bayesian methods in all of the bins.

4.3 Results for Binary Confounding Variables

In addition to continuous simulations, we also ran Binary confounder simulations, where each covariate was a random binary vector. For these simulations, in $n = 1000$ 1 simulation was flagged for separation, in $n = 500$, 12 simulations were flagged for separation. In $n = 200$, 41 simulations were flagged for separation, and in $n = 100$, 153 simulations were flagged for separation. We removed these scenarios rather than regenerate them. For the binary simulations, we present figure 5 which is similar to the one in the overall section, looking at ATE bias, coverage probability, coverage length, and coverage ratio separated by sample size. In addition to this, other sets of results similar to those in the continuous simulations are included in this thesis as a supplement.

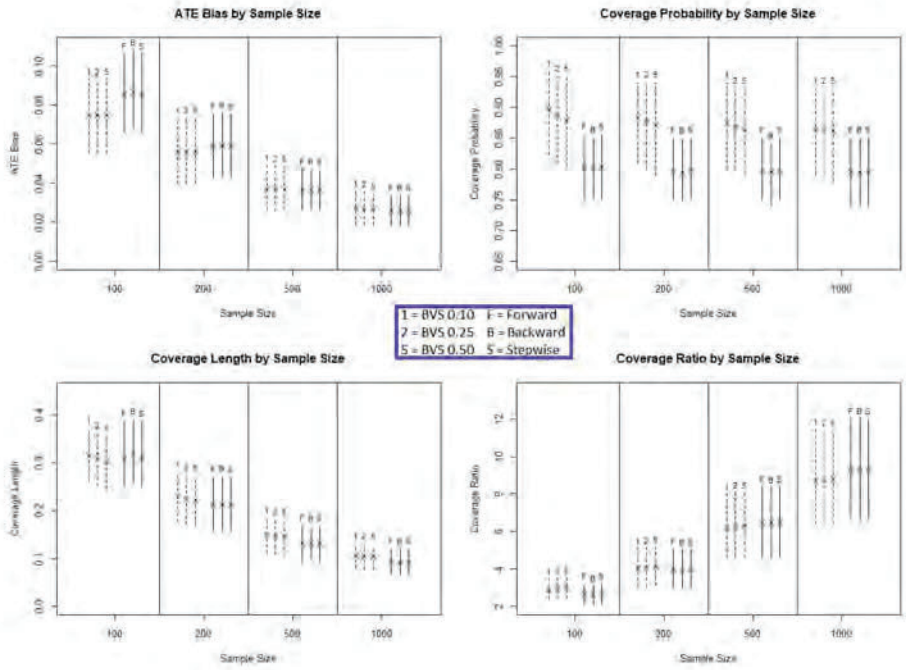


Figure 5: Plots displaying average ATE bias, coverage probability, coverage length, and coverage ratio in binary simulations for each method separated by sample size. 1 represents BVS 0.1, 2 represents BVS 0.25, and 5 represents BVS 0.5

In ATE bias, we see that the Bayesian methods outperform the frequentist methods in the smaller sample sizes, and as the sample sizes increase the methods all perform around the same. We can see this same trend in the continuous data results, however there is less variability in the binary results. In coverage probability we see the Bayesian methods outperform the frequentist methods consistently across sample sizes. This is directly opposite of the continuous results where the frequentist methods were superior in every sample size.

In coverage length, the methods perform similarly within each sample size, with an overall downward trend as sample size increases. This is in contrast to the continuous results, where the Bayesian methods outperformed the frequentist methods in every sample size category. In coverage ratio again all methods performed fairly similarly, with the frequentist methods ending slightly higher than the Bayesian methods as sample size increased. We also see an overall increasing trend as sample size increases. This is again in contrast to the continuous results, where the Bayesian methods outperformed the frequentist methods in every bin. The rest of the binary simulation results can be found in the supplemental materials in supplemental figures 8 through 14 and are largely consistent with continuous results. Tables 5 through 8 display the numerical summaries of the method performances for these three metrics.

5 Application

The methods discussed will be applied to 3 different studies. The first of these studies is large in size (n=11,660) and seeks to analyze the effect of negative pressure wound therapy (NPWT) on surgical site infection risk in c-section patients with obesity. The second study is a medium sized study, (n=307) and seeks to analyze the effect of antibiotic use on infectious complications in high-risk patients with non-operative mid-facial trauma. The final study is small in size (n=86) and seeks to analyze the effect of patient-specific implants on complications of orbital reconstruction.

5.1 Negative Pressure Wound Therapy

The first study we will apply these methods to is a retrospective cohort study for patients with obesity and negative pressure wound therapy or abdominal dressing after cesarean delivery between April 1, 2014 and January 31, 2018. In this study our binary exposure variable is whether or not a patient received negative pressure wound therapy, a method of healing wounds that involves using suction to remove excess fluid and debris, promote blood flow, and help close the wound [28]. This method's efficacy is being investigated as it relates to post-operative infection in cesarean delivery patients with obesity.

Each variable selection method discussed was used to analyze provided data and return the optimal model. The results

are shown in the following table, which contains whether or not each considered variable was included in the optimal model returned by each frequentist method, as well as the posterior probability of inclusion of each variable for the Bayesian methods.

Variable	F	B	S	BVS 0.1	BVS 0.25	BVS 0.5
ATE	.016	.012	.015	.020	.018	.018
Inclusion Probabilities						
Suture vs Staples	0	0	0	0.20	0.32	0.61
Black vs White Race	0	1	0	0.08	0.38	0.54
Diabetes	0	0	0	0.06	0.24	0.49
Private Insurance	1	1	1	0.42	0.78	0.99
Scheduled Delivery	1	1	1	0.97	0.98	1.00
Artificial Rupture	0	0	0	0.23	0.28	0.65
Previous C Section	0	0	0	0.10	0.25	0.61
BMI 35-40 vs BMI 30-35	0	0	0	0.09	0.37	0.58
BMI 40+ vs BMI 30-35	1	1	1	0.87	0.96	0.93
Cut to Close Min.	1	1	1	1.00	1.00	1.00
Maternal Age	0	0	0	0.01	0.04	0.05
Delivery Year	0	1	0	0.00	0.00	0.00

Table 1: Average treatment effects and inclusion probability of each variable in the NPWT application. F, B, and S stand for Forward, Backward, and Stepwise respectively For frequentist methods, variables are either included (probability=1) or excluded (probability=0) from the final selected models.

The *ATE* for the Bayesian methods were close to the freuqntist methods, but more positive by an average of .004. Here ATE decreased as π_η increased (favoring more variables in the model) and when removal of variables was allowed in frequentist selection (also favoring more variables in the model). As we can see the frequentist methods performed fairly similarly. Private insurance, scheduled delivery, 40+ BMI vs 30-35 BMI, and minutes from cut to close were included in the resulting models from every method, while race and delivery year were included only in the backward model. In the Bayesian methods we see mostly the same trends, where the variables included in the frequentist methods all had much higher posterior probabilities of inclusion than those not included, though it should be noted that even in the variables not included in any of the frequentist methods had relatively high posterior probabilities of inclusion (around 50%-65%) for BVS 0.5, which we might expect given its high prior probability of inclusion.

5.2 Antibiotics in Non-Operative Mid-Facial Trauma

The next study we will apply these methods to discusses the use of antibiotics in preventing infection for non-operative mid-facial trauma [29]. For context, previous studies show that antibiotics are generally not needed for non-operative mid-facial fractures. However, patients who are critically-injured with mid-facial fractures may be at higher risk of infectious complications, such as pneumonia or sinusitis. In this study our binary exposure variable is whether or not the patient used antibiotics, which are generally not needed for non-operative mid-facial fractures, however this may not be the case in patients who are at a higher risk of developing certain infections that may be exacerbated by facial features such as sinusitis or pneumonia. The antibiotics' efficacy in preventing these infections is what is being studied.

The results from analyzing these data using the methods discussed in this thesis are presented in the same manner as in the previous section.

The *ATE* for the Bayesian methods were on average .021 larger than those for the frequentist methods. *ATE* was decreased in backwards selection and as π_η increased, both of which led to more variables included. We see the frequentist methods all returned the same optimal model. The variables included were race, ED.GCS, total PRBC, and hospital length of stay in days. In this case the Bayesian methods agreed very closely with the frequentist methods as well, as we can see a large jump in posterior probability of inclusion from variables not included in the frequentist models to those that were. Again, even in the non-included variables we do see an abnormally high posterior probability of inclusion in BVS 0.5 in some instances.

5.3 Patient-Specific Implants

The last study we will apply these methods to discusses patient-specific titanium implants for use in treating orbital fractures. The key goal in orbital reconstruction is to restore orbital volume to premorbid dimensions, which can cause problems such as diplopia or orbital dystopia if not done correctly. There are many methods of orbital reconstruction used, most of them centered on pre-formed titanium mesh implants. These methods are effective, but it is challenging to contour and apply the

Variable	F	B	S	BVS 0.1	BVS 0.25	BVS 0.5
ATE	-.098	-.100	-.098	-.079	-.076	-.077
Inclusion Probabilities						
Male Gender	0	0	0	0.24	0.49	0.80
African American Race	1	1	1	0.68	0.92	0.90
Lacerations Present	0	0	0	0.14	0.40	0.60
Foreign Bodies	0	0	0	0.23	0.35	0.52
Age	0	0	0	0.00	0.03	0.12
Glascow Coma Scale	1	1	1	1.00	1.00	0.90
Injury Severity Score	0	0	0	0.04	0.06	0.18
Blood Units	1	1	1	0.54	0.86	1.00
Hospital Length of Stay	1	1	1	1.00	1.00	1.00

Table 2: Average treatment effects and inclusion probability of each variable in the trauma and antibiotics application. F, B, and S stand for Forward, Backward, and Stepwise respectively For frequentist methods, variables are either included (probability=1) or excluded (probability=0) from the final selected models.

plates with the limited surgical visibility necessitated by the anatomy of this part of the body. The study compared the complication rate (our binary outcome) and reconstruction accuracy between patients with pre-formed and patients-specific titanium implants [30]. The purpose of this study is to compare the rate of complications and reconstruction accuracy between patients with pre-formed and patient-specific titanium implants.

The data from this study will be analyzed in the same way as in the previous sections. What follows is a table of the results from the analysis.

Variable	F	B	S	BVS 0.1	BVS 0.25	BVS 0.5
ATE	-.179	-.180	-.179	-.111	-.128	-.131
Inclusion Probabilities						
Male Gender	0	0	0	0.10	0.35	0.60
ASA Score	0	1	0	0.20	0.27	0.55
Diplopia/Dystopia	0	0	0	0.15	0.35	0.50
# Surfaces	0	0	0	0.19	0.34	0.61
Prolapsed Volume	0	1	0	0.12	0.21	0.43
Defect Size	0	0	0	0.12	0.34	0.52

Table 3: Average treatment effects and inclusion probability of each variable in the trauma and patient-specific implants application. F, B, and S stand for Forward, Backward, and Stepwise respectively For frequentist methods, variables are either included (probability=1) or excluded (probability=0) from the final selected models.

The *ATE* for the Bayesian methods were on average .056 larger than those for the frequentist methods. *ATE* was decreasted in backwards selection and as π_η increased, both of which led to more variables included. We see here both the forward and backwards methods did not select any variables to be included, and the backward method selected only ASA score and prolapsed volume. We see the same trend in the Bayesian methods, as most of the variables had posterior probabilities of inclusion on average around 15% for BVS 0.1 and 28% for BVS 0.25. As previously discussed, because of the high prior, these probabilities were fairly high relative to the other methods in BVS 0.5. One reason for these odd results may be that this was the smallest study analyzed at only 86 patients, and the methods are not as reliable with such small sample sizes.

6 Discussion

In this paper we studied how deviance based confounding variable selection (forward, backwards, stepwise) and *spike-and-slab* Bayesian variable selection (with prior inclusion probabilities of .10, .25, .50) performed in terms of estimating average treatment effects (ATE). To do this we simulated 1000 different sets of parameters from which to generate a "true" ATE value, and 100 random data sets. We then used six methods of variable selection on each data set, forward, backward, stepwise, and Bayesian variable selection with prior probabilities of inclusion of 10%, 25%, and 50%, to choose an optimal model. We developed an R package called `VARSELECTEXPOSURE` containing multiple specialized functions to perform these methods. We then used that model to calculate an ATE estimate and store several features of the data as well as four measures to assess

the effectiveness of each method, ATE bias, coverage probability, coverage length, and coverage ratio. We repeat this process for data sets of 1000, 500, 200, and 100 subjects.

From these results, we see that in most settings the BVS methods perform similar to or better than the frequentist methods. Overall, the BVS methods perform better in ATE bias in the 100 and 200 subject simulations, and all methods performed about the same in the 500 and 100 subject simulations. In coverage probability the frequentist methods outperformed the Bayesian methods, averaging about 95% while the Bayesian methods averaged about 85% across the different sample sizes. In coverage length the Bayesian methods performed better in every sample size, having a lower average coverage length and about the same amount of variability. Finally in coverage ratio the Bayesian methods once again performed better than the frequentist methods, having a higher coverage ratio across sample sizes. Results largely held for binary confounding variables but did not hold in rare outcomes and small sample sizes (¡10% occurance).

We also applied these methods to three studies of varying sizes. Posterior mean estimates of the ATE were closer in larger sample sizes, but only differed by about .5 % on average in the smallest sample size of around 100 patients. This difference is due to the differential inclusion weights of the confounding variables in the two approaches. Bayesian methods include variables in some proportion of posterior samples (weights between 0 and 1), while the frequentist methods either include a variable or not (weights of 1 or 0). We highlighted differences in these weights in each application, showing how they affected ATE estimates for varying prior probabilities of inclusion and variable selection approaches.

Many different scenarios were considered for this study, but there were some things that were not explored, such as the number of covariates considered. For this study, only data sets with at most 10 covariates were analyzed, so we cannot confidently say that our results would hold for very large data sets for example ones with 25+ covariates. In addition, specifically in the Bayesian methods, we only tested the methods with our three prior probabilities of inclusion, so the results may differ with different values, for example 60% or 5%. We also used 2000 MCMC iterations, burning in the first 1000. Increasing or decreasing this may change results slightly as well.

References

[1] M. Z. I. Chowdhury and T. C. Turin, "Variable selection strategies and its importance in clinical prediction modelling," *Family medicine and community health*, vol. 8, no. 1, 2020.

[2] G. Claeskens, C. Croux, and J. Van Kerckhoven, "Variable selection for logistic regression using a prediction-focused information criterion," *Biometrics*, vol. 62, no. 4, pp. 972–979, 2006.

[3] M. Kuhn, K. Johnson, *et al.*, *Applied predictive modeling*, vol. 26. Springer, 2013.

[4] M. A. Babyak, "What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models," *Psychosomatic medicine*, vol. 66, no. 3, pp. 411–421, 2004.

[5] S. Bleeker, G. Derksen-Lubsen, D. Grobbee, A. Donders, K. Moons, and H. Moll, "Validating and updating a prediction rule for serious bacterial infection in patients with fever without source," *Acta paediatrica*, vol. 96, no. 1, pp. 100–104, 2007.

[6] J. Geweke, "Variable selection and model comparison in regression," *Bayesian Statistics*, vol. 5, pp. 609–620, 1996.

[7] E. I. George and R. E. McCulloch, "Approaches for bayesian variable selection," *Statistica sinica*, pp. 339–373, 1997.

[8] P. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate behavioral research*, vol. 46, no. 3, p. 399–424, 2011.

[9] J. M. Robins, A. Rotnitzky, and L. P. Zhao, "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, vol. 89, no. 427, pp. 846–866, 1994.

[10] J. Pearl, "On the interpretation of do(x)," *Journal of Causal Inference*, vol. 7, no. 1, p. 20192002, 2019.

[11] Z. Zhang, M. Uddin, J. Cheng, and T. Huang, "Instrumental variable analysis in the presence of unmeasured confounding.," *Annals of translational medicine*, no. 10, 2018.

[12] F. Elwert and C. Winship, "Endogenous selection bias: The problem of conditioning on a collider variable," *Annual Review of Sociology*, vol. 40, no. 1, pp. 31–53, 2014. PMID: 30111904.

[13] R. Moret and A. Chapple, "Analysis of the effects of adjusting for binary non-confounders in a logistic regression model after all true confounders have been accounted for: A simulation study," *Economics and Econometrics*, no. EERI RP 2022/05, 2022.

[14] Z. Ye, Y. Zhu, and D. Coffman, "Variable selection for causal mediation analysis using lasso-based methods," *Statistical Methods in Medical Research*, vol. 30, no. 6, pp. 1413–1427, 2021. PMID: 33755518.

[15] C. Rosenbaum, Q. Yu, S. Buzhardt, E. Sutton, and A. G. Chapple, "Inclusion of binary proxy variables in logistic regression improves treatment effect estimation in observational studies in the presence of binary unmeasured confounding variables," *Pharmaceutical Statistics*, vol. n/a, no. n/a.

[16] P. W. Holland, "Statistics and causal inference," *Journal of the American statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.

[17] M. A. Hernán, "A definition of causal effect for epidemiological research," *Journal of Epidemiology & Community Health*, vol. 58, no. 4, pp. 265–271, 2004.

[18] B. Ratner, "Variable selection methods in regression: Ignorable problem, outing notable solution," *Journal of Targeting, Measurement and Analysis for Marketing*, vol. 18, no. 1, pp. 65–75, 2010.

[19] E. Steyerberg, "Clinical prediction models: a practical approach to development, validation, and updating: Springer science & business media," *New York*, 2008.

[20] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, vol. 398. John Wiley & Sons, 2013.

[21] H. Ishwaran and J. Rao, "Spike and slab variable selection: Frequentis and bayesian strategies," *Annals of Statistics*, vol. 33, no. 2, pp. 730–773, 2005.

[22] A. Chapple, M. Vannucci, P. Thall, and S. Lin, "Bayesian variable selection for a semi-competing risks model with three hazard functions," *Computational Statistics and Data Analysis*, vol. 112, pp. 170–185, 2017.

[23] R. B. O'Hara and M. J. Sillanpää, "A review of bayesian variable selection methods: what, how and which," *Bayesian analysis*, vol. 4, no. 1, pp. 85–117, 2009.

[24] C. Rainey, "Dealing with separation in logistic regression models," *Political Analysis*, vol. 24, no. 3, pp. 339–355, 2016.

[25] S. Devika, L. Jeyaseelan, and G. Sebastian, "Analysis of sparse data in logistic regression in medical research: A newer approach," *Journal of postgraduate medicine*, vol. 62, no. 1, p. 26, 2016.

[26] R. Pordes *et al.*, "The open science grid," in *J. Phys. Conf. Ser*, vol. 78, p. 012057, 2007.

[27] I. Sfiligoi, D. C. Bradley, B. Holzman, P. Mhashilkar, S. Padhi, and F. Wurthwein, "The pilot way to grid resources using glideinwms," in *2009 WRI World congress on computer science and information engineering*, vol. 2, pp. 428–432, IEEE, 2009.

[28] C. Huang, T. Leavitt, L. R. Bayer, and D. P. Orgill, "Effect of negative pressure wound therapy on wound healing," *Current problems in surgery*, vol. 51, no. 7, pp. 301–331, 2014.

[29] D. Hyneman, J. Coburn, L. Bellfi, A. G. Chapple, and B. J. Christensen, "Do antibiotics prevent infectious complications in critically injured patients with blunt non-operative mid-facial trauma?," *Journal of Oral and Maxillofacial Surgery*, 2023.

[30] T. Gander, H. Essig, P. Metzler, D. Lindhorst, L. Dubois, M. Rücker, and P. Schumann, "Patient specific implants (psi) in reconstruction of orbital floor and wall fractures," *Journal of Cranio-Maxillofacial Surgery*, vol. 43, no. 1, pp. 126–130, 2015.

Supplement for: A Comparison of Bayesian and Frequentist Variable Selection Methods for Estimating Average Treatment Effects in Logistic Regression

Alex H. Martinez, Brian Christiansen, Elizabeth F. Sutton, Andrew G. Chapple

September 28, 2023

This supplemental material contains additional tables and figures referenced in the primary manuscript. It also contains a user manual for how to use the R package *V*ARSELECTEXPOSURE .

1 User Manual for *VARSELECTEXPOSURE* Package in R

For this research we built an R package called `VARSELECTEXPOSURE` [**?**]. This package overall executes the methods described in this thesis, but we discuss them in detail below:

- `FORWARD_EXPOSURE`: Returns the estimated Average Treatment Effect calculated by the optimal model chosen via forward selection including an exposure variable, as well as the optimal model itself. The function begins with an empty model and iteratively chooses a candidate variable for addition based on the variable whose addition results in the model with the lowest deviance, relative to the deviance resulting from addition of the other variables. As an argument the function only needs a data frame containing the binary outcome variable Y, the binary exposure variable E, and the candidate covariates.

- `BACKWARD_EXPOSURE`: Returns the estimated Average Treatment Effect calculated by the optimal model chosen via backward selection including an exposure variable, as well as the optimal model itself. The function begins with a full model and iteratively chooses a candidate variable for deletion based on the variable whose deletion results in the model with the lowest deviance, relative to the deviance resulting from deletion of the other variables. As an argument the function only needs a data frame containing the binary outcome variable Y, the

binary exposure variable E, and the candidate covariates.

- `STEPWISE_EXPOSURE`: Returns the estimated Average Treatment Effect calculated by the optimal model chosen via backward selection including an exposure variable, as well as the optimal model itself. The function begins with an empty model and performs an addition step similar to forward selection, it then performs a deletion step similar to backward selection, and continues iteratively until a variable is not added/deleted in two consecutive steps. As an argument the function only needs a data frame containing the binary outcome variable Y, the binary exposure variable E, and the candidate covariates.

- `MCMC_LOGIT_KEEP`: Returns the posterior distributions of the Average Treatment, covariate parameters, and $\boldsymbol{\eta}$, a vector of binary values indicating whether or not a certain variable was deleted or included for a given MCMC iteration. The function performs many iterations of Metropolis-Hastings sampling of model covariate parameters, and stores them for posterior analysis. As arguments the function takes a vector of the binary outcome Y, a matrix of covariate data Z, a prior probability of inclusion `PIN`, the maximum number of covariates desired in the model `MAX_COV`, the prior standard deviation for the parameters `SdBeta`, and the desired number of MCMC simulations `NUM_REPS`.

1.0.1 `FORWARD_EXPOSURE`

The `FORWARD_EXPOSURE` function is used to perform forward variable selection on a provided data set containing a binary outcome and binary exposure variable. It is used in the simulation trial to provide estimated ATE values used for comparison to the truth. Below is an example case using a data set simulated using the same methods as in *Chapter 2* with 7 covariates and $n = 750$.

```
head(testdata)
Y E  X1     X2     X3     X4     X5     X6     X7
1 0  1.507 -1.078 -1.000 -0.144  0.168  0.873 -0.035
1 0  0.064  0.024  0.715  2.237 -0.519 -0.052 -0.087
1 0 -0.612  0.274  0.305  0.443  0.147  1.271 -0.967
0 0  1.832  0.102 -1.083  1.034  0.540  0.939  1.512
1 0 -1.382 -1.078 -0.028 -1.009  0.409 -1.059  0.683
1 1 -0.322  2.517 -0.616 -0.723  1.098 -0.024 -1.828
```

As we can see the example `testdata` set has our binary outcome variable, binary exposure variable, and covariate data. The data set used for these functions must be of this form for the method to compute properly.

```
Z = FORWARD_EXPOSURE(testdata)
```

Z contains a list with the estimated Average Treatment Effect, the first 6 lines of the optimal chosen data set, and a summary of the regression model fit using the selected data set. The following is the resulting estimated Average Treatment Effect:

```
Z$ATE
[1] -0.03941409
```

Here we see the function calculated an estimated Average Treatment Effect of -0.03941409 using the selected model. This can be interpreted as a 3.94% decrease in probability of success from the control group to the treatment group. Next we see the chosen raw data set:

```
Z$DATA
 Y E          X1         X3          X6         X5          X7
 1 0  1.50650896 -1.0004223  0.87348380  0.1678093 -0.03529093
 1 0  0.06389924  0.7145363 -0.05150836 -0.5194347 -0.08717419
 1 0 -0.61233229  0.3051169  1.27046179  0.1471311 -0.96670048
 0 0  1.83244596 -1.0828616  0.93857902  0.5396228  1.51216611
 1 0 -1.38211077 -0.0271911 -1.05900546  0.4090515  0.68254848
 1 1 -0.32229316 -0.6156629 -0.02425986  1.0982293 -1.82762380
```

We can see we have our binary outcome and exposure variables, as well as the chosen covariates. The function chose a model containing the 1^{st}, 3^{rd}, 6^{th}, 5^{th}, and 7^{th} covariates. This model was then fit on these data:

```
Z$MOD

Call:
glm(formula = Y ~ ., family = "binomial", data = FORWARD.DATA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5185  -0.7445   0.3156   0.7188   3.0450

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.9065     0.1493   6.074 1.25e-09 ***
E            -0.2685     0.2154  -1.247    0.213
X1           -1.4372     0.1292 -11.128  < 2e-16 ***
X3           -0.8720     0.1113  -7.832 4.80e-15 ***
X6            0.5726     0.1032   5.547 2.91e-08 ***
X5           -0.5743     0.1056  -5.440 5.32e-08 ***
X7           -0.5832     0.1083  -5.384 7.30e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 990.02  on 749  degrees of freedom
Residual deviance: 683.49  on 743  degrees of freedom
AIC: 697.49

Number of Fisher Scoring iterations: 5
```

Here we once again see that the method chose a model containing the 1^{st}, 3^{rd}, 6^{th}, 5^{th}, and 7^{th}. The output for the other frequentist methods included is fairly similar.

1.0.2 BACKWARD_EXPOSURE

The BACKWARD_EXPOSURE function is used to perform backward variable selection on a provided data set containing a binary outcome and binary exposure variable. It is used in the simulation trial to provide estimated ATE values used for comparison to the truth. Below is an example case using the same data set used in the example for FORWARD_EXPOSURE.

```
Z = BACKWARD_EXPOSURE(testdata)
```

Similar to the previous example, Z contains a list with the estimated Average Treatment Effect, the first 6 lines of the optimal chosen data set, and a summary of the model fit using using the selected data set. The estimated Average Treatment Effect was:

```
Z$ATE
[1] -0.03941409
```

Here we see the backward method calculated an estimated Average Treatment Effect of -0.03941409. This can be interpreted as a 3.94% decrease in probability of success from the control group to the treatment group. This is the same as the forward method because they chose the same model, which can happen with small numbers of covariates. As such, the output of the BACKWARD_EXPOSURE function is similar to the previous output, containing the estimated Average Treatment Effect, the first 6 lines of the chosen data set, and a summary of the resulting model.

1.0.3 STEPWISE_EXPOSURE

Again similar to the previous examples, the STEPWISE_EXPOSURE function is used to perform stepwise variable selection on a provided data set containing a binary outcome and a binary exposure variable. It's used in the simulation trial to provide estimated ATE values used for comparison to the truth. Below is an example case using the same data set used in the previous examples.

```
Z = STEPWISE_EXPOSURE(testdata)
```

As in the previous examples, Z contains a list with the estimated Average Treatment Effect, the first 6 lines of the optimal chosen data set, and a summary of the model fit using using the selected data set. The estimated Average Treatment Effect calculated was:

```
Z$ATE
[1] -0.03941409
```

Here we see the backward method calculated an estimated Average Treatment Effect of -0.03941409. This can be interpreted as a 3.94% decrease in probability of success from the control group to the treatment group. The output of the stepwise function takes the same form as in the previous functions.

1.0.4 MCMC_LOGIT_KEEP

The MCMC_LOGIT_KEEP function performs a series of MCMC iterations to estimate the covariate parameters for a provided data set and returns their posterior distributions. Below is a brief explanation of the function arguments and the example values used.

- Y: Vector of binary outcome variable. For this example we will use the same randomly generated outcome as in the previous examples.
- Z: Matrix of covariate data. Again for this example we will use the same randomly generated data set.
- PIN: The prior probability of inclusion for each parameter. In this example we will use $PIN = 0.1$.
- MAX_COV: The maximum number of covariates desired. In this example we will use MAX_COV = 7.
- SdBeta: The prior standard deviation for the parameters. The prior distribution for the intercept parameter β_0 is assumed to be flat, so this will only be used for the covariate parameters. For this example we will use SdBeta = 1.
- NUM_REPS: Number of MCMC iterations to run. Because this is adaptive MCMC, we burn-in the first half of the iterations, so the output matrices will have half as many rows and the resulting vectors will be half as long as is specified in this argument. For this example we will use NUM_REPS = 2000.

This function performs differently to the functions for frequentist methods, as Bayesian Variable Selection approaches variable selection by estimating parameters rather than searching for the single optimal model. Below is an example case using the parameters previously specified.

```
out = MCMC_LOGIT_KEEP(Y, Z, 0.1, 7, 1, 2000)
```

The list `out` contains two vectors and two matrices of data. The first vector contains the storage of Average Treatment effects calculated in each iteration. The second vector contains the posterior distribution of the intercept parameter β_0. The first matrix contains the posterior distribution of the other covariate parameters $\beta_1, \beta_2, ..., \beta_n$, and the second matrix contains the posterior distribution of η. Following is a plot of the density of estimated ATE's:

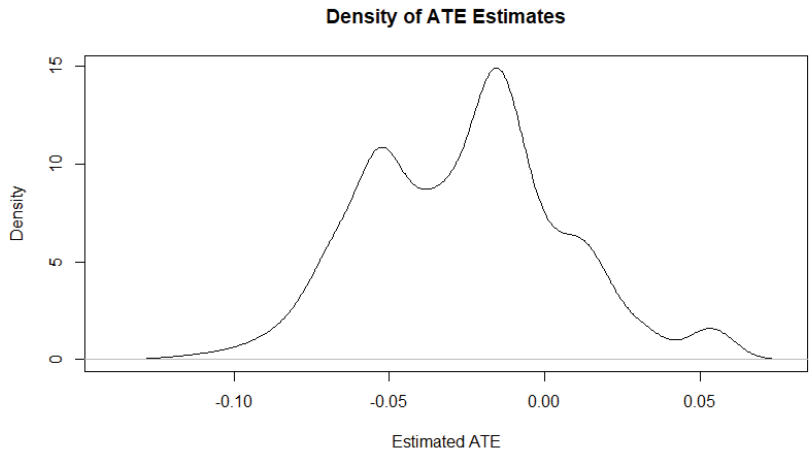


Figure 1: Density plot of Average Treatment Effects estimated by the *MCMC_LOGIT_KEEP* function

As we can see the density peaks just on the negative side of 0, if we wanted the posterior mean of the estimated ATE we could use the function output:

```
mean(out[[1]])
[1] -0.02586305
```

The posterior mean for the estimated ATE is -0.02586305, which can be interpreted as a 2.59% decrease in probability of success from the control group to the treatment group. Additionally, we can generate a credible interval for the estimated ATE by:

```
quantile(out[[1]], c(0.025, 0.975))
       2.5%       97.5%
-0.07998833  0.05054302
```

The posterior means for the parameter estimates as well as η can be obtained using a similar strategy.

```
 colMeans(out[[4]])
[1] 1.000 1.000 0.118 1.000 0.087 1.000 1.000 1.000
```

Computing the column means from the posterior distribution of η we can see the percentage of times each covariate was included. We see all but the 2^{nd} and 4^{th} variables, which were included in 11.8% and 8.7% of models respectively, were included in every model. This aligns with the results from the frequentist methods.

2 Supplemental Tables and Figures

These supplemental results are tables containing mean ATE bias and standard deviation, mean coverage probability and 10%/90% quantile, and mean coverage length with 10%/90% quantile for each method. The method-wise results are split up into three groups, one for the overall data, one for rare data (where the event occurs $<$25% of the time), and common data (where the event occurs $>$25% of the time). There is one table for each of the four sample sizes. We additionally show the results for binary confounding variables.

Method	ATE Bias Mean (SD)	Coverage Probability Mean (10%/90% Quantile)	Coverage Length Mean (10%/90% Quantile)
		Overall	
BVS 0.1	0.0230 (0.0052)	0.8515 (0.77,0.92)	0.0879 (0.07,0.11)
BVS 0.25	0.0230 (0.0052)	0.8551 (0.79,0.92)	0.0892 (0.07,0.11)
BVS 0.5	0.0235 (0.0055)	0.8546 (0.78,0.92)	0.0907 (0.07,0.12)
Forward	0.0223 (0.0048)	0.9440 (0.91,0.97)	0.1104 (0.08,0.14)
Backward	0.0230 (0.0052)	0.9415 (0.91,0.97)	0.1120 (0.08,0.15)
Stepwise	0.0223 (0.0048)	0.9443 (0.91,0.97)	0.1106 (0.08,0.14)
		Rare	
BVS 0.1	0.0221 (0.0052)	0.8536 (0.77,0.92)	0.0856 (0.06,0.11)
BVS 0.25	0.0225 (0.0054)	0.8617 (0.80,0.93)	0.0876 (0.06,0.11)
BVS 0.5	0.0232 (0.0060)	0.8569 (0.78,0.91)	0.0901 (0.06,0.12)
Forward	0.0216 (0.0046)	0.9467 (0.91,0.97)	0.1087 (0.08,0.14)
Backward	0.0226 (0.0052)	0.9426 (0.91,0.97)	0.1115 (0.08,0.15)
Stepwise	0.0215 (0.0045)	0.9473 (0.92,0.97)	0.1088 (0.08,0.14)
		Common	
BVS 0.1	0.0231 (0.0052)	0.8511 (0.77,0.92)	0.0882 (0.07,0.11)
BVS 0.25	0.0231 (0.0052)	0.8541 (0.78,0.92)	0.0895 (0.07,0.11)
BVS 0.5	0.0236 (0.0054)	0.8543 (0.78,0.92)	0.0908 (0.07,0.12)
Forward	0.0224 (0.0048)	0.9436 (0.91,0.97)	0.1107 (0.08,0.14)
Backward	0.0231 (0.0052)	0.9413 (0.91,0.97)	0.1121 (0.08,0.15)
Stepwise	0.0224 (0.0048)	0.9439 (0.91,0.97)	0.1108 (0.08,0.14)

Table 1: Results and SD/quantiles for $n = 1000$

Method	ATE Bias Mean (SD)	Coverage Probability Mean (10%/90% Quantile)	Coverage Length Mean (10%/90% Quantile)
		Overall	
BVS 0.1	0.0314 (0.0077)	0.8647 (0.79,0.93)	0.1240 (0.10,0.16)
BVS 0.25	0.0314 (0.0077)	0.8698 (0.80,0.93)	0.1254 (0.10,0.16)
BVS 0.5	0.0321 (0.0083)	0.8666 (0.79,0.93)	0.1270 (0.10,0.17)
Forward	0.0320 (0.0071)	0.9456 (0.91,0.97)	0.1583 (0.12,0.20)
Backward	0.0330 (0.0076)	0.9424 (0.91,0.97)	0.1604 (0.12,0.21)
Stepwise	0.0320 (0.0071)	0.9463 (0.91,0.97)	0.1586 (0.12,0.20)
		Rare	
BVS 0.1	0.0303 (0.0080)	0.8661 (0.79,0.93)	0.1204 (0.09,0.16)
BVS 0.25	0.0307 (0.0083)	0.8705 (0.80,0.93)	0.1230 (0.09,0.16)
BVS 0.5	0.0319 (0.0092)	0.8580 (0.78,0.93)	0.1249 (0.09,0.17)
Forward	0.0306 (0.0064)	0.9481 (0.92,0.98)	0.1552 (0.11,0.20)
Backward	0.0322 (0.0071)	0.9426 (0.91,0.97)	0.1591 (0.11,0.21)
Stepwise	0.0306 (0.0064)	0.9492 (0.92,0.98)	0.1554 (0.11,0.20)
		Common	
BVS 0.1	0.0315 (0.0077)	0.8645 (0.79,0.93)	0.1246 (0.10,0.16)
BVS 0.25	0.0315 (0.0076)	0.8697 (0.80,0.93)	0.1258 (0.10,0.16)
BVS 0.5	0.0322 (0.0081)	0.8678 (0.79,0.93)	0.1273 (0.10,0.17)
Forward	0.0322 (0.0072)	0.9452 (0.91,0.97)	0.1587 (0.12,0.20)
Backward	0.0331 (0.0077)	0.9424 (0.91,0.97)	0.1606 (0.12,0.21)
Stepwise	0.0322 (0.0072)	0.9459 (0.91,0.97)	0.1591 (0.12,0.20)

Table 2: Results and SD/quantiles for $n = 500$

Method	ATE Bias Mean (SD)	Coverage Probability Mean (10%/90% Quantile)	Coverage Length Mean (10%/90% Quantile)
		Overall	
BVS 0.1	0.0458 (0.0131)	0.8765 (0.79,0.95)	0.1862 (0.14,0.24)
BVS 0.25	0.0458 (0.0136)	0.8759 (0.79,0.95)	0.1861 (0.14,0.25)
BVS 0.5	0.0468 (0.0149)	0.8683 (0.77,0.95)	0.1860 (0.14,0.25)
Forward	0.0524 (0.0110)	0.9462 (0.91,0.98)	0.2600 (0.20,0.33)
Backward	0.0540 (0.0116)	0.9434 (0.91,0.97)	0.2657 (0.21,0.34)
Stepwise	0.0525 (0.0110)	0.9468 (0.91,0.98)	0.2605 (0.20,0.33)
		Rare	
BVS 0.1	0.0458 (0.0159)	0.8685 (0.74,0.95)	0.1817 (0.12,0.24)
BVS 0.25	0.0471 (0.0171)	0.8612 (0.75,0.95)	0.1825 (0.12,0.24)
BVS 0.5	0.0488 (0.0189)	0.8531 (0.72,0.95)	0.1839 (0.12,0.25)
Forward	0.0497 (0.0098)	0.9506 (0.92,0.98)	0.2543 (0.20,0.32)
Backward	0.0527 (0.0115)	0.9471 (0.92,0.98)	0.2647 (0.21,0.33)
Stepwise	0.0498 (0.0098)	0.9511 (0.92,0.98)	0.2548 (0.20,0.32)
		Common	
BVS 0.1	0.0458 (0.0126)	0.8776 (0.79,0.95)	0.1869 (0.14,0.25)
BVS 0.25	0.0456 (0.0130)	0.8781 (0.79,0.95)	0.1866 (0.14,0.25)
BVS 0.5	0.0465 (0.0143)	0.8706 (0.78,0.95)	0.1863 (0.14,0.25)
Forward	0.0528 (0.0111)	0.9456 (0.91,0.98)	0.2608 (0.20,0.33)
Backward	0.0542 (0.0116)	0.9429 (0.91,0.97)	0.2659 (0.21,0.34)
Stepwise	0.0529 (0.0111)	0.9462 (0.91,0.98)	0.2614 (0.20,0.33)

Table 3: Results and SD/quantiles for $n = 200$

Method	ATE Bias Mean (SD)	Coverage Probability Mean (10%/90% Quantile)	Coverage Length Mean (10%/90% Quantile)
		Overall	
BVS 0.1	0.0588 (0.0211)	0.8787 (0.75,0.97)	0.2412 (0.18,0.32)
BVS 0.25	0.0588 (0.0228)	0.8709 (0.72,0.97)	0.2356 (0.17,0.32)
BVS 0.5	0.0600 (0.0251)	0.8548 (0.68,0.97)	0.2300 (0.16,0.32)
Forward	0.0779 (0.0152)	0.9559 (0.93,0.98)	0.3959 (0.32,0.47)
Backward	0.0821 (0.0151)	0.9586 (0.93,0.99)	0.4245 (0.35,0.50)
Stepwise	0.0780 (0.0153)	0.9562 (0.93,0.98)	0.3965 (0.32,0.47)
		Rare	
BVS 0.1	0.0611 (0.0265)	0.8567 (0.67,0.97)	0.2350 (0.16,0.32)
BVS 0.25	0.0627 (0.0291)	0.8368 (0.61,0.96)	0.2299 (0.15,0.32)
BVS 0.5	0.0653 (0.0325)	0.8142 (0.57,0.96)	0.2261 (0.15,0.32)
Forward	0.0742 (0.0135)	0.9584 (0.93,0.99)	0.3847 (0.32,0.46)
Backward	0.0810 (0.0150)	0.9599 (0.93,0.99)	0.4215 (0.35,0.50)
Stepwise	0.0744 (0.0136)	0.9586 (0.93,0.99)	0.3852 (0.32,0.46)
		Common	
BVS 0.1	0.0585 (0.0201)	0.8820 (0.76,0.97)	0.2421 (0.18,0.33)
BVS 0.25	0.0582 (0.0217)	0.8760 (0.73,0.97)	0.2365 (0.17,0.32)
BVS 0.5	0.0592 (0.0238)	0.8608 (0.69,0.97)	0.2306 (0.17,0.32)
Forward	0.0785 (0.0154)	0.9555 (0.93,0.98)	0.3976 (0.32,0.48)
Backward	0.0823 (0.0151)	0.9585 (0.93,0.99)	0.4249 (0.35,0.50)
Stepwise	0.0786 (0.0154)	0.9559 (0.93,0.98)	0.3982 (0.32,0.48)

Table 4: Results and SD/quantiles for $n = 100$

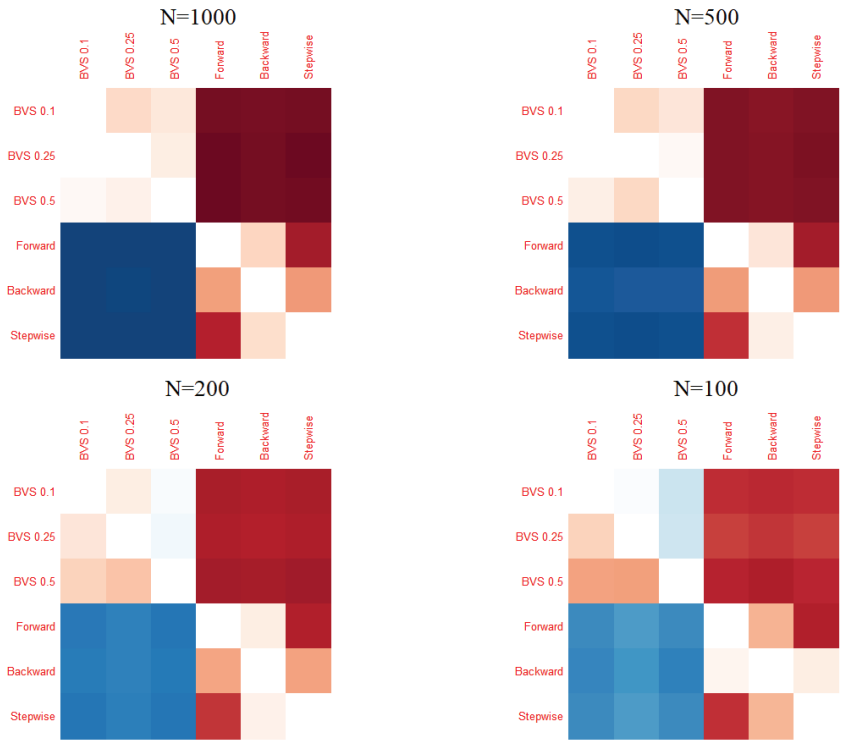


Figure 2: Continuous Simulation Results: Heat maps of method-wise average coverage probability comparisons separated by sample size, where a blue shade represents superiority of the row method in a higher proportion of simulations, and a red shade represents inferiority of the row method in a higher proportion of simulations than the column method.

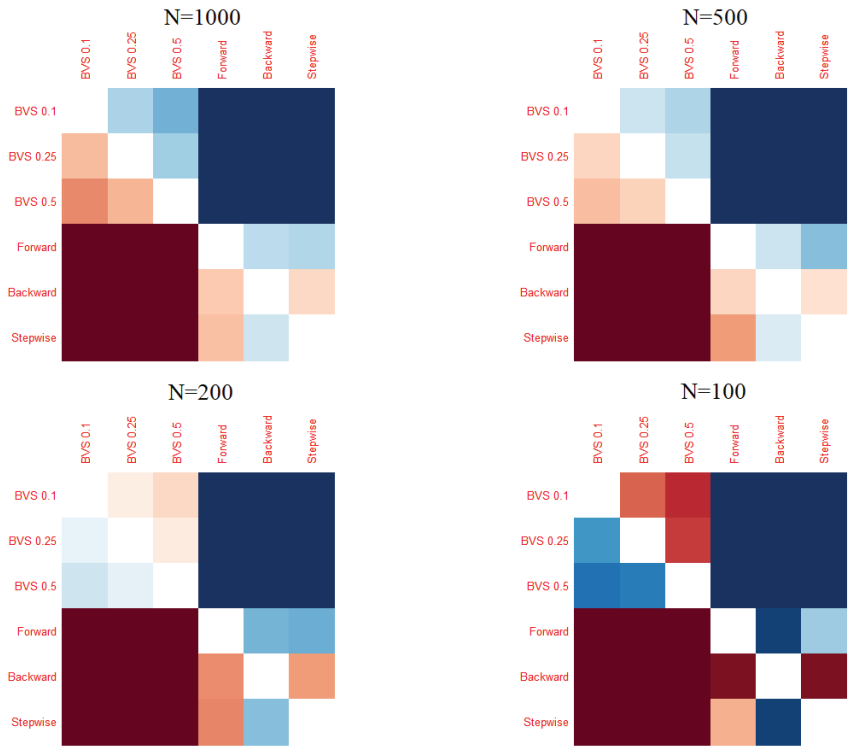


Figure 3: Continuous Simulation Results: Heat maps of method-wise average coverage length comparisons separated by sample size, where a blue shade represents superiority of the row method in a higher proportion of simulations, and a red shade represents inferiority of the row method in a higher proportion of simulations than the column method.

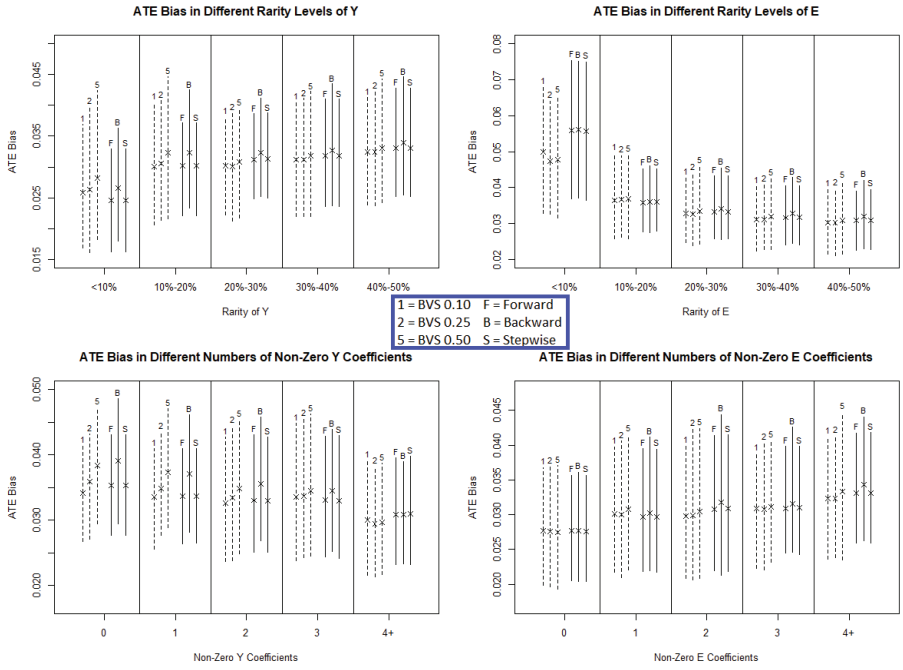


Figure 4: Continuous Simulation Results: Plots of ATE bias and 10^{th}/90^{th} quantile bars separated by rarity of Y, rarity of E, non-zero Y coefficients, and non-zero E coefficients for $n = 500$

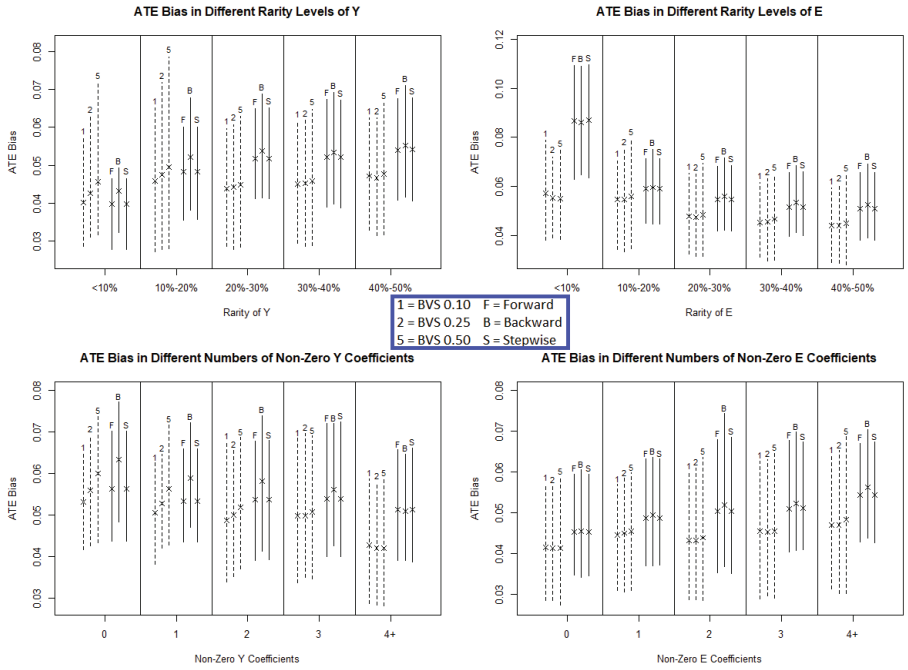


Figure 5: Continuous Simulation Results: Plots of ATE bias and 10^{th}/90^{th} quantile bars separated by rarity of Y, rarity of E, non-zero Y coefficients, and non-zero E coefficients for $n = 200$

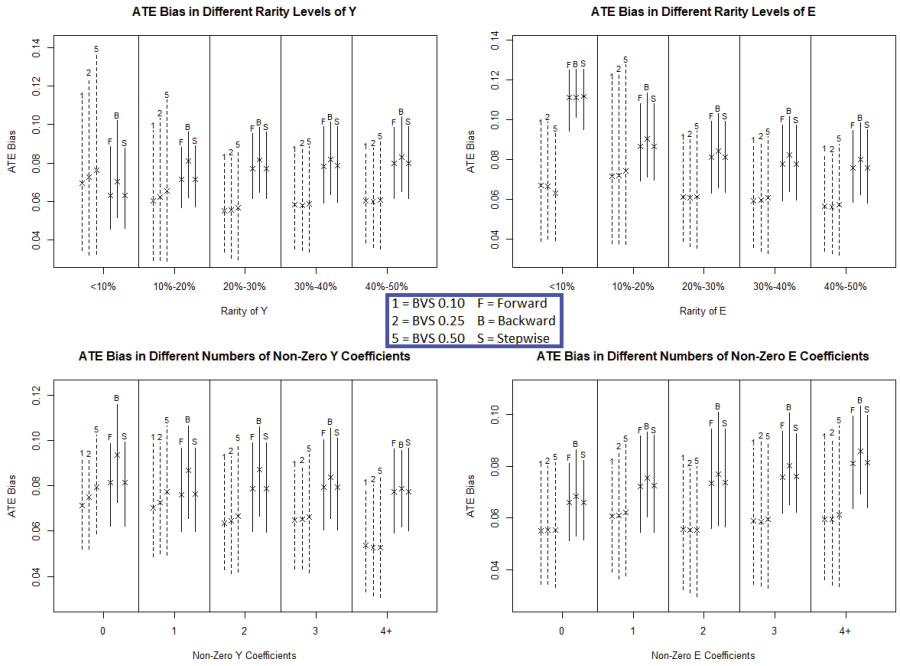


Figure 6: Continuous Simulation Results: Plots of ATE bias and $10^{th}/90^{th}$ quantile bars separated by rarity of Y, rarity of E, non-zero Y coefficients, and non-zero E coefficients for $n = 100$

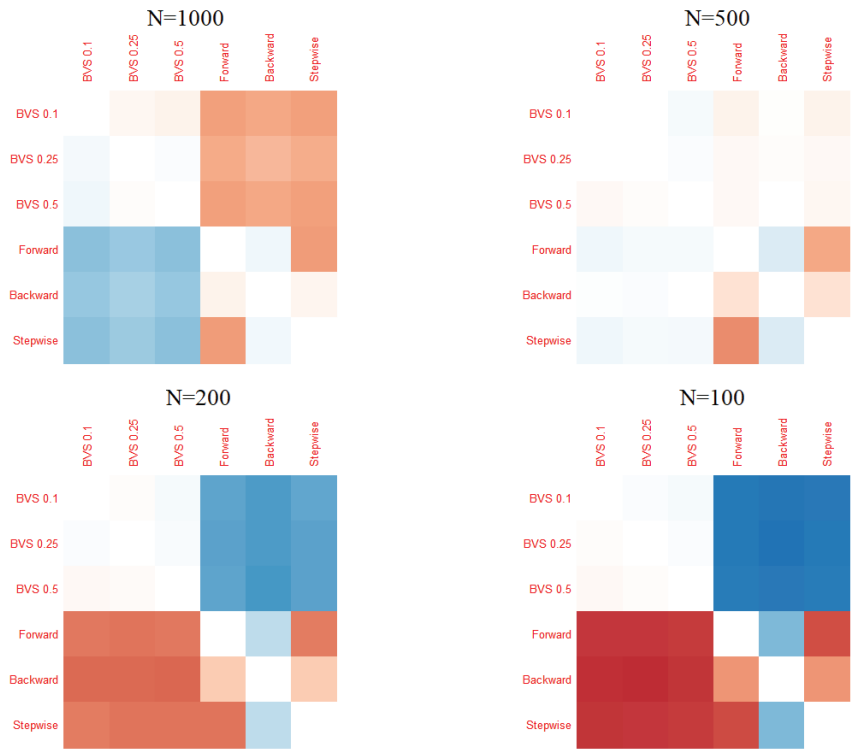


Figure 7: Binary Simulation Results: Heat maps of method-wise bias comparisons in binary simulations separated by sample size, where a blue shade represents superiority of the row method in a higher proportion of simulations

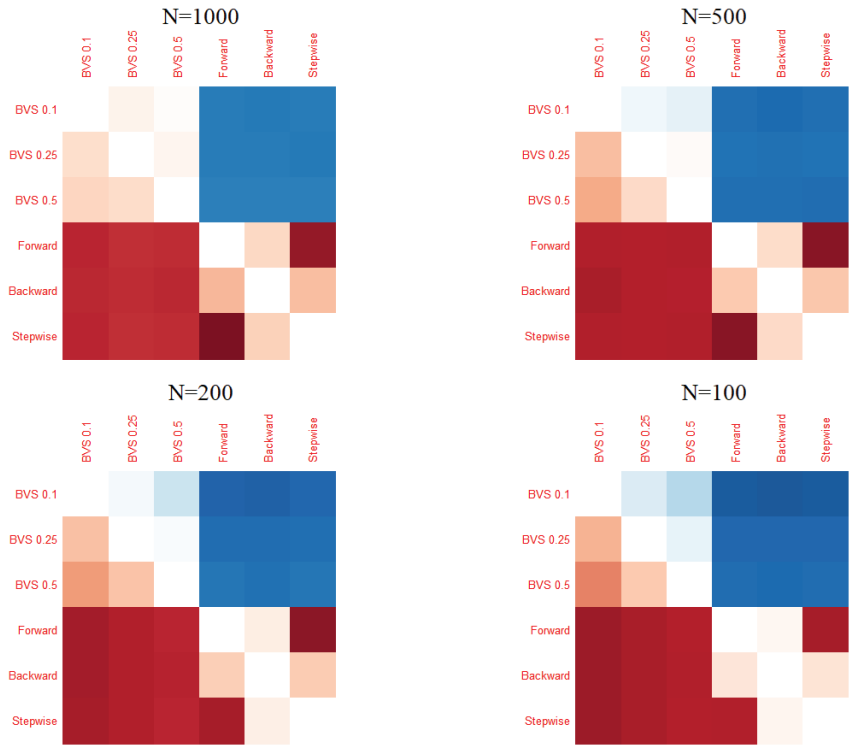


Figure 8: Binary Simulation Results: Heat maps of method-wise average coverage probability comparisons in binary simulations separated by sample size, where a blue shade represents superiority of the row method in a higher proportion of simulations

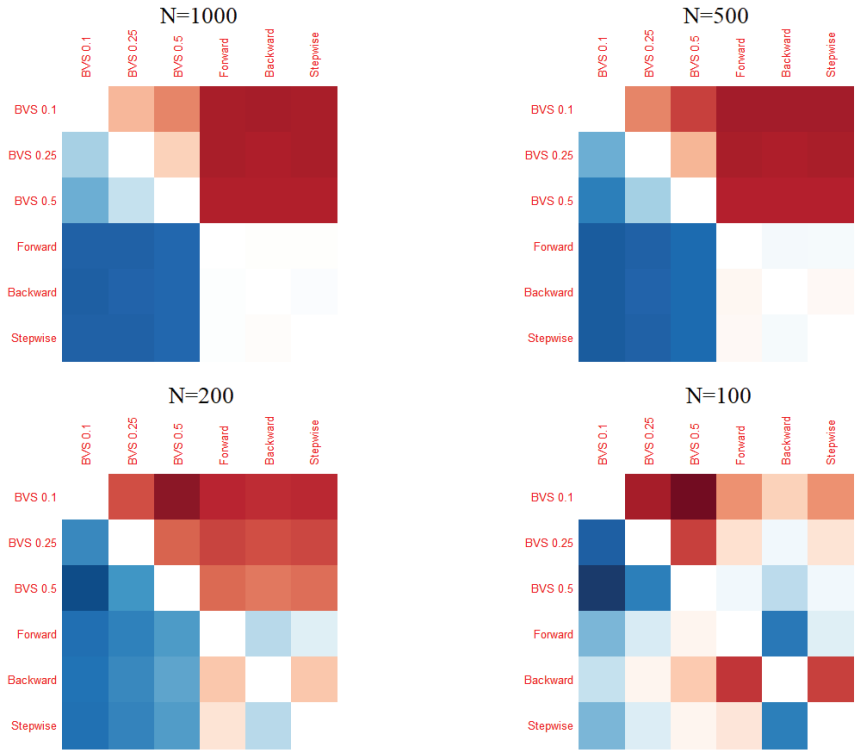


Figure 9: Binary Simulation Results: Heat maps of method-wise average coverage length comparisons in binary simulations separated by sample size, where a blue shade represents superiority of the row method in a higher proportion of simulations

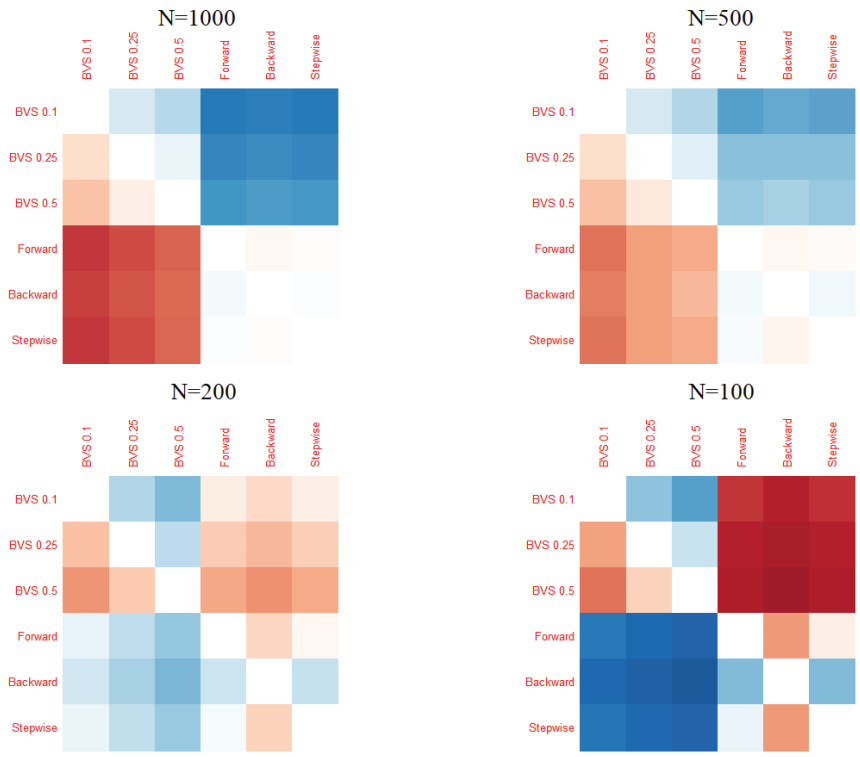


Figure 10: Binary Simulation Results: Heat maps of method-wise average coverage ratio comparisons in binary simulations separated by sample size, where a blue shade represents superiority of the row method in a higher proportion of simulations

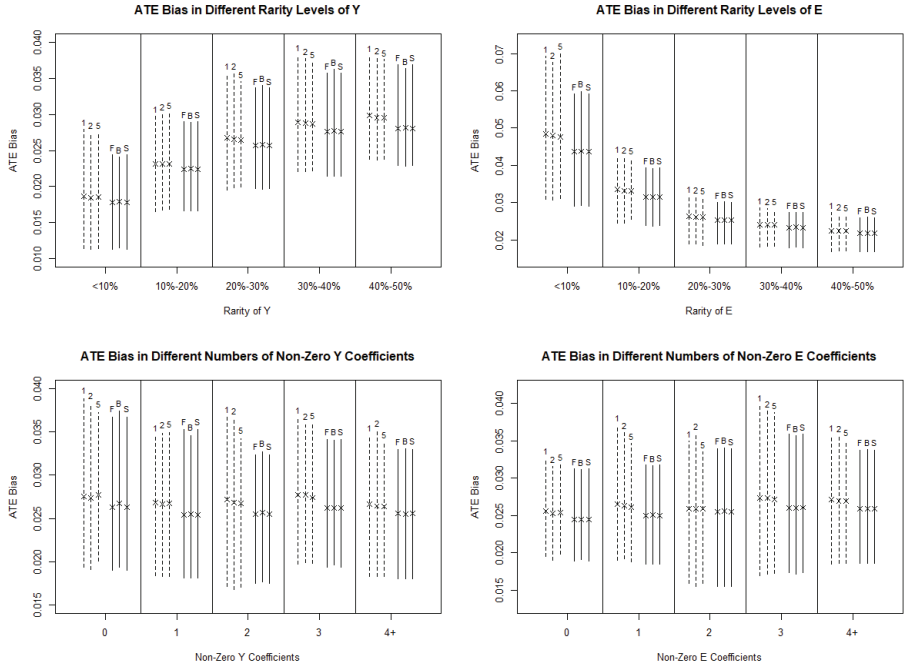


Figure 11: Binary Simulation Results: Plots of ATE bias and $10^{th}/90^{th}$ quantile bars separated by rarity of Y, rarity of E, non-zero Y coefficients, and non-zero E coefficients for $n = 1000$ binary simulations

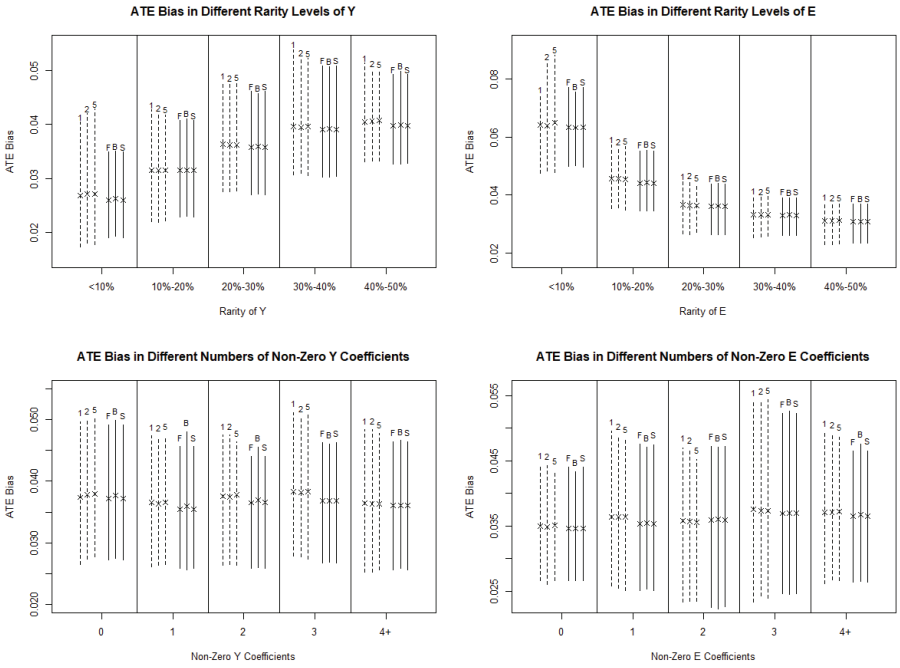


Figure 12: Binary Simulation Results: Plots of ATE bias and $10^{th}/90^{th}$ quantile bars separated by rarity of Y, rarity of E, non-zero Y coefficients, and non-zero E coefficients for $n = 500$ binary simulations

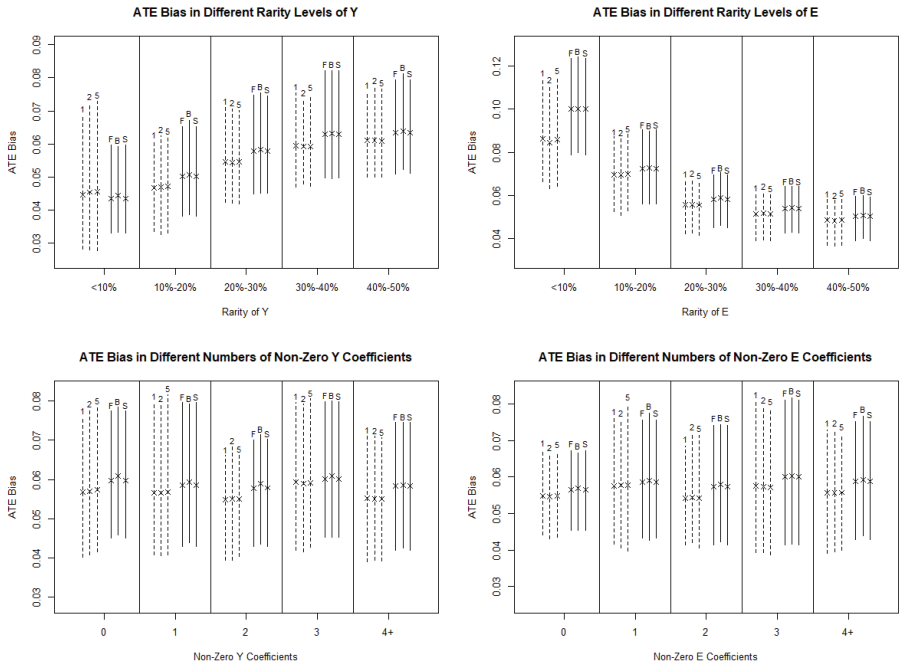


Figure 13: Plots of ATE bias and $10^{th}/90^{th}$ quantile bars separated by rarity of Y, rarity of E, non-zero Y coefficients, and non-zero E coefficients for $n = 200$ binary simulations

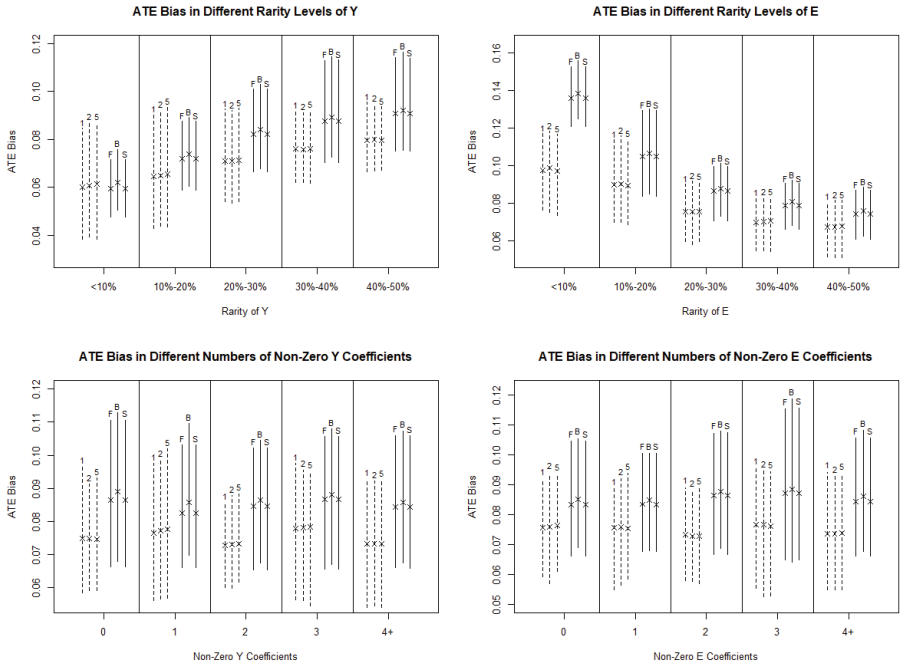


Figure 14: Binary Simulation Results: Plots of ATE bias and $10^{th}/90^{th}$ quantile bars separated by rarity of Y, rarity of E, non-zero Y coefficients, and non-zero E coefficients for $n = 100$ binary simulations

Method	ATE Bias Mean (SD)	Coverage Probability Mean (10%/90% Quantile)	Coverage Length Mean (10%/90% Quantile)
		Overall	
BVS 0.1	0.0269 (0.0086)	0.8644 (0.79,0.93)	0.1055 (0.08,0.13)
BVS 0.25	0.0267 (0.0084)	0.8629 (0.78,0.93)	0.1045 (0.08,0.13)
BVS 0.5	0.0267 (0.0083)	0.8614 (0.78,0.93)	0.1038 (0.08,0.13)
Forward	0.0257 (0.0073)	0.7957 (0.74,0.85)	0.0919 (0.07,0.12)
Backward	0.0257 (0.0074)	0.7945 (0.74,0.85)	0.0919 (0.07,0.12)
Stepwise	0.0257 (0.0073)	0.7953 (0.74,0.85)	0.0919 (0.07,0.12)
		Rare	
BVS 0.1	0.0246 (0.0084)	0.8716 (0.79,0.93)	0.0981 (0.07,0.13)
BVS 0.25	0.0245 (0.0084)	0.8697 (0.79,0.94)	0.0971 (0.07,0.13)
BVS 0.5	0.0245 (0.0081)	0.8703 (0.81,0.94)	0.0964 (0.07,0.13)
Forward	0.0235 (0.0071)	0.7973 (0.75,0.85)	0.0849 (0.06,0.11)
Backward	0.0236 (0.0071)	0.7951 (0.74,0.85)	0.0849 (0.06,0.11)
Stepwise	0.0235 (0.0071)	0.7970 (0.75,0.85)	0.0849 (0.06,0.11)
		Common	
BVS 0.1	0.0277 (0.0086)	0.8619 (0.78,0.93)	0.1081 (0.08,0.13)
BVS 0.25	0.0275 (0.0084)	0.8606 (0.78,0.93)	0.1070 (0.08,0.13)
BVS 0.5	0.0274 (0.0083)	0.8584 (0.78,0.93)	0.1063 (0.08,0.13)
Forward	0.0264 (0.0073)	0.7951 (0.74,0.85)	0.0943 (0.07,0.12)
Backward	0.0265 (0.0073)	0.7943 (0.74,0.85)	0.0943 (0.07,0.12)
Stepwise	0.0264 (0.0073)	0.7948 (0.74,0.85)	0.0943 (0.07,0.12)

Table 5: Binary Simulation Results: Display table containing mean ATE bias and standard deviation, mean coverage probability and 10%/90% Quantile, and mean coverage length with 10%/90% Quantile separated by overall data, rare data (event occured $<25\%$ of the time), and common data (event occured $>25\%$ of the time) for $n = 1000$ binary simulations

Method	ATE Bias Mean (SD)	Coverage Probability Mean (10%/90% Quantile)	Coverage Length Mean (10%/90% Quantile)
		Overall	
BVS 0.1	0.0368 (0.0105)	0.8749 (0.80,0.94)	0.1486 (0.11,0.18)
BVS 0.25	0.0368 (0.0105)	0.8681 (0.80,0.93)	0.1464 (0.11,0.18)
BVS 0.5	0.0368 (0.0107)	0.8649 (0.79,0.93)	0.1450 (0.11,0.18)
Forward	0.0363 (0.0099)	0.7965 (0.75,0.85)	0.1308 (0.10,0.17)
Backward	0.0364 (0.0099)	0.7955 (0.74,0.84)	0.1309 (0.10,0.17)
Stepwise	0.0363 (0.0099)	0.7965 (0.75,0.85)	0.1308 (0.10,0.17)
		Rare	
BVS 0.1	0.0337 (0.0098)	0.8814 (0.82,0.94)	0.1365 (0.10,0.18)
BVS 0.25	0.0337 (0.0098)	0.8729 (0.81,0.93)	0.1342 (0.10,0.17)
BVS 0.5	0.0337 (0.0099)	0.8688 (0.80,0.93)	0.1329 (0.10,0.17)
Forward	0.0331 (0.0087)	0.7911 (0.75,0.84)	0.1193 (0.09,0.16)
Backward	0.0332 (0.0088)	0.7902 (0.74,0.84)	0.1194 (0.09,0.16)
Stepwise	0.0331 (0.0087)	0.7913 (0.75,0.84)	0.1194 (0.09,0.16)
		Common	
BVS 0.1	0.0379 (0.0105)	0.8727 (0.79,0.94)	0.1526 (0.12,0.19)
BVS 0.25	0.0378 (0.0105)	0.8665 (0.79,0.93)	0.1506 (0.12,0.18)
BVS 0.5	0.0379 (0.0107)	0.8635 (0.79,0.93)	0.1491 (0.11,0.18)
Forward	0.0373 (0.0100)	0.7983 (0.75,0.85)	0.1347 (0.10,0.17)
Backward	0.0375 (0.0100)	0.7973 (0.75,0.85)	0.1348 (0.10,0.18)
Stepwise	0.0373 (0.0100)	0.7983 (0.75,0.85)	0.1347 (0.10,0.18)

Table 6: Binary Simulation Results: Display table containing mean ATE bias and standard deviation, mean coverage probability and 10%/90% Quantile, and mean coverage length with 10%/90% Quantile separated by overall data, rare data (event occured $<25\%$ of the time), and common data (event occured $>25\%$ of the time) for $n = 500$ binary simulations

Method	ATE Bias Mean (SD)	Coverage Probability Mean (10%/90% Quantile)	Coverage Length Mean (10%/90% Quantile)
		Overall	
BVS 0.1	0.0558 (0.0142)	0.8841 (0.81,0.94)	0.2296 (0.18,0.28)
BVS 0.25	0.0557 (0.0140)	0.8766 (0.80,0.94)	0.2247 (0.17,0.27)
BVS 0.5	0.0558 (0.0142)	0.8703 (0.79,0.94)	0.2209 (0.17,0.27)
Forward	0.0586 (0.0147)	0.7974 (0.75,0.85)	0.2114 (0.16,0.27)
Backward	0.0590 (0.0147)	0.7946 (0.75,0.85)	0.2124 (0.16,0.27)
Stepwise	0.0586 (0.0147)	0.7979 (0.75,0.85)	0.2115 (0.16,0.27)
		Rare	
BVS 0.1	0.0510 (0.0156)	0.8857 (0.80,0.95)	0.2090 (0.16,0.26)
BVS 0.25	0.0512 (0.0156)	0.8762 (0.80,0.95)	0.2042 (0.15,0.26)
BVS 0.5	0.0512 (0.0157)	0.8695 (0.77,0.94)	0.2003 (0.15,0.26)
Forward	0.0529 (0.0136)	0.7974 (0.74,0.85)	0.1931 (0.14,0.25)
Backward	0.0534 (0.0135)	0.7969 (0.74,0.85)	0.1945 (0.14,0.25)
Stepwise	0.0529 (0.0136)	0.7978 (0.74,0.85)	0.1932 (0.14,0.25)
		Common	
BVS 0.1	0.0574 (0.0133)	0.8836 (0.81,0.94)	0.2365 (0.19,0.28)
BVS 0.25	0.0573 (0.0131)	0.8768 (0.80,0.94)	0.2315 (0.19,0.28)
BVS 0.5	0.0573 (0.0133)	0.8706 (0.80,0.94)	0.2278 (0.18,0.28)
Forward	0.0605 (0.0145)	0.7974 (0.75,0.85)	0.2175 (0.17,0.28)
Backward	0.0609 (0.0146)	0.7938 (0.75,0.84)	0.2183 (0.17,0.28)
Stepwise	0.0605 (0.0145)	0.7979 (0.75,0.85)	0.2176 (0.17,0.28)

Table 7: Binary Simulation Results: Display table containing mean ATE bias and standard deviation, mean coverage probability and 10%/90% Quantile, and mean coverage length with 10%/90% Quantile separated by overall data, rare data (event occured $<25\%$ of the time), and common data (event occured $>25\%$ of the time) for $n = 200$ binary simulations

Method	ATE Bias Mean (SD)	Coverage Probability Mean (10%/90% Quantile)	Coverage Length Mean (10%/90% Quantile)
		Overall	
BVS 0.1	0.0742 (0.0167)	0.8972 (0.82,0.96)	0.3160 (0.26,0.37)
BVS 0.25	0.0742 (0.0170)	0.8871 (0.81,0.95)	0.3071 (0.25,0.36)
BVS 0.5	0.0743 (0.0169)	0.8791 (0.80,0.95)	0.3010 (0.24,0.36)
Forward	0.0847 (0.0177)	0.8032 (0.75,0.86)	0.3113 (0.25,0.39)
Backward	0.0864 (0.0178)	0.8018 (0.75,0.85)	0.3170 (0.26,0.39)
Stepwise	0.0847 (0.0177)	0.8034 (0.75,0.86)	0.3115 (0.25,0.39)
		Rare	
BVS 0.1	0.0692 (0.0191)	0.8958 (0.82,0.96)	0.2907 (0.24,0.35)
BVS 0.25	0.0695 (0.0195)	0.8841 (0.80,0.95)	0.2818 (0.23,0.34)
BVS 0.5	0.0701 (0.0199)	0.8728 (0.79,0.95)	0.2755 (0.22,0.33)
Forward	0.0769 (0.0166)	0.8051 (0.75,0.86)	0.2863 (0.23,0.36)
Backward	0.0786 (0.0166)	0.8060 (0.75,0.86)	0.2943 (0.24,0.36)
Stepwise	0.0769 (0.0166)	0.8053 (0.75,0.86)	0.2864 (0.23,0.36)
		Common	
BVS 0.1	0.0756 (0.0156)	0.8976 (0.83,0.96)	0.3232 (0.27,0.38)
BVS 0.25	0.0756 (0.0160)	0.8880 (0.82,0.95)	0.3144 (0.27,0.37)
BVS 0.5	0.0756 (0.0158)	0.8809 (0.81,0.95)	0.3083 (0.26,0.36)
Forward	0.0869 (0.0174)	0.8027 (0.75,0.85)	0.3185 (0.26,0.40)
Backward	0.0886 (0.0175)	0.8006 (0.75,0.85)	0.3235 (0.27,0.40)
Stepwise	0.0869 (0.0174)	0.8028 (0.75,0.85)	0.3186 (0.26,0.40)

Table 8: Binary Simulation Results: Display table containing mean ATE bias and standard deviation, mean coverage probability and 10%/90% Quantile, and mean coverage length with 10%/90% Quantile separated by overall data, rare data (event occured $<25\%$ of the time), and common data (event occured $>25\%$ of the time) for $n = 100$ binary simulations