The Use of Propensity Scores to Estimate Sample Selection Error in Observational Data

Thesis

Presented in Partial Fulfillment of the Requirements for the Degree of Master of Science in the Graduate School of The Ohio State University

By
Taylor Renee Pressler, A.B.
Graduate Program in Statistics

The Ohio State University
2011

Thesis Committee:
Eloise Kaizar, Advisor
William Gardner
Copyright by
Taylor R. Pressler
2011
Abstract

While randomized controlled trials (RCT) are considered the “gold standard” for clinical studies, the use of exclusion criteria may impact the external validity of the results. It is unknown whether estimators of effect size are biased by excluding a portion of the population. However, it may be possible to use data from observational studies to estimate a difference between the population average treatment effect (PATE) of the included and excluded portions of the population, the sample selection error (SSE). We propose an estimator for the SSE and use simulation to study its properties while considering a non-constant treatment effect. We find that a doubly robust estimator that uses both propensity scores and a model for the outcome generally outperforms an estimator that solely relies on the use of propensity scores, even when model elements are misspecified.
Dedication

I would like to dedicate this thesis to the memory of my father, Eric S. Pressler and in honor of my son, William L. Durishin. I hope to be able to pass along the enjoyment of mathematical learning to William as it was passed to me by my father.
I would like to acknowledge Dr. Elly Kaizar, for her constant support and guidance, as well as all of the faculty in the Department of Statistics. The completion of this thesis would not have been possible without the education and knowledge I received there.

I owe Dr. Philip Payne an immeasurable amount of gratitude for his continuous support and mentorship during my tenure at The Ohio State University. I also owe many thanks to Dr. Phil Binkley, as he has provided me many opportunities for leadership and teaching the importance of statistics to others.

And finally, I owe a great deal of gratitude and appreciation to Gabriel Vydra, for without the love, support, and shoulder massages he has graciously provided, this thesis may have not been possible.
Vita

2000 ......................................................... Washington High School
2008 ......................................................... A.B., Mathematics, Mount Holyoke College

Awards and Honors

2007 ........................................ Howard Hughes Medical Institute Research Fellowship
2007 ........................................ MIT Lincoln Laboratory Travel Fellowship
2008 ........................................ Student Poster Award, Joint Mathematics Meeting
2008 ................................................. Carol Loeb ’63 Award
2008 ................................................. Mary Lyon Scholar
2008 ................................................. OSU Graduate Enrichment Fellowship
2008 ................................................. Lubrizol Fellowship
2008 ................................................. Getrude Cox Award, Honorable Mention

Academic and Professional Appointments

2008-2009 ................................. University Fellow, The Ohio State University
2009 ........................................... Graduate Teaching Assistant, Department of Statistics
2009-present .............................. Research Specialist, Department of Biomedical Informatics

Publications


Fields of Study

Major Field: Statistics
# Table of Contents

Abstract ......................................................... ii
Dedication ....................................................... iii
Acknowledgements .............................................. iv
Vita ............................................................... v
List of Symbols ................................................ x

1 Introduction .................................................... 1

2 Background .................................................... 2
   2.1 Randomized Controlled Trials ................................ 2
   2.2 Observational Studies ........................................ 2
   2.3 Treatment Effect ............................................. 3
   2.4 Propensity Scores ............................................ 3
   2.5 Sample Selection Error ...................................... 4
   2.6 Study Aims ................................................... 5

3 Estimation of Parameters of Interest ......................... 6
   3.1 Response Model .............................................. 6
   3.2 Treatment Effect ............................................. 6
      3.2.1 Estimation of Treatment Effect ......................... 7
   3.3 Sample Selection Error .................................... 8

4 Simulation Methods ........................................... 10
   4.1 Data Generation ............................................ 10
      4.1.1 Covariate Variable Simulation ......................... 10
      4.1.2 Treatment Variable Simulation ......................... 10
      4.1.3 Response Variable Simulation ......................... 10
   4.2 Exclusion Criteria Selection .............................. 10
   4.3 Sampling Procedure ....................................... 11
      4.3.1 Model Misspecification ................................. 11

5 Evaluation of Estimators ...................................... 12
   5.1 Bias and Mean Squared Error ............................... 12
   5.2 Confidence Interval Coverage Probability ................ 12
### Table of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Population values for the PATE for the excluded and the included subgroups of the population.</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>The population differences in the values of average treatment effect between the excluded and included subgroups.</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>Population values of SSE.</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>Magnitude of bias of $\hat{\Delta}_{str}$ in estimating $PATE^I$.</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>Magnitude of bias of $\hat{\Delta}_{dr}$ in estimating $PATE^I$.</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>Difference in the bias of $\hat{\Delta}_{str}$ of subpopulation average treatment effect for the included and excluded subgroups.</td>
<td>17</td>
</tr>
<tr>
<td>7</td>
<td>Difference in the bias of $\hat{\Delta}_{dr}$ of subpopulation average treatment effect for the included and excluded subgroups.</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>Bias of $\hat{\gamma}_{str}$.</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td>Bias of $\hat{\gamma}_{dr}$.</td>
<td>19</td>
</tr>
<tr>
<td>10</td>
<td>MSE of $\hat{\gamma}_{str}$.</td>
<td>20</td>
</tr>
<tr>
<td>11</td>
<td>MSE of $\hat{\gamma}_{dr}$.</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>Confidence Interval Coverage of $\hat{\gamma}_{str}$.</td>
<td>21</td>
</tr>
<tr>
<td>13</td>
<td>Confidence Interval Coverage of $\hat{\gamma}_{dr}$.</td>
<td>21</td>
</tr>
</tbody>
</table>
List of Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta$</td>
<td>Population average treatment effect (PATE)</td>
</tr>
<tr>
<td>$\Delta^E$</td>
<td>Subpopulation average treatment effect for the excluded subgroup ($PATE^E$)</td>
</tr>
<tr>
<td>$\Delta'$</td>
<td>Subpopulation average treatment effect for the included subgroup ($PATE'$)</td>
</tr>
<tr>
<td>$\pi_E$</td>
<td>Proportion of excluded units in the population</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Sample selection error (SSE)</td>
</tr>
</tbody>
</table>
1 Introduction

In many areas of clinical research, the research question to be answered centers around the existence of a treatment effect. If a new medication becomes available for the treatment of some disorder, a study would be designed to gather information as to whether or not the new treatment offers a greater therapeutic effect than existing therapies. In general, randomized controlled trials (RCTs) are considered the “gold standard” for such evaluations [Rothwell, 2005]. RCTs are purposefully designed so that potential imbalances between the groups are minimized over the course of the study and therefore have strong internal validity [Rothwell, 2005]. However, there may be concern over the external validity of RCTs since the sample population may not always represent the population of interest. Many clinical trials have exclusion criteria, and these criteria can exclude a large portion of the population of interest. Exclusion criteria often exclude potential participants that have a severe form of the disease, women of childbearing age, children, or patients with multiple co-morbidities [Wiltsey, 2003]. It is unclear whether the treatment effect estimated in an RCT would be the same if there was no exclusion in the trial.

In this thesis, we propose a class of estimators of the bias induced by exclusion criteria when there is a non-constant treatment effect, and evaluate its properties using a simulation study. Our proposed class of estimators uses observational data to estimate the sample selection error (SSE) due to excluding a portion of the population. We will examine the ability of our proposed estimator of SSE to estimate this value as covariate correlation and exclusion rates change.
2 Background

2.1 Randomized Controlled Trials

RCTs have strong internal validity because they are designed in such a way that all participants within the study have a random allocation of treatment or control. This random allocation of treatment versus control allows for causal inference to be made with respect to an effect of treatment since the randomization helps to ensure that the treatment and control groups are on average similar in makeup.

In an ideal RCT, an effect of treatment estimated from the trial data is directly generalizable to the population. This is not generally the way RCTs are designed since many key portions of a population can be excluded from participating in a trial, yet an estimated effect of treatment is still often generalized to the population. For example, a common exclusion criteria prevents patients with multiple co-morbidities from participating in the RCT, yet it is common to find such patients in a clinical setting for which the treatment is intended [Zimmerman, 2007].

The exclusion criteria define a subset of the population that would not be eligible for participation in a clinical trial. Exclusion criteria often include pregnancy, young age, or multiple co-morbidities. What is not clear is whether bias is incurred from exclusion in RCTs. One would expect bias if the portion of the population that was excluded would react to treatment differently than the portion of the population that was eligible for the study. If the effect of treatment is constant over the population, no bias would be expected because the exclusion would have no impact on the effect measured in the study.

2.2 Observational Studies

Observational studies differ from RCTs as there is no random assignment of the treatment or control to the participants of the study. Rather, participants are free to elect the experimental factor. This can make causal inference difficult as there may be other factors, both observed and unobserved, that are on average different between the control and treatment groups, and could impact the outcome. When trying to measure an effect of treatment, it becomes difficult to know whether any difference in outcome was caused by the treatment alone or by other factors. That is, the effect of treatment may be confounded. Variables can confound an estimate of the effect of treatment when they are related to the treatment variable of interest and the outcome variable.

For example, consider a study of a new breast cancer treatment. There are several factors that are considered for each subject and included in the model, such as age, presence of
BRAC1 and BRAC2 genes, family history, number of pregnancies, race, and ethnicity. However, the study does not take smoking or nutrition into account. If those who do not smoke and eat more healthily are more likely to try the new breast cancer treatment, it will be impossible to distinguish the effect of smoking and nutrition and the effect of breast cancer treatment. We say the effect of the treatment is confounded with the effect of smoking and nutrition.

2.3 Treatment Effect

A treatment effect is the difference in response between a treatment and a control condition for an individual and is the underlying quantity of interest in most clinical research. The treatment effect is $TE = Y_i(1) - Y_i(0)$, where $Y_i(1)$ is the response after receiving treatment and $Y_i(0)$ is the response after receiving control for the $i^{th}$ individual in a population of size $N$. It is not usually possible to measure the treatment effect for an individual, since one is generally unable to measure a single individual’s response to both treatment and control conditions. The treatment effect for individuals can vary from person to person for many reasons, both observed and unobserved [Imai, 2007]. For this reason, much clinical research attempts to estimate the Population Average Treatment Effect (PATE), which is the average treatment effect over all individuals in a defined population:

$$PATE = \frac{1}{N} \sum_{i=1}^{N} [Y_i(1) - Y_i(0)] = \frac{1}{N} \sum_{i=1}^{N} TE_i$$ (1)

For an RCT, PATE is often estimated using the difference between the sample average response in the treatment and the control groups, $\hat{PATE} = \frac{1}{nT} \sum_{i=1}^{nT} Y_iT_i - \frac{1}{nC} \sum_{i=1}^{nC} Y_i(1 - T_i)$, where $T_i$ is 1 if a subject received treatment and 0 otherwise. This estimate is not appropriate in observational studies, due to the possibility of confounding. In order to estimate an effect of treatment in an observational study, we must use methods that reduce or eliminate confounding. One such method is to use weighted estimators based on propensity score models.

2.4 Propensity Scores

In order to make a comparison across two groups to measure an effect of treatment, one approach a researcher could take is to be certain that the two groups being compared are similar for any factors that could have an impact on the response. Assuming that all possibly confounding covariates are observed and measured, a researcher could match up participants who elected the control with similar participants who elected treatment, either through direct matching or by use of propensity score weighting. In this study, we focus on the use of weights based on the propensity score.

The use of propensity scores can balance the covariates between the treatment groups in observational studies to allow causal inferences. A propensity score is the probability of receiving treatment given a set of covariates [Rosenbaum and Rubin, 1983]. The propensity
score can be estimated using a logistic regression model, such as the following for three covariates:

\[
e(X) = P(T = 1|X, \alpha) = \frac{e^{\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3}}{1 + e^{\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3}}, \text{ where (2)}
\]

\[
T_i \sim Bernoulli(e(X_i))
\]

Here, \(e(X)\) is the true propensity score for a subject with covariate vector \(X\). Using this model, we estimate a propensity score for each subject in the study, and subjects in treatment and control groups can be matched or weighted using the estimated scores. Matching and weighting helps to eliminate confounding in the study [Rosenbaum and Rubin, 1983]. If matching or weighting are used correctly, then one would expect an estimated treatment effect to be free of confounding since each treatment unit is matched with a control unit on factors that could otherwise confound the treatment effect or weighted to reflect an appropriately similar population. However, the effective use of propensity scores relies on including all potential confounding variables in the propensity score model [Rosenbaum and Rubin, 1983]. This should be considered carefully in a practical setting since it is not always possible to know every possible confounding variable nor to be able to measure every known confounding factor. If confounding factors are omitted from the model, then bias will not be eliminated in estimates of PATE, and in fact the resulting estimators could be even more biased than the naïve estimators.

2.5 Sample Selection Error

When considering the use of RCTs, one should not ignore the potential bias that may be incurred when using exclusion criteria. This bias, also known as the Sample Selection Error (SSE), would occur if there was a non-constant treatment effect and the average effect of treatment of the excluded group is different from the average effect of treatment of the group included in the study.

SSE is the difference in the PATE for the entire population and PATE for the included portion of the population (\(PATE^I\)):

\[
\gamma = PATE - PATE^I = \frac{1}{N} \sum_{i=1}^{N} TE_i - \frac{1}{N_I} \sum_{i=1}^{N} TE_i (1 - I_i)
\]  

(3)

where \(I_i = \begin{cases} 0 & \text{if exclusion criteria are satisfied for subject } i; \\ 1 & \text{otherwise.} \end{cases}\)

and \(N_I = \sum_{i=1}^{N} I_i.\)
2.6 Study Aims

Many studies have examined the use of propensity scores to estimate PATE. However, most studies have been designed under the assumption that the treatment effect is constant. We study the effect of exclusion criteria on the estimation of PATE when there is a non-constant treatment effect present. We propose an estimator for $\gamma$, the SSE, and explore its properties via simulation.
3 Estimation of Parameters of Interest

3.1 Response Model

We consider a linear response model with three covariates and interaction with treatment:

\[
E[Y_i | (X, \beta, T, \delta)] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i} t + \delta t_i
\]  

(4)

The linear model was used to mimic models typically used in practice. This study considers a non-constant treatment effect in the response model, which differs from most studies.

3.2 Treatment Effect

Recall that the treatment effect is the difference in response under treatment and control for the \(i^{th}\) individual. For simplicity, we assumed that the treatment effect is constant for each individual. Under this assumption, PATE can be calculated directly as the difference in the expected response conditional on the values of the three covariates. For the proposed linear response model (4), the TE reduces as follows:

\[
TE_i = Y_i(1) - Y_i(0) = E[Y_i | t = 1] - E[Y_i | t = 0] = [\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i} t + \delta t_i] - [\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}] = \beta_4 X_{1i} + \delta
\]  

(5)

As seen above, the treatment effect is a function of both treatment and a covariate that moderates the treatment. To date, most studies of propensity scores do not consider such a non-constant treatment effect.

In real observational studies, it is not possible to calculate PATE directly. However, in this simulation study, the PATE can be calculated from the individual treatment effects:

\[
PATE = \beta_4 \frac{1}{N} \sum_{i=1}^{N} X_{i1} + \delta
\]  

(6)
This study also looks at PATE for each separate population of those who would be included, \( PATE^I \), or excluded, \( PATE^E \), in a study of interest.

3.2.1 Estimation of Treatment Effect

Two estimators will be used in the course of this study to estimate PATE. The first uses stratification based on the estimated propensity scores proposed by Rosenbaum and Rubin (1983). The second estimator uses both regression of the response variable on the covariates and the estimated propensity scores to estimate PATE, thus making a doubly robust estimator. This estimator was defined by Lunceford and Davidian (2004). Each estimator, \( \Delta \), has properties such that \( \frac{1}{\sqrt{n}}(\hat{\Delta} - \Delta_0) \) converges in distribution to a \( N(0, \Sigma) \) random variable, where \( \Delta_0 \) is the true PATE, and \( \Sigma \) is the variance of the estimator [Lunceford, 2004]. Thus, the approximate large sample distribution of each estimator will be \( N(\Delta_0, \frac{1}{n} \Sigma) \), although \( \Sigma \) is not generally known. Confidence intervals can be constructed using an estimator of the variance, \( \hat{\Sigma} \).

**Stratified Estimator** The stratified estimator, \( \hat{\Delta}_{str} \), uses estimated propensity scores to stratify the sample and then estimates the treatment effect by a weighted average of the difference in the average response between the units that chose treatment and control within each stratum. While different stratification schemes can be used, the most common is to define the strata with the quintiles of the estimated propensity scores. After stratification, the estimator can be calculated:

\[
\hat{\Delta}_{str} = \sum_{j=1}^{5} \frac{n_j}{n} \left[ \frac{1}{n_{1j}} \sum_{i:e_i \in Q_j} Y_i T_i - \left( \frac{1}{n_j - n_{1j}} \right) \sum_{i:e_i \in Q_j} Y_i (1 - T_i) \right]
\]

where \( n \) is the total sample size, \( n_j \) is the sample size of the \( j \)th stratum, \( j = 1, 2, ..., 5 \), \( n_{1j} \) is number of treated units in the \( j \)th stratum, and \( Q_j \) is the range of estimated propensity scores that defines stratum \( j \). The variance of this estimator is commonly estimated:

\[
\hat{\Sigma}_{str} = \frac{1}{25} \sum_{j=1}^{5} \hat{\sigma}_j^2, \quad \text{where}
\]

\[
\hat{\sigma}_j^2 = \frac{1}{n_{1j}} s_{1j}^2 + \frac{1}{n_j - n_{1j}} s_{0j}^2,
\]

\[
s_{1j}^2 = \frac{1}{n_{1j}} \sum_{i=1}^{n} I(\hat{e}_i \in Q_j) T_i (Y_i - \bar{y}_i)^2,
\]

and

\[
s_{0j}^2 = \frac{1}{n_j - n_{1j}} \sum_{i=1}^{n} I(\hat{e}_i \in Q_j) (1 - T_i) (Y_i - \bar{y}_0)^2.
\]
Doubly Robust Estimator  The doubly robust estimator, $\hat{\Delta}_{dr}$, is robust to mis-specification of the response model and to mis-specification of the propensity model. The estimator we use also has the smallest variance in the class of PATE estimators (Lunceford, 2004). It is defined as:

$$
\hat{\Delta}_{dr} = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i T_i - (T_i - \hat{e}_i) m_1(X_i, \hat{\beta}_1)}{\hat{e}_i} - \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i (1 - T_i) + (T_i - \hat{e}_i) m_0(X_i, \hat{\beta}_0)}{1 - \hat{e}_i},
$$

where $m_t(X_i, \hat{\beta}_t) = \hat{E}[Y|T=t,X_i]$ is the predicted response from a regression of the response $Y$ on the covariates $X_i$, where $\hat{\beta}_t$ are estimated separately for $t = 0, 1$, using only the units in the control group and the units in the treatment group. The form of the estimator can be re-rewritten:

$$
\hat{\Delta}_{dr} = \left[ \frac{1}{n} \sum_{i=1}^{n} m_{1i}(X_i, \hat{\beta}_1) + \sum_{i=1}^{n} \frac{T_i (Y_i - m_{1i}(X_i, \hat{\beta}_1))}{\hat{e}_i} \right] - \left[ \frac{1}{n} \sum_{i=1}^{n} m_{0i}(X_i, \hat{\beta}_0) + \sum_{i=1}^{n} \frac{(1 - T_i) (Y_i - m_{0i}(X_i, \hat{\beta}_0))}{(1 - \hat{e}_i)} \right].
$$

When the response model is correct, the value of the error term, $\hat{E} \left( Y_i - m_{ti}(X_i, \hat{\beta}_t) \right)$, goes to zero in expectation. If the response model is not correct, the second term for each component provides a correction to the expected response, which is weighted by the estimated propensity score, $\hat{e}_i$ (Kang and Shafer, 2007). Thus, this estimator is doubly robust to misspecifications to the response model and the propensity score model.

The variance for this estimator can be estimated:

$$
\hat{\Sigma}_{dr} = \frac{1}{n^2} \sum_{i=1}^{n} \left[ \frac{Y_i T_i - (T_i - \hat{e}_i) m_1(X_i, \hat{\beta}_1)}{\hat{e}_i} - \frac{Y_i (1 - T_i) + (T_i - \hat{e}_i) m_0(X_i, \hat{\beta}_0)}{1 - \hat{e}_i} - \hat{\Delta}_{dr} \right]^2.
$$

3.3 Sample Selection Error

Recall that SSE is defined to be the difference between $PATE$ and $PATE^I$, as defined in Equation 3, for the portion of the population that would be eligible for trial participation. We propose the following class of estimators to estimate the sample selection error of a sampling procedure:

$$
\hat{\gamma} = \hat{\pi}_E \left( \hat{\Delta}_E - \hat{\Delta}_I \right)
$$
where $\pi_E$ is the proportion of excluded units in the population, $\hat{\pi}_E$ is the proportion of excluded units in the sample, $\hat{\Delta}^E$ is the estimated PATE for the portion of the sample that would have been excluded from participation, and similarly, $\hat{\Delta}^I$ is the estimated treatment effect for the included portion of the population. $\hat{\Delta}^I$ and $\hat{\Delta}^E$ are calculated using Equations 7 and 9 for only the subsets of the sample that would have been included and excluded in an RCT, respectively.

We approximate the variance of $\hat{\gamma}$ using the Law of Total Variance as follows:

\[
\Sigma_\gamma = \text{Var} \left[ \hat{\pi}_E (\hat{\Delta}^E - \hat{\Delta}^I) \right] \\
= \text{Var}_\hat{\pi}_E \left[ E \left( \hat{\pi}_E (\hat{\Delta}^E - \hat{\Delta}^I | \hat{\pi}_E) \right) \right] + E \text{Var}_\hat{\pi}_E \left[ \text{Var} \left( \hat{\pi}_E (\hat{\Delta}^E - \hat{\Delta}^I | \hat{\pi}_E) \right) \right] \\
= \text{Var}_\hat{\pi}_E \left[ \hat{\pi}_E E \left( \hat{\Delta}^E - \hat{\Delta}^I | \hat{\pi}_E \right) \right] + E \text{Var}_\hat{\pi}_E \left[ \hat{\pi}_E^2 \text{Var} \left( \hat{\Delta}^E - \hat{\Delta}^I | \hat{\pi}_E \right) \right] \\
\approx \text{Var}_\hat{\pi}_E \left[ \hat{\pi}_E (\hat{\Delta}^E - \hat{\Delta}^I) \right] + E \text{Var}_\hat{\pi}_E \left[ \hat{\pi}_E^2 \text{Var} \left( \hat{\Delta}^E - \hat{\Delta}^I \right) \right] \\
\approx (\Delta^E - \Delta^I)^2 \text{Var}(\hat{\pi}_E) + \left( \text{Var}(\hat{\pi}_E) + E[\hat{\pi}_E]^2 \right) [\text{Var}(\hat{\Delta}^E) + \text{Var}(\hat{\Delta}^I)] \\
= \frac{1}{n} \pi_E (1 - \pi_E) (\Delta^E - \Delta^I)^2 + \left( \frac{\pi_E (1 - \pi_E)}{n} + \pi_E^2 \right) [\text{Var}(\hat{\Delta}^E) + \text{Var}(\hat{\Delta}^I)] \\
= \frac{(1 - \pi_E)}{n \pi_E} \gamma^2 + \left( \frac{\pi_E (1 - \pi_E)}{n} + \pi_E^2 \right) [\text{Var}(\hat{\Delta}^E) + \text{Var}(\hat{\Delta}^I)]
\]

We then use the estimated values for the true values to estimate the approximate variance:

\[
\hat{\Sigma}_\gamma = \frac{(1 - \hat{\pi}_E)}{n \hat{\pi}_E} \gamma^2 + \left( \frac{\hat{\pi}_E (1 - \hat{\pi}_E)}{n} + \hat{\pi}_E^2 \right) [\text{Var}(\hat{\Delta}^E) + \text{Var}(\hat{\Delta}^I)]
\]

(14)
4 Simulation Methods

4.1 Data Generation

We consider hypothetical data sets of size \(n=1,000,000\) with a single continuous outcome, a binary treatment, and three continuous covariates. All simulations were completed using R statistical software and the computer code for the generated analysis can be found in the Appendix.

4.1.1 Covariate Variable Simulation

All three covariates are simulated using a multivariate normal random number generator with mean zero and variance of one. We consider only positive correlation values.

4.1.2 Treatment Variable Simulation

In observational studies, treatment is not randomly assigned, but rather elected by study participants. Thus, the treatment election is not independent of the covariates in the simulated data. The treatment variable is simulated using a Bernoulli random number generator, where the probability of receiving treatment is defined in Equation 3.

The values of \(\alpha\) for the model were chosen so that in each simulated population, approximately 30% of the population elected treatment. The value of \(\alpha_0\) was set equal to \(-0.7\) and the values of \(\alpha_i\) for \(i = 1, 2, 3\) equal to 0.5.

4.1.3 Response Variable Simulation

The simulation of the response variable is based upon a normal random number generator, with a variance of one and a mean dependent upon the three covariates, treatment, and an interaction between treatment and one covariate, as shown in Equation 4. Each coefficient is set equal to one: \(\beta_j = 1, j = 0 \ldots 4\) and \(\delta = 1\).

4.2 Exclusion Criteria Selection

The exclusion criteria specifies whether a particular unit is eligible for RCT participation. The exclusion criterion for this study is a binary variable that is determined by values of
one covariate that does not directly interact with treatment ($X_2$). The exclusion criterion was set to exclude 25%, 50%, and 75% of the population respectively. It was calculated by comparing $X_2$ to a Normal quartile $q_{0.25}$, $q_{0.5}$, or $q_{0.75}$.

4.3 Sampling Procedure

The sampling and estimation procedures were repeated 5,000 times for this simulation study. The sampling and estimation procedure for each iteration consists of the following:

- A sample of size 10,000 is taken without replacement from the simulated population.
- The propensity score is the estimated for each sample unit, using Equation 3.
- The stratified estimator $\hat{\Delta}_{str}$ and its estimated variance are calculated using Equations 7 and 8.
- The doubly robust estimator $\hat{\Delta}_{dr}$ and its estimated variance are calculated using Equations 9 and 12.
- The SSE $\hat{\gamma}$ and its estimated variance are calculated using Equations 13 and 14.

4.3.1 Model Misspecification

We first study the behavior of the estimators under the ideal situation when the models for the propensity scores and the outcome are specified correctly (referenced as ‘correct’ in the results section). We also wish to evaluate the behavior of the estimators when the models are misspecified. Recall the important consideration that all potential confounding variables must be included in the model for the propensity scores so that the use of the scores can be effective (Rosenbaum and Rubin, 1983). The covariate involved in the interaction term, $X_1$, is in turn dropped from the model for the propensity scores (‘propensity’), the model for the outcome (‘regression’), and both of the models simultaneously (‘both’).
5 Evaluation of Estimators

For each of the estimators, evaluation estimates are calculated including bias, mean squared error, and confidence interval coverage probability.

5.1 Bias and Mean Squared Error

Bias is described as the deviation of an estimator’s expected value from the true value of the estimand, \( \text{Bias}(\hat{\Delta}) = E(\hat{\Delta} - \Delta) \). Mean squared error is the square of the deviation of an estimator from the expected value, \( \text{MSE}(\hat{\Delta}) = E((\hat{\Delta} - \Delta)^2) \). The \( \text{Bias}(\hat{\Delta}) \) can be estimated by \( \hat{\text{Bias}}(\hat{\Delta}) = \frac{1}{s} \sum_{k=1}^{s} (\hat{\Delta}_k - \Delta) \), where \( s \) is equal to the number of simulations, 5,000. Similarly, the \( \text{MSE}(\hat{\Delta}) \) can be estimated by \( \hat{\text{MSE}}(\hat{\Delta}) = \frac{1}{s} \sum_{k=1}^{s} (\hat{\Delta}_k - \Delta)^2 \). These are reasonable estimators since all samples are equal in probability during the selection process.

5.2 Confidence Interval Coverage Probability

Using the approximate normal distribution of each estimator, the confidence interval for PATE under each estimator \( \hat{\Delta} \) is calculated and the proportion of times that PATE is included in each respective confidence interval is recorded. The confidence intervals for PATE under \( \Delta \) are defined as:

\[
\hat{\Delta} \pm z_{\alpha/2} [\hat{\Sigma}]^{1/2}
\]

The confidence interval for the SSE, \( \gamma \), were constructed in a similar way:

\[
\hat{\gamma} \pm z_{\alpha/2} [\hat{\Sigma}_\gamma]^{1/2}
\]
6 Results

6.1 Simulation Results for the Population

For each of the three simulated populations, the overall $PATE$, $PATE^E$, and $PATE^I$ were calculated. The results, as displayed in Figure 1, illustrate how the values of subpopulation average treatment effect change in comparison to the true PATE. As the correlation between covariates increases, the deviation between the values for the subpopulation treatment effect and the true PATE also increases. This is expected since the covariate used for the exclusion criteria, $X_2$, not only effects the outcome directly, but also through the other correlated covariates. This deviation also increases as the relative size of the subgroup increases. In Figure 2, the magnitude of the difference between the subpopulation treatment effects ($PATE^E - PATE^I$) also increases with correlation. This is expected, since the average subpopulation treatment effects should be identical when the exclusion criterion $X_2$ and the treatment moderator $X_1$ are independent. However, as the correlation between $X_1$ and $X_2$ increases in magnitude, the average treatment effects should become quite different.

It is then expected that SSE will follow the same trend as $PATE^E - PATE^I$ and increase as correlation increases. This is confirmed by Figure 3. Furthermore, SSE increases not only with correlation, but also as exclusion increases, as SSE is the difference in the subpopulation average treatment effect scaled by the relative size of the excluded subpopulation.

These results confirm that the presence of a non-constant treatment effect would introduce a bias when a subpopulation average treatment effect is used to estimate the total population PATE and the exclusion criterion is correlated with a treatment moderator.
Figure 1: Population values for the PATE for the excluded and the included subgroups of the population. The red line represents population values when the exclusion is 25%. The green line represents results when the exclusion is 50%. The blue line represents results when the exclusion is 75%.

Figure 2: The population differences in the values of average treatment effect between the excluded and included subgroups, denoted by $\Delta^E - \Delta^I$. The red line represents population values when the exclusion is 25%. The green line represents results when the exclusion is 50%. The blue line represents results when the exclusion is 75%.

6.2 Simulation Results for Estimation of $\Delta^E$ and $\Delta^I$

In this section we evaluate the ability of each subsample based estimator to estimate the appropriate subpopulation quantity, $\Delta^E$ and $\Delta^I$, using the appropriate subsamples.
6.2.1 Bias of Estimators of a Subpopulation Based Average Treatment Effect

The bias for estimating PATE with only the included subgroups, using $\hat{\Delta}_{str}^I$ and $\hat{\Delta}_{dr}^I$, is graphed in Figures 4 and 5, respectively. The estimator $\hat{\Delta}_{str}^I$ is biased, regardless of the specification of the propensity model. When the models are correct (‘correct’), the bias increases slightly when covariate correlation increases. This trend is not altered when the model for the outcome is misspecified (‘regression’). However, when the propensity model is incorrect (‘propensity’ and ‘both’), $\hat{\Delta}_{str}^I$ becomes substantially more biased. The estimator becomes more robust to our propensity model misspecification as the correlation increases between the covariates, since much of the information about the treatment choice in the omitted covariate is also carried by the remaining two covariates.

When the correct models are used for estimation, $\hat{\Delta}_{dr}^I$ is relatively unbiased (~1%) and unbiased for all specified correlation structures. Since $\hat{\Delta}_{dr}^I$ is robust to misspecifications in both the propensity and response models, $\hat{\Delta}_{dr}^I$ is only substantially biased when both models are misspecified (‘both’). Again, this bias decreases with an increase in correlation between the covariates.
6.2.2 Difference in Bias Between the Included and Excluded Subgroups

The differences in the bias between the included and excluded subgroups, 
\[ \left( \hat{\Delta}^E - \Delta^E \right) - \left( \hat{\Delta}^I - \Delta^I \right) \], is an important factor in the estimation of SSE, based on Equation 13. If the bias of the estimators in the two subgroups is relatively equal, then the bias of the estimator of SSE would be small.

Figures 6 and 7 illustrate the trends in the differences of the bias between subgroup estimators. Regardless of model specification, the difference in the bias between subgroups for \( \hat{\Delta}_{str} \) increases as the size of the excluded subgroup increases. Under the correct propensity model, the difference increases in magnitude as the correlation between the covariates increases. The difference in the bias of \( \hat{\Delta}_{str} \) is much larger when the propensity model is
misspecified. This difference is smaller when correlation between the covariates is small. The difference in bias increases as correlation increases, and peaks at moderate correlation. The difference is bias decreases at high correlation. This causes a “U” shape in the graph that is present when the model for the propensity score is incorrect.

The difference in the bias between subgroups for $\hat{\Delta}_{dr}$ is very small (~1%) when only one model is specified incorrectly. This difference remains small over all exclusion rates and correlations. The difference in the bias of $\hat{\Delta}_{dr}$ increases when both the models are wrong. The difference demonstrates a similar “U” pattern when both models are misspecified, where the bias peaks for moderate correlation and decreases for high correlation. Unlike the difference in bias for $\hat{\Delta}_{str}$, the magnitudes of the differences in the bias for $\hat{\Delta}_{dr}$ remain relatively consistent across the three levels of exclusion.

Figure 6: Difference in the bias of $\hat{\Delta}_{str}$ of subpopulation average treatment effect for the included and excluded subgroups.

Figure 7: Difference in the bias of $\hat{\Delta}_{dr}$ of subpopulation average treatment effect for the included and excluded subgroups.
6.3 Simulation Results for Estimation of SSE

6.3.1 Bias of $\hat{\gamma}_{str}$

As illustrated in Figure 8, when the correct propensity model is used, the estimator $\hat{\gamma}_{str}$ is relatively unbiased for small correlation and exclusion. However, it becomes biased as the covariate correlation and rate of exclusion increases. As correlation increases, the bias of $\hat{\gamma}_{str}$ follows the same “U” shaped pattern observed in Figure 6 for the difference in bias in the subpopulation estimates, $PATE^E - PATE^I$. The continuation of this pattern is logical, since SSE is scaled difference between $\hat{\Delta}^E - \hat{\Delta}^I$ scaled by the proportion of the excluded subgroup. When the propensity model is wrong, the estimator is generally more biased and the bias increases in magnitude as the size of the excluded subgroup increases. However, as shown in Figure 8a, when there is high correlation and 25% exclusion, this “U” shaped pattern of the bias allows the estimator with the wrong propensity model to become less biased than the estimator with the right model.

As shown in Figure 9, the estimator $\hat{\gamma}_{dr}$ is relatively unbiased (<1%) when only one of the models is misspecified, and the estimator remains unbiased over all correlations and levels of exclusion. When both the models for the outcome and the propensity score are mis-specified, $\hat{\gamma}_{dr}$ becomes biased. The bias incurred with model misspecification also demonstrates the same “U” shaped pattern, though the bias of $\hat{\gamma}_{dr}$ is generally equal to or smaller in magnitude that the bias of $\hat{\gamma}_{str}$.

![Graphs showing bias of $\hat{\gamma}_{str}$](image)

Figure 8: Bias of $\hat{\gamma}_{str}$. 
6.3.2 Mean Squared Error of $\hat{\gamma}$

We also consider the mean squared error of each estimator. In general, since the samples sizes were so large ($n=10,000$), the MSE $\approx (\text{Bias})^2$, since the Var($\hat{\gamma}$) $\approx 0$, and thus the MSE is small ($<1\%$) when the models are correct and the exclusion is small to moderate. The MSE for $\hat{\gamma}_{str}$ incurs error as level of exclusion increases and as correlation increases, as shown in Figure 10. When the propensity model is not correct, $\hat{\gamma}_{str}$ incurs greater error as the exclusion and correlation increase. When the exclusion level is 25%, the difference in MSE incurred by propensity model misspecification is minimal when compared to the MSE where the model is correct. The increased error that is incurred through the model misspecification follows the same “U” shaped pattern that was discussed in the bias of $\hat{\gamma}$.

When only the model for the outcome or the model for the propensity score is misspecified, the MSE of $\hat{\gamma}_{dr}$ is small ($\sim 1\%$), as shown in Figure 11. When both the model for the propensity score and the model for the outcome are misspecified, $\hat{\gamma}_{dr}$ incurs more error. Under model misspecification, the MSE of $\hat{\gamma}_{dr}$ follows a similar trend where the error has a “U” shaped pattern and the magnitude of the error increases with the level of exclusion. The error incurred by model misspecification is minimal when exclusion is 25%, as seen in Figure 11c.

These results are expected since the large sample size results in small variances. The trends of the MSE indicate that when the propensity model is wrong, $\hat{\gamma}_{str}$ incurs more error. This same trend occurs for $\hat{\gamma}_{dr}$ when both models are misspecified. Generally, $\hat{\gamma}_{str}$ incurs greater error than $\hat{\gamma}_{dr}$ under all model specifications.
6.3.3 Confidence Interval Coverage of $\hat{\gamma}$

Even when the MSE is small, an estimator may perform poorly for practical inference if estimated confidence intervals have incorrect coverage. Simulation results for estimated 95% confidence intervals are shown in Figures 12 and 13.

When the propensity score model is correct, the 95% confidence interval coverage for $\hat{\gamma}_{str}$ is approximately equal to the nominal level at low correlation, though the coverage decreases as both correlation and exclusion increase. When there is high exclusion, coverage of $\hat{\gamma}_{str}$ becomes poor as correlation increases, as shown in Figure 12c. When the propensity score model is incorrect, then the coverage follows the same “U”-shaped pattern that was described in the bias and MSE of $\hat{\gamma}_{str}$. The confidence interval coverage of $\hat{\gamma}_{str}$ when the propensity model is misspecified increases at high correlation, and as shown in Figure 12a, the coverage under propensity model misspecification is closer to the nominal 95% level than coverage without model misspecification when exclusion is low and correlation is high. This same phenomenon was seen in the bias of $\hat{\gamma}_{str}$ (Figure 8a). With this in mind, it seems that the bias is driving the confidence interval coverage of $\hat{\gamma}_{str}$.
The confidence interval coverage for $\hat{\gamma}_{dr}$ is approximately at the nominal level when both the outcome model and the propensity model are correct or when only the propensity model is wrong. This is true for all correlations and exclusions. However, when the regression model is wrong, the confidence interval coverage for $\hat{\gamma}_{dr}$ becomes conservative. It is less conservative with high correlation than with low correlation. The misspecification of the regression model leads to an over estimation of the variance of $\hat{\gamma}_{dr}$, causing the intervals to widen and the coverage to be conservative. When both the models for the outcome and the propensity score are incorrect, the coverage is generally poor and has the same “U”-shaped pattern seen in the bias of $\hat{\gamma}_{dr}$ under model misspecification.

Figure 12: Confidence Interval Coverage of $\hat{\gamma}_{str}$.

Figure 13: Confidence Interval Coverage of $\hat{\gamma}_{dr}$.
7 Discussion

The confidence interval coverage of $\gamma_{dr}$ indicates that the doubly robust estimator, which uses the response and propensity score model, is useful for estimating the bias caused by exclusion when a non-constant effect of treatment is present. The confidence interval coverage of the estimator was at the approximate nominal level when both the response model and the propensity score model are specified correctly and when the model for the propensity score is misspecified. The misspecification of the response model does lead to conservative coverage, due to an overestimation of the variance that causes the confidence intervals to widen. The estimator can have poor coverage if both the models are wrong, depending on the level of exclusion and covariate correlation. This is driven by the bias of $\gamma_{dr}$ when both models are misspecified. Since the estimator is robust to misspecification, only minimal bias and error are incurred when only one model is incorrect.

The confidence interval coverage of $\gamma_{str}$ indicates that the stratified estimator, which relies only on the model for the propensity score, may not be a useful estimator of the SSE. The confidence interval coverage varies from nominal to poor for this estimator, dependent on the size of the excluded subgroup relative to the population and the covariate correlation, even when the propensity model is correct. Since the estimator is biased for moderate and high correlation and for increasing exclusion, the coverage generally decreases. Propensity model misspecification increases the magnitude of the bias incurred by this estimator for most exclusion and correlation scenarios. This leads to generally poor coverage, except when there is high covariate correlation and small exclusion.

When the two estimators are compared, $\gamma_{dr}$ generally outperforms $\gamma_{str}$ for estimating the bias caused by exclusion. Even when model misspecification leads to increases in bias and MSE, the magnitude of the increase in each is smaller for $\gamma_{dr}$ than $\gamma_{str}$. Additionally, because $\gamma_{dr}$ is robust to misspecification of either the response or propensity score model, misspecifying one model does not increase the bias or error incurred, whereas misspecifying only the propensity score model has a substantial effect on the amount of bias and error incurred by $\gamma_{str}$. As such, the estimator $\gamma_{dr}$ appears to be a better estimator of SSE.

This study does have limitations. The results of this study are specific to the simulation that was created for the purposes of this report. This study only considers one model for the response, which is a linear model with three continuous covariates. Changing the distribution of the covariates or including a response model that is non-linear will likely lead to different results. Additionally, we only consider a positive correlation structure between
the covariates. Further studies should consider negatively correlated covariates.

When considering the estimation of an effect of treatment, which is the central question to most clinical research, a RCT is the design most often used and trusted as the random allocation of treatment allows for causal inference to be made. However, since most clinical trials exclude a portion of the population for scientific or ethical reasons, it is unclear if the effect of treatment measured in the sample population is generalizable to the population as a whole or if bias was incurred by excluding members of that population with inconvenient genotypic or phenotypic traits. It would be useful to analyze broadly representative observational data, where each subject is free to elect the experimental factors, to assess the effect of treatment for a population. However, due to the possibility of confounding, observational data does not permit causal inference from naive analyses. However, we have shown that observational data can be useful in estimating sample selection error, even in the presence of confounding. In the future, combining RCT and observational data could be very useful and important for providing greater external validity to the results of the RCT.
References


A Appendix: Simulation Code

All simulations were run in the R statistical software package.

A.1 Creation of the Simulated Population

makepop.fun=function(mu=rep(0,3), sigsq=matrix(c(1,0,0,0,1,0,0,0,1),nrow=3,ncol=3),
                      n=10000000, alpha=c(-1, -1/2, 1/2, 1/2), bet=c(1,1,1,1,1,1), sig.resp=1,excov="X2"){

  #mu is the mean of the covariates X
  #sigsq is the covariance matrix of the covariates X
  #n is the population size
  #alpha is the coefficients for the treatment logistic regression (propensity model)
  #bet is the coefficients for the response model
  #sig.resp is the additive conditional variability in the response model

  #load the necessary libraries
  require(mvtnorm)
  require(gtools)

  #check to be sure everything has the right dimensions
  if(length(mu) != length(alpha)-1) stop("your data matrix is not the same size as your alpha vector")
  if(length(mu) != length(bet)-3) stop("your data matrix is not the same size as your bet vector,
including treatment and one interaction")

  #creating the covariates
  x<-rmvnorm(n,mu,sigm)
  #make reasonable names for them
  dimnames(x)[[2]]=paste("X",1:length(mu), sep="")
#Creating the treatment variables.
#make the logistic regression mean
pit<-alpha[1]+x%*%cbind(alpha[-1])
#assign treatment using this model
t<-rbinom(n,1,inv.logit(pit))
#print out the proportion who get treatment
#print(paste("proportion who get treatment:", mean(t)) )

#consider an interaction term
int<-x[,1]*t

#Creating the Response variable.#
#design matrix -- three x values, one treatment, one interaction between
#x1 and t (therefore bet has 1+5 length)
new.data<-cbind(x,t,int)
#find the mean
resp.mean<-as.vector(bet[1]+new.data%*%cbind(bet[-1]))
#add random noise
resp<-rnorm(n,resp.mean,sig.resp)

#Creating a Counterfactual Response variable.#
new.data.counter<-cbind(x, (1-t), x[,1]*(1-t))
resp.counter.mean<-as.vector(bet[1]+new.data.counter%*%cbind(bet[-1]))
TEi<-(resp.counter.mean - resp.mean)*c(-1)^(t) #TEi is :
  #for those receiving treatment, -(control - treatment)
  #for those receiving control, (treatment - control)

#The true propensity scores#
 scores.true<-as.vector(pit)

#Calculation of PATE for whole population#
 Delta.pop<-PATE.func(new.data, bet=bet)
 Delta.TE.popmean<-mean(TEi)

#new.data<-as.data.frame(cbind(resp,x,t,int,TEi, scores.true, treat.effect, treat.effect.finite))
new.data<-as.data.frame(cbind(resp,x,t,TEi, scores.true, Delta.pop , Delta.TE.popmean))

#The exclusion criteria
#The number corresponds as follows:
#number% have 0, 100-number% have 1.
#according to samp.error, number% are EXCLUDED
new.data$inex.25=ifelse(x[,excov] >= qnorm(0.25), 1, 0) #0.25 are EXCLUDED
new.data$inex.50=ifelse(x[,excov] >= qnorm(0.50), 1, 0) #0.5 are EXCLUDED
new.data$inex.75=ifelse(x[,excov] >= qnorm(0.75), 1, 0) #0.75 are EXCLUDED

#find the subgroup values, as well as the true sample selection error
samp.error.25<samp.error(new.data.samp.error=new.data,inex="inex.25", bet=bet)
samp.error.50<samp.error(new.data.samp.error=new.data,inex="inex.50", bet=bet)
samp.error.75<samp.error(new.data.samp.error=new.data,inex="inex.75", bet=bet)

#compile a data frame to return
# new.data<-data.frame(resp, x, t, int,
# treat.effect, scores.true, 
# new.data$inex.25, matrix(samp.error.25, ncol=ncol(samp.error.25), nrow=n, byrow=TRUE, dimnames=list(NULL, paste(colnames(samp.error.25),".25",sep=""))),
# new.data$inex.50, matrix(samp.error.50, ncol=ncol(samp.error.50), nrow=n, byrow=TRUE, dimnames=list(NULL, paste(colnames(samp.error.50),".50",sep=""))),
# new.data$inex.75, matrix(samp.error.75, ncol=ncol(samp.error.75), nrow=n, byrow=TRUE, dimnames=list(NULL, paste(colnames(samp.error.75),".75",sep=""))))

#compile a data frame to return
new.data<-data.frame(new.data, 
matrix(samp.error.25, ncol=ncol(samp.error.25), nrow=n, byrow=TRUE, dimnames=list(NULL, paste(colnames(samp.error.25),".25",sep=""))),
matrix(samp.error.50, ncol=ncol(samp.error.50), nrow=n, byrow=TRUE, dimnames=list(NULL, paste(colnames(samp.error.50),".50",sep=""))))
A.2 Summary of the Simulated Population

summ.fun<-function(popmatname, names.df){ #NOTE THIS FUNCTION IS NOT PORTABLE!
  popmat<-get(popmatname)
  prop.treat<-mean(popmat[,"t"])
  treat.effect<-unique(popmat[,"treat.effect"])
  treat.effect.finite<-unique(popmat[,"treat.effect.finite"])
  samp.err.25<-unique(popmat[,"sample.error.25"])
  treat.effect.finite.inc.25<-unique(popmat[,"treat.effect.finite.inc.25"])
  treat.effect.finite.exc.25<-unique(popmat[,"treat.effect.finite.exc.25"])
  samp.err.50<-unique(popmat[,"sample.error.50"])
  treat.effect.finite.inc.50<-unique(popmat[,"treat.effect.finite.inc.50"])
  treat.effect.finite.exc.50<-unique(popmat[,"treat.effect.finite.exc.50"])
  samp.err.75<-unique(popmat[,"sample.error.75"])
  treat.effect.finite.inc.75<-unique(popmat[,"treat.effect.finite.inc.75"])
  treat.effect.finite.exc.75<-unique(popmat[,"treat.effect.finite.exc.75"])
  stopifnot(length(treat.effect)==1, length(samp.err.25)==1, length(samp.err.50)==1, length(samp.err.75)==1,
            length(treat.effect.finite)==1,
            length(treat.effect.finite.inc.25)==1, length(treat.effect.finite.exc.25)==1,
            length(treat.effect.finite.inc.50)==1, length(treat.effect.finite.exc.50)==1,
            length(treat.effect.finite.inc.75)==1, length(treat.effect.finite.exc.75)==1)

  summ.mat<-cbind(prop.treat, treat.effect, treat.effect.finite, samp.err.25,
                   treat.effect.finite.inc.25,treat.effect.finite.exc.25,
                   samp.err.50,treat.effect.finite.inc.50,treat.effect.finite.exc.50,
                   samp.err.75,treat.effect.finite.inc.75,treat.effect.finite.exc.75)

  dimnames(summ.mat)[[2]]<-names.df$english[match(dimnames(summ.mat)[[2]], names.df$variable)]
  dimnames(summ.mat)[[1]]<-popmatname

return(new.data)
}
A.3 Estimation

The following code is used for the estimation of various parameters.

A.3.1 Estimation of PATE

```r
est.function<-function(i=1, new.data, tname="t",
propcovnames=paste("X",1:3, sep=""),
respname="resp", prop.scores="scores.true",
respcovnames=c(paste("X",1:3, sep="")), numstrata=5, usetruePS=FALSE,
unbiasedvar=TRUE, trueTEi="TEi")
{

#i is an id to identify the simulation number
#new.data is the *population* data frame, which contains (in order)
#response, the X covariates, the treatment chosen, treatment-X1 interaction (plus other stuff)
#tname identifies the treatment variable in the new.data dataframe
#propcovnames identifies the names of the predictors for treatment
# in the new.data dataframe to be used in estimating ps
#respname identifies the response variable in the new.data dataframe
#prop.scores is the name of the *
#true* propensity scores (can be NULL)
#respcovnames identifies the names of the predictors for response in the new.data dataframe
#to be used in the regression model
#numstrata defines the number of strata that you want to use in the stratified estimate
#usetruePS identifies whether or not we should use the known or estimated scores
#unbiasedvar identifies whether or not to used unbiased version of variance (s)
#trueTEi is the name of the column that contains the true individual TE for each member of the sample

#FIND CONSTANTS

#ns is the sample size
```
ns<-nrow(new.data)
n.inv<-1/ns

#A NAIVE ESTIMATE

DeltaNaive<-mean(new.data[new.data[,tname]==1,respname]) - mean(new.data[new.data[,tname]==0,respname])
SigmaNaive<-var(new.data[new.data[,tname]==1, respname]) + var(new.data[new.data[,tname]==0,respname])
CI.Naive.lower<-DeltaNaive-qnorm(0.975)*sqrt(SigmaNaive)
CI.Naive.upper<-DeltaNaive+qnorm(0.975)*sqrt(SigmaNaive)

Naive.vals <- c(DeltaNaive, SigmaNaive, CI.Naive.lower, CI.Naive.upper)

#THE PROPENSITY MODEL

#if we should estimate the true propensity scores, then we estimate them
if(!usetruePS){
  #generate the formula:
  propform = paste(tname, "~", paste(propcovnames, collapse=" + "))

  #estimate the parameters of the model:
  sample.logit=glm(formula(propform), data=new.data, family=binomial(link="logit"))

  #The estimated propensity scores of the sample, **OVERWRITTEN** in the data frame
  new.data[,prop.scores]<-predict(sample.logit,newdata=NULL,type="response")
}

#STRATIFIED ESTIMATION

#find PS quantiles, and convert to factor variable
quant<-quantile(new.data[,prop.scores],probs=seq(from=0,to=1,length=numstrata+1),names=TRUE)

#If the sample isn't big enough to have quantiles, don't bother with the stratified estimator
if(length(unique(quant))!=(numstrata+1)){
  StratEstWorked<-0
StratVarWorked<-0

# mean-based quantities
DeltaS<-NA
# strat.delta.vs.TE<-NA
# strat.delta.vs.TE.sq<-NA

# variance-based quantities
SigmaS<-NA
CI.S.lower <- NA
CI.S.upper <- NA
# ci.capture.strat<-NA
}

else {

sample.cut<-cut(new.data[,prop.scores],quant,labels=paste("Q",c(1:numstrata), sep=""), include.lowest=TRUE)

# Get quantile size and group size within each quantile.
quant.size<-tapply(rep(1,nrow(new.data)),sample.cut,sum)
quant.size.grouped<-tapply(rep(1,nrow(new.data)),list(sample.cut,new.data[,tname]),sum)
quant.size.grouped.inv<-1/(tapply(rep(1,nrow(new.data)),list(sample.cut,new.data[,tname]),sum))

### if any group is size zero, don’t bother calculating anything
if(any(is.na(quant.size.grouped)) | any(quant.size.grouped==0)){

StratEstWorked<-0
StratVarWorked<-0

# mean-based quantities
DeltaS<-NA
# strat.delta.vs.TE<-NA
# strat.delta.vs.TE.sq<-NA

# variance-based quantities
SigmaS<-NA
CI.S.lower <- NA
CI.S.upper <- NA
#ci.capture.strat<-NA

} else {  #otherwise we can calculate an estimate, difference and squared difference

StratEstWorked<-1

#find the average response within PS quantiles and groups
Qmeans<-tapply(new.data[,respname],list(sample.cut,new.data[,tname]),mean)

#Stratified estimate of PATE -- using equation 4 of Lunceford and Davidian
DeltaS<-sum(as.vector(quant.size/ns)*diff(t(Qmeans))

#calculate difference and squared difference (compared to the truth) [used to find bias and MSE later]
#strat.delta.vs.TE<-unique(DeltaS-new.data[,treat.effect])
#strat.delta.vs.TE.sq<-unique((DeltaS-new.data[,treat.effect])^2)
#stop if the difference and squared difference are not length 1
#stopifnot(length(strat.delta.vs.TE)==1, length(strat.delta.vs.TE.sq)==1)

###if any group is size zero or one, don’t bother calculating variance
if(any(is.na(quant.size.grouped)) | any(quant.size.grouped<=1)){

StratVarWorked<-0

#variance-based quantities
SigmaS<-NA
CI.S.lower <- NA
CI.S.upper <- NA
#ci.capture.strat<-NA

} else {  #otherwise we can calculate a variance estimate and CI (with coverage)

StratVarWorked<-1


#find the variance -- using equation 29 of Lunceford and Davidian
Qvars<-tapply(new.data[,respname],list(sample.cut,new.data[,tname]),var)

#to use the biased estimator s:
if(!unbiasedvar) Qvars<-Qvars*(quant.size.grouped-1)/quant.size.grouped

#sum the individual $S\bar{E}means^2$
tot.var.within.strat<-sum(quant.size.grouped.inv*Qvars)

#divide by # strata
SigmaS<-(1/numstrata)*tot.var.within.strat

#calculate the confidence limits
CI.S.lower<-DeltaS-qnorm(0.975)*sqrt(SigmaS)
CI.S.upper<-DeltaS+qnorm(0.975)*sqrt(SigmaS)

#determine if the confidence limits capture the "truth"
#ci.capture.strat = ifelse((new.data[1,treat.effect] >= conf.int.strat.lower) &
#(new.data[1,treat.effect] <= conf.int.strat.upper), 1, 0)

S.vals <- c(DeltaS, SigmaS, CI.S.lower, CI.S.upper)
S.worked<-cbind(StratEstWorked, StratVarWorked)

#DOUBLE ROBUST ESTIMATION -- using Lunceford and Davidian equations 9 and 21

#get response and treatment (for ease of programming)
Zi<-new.data[,tname]
Yi<-new.data[,respname]

#Estimate based on equation 9

#first find the regression estimates
#response formula
respform=paste(respname, " ~", paste(respcovnames, collapse=" + "))
#regress response on X for the treated, find mean for all data
sample.regress.1<-lm(formula(respform),data=new.data[new.data[,tname]==1,])
new.data$response.hat.1<-predict(sample.regress.1,newdata=new.data,type="response")

#regress response on X for the untreated, find mean for all data
sample.regress.0<-lm(formula(respform),data=new.data[new.data[,tname]==0,])
new.data$response.hat.0<-predict(sample.regress.0,newdata=new.data,type="response")

#first term (treated term) of equation 9
treated.term<-n.inv*sum((Zi*Yi-(Zi-new.data[,prop.scores])*new.data$response.hat.1)/new.data[,prop.scores])

#second term (untreated term) of equation 9
untreated.term<-n.inv*sum(((1-Zi)*Yi+(Zi-new.data[,prop.scores])*new.data$response.hat.0)/(1-new.data[,prop.scores]))

#put it all together
DeltaDR<-treated.term - untreated.term

#calculate difference and squared difference (compared to the truth) [used to find bias and MSE later]
#dr.delta.vs.TE<-unique(DeltaDR-new.data[,treat.effect])
#dr.delta.vs.TE.sq<-unique((DeltaDR-new.data[,treat.effect])^2)
#stop if the difference and squared difference are not length 1
#stopifnot(length(dr.delta.vs.TE)==1, length(dr.delta.vs.TE.sq)==1)

#Variance - equation 21
I.hat.dr<-((Zi*Yi-new.data$response.hat.1*(Zi-new.data[,prop.scores]))/(new.data[,prop.scores])-
((1-Zi)*Yi)+new.data$response.hat.0*(Zi-new.data[,prop.scores]))/(1-new.data[,prop.scores])-
DeltaDR

SigmaDR<-(n.inv)^2*sum((I.hat.dr)^2)

#calculate the confidence limits
CI.DR.lower<-DeltaDR-qnorm(0.975)*sqrt(SigmaDR)
CI.DR.upper<-DeltaDR+qnorm(0.975)*sqrt(SigmaDR)
# ci.capture.dr = ifelse((new.data[1,treat.effect] >= conf.int.dr.lower) &
(new.data[1,treat.effect] <= conf.int.dr.upper), 1, 0)

DR.vals <- c(DeltaDR, SigmaDR, CI.DR.lower, CI.DR.upper)

estimate.mat<-rbind(Naive.vals, S.vals, DR.vals)
dimnames(estimate.mat)[[2]]<-c("Delta.hat", "Sigma.hat", "CI.lower", "CI.upper")
dimnames(estimate.mat)[[1]]<-c("Naive", "S", "DR")

#########################
# Calculate true sample average TE

SATE<-mean(new.data[,"TEi"])

return(list(estimate.mat=estimate.mat, S.worked=S.worked, SATE=SATE))
}

PATE.func=function(new.data.PATE,bet)
{
#new.data is the data set of interest
#bet is the betas in the true model, last value is interaction, second to last is treatment effect

treat.effect.PATE=bet[length(bet)]*mean(new.data.PATE[,"X1"])+bet[length(bet)-1]

return(treat.effect.PATE)
}

A.3.2 Estimation of Sample Selection Error

samp.error=function(new.data.samp.error,inex="inex.25",n=n, bet=c(1,1,1,1,1,1))
{
#new.data is the population
#inex is the name of the variable that contains a dummy variable
#number% are 0, 100-number% are 1
#inex == 1 means include, inex==0 means exclude
#bet are the betas in the true model

#calculate the percent of 0s (should be close to 100-number%)
pi.ex = 1-mean(new.data.samp.error[,inex])


#calculate the average treatment effect for the INCLUDED (inex==1)
new.data.included <- new.data.samp.error[new.data.samp.error[,inex]==1,]
if(nrow(new.data.included)>0) Delta.inc.pop <- PATE.func(new.data.included, bet=bet) else Delta.inc.pop = NA

#calculate the average treatment effect for the EXCLUDED (inex==0)
new.data.excluded<-new.data.samp.error[new.data.samp.error[,inex]==0,]
if(nrow(new.data.excluded)>0) Delta.exc.pop <- PATE.func(new.data.excluded, bet=bet) else Delta.exc.pop = NA

#calculate the sample selection error = proportion excluded * difference
gamma.pop <- pi.ex*(Delta.exc.pop-Delta.inc.pop)

#also calculate the raw difference in treatment effects
diff.pop <- Delta.exc.pop-Delta.inc.pop

#calculate from the means
Delta.inc.TE.popmean <- mean(new.data.included$TEi)
Delta.exc.TE.popmean <- mean(new.data.excluded$TEi)

diff.popmean <- Delta.exc.TE.popmean - Delta.inc.TE.popmean

Delta.TE.popmean <- unique(new.data.samp.error[,"Delta.TE.popmean"])
stopifnot(length(Delta.TE.popmean)==1)
gamma.popmean <- Delta.TE.popmean - Delta.inc.TE.popmean

return.results <- cbind(pi.ex, Delta.inc.pop, Delta.exc.pop, diff.pop, gamma.pop, Delta.inc.TE.popmean,
Delta.exc.TE.popmean, diff.popmean, gamma.popmean)

return(return.results)