THE USE OF EFFECT SIZE ESTIMATES TO EVALUATE COVARIATE SELECTION, GROUP SEPARATION, AND SENSITIVITY TO HIDDEN BIAS IN PROPENSITY SCORE MATCHING

Forrest C. Lane, B.S., M.S.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

December 2011

APPROVED:

Robin K. Henson, Major Professor
Darrell Hull, Committee Member
Prathiba Natesan, Committee Member
DeMarquis Hayes, Committee Member
Alexander Beaujean, Committee Member
Abbas Tashakkori, Chair of the Department of Educational Psychology
James D. Meernik, Acting Dean of the Toulouse Graduate School
Lane, Forrest C. The use of effect size estimates to evaluate covariate selection, group separation, and sensitivity to hidden bias in propensity score matching. Doctor of Philosophy (Educational Research), December 2011, 106 pp., 15 tables, 3 figures, references, 83 titles.

Covariate quality has been primarily theory driven in propensity score matching with a general adversity to the interpretation of group prediction. However, effect sizes are well supported in the literature and may help to inform the method. Specifically, $I$ index can be used as a measure of effect size in logistic regression to evaluate group prediction. As such, simulation was used to create 35 conditions of $I$, initial bias and sample size to examine statistical differences in (a) post-matching bias reduction and (b) treatment effect sensitivity. The results of this study suggest these conditions do not explain statistical differences in percent bias reduction of treatment likelihood after matching. However, $I$ and sample size do explain statistical differences in treatment effect sensitivity. Treatment effect sensitivity was lower when sample sizes and $I$ increased. However, this relationship was mitigated within smaller sample sizes as $I$ increased above $I = .50$. 
TABLE OF CONTENTS

LIST OF TABLES ............................................................................................................ v
LIST OF FIGURES ......................................................................................................... vii

THE USE OF EFFECT SIZES ESTIMATES TO EVALUATE COVARIATE SELECTION
AND BIAS REDUCTION IN PROPENSITY SCORE MATCHING ................................. 1

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Propensity Score Matching</td>
<td>5</td>
</tr>
<tr>
<td>Covariate Selection and Propensity Score Estimation</td>
<td>6</td>
</tr>
<tr>
<td>Propensity Score Matching and Diagnostics</td>
<td>7</td>
</tr>
<tr>
<td>Assumption of Strongly Ignorable Treatment Assignment</td>
<td>8</td>
</tr>
<tr>
<td>Sensitivity Analysis as a Post-Hoc Analysis of Hidden Bias</td>
<td>9</td>
</tr>
<tr>
<td>Propensity Score Matching Limitations and Considerations</td>
<td>10</td>
</tr>
<tr>
<td>Effect Size Estimates</td>
<td>11</td>
</tr>
<tr>
<td>Research Questions</td>
<td>14</td>
</tr>
<tr>
<td>Methodology</td>
<td>15</td>
</tr>
<tr>
<td>Data Simulation and Parameter Specification</td>
<td>16</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>19</td>
</tr>
<tr>
<td>Results</td>
<td>21</td>
</tr>
<tr>
<td>Simulation</td>
<td>21</td>
</tr>
<tr>
<td>Descriptives of the / Index and Average Treatment Effects across Samples</td>
<td>22</td>
</tr>
<tr>
<td>Percent Reduction in Propensity Score Separation after Matching</td>
<td>23</td>
</tr>
<tr>
<td>Sensitivity Analysis</td>
<td>28</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Simulated Conditions in the Population</td>
<td>15</td>
</tr>
<tr>
<td>2.</td>
<td>Covariate Weights and Simulated Models in the Estimation of Propensity Scores.</td>
<td>18</td>
</tr>
<tr>
<td>3.</td>
<td>Initial Bias Estimates and Percent Overlap in Propensity Scores within Simulated Conditions of I</td>
<td>19</td>
</tr>
<tr>
<td>5.</td>
<td>Sample Means, Standard Deviations and Mean Differences from the Population of the I Index for Conditions of I Index and Sample Size</td>
<td>25</td>
</tr>
<tr>
<td>6.</td>
<td>Sample Means, Standard Deviations and Mean Differences from the Population Average Treatment Effect ($\tau^p = .50$) for Conditions of I Index and Sample Size</td>
<td>26</td>
</tr>
<tr>
<td>7.</td>
<td>ANOVA Summary Table of Differences in Percent Reduction in Group Separation for Conditions of I and Sample Size</td>
<td>27</td>
</tr>
<tr>
<td>9.</td>
<td>Means and Standard Deviations of $\Gamma$ for Conditions of I Index and Sample Size</td>
<td>29</td>
</tr>
<tr>
<td>10.</td>
<td>ANOVA Summary Table of Differences in Sensitivity ($\Gamma$) for Conditions of I Index and Sample Size</td>
<td>30</td>
</tr>
<tr>
<td>11.</td>
<td>Homogeneous Subsets from I for Conditions of I</td>
<td>31</td>
</tr>
<tr>
<td>12.</td>
<td>Planned Contrast Summary of I Index for Conditions of I</td>
<td>31</td>
</tr>
</tbody>
</table>
13. Homogenous Subsets for Conditions of Sample Size ........................................ 32
14. Planned Contrast Summary for Conditions of Sample Size............................. 32
15. Mean Sensitivity ($\bar{\Gamma}$) and Mean Standardized Differences Pre- ($\bar{d}_I$) and Post- ($\bar{d}_P$) Matching for Simulated Conditions ................................................................. 39
LIST OF FIGURES

1. Illustration of the propensity score matching process ........................................... 6
2. Mean sensitivity for conditions of sample size and I /index................................. 33
3. Plot of sensitivity and post matching bias among conditions of I index for sample size of 700 ........................................................................................................... 39
Researchers are often interested in examining the impact of various programs and treatment interventions (Stuart & Rubin, 2007). Experimental design has historically been considered the “gold standard” for statistically evaluating these effects (West, 2009). This is a result of the design’s use of random assignment to control for pre-group differences (Johnson & Christensen, 2009). Over the long term, random assignment ensures that systematic differences between groups have been removed and provides an ability to make causal inferences (Cook, 2002; Rubin, 1974).

The ability to make causal inference has become increasingly important for educational researchers given a growing era of accountability, particularly as a result of new legislation such as the No Child Left Behind Act. (Odom, Brantlinger, Gersten, Horner, Thompson, & Harris, 2005; Rudd & Johnson, 2008. As Schneider, Carnoy, Kilpatrick, Schmidt, and Shavelson (2007) suggest,

The concern is fundamentally about having better evidence for making decisions about what programs and practices do or do not work. The need for such evidence leads to causal questions, such as whether particular programs and practices improve student academic achievement, social development, and educational attainment. (p. 1)

In response, the Institute for Education Sciences (IES) now encourages the use of advanced statistical techniques to meet evidence based practices. IES suggests that practices should be rigorous and systematic, using both experimental and quasi-
experimental methods with a particular emphasis on random assignment (Slavin, 2002). In addition, IES has tied substantial fiscal incentives, with as much as $150 million distributed annually to programs that demonstrate evidence-based research (Slavin, 2002; Whitehurst, 2002; U.S. Department of Education, 2003).

The problem is that educational researchers do not always favor large scale experimental design and true randomization (Grunwald & Mayhew, 2008; Mosteller, Light, & Sachs 1996). They can be logistically difficult to implement in schools, come at a high costs to external validity of statistical findings, and facilitate model development that is not reflective of the ways schools work in actuality (Cook, 2002). Perhaps even more disconcerting is that the few studies utilizing experimental design are often done by those outside the field as random assignment tends to go against the intellectional zeitgeist within educational literature (Linblom & Cohen, 1980).

As a result, findings reported may be confounded by systematic group differences based on number of covariates (Rosenbaum & Rubin, 1983). For example, students participating in an advanced curriculum program may do so as a result of familial influences, intrinsic motivation, or academic self-efficacy. However, these conditions of treatment must be accounted for in the research design. This is because counterfactual statements rely on well-defined causal states and an assumption that confounding variables have not contributed to the outcome of the effect (Morgan & Winship, 2007). The absence of this can lead to a biased interpretation of any treatment effects.

Propensity score matching is a statistical technique that controls for self-selection bias and allows for the estimation of average treatment effects (ATE) in non-randomized or quasi-experimental studies (Rosenbaum & Rubin, 1983). As such, it affords an
opportunity to extend causal inference into these designs. Disciplines outside the field of education have been using propensity score matching for nearly two decades with great success (D'Agostino, 1998; Dehejia & Wahba, 2002; Grunwald & Mayhew, 2008; Morgan, 2001; Schafer & Kang, 2008; Schnider, Carnoy, Kilpatrick, & Shavelson, 2007; Shadish, Luellen, & Clark, 2006). As such, the U.S. Department of Education (2003) supports propensity score matching as a method suitable for evidence-based research:

> We believe that such well-matched studies can play a valuable role in education, as they have in medicine and other fields, in establishing ‘possible’ evidence of an intervention’s effectiveness, and thereby generating hypotheses that merit confirmation in randomized controlled trials. (p. 5)

In spite of these calls, propensity score matching remains highly underutilized in the educational research literature. As Cook (2002) suggests, “we lament the dearth of interrupted time-series studies, of regression-discontinuity studies and of nonequivalent control group designs with more than one pretest measurement wave on matched cohort samples” (p. 193). Thoemmes and Kim (2011) report only 111 studies published in the social sciences using propensity score matching. Of these, only 34 of these studies were within the field of education. That being said, the use of propensity score matching in the literature has increased in recent years. However, the total number of studies remain relatively small at about 18-25 articles per year (Thoemmes & Kim, 2011).

The lack of propensity score matching may be a result of the ambiguity and complexity around the evaluation of the method. Many studies report insufficient information regarding covariate selection, matching techniques, and bias reduction (Astin, 2007; Baser, 2006). Even when this information is presented, propensity score
matching can be impacted by hidden bias (e.g., relevant covariates not included in the model) in the estimation of treatment effects (Luellen, Shadish, & Clark, 2005). Sensitivity analysis (Rosenbaum, 2010) can be used to examine the influence of hidden bias but “few applied studies take them into account” (Caliendo & Kopeinig, 2008, p. 57).

This study suggests that effect sizes may help to practically inform covariate selection, group overlap in propensity scores and sensitivity to hidden bias in the model. Effect sizes are well supported in the literature and are increasingly encourage within educational research (Henson, 2006; Vacha-Haase & Thompson, 2004; Wilkinson & APA Task Force on Statistical Inference, 1999) Specifically, the $I$ index (Huberty & Lowman, 2000) is a measure of effect size that can be used in logistic regression to evaluate group overlap between the observed and predicted values of group assignment. As such, $I$ may help educational researchers to better evaluate covariates used in propensity score matching, improving clarity in the method and promoting its use within educational research.

The purpose of this study is to present the theoretical framework for propensity score matching and how $I$ may be used in the estimation of propensity scores to evaluate covariate selection. The $I$ index (Huberty, 1984; Huberty & Lowman, 2000; Natesan & Thompson, 2007), along with differences in sample size and initial group overlap, were examined for their ability to explain differences in (a) reduction of group separation after matching and (b) critical values of hidden bias needed to threaten the interpretation of treatment effects. These results and their implications for theory and practice are discussed.
Propensity Score Matching

The literature on use of propensity scores can be attributed to the seminal work of Rosenbaum and Rubin (1983). Their aim was to address limitations of non-randomization by deriving a mathematical solution to account for group differences as a result of self-selection into a treatment condition. In true randomization, we would expect an equal likelihood or probability of group membership. In experimental designs, participants are randomly assigned and have an equal probability of being in either group (e.g., treatment, control). As such, these groups can be directly compared because systematic differences have been controlled through study’s design. Conversely, quasi-experimental designs are subject to non-random assignment which introduces statistical bias into group comparisons. In other words, the probability of group assignment is unknown and instead must be estimated.

Propensity score matching is a multi-step strategy to equate groups and reduce bias in the model as a result of non-randomization (Figure 1). A propensity score (e) is the expected value of treatment for an individual (i) and defined as the conditional probability (P) of getting treatment (T) given a vector of covariates (X) (Rosenbaum & Rubin, 1983), expressed as,

\[ e_i(X_i) = P(T_i = 1|X_i). \]  

First, theoretically relevant pretreatment variables are used to derive probabilities of group membership, most commonly through use of logistic regression (Guo & Fraser, 2010). Then, predicted probabilities of treatment are used to match participants across groups (Rosenbaum & Rubin, 1983). Once matched, treatments effects can be estimated within the newly matched sample and should be more reflective of the
treatment effect from randomized designs (Austin, 2008; D’Agostino, 1998; Shadish, Luellen, & Clark, 2006). Steps in propensity score matching are discussed below.

**Covariate Selection and Propensity Score Estimation**

In non-randomized studies, group selection can be influenced by covariates that confound interpretation of treatment effects (Luellen, Shadish, & Clark, 2005). Therefore, theoretically relevant covariates grounded in literature and likely to influence group selection should be included in the estimation of propensity scores. Accounting for these covariates provides a more meaningful and statistical approximation of group membership (Rubin & Rosenbaum, 1983; Rubin & Rosenbaum 1984; Rubin, 2001). There are no limits to the number of covariates that can be used in the estimation of propensity scores. Literature suggests that any covariate improving predictability ought to be included in the model (Luellen, Shadish, & Clark, 2005). However, covariates should not influence on the outcome itself.

![Propensity Score Matching Process](Figure 1)

**Step 1a:** Covariate Selection and Propensity Score Estimation
- Logistic Regression
- Classification Trees
- Ensemble Methods

**Step 1b:** Covariate Evaluation
- Theory Driven

**Step 2a:** Matching on the Propensity Score
- Nearest Neighbor
- Calipers
- Mahalanobis Dis.
- Optimal Matching

**Step 2b:** Matching Evaluation
- Bias Reduction
- Stratification on Covariates

**Step 3:** Estimation of the Treatment Effect
- T-test
- Repeated Measures ANOVA

**Step 3b:** Hidden Bias Evaluation
- Sensitivity Analysis

*Figure 1. Illustration of the propensity score matching process.*
Once covariates are identified, probabilities of treatment are estimated using covariates in the model. Logistic regression tends to be the most commonly used method in the literature (Guo & Fraser, 2010) and is the easiest to interpret given the predicted probabilities ($P$) of treatment ($T$) are the propensity scores ($e$) for a given set of covariates ($X$).

$$e_i (X_i) = P(T_i = 1|X_i) = \frac{1}{1+e^{-b_i}}$$  \hspace{1cm} (2)

However, propensity scores may also be estimated through classification trees or ensemble methods such as bagging or boosted regression trees (Austin, 2008; Shadish, Luellen, & Clark, 2006).

**Propensity Score Matching and Diagnostics**

The aim of propensity score matching is then to produce groups that share approximately the same probability of group assignment through matching (Rosenbaum & Rubin, 1984). Some of the more commonly used matching techniques include the use of greedy, optimal and fine balance matching. Greedy matching selects participants from the treatment group and matches them, one at a time, to participants in the control group based on (a) the nearest possible propensity score, (b) a pre-specified caliper, (c) Mahalanobis distance or (d) Mahalanobis distance based on the average of the variances within treatment groups. Optimal matching considers the impact of earlier matches on subsequent ones and allows for participants to be matched across treatment groups more than once (Guo & Fraser, 2010). Lastly, fine balancing is a technique which does not require individually matching on the propensity score but uses this score to balance participants instead on some meaningful nominal variable.

Regardless of the technique used match participants, Rubin (2001) suggests assessing the balance between groups pre and post matching. Balance refers to the
equality of propensity scores between groups. This can be accomplished by examining the standardized difference in the mean propensity score between the two groups before and after matching. This difference should be at or near 0 after matching. Additionally, the ratio of the variances of propensity scores and covariates between the two groups should be approximately equal to one. Shadish and Steiner (2010) "suggest ratios between 0.80 and 1.25 are desirable and those smaller than 0.50 or greater than 2.0 are far too extreme" (p. 22).

Stratification or sub classification may also be used to assess the magnitude and statistical significance of the predicted probabilities of group assignment. This can be accomplished by stratifying across quintiles and testing covariates in a 2 x 5 (treatment group by quintile) ANOVA (Rosenbaum & Rubin, 1984). Groups are assumed to be balanced when $F$-values for the main treatment and interaction effect are small and there are no statistically significant interaction effects (Rubin, 2001). Stratification across quintiles has been shown to reduce approximately 90% of bias due to covariates (Rubin & Rosenbaum, 1983; Rubin & Rosenbaum, 1984; Shadish, Luellen, & Clark, 2005). Once the model is balanced, treatments effects are estimated in the newly matched sample on the outcome variable(s) through a $t$-test or multi-group equivalent.

Assumption of Strongly Ignorable Treatment Assignment

Propensity score matching assumes that once groups are balanced, systematic differences in group assignment due to observable covariates have been removed. This allows causal inference to be extended into any treatment effects. However, "only when the strongly ignorable treatment assumption is met does propensity score methodology produce approximate unbiased treatment effect estimates" (Yanovitzky, Zanutto, &
Hornik, 2005, p. 211). Therefore this assumption should be tested prior to interpreting results from propensity score matching.

Strongly ignorable treatment assignment is conceptually similar to an assumption in ordinary least squares (OLS) regression regarding the independence of predictor variables. When this assumption is false, spurious relationships have been ignored in the model and suggest group assignment is confounded within those relationships. Those confounders are also known as hidden bias or the information not recorded in the analysis that may threaten the interpretation of treatment effects (Guo & Fraser, 2010). For example, two participants, k and l, measured on the same covariates (x), should have the same probability (P) of group assignment. When true, the ratio of the probability for group assignment relative to non-group assignment should be close to one. If false, probability of group assignment differs by a multiplier or factor of $\Gamma$,

$$\frac{1}{\Gamma} \leq \frac{P_k/(1-P_k)}{P_l/(1-P_l)} \leq \Gamma \text{ whenever } x_k = x_l.$$  (3)

Odds ratios greater than one suggest the level of hidden bias present in the propensity score matching model.

**Sensitivity Analysis as a Post-Hoc Analysis of Hidden Bias**

The problem in non-randomized studies is that the true probability of group assignment (P) is not known and instead an estimate of the true value. Rosenbaum (2010) suggests a Wilcoxon signed rank test may be used to statistically test the impact of various levels of $\Gamma$ on the interpretation of the treatment effect (i.e., sensitivity analysis). This test has been shown to be a useful tool in evaluating the assumption of strongly ignorable treatment assignment (Austin, 2008; Caliendo & Kopeinig, 2008; Fraser & Guo, 2010; Rosenbaum, 2010; Yanovistzky, Zanutto, & Hornik, 2005).
In the Wilcoxon signed rank test, a confidence interval is computed around the statistical probability ($\rho$) of the treatment effect for various levels of hidden bias ($\Gamma'$). This allows the researcher to examine the impact of various levels of $\Gamma'$ on the interpretation of treatment effects. There is greater confidence in treatment effect when larger values of $\Gamma'$ result in confidence intervals ($\rho_{\text{min}}, \rho_{\text{max}}$) of the treatment effect less than the specified $\alpha$ (i.e., 95% CI $\Gamma' = [.001, .025]$). This interpretation becomes less certain as smaller levels of $\Gamma'$ are required to move the upper bound of the confidence interval above specified $\alpha$ (e.g. 95% CI $\Gamma' = [.01, .20]$).

**Propensity Score Matching Limitations and Considerations**

Several issues should be considered in propensity score matching with respect to bias reduction and sensitivity analysis. First, matching may be influenced in part by the degree of overlap (e.g., initial bias) in the probability of group assignment (Dehejia & Wahba, 2002). Greater shared area in group propensity score distributions may increase the likelihood of obtaining exact matches. Conversely, less shared area can increase matching difficulty as larger distance measures may be necessary to equate groups. For example, a standardized difference (bias) of $d = .5$ results in approximately 67% shared overlap where as $d = .8$ results in only 52.6% (Henson, 2006).

Additionally, the problem of overlap may also be impacted by the sample size in the study. Although matching is still possible within a smaller matching pool, there is an increase likelihood of bias between the groups and uncertainty about the influence of error in the model. Schafer and Kang (2008) noted:

If the samples are far apart in the sense that few treated individuals have covariate values resembling those of untreated individuals or vice-versa, the estimate of either surface near $E(X_i)$ may require extrapolation. (p. 289)
As such, both sample size and initial bias should be considered in the context of sensitivity analysis.

Holding sample size and initial bias constant, information on the relative contribution of a set of covariates to predict treatment (e.g., effect sizes) may help inform researchers about the level of bias reduction and hidden bias earlier in matching process. This is because effect size and error of prediction are related to one another. As the ability of a set of covariates to predict treatment goes up, the contribution of residual error in prediction must go down. In this way, effect sizes theoretically share a relationship to hidden bias.

*Effect Size Estimates*

Effect sizes are defined here as any “statistic that quantifies the degree to which sample results diverge from the expectations (e.g., no difference in group medians, no relationship between two variables) specified in the null hypothesis” (Vacha-Hasse & Thompson, 2004, p. 473). Effect sizes capture how sample results diverge from null expectations and this divergence can be useful in interpreting results (Henson, 2006; Synder & Lawson, 1993; Vacha-Haase & Thompson, 2004). Specifically, they allow the researcher to examine the relative contribution of a set of variables to explain changes in model prediction.

A variety of effect size measures are available (Henson, 2006; Vacha-Hasse & Thompson, 2004). Cohen’s $d$ is the most commonly reported effect size in propensity score matching and is used to assess balance between groups. However, the use of Cohen’s $d$ is generally intended to evaluate the adequacy of matches and not the contribution of covariates in the model. Although one may influence the other, they assess distinctly different parts of the analysis. Pearson $\chi^2$ goodness of fit, Hosmer-
Lemeshow goodness-of-fit test and pseudo $R^2$ have also been reported in propensity score estimation (Caliendo, 2006; Campbell, Nayga, Silvia, & Park, 2009). However, these are usually reported only as a matter of transparency to the reader with little interpretation relative to prediction. Guo and Fraser (2010) suggest “none of these statistics indicates whether the estimated propensity scores are representative of the true propensity scores” (p. 137). Lastly, the classification rate may also be used as it provides a measure of the logistic regression model’s ability to predict group assignment. However, literature suggests that the “main purpose of the propensity score estimation is not to predict selection into treatment as good as possible but to balance all covariates” (Caliendo & Kopeinig, 2008, p. 39). Focusing only on the classification rate may not effectively ensure covariate balance (Shadish & Steiner, 2010).

While the arguments against classification rate may have merit, they confuse (a) the practical problems associated with matching with (b) the contribution of covariates to the quality of those matches. Covariate balance is important in propensity score matching. However, a propensity score model that enables covariate balancing but does not predict treatment assignment would seem to contradict the purpose of using predictive analyses in the estimation of propensity scores. A better argument against the use of classification rates may be their inability to reflect the change or improvement in the classification rate relative to the null.

An easily estimated and interpretable effect size measure which captures this change relative to the null for a set of covariates is the improvement-over-chance index (Huberty & Holmes, 1983; Huberty & Lowman, 2000). This index ($I$) is a ratio of
improvement in the model from the null ($H_O - H_E$) given the maximum improvement that could have been obtained by chance ($1 - H_E$),

$$I = \frac{[(1-H_E)-(1-H_O)]}{1-H_e} = \frac{H_O-H_E}{1-H_E}$$  \hspace{1cm} (4)

Improvement-over-chance values ranges from 0 to 1 with larger values representing greater predictability of group assignment in the model. Larger values of $I$ also represent less group separation in the predicted probability of group assignment from the observed probability. For example, a null model with no covariates and equal group sizes would accurately predict group assignment 50% of the time, resulting in a classification rate of 50% or $I = .00$. As theoretically driven covariates are added to the logistic regression model, the classification rate should increase and is then compared to the maximum improvement possible. For example, a classification rate of 75% relative to a null model classification rate of 50% yields an effect size of $I = .50$ and would suggest half of the model’s ability to predict treatment beyond the rate expected by chance is attributable to covariates in the model.

Although a value of $I = .50$ may be a reasonable effect in educational research, there is an equal amount of unexplained variability in group assignment. If $I$ improves to values of .75 or .90, the logistic regression model better predicts group assignment, has less error due to observed covariates and should be less susceptible to hidden bias. As a result, the level of $\Gamma$ needed to threaten the interpretation of results should increase. In this way, effect size estimates may help to inform both covariate selection and sensitivity to hidden bias earlier in the propensity score estimation process.
Research Questions

Previous studies have examined various methods of propensity score matching (e.g., ensemble, random forest, stratification) in the literature (Berk, Li, & Hickman, 2005; Shadish, Luellen, & Clark, 2005, 2006; McCaffrey, Ridgeway, & Morral, 2004). Two have used simulation techniques in their approach (Akers, 2010; Luellen, 2007). However, all focus on specific bias reduction techniques which ignore the effect of group prediction from covariate selection and the potential hidden bias present in propensity score after matching. Examining this relationship may provide a practical approach to evaluating results earlier and across multiple levels of the analysis.

The purpose of this study is to build on previous literature by examining the relationship of effect size to changes in propensity score overlap and sensitivity analysis. Simulation will be used to examine differences in (a) propensity score group overlap after matching and (b) treatment effect sensitivity as a result of differences in $I$, sample size and initial group bias. The following questions will be used to guide the analysis:

1. Do $I$ index, sample size, and initial group overlap in the propensity score explain statistical differences in changes to group overlap after matching?
2. Do $I$ index, sample size, and initial group overlap in the propensity score explain statistical differences in treatment effect sensitivity?
3. Under what conditions of $I$, sample size, and initial group overlap are these differences explained?
Methodology

Participant data were simulated under various conditions for propensity score matching and sensitivity analyses (Table 1). The procedures used in the simulation of data were guided by previous studies in the literature (Skidmore & Thompson, 2011; Taylor, West, & Aiken, 2006). All simulations and statistical analyses were conducted in the free statistical package R (version 2.13.1) using simulation syntax discussed by Jones, Maillardet, and Robinson (2009). Once conditions had been simulated in the data, random samples of 1,000 were drawn from each population of 1 million to conduct propensity score matching and sensitivity analyses. Finally, statistical differences in (a) percent change in the propensity score overlap between groups after matching and (b) the treatment effect sensitivity were examined using ANOVA under varying conditions of $I$, sample size, and initial group overlap. Each of these conditions were selected given the purported value of effect sizes in the literature (Henson, 2006; Synder & Lawson, 1993; Vacha-Haase & Thompson, 2004) and the potential impact of sample size and initial group overlap in propensity score matching (Dehajia & Wahba, 2002; Schafer & Kang, 2008).

Table 1

*Simulated Conditions in the Population*

<table>
<thead>
<tr>
<th>Manipulated Independent Variables</th>
<th>Simulated Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I$ Index</td>
<td>.00, .25, .50, .75, 1.00</td>
</tr>
<tr>
<td>Initial Group Bias $(d)$</td>
<td>.25, .50, 1.00</td>
</tr>
</tbody>
</table>
Data Simulation and Parameter Specification

Step 1: Simulation of populations with continuous predictors of group assignment and a dichotomous outcome variable. First, populations of 1 million cases were simulated with three normally distributed covariates (X1, X2, X3) and one dichotomous outcome variable (Y). Shadish and Steiner (2010) suggest all variables “plausibly related to treatment and potential outcomes should be in the initial propensity score model” (p. 21). However, effect size ($I$) can be simulated independently of the number of covariates used in the estimation of propensity scores. Therefore, only three covariates were included in this study. Each covariate was specified to be univariate normal using the rnorm function in R with no correlation between covariates. The dichotomous outcome variable (Y) was specified to be a log linear function of the covariates in the analysis.

Step 2: Simulation of the Average Treatment Effect. Within each population, one normally distributed treatment variable (T) was also included. Differences in treatment effects ($\tau$) between groups (Y) are possible within the literature. However, there is little guidance regarding this effect on group overlap after matching or sensitivity analysis. Therefore, several meta-analytic studies were examined among non-randomized designs within the field of education. Weighted average effect sizes were found to range between .25 and .82 (Bernard, Abrami, Borokhovski, Wade, Tamin, Surkes, & Bethel, 2009; Graham & Perin 2007). Therefore, the treatment effect of $\tau = .50$ was specified as a constant to ensure sensitivity analyses were not confounded among differences in treatment effects. This effect was specified as the standardized mean difference in treatment scores (T) and achieved by using different multiplicative weights for both groups (Y) in the study.
\[
\tau = \frac{|\bar{\tau}_0 - \bar{\tau}_1|}{s_r}
\] (5)

**Step 3: Simulation of I.** Huberty and Lowman (2000) provide some guidance for interpretation of the magnitude of \(I\) using Cohen (1988). However, there are no guidelines in propensity score literature regarding the magnitude of the effect from treatment prediction to obtain adequate group overlap in the propensity scores. Therefore, a null model with no effect beyond prediction by chance (\(I = .00\)) was specified first as a basis for meaningful comparison. Increasing values of \(I = .25\) were then specified up to perfect prediction of treatment group classification (\(I = 1.00\)) to evaluate the change in propensity score overlap and treatment effect sensitivity across a range of possible effect sizes (Table 2).

Differences in each condition of \(I\) were generated using variable weights (Table 2). These weights were generated through a trial and error process. However, the proportional weight of each covariate in the model relative to other covariates was maintained for each condition. For example, \(X_1\) was specified such that it contributed twice as much to the propensity score model as \(X_2\).

A third covariate \((X_3)\) was assigned a zero weight but was included as a covariate related to the treatment outcome. Some have suggested it may be appropriate to disregard such covariates (Augurzky & Schmidt, 2001; Bryson, Dorsett, & Purdon, 2002). However, others suggest these covariates should be retained unless shown to be unrelated to the outcome (Rubin & Thomas, 1996). As such, \(X_3\) was included in the model given the lack of clarity in the literature.
Following the specification of $I$, differences in the initial group overlap were simulated in the propensity scores. However, several problems were encountered when simulating this condition. First, $I$ was positively and statistically correlated ($\alpha < .05$) with initial group overlap ($r_s = .944$). Given this relationship, $I$ and initial bias could be manipulated independently. As such, a decision was made to focus on the primary purpose of exploring if differences in effect size (i.e., $I$) with respect to group overlap after matching and treatment effect sensitivity. Therefore, only conditions $I$ and sample size were simulated. Secondly, there was an increased separation in the predicted probabilities of group membership as prediction ($I$) of group assignment improved. This phenomenon is discussed as “perfect separation” in logistic regression literature (Heinze & Schumper, 2002; Menard, 2002; O’Connell & Gray, 2011; Rice, 1994) and occurs when covariates used in the analysis result in near perfect or perfect prediction. In other words, as prediction of assignment into control and treatment groups increases, the means of predicted probabilities within each group move towards 0 or 1. As a result, the ability to match is impeded by a decreased overlap in the propensity for treatment and increased bias in the mean propensity scores.
Using Cohen’s (1988) non-overlap distribution estimates as a comparison to the levels of initial group overlap recorded for conditions of $I$ in this study, overlap decreased considerably as $I$ improved (Table 3). For example, the shared overlap in the distributions of the propensity scores was 33% when $\bar{I}P = .25$. Shared overlap decreased even further to 4% when $\bar{I}P = .50$ and less than 1% at $\bar{I}P = .75$. This resulted in a complete inability to match at higher levels of $I$, particularly in cases of near perfect or perfect prediction (e.g., $\bar{I}P = 1.00$). As a result, matching was only possible up to $\bar{I}P = .90$ and became the highest condition of $I$ used in the study.

Table 3

<table>
<thead>
<tr>
<th>Initial Bias Estimates and Percent Overlap in Propensity Scores within Simulated Conditions of $I$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pop. $I$ Index</td>
</tr>
<tr>
<td>$\bar{I}P = .00$</td>
</tr>
<tr>
<td>$\bar{I}P = .25$</td>
</tr>
<tr>
<td>$\bar{I}P = .50$</td>
</tr>
<tr>
<td>$\bar{I}P = .75$</td>
</tr>
<tr>
<td>$\bar{I}P = 1.00$</td>
</tr>
</tbody>
</table>

Step 4: Simulation of Conditions of Sample Size. Once each of the five populations were generated for conditions of $I$ up to of $\bar{I}P = .90$, conditions of sample size were simulated in the study. Guidance on propensity score matching is that larger samples are generally better (Luellen, Shadish, & Clark, 2005; Yanovitzky, Zanutto, & Hornik, 2005). Although such guidance is true for most statistics, the study replicated a range of possible sample sizes generally found across studies using propensity score matching to provide more detailed guidance. This review resulted in a variety of small $N < 100$ (Stuart, 2007), moderate $N = 200-400$ (Grunwald & Mayhew, 2008; Steiner, Cook, Shadish, Clark, 2009) and large scale multi-institutional studies $N > 1000$ (Yanovitzky,
Zanutto, & Hornik, 2005; Titus, 2006). Given this range, sample sizes of \( N = 100, 400, 700, 1000, 1500, 2000, \text{ and } 2500 \) were simulated in the study.

**Data Analysis**

For each condition, propensity scores were calculated using logistic regression and all three covariates (\( X_1, X_2, X_3 \)) in the model. Groups were matched with replacement on their propensity scores using the Matching program in R (Sekhon, 2011) and a distance caliper of .25. The standardized mean differences in the propensity scores between groups (i.e., group separation) were then saved pre- and post-matching using a loop function in R. The percent reduction in standardized mean differences after matching served as the first dependent variable in the analysis.

Following propensity score matching, sensitivity analysis was conducted on the treatment effect of matched samples using the rbounds package in R (Keele, 2010). This package is based on the Wilcoxon’s signed rank test suggested by Rosenbaum (2002) for sensitivity analysis in propensity score matching. Values of hidden bias (\( \Gamma' \)) were specified within a range of 1-2 (0.1 increments) in order to obtain confidence intervals on the treatment effect using guidance in the literature (Keele, 2010). The smallest value of \( \Gamma' \) needed to move the upper limit of confidence intervals above \( \alpha = .05 \) was recorded as a point estimate of the hidden bias needed to threaten treatment effects and used as the second dependent variable in the analysis.

The range of \( \Gamma' \) was generally adequate for sample sizes drawn from the population with a zero \( I \) (\( \bar{I}^P = .00 \)). However, a wider range of \( \Gamma' \) was necessary to obtain non zero \( p \) values in the upper bound of confidence intervals as \( I \) increased (\( \bar{I}^P \geq .00 \)). Therefore, \( \Gamma' \) was adjusted to 6.0 for all conditions in the simulation syntax (Appendix A).
As such, the upper bounds of the confidence intervals associate with increments of $\Gamma$ were evaluated over a range of 60 different points rather than 10.

After the simulation was conducted and both matching and sensitivity analyses performed, two separate 2-way (5 x 7) Analysis of Variance (ANOVA) were conducted to examine group differences (sample size, $I$) on both dependent variables. Homegeneity of variance was examined prior to the analysis and revealed no statistical differences ($\alpha < .05$). Both main effects and cross-level interactions were examined, considering both $p$ values and effect sizes in the interpretation of effects (Vacha-Haase & Thompson, 2004; Wilkinson & APA Task Force on Statistical Inference, 1999). Specifically, $\eta^2$ was examined because the interpretation of this value is similar to $r^2$. However, $\omega^2$ was also reported as this value has been shown to temper statistical conclusions when there is considerable statistical power (Hinkle, Wiersma, Jurs, 2003).

Results

Simulation

Analyses were repeated 1,000 times for each condition in the study. This resulted in 35,000 simulations of the percent change in propensity score overlap and critical values of $\Gamma$ through sensitivity analyses. Propensity score matching and sensitivity analysis took approximately 30 minutes for each 1,000 replications of smaller sample sizes and up to 90 minutes for larger sample sizes. The replication of all conditions took approximately 24 hours to complete using an Intel Core i3 computer with a 2.13 GHz processor and 4GM of RAM.

Matching generally worked well across larger samples (i.e., > 400) and smaller effect sizes ($\bar{\Gamma}_p < .75$). However, only one matched pair would be found when sample
sizes of 100 were drawn from conditions of $\bar{I}^P \geq .75$. This was problematic as at least two matches are necessary to calculate mean group overlap and conduct sensitivity analyses. As such, it was necessary to relax the matching caliper to a level of 3.0 for sample sizes of 100 drawn from conditions of $\bar{I}^P = .75$ and $\bar{I}^P = .90$. The change in caliper may have provided justification to remove these specific conditions from the analysis. However, these conditions were included given the use of small sample sizes within propensity score matching literature.

Descriptives of $I$ and the Average Treatment Effects across Samples

Within each of the 35 matching conditions, values of $I$ and the average treatment effect ($\tau$) were first examined for their replicability of the population parameters (Table 4-5). The mean $I$ index was generally replicated to within .05 units of the true population parameter for all conditions of sample size (Table 4). Larger sample sizes replicated the true population parameter best and with smaller variability around those mean estimates. Replicated values of $I$ were least accurate across sample sizes of 100, particularly when drawn from populations where $I$ was zero ($\bar{I}^P = .00$). A one-way ANOVA of the replicated values of $I$ for conditions of sample size suggested sample sizes of 100 were statistically different from all other sample sizes $F(6, 34993) = 30.998, p < .001$. The average mean difference within these conditions was approximately .05 of $I$. However, the variance explained by this difference was near zero ($\eta^2 < .01$). Thus, replicated values of $I$ were considered stable across conditions of sample size.

The average treatment effect across all conditions was generally replicated to within .02 units of the true population parameter ($\tau_P = .50$). However, variability around
the mean for each condition increased with larger sample sizes and values of \( I \) (Table 5). A two-way ANOVA on the replicated average treatment effects were conducted to test for statistical differences across conditions of sample size and \( I \). No statistical differences were found between conditions of effect size \( F(4, 34965) = .149, p = .964 \) and sample size \( F(6, 34965) = .282, p = .946 \). There was a statistically significant interaction effect \( F(24, 34965) = 2.111, p = .001 \). However, the variance explained by the average treatment effect in this interaction was near zero \( (\eta^2 < .01) \) suggesting the treatment effect was relatively stable across all conditions.

**Percent Reduction in Propensity Score Separation after Matching**

Given the population parameters were reasonably reproduced for conditions of \( I \) and the average treatment effect, differences in propensity score group overlap after matching were examined. This difference was recorded as the percent reduction in the standardized mean difference of propensity scores between groups after matching (Table 6). This reduction generally ranged between 96-99% of the initial group separation prior to matching. There was a statistically significant (\( \alpha < .05 \)) and negative relationship \( (r_s = -.425) \) between percent reduction in the group separation and \( I \). There was also a statistically significant (\( \alpha < .05 \)) positive relationship \( (r_s = .416) \) between percent reduction in the group separation and sample size. The reduction in propensity score group separation was smallest among sample sizes of 100 when \( I \) values were large \( (\bar{I}^P \geq .75) \). For example, only 19% of the initial separation in the propensity for group assignment was reduced when \( I \) increased to \( \bar{I}^P = .75 \). Percent reduction in the group separation became negative when \( \bar{I}^P = .90 \). This suggested that the percent
reduction in the group separation may be affected by larger levels of $I$, particularly when smaller sample sizes were used in propensity score matching.

Table 4

Means and Standard Deviations of Percent Reduction in Group Separation for Conditions of $I$ and Sample Size ($N = 35,000$)

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>$I_p = .00$</th>
<th>$I_p = .25$</th>
<th>$I_p = .50$</th>
<th>$I_p = .75$</th>
<th>$I_p = .90$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>100</td>
<td>.97</td>
<td>.03</td>
<td>.96</td>
<td>.03</td>
<td>.90</td>
<td>.71</td>
</tr>
<tr>
<td>400</td>
<td>.99</td>
<td>.02</td>
<td>.98</td>
<td>.02</td>
<td>.97</td>
<td>.03</td>
</tr>
<tr>
<td>700</td>
<td>.99</td>
<td>.01</td>
<td>.98</td>
<td>.01</td>
<td>.98</td>
<td>.02</td>
</tr>
<tr>
<td>1000</td>
<td>.99</td>
<td>&lt;.01</td>
<td>.99</td>
<td>&lt;.01</td>
<td>.98</td>
<td>.01</td>
</tr>
<tr>
<td>1500</td>
<td>.99</td>
<td>&lt;.01</td>
<td>.99</td>
<td>&lt;.01</td>
<td>.99</td>
<td>.01</td>
</tr>
<tr>
<td>2000</td>
<td>.99</td>
<td>&lt;.01</td>
<td>.99</td>
<td>&lt;.01</td>
<td>.99</td>
<td>.01</td>
</tr>
<tr>
<td>2500</td>
<td>.99</td>
<td>&lt;.01</td>
<td>.99</td>
<td>&lt;.01</td>
<td>.99</td>
<td>.01</td>
</tr>
<tr>
<td>Total</td>
<td>.99</td>
<td>.02</td>
<td>.99</td>
<td>.02</td>
<td>.97</td>
<td>.27</td>
</tr>
</tbody>
</table>

Note: $N$ refers to the population of all simulation data.

*A caliper of 3.00 was used to obtain matches for sample sizes of 100 with $I$ values of .75 and .90. A caliper of .25 was used for all other conditions.*
Table 5

Sample Means, Standard Deviations and Mean Differences from the Population 1 for Conditions of 1 and Sample Size (N = 35,000)

<table>
<thead>
<tr>
<th>Population Parameters</th>
<th>( I^p = .00 )</th>
<th>( I^p = .25 )</th>
<th>( I^p = .50 )</th>
<th>( I^p = .75 )</th>
<th>( I^p = .90 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>( I^s )</td>
<td>SD</td>
<td>( I^s - I^p )</td>
<td>SD</td>
<td>( I^s - I^p )</td>
</tr>
<tr>
<td>100</td>
<td>.146</td>
<td>.115</td>
<td>.146</td>
<td>.294</td>
<td>.254</td>
</tr>
<tr>
<td>400</td>
<td>.074</td>
<td>.059</td>
<td>.074</td>
<td>.264</td>
<td>.090</td>
</tr>
<tr>
<td>700</td>
<td>.055</td>
<td>.046</td>
<td>.055</td>
<td>.258</td>
<td>.068</td>
</tr>
<tr>
<td>1000</td>
<td>.045</td>
<td>.040</td>
<td>.045</td>
<td>.256</td>
<td>.057</td>
</tr>
<tr>
<td>1500</td>
<td>.039</td>
<td>.032</td>
<td>.039</td>
<td>.257</td>
<td>.047</td>
</tr>
<tr>
<td>2000</td>
<td>.035</td>
<td>.029</td>
<td>.035</td>
<td>.256</td>
<td>.039</td>
</tr>
<tr>
<td>2500</td>
<td>.030</td>
<td>.025</td>
<td>.030</td>
<td>.254</td>
<td>.035</td>
</tr>
<tr>
<td>All Samples</td>
<td>.061</td>
<td>.069</td>
<td>.061</td>
<td>.262</td>
<td>.111</td>
</tr>
</tbody>
</table>

Note: \( N \) refers to the population of all simulation data.
Table 6

Sample Means, Standard Deviations and Mean Differences from the Population Average Treatment Effect ($\tau^P = .50$) for Conditions of $I$ and Sample Size ($N = 35,000$)

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>$\bar{I}^P = .00$</th>
<th>$\bar{I}^P = .25$</th>
<th>$\bar{I}^P = .50$</th>
<th>$\bar{I}^P = .75$</th>
<th>$\bar{I}^P = .90$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>.494</td>
<td>.517</td>
<td>.502</td>
<td>.488</td>
<td>.500</td>
</tr>
<tr>
<td>400</td>
<td>.494</td>
<td>.489</td>
<td>.470</td>
<td>.547</td>
<td>.503</td>
</tr>
<tr>
<td>700</td>
<td>.504</td>
<td>.501</td>
<td>.498</td>
<td>.502</td>
<td>.500</td>
</tr>
<tr>
<td>1000</td>
<td>.502</td>
<td>.504</td>
<td>.501</td>
<td>.501</td>
<td>.500</td>
</tr>
<tr>
<td>1500</td>
<td>.497</td>
<td>.480</td>
<td>.518</td>
<td>.529</td>
<td>.503</td>
</tr>
<tr>
<td>2000</td>
<td>.498</td>
<td>.503</td>
<td>.503</td>
<td>.500</td>
<td>.500</td>
</tr>
<tr>
<td>2500</td>
<td>.502</td>
<td>.506</td>
<td>.503</td>
<td>.503</td>
<td>.500</td>
</tr>
<tr>
<td>All Samples</td>
<td>.499</td>
<td>.500</td>
<td>.503</td>
<td>.505</td>
<td>.500</td>
</tr>
</tbody>
</table>

Note: $N$ refers to the population of all simulation data.
To test the statistical significance of these differences, percent reduction in the group separation of propensity scores across conditions of \( I \) and sample size were examined using a two-way (5x7) ANOVA. \((\alpha = .05)\). Assumptions of homogeneity were examined prior to interpreting results. This test revealed statistical differences between group variances \((F = 4.00, p < .001)\). However, these results can be impacted by the sample size which was relatively large in this study. Furthermore, heterogeneity is generally minimal on Type I error when sample sizes are equal (Hinkle, Weirsma, & Jurs, 2003). Therefore, the results of the two-way ANOVA were interpreted despite statistical differences in group variances (Table 7).

Table 7

ANOVA Summary Table of Differences in Percent Reduction in Group Separation for Conditions of \( I \) and Sample Size \((N = 35,000)\)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>( F )</th>
<th>( p )</th>
<th>( \eta^2 )</th>
<th>( \omega^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( I )</td>
<td>19989.08</td>
<td>4</td>
<td>4997.27</td>
<td>1.09</td>
<td>.36</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Sample Size</td>
<td>34497.89</td>
<td>6</td>
<td>5749.65</td>
<td>1.25</td>
<td>.28</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Interaction</td>
<td>116905.80</td>
<td>24</td>
<td>4871.08</td>
<td>1.06</td>
<td>.38</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Error</td>
<td>160397986.06</td>
<td>34965</td>
<td>4587.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>160569378.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: \( N \) refers to the population of all simulation data.

Differences in the percent reduction of group separation in the propensity scores for conditions \( I \) \((F=1.09 [4, 34965], p = .36, \eta^2<.001)\) and sample size \((F=1.25 [6, 34965], p = .28, \eta^2<.001)\) were not statistically significant \((\alpha = .05)\). Additionally, there was no statistically significant interaction effect \((F=1.06 [24, 34965], p = .38, \eta^2<.001)\). These results were further supported by the lack of variance explained in the model.

Therefore, \( I \) and sample size were interpreted to have little statistical value in explaining differences in the percent reduction in the standardized mean differences in propensity scores.
However, Rubin (2001) also recommends that the standardized difference in the mean propensity score should be at or near 0 after matching ($d < .20$). An examination of those means suggested not all conditions of sample size and $I$ resulted in near zero differences post matching, despite of relatively high reduction to initial group separation (Table 8). For example, standardized mean differences in the propensity scores remained above $d > .20$ for sample sizes of 100 when drawn from $\bar{I}^P \geq .50$. Similar results were also found for all sample sizes drawn from $\bar{I}^P = .90$. Therefore, some conditions of $I$ and sample size may not result in adequately matched groups with respect to their standardized mean differences despite large overall reductions in percent group separation.

Table 8

*Means and Standard Deviations of the Standardized Mean Difference in Propensity Scores Post Matching for Conditions of $I$ and Sample Size ($N = 35,000$)*

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>$I^P = .00$</th>
<th>$I^P = .25$</th>
<th>$I^P = .50$</th>
<th>$I^P = .75$</th>
<th>$I^P = .90$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>$M$</td>
<td>$SD$</td>
<td>$M$</td>
<td>$SD$</td>
</tr>
<tr>
<td>100</td>
<td>.01</td>
<td>.01</td>
<td>.09</td>
<td>.09</td>
<td>.49</td>
<td>4.87</td>
</tr>
<tr>
<td>400</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>.03</td>
<td>.03</td>
<td>.12</td>
<td>.11</td>
</tr>
<tr>
<td>700</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>.02</td>
<td>.02</td>
<td>.08</td>
<td>.08</td>
</tr>
<tr>
<td>1000</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>.01</td>
<td>.01</td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td>1500</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>.01</td>
<td>.01</td>
<td>.04</td>
<td>.04</td>
</tr>
<tr>
<td>2000</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>.01</td>
<td>.01</td>
<td>.03</td>
<td>.03</td>
</tr>
<tr>
<td>2500</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>.02</td>
<td>.05</td>
<td>.12</td>
<td>1.83</td>
</tr>
</tbody>
</table>

Note: $N$ refers to the population of all simulation data.

A caliper of 3.00 was used to obtain matches for sample sizes of 100 with $I$ values of .75 and .90. A caliper of .25 was used for all other conditions.

*Sensitivity Analysis*

Following propensity score matching, sensitivity analyses were also conducted on matched samples in order to examine the impact of $I$ and sample size on the treatment effect sensitivity to hidden bias ($\Gamma$). Means and standard deviations for $\Gamma$ were
found to range between 1.38 and 3.32 units (Table 9) for treatment effects of $\tau_P = .50$.

Larger values of $\Gamma$ were generally associated with larger values of $I$ up to $I_P = .75$. This increase ranged between .27 and .53 units of $\Gamma$. However, $\Gamma$ began to decline as $I$ increased to $I_P = .90$. The Spearman rho correlation coefficient between $I$ and $\Gamma$ was near zero ($r_s = .012$). Additionally, $\Gamma$ was found to increase for larger sample sizes used in propensity score matching. This increase ranged between .02 and .07 units of $\Gamma$ across all levels of $I$. However, $\Gamma$ declined as early as $I_P = .50$ for smaller sample sizes used in the analysis. The relationship between sample size and $\Gamma$ was moderate ($r_s = .227$) and statistically significant ($\alpha < .05$).

Table 9

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>$I^P = .00$</th>
<th>$I^P = .25$</th>
<th>$I^P = .50$</th>
<th>$I^P = .75$</th>
<th>$I^P = .90$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Gamma$</td>
<td>$SD$</td>
<td>$\Gamma$</td>
<td>$SD$</td>
<td>$\Gamma$</td>
<td>$SD$</td>
</tr>
<tr>
<td>100</td>
<td>1.38</td>
<td>.41</td>
<td>2.02</td>
<td>1.35</td>
<td>1.76</td>
<td>1.34</td>
</tr>
<tr>
<td>400</td>
<td>1.71</td>
<td>.32</td>
<td>2.16</td>
<td>1.14</td>
<td>2.66</td>
<td>1.67</td>
</tr>
<tr>
<td>700</td>
<td>1.86</td>
<td>.27</td>
<td>2.16</td>
<td>1.01</td>
<td>2.83</td>
<td>1.71</td>
</tr>
<tr>
<td>1000</td>
<td>1.92</td>
<td>.24</td>
<td>2.16</td>
<td>.87</td>
<td>2.79</td>
<td>1.70</td>
</tr>
<tr>
<td>1500</td>
<td>1.99</td>
<td>.20</td>
<td>2.08</td>
<td>.74</td>
<td>2.88</td>
<td>1.70</td>
</tr>
<tr>
<td>2000</td>
<td>2.03</td>
<td>.17</td>
<td>2.15</td>
<td>.68</td>
<td>2.83</td>
<td>1.63</td>
</tr>
<tr>
<td>2500</td>
<td>2.07</td>
<td>.17</td>
<td>2.15</td>
<td>.57</td>
<td>2.79</td>
<td>1.62</td>
</tr>
<tr>
<td>Total</td>
<td>1.85</td>
<td>.35</td>
<td>2.12</td>
<td>.95</td>
<td>2.65</td>
<td>1.67</td>
</tr>
</tbody>
</table>

Note: $N$ refers to the population of all simulation data

* A caliper of 3.00 was used to obtain matches for sample sizes of 100 with $I$ values of .75 and .90. A caliper of .25 was used for all other conditions.

Given the changes above, values of $\Gamma$ were then tested in two-way (5x7) ANOVA to examine the statistical significance of these differences for conditions of $I$ and sample size. Levene’s test was used to test assumptions of homogeneity of variance ($\alpha = .05$) and revealed statistical differences between group variances ($F = 534.16$, $p < .001$). However, the ANOVA was interpreted despite these differences due to the use of large
sample sizes in this study and minimal impacts on Type 1 error (Hinkle, Weirsm, & Jurs, 2003). Results suggested that the main effects for $\mathcal{I}$ ($F[4, 34965] = 649.78, p < .001, \eta^2 = .066$) and sample size ($F[6, 34965] = 147.89, p < .001, \eta^2 = .023$) were statistically significant (Table 9). Additionally, the interaction effect between $\mathcal{I}$ and sample size was ($F[24, 34965] = 29.37, p < .001, \eta^2 < .018$) was statistically significant. Both $\eta^2$ and $\omega^2$ values were small but suggested that approximately 10.6% of the variance in $\Gamma$ was explained by sample size, $\mathcal{I}$, and their interaction. Given the lack of comparative literature from which to evaluate $\eta^2$ and $\omega^2$ values in this study, the results of this two-way (5x7) ANOVA were interpreted using a Tukey HSD test (Hinkle, Wiersma, & Jurs, 2003).

Table 10

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>$F$</th>
<th>$p$</th>
<th>$\eta^2$</th>
<th>$\omega^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{I}$</td>
<td>5139.43</td>
<td>4</td>
<td>1284.86</td>
<td>649.78</td>
<td>&lt;.001</td>
<td>.066</td>
<td>.066</td>
</tr>
<tr>
<td>Sample Size</td>
<td>1754.68</td>
<td>6</td>
<td>292.44</td>
<td>147.89</td>
<td>&lt;.001</td>
<td>.023</td>
<td>.023</td>
</tr>
<tr>
<td>Interaction</td>
<td>1393.72</td>
<td>24</td>
<td>58.07</td>
<td>29.37</td>
<td>&lt;.001</td>
<td>.018</td>
<td>.017</td>
</tr>
<tr>
<td>Error</td>
<td>69138.93</td>
<td>34965</td>
<td>1.977</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>77426.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: $N$ refers to the population of all simulation data.

Post-hoc results indicated that all levels of $\mathcal{I}$ were statistically different from one another in the study ($\alpha < .05$). Greater levels of $\Gamma$ were generally required to threaten the interpretation of results for $\tau^p = .50$ up to $\mathcal{I}$ values of $\bar{\mathcal{I}}^p = .75$ (Table 11). However, $\bar{\Gamma}$ dropped to below sensitivity levels recorded for samples drawn from $\bar{\mathcal{I}}^p = .50$ when $\bar{\mathcal{I}}^p = .90$. To help inform these results, planned contrasts were also conducted to examine model effects (Table 12). These contrasts suggest most of the effect in this model
appears to come between conditions of no prediction ($\bar{I}^p = .00$) and moderate prediction of group assignment ($\bar{I}^p = .50, .75$).

Table 11

**Homogeneous Subsets from for Conditions of $I$ (n = 7000)**

<table>
<thead>
<tr>
<th>$I$</th>
<th>Subset</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00</td>
<td>1.851221</td>
</tr>
<tr>
<td>.25</td>
<td>2.127421</td>
</tr>
<tr>
<td>.50</td>
<td>2.650307</td>
</tr>
<tr>
<td>.75</td>
<td>2.943879</td>
</tr>
<tr>
<td>.90</td>
<td>2.417571</td>
</tr>
</tbody>
</table>

Note: $n$ refers to the sample size for conditions of $I$.

Table 12

**Planned Contrast Summary of $I$ Index for Conditions of $I$ (n = 7000)**

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>$F$</th>
<th>$p$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>267.00</td>
<td>1</td>
<td>267.00</td>
<td>129.24</td>
<td>&lt;.001</td>
<td>.003</td>
</tr>
<tr>
<td>C2</td>
<td>2038.88</td>
<td>1</td>
<td>2038.88</td>
<td>986.87</td>
<td>&lt;.001</td>
<td>.026</td>
</tr>
<tr>
<td>C3</td>
<td>2830.23</td>
<td>1</td>
<td>2830.23</td>
<td>1369.90</td>
<td>&lt;.001</td>
<td>.037</td>
</tr>
<tr>
<td>C4</td>
<td>3.32</td>
<td>1</td>
<td>3.32</td>
<td>1.606</td>
<td>.205</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Residual</td>
<td>72287.27</td>
<td>34995</td>
<td>2.066</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>77426.71</td>
<td>34999</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: $n$ refers to the sample size for conditions of $I$.

\(^a\)C1: Contrast of $I$ = .00 and $I$ = .25  
\(^b\)C2: Contrast of $I$ = .00 and $I$ = .50  
\(^c\)C3: Contrast of $I$ = .00 and $I$ = .75  
\(^d\)C4: Contrast of $I$ = .00 and $I \geq .25$

With respect to the second main effect, larger values of $I$ were obtained when larger sample sizes were used in the propensity score matching process. As such, $\bar{I}$ for larger sample sizes were statistically different from smaller sample sizes and less sensitivity to hidden bias (Table 12). However, larger sample sizes tended to be more homogenous than smaller sample sizes in terms of $\bar{I}$. For example, there was no statistical difference was observed between sample sizes of 700 and 1,000 or between
1,500 and 2,000. Again, planned contrasts were also conducted to help inform these results (Table 14). The magnitude of these group differences appear to be about the same for all contrasts examined in the study.

Table 13

Homogeneous Subsets for Conditions of Sample Size (n = 5000)

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Subset 1</th>
<th>Subset 2</th>
<th>Subset 3</th>
<th>Subset 4</th>
<th>Subset 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>2.021730</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>400</td>
<td></td>
<td>2.146980</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>700</td>
<td></td>
<td></td>
<td>2.353830</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td></td>
<td></td>
<td>2.433050</td>
<td></td>
</tr>
<tr>
<td>1500</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.555130</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.597040</td>
</tr>
<tr>
<td>2500</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.678800</td>
</tr>
</tbody>
</table>

Note: n refers to the sample size for conditions for conditions of sample size.

Table 14

Planned Contrast Summary for Conditions of Sample Size (n = 5000)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C11</td>
<td>39.22</td>
<td>1</td>
<td>39.22</td>
<td>18.16</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>C12</td>
<td>242.06</td>
<td>1</td>
<td>242.06</td>
<td>112.06</td>
<td>&lt;.001</td>
<td>.003</td>
</tr>
<tr>
<td>C13</td>
<td>251.30</td>
<td>1</td>
<td>251.30</td>
<td>116.34</td>
<td>&lt;.001</td>
<td>.003</td>
</tr>
<tr>
<td>C14</td>
<td>400.01</td>
<td>1</td>
<td>400.01</td>
<td>185.18</td>
<td>&lt;.001</td>
<td>.005</td>
</tr>
<tr>
<td>C15</td>
<td>362.35</td>
<td>1</td>
<td>362.35</td>
<td>167.91</td>
<td>&lt;.001</td>
<td>.005</td>
</tr>
<tr>
<td>C16</td>
<td>459.69</td>
<td>1</td>
<td>459.69</td>
<td>217.57</td>
<td>&lt;.001</td>
<td>.006</td>
</tr>
<tr>
<td>Residual</td>
<td>75672.08</td>
<td>34993</td>
<td>2.16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>77426.71</td>
<td>34999</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: n refers to the sample size for conditions for conditions of sample size

$^a$C1: Contrast of n = 100 and n = 400
$^b$C2: Contrast of n = 100 and n = 700
$^c$C3: Contrast of n = 100 and n = 1000
$^d$C4: Contrast of n = 100 and n = 1500
$^e$C5: Contrast of n = 100 and n = 2000
$^f$C6: Contrast of n ≤ 1000 and n ≥ 1000.

In examining the interaction effect, there was a difference in $\bar{\Gamma}$ for sample sizes of 100 relative to all other sample sizes when $\bar{I}^p = .50$. Specifically, $\bar{\Gamma}$ for sample sizes of
100 was 1.76 and decreased .26 units relative to all other sample sizes when $\bar{\Gamma} = .50$ (Figure 2). In contrast, $\bar{\Gamma}$ of all other sample sizes increased .65 units to $\bar{\Gamma} = 2.79$ for the same level of $\bar{\Gamma}$. There was also a greater decrease in $\bar{\Gamma}$ as smaller sample sizes were used in propensity score matching for $\bar{\Gamma} = .90$. For example, sample sizes of less than 1500 resulted in only a $\bar{\Gamma}$ decrease of -.31 units whereas sample sizes greater than or equal to 1000 decreased -.68 units. This seemed to suggest that $\bar{\Gamma}$ may be more stable for larger samples sizes at higher levels of $\bar{\Gamma}$.

![Figure 2. Mean sensitivity for conditions of sample size and $\bar{I}$.](image)

Discussion

One of the aims of this study was to explore changes in $I$ and initial group overlap relative to the percent reduction is propensity score separation and treatment effect sensitivity. The problem was that intial group overlap and $I$ could not be explored independently because of their positive relationship to one another. Initial group
separation ($d = 2.0$) was large at relatively low levels of $I$ ($\bar{I} = .25$) resulting in a decreased overlap in the distribution of propensity scores. This made matching prohibitive as $I$ neared 1.0 due to issues of near-perfect or perfect separation.

Although the problem of separation should have been identified a priori in the literature review, it is not always discussed by those who utilize this logistic regression. For example, “neither Maddala (1983) nor Long (1997) raise the issue of separation in any form” (Zorn 2005, p. 161). Those that do, discuss it mostly as a “numerical coincidence” and as a problem that must be worked around (Hosmer & Lemeshow, 2000, p. 140). Some have suggested a penalized maximum likelihood correction to the standard binomial GLM score functions as a possible solution (Firth, 2005; Firth 2000; Zorn 2005). However, the impact of this specific correction in propensity score matching is not clear.

Within propensity score matching literature, the discussion on separation is also limited. For example, the potential impact of decreased overlap or a “common support region” is only briefly discussed by Guo and Fraser (2010) and primarily with a purpose of introducing other estimation methods (p. 148). It was not clear from this discussion that overlap was considerably impacted by increased prediction. Others have suggested that separation may appear to make control and treatment groups look quite different when they are actually more comparable in the context of predicting treatment assignment (Hill, Weiss, & Zhai, 2011). As such, several alternative approaches have been suggested (Hill, Weiss, & Zhai, 2011). First, overlap in the propensity score of the two distributions can be tested using both a fake and true treatment variable with the same rate of success. Differences between the two may help evaluate the adequacy of the covariate model for propensity score estimation. Another alternative is to use a
Bayesian approach to generalized linear models which constrains coefficients in the model to within a reasonable range of the propensity score. This restricts predicted probabilities from reaching values of 0 or 1. Lastly, generalized boosted models (GBM) and Bayesian Additive Regression Trees (BART) may also be employed. However, all of these models were still susceptible to overfitting and with few differences in the prediction of treatment assignment.

These alternatives may provide some improvement to separation but their findings should be taken with caution. Each of alternative approaches presented by Hill, Weiss, and Zhai (2011) were modeled using more than 500 covariates in the estimation of propensity scores. This can be problematic as separation is more likely when then number of predictors begins to equal the sample size in the analysis (O’Connell & Gray, 2011). That being said, only three predictors were necessary to create perfect separation in the random samples drawn from conditions of $\hat{IP} > .50$ in this study. Therefore, it may not be reasonable to assume that this problem exists only under a large number of covariates.

It is unclear the extent or likelihood of perfect separation in practice. However, the robustness of matching techniques to issues of separation should be considered if propensity score matching is to be more widely used in the literature, particularly given an assumption that all relevant covariates have been included in the model. If the covariates included in the estimation of treatment assignment are so theoretically driven that they cannot be removed to improve issues of separation, then perhaps matching is inappropriate for selected participants due to real differences in the probability of group assignment.
In spite of issues to manipulate initial group separation, both $I$ and sample size were simulated and explained statistical differences in $\Gamma$. Treatment effects of $\tau^P = .50$ were less sensitive to hidden bias as both the sample size and $I$ increased. On average, treatment effects were .11 units less sensitive to hidden bias for increases in sample size and .36 units less sensitive for positive changes in $I$ up to values of $\bar{I}^P = .75$. Given that $\Gamma$ is the factor or odds by which control and treatment groups would need to differ in their propensity for treatment in order to threaten treatment effects, these findings suggest some conditions result in as much as a 50% decrease in model sensitivity. For example, a sample size of $n = 2500$ required a 50% increase in the level of hidden bias relative to sample size of $n = 100$ for $\bar{I}^P = .00$.

$$\Delta\Gamma = \frac{(\Gamma_{n=2500} - \Gamma_{n=100})}{\Gamma_{n=100}} = \frac{(2.07 - 1.38)}{1.38} = .50$$

When sample size was held constant (e.g., $n = 100$) and $\bar{I}^P$ increased from .00 to .25, a 46% increase in the level of hidden bias was needed to threaten the interpretation of treatment effects. These results seemed notable. However, other possible conditions including differences in propensity score matching techniques may be more robust to hidden bias and model sensitivity.

Values of $\Gamma$ in this study were compared to the literature to examine the validity of findings in this study. Few studies were found having used Rosenbaum’s sensitivity analysis but this was anticipated given the general lack of sensitivity analysis in propensity score matching literature (Caliendo & Kopeinig, 2008). Of those studies identified, $\Gamma$ was generally reported to range between between 1.5 and 2.0 (Haviland, Nagin, & Rosenbaum, 2007; Payne, DiGiuseppe, & Tilahun, 2002). The range of values in this study was slightly larger ($1.38 < \bar{\Gamma} < 3.32$) but also appeared to replicate values of
Another aim of this study was to examine the statistical differences in the reduction of group separation for conditions of \( I \) and sample size. It was believed prior to this study that increases in both of these conditions would improve the accuracy of propensity scores which would then correspond with greater reduction in group separation. That relationship was positive for sample size but negative across conditions of \( I \). However, neither sample size nor \( I \) statistically explained differences in the reduction of group separation. The percent reduction to group separation in the propensity scores remained above 90% for all but one of the conditions examined in this study (of \( \bar{I}^P = .90, n = 100 \)). In addition, the standardized mean differences in propensity scores between group generally remained below \( d \leq .20 \) (Table 12). As such, these findings may suggest the robustness of propensity score matching to reduce separation across conditions of sample sizes and \( I \).

Lastly, the results of this study may indicate a multivariate relationship between \( I \), \( I^r \), and the reduction to initial propensity score group separation. Rosenbaum (2010) suggests that treatment effects should be less sensitive to hidden bias as theoretically relevant covariates have been included in the model. In this study, greater levels of \( I \) were found to correspond to higher levels of \( I^r \) which supports this theory. However, that relationship was found to exist only to the point where the post matching bias was less than \( d \leq .25 \) (Table 13). Even though the propensity score model may have a higher effect size relative to prediction of group assignment (i.e., \( I \)), groups that cannot be adequately equated should more critically evaluated. This finding may suggest these
variables should be considered multivariately. However, that relationship is not clear based on current propensity score matching literature.
Table 15.

*Mean Sensitivity (\( \Gamma \)) and Standardized Mean Differences Pre- (\( \bar{d}_i \)) and Post- (\( \bar{d}_p \)) Matching (N = 35,000)*

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>( \bar{d}_i )</th>
<th>( \bar{d}_p )</th>
<th>( \bar{d}_i )</th>
<th>( \bar{d}_p )</th>
<th>( \bar{d}_i )</th>
<th>( \bar{d}_p )</th>
<th>( \bar{d}_i )</th>
<th>( \bar{d}_p )</th>
<th>( \bar{d}_i )</th>
<th>( \bar{d}_p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1.38</td>
<td>.33</td>
<td>.01</td>
<td>2.00</td>
<td>2.38</td>
<td>.09</td>
<td>1.76</td>
<td>3.95</td>
<td>.49</td>
<td>2.73</td>
</tr>
<tr>
<td>400</td>
<td>1.71</td>
<td>.16</td>
<td>&lt;.01</td>
<td>2.16</td>
<td>2.20</td>
<td>.03</td>
<td>2.66</td>
<td>4.14</td>
<td>.12</td>
<td>2.49</td>
</tr>
<tr>
<td>700</td>
<td>1.86</td>
<td>.12</td>
<td>&lt;.01</td>
<td>2.16</td>
<td>2.16</td>
<td>.02</td>
<td>2.83</td>
<td>4.05</td>
<td>.08</td>
<td>2.90</td>
</tr>
<tr>
<td>1000</td>
<td>1.92</td>
<td>.09</td>
<td>&lt;.01</td>
<td>2.16</td>
<td>2.15</td>
<td>.01</td>
<td>2.79</td>
<td>4.02</td>
<td>.05</td>
<td>2.93</td>
</tr>
<tr>
<td>1500</td>
<td>1.99</td>
<td>.08</td>
<td>&lt;.01</td>
<td>2.08</td>
<td>2.15</td>
<td>&lt;.01</td>
<td>2.88</td>
<td>4.00</td>
<td>.04</td>
<td>3.06</td>
</tr>
<tr>
<td>2000</td>
<td>2.03</td>
<td>.07</td>
<td>&lt;.01</td>
<td>2.15</td>
<td>2.15</td>
<td>&lt;.01</td>
<td>2.83</td>
<td>4.00</td>
<td>.03</td>
<td>3.17</td>
</tr>
<tr>
<td>2500</td>
<td>2.07</td>
<td>.06</td>
<td>&lt;.01</td>
<td>2.15</td>
<td>2.14</td>
<td>&lt;.01</td>
<td>2.79</td>
<td>3.99</td>
<td>.02</td>
<td>3.32</td>
</tr>
</tbody>
</table>

*A caliper of 3.00 was used to obtain matches for sample sizes of 100 with \( I \) values of .75 and .90. A caliper of .25 was used for all other conditions.

*Figure 3. Plot of Sensitivity and Post Matching Bias among Conditions of \( I \) for sample size of 700.*
Conclusions

The purpose of this study was to examine the ability of $I$ and sample size to explain differences in (a) the reduction of group separation after matching and (b) treatment effect sensitivity. Covariate quality has been primarily theory driven in propensity score matching with a general adversity to the interpretation of treatment prediction in the literature (Caliendo & Kopeinig, 2008; Guo & Fraser, 2010). There is no argument that covariate selection should be theory driven, as in any statistical analysis. However, a great deal of literature has been published on the benefits of effect sizes (Henson, 2006; Synder & Lawson, 1993; Vacha-Haase & Thompson, 2004) and the results of this study seem to support that the magnitude of group prediction ($I$) is related to a reduction in model sensitivity to hidden bias ($\Gamma'$). As such, researchers may want to include covariates which increase group prediction or $I$. This may help researchers to more effectively identify covariates that are more robust to hidden bias in propensity score matching models.

Limitations and Recommendations for Future Research

Interpretation of results should be done in the context of limitations when generalizing results to the broader literature. These limitations point to future research that may be helpful in exploring the relationship of $I$ to the reduction of group separation in propensity scores and model sensitivity. First, the use of varied calipers may have confounded the results in this study. As $I$ increased to values of $\bar{I}_P \geq 75$, larger distance calipers were needed to identify propensity score matches between groups. Although this improved matching ability in the study, both the percent reduction in group separation and standardized mean differences between groups remained adequate.
Additionally, covariates used in the estimation of propensity scores were specified as orthogonal to one another in this study. This may not be reflective of the multivariate nature of these variables in practice. A multivariate relationship between covariates should be explored with respect to a reduction in group separation of propensity scores and sensitivity to hidden bias.

The decision to include small sample sizes in propensity score matching (e.g. \( n = 100 \)) may also be problematic. This moved the grand mean of percent reduction in group separation downward. As a result, statistical differences between groups and the variance explained by conditions in the model were likely mitigated. Removing this condition may have increased model performance. However, the practical significance of a 1-2% difference in the reduction of group separation is probably negligible.

Additional conditions of \( I \) may also need to be examined with regard to treatment effect sensitivity, particularly within areas of greater overlap in the distribution of propensity scores. There was very little guidance on this topic as few studies report initial bias. However, the first non-zero \( I \) (\( \bar{I}^p = .25 \)) resulted in a standardized mean difference of \( d = 2.0 \) or approximately 32% shared overlap in propensity scores. This level of initial bias may be too large for some educational and social science research. Examining these effects at smaller levels of overlap in the distributions of propensity scores may be informative, particularly when prediction is small.

Other methods of propensity score estimation have also been suggested and may be advantageous to logistic regression. For example, classification trees, bagging, and boosted regression trees have all been suggested for propensity score estimation (Austin, 2008; Shadish, Luellen, & Clark, 2006). There is no consensus in the literature as to which approaches work best. However, each of these estimation methods were
created to help better inform covariate selection, particularly as treatment effects can be sensitive to model specification. Therefore, the impact of these alternative approaches on percent bias reduction and treatment effect sensitivity should be considered.

Similarly, other matching methods have also been suggested that may help with issues of separation. The literature on optimal matching is relatively new and extends only 10 years (Guo & Fraser, 2010). However, optimal matching was designed to minimize the total distance between propensity scores. This is accomplished by matching participants to one or more participants in other groups (Rosenbaum, 2002). It is not clear if optimal matching would alleviate problems of near perfect or perfect separation. However, the use of caliper matching likely reduced the number of matches between groups in this study.

The use other effect sizes in logistic regression including Pearson $\chi^2$ goodness of fit, Hosmer-Lemeshow goodness-of-fit test, and Pseudo $R^2$ should be explored with respect to their relationship to $\Gamma$. These measures evaluate model fit and available in most statistical software. However, there are some limitations to these measures which should be considered. First, goodness of fit statistics can be sensitive to sample size which may impact model fit. Secondly, pseudo $R^2$ statistics cannot be interpreted independently or compared across data sets (Guo & Fraser, 2010). These statistics have meaning only “when compared to another pseudo R-square of the same type, on the same data, predicting the same outcome” (p. 138). This seemed to support the use of $\Gamma$ as measure of group prediction in this study. However, other effect size measures may be more generalizable than others, particularly within some areas of research.

Finally, the results of this study were limited to treatment effects of $\tau^p = .50$. Again, this treatment effect seemed appropriate given treatment effects generally
reported in educational literature. However, Rosenbaum (2010) suggests $\Gamma$ may increase with higher levels of treatment effect. Given larger treatment effects are possible, simulations of these conditions under varying degrees of $\tau^p$ may be warranted.
APPENDIX

R SYNTAX FOR DATA SIMULATION, PROPENSITY SCORE MATCHING, AND SENSITIVITY ANALYSES
install.packages("QuantPsyc")
install.packages("Matching")
install.packages("rbounds")
library(QuantPsyc)
library(Matching)
library(rbounds)

###Create a vector of 1 dichotomous treatment variable
###1 outcome variable
###and 3 covariates predicting treatment with 1,000,000 scores
###with a mean of 0 and sd of 1.

intercept = 0
beta1 = 0
beta2 = 0
beta3 = 0
x1 = rnorm(1000000,1,1)
x2 = rnorm(1000000,1,1)
x3 = rnorm(1000000,1,1)
Tr = rnorm(1000000,1,1)
linpred = intercept + x1*beta1+x2*beta2+x3*beta3
prob = exp(linpred)/(1 + exp(linpred))
runis = runif(1000000,0,1)
y = ifelse(runis < prob,1,0)

###Creating an ATE of .50 in the population
Tr.1<-ifelse(y==1,Tr*1.25,Tr*.75)
massive<-data.frame (x1=x1, x2=x2, x3=x3, Tr=Tr,Tr.1=Tr.1, y=y)
t.test(Tr.1~y, alternative="two.sided",var.equal=TRUE, conf.level=.95)
T.effect<-cohen.d(Tr.1,y)$CohenD
T.effect

###Propensity scores of all cases in the population
glm.massive<-glm(y~x1+x2+x3, massive, family=binomial(link="logit"))
summary(glm.massive)

###Calculating predicted values from the population.
pred(glm.massive) <- predict(glm.massive,type="response")
### Writing Cohen's d function
### Cohen's d (pooled within group SD)
cohen.d <- function(DV, groupv)
{
  varg1 <- by(DV, groupv, var)
  varg2 <- by(DV, groupv, var)
  ng1 <- by(DV, groupv, length)
  ng2 <- by(DV, groupv, length)
  pooledvar <- ((ng1 - 1) * varg1 + (ng2 - 1) * varg2) / (ng1 + ng2 - 2)
  pooledsd <- sqrt(pooledvar)
  meang1 <- by(DV, groupv, mean)
  meang2 <- by(DV, groupv, mean)
  meandif <- meang1 - meang2
  cohd <- meandif / pooledsd
  varratio <- varg1 / varg2
  list(pooled.SD = pooledsd, Group.Mean.Dif = meandif, CohenD = cohd, VarRatio = varratio)
}

### Calculation of initial bias in propensity scores in the population
Initial.bias <- cohen.d(pred.glm.massive, glm.massive$y)$CohenD

### Calculation of the hit rate given a set of covariates
ClassLog (glm.massive, massive$y)
b <- ClassLog (glm.massive, massive$y)
b$overall
b$mcFadden

### Calculation of the I Index in the population given the improvement over the null hit rate
outcome <- b$rawtab[1, 2] + b$rawtab[2, 2]
initialHR <- (outcome / 1000000)
Improvement <- (b$overall - initialHR)
ChanceImprove <- (1 - initialHR)
lindex <- (Improvement[1] / ChanceImprove[1])
lindex

### Propensity scores of all cases in the population
glm.massive <- glm(y ~ x1 + x2 + x3, data, family = binomial(link = "logit"))

### Test Sample for sample size of 100
matr <- sample(nrow(massive), 100, replace = F)
samp.100 <- data.frame(massive[matr])
```r
### Run the logistic regression on the sample
glm1 <- glm(y ~ x1 + x2 + x3, samp.100, family = binomial(link = "logit"))
summary(glm1)

### Calculating predicted values from the sample.
pred <- predict(glm1, type = "response")
pred

### T-test of propensity scores
t.test(pred ~ glm1$y, alternative = "two.sided", var.equal = TRUE, conf.level = .95)

### Calculation of initial bias in propensity scores
Initial.bias <- cohen.d(pred, glm1$y)$CohenD
Initial.bias

### Calculation of the hit rate given a set of covariates
ClassLog (glm1, samp.100$y)
b <- ClassLog (glm1, samp.100$y)
b$overall
b$mcFadden

### Calculation of the $I$ Index given the improvement over the null hit rate
outcome <- b$rawtab[1,2] + b$rawtab[2,2]
initialHR <- (outcome/1000000)
Improvement <- (b$overall - initialHR)
ChancelImprove <- (1 - initialHR)
index <- (Improvement[1]/ChancelImprove[1])
index

### Caliper Matching on covariates with replacement and a caliper of .25
### P values of the ATE were replicated 1000 times to best estimate the true ATE.
X <- glm1$fitted
Y <- samp.100$Tr.1
Tr <- samp.100$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
summary(rr)
mb <- MatchBalance(y ~ x1 + x2 + x3, data = samp.100, match.out = rr, nboots = 100)
new <- c(rr$mdata$X[1,], rr$mdata$X[2,])

### T-test of propensity scores on the matched sample
t.test(new ~ rr$mdata$Tr, alternative = "two.sided", var.equal = TRUE, conf.level = .95)

### Calculation of bias in the propensity scores after matching
```
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
bias.red
abs(Initial.bias)

####Sensitivity Analysis with a Gamma of 2.5 evaluated at increments of .05
psens(rr, Gamma=2.5, GammaInc=.05)
a<-psens(rr, Gamma = 2.5, GammaInc = 0.05)
a$bounds
t(a$bounds)

####Place data into data file.
output<-data.frame((t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1], Initial.bias, Post.bias, bias.red, t(glm1$coefficients), rr[15])
write.table (output, file = "output100.csv", sep ="," , col.names =,)

######################################################################
####      Simulation of 1000 draws from the population for sample size of 100    #######
######################################################################
set.seed(13509)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 100, replace=T)
samp.100<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.100, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred(glm1$y))$CohenD
ClassLog (glm1, samp.100$y)
b<-ClassLog (glm1, samp.100$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/100)
Improvement<-(b$overall-initialHR)
ChanceImprove<-(1-initialHR)
Iindex<-(Improvement[1]/ChanceImprove[1])
X <- glm1$fitted
Y <- samp.100$Tr.1
Tr <- samp.100$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.100, match.out=rr, nboots=100)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=2.5, GammaInc=.05)
a<-psens(rr, Gamma = 2.5, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1], abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
write.table (output, file = "output100.csv", sep =",", col.names =,)

######################################################################
#####      Simulation of 1000 draws from the population for Sample Size of 400    ######
######################################################################
set.seed(14510)
for(i in 1:1000) {
matr<-sample(nrow(massive), 400, replace=T)
samp.400<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.400, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred,glm1$y)$CohenD
ClassLog (glm1, samp.400$y)
b<-ClassLog (glm1, samp.400$y)
if (nrow(b$rawtab)>1) outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/400)
Improvement<-(b$overall-initialHR)
ChancelImprove<-(1-initialHR)
lindex<-(Improvement[1]/ChancelImprove[1])
X <- glm1$fitted
Y <- samp.400$Tr.1
Tr <- samp.400$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.400, match.out=rr, nboots=100)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=2.5, GammaInc=.05)
a<-psens(rr, Gamma = 2.5, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], lindex, b$mcFadden[1],
abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
colnames(output) <-c("1.0", "1.05", "1.10", "1.15", "1.20", "1.25", "1.30", "1.35", "1.40",
"1.45", "1.50", "1.55", "1.60", "1.65", "1.70", "1.75", "1.80", "1.85", "1.90", "1.95", "2.00",
"2.05", "2.10", "2.15", "2.20", "2.25", "2.30", "2.35", "2.40", "2.45", "2.50", "2.55", "2.60",
"2.65", "2.70", "2.75", "2.80", "2.85", "2.90", "2.95", "3.00", "3.05", "3.10", "3.15", "3.20",
"5.05", "5.10", "5.15", "5.20", "5.25", "5.30", "5.35", "5.40", "5.45", "5.50", "5.55", "5.60",
"5.65", "5.70", "5.75", "5.80", "5.85", "5.90", "5.95", "6.00", "ATE", "lindex", "mcFadden",
"InitialBias", "PostBias", "BiasReduce", "intercept", "X1", "X2", "X3", "numobs")
write.table (output, file = "output400.csv", sep =",", col.names =,)

set.seed(15511)
for(i in 1:1000) {
    matr<-sample(nrow(massive), 700, replace=T)
samp.700<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.700, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred,glm1$y)$CohenD
ClassLog (glm1, samp.700$y)
b<-ClassLog (glm1, samp.700$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/700)
Improvement<-(b$overall-initialHR)
ChanceImprove<-(1-initialHR)
lindex<-(Improvement[1]/ChanceImprove[1])
X <- glm1$fitted
Y <- samp.700$Tr.1
Tr <- samp.700$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.700, match.out=rr, nboots=100)
new<-c(rr$mdata$X[,1],rr$mdata$X[,2])
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=2.5, Gammalnc=.05)
a<-psens(rr, Gamma = 2.5, Gammalnc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[3])),rr[1], Iindex, b$mcFadden[1],
abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
colnames(output) <-c("1.0", "1.05", "1.10", "1.15", "1.20", "1.25", "1.30", "1.35", "1.40",
"1.45", "1.50", "1.55", "1.60", "1.65", "1.70", "1.75", "1.80", "1.85", "1.90", "1.95", "2.00",
"2.05", "2.10", "2.15", "2.20", "2.25", "2.30", "2.35", "2.40", "2.45", "2.50", "2.55", "2.60",
"2.65", "2.70", "2.75", "2.80", "2.85", "2.90", "2.95", "3.00", "3.05", "3.10", "3.15", "3.20",
"5.05", "5.10", "5.15", "5.20", "5.25", "5.30", "5.35", "5.40", "5.45", "5.50", "5.55", "5.60",
"5.65", "5.70", "5.75", "5.80", "5.85", "5.90", "5.95", "6.00", "ATE", "Iindex", "mcFadden",
"InitialBias", "PostBias", "BiasReduc", "intercept", "X1", "X2", "X3", "numobs")
write.table (output, file = "output700.csv", sep ="," , col.names =,)

###########################################################################
#####      Simulation of 1000 draws from the population for Sample Size of 1000    ####
###########################################################################

set.seed(16512)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 1000, replace=T)
samp.1000<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.1000, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred,glm1$y)$CohenD
ClassLog (glm1, samp.1000$y)
b<-ClassLog (glm1, samp.1000$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/1000)
Improvement<-(b$overall-initialHR)
ChanceImprove<-(1-initialHR)
Iindex<-(Improvement[1]/ChanceImprove[1])
X <- glm1$fitted
Y <- samp.1000$Tr.1
Tr <- samp.1000$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.1000, match.out=rr, nboots=100)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(-abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(-bias.diff/abs(Initial.bias))
psens(rr, Gamma=2.5, GammaInc=.05)
a<-psens(rr, Gamma = 2.5, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,.]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1], abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
Tr <- samp.1500$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.1500, match.out=rr, nboots=100)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=2.5, GammaInc=.05)
a<-psens(rr, Gamma = 2.5, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1], abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
write.table (output, file = "output1500.csv", sep =",", col.names =,)

######################################################################
#####      Simulation of 1000 draws from the population for Sample Size of 2000    ####
######################################################################
set.seed(19002)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 2000, replace=T)
samp.2000<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.2000, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred,glm1$y)$CohenD
ClassLog (glm1, samp.2000$y)
b<-ClassLog (glm1, samp.2000$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/2000)
Improvement<-(b$overall-initialHR)
ChancelImprove<-(1-initialHR)
Iindex<-(Improvement[1]/ChancelImprove[1])
X <- glm1$fitted
Y <- samp.2000$Tr.1
Tr <- samp.2000$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.2000, match.out=rr, nboots=100)
new<-c(rr$mdata$X[,1],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=2.5, GammalInc=.05)
a<-psens(rr, Gamma = 2.5, GammalInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[3])),rr[1], lindex, b$mcFadden[1],
abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
colnames(output) <-c("1.0", "1.05", "1.10", "1.15", "1.20", "1.25", "1.30", "1.35", "1.40",
"1.45", "1.50", "1.55", "1.60", "1.65", "1.70", "1.75", "1.80", "1.85", "1.90", "1.95", "2.00",
"2.05", "2.10", "2.15", "2.20", "2.25", "2.30", "2.35", "2.40", "2.45", "2.50", "2.55", "2.60",
"2.65", "2.70", "2.75", "2.80", "2.85", "2.90", "2.95", "3.00", "3.05", "3.10", "3.15", "3.20",
"5.05", "5.10", "5.15", "5.20", "5.25", "5.30", "5.35", "5.40", "5.45", "5.50", "5.55", "5.60",
"5.65", "5.70", "5.75", "5.80", "5.85", "5.90", "5.95", "6.00", "ATE", "lindex", "mcFadden",
"InitialBias", "PostBias", "BiasReduc", "intercept", "X1", "X2", "X3", "numobs")
write.table (output, file = "output2000.csv", sep =",", col.names =)

######################################################################
#####      Simulation of 1000 draws from the population for Sample Size of 2500    ####
######################################################################
set.seed(20001)
for(i in 1:1000) {
matr<-sample(nrow(massive), 2500, replace=T)
samp.2500<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.2500, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred,glm1$y)$CohenD
ClassLog (glm1, samp.2500$y)
b<-ClassLog (glm1, samp.2500$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/2500)
Improvement<-((b$overall-initialHR)
ChanceImprove<-(1-initialHR)
lindex<-(Improvement[1]/ChanceImprove[1])
X <- glm1$fitted
Y <- samp.2500$Tr.1
Tr <- samp.2500$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.2500, match.out=rr, nboots=100)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=2.5, GammaInc=.05)
a<-psens(rr, Gamma = 2.5, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1], abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
write.table (output, file = "output2500.csv", sep ="," , col.names =)
### Create a vector of 1 dichotomous treatment variable
### 1 outcome variable
### and 3 covariates predicting treatment with 1,000,000 scores
### with a mean of 0 and sd of 1.

intercept = 0
beta1 = 2
beta2 = 1
beta3 = 0
x1 = rnorm(1000000,1,1)
x2 = rnorm(1000000,1,1)
x3 = rnorm(1000000,1,1)
Tr = rnorm(1000000,1,1)
linpred = intercept + x1*beta1+x2*beta2+x3*beta3
prob = exp(linpred)/(1 + exp(linpred))
runis = runif(1000000,0,1)
y = ifelse(runis < prob,1,0)

### Writing cohen's d function
### Cohen's d (pooled within group SD)
cohen.d<-function(DV,groupv)
{
  varg1<-by(DV, groupv,var)[[1]]
  varg2<-by(DV, groupv,var)[[2]]
  ng1<-by(DV, groupv,length)[[1]]
  ng2<-by(DV, groupv,length)[[2]]
  pooledvar<-(ng1-1)*varg1+(ng2-1)*varg2)/(ng1 + ng2 -2)
  pooledsd<-sqrt(pooledvar)
  meang1<-by(DV, groupv,mean)[[1]]
  meang2<-by(DV, groupv,mean)[[2]]
  meandif<-meang1-meang2
  cohd<-meandif/pooledsd
  varratio<-varg1/varg2
  list(pooled.SD=pooledsd, Group.Mean.Dif=meandif,CohenD=cohd,VarRatio=varratio)
}

### Creating an ATE of .50 in the population
Tr.1<-ifelse(y==1,Tr*1.25,Tr*.75)
massive<-data.frame (x1=x1, x2=x2, x3=x3, Tr=Tr,Tr.1=Tr.1, y=y)
write.table (massive, file = "massive.25.csv", sep ="," , col.names =)

t.test(Tr.1~y, alternative="two.sided",var.equal=TRUE, conf.level=.95)
T.effect<-cohen.d(Tr.1,y)$CohenD
### Propensity scores of all cases in the population

glm.massive<-glm(y~x1+x2+x3, massive, family=binomial(link="logit"))  
summary(glm.massive)

### Calculating predicted values from the population.
pred.glm.massive <- predict(glm.massive,type="response")

### Calculation of initial bias in propensity scores in the population
Initial.bias<-cohen.d(pred.glm.massive,glm.massive$y)$CohenD

### Calculation of the hit rate given a set of covariates
ClassLog (glm.massive, massive$y)
b<-ClassLog (glm.massive, massive$y)
b$overall
b$mcFadden

### Calculation of the I Index in the population given the improvement over the null hit rate
outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/1000000)
Improvement<-(b$overall-initialHR)
ChancelImprove<-(1-initialHR)
lindex<-(Improvement[1]/ChancelImprove[1])
lindex

#### Test Sample for sample size 100

#### Draw a random sample of 100 people from the massive data set
matr<-sample(nrow(massive), 100, replace=T)
samp.100<-data.frame(massive[matr,])

#### Run the logistic regression on the sample
glm1<-glm(y~x1+x2+x3, samp.100, family=binomial(link="logit"))  
summary(glm1)

#### Calculating predicted values from the sample.
pred <- predict(glm1,type="response")
pred

#### T-test of propensity scores
t.test(pred~glm1$y, alternative="two.sided",var.equal=TRUE, conf.level=.95)
## Calculation of initial bias in propensity scores

```r
Initial.bias <- cohen.d(pred, glm1$y)$CohenD
Initial.bias
```

## Calculation of the hit rate given a set of covariates

```r
ClassLog (glm1, samp.100$y)
b <- ClassLog (glm1, samp.100$y)
b$overall
b$mcFadden
```

## Calculation of the I Index given the improvement over the null hit rate

```r
outcome <- b$rawtab[1,2] + b$rawtab[2,2]
initialHR <- (outcome/100)
Improvement <- (b$overall - initialHR)
ChanceImprove <- (1 - initialHR)
Iindex <- Improvement[1] / ChanceImprove[1]
Iindex
```

## Caliper Matching on covariates with replacement and a caliper of .25

### P values of the ATE were replicated 1000 times to best estimate the true ATE.

```r
X <- glm1$fitted
Y <- samp.100$Tr.1
Tr <- samp.100$y
rr <- Match(Y = Y, Tr = Tr, X = X, caliper = .25, M = 1, replace = TRUE)
summary(rr)
mf <- MatchBalance(y ~ x1 + x2 + x3, data = samp.100, match.out = rr, nboots = 1000)
new <- c(rr$mdata$X[1,], rr$mdata$X[2,])
```

## T-test of propensity scores on the matched sample

```r
t.test(new ~ rr$mdata$Tr, alternative = "two.sided", var.equal = TRUE, conf.level = .95)
```

## Calculation of bias in the propensity scores after matching

```r
Post.bias <- cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff <- (abs(Initial.bias) - (abs(Post.bias)))
bias.red <- bias.diff / abs(Initial.bias)
bias.red
abs(Initial.bias)
```

## Sensitivity Analysis with a Gamma of 2.5 evaluated at increments of .05

```r
psens(rr, Gamma = 6.0, GammaInc = .05)
a <- psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)
```

## Place data into data file.

```r
output <- data.frame(t(a$bounds[,3]), rr[1], Iindex, b$mcFadden[1], Initial.bias, Post.bias,
```

```r
```
bias.red, t(glm1$coefficients), rr[15])
write.table (output, file = "output.csv", sep = "", col.names =,)

######################################################################
#####    Simulation of 1000 draws from the population for sample size 100   #########
######################################################################
set.seed(16664)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 100, replace=T)
samp.100<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.100, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred, glm1$y)$CohenD
ClassLog (glm1, samp.100$y)
b<-ClassLog (glm1, samp.100$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/100)
Improvement<-(b$overall-initialHR)
ChanceImprove<-(1-initialHR)
Iindex<-(Improvement[1]/ChanceImprove[1])
X <- glm1$fitted
Y <- samp.100$Tr.1
Tr <- samp.100$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.100, match.out=rr, nboots=100)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided",var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma = 6.0, GammaInc = .05)
a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1],


abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
write.table (output, file = "output100.csv", sep ="", col.names =)

######################################################################
#####      Simulation of 1000 draws from the population for Sample Size of 400    #####
######################################################################

set.seed(17665)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 400, replace=T)
samp.400<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.400, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred,glm1$y)$CohenD
ClassLog (glm1, samp.400$y)
b<-ClassLog (glm1, samp.400$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/400)
Improvement<-(b$overall-initialHR)
ChanceImprove<-(1-initialHR)
Iindex<-(Improvement[1]/ChanceImprove[1])
X <- glm1$fitted
Y <- samp.400$Tr.1
Tr <- samp.400$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.400, match.out=rr, nboots=100)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided",var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=6.0, GammaInc=.05)
a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a.bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1],

60
abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
write.table (output, file = "output400.csv", sep =",", col.names =,)

########################################################################
#####      Simulation of 1000 draws from the population for Sample Size of 700      ####
########################################################################
set.seed(18666)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 700, replace=T)
samp.700<-data.frame(massive[matr,])
  glm1<-glm(y~x1+x2+x3, samp.700, family=binomial(link="logit"))
  pred <- predict(glm1,type="response")
  Initial.bias<-cohen.d(pred,glm1$y)$CohenD
  ClassLog (glm1, samp.700$y)
  b<-ClassLog (glm1, samp.700$y)
  if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
  initialHR<-(outcome/700)
  Improvement<-(b$overall-initialHR)
  ChancelImprove<-(1-initialHR)
  Iindex<-(Improvement[1]/ChancelImprove[1])
  X <- glm1$fitted
  Y <- samp.700$Tr.1
  Tr <- samp.700$y
  rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
  mb <- MatchBalance(y~x1+x2+x3, data=samp.700, match.out=rr, nboots=100)
  new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
  t.test(new~rr$mdata$Tr, alternative="two.sided",var.equal=TRUE, conf.level=.95)
  Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
  bias.diff<-(abs(Initial.bias)-(abs(Post.bias))
  bias.red<-(bias.diff/abs(Initial.bias))
  psens(rr, Gamma=6.0, Gammalnc=.05)
a<-psens(rr, Gamma = 6.0, Gammalnc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1],
abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
colnames(output) <-c("1.0", "1.05", "1.10", "1.15", "1.20", "1.25", "1.30", "1.35", "1.40",
"1.45", "1.50", "1.55", "1.60", "1.65", "1.70", "1.75", "1.80", "1.85", "1.90", "1.95", "2.00",
"2.05", "2.10", "2.15", "2.20", "2.25", "2.30", "2.35", "2.40", "2.45", "2.50", "2.55", "2.60",
"2.65", "2.70", "2.75", "2.80", "2.85", "2.90", "2.95", "3.00", "3.05", "3.10", "3.15", "3.20",
"5.05", "5.10", "5.15", "5.20", "5.25", "5.30", "5.35", "5.40", "5.45", "5.50", "5.55", "5.60",
"5.65", "5.70", "5.75", "5.80", "5.85", "5.90", "5.95", "6.00", "ATE", "Iindex", "mcFadden",
"InitialBias", "PostBias", "BiasReduc", "intercept", "X1", "X2", "X3", "numobs")
write.table (output, file = "output700.csv", sep = ",", col.names =,)

########################################################################
###     Simulation of 1000 draws from the population for Sample Size of 1000   ###
########################################################################

set.seed(19667)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 1000, replace=T)
samp.1000<-data.frame(massive[matr,])
glm1<glm(y~x1+x2+x3, samp.1000, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred,glm1$y)$CohenD
ClassLog (glm1, samp.1000$y)
b<-ClassLog (glm1, samp.1000$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/1000)
Improvement<-(b$overall-initialHR)
ChanceImprove<-(1-initialHR)
Iindex<-(Improvement[1]/ChanceImprove[1])
X <- glm1$fitted
Y <- samp.1000$Tr.1
Tr <- samp.1000$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.1000, match.out=rr, nboots=100)
new<-(rr$mdata$X[1,],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided",var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=6.0, GammaInc=.05)
a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1],

62
set.seed(40001)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 1500, replace=T)
  samp.1500<-data.frame(massive[matr,])
  glm1<-glm(y~x1+x2+x3, samp.1500, family=binomial(link="logit"))
  pred <- predict(glm1,type="response")
  Initial.bias<-cohen.d(pred,glm1$y)$CohenD
  ClassLog (glm1, samp.1500$y)
  b<-ClassLog (glm1, samp.1500$y)
  if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
  initialHR<-(outcome/1500)
  Improvement<-(b$overall-initialHR)
  ChanceImprove<-(1-initialHR)
  Iindex<-(Improvement[1]/ChanceImprove[1])
  X <- glm1$fitted
  Y <- samp.1500$Tr.1
  Tr <- samp.1500$y
  rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
  mb <- MatchBalance(y~x1+x2+x3, data=samp.1500, match.out=rr, nboots=100)
  new<-c(rr$mdata$X[1,,],rr$mdata$X[2,,])
  t.test(new~rr$mdata$Tr, alternative="two.sided",var.equal=TRUE, conf.level=.95)
  Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
  bias.diff<-(abs(Initial.bias)-(abs(Post.bias))
  bias.red<-(bias.diff/abs(Initial.bias))
  psens(rr, Gamma=6.0, GammaInc=.05)
  a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
  a$bounds
  t(a$bounds)
  output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], lindex, b$mcFadden[1],
  abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
colnames(output) <-c("1.0", "1.05", "1.10", "1.15", "1.20", "1.25", "1.30", "1.35", "1.40", 
"1.45", "1.50", "1.55", "1.60", "1.65", "1.70", "1.75", "1.80", "1.85", "1.90", "1.95", "2.00", 
"2.05", "2.10", "2.15", "2.20", "2.25", "2.30", "2.35", "2.40", "2.45", "2.50", "2.55", "2.60", 
"2.65", "2.70", "2.75", "2.80", "2.85", "2.90", "2.95", "3.00", "3.05", "3.10", "3.15", "3.20", 
"5.05", "5.10", "5.15", "5.20", "5.25", "5.30", "5.35", "5.40", "5.45", "5.50", "5.55", "5.60", 
"5.65", "5.70", "5.75", "5.80", "5.85", "5.90", "5.95", "6.00", "ATE", "Iindex", "mcFadden", 
"InitialBias", "PostBias", "BiasReduc", "intercept", "X1", "X2", "X3", "numobs")
write.table (output, file = "output1000.csv", sep ="," , col.names =,)

### Simulation of 1000 draws from the population for Sample Size of 2000

```r
set.seed(41002)
for(i in 1:1000) {
  matr <- sample(nrow(massive), 2000, replace = T)
  samp.2000 <- data.frame(massive[matr,])
  glm1 <- glm(y ~ x1 + x2 + x3, samp.2000, family = binomial(link = "logit"))
  pred <- predict(glm1, type = "response")
  Initial.bias <- cohen.d(pred, glm1$y)$CohenD
  b <- ClassLog(glm1, samp.2000$y)
  if (nrow(b$rawtab) > 1) outcome <- b$rawtab[1, 2] + b$rawtab[2, 2]
  initialHR <- (outcome / 2000)
  Improvement <- (b$overall - initialHR)
  ChancelImprove <- (1 - initialHR)
  Iindex <- (Improvement[1] / ChanceImprove[1])
  X <- glm1$fitted
  Y <- samp.2000$Tr.1
  Tr <- samp.2000$y
  rr <- Match(Y = Y, Tr = Tr, X = X, caliper = .25, M = 1, replace = TRUE)
  mb <- MatchBalance(y ~ x1 + x2 + x3, data = samp.2000, match.out = rr, nboots = 100)
  new <- c(rr$mdata$X[1,], rr$mdata$X[2,])
  t.test(new ~ rr$mdata$Tr, alternative = "two.sided", var.equal = TRUE, conf.level = .95)
  Post.bias <- cohen.d(new, rr$mdata$Tr)$CohenD
  bias.diff <- (abs(Initial.bias) - abs(Post.bias))
  bias.red <- (bias.diff / abs(Initial.bias))
  psens(rr, Gamma = 6.0, GammaInc = .05)
  a <- psens(rr, Gamma = 6.0, GammaInc = 0.05)
  a$bounds
  t(a$bounds)
  output[i,] <- data.frame(cbind(t(a$bounds[, 3])), rr[1], Iindex, b$mcFadden[1],
```
abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])


write.table(output, file = "output2000.csv", sep = ",", col.names =)

####################################################################
#####      Simulation of 1000 draws from the population for Sample Size of 2500    ####
####################################################################

set.seed(42003)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 2500, replace=T)
  samp.2500<-data.frame(massive[matr,])
  glm1<-glm(y~x1+x2+x3, samp.2500, family=binomial(link="logit"))
  pred <- predict(glm1,type="response")
  Initial.bias<-cohen.d(pred,glm1$y)$CohenD
  ClassLog (glm1, samp.2500$y)
  b<-ClassLog (glm1, samp.2500$y)
  if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
  initialHR<-(outcome/2500)
  Improvement<-(b$overall-initialHR)
  ChancelImprove<-(1-initialHR)
  Iindex<-(Improvement[1]/ChancelImprove[1])
  X <- glm1$fitted
  Y <- samp.2500$Tr.1
  Tr <- samp.2500$y
  rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
  mb <- MatchBalance(y~x1+x2+x3, data=samp.2500, match.out=rr, nboots=100)
  new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
  t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
  Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
  bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
  bias.red<-(bias.diff/abs(Initial.bias))
  psens(rr, Gamma=6.0, GammaInc=.05)
  a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
  a$bounds
  t(a$bounds)
  output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1],
abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
write.table(output, file = "output2500.csv", sep ="," , col.names =)

######################################################################
#######      Set population parameters for I Index of 0.25     ##############
######################################################################

###Create a vector of 1 dichotomous treatment variable
###1 outcome variable
###and 3 covariates predicting treatment with 1,000,000 scores
###with a mean of 0 and sd of 1.

intercept = 0
beta1 = 4
beta2 = 2
beta3 = 0
x1 = rnorm(1000000,1,1)
x2 = rnorm(1000000,1,1)
x3 = rnorm(1000000,1,1)
Tr = rnorm(1000000,1,1)
linpred = intercept + x1*beta1+x2*beta2+x3*beta3
prob = exp(linpred)/(1 + exp(linpred))
runis = runif(1000000,0,1)
y = ifelse(runis < prob,1,0)

###Writing cohen's d function
###Cohen's d (pooled within group SD)
cohen.d<-function(DV,groupv)
{
  varg1<-by(DV, groupv, var)
  varg2<-by(DV, groupv, var)
  ng1<-by(DV, groupv, length)
  ng2<-by(DV, groupv, length)
pooledvar<-mean((ng1-1)*varg1 + (ng2-1)*varg2)/(ng1 + ng2 -2)
pooledsd<-sqrt(pooledvar)
meanv1<-by(DV, groupv, mean)
}
mean2 <- by(DV, groupv, mean)[2]
meandif <- mean1 - mean2
cohd <- meandif / pooledsd
varratio <- var1 / var2
list(pooled.SD = pooledsd, Group.Mean.Dif = meandif, CohenD = cohd, VarRatio = varratio)

### Creating an ATE of .50 in the population
Tr.1 <- ifelse(y == 1, Tr * 1.25, Tr * .75)
massive <- data.frame (x1 = x1, x2 = x2, x3 = x3, Tr = Tr, Tr.1 = Tr.1, y = y)
write.table (massive, file = "massive50.csv", sep = ",", col.names = ,)

t.test(Tr.1 ~ y, alternative = "two.sided", var.equal = TRUE, conf.level = .95)
T.effect <- cohen.d(Tr.1, y)$CohenD
T.effect

### Propensity scores of all cases in the population
glm.massive <- glm(y ~ x1 + x2 + x3, massive, family = binomial(link = "logit"))
summary(glm.massive)

### Calculating predicted values from the population.
pred.glm.massive <- predict(glm.massive, type = "response")

### Calculation of initial bias in propensity scores in the population
Initial.bias <- cohen.d(pred.glm.massive, glm.massive$y)$CohenD

### Calculation of the hit rate given a set of covariates
ClassLog (glm.massive, massive$y)
b <- ClassLog (glm.massive, massive$y)
b$overall
b$mcFadden

### Calculation of the I Index in the population given the improvement over the null hit rate
outcome <- b$rawtab[1,2] + b$rawtab[2,2]
initialHR <- (outcome / 1000000)
Improvement <- (b$overall - initialHR)
ChanceImprove <- (1 - initialHR)
Iindex <- (Improvement[1] / ChanceImprove[1])
Iindex

########################################################################
#######       Test Sample for sample size 100        ###############
########################################################################
### Draw a random sample of 100 people from the massive data set

```r
matr <- sample(nrow(massive), 700, replace = T)
samp.100 <- data.frame(massive[matr,])
```

### Run the logistic regression on the sample

```r
glm1 <- glm(y ~ x1 + x2 + x3, samp.100, family = binomial(link = "logit"))
summary(glm1)
```

### Calculating predicted values from the sample.

```r
pred <- predict(glm1, type = "response")
pred
```

### T-test of propensity scores

```r
t.test(pred ~ glm1$y, alternative = "two.sided", var.equal = TRUE, conf.level = .95)
```

### Calculation of initial bias in propensity scores

```r
Initial.bias <- cohen.d(pred, glm1$y)$CohenD
Initial.bias
```

### Calculation of the hit rate given a set of covariates

```r
ClassLog (glm1, samp.100$y)
b <- ClassLog (glm1, samp.100$y)
b$overall
b$mcFadden
```

### Calculation of the I Index given the improvement over the null hit rate

```r
outcome <- b$rawtab[1, 2] + b$rawtab[2, 2]
initialHR <- (outcome / 100)
Improvement <- (b$overall - initialHR)
ChanceImprove <- (1 - initialHR)
Iindex <- (Improvement[1] / ChanceImprove[1])
Iindex
```

### Caliper Matching on covariates with replacement and a caliper of .25

```R
X <- glm1$fitted
Y <- samp.100$Tr.1
Tr <- samp.100$y
rr <- Match(Y = Y, Tr = Tr, X = X, caliper = .25, M = 1, replace = TRUE)
summary(rr)
mb <- MatchBalance(y ~ x1 + x2 + x3, data = samp.100, match.out = rr, nboots = 1000)
new <- c(rr$mdata$X[1,], rr$mdata$X[2,])
```

### T-test of propensity scores on the matched sample

```r
t.test(new ~ rr$mdata$Tr, alternative = "two.sided", var.equal = TRUE, conf.level = .95)
```

### Calculation of bias in the propensity scores after matching
Post.bias <- cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff <- (abs(Initial.bias) - abs(Post.bias))
bias.red <- (bias.diff / abs(Initial.bias))
bias.red
abs(Initial.bias)

#### Sensitivity Analysis with a Gamma of 2.5 evaluated at increments of .05
psens(rr, Gamma = 6.0, GammaInc = .05)
a <- psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)

#### Place data into data file.
output <- data.frame(t(a$bounds[,3]), rr[1], Iindex, b$mcFadden[1], Initial.bias, Post.bias, bias.red, t(glm1$coefficients), rr[15])
write.table(output, file = "output100.csv", sep = "", col.names = ,)

######################################################################
#####      Simulation of 1000 draws from the population for Sample Size of 100      ####
####################################################################
set.seed(21000)
for(i in 1:1000) {
  matr <- sample(nrow(massive), 100, replace = T)
samp.100 <- data.frame(massive[matr,])
glm1 <- glm(y ~ x1 + x2 + x3, samp.100, family = binomial(link = "logit"))
pred <- predict(glm1, type = "response")
Initial.bias <- cohen.d(pred, glm1$y)$CohenD
ClassLog (glm1, samp.100$y)
b <- ClassLog(glm1, samp.100$y)
if (nrow(b$rawtab) > 1) outcome <- b$rawtab[,1] + b$rawtab[,2]
initialHR <- (outcome / 100)
Improvement <- (b$overall - initialHR)
ChancelImprove <- (1 - initialHR)
Iindex <- (Improvement[1] / ChancelImprove[1])
X <- glm1$fitted
Y <- samp.100$Tr.1

69
Tr <- samp.100$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.100, match.out=rr, nboots=100)
new<-c(-rr$mdata$X[1,],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=6.0, GammaInc=.05)
a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], lindex, b$mcFadden[1],
abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
colnames(output) <-c("1.0", "1.05", "1.10", "1.15", "1.20", "1.25", "1.30", "1.35", "1.40",
"1.45", "1.50", "1.55", "1.60", "1.65", "1.70", "1.75", "1.80", "1.85", "1.90", "1.95", "2.00",
"2.05", "2.10", "2.15", "2.20", "2.25", "2.30", "2.35", "2.40", "2.45", "2.50", "2.55", "2.60",
"2.65", "2.70", "2.75", "2.80", "2.85", "2.90", "2.95", "3.00", "3.05", "3.10", "3.15", "3.20",
"5.05", "5.10", "5.15", "5.20", "5.25", "5.30", "5.35", "5.40", "5.45", "5.50", "5.55", "5.60",
"5.65", "5.70", "5.75", "5.80", "5.85", "5.90", "5.95", "6.00", "ATE", "lindex", "mcFadden",
"InitialBias", "PostBias", "BiasReduc", "intercept", "X1", "X2", "X3", "numobs")
write.table (output, file = "output100.csv", sep =",", col.names =,)

######################################################################
#####      Simulation of 1000 draws from the population for Sample Size of 400      ####
######################################################################
set.seed(22001)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 400, replace=T)
samp.400<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.400, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred,glm1$y)$CohenD
ClassLog (glm1, samp.400$y)
b<-ClassLog (glm1, samp.400$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-<outcome/400)
Improvement<-(b$overall-initialHR)
ChanceImprove<-(1-initialHR)
lindex<-<Improvement[1]/ChanceImprove[1])
X <- glm1$fitted
Y <- samp.400$Tr.1

Tr <- samp.400$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.400, match.out=rr, nboots=100)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=6.0, GammaInc=.05)
a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds

t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1], abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])

write.table (output, file = "output400.csv", sep =",", col.names =,)

set.seed(23002)
for(i in 1:1000) {
    matr<-sample(nrow(massive), 700, replace=T)
samp.700<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.700, family=binomial(link="logit"))
pred <- predict(glm1, type="response")
Initial.bias<-cohen.d(pred, glm1$y)$CohenD
ClassLog (glm1, samp.700$y)
b<-ClassLog (glm1, samp.700$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/700)
Improvement<-(b$overall-initialHR)
ChancelImprove<-(-1-initialHR)
Iindex<-(Improvement[1]/ChancelImprove[1])
X <- glm1$fitted
Y <- samp.700$Tr.1

71
Tr <- samp.700$y
ger <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.700, match.out=rr, nboots=100)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=6.0, GammaInc=.05)
a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[3])),rr[1], lindex, b$mcFadden[1], abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
Tr <- samp.1000$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.1000, match.out=rr, nboots=100)
new<-c(rr$mdata$X[,1], rr$mdata$X[,2])
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=6.0, GammaInc=.05)
a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], lindex, b$mcFadden[1], abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
write.table(output, file = "output1000.csv", sep =",", col.names =,)

###########################################################################
####      Simulation of 1000 draws from the population for Sample Size of 1500    ####
###########################################################################
set.seed(25004)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 1500, replace=T)
samp.1500<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.1500, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred, glm1$y)$CohenD
ClassLog (glm1, samp.1500$y)
b<-ClassLog (glm1, samp.1500$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/1500)
Improvement<-(b$overall-initialHR)
ChanceImprove<-(1-initialHR)
lindex<-(Improvement[1]/ChanceImprove[1])
X <- glm1$fitted
Y <- samp.1500$Tr.1
Tr <- samp.1500$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.1500, match.out=rr, nboots=100)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<--(abs(Initial.bias)-(abs(Post.bias)))
bias.red<--(bias.diff/abs(Initial.bias))
psens(rr, Gamma=6.0, GammaInc=.05)
a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[,i,]<-data.frame(cbind(t(a$bounds[3,])),rr[1], lindex, b$mcFadden[1], abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
write.table (output, file = "output1500.csv", sep =",", col.names =)

###############################################################################
#####      Simulation of 1000 draws from the population for Sample Size of 2000    ####
###############################################################################
set.seed(26005)
for(i in 1:1000) {
matr<-sample(nrow(massive), 2000, replace=T)
samp.2000<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.2000, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred,glm1$y)$CohenD
ClassLog (glm1, samp.2000$y)
b<-ClassLog (glm1, samp.2000$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/2000)
Improvement<-(b$overall-initialHR)
ChancelImprove<-(1-initialHR)
lindex<-(Improvement[1]/ChancelImprove[1])
X <- glm1$fitted
Y <- samp.2000$Tr.1
Tr <- samp.2000$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.2000, match.out=rr, nboots=100)
new<-.c(rr$mdata$X[1,],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-.cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=6.0, GammaInc=.05)
a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1], abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
write.table (output, file = "output2000.csv", sep =",", col.names =,)

##########################################################################
#####      Simulation of 1000 draws from the population for Sample Size of 2500    ####
##########################################################################
set.seed(27006)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 2500, replace=T)
samp.2500<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.2500, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-.cohen.d(pred,glm1$y)$CohenD
ClassLog (glm1, samp.2500$y)
b<-ClassLog (glm1, samp.2500$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/2500)
Improvement<-(b$overall-initialHR)
ChanceImprove<-(1-initialHR)
Iindex<-(Improvement[1]/ChanceImprove[1])
X <- glm1$fitted
Y <- samp.2500$Tr.1


Tr <- samp.2500$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.2500, match.out=rr, nboots=100)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=6.0, GammaInc=.05)
a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], lindex, b$mcFadden[1],
abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
colnames(output) <-c("1.0", "1.05", "1.10", "1.15", "1.20", "1.25", "1.30", "1.35", "1.40",
"1.45", "1.50", "1.55", "1.60", "1.65", "1.70", "1.75", "1.80", "1.85", "1.90", "1.95", "2.00",
"2.05", "2.10", "2.15", "2.20", "2.25", "2.30", "2.35", "2.40", "2.45", "2.50", "2.55", "2.60",
"2.65", "2.70", "2.75", "2.80", "2.85", "2.90", "2.95", "3.00", "3.05", "3.10", "3.15", "3.20",
"5.05", "5.10", "5.15", "5.20", "5.25", "5.30", "5.35", "5.40", "5.45", "5.50", "5.55", "5.60",
"5.65", "5.70", "5.75", "5.80", "5.85", "5.90", "5.95", "6.00", "ATE", "lindex", "mcFadden",
"InitialBias", "PostBias", "BiasReduc", "intercept", "X1", "X2", "X3", "numobs")
write.table (output, file = "output2500.csv", sep ="", col.names =)
### Create a vector of 1 dichotomous treatment variable
### # 1 outcome variable
### and 3 covariates predicting treatment with 1,000,000 scores
### with a mean of 0 and sd of 1.

intercept = 0
beta1 = 8
beta2 = 4
beta3 = 0
x1 = rnorm(1000000,1,1)
x2 = rnorm(1000000,1,1)
x3 = rnorm(1000000,1,1)
Tr = rnorm(1000000,1,1)
linpred = intercept + x1*beta1+x2*beta2+x3*beta3
prob = exp(linpred)/(1 + exp(linpred))
runis = runif(1000000,0,1)
y = ifelse(runis < prob,1,0)

#### Writing cohen’s d function
#### Cohen’s d (pooled within group SD)
cohen.d<-function(DV,groupv)
{
varg1<-by(DV, groupv,var)[[1]]
varg2<-by(DV, groupv,var)[[2]]
ng1<-by(DV, groupv,length)[[1]]
ng2<-by(DV, groupv,length)[[2]]
pooledvar<-(ng1-1)*varg1+(ng2-1)*varg2)/(ng1 + ng2 -2)
pooledsd<-sqrt(pooledvar)
meang1<-by(DV, groupv,mean)[[1]]
meang2<-by(DV, groupv,mean)[[2]]
meandif<-meang1-meang2
cohd<-meandif/pooledsd
varratio<-varg1/varg2
list(pooled.SD=pooledsd, Group.Mean.Dif=meandif,CohenD=cohd,VarRatio=varratio)
}

### Creating an ATE of .50 in the population
Tr.1<-ifelse(y==1,Tr*1.25,Tr*.75)
massive<-data.frame (x1=x1, x2=x2, x3=x3, Tr=Tr,Tr.1=Tr.1, y=y)
write.table (massive, file = "massive75.csv", sep =" ", col.names =)
t.test(Tr.1~y, alternative="two.sided",var.equal=TRUE, conf.level=.95)
T.effect<-cohen.d(Tr.1,y)$CohenD
### Propensity scores of all cases in the population

```r
glm.massive <- glm(y ~ x1 + x2 + x3, massive, family = binomial(link = "logit"))
summary(glm.massive)
```

#### Calculating predicted values from the population.

```r
pred.glm.massive <- predict(glm.massive, type = "response")
```

#### Calculation of initial bias in propensity scores in the population

```r
Initial.bias <- cohen.d(pred.glm.massive, glm.massive$y)$CohenD
```

#### Calculation of the hit rate given a set of covariates

```r
ClassLog (glm.massive, massive$y)
b <- ClassLog (glm.massive, massive$y)
b$overall
b$mcFadden
```

#### Calculation of the I Index in the population given the improvement over the null hit rate

```r
outcome <- b$rawtab[1,2] + b$rawtab[2,2]
initialHR <- (outcome / 1000000)
Improvement <- (b$overall - initialHR)
ChancelImprove <- (1 - initialHR)
lindex <- (Improvement[1] / ChancelImprove[1])
lindex
```

### Test Sample for sample size 100

#### Draw a random sample of 100 people from the massive data set

```r
matr <- sample(nrow(massive), 1000, replace = T)
samp.100 <- data.frame(massive[matr,])
```

#### Run the logistic regression on the sample

```r
glm1 <- glm(y ~ x1 + x2 + x3, samp.100, family = binomial(link = "logit"))
summary(glm1)
```

#### Calculating predicted values from the sample.

```r
pred <- predict(glm1, type = "response")
pred
```

#### T-test of propensity scores

```r
t.test(pred ~ glm1$y, alternative = "two.sided", var.equal = TRUE, conf.level = .95)
```

#### Calculation of initial bias in propensity scores
Initial.bias<-cohen.d(pred,glm1$y)$CohenD

####Calculation of the hit rate given a set of covariates
ClassLog (glm1, samp.100$y)
b<-ClassLog (glm1, samp.100$y)
b$overall
b$mcFadden

####Calculation of the I Index given the improvement over the null hit rate
outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/100)
Improvement<-(b$overall-initialHR)
ChanceImprove<-(1-initialHR)
lindex<-(Improvement[1]/ChanceImprove[1])
lindex

####Caliper Matching on covariates with replacement and a caliper of .25
####P values of the ATE were replicated 1000 times to best estimate the true ATE.
X <- glm1$fitted
Y <- samp.100$Tr.1
Tr <- samp.100$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
summary(rr)
mb <- MatchBalance(y~x1+x2+x3, data=samp.100, match.out=rr, nboots=1000)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])

####T-test of propensity scores on the matched sample
t.test(new~rr$mdata$Tr, alternative="two.sided",var.equal=TRUE, conf.level=.95)

####Calculation of bias in the propensity scores after matching
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
bias.red

####Sensitivity Analysis with a Gamma of 2.5 evaluated at increments of .05
psens(rr, Gamma=6.0, GammaInc=.05)
a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)

####Place data into data file.
output<-data.frame((t(a$bounds[,3])),rr[1], lindex, b$mcFadden[1], Initial.bias, Post.bias, bias.red, t(glm1$coefficients), rr[15])
colnames(output) <-c("1.0", "1.05", "1.10", "1.15", "1.20", "1.25", "1.30", "1.35", "1.40", "1.45")
```r
set.seed(30000)
for (i in 1:1000) {
  matr <- sample(nrow(massive), 100, replace = T)
  samp.100 <- data.frame(massive[matr,])
  glm1 <- glm(y ~ x1 + x2 + x3, samp.100, family = binomial(link = "logit"))
  pred <- predict(glm1, type = "response")
  Initial.bias <- cohen.d(pred, glm1$y)$CohenD
  ClassLog (glm1, samp.100$y)
  b <- ClassLog (glm1, samp.100$y)
  if (nrow(b$rawtab) > 1) outcome <- b$rawtab[1, 2] + b$rawtab[2, 2]
  initialHR <- (outcome / 100)
  Improvement <- (b$overall - initialHR)
  ChancelImprove <- (1 - initialHR)
  lindex <- (Improvement[1] / ChancelImprove[1])
  X <- glm1$fitted
  Y <- samp.100$Tr.1
  Tr <- samp.100$y
  rr <- Match(Y = Y, Tr = Tr, X = X, M = 1, caliper = .25, replace = TRUE)
  mb <- MatchBalance(y ~ x1 + x2 + x3, data = samp.100, match.out = rr, nboots = 100)
  new <- c(rr$data$X[1,], rr$data$X[2,])
  t.test(new ~ rr$data$Tr, alternative = "two.sided", var.equal = TRUE, conf.level = .95)
  Post.bias <- cohen.d(new, rr$data$Tr)$CohenD
  bias.diff <- (abs(Initial.bias) - abs(Post.bias))
  bias.red <- (bias.diff / abs(Initial.bias))
  a <- psens(rr, Gamma = 6.0, GammaInc = .05)
  a$bounds
  t(a$bounds)
  output[i,] <- data.frame(cbind(t(a$bounds[, 3]), rr[1], lindex, b$mcFadden[1],
          abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
  }
colnames(output) <- c("1.0", "1.05", "1.10", "1.15", "1.20", "1.25", "1.30", "1.35", "1.40",
```

80

---

The text above contains R code for simulating data from a population, performing logistic regression, calculating Cohen's d for initial and post-bias, calculating other measures of effect size, and writing the results to a CSV file.
set.seed(31001)
for(i in 1:2) {
  matr <- sample(nrow(massive), 400, replace=T)
  samp.400 <- data.frame(massive[matr,])
  glm1 <- glm(y~x1+x2+x3, samp.400, family=binomial(link="logit"))
  pred <- predict(glm1, type="response")
  Initial.bias <- cohen.d(pred, glm1$y)$CohenD
  ClassLog (glm1, samp.400$y)
  b <- ClassLog (glm1, samp.400$y)
  if (nrow(b$rawtab)>1) outcome <- b$rawtab[1,2]+b$rawtab[2,2]
  initialHR <- (outcome/400)
  Improvement <- (b$overall-initialHR)
  ChanceImprove <- (1-initialHR)
  Iindex <- (Improvement[1]/ChanceImprove[1])
  X <- glm1$fitted
  Y <- samp.400$Tr.1
  Tr <- samp.400$y
  rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
  mb <- MatchBalance(y~x1+x2+x3, data=samp.400, match.out=rr, nboots=100)
  new <- c(rr$mdata$X[1,], rr$mdata$X[2,])
  t.test(new ~ rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
  Post.bias <- cohen.d(new, rr$mdata$Tr)$CohenD
  bias.diff <- (abs(Initial.bias)-(abs(Post.bias)))
  bias.red <- (bias.diff/abs(Initial.bias))
  psens(rr, Gamma=6.0, GammaInc=.05)
  a <- psens(rr, Gamma = 6.0, GammaInc = 0.05)
  a$bounds
  t(a$bounds)
  output[i, ] <- data.frame(cbind(t(a$bounds[,3]), rr[1], Iindex, b$mcFadden[1], abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
} 
write.table(output, file = "output100.csv", sep = ",", col.names =)
set.seed(33002)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 700, replace=T)
  samp.700<-data.frame(massive[matr])
  glm1<-glm(y~x1+x2+x3, samp.700, family=binomial(link="logit"))
  pred <- predict(glm1,type="response")
  Initial.bias<-cohen.d(pred,glm1$y)$CohenD
  ClassLog (glm1, samp.700$y)
  b<-ClassLog (glm1, samp.700$y)
  if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
  initialHR<-(outcome/700)
  Improvement<-(b$overall-initialHR)
  ChanceImprove<-(1-initialHR)
  Iindex<-(Improvement[1]/ChanceImprove[1])
  X <- glm1$fitted
  Y <- samp.700$Tr.1
  Tr <- samp.700$y
  rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
  mb <- MatchBalance(y~x1+x2+x3, data=samp.700, match.out=rr, nboots=100)
  new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
  t.test(new~rr$mdata$Tr, alternative="two.sided",var.equal=TRUE, conf.level=.95)
  Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
  bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
  bias.red<-(bias.diff/abs(Initial.bias))
  psens(rr, Gamma=6.0, GammaInc=.05)
  a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
  a$bounds
  t(a$bounds)
  output[i,]<-data.frame(cbind(t(a$bounds[3])),rr[1], lindex, b$mcFadden[1], abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}

write.table (output, file = "output700.csv", sep = ",", col.names =)

######################################################################
#####      Simulation of 1000 draws from the population for Sample Size of 1000    ####
######################################################################

set.seed(34003)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 1000, replace=T)
  samp.1000<-data.frame(massive[matr,])
  glm1<-glm(y~x1+x2+x3, samp.1000, family=binomial(link="logit"))
  pred <- predict(glm1,type="response")
  Initial.bias<-cohen.d(pred,glm1$y)$CohenD
  ClassLog (glm1, samp.1000$y)
  b<-ClassLog (glm1, samp.1000$y)
  if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
  initialHR<-(outcome/1000)
  Improvement<-(b$overall-initialHR)
  ChancelImprove<-(1-initialHR)
  Iindex<-(Improvement[1]/ChancelImprove[1])
  X <- glm1$fitted
  Y <- samp.1000$Tr.1
  Tr <- samp.1000$y
  rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
  mb <- MatchBalance(y~x1+x2+x3, data=samp.1000, match.out=rr, nboots=100)
  new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
  t.test(new~rr$mdata$Tr, alternative="two.sided",var.equal=TRUE, conf.level=.95)
  Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
  bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
  bias.red<-(bias.diff/abs(Initial.bias))
  psens(rr, Gamma=6.0, Gammalnc=0.05)
  a<-psens(rr, Gamma = 6.0, Gammalnc = 0.05)
  a$bounds t(a$bounds)
  output[i,]<-data.frame(cbind(t(a$bounds[3])),rr[1], lindex, b$mcFadden[1], abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}


write.table (output, file = "output1000.csv", sep =",", col.names =,)

#########################################################################
#####      Simulation of 1000 draws from the population for Sample Size of 1500    ####
#########################################################################

set.seed(35004)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 1500, replace=T)
samp.1500<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.1500, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred,glm1$y)$CohenD
ClassLog (glm1, samp.1500$y)
b<-ClassLog (glm1, samp.1500$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/1500)
Improvement<-(b$overall-initialHR)
ChancelImprove<-(1-initialHR)
lindex<-(Improvement[1]/ChancelImprove[1])
X <- glm1$fitted
Y <- samp.1500$Tr.1
Tr <- samp.1500$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.1500, match.out=rr, nboots=100)
new<-c(rr$mdata$X[,1],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided",var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(-(abs(Initial.bias)-(abs(Post.bias))))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=6.0, GammaInc=.05)
a<-(psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[3])),rr[1], lindex, b$mcFadden[1], abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
### Simulation of 1000 draws from the population for Sample Size of 2000 ###

```r
set.seed(36005)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 2000, replace=T)
  samp.2000<-data.frame(massive[matr,])
  glm1<-glm(y~x1+x2+x3, samp.2000, family=binomial(link="logit"))
  pred <- predict(glm1,type="response")
  Initial.bias<-cohen.d(pred,glm1$y)$CohenD
  ClassLog (glm1, samp.2000$y)
  b<-ClassLog (glm1, samp.2000$y)
  if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
  initialHR<-(outcome/2000)
  Improvement<-(b$overall-initialHR)
  ChanceImprove<-(1-initialHR)
  Iindex<-(Improvement[1]/ChanceImprove[1])
  X <- glm1$fitted
  Y <- samp.2000$Tr.1
  Tr <- samp.2000$y
  rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
  mb <- MatchBalance(y~x1+x2+x3, data=samp.2000, match.out=rr, nboots=100)
  new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
  t.test(new~rr$mdata$Tr, alternative="two.sided",var.equal=TRUE, conf.level=.95)
  Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
  bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
  bias.red<-(bias.diff/abs(Initial.bias))
  psens(rr, Gamma=6.0, GammaInc=.05)
  a<--psens(rr, Gamma = 6.0, GammaInc = 0.05)
  a$bounds
  t(a$bounds)
  output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], lindex, b$mcFadden[1],
  abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
```

write.table(output, file = "output2000.csv", sep =",", col.names =,

######################################################################
#####      Simulation of 1000 draws from the population for Sample Size of 2500    ####
######################################################################
set.seed(37006)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 2500, replace=T)
  samp.2500<-(data.frame(massive[matr,]))
  glm1<-(glm(y~x1+x2+x3, samp.2500, family=binomial(link="logit")))
  pred <- predict(glm1,type="response")
  Initial.bias<-(cohen.d(pred,glm1$y)$CohenD)
  ClassLog (glm1, samp.2500$y)
  b<-ClassLog (glm1, samp.2500$y)
  if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
  initialHR<-(outcome/2500)
  Improvement<-(b$overall-initialHR)
  ChancelImprove<-(1-initialHR)
  Iindex<-(Improvement[1]/ChancelImprove[1])
  X <- glm1$fitted
  Y <- samp.2500$Tr.1
  Tr <- samp.2500$y
  rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
  mb <- MatchBalance(y~x1+x2+x3, data=samp.2500, match.out=rr, nboots=100)
  new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
  t.test(new~rr$mdata$Tr, alternative="two.sided",var.equal=TRUE, conf.level=.95)
  Post.bias<-(cohen.d(new, rr$mdata$Tr)$CohenD)
  bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
  bias.red<-(-bias.diff/abs(Initial.bias))
  psens(rr, Gamma=6.0, GammaInc=.05)
  a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
  a$bounds
  t(a$bounds)
  output[i,]<-data.frame(cbind(t(a$bounds[3]),rr[1], lindex, b$mcFadden[1], abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
write.table (output, file = "output2500.csv", sep =",", col.names =,)

######################################################################
#############        Set population parameters for I Index of 0.90       ##############
######################################################################
###Create a vector of 1 dichotomous treatment variable
###1 outcome variable
###and 3 covariates predicting treatment with 1,000,000 scores
###with a mean of 0 and sd of 1.

intercept = 0
beta1 = 20
beta2 = 10
beta3 = 0
x1 = rnorm(1000000,1,1)
x2 = rnorm(1000000,1,1)
x3 = rnorm(1000000,1,1)
Tr = rnorm(1000000,1,1)
linpred = intercept + x1*beta1 + x2*beta2 + x3*beta3
prob = exp(linpred)/(1 + exp(linpred))
runis = runif(1000000,0,1)
y = ifelse(runis < prob,1,0)

###Writing cohen’s d function
###Cohen’s d (pooled within group SD)
cohen.d<-function(DV,groupv)
{
varg1<-by(DV, groupv,var)[[1]]
varg2<-by(DV, groupv,var)[[2]]
ng1<-by(DV, groupv,length)[[1]]
ng2<-by(DV, groupv,length)[[2]]
pooledvar<=((ng1-1)*varg1+(ng2-1)*varg2)/(ng1 + ng2 -2)
pooledsd<=sqrt(pooledvar)
meang1<=by(DV, groupv,mean)[[1]]
meang2<=by(DV, groupv,mean)[[2]]
meandif<=meang1-meang2
cohd<-meandif/pooledsd
varratio<-varg1/varg2
list(pooled.SD=pooledsd, Group.Mean.Dif=meandif,CohenD=cohd,VarRatio=varratio)

###Creating an ATE of .50 in the population
Tr.1<-ifelse(y==1,Tr*1.25,Tr*.75)
massive<-data.frame (x1=x1, x2=x2, x3=x3, Tr=Tr,Tr.1=Tr.1, y=y)
write.table (massive, file = "massive90.csv", sep ="", col.names =)

t.test(Tr.1~y, alternative="two.sided",var.equal=TRUE, conf.level=.95)
T.effect<-cohen.d(Tr.1,y)$CohenD
T.effect

###Propensity scores of all cases in the population
glm.massive<-glm(y~x1+x2+x3, massive, family=binomial(link="logit"))
summary(glm.massive)

####Calculating predicted values from the population.
pred.glm.massive <- predict(glm.massive,type="response")

####Calculation of initial bias in propensity scores in the population
Initial.bias<-cohen.d(pred.glm.massive,glm.massive$y)$CohenD

####Calculation of the hit rate given a set of covariates
ClassLog (glm.massive, massive$y)
b<-ClassLog (glm.massive, massive$y)
b$overall
b$mcFadden

####Calculation of the I Index in the population given the improvement over the null hit rate
outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/1000000)
Improvement<-(b$overall-initialHR)
ChanceImprove<-(1-initialHR)
lindex<-(Improvement[1]/ChanceImprove[1])
lindex

################################################################################
# Test Sample for sample size 100 #
################################################################################

####Draw a random sample of 100 people from the massive data set
matr<-sample(nrow(massive), 1500, replace=T)
samp.100<-data.frame(massive[matr,])

88
##### Run the logistic regression on the sample

```r
glm1 <- glm(y ~ x1 + x2 + x3, samp.100, family = binomial(link = "logit"))
summary(glm1)
```

##### Calculating predicted values from the sample.

```r
pred <- predict(glm1, type = "response")
pred
```

##### T-test of propensity scores

```r
t.test(pred ~ glm1$y, alternative = "two.sided", var.equal = TRUE, conf.level = .95)
```

##### Calculation of initial bias in propensity scores

```r
Initial.bias <- cohen.d(pred, glm1$y)$CohenD
Initial.bias
```

##### Calculation of the hit rate given a set of covariates

```r
ClassLog (glm1, samp.100$y)
b <- ClassLog (glm1, samp.100$y)
b$overall
b$mcFadden
```

##### Calculation of the I Index given the improvement over the null hit rate

```r
outcome <- b$rawtab[1, 2] + b$rawtab[2, 2]
initialHR <- (outcome / 100)
Improvement <- (b$overall - initialHR)
ChancelImprove <- (1 - initialHR)
Iindex <- (Improvement[1] / ChancelImprove[1])
Iindex
```

##### Caliper Matching on covariates with replacement and a caliper of .25

##### P values of the ATE were replicated 1000 times to best estimate the true ATE.

```r
X <- glm1$fitted
Y <- samp.100$Tr.1
Tr <- samp.100$y
rr <- Match(Y = Y, Tr = Tr, X = X, caliper = .25, M = 1, replace = TRUE)
summary(rr)
mb <- MatchBalance(y ~ x1 + x2 + x3, data = samp.100, match.out = rr, nboots = 1000)
new <- c(rr$mdata$X[1, ], rr$mdata$X[2, ])
```

##### T-test of propensity scores on the matched sample

```r
t.test(new ~ rr$mdata$Tr, alternative = "two.sided", var.equal = TRUE, conf.level = .95)
```

##### Calculation of bias in the propensity scores after matching

```r
Post.bias <- cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff <- (abs(Initial.bias) - abs(Post.bias))
bias.red <- (bias.diff / abs(Initial.bias))
bias.red
```
abs(Initial.bias)

#### Sensitivity Analysis with a Gamma of 2.5 evaluated at increments of .05
psens(rr, Gamma=6.0, GammaInc=.05)
a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)

#### Place data into data file.
output<-data.frame((t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1], Initial.bias, Post.bias, bias.red, t(glm1$coefficients), rr[15])
write.table (output, file = "output100.csv", sep = ",", col.names =,)

# Simulation of 1000 draws from the population for Sample Size of 100 #
set.seed(38006)
for(i in 1:1000) {
  matr<sample(nrow(massive), 100, replace=T)
samp.100<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.100, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred,glm1$y)$CohenD
ClassLog (glm1, samp.100$y)
b<-ClassLog (glm1, samp.100$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/100)
Improvement<-(b$overall-initialHR)
ChanceImprove<-(1-initialHR)
lindex<-(Improvement[1]/ChanceImprove[1])
X <- glm1$fitted
Y <- samp.100$Tr.1
Tr <- samp.100$y
rr <- Match(Y=Y, Tr=Tr, X=X, M=1,caliper = .25, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.100, match.out=rr, nboots=100)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])

t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohend(new, rr$mdata$Tr)$CohenD
bias.diff<-(-abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(-bias.diff/abs(Initial.bias))
psens(rr, Gamma = 6.0, GammaInc = .05)
a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1],
abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
colnames(output) <-c("1.0", "1.05", "1.10", "1.15", "1.20", "1.25", "1.30", "1.35", "1.40",
"1.45", "1.50", "1.55", "1.60", "1.65", "1.70", "1.75", "1.80", "1.85", "1.90", "1.95", "2.00",
"2.05", "2.10", "2.15", "2.20", "2.25", "2.30", "2.35", "2.40", "2.45", "2.50", "2.55", "2.60",
"2.65", "2.70", "2.75", "2.80", "2.85", "2.90", "2.95", "3.00", "3.05", "3.10", "3.15", "3.20",
"5.05", "5.10", "5.15", "5.20", "5.25", "5.30", "5.35", "5.40", "5.45", "5.50", "5.55", "5.60",
"5.65", "5.70", "5.75", "5.80", "5.85", "5.90", "5.95", "6.00", "ATE", "Iindex", "mcFadden",
"InitialBias", "PostBias", "BiasReduc", "intercept", "X1", "X2", "X3", "numobs")
write.table (output, file = "output100.csv", sep =",", col.names =,)

# Simulation of 1000 draws from the population for Sample Size of 400 #

set.seed(39007)
for(i in 1:2) {
matr<-sample(nrow(massive), 400, replace=T)
samp.400<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.400, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohend(pred,glm1$y)$CohenD
ClassLog (glm1, samp.400$y)
b<-ClassLog (glm1, samp.400$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/400)
Improvement<-(b$overallsd-initialHR)
ChanceImprove<-(1-initialHR)
lindex<-(-Improvement[1]/ChanceImprove[1])
X <- glm1$fitted
Y <- samp.400$Tr.1
Tr <- samp.400$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.400, match.out=rr, nboots=100)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])

# # # Simulation of 1000 draws from the population for Sample Size of 400 #
# # #
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(-abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=6.0, Gammalnc=.05)
a<-psens(rr, Gamma = 6.0, Gammalnc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1],
abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
colnames(output) <-c("1.0", "1.05", "1.10", "1.15", "1.20", "1.25", "1.30", "1.35", "1.40",
"1.45", "1.50", "1.55", "1.60", "1.65", "1.70", "1.75", "1.80", "1.85", "1.90", "1.95", "2.00",
"2.05", "2.10", "2.15", "2.20", "2.25", "2.30", "2.35", "2.40", "2.45", "2.50", "2.55", "2.60",
"2.65", "2.70", "2.75", "2.80", "2.85", "2.90", "2.95", "3.00", "3.05", "3.10", "3.15", "3.20",
"5.05", "5.10", "5.15", "5.20", "5.25", "5.30", "5.35", "5.40", "5.45", "5.50", "5.55", "5.60",
"5.65", "5.70", "5.75", "5.80", "5.85", "5.90", "5.95", "6.00", "ATE", "Iindex", "mcFadden",
"InitialBias", "PostBias", "BiasReduce", "intercept", "X1", "X2", "X3", "numobs")
write.table (output, file = "output400.csv", sep =",", col.names =)

set.seed(40008)
for(i in 1:1000) {
matr<-sample(nrow(massive), 700, replace=T)
samp.700<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.700, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred,glm1$y)$CohenD
ClassLog (glm1, samp.700$y)
b<-ClassLog (glm1, samp.700$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/700)
Improvement<-(b$overall-initialHR)
ChancelImprove<-(1-initialHR)
lindex<-(Improvement[1]/ChancelImprove[1])
X <- glm1$fitted
Y <- samp.700$Tr.1
Tr <- samp.700$Tr
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.700, match.out=rr, nboots=100)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-((abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=6.0, GammaInc=.05)
a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1],
abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
colnames(output) <-c("1.0", "1.05", "1.10", "1.15", "1.20", "1.25", "1.30", "1.35", "1.40",
"1.45", "1.50", "1.55", "1.60", "1.65", "1.70", "1.75", "1.80", "1.85", "1.90", "1.95", "2.00",
"2.05", "2.10", "2.15", "2.20", "2.25", "2.30", "2.35", "2.40", "2.45", "2.50", "2.55", "2.60",
"2.65", "2.70", "2.75", "2.80", "2.85", "2.90", "2.95", "3.00", "3.05", "3.10", "3.15", "3.20",
"5.05", "5.10", "5.15", "5.20", "5.25", "5.30", "5.35", "5.40", "5.45", "5.50", "5.55", "5.60",
"5.65", "5.70", "5.75", "5.80", "5.85", "5.90", "5.95", "6.00", "ATE", "Iindex", "mcFadden",
"InitialBias", "PostBias", "BiasReduc", "intercept", "X1", "X2", "X3", "numobs")
write.table (output, file = "output700.csv", sep =",", col.names =,)

#################################################################################
# Simulation of 1000 draws from the population for Sample Size of 1000  #
#################################################################################

colnames(output) <-c("1.0", "1.05", "1.10", "1.15", "1.20", "1.25", "1.30", "1.35", "1.40",
"1.45", "1.50", "1.55", "1.60", "1.65", "1.70", "1.75", "1.80", "1.85", "1.90", "1.95", "2.00",
"2.05", "2.10", "2.15", "2.20", "2.25", "2.30", "2.35", "2.40", "2.45", "2.50", "2.55", "2.60",
"2.65", "2.70", "2.75", "2.80", "2.85", "2.90", "2.95", "3.00", "3.05", "3.10", "3.15", "3.20",
"5.05", "5.10", "5.15", "5.20", "5.25", "5.30", "5.35", "5.40", "5.45", "5.50", "5.55", "5.60",
"5.65", "5.70", "5.75", "5.80", "5.85", "5.90", "5.95", "6.00", "ATE", "Iindex", "mcFadden",
"InitialBias", "PostBias", "BiasReduc", "intercept", "X1", "X2", "X3", "numobs")
write.table (output, file = "output700.csv", sep =",", col.names =,)

set.seed(41009)
for(i in 1:1000) {
matr<-sample(nrow(massive), 1000, replace=T)
samp.1000<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.1000, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred,glm1$y)$CohenD
ClassLog (glm1, samp.1000$y)
b<-ClassLog (glm1, samp.1000$y)
if (nrow(b$rawtab)>1)outcome<b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/1000)
Improvement<-(b$overall-initialHR)
ChancelImprove<-(1-initialHR)
lindex<-(Improvement[1]/ChancelImprove[1])
X <- glm1$fitted
Y <- samp.1000$Tr.1
Tr <- samp.1000$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.1000, match.out=rr, nboots=100)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
}
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-((abs(Initial.bias)-(abs(Post.bias)))
bias.red<-((bias.diff)/(abs(Initial.bias)))
psens(rr, Gamma=6.0, GammaInc=.05)
a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1],
abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
colnames(output) <-c("1.0", "1.05", "1.10", "1.15", "1.20", "1.25", "1.30", "1.35", "1.40",
"1.45", "1.50", "1.55", "1.60", "1.65", "1.70", "1.75", "1.80", "1.85", "1.90", "1.95", "2.00",
"2.05", "2.10", "2.15", "2.20", "2.25", "2.30", "2.35", "2.40", "2.45", "2.50", "2.55", "2.60",
"2.65", "2.70", "2.75", "2.80", "2.85", "2.90", "2.95", "3.00", "3.05", "3.10", "3.15", "3.20",
"5.05", "5.10", "5.15", "5.20", "5.25", "5.30", "5.35", "5.40", "5.45", "5.50", "5.55", "5.60",
"5.65", "5.70", "5.75", "5.80", "5.85", "5.90", "5.95", "6.00", "ATE", "Iindex", "mcFadden",
"InitialBias", "PostBias", "BiasReduc", "intercept", "X1", "X2", "X3", "numobs")
write.table (output, file = "output1000.csv", sep =","", col.names =,)

######################################################################
#####      Simulation of 1000 draws from the population for Sample Size of 1500    ####
######################################################################
set.seed(42010)
for(i in 1:1000) {
matr<-sample(nrow(massive), 1500, replace=T)
samp.1500<-data.frame(massive[matr,])
glm1<-glm(y~x1+x2+x3, samp.1500, family=binomial(link="logit"))
pred <- predict(glm1,type="response")
Initial.bias<-cohen.d(pred,glm1$y)$CohenD
ClassLog (glm1, samp.1500$y)
b<-ClassLog (glm1, samp.1500$y)
if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
initialHR<-(outcome/1500)
Improvement<-(b$overall-initialHR)
ChancelImprove<-(1-initialHR)
Iindex<-(Improvement[1]/ChancelImprove[1])
X <- glm1$fitted
Y <- samp.1500$Tr.1
Tr <- samp.1500$y
rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
mb <- MatchBalance(y~x1+x2+x3, data=samp.1500, match.out=rr, nboots=100)
new<-c(rr$mdata$X[1,],rr$mdata$X[2,])}
```r
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=6.0, Gammalnc=.05)
a<-psens(rr, Gamma = 6.0, Gammalnc = 0.05)
a$bounds
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1],
abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
write.table (output, file = "output1500.csv", sep =",", col.names =)
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(-abs(Initial.bias)-abs(Post.bias))
bias.red<-bias.diff/abs(Initial.bias)
psens(rr, Gamma=6.0, GammaInc=.05)
a<-psens(rr, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1],
abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
colnames(output) <-c("1.0", "1.05", "1.10", "1.15", "1.20", "1.25", "1.30", "1.35", "1.40",
"1.45", "1.50", "1.55", "1.60", "1.65", "1.70", "1.75", "1.80", "1.85", "1.90", "1.95", "2.00",
"2.05", "2.10", "2.15", "2.20", "2.25", "2.30", "2.35", "2.40", "2.45", "2.50", "2.55", "2.60",
"2.65", "2.70", "2.75", "2.80", "2.85", "2.90", "2.95", "3.00", "3.05", "3.10", "3.15", "3.20",
"5.05", "5.10", "5.15", "5.20", "5.25", "5.30", "5.35", "5.40", "5.45", "5.50", "5.55", "5.60",
"5.65", "5.70", "5.75", "5.80", "5.85", "5.90", "5.95", "6.00", "ATE", "Iindex", "mcFadden",
"InitialBias", "PostBias", "BiasReduc", "intercept", "X1", "X2", "X3", "numobs")
write.table (output, file = "output2000.csv", sep =",", col.names =,)

#########################################################################
####      Simulation of 1000 draws from the population for Sample Size of 2500    ####
#########################################################################
set.seed(44012)
for(i in 1:1000) {
  matr<-sample(nrow(massive), 2500, replace=T)
samp.2500<-data.frame(massive[matr,])
 glm1<-glm(y~x1+x2+x3, samp.2500, family=binomial(link="logit"))
  pred <- predict(glm1,type="response")
  Initial.bias<-cohen.d(pred,glm1$y)$CohenD
  ClassLog (glm1, samp.2500$y)
  if (nrow(b$rawtab)>1)outcome<-b$rawtab[1,2]+b$rawtab[2,2]
  initialHR<-(outcome/2500)
  Improvement<-(b$overall-initialHR)
  ChancelImprove<-(1-initialHR)
  lindex<-(Improvement[1]/ChancelImprove[1])
  X <- glm1$fitted
  Y <- samp.2500$Tr.1
  Tr <- samp.2500$y
  rr <- Match(Y=Y, Tr=Tr, X=X, caliper = .25, M=1, replace=TRUE)
  mb <- MatchBalance(y~x1+x2+x3, data=samp.2500, match.out=rr, nboots=100)
  new<-c(rr$mdata$X[1,],rr$mdata$X[2,])
}
t.test(new~rr$mdata$Tr, alternative="two.sided", var.equal=TRUE, conf.level=.95)
Post.bias<-cohen.d(new, rr$mdata$Tr)$CohenD
bias.diff<-(abs(Initial.bias)-(abs(Post.bias)))
bias.red<-(bias.diff/abs(Initial.bias))
psens(rr, Gamma=6.0, GammaInc=.05)
a<-psens(r, Gamma = 6.0, GammaInc = 0.05)
a$bounds
t(a$bounds)
output[i,]<-data.frame(cbind(t(a$bounds[,3])),rr[1], Iindex, b$mcFadden[1],
abs(Initial.bias), abs(Post.bias), bias.red, t(glm1$coefficients), rr[15])
}
colnames(output) <-c("1.0", "1.05", "1.10", "1.15", "1.20", "1.25", "1.30", "1.35", "1.40", 
"1.45", "1.50", "1.55", "1.60", "1.65", "1.70", "1.75", "1.80", "1.85", "1.90", "1.95", "2.00", 
"2.05", "2.10", "2.15", "2.20", "2.25", "2.30", "2.35", "2.40", "2.45", "2.50", "2.55", "2.60", 
"2.65", "2.70", "2.75", "2.80", "2.85", "2.90", "2.95", "3.00", "3.05", "3.10", "3.15", "3.20", 
"5.05", "5.10", "5.15", "5.20", "5.25", "5.30", "5.35", "5.40", "5.45", "5.50", "5.55", "5.60", 
"5.65", "5.70", "5.75", "5.80", "5.85", "5.90", "5.95", "6.00", "ATE", "Iindex", "mcFadden", 
"InitialBias", "PostBias", "BiasReduc", "intercept", "X1", "X2", "X3", "numobs")
write.table (output, file = "output2500.csv", sep =",", col.names =)
REFERENCES


