AN ABSTRACT OF THE THESIS OF

Daniel K. Yang for the degree of Doctor of Philosophy in Statistics presented on December 9, 2011.
Title: Propensity Score Adjustments Using Covariates in Observational Studies

Abstract approved:

_____

Virginia M. Lesser Alix I. Gitelman

In this thesis we develop a theoretical framework for the identification of situations where the equal frequency (EF) or equal variance (EV) subclassification may produce lower bias and/or variance of the estimator. We conduct simulation studies to examine the EF and EV approaches under different types of model misspecification. We apply two weighting schemes in our simulations: equal weights (EW) and inverse variance (IV) weights. Our simulation results indicate that under the quadratic term misspecification, the EF-IV estimator provides the lowest bias and root mean square error as compared to the ordinary least square estimator and other propensity score estimators. Our theorem development demonstrates that if higher variation occurs with larger bias for within subclass treatment effect estimates then the EF-IV estimator has a smaller overall bias than the EF-EW estimator. We show that the EF-IV estimator always has a smaller variance than the EF-EW estimator. We also propose a novel method of subclassification that focuses on creating homogeneous propensity score subclasses to produce an estimator with reduced biased in some circumstances. We feel our research contributes to the field of propensity score adjustments by providing new theorems to compare the overall bias and variance between different propensity score estimators.

Propensity Score Adjustments Using Covariates in Observational Studies

by
Daniel K. Yang

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented December 9, 2011
Commencement June 2012

Doctor of Philosophy thesis of Daniel K. Yang presented on December 9, 2011.

APRROVED:

Co-Major Professor, representing Statistics

Co-Major Professor, representing Statistics

Chair of the Department of Statistics

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Daniel K. Yang, Author

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation and gratitude to my advisors, Dr. Virginia M. Lesser and Dr. Alix I. Gitelman for their patience, persistence, support, and guidance during the course of this dissertation research. I especially thank Dr. Lesser for her tireless mentorship and advice throughout my doctoral program and for the opportunity to work on real surveys and sampling problems in the Survey Research Center. Without her unreserved help and encouragement, this dissertation would not have been finished. I cannot thank her enough for giving me the opportunity to attend and benefit from the 2010 Joint Statistical Meetings in Vancouver, BC. I also sincerely thank Dr. Gitelman, for her valuable suggestions, constant advice and help during the preparation of this dissertation. I am indebted to both of my advisers for their mentorship, extraordinary helpfulness and availability during my research and preparation of this dissertation.

I would also like to heartily express my thanks to Dr. David S. Birkes, who taught most of the Statistics PhD classes, for his support, unreserved advice and help on my research and dissertation preparation, and for serving on my committee. I offer special thanks to Dr. Daniel W. Schafer for taking the time to read this dissertation and for serving on my committee. It is my pleasure to thank Dr. Katherine Gunter for reviewing the dissertation and for serving as my graduate council representative. I would like to thank the entire faculty, staff and students of the Statistics Department at Oregon State University for their contribution to my education and research.

I wholeheartedly thank my family for their love, faith and support, especially my mother. I am grateful for the sacrifices she has made during my time as a graduate student. Her unequivocal encouragement and endless support has been indispensable to the successful completion of my doctoral studies.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF APPENDIX FIGURES

LIST OF APPENDIX TABLES

LIST OF APPENDIX TABLES (Continued)

Table | Page

Propensity Score Adjustments for Covariates in Observational Studies

Chapter 1. INTRODUCTION

This chapter provides a general introduction of the thesis. In section 1.1, we describe missing data and the nonresponse problem. In section 1.2, we introduce the response propensity score adjustment as a method to adjust for nonresponse. In section 1.3, we provide a brief review of propensity score methods, and in section 1.4 we summarize the contribution and organization of this thesis.

1.1 Nonresponse and its mechanism

Missing data is a common problem in the data collection process when the designated information of interest cannot be collected for a portion of the sampling units (Little and Rubin 2002). In survey sampling, this problem is called nonresponse. For example, in a mail survey, an invalid mailing address or a sampled subject choosing not to answer the questionnaire will both lead to nonresponse. In an economic status survey, subjects may be sensitive and reluctant to answer questions relating to income level, which leads to nonresponse. In a medical therapy study, patients may drop out during the study period due to adverse side effects or lack of treatment benefits, leading to missing data. Ignoring these observations can lead to biased results. This thesis investigates ways to adjust for nonresponse.

In survey sampling, there are two types of nonresponse: unit nonresponse and item nonresponse. Unit nonresponse implies that the entire unit or subject is missing. In item nonresponse, the subject has not answered all the questions asked. Only a portion of completed questions are obtained for that specific subject.

Nonresponse or missing data can be classified into three mechanisms. They are MCAR (missing complete at random), MAR (missing at random or ignorable) and nonignorable (informative) missing (Lohr 1999, Little and Rubin 2002). In order to describe these mechanisms, let $\mathbf{Y}$ be the complete data matrix for the outcome variable of the sample, and it can be partitioned as $\mathbf{Y} = (\mathbf{Y}^o, \mathbf{Y}^m)$, where $\mathbf{Y}^o$ represents the observed outcome and $\mathbf{Y}^m$ represents the missing outcome; let $\mathbf{X}$ be the data matrix of covariates; let $\mathbf{R}$ be a vector of response/nonresponse indicator; let θ and φ be two sets of distinguish parameters corresponding to $\mathbf{Y}^o$, $\mathbf{R}$, respectively. The three mechanisms can be described as follows.

Under MCAR, the nonresponses are independent of the outcome variable and the covariates. This implies that respondents are representative of the selected sample. The conditional probability density function (pdf.) of $\mathbf{R}$ can be simplified as $f_\varphi(\mathbf{R} \mid \mathbf{Y}, \mathbf{X}) = f_\varphi(\mathbf{R})$, which implies the joint pdf. $f_{\theta,\varphi}(\mathbf{Y}^o, \mathbf{R}) = f_\varphi(\mathbf{R}) f_\theta(\mathbf{Y}^o)$. For example, if a severe weather condition makes it impossible for a survey interviewer to meet with a sampled subject, or an electronic glitch causes some records to be lost in a medical study, this kind of missingness may have no relation to the outcome or covariates. Therefore, the probability of response is independent of the observed outcomes, and the researcher can

analyze the completed data with no adjustment needed. The sample results are representative of the population.

Under MAR, the nonresponses depend on the covariates but not the missing outcome variable (Lohr 1999). The conditional pdf. of **R** is then $f_\varphi(\mathbf{R} \mid \mathbf{Y}, \mathbf{X}) = f_\varphi(\mathbf{R} \mid \mathbf{X})$. For example, the younger sampling subjects tend to refuse participation more often than older subjects. In some cases, subjects with a higher level of education are more willing to participate than subjects with a lower level of education. This type of missing mechanism can be explained by the corresponding covariates (e.g., age, education). MAR is also referred to as ignorable. If the missingness has no relation to the missing outcome then the covariates can be used to make nonresponse adjustments. Once the adjustment is implemented, the missing mechanism resembles MCAR.

The nonignorable missing mechanism is the most difficult situation because the nonresponse mechanism depends on the missing outcome and cannot be fully explained by the covariates. The conditional pdf. of **R**, $f_\varphi(\mathbf{R} \mid \mathbf{Y}^o, \mathbf{Y}^m)$ does not have a simple expression, since the mechanism of nonresponse is more complex. An example of this can be illustrated in a state tax survey. Upper income level households may be less likely to respond to the survey and reveal their actual income than middle or lower income households, even though they have similar demographics in age category, race and gender. For nonignorable missingness, strong modeling assumptions are necessary to account for nonresponse. The analysis based only on the completed data or based on the

observed outcome and covariates will generate bias in the estimates (Little and Rubin 2002).

1.2 The response propensity scores adjustment

Recently, some scholars suggest that MAR is the mechanism of nonresponse often found in most situations when missing data occurs (Bjørnstad 2007 and Chambers 2007). We will examine methods to adjust for the unit nonresponse under the MAR (ignorable) mechanism. Lohr (1999) describes three traditional methods to deal with MAR in survey sampling: (1) weighting class adjustment (WCA), (2) post-stratification, and (3) imputation. For the weighting class adjustment, demographic information is available on the entire sample. Weighting classes are created, often using the demographic information. The weighting class adjustment assumes that within a weighting class, the responses obtained from the responding subjects and nonresponding subjects are similar. In addition, there is a differential response rate across demographic groups. The idea of this method is to develop adjustment weights for the respondents in order to bring the balance of the demographics of the sample more in line to the population demographics. Since the responses from the nonresponding individuals are not available, the WCA method uses demographic information obtained for the "original" sample (respondents and nonrespondents) to form non-overlapping subsets or weighting classes. For example, age can be used to create non-overlapping weighting classes. Response rates often differ across age groups, with older individuals being more likely to respond than younger individuals. Let π_i be the selection probability of subject i in the

original sample, then the sampling weight of subject i is $w_i = 1 / \pi_i$. Let c be a weighting class ($c \in Z^+$, i.e. positive integers); let $sum(w_c^r)$ denote the summation of sampling weights for respondents in weighting class c, and let $sum(w_c)$ denote the summation of weights for the original sample in weighting class c. The response probability for each weighting class is estimated by $\hat{p}_c = sum(w_c^r) / sum(w^r)$. Hence, the WCA method adjusts the weight of a respondent in weighting class c by multiplying $1 / \hat{p}_c$ (Lohr 1999).

The post-stratification adjustment (PSA) is used to bring the sample more in line with the population, as what was done with the WCA. However, in this case, no demographic information is available on the nonrespondents. Population information obtained from an external source is used to adjust the weights. For example, Census data is often used in national population surveys to calibrate the sample results to population numbers. Post-strata are created based on demographic information collected on the questionnaire from the respondents along with information available on the same demographic variables from the Census. For example, age groups can be used to create the post-strata. The number of individuals in the sample may be out of balance with proportions of these age groups in the population. The post-strata are created to make weighting adjustments to bring the sample more in line with the population proportions. Details on this adjustment are provided in Lohr (1999).

The imputation technique is used to assign or impute values to the values or items that are missing. For example, if the income variable has some missing values, the

researcher can replace the missing values by substituting in the mean of the observed income (mean imputation). Another approach substitutes the missing values with predicted income values from a regression model of the observed income vs. observed covariates. For interested readers, Lohr (1999) and Heeringa et al. (2010) provide more detailed descriptions about the imputation techniques.

David et al. (1983) provide the fundamental assumption for response propensity, which is that "the response (or nonresponse) mechanism is ignorable." Little (1986) develops the response propensity score method to adjust the estimated mean. He defines the response propensity score as the conditional probability of a subject responding to the survey, given the observed covariates. The response propensity scores method can be viewed as an alternative to the WCA; it has a model-based element, since the response propensity scores are generally estimated by a logistic model on all observed covariates. Then the weighting classes (a.k.a subclasses) are formed on the estimated response propensity scores. The majority of researchers currently choose to form equally-divided weighting classes or subclasses on the estimated response propensity scores, and to apply equal weights (EW) to adjust the nonresponse when estimating the outcome.

1.3 The propensity score methods

To understand propensity score methods, we will provide a general review. Cochran (1968) suggests that for an observational study with only one covariate and a binary treatment indicator (e.g. treatment vs. control), a five-20%-quantile equal

frequency (EF) subclassification method is an efficient procedure to eliminate 90% of the bias of the estimated treatment effect induced by this covariate. However, this procedure may not be feasible if there are multiple covariates collected in a study. Let the total number of covariates be p, where $p \in Z^+$. If there was a dichotomous subclassification (EF with two subclasses) for each covariate, using Cochran's (1968) approach, 2^p subclasses are produced. As p increases, the data may not be able to support the total number of 2^p subclasses.

For multiple covariates, one can also consider using regression adjustment to estimate the treatment effect, but Rubin (1979) questions that the regression estimate may not be appropriate if the linear model is incorrect. Rosenbaum and Rubin (1983) indicate that the expectation of the quadratic term of the conditional bias ("expected squared bias") for the regression estimate of the treatment effect increases if the covariate variances (or "covariance matrices") of the treatment and control group are different. Further, Rosenbaum and Rubin (1983) define a subject's propensity score as the conditional probability for each subject to receive the treatment assignment given the observed covariates. They provide the theory of the propensity scores method to adjust the bias of the estimated treatment effect due to the observed multiple covariates. In Rosenbaum and Rubin (1984), they provide another theorem corresponding to Cochran's (1968) results. This theorem implies that by using five-20%-quantile subclassification groups with equal weights (EF-EW) on the estimated propensity scores reduces 90% of the bias of the estimated treatment effect caused by all observed covariates in a observational study with binary treatments (e.g., treatment vs. control, participants vs.

non-participants). Rosenbaum (1987) introduces a propensity scores method for post-stratification adjustment. If a study has an overwhelmingly larger number of control units than treatment units, researchers may want to select a portion of the control units to compare with the treatment units to estimate the treatment effect. Rosenbaum and Rubin (1985) develop the propensity scores matching approach to select control units, where their estimated propensity scores are similar to treatment units based on certain criteria. Rubin and Thomas (1992) describe the propensity scores matching approach using discriminant analysis and logistic regression to estimate the propensity scores.

Drake (1993) introduces a simulation approach to evaluate the propensity score estimators and the ordinary least square (OLS) estimators by omitting an independent covariate and misspecifying a quadratic term. Dehejia and Wahba (1999, 2002) propose to further split EF subclasses until the difference of the means of the propensity scores between the dichotomous treatment groups in a split subclass is homogeneous (or "balanced" in Imbens 2004) from a two-sample t-test. Hullsiek and Louis (2002) propose "equal variance" (EV) subclassification on the propensity scores, which involves inverse variance (IV) weighting of the treatment effect estimator. Caliendo and Kopeinig (2008) provide a recent review for the propensity score methodology, including: matching (nearest neighbor, caliper and radius), subclassification (a.k.a. interval matching, blocking or stratification) and inverse propensity scores weighting.

1.4 Summary of the contribution and organization of the thesis

The objective of this thesis work is to develop a theoretical framework to identify situations in which EF or EV approaches might be better in terms of reducing the bias and/or variance of the estimator, and to propose a novel approach of forming the subclasses with homogeneous propensity scores.

The first contribution of the thesis is to evaluate the equal variance subclassification method under model misspecification. The second contribution is a theoretical investigation on propensity score subclassification adjustment estimators to identify at what condition the EF-IV estimator has smaller bias than the EF-EW estimator, and at what condition the EF-IV estimator has no larger variance than the EV-IV estimator. The third contribution is the novel propensity scores balancing subclassification approach. We propose a new approach of subclassification with homogeneous or identical propensity scores between treatment subjects and control subjects (at least for most of the subclasses). This approach tends to produce a large number of subclasses where the data would allow in order to reduce more bias, which also reflects the suggestion by Cochran (1968), Imbens (2004) and Myers and Louis (2007).

The organization of the thesis is the following: In Chapter Two, we will provide a literature review of the background of propensity score methodology, response propensity scores adjustment on survey nonresponse, details of the contribution and organization of the thesis. In Chapter Three, we will provide a simulation study to

evaluate the propensity score adjustment estimators, with respect to EF and EV subclassification approaches; we will use different weighting methods, such as EW or IV weights, of the treatment effect. In Chapter Four, we will provide the theoretical work of the bias, variance and root mean square error (RMSE) of the propensity score subclassification adjustment estimators with different weighting methods. In Chapter Five, we propose our novel propensity scores balancing (PSB) subclassification approach. We then simulate a situation with a proportion of control subjects that have low estimated propensity scores. We also provide an evaluation for the PSB approach and other propensity score estimators in terms of average bias, variance and corresponding RMSE. In Chapter Six, we will summarize the major findings from our simulation results, theory work and our newly proposed PSB subclassification approach and discuss future research interests.

Chapter 2. LITERATURE REVIEW

There are often respondents and nonrespondents in surveys. If respondents provide different responses as that expected from nonrespondents, then analyzing only the completed data will produce biased results. If the nonresponse mechanism can be explained by the observed covariates, the analysis should take into account the observed covariates to adjust for nonresponses. This can be viewed as an analog to observational studies that include a treatment group and a control group. The objective of this thesis is to seek an adjustment method using covariates information available in observational studies or surveys that encounter nonresponse. To discuss the adjustment for survey nonresponse, we will first examine adjustment methods in the context of observational studies with a treatment group and a control group. We will then extend these methods to survey nonresponse adjustment. This chapter reviews the literature on single covariate subclassification methods as well as propensity score subclassification methods and their applications to survey nonresponse. In section 2.1, we describe the equal frequency subclassification adjustment on a single covariate. In section 2.2, we discuss the adjustment using propensity scores subclassification. Following this development, section 2.3 reviews applications of the propensity scores method, including a survey nonresponse adjustment. Section 2.4 reviews the evaluation of the propensity scores equal frequency subclassification under model misspecification and the equal variance subclassification approach. Section 2.5 outlines the contribution of this thesis, and section 2.6 describes the organization of the thesis.

2.1 The single covariate quintile subclassification adjustment

In observational studies with a treatment group and a control group, researchers are often interested in comparing the means of the outcomes between the treatment group and the control group to estimate the treatment effect. There may be an imbalance in those covariates across the treatment group and the control group. A direct comparison of the means of the outcomes between the treatment group and the control group may be biased in estimating the treatment effect due to this imbalance. The effect of the covariates cannot be separated from the effect of the treatment on the outcome. In surveys, the distribution of some demographics (e.g. age or education level) may be very different between the respondent and nonrespondent groups. If this imbalance is not taken into account, the estimates for the respondents are biased, since these estimates also reflect demographic differences.

Cochran (1968) initially examined this problem in observational studies involving a treatment group, and a control group and one covariate related to the outcome. He was interested in developing an adjustment to account for the imbalance due to the covariate. His solution is described as the single covariate equal frequency (EF) subclassification adjustment. This method equally divides the covariate along the percentiles of its empirical distribution based on a predetermined number of subclasses. The subclasses are non-overlapping. Within each subclass, the distribution of the covariate is relatively similar between the treatment group and the control group. The difference in the means

of the outcome between the treatment group and the control group is obtained for each subclass. These subclass-specific means are combined with weights to compute the overall estimate of the treatment effect. A common weighting scheme used in this application is the equal weights (EW) approach, in which the weight is the reciprocal of the number of subclasses. Below, a brief description of the single covariate equal frequency subclassification adjustment developed by Cochran (1968) is presented.

Notionally, let y_{0j} denote the outcome for subject j in the control group, where j = 1, 2, …, n_0; n_0 is the number of subjects from the control group ($n_0 > 1$). Let y_{1i} denote the outcome for subject i in the treatment group, where i = 1, 2, …, n_1; n_1 is the number of subjects from the treatment group ($n_1 > 1$). Let $x_{01}, x_{02}, ..., x_{0n_0}$ and $x_{11}, x_{12}, ..., x_{1n_1}$ denote the corresponding covariate in the control group and the treatment group, respectively. For now, we assume that both the outcome and covariate are continuous. Suppose the covariate relates to the outcome such that,

$$y_{1i} = \alpha_1 + u(x_{1i}) + e_{1i}\ ,\ \ y_{0j} = \alpha_0 + u(x_{0j}) + e_{0j}\ ,$$

where u is a regression function (e.g., $u(x_{1i}) = \beta x_{1i}$, $u(x_{0j}) = \beta x_{0j}$, and β is a regression coefficient). Here, α_1 and α_0 are the true means of the treatment group and the control group, respectively, and e_{1i} and e_{0j} are the corresponding zero mean independent random error terms of the treatment group and the control group. The goal is to estimate the treatment effect $\alpha_1 - \alpha_0$. The unadjusted means of the outcomes of the treatment group and the control group can be obtained by

$$E(\bar{y}_1) = \alpha_1 + E(\bar{u}_1),\ E(\bar{y}_0) = \alpha_0 + E(\bar{u}_0),$$

where $\bar{u}_1 = \sum_{i=1}^{n_1} u(x_{1i})/n_1,\ \bar{u}_0 = \sum_{j=1}^{n_0} u(x_{0j})/n_0$. Thus, a direct comparison of these means will lead to the bias caused by the covariate, $E(\bar{u}_1) - E(\bar{u}_0)$. For the sake of simplicity, let the number of subclasses be five, and let c index the subclasses. Within subclass c, the means of the outcomes of the treatment group and the control group can be obtained by

$$E(\bar{y}_1^{(c)}) = \alpha_1 + E(\bar{u}_1^{(c)}),\ E(\bar{y}_0^{(c)}) = \alpha_0 + E(\bar{u}_0^{(c)}),$$

where $\bar{u}_1^{(c)} = \sum_{i=1}^{n_1^{(c)}} u(x_{1i}^{(c)})/n_1^{(c)},\ \bar{u}_0^{(c)} = \sum_{j=1}^{n_0^{(c)}} u(x_{0j}^{(c)})/n_0^{(c)}$; $n_1^{(c)}$ and $n_0^{(c)}$ are corresponding subclass-specific numbers of subjects from the treatment group and the control group, respectively; and $n_1^{(c)}$, $n_0^{(c)} > 1$. The subclass-specific bias caused by the covariate is then $E(\bar{u}_1^{(c)}) - E(\bar{u}_0^{(c)})$. Suppose the weight of subclass c is w_c, denoting the single covariate equal frequency subclassification adjustment. Then, the overall weighted bias caused by the covariate is

$$\sum w_c [E(\bar{u}_1^{(c)}) - E(\bar{u}_0^{(c)})].$$

Therefore, the proportion of bias removed by the single covariate subclassification adjustment method from the covariate is

$$1 - \sum w_c [E(\bar{u}_1^{(c)}) - E(\bar{u}_0^{(c)})]/[E(\bar{u}_1) - E(\bar{u}_0)]. \qquad (2.1)$$

If using equal weights, then $w_c = 1/5$.

To illustrate the influence of the observed covariate, let us look at an example in Cochran (1968). Three studies are conducted in the U.K., Canada, and the U.S. to investigate the impact of smoking on patient fatality rates. Each study involves three groups: non-smokers, cigarette smokers and cigar/pipe smokers. The unadjusted estimates from these three studies show that the fatality rates of non-smokers and cigarette smokers are similar, while cigar/pipe smokers have higher fatality rates. Across all three studies, the average age of the cigar/pipe smokers is higher than the average age of the cigarette smokers and non-smokers. Obviously, a patient's age, especially for the elderly patients, is also related to a higher fatality rate. An adjustment to account for the covariate (i.e. age) would be advantageous in this example. Using equal weights, the adjusted estimates from all three studies indicate that the fatality rates of cigarette smokers are consistently higher than non-smokers. For cigar/pipe smokers, fatality rates show no increase when compared to non-smokers. It is possible that because of higher ages of cigar/pipe smokers, other medical factors (e.g., patient elevated cholesterol numbers, chronic complication, etc.) may also contribute to their fatalities. Thus, additional covariates from these studies may be needed to further adjust the fatality rates of cigar/pipe smokers.

If using equal weights, equation (2.1) above can be viewed as a function of the covariate. That is, the percentage of bias reduction by using the covariate equal frequency subclassification adjustment will depend on the covariate, not on the outcome. Therefore, equation (2.1) quantifies how effective the single covariate equal frequency subclassification adjustment is in removing the bias introduced by the covariate. The

evaluation of this method was achieved by Cochran's simulations, where certain empirical distributions were assumed for the covariate: Normal, t, Chi-square and Beta. The results provided by Cochran (1968) indicate that for observational studies with a treatment group, a control group and a single covariate related to the outcome variable, using five subclasses with equal weights, the single covariate equal frequency subclassification adjustment removes 90% of the bias introduced by this covariate.

For just one covariate, the equal frequency subclassification adjustment works well to eliminate the bias induced by the covariate. However, when considerable numbers of covariates are observed in observational studies, this method is impractical to implement. In some cases, there is a drastically increased number of subclasses. For example, Rosenbaum and Rubin (1983) discuss a coronary artery bypass surgery study of patients including 74 covariates. If five EF subclasses for each covariate are selected, then a total number of 5^{74} subclasses are needed. If just two EF subclasses for each covariate are selected, the total number of subclasses will be 2^{74}. Under either scenario, the method of subclassification adjustment on each covariate would be very difficult to implement due to the limited number of subjects in each subclass (i.e. very small sample sizes). Therefore, for observational studies with large numbers of covariates, a method incorporating the observed covariates into a univariate measure is desirable. The propensity scores method is developed to deal with this situation.

2.2 The propensity scores equal frequency subclassification adjustment

In observational studies where many covariates are collected, Rosenbaum and Rubin (1983) propose the propensity score method to estimate the treatment effect. The propensity score is defined as the conditional probability of a subject being assigned to the treatment group, given the observed covariates. In a randomized experiment, the propensity score of each subject is known to the researcher at the design stage. For example, in a randomized experiment involving an equal number of individuals in both the treatment group and the control group, the propensity score is 1/2 for each subject. In an observational study, however, treatment is not randomly assigned to subjects. Subsequently, there may be an imbalance in the covariates across the treatment group and the control group. Comparing the outcome means between the treatment group and the control group may be confounded by the covariates. The propensity score summarizes the information from all observed covariates into a univariate measurement. By conditioning on the propensity scores, the subjects can be placed into subclasses with similar covariate information across the treatment group and the control group. Thus, the treatment effect can be estimated within subclasses, thereby reducing the influence of the observed covariates.

For observational studies, another method to account for the covariates when estimating the treatment effect is multiple linear regression incorporating covariates. Rubin (1979) indicates, however, that the estimates using multiple regression may be inappropriate if the linear model is not correct. Rosenbaum and Rubin (1983) describe a

covariate imbalance. For example, if a model for the mean outcome is not a linear function of the covariates, and the covariate variances ("covariance matrices") between the treatment group and the control group are different, then there is a covariate imbalance. Under this condition, they indicate that this covariate imbalance increases the expectation of the quadratic term of the conditional bias ("expected squared bias") of the regression estimate of the treatment effect. When some of the covariates are imbalanced, the propensity score method is an option to consider removing the confounding effect from the observed covariates to adequately estimate the treatment effect.

To further explain the propensity score method, we provide several definitions. A sample is called "*balanced*" if the subjects of the treatment group and the control group can be subclassified, such that the confounding effects of covariates on the treatment effect are removed. For each subject i, i = 1, 2, …, n, the treatment assignment indicator, z_i, is defined as the following: let $z_i = 1$ denote that the subject i is in the treatment group, and $z_i = 0$ denote that the subject i is in the control group. Let $\mathbf{x}_i$ be the vector of observed covariates for one subject. The propensity score, $e(\mathbf{x}_i)$ of a subject is defined as

$$e(\mathbf{x}_i) = \Pr\{z_i = 1 \mid \mathbf{x}_i\} \tag{2.2}$$

As Rosenbaum and Rubin (1983) state, a "*balancing score*," $b(\mathbf{x}_i)$, is a function of the covariates, such that the conditional distribution of $\mathbf{x}_i$ given $b(\mathbf{x}_i)$ is the same for treatment and control subjects. The influence of the covariates on the treatment effect has been removed by subclassifying the subjects based on the balancing score. This

provides the capability to obtain an estimate of the treatment effect without the influence of the covariates.

Another definition used when discussing the propensity scores is the "*strong ignorable treatment assignment*" assumption. When the outcome of interest is conditionally independent of the treatment assignment indicator given the observed covariates, this assumption is met. This assumption is the analog to the missing at random (MAR) assumption (in Chapter One) in the situation when we discuss surveys and nonresponse.

Rosenbaum and Rubin (1983) provide four major theorems to give the theoretical foundation for propensity scores adjustment to reduce bias due to observed covariates in an observational study. The first theorem states that, given the propensity score, the observed covariates and the treatment assignment are conditionally independent. The imbalance in observed covariates is incorporated into the propensity score as a single measurement. This theorem also implies that if a subclass of subjects or a matched treatment-control pair is homogeneous or identical in the propensity scores, then the subjects of the treatment group and the control group in this subclass or matched pair will have the same distribution of covariates. Therefore, the imbalance (distributional differences) of the covariates between the treatment group and the control group are eliminated within this subclass or matched pair. The second theorem implies that the propensity score is a balancing score. In the third theorem, if the treatment assignment is

strongly ignorable given the observed covariates, then the treatment assignment is strongly ignorable given the propensity score.

The fourth theorem states that if we assume a strongly ignorable treatment assignment, then the difference between the conditional expectation of the outcome given the propensity score for the treatment group and the control group equals the expectation of the average treatment effect given the propensity score. That is, under the assumption of the strong ignorable treatment assignment, conditioning on the propensity score is adequate to provide relatively unbiased (i.e. Cochran's 90% bias reduction) estimates of the treatment effect by pair-matching or subclassifying the propensity scores. Subclassification can be implemented by forming subclasses with homogeneous propensity scores. Within each subclass, the treatment effect is estimated by comparing the outcomes between the treatment group and the control group. A graphic display of equal frequency subclassification on the propensity scores is presented in Figure 2.1.

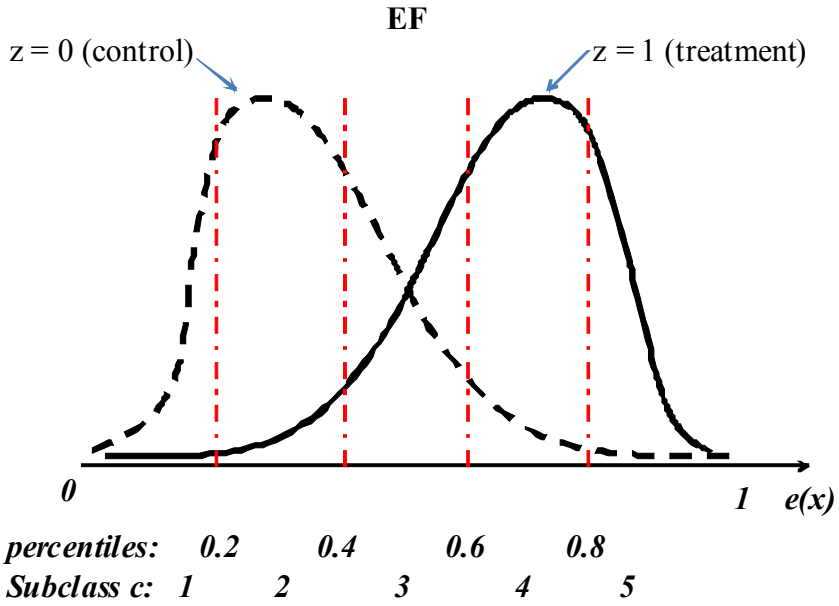

Figure 2. 1 Equal frequency subclassification on the propensity scores

Recall that Cochran (1968) found that the single covariate subclassification adjustment using five subclasses and equal weights reduces 90% of the bias contributed by the covariate. Rosenbaum and Rubin (1984) provide an additional theorem to extend Cochran's results to the observational studies with multiple covariates. The above theorems provide the foundation of the propensity scores adjustment method using pair-matching or subclassification. The applications of the propensity scores method will be introduced in the next section.

2.3 Applications of propensity scores adjustment

2.3.1 Propensity scores matching approach

In observational studies, assume a small number of subjects in the treatment group and a large number of subjects in the control group. In these cases, obtaining the outcome of interest for all control subjects may be too expensive (e.g., locating large numbers of control subjects may be difficult and measuring their outcome may also be very costly). Rosenbaum and Rubin (1985) describe the matching method for selecting a subset of control subjects that are similar to the subjects in the treatment group with respect to observed covariates. The propensity scores matching method matches a subject in the treatment group with a subject in the control group who has a similar propensity score.

Propensity score matching methods have been widely applied in clinical and econometrics studies. D'Agostino Jr. (1998) examines the efficiency of removing covariates imbalance by propensity scores matching method for a clinical study. In a study using matching, Rubin and Thomas (2000) introduce an extension of propensity scores matching for a clinical study; the matching method uses both the estimated propensity scores and a collection of observed covariates that are closely related ("prognostic") to the outcome of interest. Similarly, Dehejia and Wahba (2002) illustrate propensity score matching methods on a labor program data to estimate the impact of the training.

2.3.2 Equal frequency subclassification and equal weights approach

The propensity scores matching approach selects a subset of the subjects in the control group to compare with the subjects in the treatment group. Thus, matching uses a reduced number of samples, which costs less than measuring all control subjects. However, in some studies, researchers may be interested in using the full sample. Under these circumstances, the propensity scores subclassification method is an option to consider. The propensity scores subclassification is also known as interval matching, blocking or stratification (Caliendo and Kopeinig 2008).

The propensity scores subclassification method is a nonparametric procedure, which does not depend on a specified regression function relating the outcome to covariates. Rubin (1997) states that this feature of the propensity scores subclassification is an advantage in estimating the treatment effect compared to multiple regression, because the regression estimate may not be reliable if its model specification is incorrect. One type of the propensity scores subclassification is Rosenbaum and Rubin's (1983, 1984) propensity scores equal frequency quintile subclassification method. This can be described as the following (D'Agostino Jr. 1998):

(1) Estimate the propensity scores by a logistic regression model of the treatment assignment indicator on the observed covariates;

(2) Sort the estimated propensity scores in an ascending order;

(3) Use quintiles of the estimated propensity scores as boundaries to form five subclasses.

If equal weights are used, then the overall treatment effect estimate is equal to the average of the mean differences of the outcome between the treatment and the control group among the five subclasses. Let c index the subclasses, and assume the subclass-specific means of the outcome are $\bar{y}_1^{(c)}$ for the treatment group and $\bar{y}_0^{(c)}$ for the control group, then the estimated treatment effect is $\sum_{c=1}^{5}(1/5)(\bar{y}_1^{(c)} - \bar{y}_0^{(c)})$.

The first theorem of the propensity scores method in Rosenbaum and Rubin (1983) states that, given the propensity scores, the observed covariates and the treatment assignment are conditionally independent. This implies that if a subclass of subjects is "homogeneous" in the propensity scores, then the subjects of the treatment group and the control group in this subclass will have the same distribution of covariates. Therefore, the degree of imbalance of the covariates between the treatment group and the control group can be evaluated by testing the mean difference of the propensity scores between the treatment group and the control group within each subclass. Dehejia and Wahba (1999, 2002) apply a two-sample t-test to assess whether the mean estimated propensity scores within each subclass are "identical" (or "balanced" in Imbens 2004) in order to create the minimal set of identical subclasses. This leads to further splitting the propensity score subclasses. The splitting procedure can be described in four steps:

(1) Form five equal frequency subclasses on the estimated propensity scores;

(2) Apply a two-sample t-test to check whether the means of the estimated propensity scores between the treatment group and the control group are identical within each subclass;

(3) If the t-test is statistically significant, this indicates the estimated propensity scores in that subclass are not balanced, then the EF subclasses will be further equally split;

(4) Repeat steps (2) and (3) until the t-tests are not statistically significant for all further split EF subclasses (additional details of this procedure are provided in Dehejia and Wahba 1999, 2002).

Propensity score equal frequency subclassification equal weights (EF-EW) methods have been broadly implemented in various studies. Little and Rubin (2000) discuss the applications of the propensity score EF-EW approach in clinical trials and epidemiology. In a study applying equal frequency subclassification, Aakvik (2001) uses 12 EF subclasses to evaluate a Norwegian labor training program.

2.3.3 Propensity scores equal frequency subclassification adjustment for survey nonresponse bias

We can apply the propensity score methodology to survey situations. Most surveys exhibit nonresponse, since not every selected subject responds to complete a questionnaire request. Instead of a treatment group and a control group that we discussed earlier for observational studies, we have a respondent group and a nonrespondent group.

David et al. (1983) provide the fundamental assumption for response propensity, which is that "the response (or nonresponse) mechanism is ignorable." This assumption is analogous to the strongly ignorable treatment assignment assumption of the propensity scores method in observational studies with a treatment group and a control group. Ignorable means that the response mechanism related to the outcome of interest can be explained by the observed covariates. For example, in a public opinion survey, the response mechanism has nothing to do with the opinion collected, although, it is associated with a known covariate, such as age.

The response propensity score is defined as the conditional probability of a subject responding to (or participating in) the survey, given the observed covariates. The subclassification can be applied on the estimated response propensity scores as in observational studies. When some demographics (e.g., age, income) are imbalanced between the respondent group and the nonrespondent group, the propensity score subclassification method has the advantage of reducing the nonresponse bias but not requiring subclassification on all covariates, as discussed in Section 2.1. It had been suggested that missing at random (MAR) is the mechanism of nonresponse often found in situations when missing data occurs (Bjørnstad 2007 and Chambers 2007). The strong ignorable treatment assignment assumption of the propensity score method is also an analog to the MAR assumption. This discussion provides some background of the response propensity scores subclassification adjustment estimator.

The response propensity scores method has been developed and utilized by several researchers. Little (1986) introduces a response propensity score method to adjust the estimated mean. Building on Little's research, the response propensity has been applied to a variety of studies with survey nonresponse. In these cases, propensity scores were used to adjust the means for nonresponse. Eltinge and Yansaneh (1997) use propensity score equal frequency subclasses to form nonresponse adjustment classes (cells). Another study using the response propensity scores, Smith et al. (2000) estimate the vaccination rates for the U.S. Carlson and Williams (2001) apply the response propensity score method to a household survey, while Diaz-Tena et al. (2002) use this method on a physician survey data. Vartivarian and Little (2003) introduce "joint classification" based on the response propensity scores and the predicted means from regressing the respondents' outcome on the covariates to improve efficiency, and to reduce the nonresponse bias. Furthermore, Harrod and Lesser (2007) propose a new response propensity score model to deal with the situation when a subsample of survey nonrespondents is collected.

2.4 Propensity scores adjustment under model misspecification; equal variance subclassification

Drake (1993) provides a simulation study to evaluate the impact of covariate omission and quadratic term misspecification on the propensity scores equal frequency subclassification adjustment under a linear regression model. The simulation results indicate that for a covariate omission in the propensity scores model, the estimated propensity scores equal frequency subclassification induces a similar bias as the ordinary

least square (OLS) estimator. For the quadratic term misspecification in the propensity scores model, the estimated propensity scores equal frequency subclassification estimator has less bias than the OLS estimator.

Often under propensity scores equal frequency subclassification, the lower subclass tends to contain a fewer number of treatment subjects, and the upper subclass tends to contain a fewer number of control subjects. As a consequence, the within subclass treatment effect estimates may have high variation. Even though some subclasses may have more variation, equal weights are generally assigned to subclasses in overall treatment effect estimation. Hullsiek and Louis (2002) introduce the propensity scores equal variance (EV) subclassification method to equalize the within subclass variances. Their approach applies an iteration procedure to subclassify the propensity scores based on the variances of the within subclass treatment effect estimator, which are approximately equivalent ("equal variance"). The variances of the within subclass treatment effect are estimated by a regression model of the outcome on the covariates within each subclass. However, after the equal variance subclasses have been formed, the within subclass treatment effect is estimated by a regression model of the outcome on the treatment assignment indicator only. Since the inverses of the equalized variances (IV) of the within subclass treatment effects are also used as the weights for the overall treatment effect estimate, the weights are theoretically equal among subclasses. They suggest that using the maximum number of propensity score subclasses, where data would allow, would enable the propensity scores subclassification method to remove

more bias. The trade-off would be a high variation of the within subclass treatment effect estimate.

Myers and Louis (2007) indicate that the propensity scores equal variance subclass boundaries could be relatively far from the equal frequency subclass boundaries. In order to equalize the variances among subclasses, the lower end subclasses may need to be wider. Although a wider subclass may achieve smaller variance, it may produce larger bias. Under equal frequency subclassification and under equal variance subclassification, Myers and Louis (2007) determine the number of subclasses that produce the smallest mean square error (MSE) of the treatment effect estimator. They conclude that under a simple linear model, the propensity scores equal frequency subclassification has an advantage over the equal variance approach. They also suggest increasing the number of subclasses to remove more bias until the overall treatment effect estimator becomes similar or until the overall variance of the estimator substantially increases.

2.5 Contribution of the thesis

2.5.1 Equal variance subclassification method under model misspecification

Hullsiek and Louis (2002) did not evaluate how the equal variance subclassification approach would perform under propensity scores model

misspecification as implemented in Drake (1993). No follow-up publications on this topic have yet been found in the literature. Using a simulation, we will assess the equal variance subclassification method under model misspecification. Additionally, Myers and Louis (2010) state that the subclassification estimator with inverse variance weighting generally underestimates the variance of the overall estimated treatment effect. We will investigate the variance and the bias of the treatment effect estimate using inverse variance weights for the equal frequency subclassification and the equal variance subclassification, under both correctly specified models and misspecified models.

2.5.2 Theoretical investigation on propensity scores subclassification adjustment estimator

We have an interest in evaluating the use of propensity score methodology on variance and bias. The propensity scores equal frequency subclassification approach may produce treatment effect estimates with a high variation in some subclasses. This method uses equal weights in the overall treatment effect estimation. Presently, under equal frequency (EF) subclassification, two weighting schemes, equal weights and inverse variance weights, can be applied; though, no evaluation has been provided on which weighting scheme would be more efficient to remove the bias of the covariates. For instance, near the lower end and the upper end quintile subclasses of the propensity scores, the sample size of the treatment group or the control group tends to be small. Thus, the variances of the near-end subclasses may be higher than those of other subclasses. We will show that under equal frequency subclassification, the inverse

variance (EF-IV) weighting estimator has smaller bias than the equal weights (EF-EW) estimator under certain condition. We will also show that under equal frequency (EF) subclassification, the EF-IV estimator always has a variance no larger than the EF-EW estimator. We will investigate the equal variance (EV) subclassification approach with an inverse variance weighting scheme.

2.5.3 New approach to propensity scores subclassification

Creating propensity scores for observational studies with treatment and control subjects has been challenging in certain conditions. Researchers exclude (1) the control subjects whose estimated propensity scores are lower than the minimum of the estimated propensity scores from the treatment group, and (2) the treatment subjects whose estimated propensity scores are higher than the maximum of the estimated propensity scores from the control group (Dehejia and Wahba 1999). For example, Stürmer et al. (2006) provide a case where the lowest and highest propensity scores are deleted when only one of the two groups (i.e., either treatment or control) is present. After the exclusion, only a subset of subjects are used for analysis, Figure 2.2 (Stürmer et al. 2006). This exclusion is also known as propensity scores trimming (Stürmer et al. 2007).

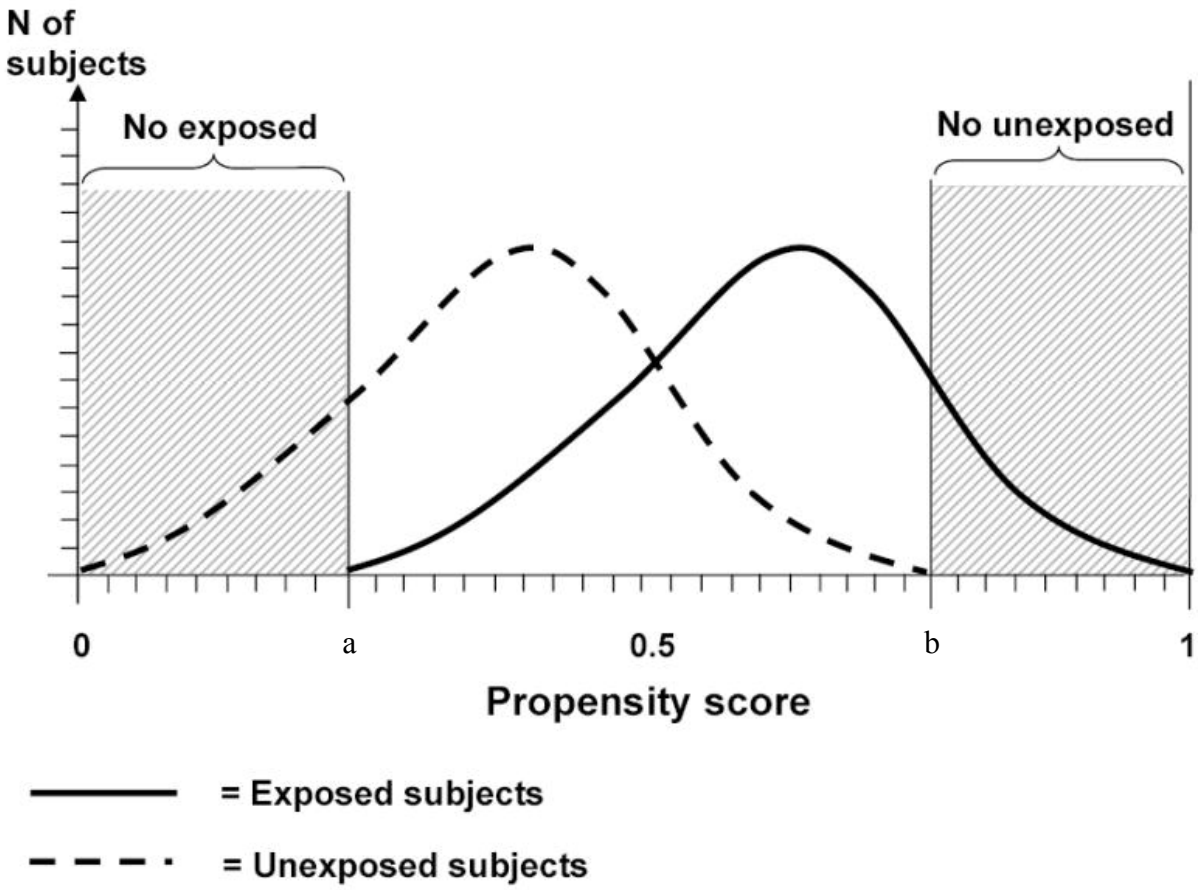

Figure 2. 2 Propensity score distributions of the treatment (exposed to the treatment or risk factor) subjects and control (unexposed) subjects

In surveys, we adopt propensity scores to adjust for nonresponse. The estimated propensity scores may occur at the lower end of the distribution; younger subjects refuse to participate more often than older subjects. This may result in some younger subjects being trimmed from the analysis. However, some researchers may be interested in analyzing the entire sample. Therefore, an alternative propensity scores subclassification method is desired to deal with this situation.

We propose a new approach of subclassification by attempting to balance the propensity scores for most subclasses. Our method will form a subclass that has homogeneous estimated propensity scores in the treatment group and the control group. Thus, the distribution of the covariates within each subclass is identical. This approach tends to create a large number of subclasses to reduce more bias as suggested by Cochran (1968), Imbens (2004) and Myers and Louis (2007).

2.6 Organization of the thesis

In Chapter Three, we will provide a simulation study to evaluate the propensity scores adjustment estimator of the treatment effect with respect to equal frequency and equal variance subclassification methods by using two weighting schemes: equal weights or inverse variance weights. We will also assess propensity score adjustment estimators under model misspecification in this simulation.

In Chapter Four, we will derive bias, variance and the root mean square error (RMSE) of the propensity scores subclassification adjustment estimator assuming equal weights and inverse variance weights. Our theoretical development will assume the outcome is generated by a linear regression model. We then develop lemmas, theorems and corresponding corollaries. We will extend the theoretical results to the multiple covariates situation.

In Chapter Five, we propose a novel propensity scores balancing subclassification, the PSB method. We implement a simulation to compare the results of the PSB estimators with other estimators including OLS, EF-EW, EF-IV and EV-IV. We also simulate a propensity score trimming situation. We will assume there is no treatment subject data near the lower end of the estimated propensity scores. We will examine our PSB method, and the EF and EV approaches under this situation.

In Chapter Six, we summarize the major findings from our simulation results, theory development and our proposed PSB subclassification approach. We also give proposals for future research work.

Chapter 3. SIMULATION STUDIES OF PROPENSITY SCORES METHOD

In this chapter, we examine the performance of several treatment effect estimators in terms of bias and root mean square error (RMSE) under different types of model misspecification. Specifically, we examine the regression or ordinary least square (OLS) estimator and three propensity score estimators (EF-EW, EF-IV and EV-IV). The differences among the propensity score estimators consist of two elements. One element is the forming of the propensity score subclasses, either by forming equal frequency (EF) subclasses or equal variance (EV) subclasses. The other element is the weighting schemes applied to combine the within subclass estimators to compute the overall estimate for the treatment effect, either by applying equal weights (EW) or inverse variance (IV) weights. We use the setting of an observational study with a treatment group and a control group for the simulations. Both the simulation model and the types of the model misspecification are adopted from Drake (1993).

For both the outcome variable and the treatment indicator variable, the independent covariates are generated as predictors for two scenarios. In the first scenario, the treatment indicator is simulated as an independent Bernoulli random variable using a logistic model involving two independent covariates. The outcome is simulated using a regression model having two independent covariates in addition to the treatment indicator. In the second scenario, the treatment indicator is simulated as an independent Bernoulli random variable using a logistic model involving a single

covariate and its squared term. The outcome is simulated using a regression model that includes the treatment indicator, a single covariate and its squared term.

For the first scenario, we will obtain the OLS estimator and propensity score estimators under correctly specified models and misspecified models. Under a correctly specified model, the OLS estimator is obtained by fitting the regression model with two independent covariates. For the propensity score estimators, the correct specification comes from also fitting the model of the propensity scores with two independent covariates. Under a misspecified model, the OLS estimator model misspecification involves fitting the regression model by excluding one of the two independent covariates. Likewise, for the propensity score estimators, the misspecification comes from fitting the model of the propensity scores without one of the two independent covariates.

Recall that the outcome was generated from a regression model including a quadratic term. In this second scenario, under the misspecified model the OLS estimator and propensity score estimators are obtained from fitting the model through the exclusion of the squared term. Table 3.1 provides a summary of simulation models and treatment effect estimators. In Hullsiek and Louis (2002), no assessment is provided regarding the performance of the EV subclassification approach under the types of model misspecification as suggested by Drake (1993). We have found no follow-up publication on this performance topic in the recent literature for EV-IV estimator under model misspecification.

Table 3. 1 Fitted models and treatment effect estimators

Model	Treatment effect estimators			
	Regression method	Propensity scores method		
	OLS (ordinary least square)	EF-EW	EF-IV	EV-IV
Correct Specification	Regression model has two predictors	Propensity scores model has two predictors		
Misspecification	Regression model has one predictor	Propensity scores model has one predictor		

In the first section below, we will introduce how data are simulated under the two scenarios. In the second section, we will describe the estimation of the regression model and the propensity scores model. The propensity score subclassification methods are described in the third section. In the fourth section, we introduce how the OLS estimator and the propensity scores estimators are obtained. The fifth section presents the results of the simulations. In Appendix Section A3.1, we provide the acronyms, notation tables and a flow chart diagram for each step in the simulation procedure.

3.1 Simulating data

In this section, we implement the simulation by generating the treatment indicator and outcome variables. The data are simulated in two scenarios. One scenario involves two independent covariates; the other scenario involves a single covariate and its squared term.

3.1.1 Scenario involving two independent covariates (x_1, x_2)

Define two independent covariates, $x_{11}, \ldots, x_{1n} \sim$ iid. N(0, 1), and independently, $x_{21}, \ldots, x_{2n} \sim$ iid. N(0, 1). In the first scenario, we simulate the treatment indicator values (z_i) as independent Bernoulli random variables under a logistic model:

$$z_i \mid x_{1i}, x_{2i} \sim \text{Bernoulli}(\pi_i)$$
$$\pi_i = \Pr(z_i = 1 \mid x_{1i}, x_{2i}) = \{1 + \exp[-(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i})]\}^{-1} \quad (3.1)$$

We generate the outcome variable using the model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \delta z_i + e_i \quad (3.2)$$

where $e_1, \ldots, e_n \sim$ iid. N(0, 1) with $\mathbf{e} \perp\!\!\!\perp (\mathbf{x}_1, \mathbf{x}_2, \mathbf{z})$, given $\mathbf{e} = (e_1, \ldots, e_n)$, $\mathbf{x}_1 = (x_{11}, \ldots, x_{1n})$, $\mathbf{x}_2 = (x_{21}, \ldots, x_{2n})$, $\mathbf{z} = (z_1, \ldots, z_n)$ and $\perp\!\!\!\perp$ denotes independence. For detailed information, see Appendix A3.1, Table A3.2.

3.1.2 Scenario involving a single covariate and its squared term (x, x^2)

Denote a single covariate, $x_1, \ldots, x_n \sim$ iid. N(0, 1). We simulate the treatment indicator values as independent Bernoulli random variables under a logistic model using x_i and its square:

$$z_i \mid x_i \sim \text{Bernoulli}(\pi_i)$$

$$\pi_i = Pr(z_i = 1 \mid x_i) = \{1 + \exp[-(\gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2)]\}^{-1} \quad (3.3)$$

We use a linear regression model with a single covariate, x_i, and its quadratic term to generate the outcome variable:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \delta z_i + e_i \quad (3.4)$$

The π_i denoted in equation (3.1) or equation (3.3) is the true propensity score for subject i. In observational studies, π_i is unknown. The true propensity scores are used to simulate the treatment indicator variable under both scenarios. We also assume either equation (3.2) or equation (3.4) to be the true outcome models.

We use the parameters provided by Drake (1993), where $\beta_0 = 1$, $\beta_1 = 1$, $\delta = (1, 3)$, $\beta_2 = (1, 2, 3)$; $\gamma_0 = 0$, $\gamma_1 = 0.4$, $\gamma_2 = (0.4, 0.7, 1.1)$. At each combination, 1000 random samples are simulated. Each sample consists of 1000 randomly generated observations.

3.2 Estimation of regression and propensity scores models

Under each of the two scenarios described in Section 3.1, we will fit the regression model to obtain the OLS estimator of the treatment effect. We will also fit the logistic regression model to obtain the propensity score estimators of the treatment effect. For both the regression model and the propensity score models, the correct specification of the model implies no omission for one of the two independent covariates under the

first scenario. The misspecified model includes a covariate omission under the first scenario or a quadratic term omission under the second scenario. Table 3.2 compares the models used to simulate and fit both the outcome models and propensity score models (subscript i is ignored for simplicity).

Table 3. 2 Comparison between simulating and fitting the outcome and propensity score models

<table>
<tr><td>Covariates</td><td colspan="2">OLS estimator</td></tr>
<tr><td rowspan="3">(x_1, x_2)</td><td>True propensity scores:
$\pi = \{1 + \exp[-(\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2)]\}^{-1}$</td><td>Outcome model:
(a) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \delta z + e$
(b) $y = \beta_0 + \beta_1 x_1 + \delta z + e$</td></tr>
<tr><td colspan="2">Propensity scores estimators (EF-EW, EF-IV, EV-IV)</td></tr>
<tr><td>Propensity scores model:
(a) $e(\mathbf{x}) = \{1 + \exp[-(\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2)]\}^{-1}$
(b) $e(\mathbf{x}) = \{1 + \exp[-(\gamma_0 + \gamma_1 x_1)]\}^{-1}$</td><td>True outcome:
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \delta z + e$</td></tr>
<tr><td rowspan="3">(x, x^2)</td><td colspan="2">OLS estimator</td></tr>
<tr><td>True propensity scores:
$\pi = \{1 + \exp[-(\gamma_0 + \gamma_1 x + \gamma_2 x^2)]\}^{-1}$</td><td>Outcome model:
(b) $y = \beta_0 + \beta_1 x + \delta z + e$</td></tr>
<tr><td colspan="2">Propensity scores estimators (EF-EW, EF-IV, EV-IV)</td></tr>
<tr><td></td><td>Propensity scores model:
(b) $e(\mathbf{x}) = \{1 + \exp[-(\gamma_0 + \gamma_1 x)]\}^{-1}$</td><td>True outcome:
$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \delta z + e$</td></tr>
</table>

Note: (a) - correctly specified regression or propensity score models,

(b) - misspecified regression or propensity score models.

3.3 Propensity score subclassification

The propensity scores estimation approach produces the estimated propensity scores by logistic regression (either with a correctly specified propensity scores model or with the misspecified propensity scores model described in Table 3.2). The estimated propensity scores are then sorted in an ascending order before applying subclassification. The number of propensity score subclasses is set at five. There are two methods used when forming propensity scores subclasses: equal frequency (EF) and equal variance (EV). The EF and EV subclassification procedures are described next.

The EF subclassification uses quintiles of the estimated propensity scores as boundaries to form five adjacent subclasses (details are in Chapter Two). We implement an iterative procedure to apply the EV subclassification provided by Hullsiek and Louis (2002). The EV subclassification starts with five EF subclasses on the estimated propensity scores. Within each EF subclass, this procedure applies a regression model of the outcome variable on the covariates to estimate the variance of the within subclass treatment effect (regression) estimator. Then, the subclass boundaries are adjusted to approximately equalize the estimated variances of the within subclass treatment effect estimators among subclasses. For extra information, see Chapter Two and Appendix-Figure A3.2.

In some cases, the EV subclassification procedure does not produce a result. This may be attributed to outliers creating wide subclasses resulting in too few observations remaining to form five subclasses. Thus, less than five EV subclasses are created in this situation. Figure A3.4 in the Appendix illustrates a situation where EV subclassification only produces three subclasses.

3.4 Treatment effect estimates

After forming the propensity score subclasses, the subclass-specific treatment effect estimator and its estimated variance are obtained within each subclass in order to compute the weighted (overall) propensity score treatment effect estimators. For each of the two scenarios described in Section 3.1, we will obtain the OLS estimator and propensity score estimators for models under correct specification or misspecifications.

3.4.1 OLS estimator and propensity score estimators

In the case of the OLS estimator for the treatment effect δ, $\hat{\delta}$ is obtained from the regression models described in Table 3.2. For the propensity score estimators, we first obtain the estimated treatment effect and its estimated variance within each subclass. The subclass-specific treatment effect within subclass c is estimated by,

$$\hat{\delta}_c = \bar{y}_1^{(c)} - \bar{y}_0^{(c)} \tag{3.5}$$

where $\bar{y}_1^{(c)} = \sum_{i=1}^{n_1^{(c)}} y_{1i}^{(c)} / n_1^{(c)}$, $\bar{y}_0^{(c)} = \sum_{i=1}^{n_0^{(c)}} y_{0i}^{(c)} / n_0^{(c)}$. For detailed information, see Appendix Table A3.2.

The estimated treatment effect within subclass c in equation (3.5) can also be obtained by fitting the regression model for subclass c data only:

$$y_i^{(c)} = \beta_0^{(c)} + \delta_c z_i^{(c)} + e_i^{(c)} \tag{3.6}$$

Hullsiek and Louis (2002) use the same regression model to obtain $\hat{\delta}_c$ and its estimated variance, $\hat{V}_c$. Using EV subclassification, the estimated variances, $\hat{V}_c$, which are obtained by fitting equation (3.6), may not necessarily be equal among subclasses.

The estimated variance for $\hat{\delta}_c$ is,

$$\hat{V}_c = \hat{Var}(\hat{\delta}_c) = \hat{Var}(\bar{y}_1^{(c)}) + \hat{Var}(\bar{y}_0^{(c)}) \tag{3.7}$$

The propensity scores estimator of the overall treatment effect is a weighted estimate,

$$\hat{\delta} = \sum_{c=1}^{5} w_c (\bar{y}_1^{(c)} - \bar{y}_0^{(c)}) \tag{3.8}$$

If equal weights are assumed, $w_c^{EW} = 1/5$. If inverse variance weights are assumed, then

$$w_c^{IV} = \frac{1/\hat{V}_c}{\sum_{c=1}^{5} 1/\hat{V}_c} \tag{3.9}$$

Additional details are provided in the Appendix table A3.2.

3.4.2 Measuring the performance of treatment effect estimates

We generated simulations to assess the performance of the treatment effect estimators. We created 1000 simulated samples, each with 1000 random observations. Among those 1000 simulated samples, we obtained $\bar{\delta}$ = mean($\underset{\sim}{\hat{\delta}}$), where $\underset{\sim}{\hat{\delta}} = (\hat{\delta}^{(1)}, \hat{\delta}^{(2)}, ..., \hat{\delta}^{(1000)})$. We computed the percentage relative bias of the mean (PRMB). We repeated this for a range of parameter values in the tables shown in Section 3.5. In addition, we computed the percentage relative bias of the median (RPBM) of $\hat{\delta}$ as shown by Drake (1993), the percentage relative standard deviation (PRSD) and percentage relative RMSE (PRRMSE). The percentage relative measure formulas are summarized in the Appendix Table A3.3, and the true value of the parameters are in Appendix Table A3.4.

3.5 Simulation results

The performance of the treatment effect estimators are evaluated in three subsections. In the first subsection, we discuss obtaining the OLS estimator and propensity score estimators under the correctly specified models. In the second subsection, we obtain estimators using misspecified models, omitting an independent covariate for estimating propensity scores and fitting the regression model. In the third subsection, we use the quadratic term misspecification for estimating propensity scores and fitting the regression model. In each subsection, the PRMB are obtained for the treatment effect estimators.

3.5.1 Correctly specified model involving covariates (x_1, x_2)

Table 3.5.1. 1 PRMB for the OLS and three propensity score estimators using correctly specified propensity score models and correctly specified regression models for covariates (x_1, x_2)

Parameters			PRMB of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	0.1461	7.9362	6.6079	7.2425
		0.7	-0.0872	9.9998	8.1127	9.3759
		1.1	-0.2448	12.3790	9.4942	12.1327
	2	0.4	0.2306	11.7691	9.4455	10.1228
		0.7	0.3096	16.7542	12.2626	13.7877
		1.1	0.2194	21.7132	15.2894	18.2459
	3	0.4	-0.0923	15.1893	12.0489	12.9778
		0.7	-0.0476	23.0067	15.7371	17.3590
		1.1	0.1624	31.1385	19.5637	22.3264
3	1	0.4	0.0887	2.6308	2.2026	2.4083
		0.7	0.0633	3.4089	2.7778	3.2000
		1.1	0.0989	4.3244	3.3734	4.2355
	2	0.4	0.0028	3.8369	3.0805	3.3032
		0.7	0.0386	5.5163	4.0365	4.4723
		1.1	0.1338	7.3883	5.2574	6.2693
	3	0.4	0.0682	5.2261	4.2395	4.6025
		0.7	0.0190	7.6604	5.2674	5.7456
		1.1	0.0108	10.3996	6.5743	7.4332

The results in Table 3.5.1.1 indicate that when using correctly specified propensity scores, the PRMB of all propensity score estimators (EF-EW, EF-IV and EV-

IV) increase as γ_2 (i.e. the influence of x_2 on the propensity scores) increases. This pattern is consistent for each level of β_2 (i.e. the influence of x_2 on the outcome). All propensity score estimators produce positive biases. The PRMB of all propensity score estimators at $\delta=3$ are smaller than they are at $\delta=1$. This is because the PRMB is computed by $100\times[(\bar{\delta} - \delta)/\delta]\%$, so at $\delta=3$, the denominator increases by a factor of three as compared to when $\delta=1$. Among propensity score estimators, the EF-IV has a lower PRMB than the other two propensity score estimators, while EV-IV has a lower PRMB than EF-EW. Given a predetermined p-value level, for each combination of δ, β_2 and γ_2 values, a paired t-test can be performed to evaluate whether there is a significant difference of mean biases between EF-IV and EV-IV. A similar test can also be applied to compare mean biases between EV-IV and EF-EW. As we expected, OLS has the lowest PRMB (around zero) among all estimators, since the estimation model is correctly specified.

3.5.2 Misspecified model with a covariate omission (excluding x_2)

Table 3.5.2. 1 PRMB for the OLS and three propensity score estimators using misspecified propensity score models and misspecified regression models

Parameters			PRMB of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	38.8223	42.5489	42.2839	42.5400
		0.7	63.2160	66.7444	66.5470	66.6874
		1.1	88.3444	91.2998	91.0594	91.1867
	2	0.4	77.3982	81.0031	80.8393	81.0293
		0.7	126.6759	130.1814	130.0417	130.1572
		1.1	176.4260	179.2282	179.1568	179.2922
	3	0.4	115.9654	119.3896	119.3800	119.5767
		0.7	189.7551	192.7523	192.7413	192.8746
		1.1	265.6911	268.3094	268.2216	268.4400
3	1	0.4	12.9411	14.1880	14.1061	14.1795
		0.7	21.1205	22.2899	22.2132	22.2739
		1.1	29.6657	30.6738	30.6110	30.6456
	2	0.4	25.8473	27.0040	26.9865	27.0471
		0.7	42.1397	43.2108	43.1476	43.2044
		1.1	59.0174	59.9629	59.9147	59.9557
	3	0.4	38.7916	39.9949	39.9302	39.9756
		0.7	63.0146	64.0501	64.0070	64.0514
		1.1	88.5442	89.4056	89.3705	89.3974

The results in Table 3.5.2.1 indicate that for a misspecified propensity scores with a covariate omission, the PRMB of all propensity score estimators increase substantially as γ_2 increases. This pattern is the same for all propensity score estimators at each level

of β_2. The misspecified OLS estimator, which regresses only on covariate x_1, follows the same pattern. As γ_2 and β_2 increase, the PRMB of all estimators increase substantially.

The PRMB of all estimators at $\delta=3$ are smaller than they are at $\delta=1$. This is because the PRMB is computed by $100\times[(\bar{\delta} - \delta)/\delta]\%$, so at $\delta=3$, the denominator increases by a factor of three as compared to when $\delta=1$. Overall, none of the estimators perform well in terms of PRMB when an independent covariate is omitted from the propensity scores model and from the regression model. This implies that under this condition, researchers should not use these estimators in the analysis. This is not surprising, because the omitted covariate is the one used in both a regression model to generate the outcome and a logistic model to generate the propensity scores. These simulation results suggest that researchers should estimate the propensity scores based on all available covariate information to increase accuracy of the analysis. Also, the propensity score model specification should include all observed covariates.

3.5.3 Model with a quadratic term misspecification (omitting x^2)

Table 3.5.3. 1 PRMB for the OLS and three propensity score estimators using misspecified propensity score models and misspecified regression models

Parameters			PRMB of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	55.0676	30.9356	13.6264	16.5840
		0.7	76.1795	39.0702	17.9736	24.7352
		1.1	89.7706	44.4252	19.5753	31.4377
	2	0.4	110.9671	59.5532	12.6512	18.1618
		0.7	151.6400	76.6421	18.7634	29.5091
		1.1	179.9240	88.7122	23.9111	42.8565
	3	0.4	166.6882	89.2934	11.8818	21.0997
		0.7	229.9696	115.8887	18.1795	34.9952
		1.1	269.3079	131.5643	23.8583	51.4120
3	1	0.4	18.5396	10.4479	4.6598	5.6252
		0.7	25.4047	12.9898	6.1071	8.2665
		1.1	29.8641	14.8887	6.6913	10.6205
	2	0.4	36.7871	19.7313	4.2933	6.2256
		0.7	51.0099	25.8933	6.3438	10.0236
		1.1	59.8128	29.5061	8.0629	14.2787
	3	0.4	55.6600	29.6158	3.8182	7.0717
		0.7	76.7866	38.7396	5.8827	11.7854
		1.1	89.6845	43.9825	7.8445	17.0990

The results in Table 3.5.3.1 indicate that under quadratic term misspecification, the PRMB of all propensity score estimators increase as γ_2 increases. This pattern is the same for all propensity score estimators at each level of β_2. The misspecified OLS

estimator, which omits a quadratic term, follows the same pattern. However, the EF-IV has the lowest PRMB of all estimators in this simulation, while the OLS estimator has the highest PRMB. These results show that as β_2 increases, the PRMB of the OLS, EF-EW and EV-IV estimators increase. However, the results obtained from the EF-IV estimator do not follow the same pattern. As in Section 3.5.1, given a predetermined p-value level, for each combination of δ, β_2 and γ_2 values, paired t-tests can be applied to compare differences in mean biases between EV-IV and EF-EW, and between EV-IV and EF-EW. The PRMB of all estimators at δ=3 are smaller than they are at δ=1. This is not surprising, since the PRMB is computed by $100\times[(\bar{\delta} - \delta)/\delta]\%$, so at δ=3, the denominator increases by a factor of three as compared to when δ=1.

3.5.4 Summary of simulation

In the investigation of different treatment effect estimators, we found that under correctly specified propensity score models, the EF-IV estimator has the lowest PRMB and PRRMSE compared to other propensity score estimators across the range of parameters investigated. We also discovered that under the quadratic term misspecification, the EF-IV estimator has the lowest PRMB, PRSD and PRRMSE among all estimators including the OLS estimator (see Appendix A3.5.1, A3.5.2 and A3.5.3). The simulation results are summarized below.

- For propensity score estimators: the PRMB and the PRRMSE generally increase as the influence of the second covariate on the propensity scores (γ_2) increases.

As the influence of the second covariate on the outcome variable (β_2) increases, the PRMB and the PRRMSE often increases.

- Under a correctly specified propensity score models and regression models using two independent covariates:
  - ❖ Among propensity score estimators, EF-IV has the lowest PRMB, PRSD and PRRMSE.
  - ❖ OLS has lower PRMB, PRSD and PRRMSE than the propensity score estimators. This is expected since the estimating model is correctly specified.
- Under the misspecified models (e.g., omitting an independent covariate): the performance of the OLS is very similar to the performance of propensity score estimators. All estimators perform poorly in terms of PRMB, PRSD and PRRMSE.
- Under the misspecified models (e.g., excluding a quadratic term): EF-IV provides the lowest PRMB, PRSD and PRRMSE. This suggests that it has the best performance among all estimators in this simulation. EV-IV has a lower PRMB, PRSD and PRRMSE than EF-EW. OLS has the highest PRMB, PRSD and PRRMSE when compared to the propensity scores estimators.

In the next chapter, we develop a theoretical derivation to specify under what conditions a lower weighted (overall) bias or variance of the estimator would be obtained using either EW or IV weights.

Chapter 4. THEORETICAL INVESTIGATION OF PROPENSITY SCORE ESTIMATORS

In this chapter, we develop a theoretical framework to compare different propensity score subclassification adjustment estimators (EF-EW, EF-IV and EV-IV). In this framework, we find the following:

1) Under the EF subclassification adjustment, if higher variation occurs with larger bias for within subclass treatment effect estimates, then the overall bias of the IV weighting estimator is smaller than that of the EW estimator.
2) The EF-IV estimator always has no larger variance than the EF-EW estimator.
3) If the variance of the treatment effect estimator within subclass in the EV subclassification is larger than the harmonic mean of the variances of the EF within subclass treatment effect estimators, then the EF-IV estimator has a lower variance than the EV-IV estimator.

The setting of our theoretical framework is an observational study with a "treatment" indicator variable and a covariate. We assume the outcome variable is generated from a linear regression model that includes this treatment indicator and the covariate. The first section provides an expression for B_c, the bias due to the covariate within subclass; an expression of the variance of the subclass-specific treatment effect estimate, V_c; and a lemma giving a condition when B_c is nonnegative. The second section develops theorems to compare the bias and variance of the two estimators under EF subclassification. The third section develops a theorem for comparing variances between the EF-IV and EV-IV estimators. In the Appendix, Section A4.1 provides

acronyms, notation tables and a flow chart diagram for each step in the theoretical derivation. We extend this theory work to the situation with multiple covariates in Appendix A4.4.

4.1 Expression of B_c, V_c and Lemma under a linear regression model

We assume the outcome variable, y_i, is generated under a linear regression model that includes a treatment indicator and a single covariate:

$$y_i = \beta_0 + \beta_1 x_i + \delta z_i + e_i \tag{4.1}$$

where $e_1, \ldots, e_n \sim$ iid. $N(0, \sigma^2_e)$ with $(\mathbf{x}, \mathbf{z}) \perp\!\!\!\perp \mathbf{e}$. For detailed information, see Appendix A4.1, Table A4.1.

To estimate variances of the propensity score estimators, we assume that there are at least two observations in both the treatment group and the control group within each subclass. By definition (Section 3.4), the propensity score estimator is given by

$$\hat{\delta}(\underset{\sim}{w}) = \sum_{c=1}^{C} w_c \hat{\delta}_c = \sum_{c=1}^{C} w_c (\bar{y}_1^{(c)} - \bar{y}_0^{(c)}) \tag{4.2}$$

where $w_c \geq 0$, $\sum_{c=1}^{C} w_c = 1$, with c = 1, 2, …, C indexing the subclasses.

Here, let s_i^c denote the indicator of whether subject i is in subclass c, such that $s_i^c = 1$ if and only if $x_i \in (a_c, b_c)$, otherwise $s_i^c = 0$, where a_c, b_c are lower bound and upper bound of subclass c, respectively; then

$$\bar{y}_1^{(c)} = \frac{\sum_{i=1}^{n} z_i s_i^c y_i}{\sum_{i=1}^{n} z_i s_i^c} = \frac{\sum_{i=1}^{n} z_i s_i^c (\beta_0 + \beta_1 x_i + \delta z_i + e_i)}{\sum_{i=1}^{n} z_i s_i^c} = \beta_0 + \beta_1 \frac{\sum_{i=1}^{n} z_i s_i^c x_i}{\sum_{i=1}^{n} z_i s_i^c} + \delta \frac{\sum_{i=1}^{n} z_i^2 s_i^c}{\sum_{i=1}^{n} z_i s_i^c} + \frac{\sum_{i=1}^{n} z_i s_i^c e_i}{\sum_{i=1}^{n} z_i s_i^c}$$

. Since, $z_i = 1, 0$ implies that $z_i^2 = z_i$, we see that $\sum_{i=1}^{n} z_i^2 s_i^c = \sum_{i=1}^{n} z_i s_i^c$. Therefore, $\bar{y}_1^{(c)}$ reduces to $\beta_0 + \beta_1 \bar{x}_1^{(c)} + \delta + \bar{e}_1^{(c)}$, where $\bar{x}_1^{(c)} = \frac{\sum_{i=1}^{n} z_i s_i^c x_i}{\sum_{i=1}^{n} z_i s_i^c}$, $\bar{e}_1^{(c)} = \frac{\sum_{i=1}^{n} z_i s_i^c e_i}{\sum_{i=1}^{n} z_i s_i^c}$. Similarly, $\bar{y}_0^{(c)} = \beta_0 + \beta_1 \bar{x}_0^{(c)} + \bar{e}_0^{(c)}$. Thus, we have

$$\bar{y}_1^{(c)} - \bar{y}_0^{(c)} = \beta_1 (\bar{x}_1^{(c)} - \bar{x}_0^{(c)}) + \delta + \bar{e}_1^{(c)} - \bar{e}_0^{(c)} \tag{4.3}$$

Taking the expectation in equation (4.2) gives

$$E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}) = \beta_1 [E(\bar{x}_1^{(c)}) - E(\bar{x}_0^{(c)})] + \delta + E(\bar{e}_1^{(c)}) - E(\bar{e}_0^{(c)}) \tag{4.4}$$

Here, we should notice that none of $\bar{x}_1^{(c)}$, $\bar{y}_1^{(c)}$, $\bar{e}_1^{(c)}$ are defined, except when $z_i = 1$ (from the treatment group) and $s_i^c = 1$ (within subclass c). Therefore, we write $E(\bar{x}_1^{(c)})$ as $E(\bar{x}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}^c)$, or simply, $E(x \mid z = 1, s^c = 1)$. Similarly, we write $E(\bar{x}_0^{(c)})$

as $E(x \mid z=0, s^c=1)$, write $E(\bar{e}_1^{(c)})$ as $E(e \mid z=1, s^c=1)$ and $E(\bar{e}_0^{(c)})$ as $E(e \mid z=0, s^c=1)$. By definition, s_i^c is a function of x_i, so that $(\mathbf{x}, \mathbf{z}) \perp\!\!\!\perp \mathbf{e}$ implies $(\mathbf{z}, \mathbf{s}^c) \perp\!\!\!\perp \mathbf{e}$. Therefore both of $E(e \mid z=1, s^c=1)$ and $E(e \mid z=0, s^c=1)$ are zero. Hence, we have

$$E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}) = \beta_1[E(x \mid z=1, s^c=1) - E(x \mid z=0, s^c=1)] + \delta \tag{4.5}$$

The bias for subclass c is then $B_c = \beta_1[E(x \mid z=1, s^c=1) - E(x \mid z=0, s^c=1)]$, wherein $E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}) = B_c + \delta$. Hence, the expectation of the propensity scores estimator becomes

$$E[\hat{\delta}(\underset{\sim}{w})] = \sum_{c=1}^{C} w_c B_c + \delta, \tag{4.6}$$

and then overall bias of the propensity score estimator is

$$Bias[\hat{\delta}(\underset{\sim}{w})] = \sum_{c=1}^{C} w_c B_c \tag{4.7}$$

We denote the within subclass variance of $\bar{y}_1^{(c)} - \bar{y}_0^{(c)}$ by $V_c = Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)})$, and the overall variance of the propensity score estimator by

$$Var[\hat{\delta}(\underset{\sim}{w})] = \sum_{c=1}^{C} w_c^2 V_c \tag{4.8}$$

(for details, see Appendix A4.1.2, A4.1.3)

Next, we give a Lemma that provides a condition under which B_c is nonnegative for all subclasses.

Lemma 4.1 Suppose P{z = 1 | x} = e(x) is a non-decreasing function of x $\in \mathcal{R}$.

Then for x $\in$ (a, b), E(x | z = 1) $\geq$ E(x | z = 0) for any a < b in $\mathcal{R}$.

Proof:

First, we provide two identities:

$$f(x \mid z=1) = \frac{f(x, z=1)}{P(z=1)} = \frac{P(z=1 \mid x) f(x)}{P(z=1)} = \frac{e(x) f(x)}{P(z=1)},$$

and

$$f(x \mid z=0) = \frac{f(x, z=0)}{P(z=0)} = \frac{P(z=0 \mid x) f(x)}{P(z=0)} = \frac{[1 - P(z=1 \mid x)] f(x)}{P(z=0)} = \frac{[1 - e(x)] f(x)}{P(z=0)}.$$

Let $f_{(a, b)}(x \mid z)$ denote the conditional pdf of x given z when x is restricted to the interval (a, b). Then $f_{(a,b)}(x \mid z) = \frac{f(x \mid z)}{\int_a^b f(x \mid z) dx}$. Similarly, let $f_{(a, b)}(x)$ denote the pdf of x when x is restricted to the interval (a, b).

Using these, we can write the conditional expectations:

$$
\begin{aligned}
E_{(a,b)}(x \mid z=1) &= \frac{\int_a^b x f(x \mid z=1)dx}{\int_a^b f(x \mid z=1)dx} \\
&= \frac{\int_a^b x \dfrac{e(x)f(x)}{P(z=1)}dx}{\int_a^b \dfrac{e(x)f(x)}{P(z=1)}dx} \\
&= \frac{\int_a^b xe(x)f(x)dx}{\int_a^b e(x)f(x)dx} \\
&= \frac{\int_a^b xe(x)f(x)dx}{\int_a^b e(x)f(x)dx} \\
&= \frac{E_{(a,b)}[xe(x)]}{E_{(a,b)}[e(x)]}
\end{aligned}
$$

and

$$
\begin{aligned}
E_{(a,b)}(x \mid z=0) &= \frac{\int_a^b x f(x \mid z=0)dx}{\int_a^b f(x \mid z=0)dx} \\
&= \frac{\int_a^b x \dfrac{[1-e(x)]f(x)}{P(z=0)}dx}{\int_a^b \dfrac{[1-e(x)]f(x)}{P(z=0)}dx} \\
&= \frac{\int_a^b x[1-e(x)]f(x)dx}{\int_a^b [1-e(x)]f(x)dx} \\
&= \frac{E_{(a,b)}[x] - E_{(a,b)}[xe(x)]}{1-E_{(a,b)}[e(x)]}.
\end{aligned}
$$

Next, since e(x) is a non-decreasing function of x, it is certainly a non-decreasing function in (a, b). By the Covariance Inequality Theorem in Casella and Berger (2001), we have $Cov_{(a,b)}[x, e(x)] \geq 0$. Hence,

$$\begin{aligned} Cov_{(a,b)}[x, e(x)] \geq 0 &\Leftrightarrow E_{(a,b)}[xe(x)] \geq E_{(a,b)}[e(x)]E_{(a,b)}[x] \\ &\Leftrightarrow E_{(a,b)}[xe(x)] - E_{(a,b)}[xe(x)]E_{(a,b)}[e(x)] \geq \\ &\quad E_{(a,b)}[e(x)]E_{(a,b)}[x] - E_{(a,b)}[e(x)]E_{(a,b)}[xe(x)] \\ &\Leftrightarrow E_{(a,b)}[xe(x)]\{1 - E_{(a,b)}[e(x)]\} \geq \\ &\quad E_{(a,b)}[e(x)]\{E_{(a,b)}[x] - E_{(a,b)}[xe(x)]\} \\ &\Leftrightarrow \frac{E_{(a,b)}[xe(x)]}{E_{(a,b)}[e(x)]} \geq \frac{E_{(a,b)}[x] - E_{(a,b)}[xe(x)]}{1 - E_{(a,b)}[e(x)]}. \end{aligned}$$

Therefore, $E_{(a,b)}(x \mid z = 1) \geq E_{(a,b)}(x \mid z = 0)$. This completes the proof.

4.2 Theorems comparing different weighting schemes

We now compare the biases of two propensity score estimators under the same subclassification but using different weights. We also examine the variances between the equal weights (EW) estimator and the inverse variance (IV) weight estimator.

4.2.1 Discordance and concordance

Consider the vector of subclass biases from a particular subclassification scheme. Two weighting schemes will provide two different vectors of weights. Assessing the bias of two propensity score estimators using the same subclassification but different weights

involves all three vectors. We first introduce the notations of discordance and concordance between two vectors.

Consider two arbitrary vectors with the same length: $\mathbf{a} = (a_1, \ldots, a_i, \ldots, a_j, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_i, \ldots, b_j, \ldots, b_n)$. By definition, **a** and **b** are discordant if $(a_i - a_j)(b_i - b_j) \leq 0$ for all i, j. And **a** and **b** are concordant if $(a_i - a_j)(b_i - b_j) \geq 0$ for all i, j. In other words, discordance indicates that as a_i increases, b_i is non-increasing, and concordance indicates that as a_i increases, b_i is non-decreasing. We provide the following Lemma of discordance inequality and concordance inequality.

Lemma 4.2 If **a** and **b** are discordant, then $\sum_{i=1}^{n} a_i \sum_{i=1}^{n} b_i \geq n\sum_{i=1}^{n} a_i b_i$; if **a** and **b** are concordant, then $\sum_{i=1}^{n} a_i \sum_{i=1}^{n} b_i \leq n\sum_{i=1}^{n} a_i b_i$.

Proof:

For concordance:

$$\begin{aligned}
\sum_{i=1}^{n} a_i \sum_{i=1}^{n} b_i - n\sum_{i=1}^{n} a_i b_i &= \sum_{i=1}^{n} a_i \sum_{j=1}^{n} b_j - n\sum_{j=1}^{n} a_j b_j \\
&= \sum_{i=1}^{n} a_i \sum_{j=1}^{n} b_j - (\sum_{i=1}^{n})\sum_{j=1}^{n} a_j b_j \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j - \sum_{i=1}^{n}\sum_{j=1}^{n} a_j b_j \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} (a_i - a_j) b_j \\
&= \sum\sum_{i<j} (a_i - a_j) b_j + \sum\sum_{i>j} (a_i - a_j) b_j \\
&= \sum\sum_{i<j} (a_i - a_j) b_j + \sum\sum_{i<j} (a_j - a_i) b_i \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{n} (a_i - a_j)(b_i - b_j) \geq 0 \text{ since } (a_i - a_j)(b_i - b_j) \leq 0 \text{ for all } i, j
\end{aligned}$$

For discordance, the proof is similar. This completes the proof.

Lemma 4.2 is based on the well-known Chebyshev's Sum Inequality (Abramowize and Stegun, 1970; Mitrinović, 1970) and is also introduced as the synchronous inequality in Toader (1996). We now introduce another vector, $\mathbf{t} = (t_1, \ldots, t_i, \ldots, t_j, \ldots, t_n)$ with $t_i \geq 0$ for all i. Now, we develop an extension of Lemma 4.2.

Lemma 4.3 If **a** and **b** are discordant and **t** = (t_1, …, t_i, …, t_j, …, t_n) with $t_i \geq 0$ for all i, then

$$\sum_{i=1}^{n} t_i a_i \sum_{i=1}^{n} t_i b_i \geq \sum_{i=1}^{n} t_i \sum_{i=1}^{n} t_i a_i b_i \ .$$

Proof:

$$\begin{aligned}
\sum_{i=1}^{n} t_i a_i \sum_{i=1}^{n} t_i b_i - \sum_{i=1}^{n} t_i \sum_{i=1}^{n} t_i a_i b_i &= \sum_{i=1}^{n}\sum_{j=1}^{n} t_i a_i t_j b_j - \sum_{i=1}^{n} t_i \sum_{j=1}^{n} t_j a_j b_j \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} t_i t_j (a_i - a_j) b_j \\
&= \sum\sum_{i<j} t_i t_j (a_i - a_j) b_j + \sum\sum_{i>j} t_i t_j (a_i - a_j) b_j \\
&= \sum\sum_{i<j} t_i t_j (a_i - a_j) b_j + \sum\sum_{i<j} t_i t_j (a_j - a_i) b_i \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{n} t_i t_j (a_i - a_j)(b_i - b_j) \geq 0
\end{aligned}$$

This completes the proof.

We use Lemma 4.3 to prove the theorem in the next subsection.

4.2.2 Comparing biases between two propensity score estimators

For a propensity score estimator using one weighting scheme, we use u_c to denote the original (unstandardized) weight for subclass c; and use **u** to denote the vector of unstandardized weights, where **u** = (u_1, …, u_C). We assume $u_c > 0$ and that its corresponding standardized weight is $w_c = \dfrac{u_c}{\sum_{c=1}^{C} u_c}$, whereby $\sum_{c=1}^{C} w_c = 1$. If equal weights

are used, then $w_c^{EW} = 1/C$; if inverse variance weights are used, then $w_c^{IV} = \frac{1}{V_c} / \sum_{c=1}^{C} \frac{1}{V_c}$.

We use **w** to denote the vector of standardized weights, where $\mathbf{w} = (w_1, \ldots, w_C)$. The bias of the propensity scores estimator using standardized weights (**w**) is then denoted by

$$Bias[\hat{\delta}(\underset{\sim}{w})] = \sum_{c=1}^{C} w_c B_c .$$

Consider another propensity score estimator using a different weighting scheme. We use u_c^* to denote its unstandardized weight for subclass c and use w_c^* to denote the corresponding standardized weight for subclass c.

Let $\boldsymbol{B}$ denote the vector of biases for all subclasses: $\boldsymbol{B} = (B_1, \ldots, B_C)$. We are interested in the biases of $\hat{\delta}(\underset{\sim}{w})$ and $\hat{\delta}(\underset{\sim}{w}*)$, when $\mathbf{u}^*/\mathbf{u}$ and $\boldsymbol{B}$ are discordant. While seeming constrained, this allows us to relate or compare these two different types of estimators under the same subclassification scheme, as we will indicate in two corollaries following the main theorem.

Theorem 4.4 Assume $B_c \geq 0$ for all subclasses. If $\mathbf{u}^*/\mathbf{u}$ and $\boldsymbol{B}$ are discordant, then $| Bias[\hat{\delta}(\underset{\sim}{w}*)] | \leq | Bias[\hat{\delta}(\underset{\sim}{w})] |$. If $\mathbf{u}^*/\mathbf{u}$ and $\boldsymbol{B}$ are concordant, then $| Bias[\hat{\delta}(\underset{\sim}{w}*)] | \geq | Bias[\hat{\delta}(\underset{\sim}{w})] |$.

Proof:

First, since $w_c = u_c / \sum_{c=1}^{C} u_c$, then

$$u_c^*/w_c = u_c^*/(u_c/\sum_{c=1}^{C} u_c)$$
$$= (u_c^*/u_c)(\sum_{c=1}^{C} u_c),$$

Hence, if $\mathbf{u}^*/\mathbf{u}$ and $\boldsymbol{B}$ are discordant, then $\mathbf{u}^*/\mathbf{w}$ and $\boldsymbol{B}$ are discordant.

Next,

$$|Bias[\hat{\delta}(\underset{\sim}{w})]| - |Bias[\hat{\delta}(\underset{\sim}{w}*)]| = (\sum_{c=1}^{C} w_c B_c - \sum_{c=1}^{C} w_c^* B_c)$$
$$= [\sum_{c=1}^{C} w_c B_c - \sum_{c=1}^{C} (u_c^*/\sum_{c=1}^{C} u_c^*) B_c]$$
$$= [(\sum_{c=1}^{C} u_c^*)\sum_{c=1}^{C} w_c B_c - \sum_{c=1}^{C} u_c^* B_c]/\sum_{c=1}^{C} u_c^*$$

Apply Lemma 4.3, let $i = c$, $a_i = \frac{u_c^*}{w_c}$, $b_i = B_c$, $t_i = w_c$, then we have

$$(\sum_{c=1}^{C} u_c^*)\sum_{c=1}^{C} w_c B_c - \sum_{c=1}^{C} u_c^* B_c = (\sum_{c=1}^{C} w_c \frac{u_c^*}{w_c})(\sum_{c=1}^{C} w_c B_c) - (\underbrace{\sum_{c=1}^{C} w_c}_{=1})(\sum_{c=1}^{C} w_c \frac{u_c^*}{w_c} B_c) \geq 0.$$

Therefore, $|Bias[\hat{\delta}(\underset{\sim}{w})]| - |Bias[\hat{\delta}(\underset{\sim}{w}*)]| \geq 0$. For concordance, the proof is similar. This completes the proof.

Theorem 4.4 provides a method to compare the biases of two propensity score estimators using different weighting schemes by determining whether the ratio of those two distinct vectors of unstandardized weights is discordant with $\boldsymbol{B}$ (e.g. Spearman's rank correlation coefficient equal to one implies concordance). Based on this theorem, we develop the following two corollaries.

Corollary 4.4.1 For any subclassification, let $\mathbf{w}^{EW} = (1/C, \ldots, 1/C)$ be equal weights. Then $|Bias[\hat{\delta}(\underset{\sim}{w}^*)]| \leq |Bias[\hat{\delta}(\underset{\sim}{w}^{EW})]|$ if $\mathbf{u}^*$ and $\boldsymbol{B}$ are discordant,

$|Bias[\hat{\delta}(\underset{\sim}{w}^*)]| \geq |Bias[\hat{\delta}(\underset{\sim}{w}^{EW})]|$ if $\mathbf{u}^*$ and $\boldsymbol{B}$ are concordant.

Corollary 4.4.2 For any subclassification, let $\mathbf{w}^{IV}$ be (standardized) inverse variance weights and $\mathbf{w}^{EW}$ be equal weights. If $\mathbf{V} = (V_1, \ldots, V_C)$ and $\boldsymbol{B}$ are concordant, then $|Bias[\hat{\delta}(\underset{\sim}{w}^{IV})]| \leq |Bias[\hat{\delta}(\underset{\sim}{w}^{EW})]|$. If $\mathbf{V}$ and $\boldsymbol{B}$ are discordant, then $|Bias[\hat{\delta}(\underset{\sim}{w}^{IV})]| \geq |Bias[\hat{\delta}(\underset{\sim}{w}^{EW})]|$.

Proof:

Let $u^* = 1/V_c$ for an IV weight. Hence, discordance between $\mathbf{u}^*$ and $\boldsymbol{B}$ is equivalent to concordance between $\mathbf{V}$ and $\boldsymbol{B}$. Apply Corollary 4.4.1, we have $|Bias[\hat{\delta}(\underset{\sim}{w}^{IV})]| \leq |Bias[\hat{\delta}(\underset{\sim}{w}^{EW})]|$. For discordance, the proof is similar. This completes the proof.

Therefore, as Corollary 4.4.2 indicates, under EF subclassification, if higher variance occurs with larger bias, then the EF-IV estimator has a smaller bias than the EF-EW estimator.

In practice, the Spearman's rank correlation between biases and variances of the within subclass treatment effect estimates is equivalent to the Spearman's rank correlation between the within subclass treatment effect estimates and their variances. Spearman's

rank correlation can be used to evaluate whether this relationship is concordant or discordant.

4.2.3 Comparing variances between IV and EW estimators

Using Lemma 4.2, we develop the second theorem to compare the variances of inverse variance weighted and equal weights estimators under the same subclassification.

Theorem 4.5: Under the same subclassification, let $\mathbf{w}^{\mathrm{IV}}$ be (standardized) inverse variance weights and $\mathbf{w}^{\mathrm{EW}}$ be equal weights, then $Var[\hat{\delta}(\underset{\sim}{w}^{IV})] \le Var[\hat{\delta}(\underset{\sim}{w}^{EW})]$.

Proof:

$$\begin{aligned} Var[\hat{\delta}(\underset{\sim}{w}^{EW})] &= \sum_{c=1}^{C} w_c^2 V_c \\ &= \sum_{c=1}^{C} \frac{1}{C^2} V_c \\ &= \frac{1}{C^2} \sum_{c=1}^{C} V_c \end{aligned}$$

and

$$
\begin{aligned}
Var[\hat{\delta}(\underset{\sim}{w}^{IV})] &= \sum_{c=1}^{C}(w_c^*)^2 V_c \\
&= \sum_{c=1}^{C}\frac{1/V_c^2}{(\sum_{c=1}^{C}1/V_c)^2}V_c \\
&= \frac{\sum_{c=1}^{C}1/V_c}{(\sum_{c=1}^{C}1/V_c)^2} \\
&= \frac{1}{\sum_{c=1}^{C}1/V_c}.
\end{aligned}
$$

We should notice that for two different subclasses, e.g. subclass c and d, we have

$\frac{1}{V_c} \le \frac{1}{V_d} \Leftrightarrow V_c \ge V_d$, this implies the discordance property,

$(\frac{1}{V_c} - \frac{1}{V_d})(V_c - V_d) \le 0$ for all $c,\ d$. Then, by the discordance Inequality in Lemma 4.2,

we have

$\sum_{c=1}^{C}\frac{1}{V_c}\sum_{c=1}^{C}V_c \ge C\sum_{c=1}^{C}\frac{1}{V_c}V_c = C^2$. Therefore,

$$
\begin{aligned}
\sum_{c=1}^{C}\frac{1}{V_c}\sum_{c=1}^{C}V_c \ge C^2 &\Leftrightarrow \frac{1}{\sum_{c=1}^{C}\frac{1}{V_c}\sum_{c=1}^{C}V_c} \le \frac{1}{C^2} \\
&\Leftrightarrow \frac{1}{\sum_{c=1}^{C}\frac{1}{V_c}} \le \frac{1}{C^2}\sum_{c=1}^{C}V_c \\
&\Leftrightarrow \sum_{c=1}^{C}(w_c^*)^2 V_c \le \sum_{c=1}^{C}w_c^2 V_c \\
&\Leftrightarrow Var[\hat{\delta}(\underset{\sim}{w}^{IV})] \le Var[\hat{\delta}(\underset{\sim}{w}^{EW})].
\end{aligned}
$$

This completes the proof.

Theorem 4.5 indicates that the EF-IV estimator always has a variance no larger than the EF-EW estimator. This theorem has the following corollary.

Corollary 4.5.1 Let $\mathbf{w}^{IV}$ be (standardized) inverse variance weights and $\mathbf{w}^{EW}$ be equal weights. If V_c is a non-decreasing function of B_c, then $|\, Bias[\hat{\delta}(\underset{\sim}{w}^{IV})]\,| \leq |\, Bias[\hat{\delta}(\underset{\sim}{w}^{EW})]\,|$ and $Var[\hat{\delta}(\underset{\sim}{w}^{IV})] \leq Var[\hat{\delta}(\underset{\sim}{w}^{EW})]$, so $RMSE[\hat{\delta}(\underset{\sim}{w}^{IV})] \leq RMSE[\hat{\delta}(\underset{\sim}{w}^{EW})]$.

4.3 Comparing variances between EF-IV and EV-IV estimators

In this section, we derive a theorem for comparing variances between the EF-IV and EV-IV estimators. Specifically, we provide a condition under which the variance of the EF-IV estimator is no larger than the variance of the EV-IV estimator. We use V^{EV} to denote the variance of the within subclass treatment effect estimator under the EV approach, and V_c^{EF} to denote the variance of the within subclass treatment effect estimate under the EF approach.

Theorem 4.6 For the EV subclassification, let $\mathbf{w}^{EV\text{-}IV} = (1/C, \ldots, 1/C)$ be inverse variance (equal) weights. For the EF subclassification, let the standardized inverse variance weights be $\mathbf{w}^{EF\text{-}IV}$. Then $Var[\hat{\delta}(\underset{\sim}{w}^{EF-IV})] \leq Var[\hat{\delta}(\underset{\sim}{w}^{EV-IV})]$ if and only if $V^{EV} \geq \dfrac{C}{\sum_{c=1}^{C} \dfrac{1}{V_c^{EF}}}$, the equality holds for $V_c^{EF} = V^{EV}$.

Proof:

From the proof of Theorem 4.5, $Var[\hat{\delta}(\underset{\sim}{w}^{EF-IV})] = \dfrac{1}{\sum_{c=1}^{C} \dfrac{1}{V_c^{EF}}}$.

For the EV-IV estimator, the weight of subclass c is

$$\begin{aligned} w_c^{EV-IV} &= \frac{1}{V^{EV}} / \sum_{c=1}^{C} \frac{1}{V^{EV}} \\ &= 1/C. \end{aligned}$$

Hence, we have

$$\begin{aligned} Var[\hat{\delta}(\underset{\sim}{w}^{EV-IV})] &= \sum_{c=1}^{C} \frac{1}{C^2} V^{EV} \\ &= \frac{1}{C^2} \sum_{c=1}^{C} V^{EV} \\ &= \frac{1}{C^2} C V^{EV} \\ &= \frac{V^{EV}}{C}. \end{aligned}$$

Thus,

$$\begin{aligned} Var[\hat{\delta}(\underset{\sim}{w}^{EF-IV})] \le Var[\hat{\delta}(\underset{\sim}{w}^{EV-IV})] &\Leftrightarrow \frac{1}{\sum_{c=1}^{C} \frac{1}{V_c^{EF}}} \le \frac{V^{EV}}{C} \\ &\Leftrightarrow \frac{1}{V^{EV}} \le \frac{1}{C} \sum_{c=1}^{C} \frac{1}{V_c^{EF}} \\ &\Leftrightarrow V^{EV} \ge \frac{C}{\sum_{c=1}^{C} \frac{1}{V_c^{EF}}} \end{aligned}$$

Here, $\dfrac{C}{\sum_{c=1}^{C} \dfrac{1}{V_c^{EF}}}$ is the harmonic mean of the V_c^{EF} (s).

This completes the proof.

Theorem 4.6 indicates that if the variance of the EV subclassification is bigger than the harmonic mean of the subclass-specific variances of the EF subclassification, than the EF-IV estimator has a smaller variance than the EV-IV estimator, and vice versa.

We develop the following corollary of Theorem 4.6 for the situation when the EV subclassification is unable to produce a lower variance than the average variance of the subclass treatment effect estimates under the EF subclassification.

Corollary 4.6.1 If $V^{EV} \geq \frac{1}{C}\sum_{c=1}^{C} V_c^{EF}$, then $Var[\hat{\delta}(\underset{\sim}{w}^{EF-IV})] \leq Var[\hat{\delta}(\underset{\sim}{w}^{EV-IV})]$.

Proof:

By Jensen's Inequality, we have $\frac{1}{C}\sum_{c=1}^{C}\frac{1}{V_c^{EF}} \geq \frac{1}{\frac{1}{C}\sum_{c=1}^{C} V_c^{EF}}$. And the inequality

$\frac{1}{\frac{1}{C}\sum_{c=1}^{C} V_c^{EF}} \geq \frac{1}{V^{EV}}$ is true if and only if $\frac{1}{C}\sum_{c=1}^{C} V_c^{EF} \leq V^{EV}$. This completes the proof.

Corollary 4.6.1 indicates that when the EV subclassification does not produce a lower variance than the average variance of the subclass treatment effect estimates under the EF subclassification, then the EF-IV estimator has a smaller variance than the EV-IV estimator.

We also extend the theory work to the situation with multiple covariates and a treatment indicator in the Appendix A4.4.

Chapter 5. PROPENSITY SCORES BALANCING SUBCLASSIFICATION

In this chapter, we propose a novel propensity scores balancing subclassification, the PSB method. Two weighting schemes will be applied to a PSB estimator, which include the inverse variance (IV) weights and proportional weights (PW). We implement simulations to examine the performance of the PSB estimators when compared with other treatment effect estimators studied in Chapter Three, in terms of bias, variance and RMSE. Our PSB estimator tends to use many more subclasses than the five that are typical of the other propensity score estimators. For this reason, we compare our PSB estimator with the other propensity score estimators when they use more than the traditional number of five subclasses. We also examine the relative performance of our estimator and the other estimators in the situation where a proportion of control subjects have propensity scores lower than the minimum propensity score in the treatment group. This situation is discussed by Dehejia and Wahba (1999) and Strumer et al. (2007).

In the first section, we describe the PSB subclassification method. In the second section, under a correctly specified regression and propensity score model, we implement a simulation to examine the performance of the PSB-PW and PSB-IV estimators in comparison with the ordinary least square (OLS) estimator and other propensity score estimators in Chapter Three. In the third section, we conduct a second simulation to model the situation of control subjects with propensity scores lower than the minimum estimated propensity score in the treatment group. We apply the PSB method by restricting the lowest subclass size to avoid trimming (Strummer et al., 2007), but such a

restriction is not feasible for the other propensity score subclassification approaches. A summary of simulation results is provided in the fourth section. In the Appendix Section A5.4, we provide a flow chart diagram of the simulation procedure.

5.1 PSB subclassification method

Rosenbaum and Rubin's (1983) first propensity score theorem states that conditional on propensity scores, the observed covariates and the treatment indicator are independent. This indicates that if a propensity score subclass is "homogeneous," implying the subjects within it have similar propensity scores, then the covariates for the treated subjects and the control subjects in that subclass should be approximately independent of their treatment assignment. Subsequently, the within subclass treatment comparison for homogeneous subclasses should be unbiased in terms of observed covariates. Our approach, therefore, is to test the difference in mean propensity scores between the treatment group and the control group to assess the balance within each subclass. Dehejia and Wahba (1999, 2002) apply a two-sample t-test to assess whether the means of the estimated propensity scores within each subclass are "identical" (or "balanced" in Imbens 2004). In addition, it has been suggested that implementing a large number of subclasses would reduce more bias from the propensity scores estimator (Cochran 1968, Imbens 2004, Myers and Louis 2007). These discussions provide some background for an alternative subclassification scheme that involves forming a large number of subclasses, each with homogeneous propensity scores.

In our propensity scores balancing (PSB) subclassification method, we form subclasses within which estimated propensity scores in the treatment group and the control group are tested for homogeneity by using a two-sample t-test. To compute the variance and perform a two-sample t-test, we require each PSB subclass be formed with at least two observations in both the treatment group and the control group. When this requirement is satisfied, the researcher can decide how many additional observations would be assigned initially to form a subclass. The PSB subclassification approach adds observations one at a time to the subclass to adjust its boundary. The subclass boundary will be formed when the two-sample t-test of the estimated propensity scores between the treatment group and the control group becomes "nonsignificant'' (a pretermined p-value can be used, such as 0.05 or 0.01). For detailed information, see Appendix-Figure A5.1 and A5.2.

According to Rosenbaum and Rubin's (1983) theorem, for each PSB subclass with homogeneous propensity scores, the distribution of the covariates within the subclass will be identical. This in turn will help to remove bias due to observed covariates from the propensity scores estimator. The PSB subclassification method uses no predetermined number of subclasses. The PSB subclasses will not necessarily, or even typically, have equal sizes. Therefore, when using subclass sizes as weights, the PSB approach uses proportional weights rather than the equal weights under the EF approach.

The algorithm of the PSB subclassification procedure is provided below:

(1) Sort the estimated propensity scores in an ascending order;

(2) From the lower end of the estimated propensity scores, assign a pre-determined number of observations (e.g., we use 10 in our simulation study) to form a subclass. If there are at least two observations in the treatment group and the control group, then move to step (3). Otherwise, the size of this subclass will be increased by one observation at a time until both the treatment group and the control group have at least two observations;

(3) Perform a two sample t-test on the estimated propensity scores between the treatment group and the control group in the subclass from step (2). Assume p-value = 0.05 is adopted, if the p-value ≥ 0.05 then the current observations will form the subclass; if the p-value < 0.05, the size of this subclass will be adjusted by adding one observation at a time until p-value ≥ 0.05, is achieved from a two sample t-test;

(4) Repeat step (2) and (3) to form additional subclasses until the 50th percentile of the estimated propensity scores is reached;

(5) From the upper end of the estimated propensity scores, assign a pre-determined number of observations (e.g., we use 10 in our simulation study). Follow a similar procedure as in step (2) to achieve a minimum of two observations from both the treatment group and the control group. Then repeat steps (3) and (4) to form subclasses until the 50th percentile of the estimated propensity scores is reached;

(6) Combine the two nearest subclasses on either side of the 50th percentile in order to form the subclass around the 50th percentile of the estimated propensity scores.

A flow chart diagram of the above PSB procedure is provided in the Appendix Section A5.1.

5.2 A simulation study under a correctly specified regression and propensity scores model

In this section, we implement a simulation study by generating the treatment indicator and the outcome variable. The bias of the mean, variance and RMSE of these estimators are obtained: OLS, equal frequency equal weights (EF-EW), equal frequency inverse variance weights (EF-IV), equal variance inverse variance weights (EV-IV), propensity scores balancing proportional weights (PSB-PW) and propensity scores balancing inverse variance weights (PSB-IV).

5.2.1 Simulating data involving two independent covariates (x_1, x_2)

We define two covariates, $x_{11}, \ldots, x_{1n} \sim$ iid. N(0, 1), and independently, $x_{21}, \ldots, x_{2n} \sim$ iid. N(0, 1). In this simulation, we take the treatment indicator values (z_i) to be simulated as independent Bernoulli random variables under a logistic model:

$$z_i \mid x_{1i}, x_{2i} \sim \text{Bernoulli}(\pi_i)$$

$$\pi_i = \Pr(z_i = 1 \mid x_{1i}, x_{2i}) = \{1 + \exp[-(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i})]\}^{-1} \quad (5.1)$$

We generate the outcome variable using the model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \delta z_i + e_i \quad (5.2)$$

where $e_1, \ldots, e_n \sim$ iid. $N(0, 1)$ with $\mathbf{e} \perp\!\!\!\perp (\mathbf{x}_1, \mathbf{x}_2, \mathbf{z})$, given $\mathbf{e} = (e_1, \ldots, e_n)$, $\mathbf{x}_1 = (x_{11}, \ldots, x_{1n})$, $\mathbf{x}_2 = (x_{21}, \ldots, x_{2n})$, $\mathbf{z} = (z_1, \ldots, z_n)$ and $\perp\!\!\!\perp$ denotes independence. For additional information, see Appendix A3.1, Table A3.2.

We assume the same parameter values that were introduced by Drake (1993), where $\beta_0 = 1$, $\beta_1 = 1$, $\delta = 1$, $\beta_2 = 1$; $\gamma_0 = 0$, $\gamma_1 = 0.4$, $\gamma_2 = 0.4$. These parameters provide a simple model to implement in the simulation. This set of coefficients gives both covariates an equal influence on generating the propensity scores and the outcome variable. Based on the performance pattern of the EF and EV approaches observed in Chapter Three, using these parameter values in an evaluation of the PSB method provides an indication of whether the values of β_2 and γ_2 should be changed under correctly specified propensity score models and regression models.

5.2.2 Estimation of regression and propensity score models

We will fit the regression model to obtain the OLS estimator of the treatment effect and fit the logistic regression model to obtain the propensity score estimators of the treatment effect. Both the regression model and the propensity scores model use the correct specification given above. Table 5.1 summarizes the models used to fit the outcome model or the propensity scores model (subscript i is ignored for simplicity).

Table 5. 1 Model specifications fitting the outcome or the propensity score models

Covariates	OLS estimator	
(x_1, x_2)	True propensity scores: $\pi = \{1 + \exp[-(\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2)]\}^{-1}$	Outcome model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \delta z + e$
	Propensity scores estimators (EF-EW, EF-IV, EV-IV, PSB-EW, PSB-IV)	
	Propensity scores model: $e(\mathbf{x}) = \{1 + \exp[-(\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2)]\}^{-1}$	True outcome: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \delta z + e$

5.2.3 Measuring the performance of treatment effect estimators

We generated samples to assess the performance of the different treatment effect estimators. We created 1000 simulated samples, each with 1000 observations. Among

those 1000 simulated samples, we obtained treatment effect estimators under each of the methods of Table 5.2. We computed bias, variance and root mean square error for all estimators. The PSB approach typically results in a very large number of subclasses. In our simulation, the PSB produced between 56 and 88 subclasses. Since each random sample contains 1000 observations, EF or EV approaches may not be able to produce 50 or more subclasses because there may not be enough observations in either the treatment group or the control group for some subclasses.

Table 5. 2 The bias of the mean, variance and RMSE for the treatment effect estimators using a correctly specified propensity scores model and a correctly specified regression model for covariates (x_1, x_2)

Estimator	Number of Subclasses	bias average	Var($\hat{\delta}$)	RMSE
OLS		0.0015	0.0046	0.0680
PSB-PW	range: (56, 88)	0.0100	0.0054	0.0742
PSB-IV		0.0045	0.0085	0.0922
EF-EW	5	0.0793	0.0054	0.1081
EF-IV		0.0660	0.0053	0.0981
EV-IV		0.0723	0.0053	0.1028
EF-EW	10	0.0323	0.0051	0.0781
EF-IV		0.0250	0.0050	0.0752
EV-IV		0.0296	0.0051	0.0771
EF-EW	20	0.0144	0.0049	0.0716
EF-IV		0.0106	0.0050	0.0718
EV-IV		0.0184	0.0051	0.0727

Table 5.2 indicates that the PSB approach removes more bias as compared to the EF and EV approaches when five subclasses are used for those approaches. The biases of the PSB estimators are similar to the OLS estimator. The variance of the PSB-IV estimator is larger than those of the other estimators. The RMSE of the PSB estimators

are smaller than those of the other propensity score estimators when five subclasses are used.

The average bias of the EF and the EV estimators improve when the five subclasses are split to create 10 subclasses, and more so for 20 subclasses, making them closer to those using the PSB approach. The variances of the EF and EV estimators using 10 or 20 subclasses remain similar to those obtained when using five subclasses. The RMSE of the EF and the EV estimators further improve when using 10 subclasses.The RMSE of the EF and EV estimators using 20 subclasses are smaller than those of the PSB estimators.

Overall, the PSB approach is similar to the more traditional approaches when more than the usual number of five subclasses is used in those approaches.

5.3 A simulation study with imbalance in the lowest subclass

5.3.1 Simulating data and estimating the treatment effect

Expanding on the model in Section 5.2, we now take one of the two independent covariates, x_2, to have a different mean and a higher SD than before; namely, let $x_{21}, \ldots, x_{2n} \sim$ iid. N(1, 2). In this simulation, the treatment indicator values (z_i) remain simulated as independent Bernoulli random variables under a logistic model as illustrated in equation (5.1). We intend to simulate the situation in which a proportion of control

subjects have propensity scores lower than the minimum estimated propensity score among subjects in the treatment group. We accomplish this by introducing a negative coefficient for covariate x_2, we choose $\gamma_2 = -1.1$. All other parameters remain the same as in Section 5.2.1.

The control subjects with lower propensity scores than the minimum propensity score among subjects in the treatment group are then excluded from the analysis (Dehejia and Wahba 1999, Stürmer et al. 2007). However, if we want to make inferences based on the entire sample, we can implement the PSB method simply by restricting the size of the lowest subclass. The adjusted algorithm for the PSB approach in this situation is provided as follows:

(1) From the lower end of the estimated propensity scores, assign a pre-determined number of observations (e.g., we use 10 in our simulation study) to form a subclass. The size of this subclass will be increased by one observation at a time until the treatment group has two observations;

(2) From the upper end of the estimated propensity scores, assign a pre-determined number of observations (e.g., we use 10 in our simulation study) to form a subclass. If there are at least two observations in the treatment group and the control group, then move to step (3). Otherwise, the size of this subclass will be increased by one observation at a time until both the treatment group and the control group have at least two observations;

(3) Perform a two sample t-test on the estimated propensity scores between the treatment group and the control group in the subclass from step (2). Assume p-value = 0.05 is adopted, if the p-value ≥ 0.05 then the current observations will form the subclass; if the p-value < 0.05, the size of this subclass will be adjusted by adding one observation at a time until a p-value ≥ 0.05, is achieved from a two sample t-test;

(4) Repeat step (2) and (3) to form additional subclasses before reaching the upper bound of the subclass formed in step (1);

(5) Adjust the upper bound of the subclass formed in step (1) to be adjacent to the PSB subclasses in step (4).

For detailed information on this PSB subclassification procedure by restricting the size of the lowest subclass, see Appendix-Figure A5.3.

We fit the same models in Table 5.1 to obtain the OLS estimator and the propensity score estimators of the treatment effect for this second simulation. We wished to obtain 10 EF subclasses and to compare estimators. However, we found that 15.6% of the simulated random samples did not have enough treatment subjects in the first quantile of the estimated propensity scores to allow further splitting of EF subclasses. We collected 1000 of those simulated samples, each sample containing 1000 random observations. We kept the number of EF subclasses at five to apply the EF approach.

5.3.2 Simulation results

Table 5. 3 The bias of the mean, variance and RMSE for the treatment effect estimators under the condition that a proportion of control subjects have propensity scores lower than the minimum propensity score among treated subjects

Estimator	Number of Subclasses	bias average	$\mathrm{Var}(\hat{\delta})$	RMSE
OLS		0.0020	0.0070	0.0839
PSB-PW	range: (39, 60)	-0.1673	0.0634	0.3024
PSB-IV		-0.0558	0.0574	0.2460
EF-EW	5	-0.3159	0.0514	0.3888
EF-IV		-0.2287	0.0377	0.3000
EV-IV		-0.2936	0.0112	0.3121

Table 5.3 indicates that the PSB-IV estimator has the lowest average bias and RMSE when compared to the other propensity score estimators. The PSB-PW estimator produces a larger average bias and RMSE than the PSB-IV estimator, but its average bias is still lower than the EF and EV estimators. The variances of the PSB estimators are larger than those of the other estimators. The EF and EV estimators all produce much larger average biases and RMSE than those of the PSB estimators. Because the regression model is correctly specified, as we expected, the OLS estimator has the lowest average bias and RMSE when compared to propensity score estimators. These simulation results suggest that under the situation of correct regression model

specification, researchers should use the OLS estimator because propensity score adjustments are not needed.

5.4 Summary of PSB simulation studies.

Our simulation studies indicate that under a correctly specified regression model and a propensity scores model:

- The PSB approach removes more bias as compared to the EF or EV approaches with five subclasses. The biases of the PSB estimators are similar to the OLS estimator. The RMSE of the PSB-PW estimator is slightly larger than that of the OLS estimator and is smaller than that of the other propensity score estimators.
- The average biases of the EF and EV estimators improve when observations are split to create 10 or 20 subclasses.
- When a proportion of the control subjects have propensity scores lower than the minimum of the estimated propensity scores in the treatment group, the PSB-IV estimator produces the lowest average bias and RMSE among propensity score estimators by restricting the size of the lowest subclass. However, the OLS estimator provides the lowest average bias and RMSE when compared to propensity score estimators, which suggests that no propensity score adjustment is needed under this condition.

In circumstances where a linear model may be inappropriate, and when a proportion of control subjects have propensity scores lower than the minimum propensity score among treated subjects, the PSB-IV estimator approach may be used to produce estimators with lower average bias than the EF and EV approaches by restricting the size of the lowest subclass.

Chapter 6. CONCLUSIONS

In this thesis, we explored propensity score adjustment methods and proposed a novel subclassification approach to estimate the "treatment" effect in observational studies. In Chapter Two, we reviewed the single covariate adjustment method using subclassification. We discuss propensity score adjustments that incorporate multiple covariates in observational studies. We reviewed propensity score methodology and approaches adopted by researchers, such as matching, equal frequency (EF) subclassification, further splitting of EF subclasses, and equal variance (EV) subclassification. Two weighting schemes used in propensity score subclassification estimators were also discussed: equal weights (EW) and inverse variance (IV) weights.

In Chapter Three, we evaluated the EV subclassification approach under model misspecification. We also examined EF subclassification estimators and the ordinary least square (OLS) estimator of the treatment effect. Our simulation results indicated that under correctly specified propensity score models, the EF-IV estimator resulted in a lower bias and root mean square error (RMSE) than the EF-EW and EV-IV estimators. After excluding a quadratic term to misspecify the propensity score and regression models, the EF-IV estimator resulted in the lowest bias and RMSE as compared to the EF-EW, EV-IV and OLS estimators (the OLS estimator generated the largest bias and RMSE as compared to the propensity score estimators). When the propensity score

models and regression models were misspecified by omitting an independent covariate, we concluded that none of these estimators work well.

Our simulation showed that the EV subclassification approach requires much more computation than the EF subclassification method. For quadratic term misspecification, depending on the coefficient of the quadratic term used to generate the propensity scores, 6-19% of the time the EV subclassification procedure did not produce five EV subclasses from 1000 random samples (Appendix A3.5.3).

In Chapter Four, we developed three theorems that provide theoretical results for comparing the different propensity score subclassification estimators. Our first theorem indicates that under EF subclassification, if higher variation occurs with larger bias for within subclass treatment effect estimates, then the EF-IV estimator has smaller overall bias than the EF-EW estimator. Our second theorem indicates that the EF-IV estimator always has a variance no larger than the EF-EW estimator. This theoretical finding provides corroboration to the statement about the IV weights underestimating the overall variance of the treatment effect estimator (Chapter Two). This underestimation leads to the EF-IV estimator with lower variance than the EF-EW estimator. Our third theorem indicates that when the variance of an EV subclassification is larger than the harmonic mean of the variances of the within subclass treatment effect estimators, under EF subclassification, then the EF-IV estimator has lower variance than the EV-IV estimator.

In Chapter Five, we proposed a new propensity score balancing (PSB) subclassification method. Our proposal is a subclassification approach to form "homogeneous" propensity score subclasses. We compared this method to the more traditional approaches when five subclasses are used. Our simulation results indicate that under a correctly specified model for propensity scores and a correctly specified regression model, the PSB approach removes more bias than the EF and EV approaches with five subclasses. The biases of both PSB estimators are similar to the OLS estimator, while the variances of the PSB estimators are larger than other estimators. This is because the PSB approach tends to generate a large number of subclasses with potentially a small number of observations assigned initially to each subclass. Smaller subclass sizes have larger subclass variation, which in turn increases the overall variances of the PSB estimators. We also simulate a situation where a proportion of the control subjects have propensity scores lower than the minimum of the estimated propensity scores in the treatment group. By restricting the size of the lowest subclass, the PSB-IV estimator produces the lowest bias and RMSE among the propensity score estimators even though it has a larger variance than the EF and EV estimators. However, due to the lowest average bias and RMSE provided by the OLS estimator in comparison with propensity score estimators, which suggests that researchers should use the OLS estimator under this condition. Our proposed PSB subclassification method can produce many subclasses conditional on the data. Cochran (1968), Imbens (2004) and Myers and Louis (2007) suggested that generating a large number of subclasses would reduce more bias from the propensity scores estimator. The trade-off is that the PSB estimators can have larger variances due to smaller subclass sizes.

6.1 Discussion

The major results of this thesis research can be summarized as follows. First, further splitting of five EF subclasses into 10 or 20 subclasses improves the overall bias of the treatment effect estimator under correctly specified propensity score models. Hence, this makes the EF-EW estimator competitive with other propensity score estimators given its ease of use under this condition.

Second, our simulation results in Chapter Three indicate that the EF-IV estimator provides the lowest bias and variance under a quadratic term misspecification when compared to the EF-EW, EV-IV and OLS estimators.

Third, the concordance or discordance between the bias and variance of the subclass treatment effect estimates should be considered when assigning weights to the propensity score estimator. In our theoretical investigation, Corollary 4.4.1 indicates that when higher variance occurs with larger bias under the EF subclassification, the inverse variance weights produce smaller bias as compared to equal weights.

Fourth, Theorem 4.4 provides a way to evaluate the overall bias between two weighting schemes by comparing the ratio of the different weights; this theorem determines the weighting scheme that produces the smallest overall bias.

Fifth, we proposed our PSB subclassification method. This method attempts to create subclasses that are homogeneous in their propensity score distributions. Our simulation results indicate that, in some circumstances, the PSB-IV estimator produces a smaller bias and RMSE than the EF-EW, EF-IV and EV-IV estimators.

6.2 Future work

Our proposed PSB subclassification method can be further modified to address trimming. In Section 5.3, we simulated control subjects with low estimated propensity scores. We could also consider treatment subjects that have high estimated propensity scores. Alternatively, we could consider both control subjects with low estimated propensity scores and treatment subjects with high estimated propensity scores. To accommodate these conditions, we can adapt a restricted PSB subclassification method (rPSB). Using rPSB subclassification, the estimated propensity scores between the treatment and the control group may not be "homogeneous" for the lowest and highest subclasses. We can expand our simulations to the three scenarios of possible trimming subjects. We can also implement model misspecifications. We can then evaluate the PSB and rPSB subclassification methods using simulations. We can compare the performances of PSB and rPSB estimators with other treatment effect estimators studied in Chapter Three.

Secondly, we propose exploring an alternative weighting scheme by using a non-parametric measure: the inverse Kolmogorov-Smirnov distance (K-S distance) of the

estimated propensity scores between the treatment group and the control group within a subclass. This is called inverse K-S distance weights. The K-S distance is a non-parametric estimate of the maximum distance between two cumulative distribution functions (cdf.). For propensity score estimators, equal weights have been adopted by most researchers and inverse variance weights have been used by some researchers; we want to provide an extra choice of weights. Propensity scores summarize the information from all observed covariates. Hence, we could use the estimated propensity scores to obtain the K-S distance for each subclass between the treatment and control groups. A relatively large K-S distance estimate may indicate that there is an imbalance of covariates between the treatment and control groups for a subclass. This will induce bias to the estimated treatment effect within the subclass. Using inverse K-S distances creates smaller weights for imbalanced subclasses. Therefore, more bias might be removed from imbalanced subclass treatment effect estimators with smaller inverse K-S distance weights, which in turn produces lower overall bias. In addition to equal weights and inverse variance weights, researchers will have another weighting scheme to choose from and can apply Theorem 4.4 to determine which weighting scheme would produce the smallest overall bias of the treatment effect estimator.

We feel our research contributes to the field of propensity score adjustments by providing new theorems to compare the overall bias and variance between propensity score estimators using different weighting schemes. We also present an alternative subclassification method that focuses on creating homogeneous propensity score subclasses for reducing the overall bias that may be used in some circumstances.

BIBLIOGRAPHY

Lesser, Virginia M. (2003). ST531 Sampling Methods Class Notes, unpublished, Oregon State University.

Qu, Annie (2004). ST599 Longitudinal Data Analysis Class Notes, unpublished, Oregon State University.

Annie Qu (2002). ST561 Statistical Theory Notes, unpublished, Oregon State University.

David A. Dr. Birkes (2006). ST663 Advanced Theory of Statistics Notes, unpublished, Oregon State University.

Lohr, Sharon L. (1999). Sampling: Design and Analysis. Textbook, Duxbury Press.

Little, J.A.R. and D.B. Rubin (2002). *Statistical Analysis with Missing Data* (2nd ed.). New Jersey: John Wiley and Sons, Inc.

Heeringa, Steven G., West, Brady T., Berglind, Patricia A. (2010). Applied Survey Data Analysis. Textbook, CRC Press.

W. G. Cochran (1968). The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*, Vol. 24, No. 2 (Jun., 1968), pp. 295-313.

Cochran, W.G., and Rubin, D.B. (1973). Controlling Bias in Observational Studies: A Review. *Sankya*, Ser. A, 35, 417-446.

D. B. Rubin (1979). Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. *Journal of the American Statistical Association*, Vol. 74, No. 366. (June 1979), pp. 318-328.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1, pp. 41-55.

Paul R. Rosenbaum; Donald B. Rubin (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, Vol. 79, No. 387. (Sep., 1984), pp. 516-524.

Donald B. Rubin (1997). Estimating Causal Effects from Large Data Sets Using Propensity Scores. 15 Octoder 1997 • *Annals of Internal Medicine* • VolunwS 127 • Number 8 (Part 2) • pp. 757-763.

D'Agostino, R. B. Jr. (1998). Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Nonrandomized Control Group. *Statistics in Medicine,* 17, 2265-2281 (1998).

Roderick J. Little and Donald B. Rubin (2000). Causal Effects in Clinical and Epidemiological Studies via Potential Outcomes, concepts and analytical Approaches. *Annu. Rev. Public Health.* 2000. 21:121–45.

Paul R. Rosenbaum; Donald B. Rubin (1985). Constructing a control group using multivariate matched sampling incorporating the propensity score. *The American Statistician*, February 1985, Vol. 39, No. 1, 33-38.

Donald B. Rubin and Neal Thomas (2000). Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates. *Journal of the American Statistical Association*, Vol. 95, No. 450. (Jun., 2000), pp. 573-585.

Rosenbaum, Paul R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82, 387–394.

Elaine Zanutto, Bo Lu, Robert Hornik (2005). Using Propensity Score Subclassification for Multiple Treatment Doses to Evaluate a National Antidrug Media Campaign. *Journal of Educational and Behavioral Statistics*, Vol. 30, No. 1, 59-73 (2005).

Little, R.J.A. (1986) p. 146. Survey nonresponse adjustments for estimates of means. *International Statistical Review*. 54(2), 139-157.

David, M.H., Little, R.J.A., Samuhel, M., and Triest, R. (1983) p. 169. Nonrandom nonresponse models based on the propensity to respond. Proceedings of the Business and Economics Section of the American Statistical Association. pp. 168-173.

Eltinge, J.L. and Yansaneh, I.S. (1997). Diagnostics for formulation of nonresponse adjustment cells with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology*, 23(1), 33-40.

Smith, Philip J., Rao, J.N.K., Battaglia, Michael P., Daniels, Danni, Ezzati-Rice, Trena, Khare, Meena. (2000). Compensating for Nonresponse Bias in the National Immunization Survey Using Response Propensities. Proceedings of the Section on Survey Research Methods of the American Statistical Association. pp. 641-646.

Barbara Lepidus Carlson and Stephen Williams (2001). A Comparison of Two Methods to Adjust Weights for Nonresponse: Propensity Modeling and Weighting Class Adjustments. Proceedings of the Section on Survey Research Methods of the American Statistical Association, 00111.

Nuria Diaz-Tena, Frank Potter, Michael Sinclair and Stephen Williams (2002). Logistic Propensity Models to Adjust for Nonresponse in Physician Surveys. Proceedings of the Section on Survey Research Methods of the American Statistical Association, 000175.

Ray Chambers (2007). Non-Bayesian Multiple Imputation Discussion. *Journal of Official Statistics*, Volume 23, No. 4, 2007, pp. 453-454. In discussion of Jan F. Bjørnstad (2007). Non-Bayesian Multiple Imputation. *Journal of Official Statistics*, Volume 23, No. 4, 2007, pp. 433-452.

Jan F. Bjørnstad (2007). Non-Bayesian Multiple Imputation Rejoinder. *Journal of Official Statistics*, Volume 23, No. 4, 2007, pp. 485-491. In rejoinder of Jan F. Bjørnstad (2007). Non-Bayesian Multiple Imputation. *Journal of Official Statistics*, Volume 23, No. 4, 2007, pp. 433-452.

Jan F. Bjørnstad (2007). Non-Bayesian Multiple Imputation. *Journal of Official Statistics*, Volume 23, No. 4, 2007, pp. 433-452.

Roderick J. Little and Donald B. Rubin (2002). Statistical Analysis with Missing Data. 2nd edition, *John Wiley & Sons, Inc.* 2002, pp.4.

Drake, Christiana (1993). Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect. *Biometrics*, Vol. 49, No. 4 (Dec., 1993), pp. 1231-1236.

Katherine Huppler Hullsiek, Thomas A. Louis (2002). Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics* (2002), Vol. 2, No. 4, pp. 179-193 - Biometrika Trust.

Jessica A. Myers and Thomas A. Louis (2007). Johns Hopkins University, Dept. of Biostatistics *Working Papers*, 2007.

Jessica A. Myers and Thomas A. Louis (2010). Regression Adjustment and Stratification by Propensity Score in Treatment Effect. *Working Papers*, Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, 2010.

Caliendo M. and Kopeinig S. (2008) Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys* (2008) Vol. 22, No. 1, pp. 31–72.

Imbens, G. (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics* 86(1): 4–29.

Dehejia, R.H. and Wahba, S. (1999) Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94(448): 1053–1062.

Dehejia, R.H. and Wahba, S. (2002) Propensity score matching methods for nonexperimental causal studies. *Review of Economics and Statistics* 84(1): 151–161.

Aakvik, A. (2001) Bounding a matching estimator: the case of a Norwegian training program. *Oxford Bulletin of Economics and Statistics* 63(1): 115–143.

Til Stürmer, Manisha Joshi, Robert J. Glynn, Jerry Avorna,Kenneth J. Rothman, Sebastian Schneeweiss (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology* 59 (2006) 437–447.

Til Stürmer, Manisha Joshi, Robert J. Glynn, Jerry Avorna,Kenneth J. Rothman, Sebastian Schneeweiss (2007). *Journal of Clinical Epidemiology* 60 (2007) 437–447.

Nitis Mukhopadhyay (2000). *Probability and Statistical Inference*. Marcel Dekker, Inc., New York, 2000. Theorem 3.5.2, p. 128.

George Casella and Roger L. Berger (2001). *Statistical Inference 2nd*. Duxbury Press, 2002. Theorem 4.7.9 (Covariance Inequality), p.192.

Milton Abramowize and Irene A. Stegun (1970). *Handbook of Mathematical Functions*. Dover Publications, Inc., New York, 1970, p. 11.

D. S. Mitrinović (1970). *Analytic Inequalities*. Springer, Berlin, 1970, p. 36.

Gh. Toader (1996). Note on Chebyshev's Inequality for Sequences, p. 317 and p. 322. *Discrete Mathematics* 161 (1996) 317-322.

S.M. Shah (1966). On Estimating the Parameter of a Doubly Truncated Binomial Distribution. *Journal of the American Statistical Association*, Vol. 61, No. 313. (Mar., 1966), pp. 259-263.

APPENDIX

A1.1 Glossary

MCAR: missing complete at random

MAR: missing at random or ignorable

pdf.: probability density function

cdf.: cumulative distribution function

EF: equal frequency (subclassification)

EV: equal variance (subclassification)

EW: equal weights

IV: inverse variance weights

OLS: ordinary least square

PR: percentage relative

BM: bias of the median

MB: mean bias (bias of the mean)

SD: standard deviation

RMSE: root mean square error

PSB: propensity score balancing subclassification

K-S distance: Kolmogorov-Smirnov distance

A3.1 Acronyms, Notations and simulation procedure diagram

Table A3.1 summarizes the acronyms used in this chapter.

Table A3. 1 Acronyms

Acronym	Description
OLS	ordinary least square
EF	equal frequency (subclassification)
EV	equal variance (subclassification)
EW	equal weights
IV	inverse variance weights
PR	percentage relative
BM	bias of the median
MB	mean bias (bias of the mean)
SD	standard deviation
RMSE	root mean square error

Table A3.2 summarizes the notations used in this chapter.

Table A3. 2 Notations

Notation	Description	Section
n	total number of subjects in the sample	§ 3.1 Simulating data
x_{1i}, x_{2i}	two independent covariates for subject i in the model involving (x_1, x_2), for i = 1, …, n	
x_i	single covariate for subject i in the model involving (x, x^2)	
γ_0, γ_1, γ_2	true logistic regression coefficients	
π_i	true propensity scores for subject i	
z_i	treatment indicator for subject i, let $z_i = 1$ if subject i is in the treatment group and $z_i = 0$ if subject i is in the control group	
δ	true treatment effect	
β_0, β_1, β_2	true regression coefficients	
e_i	error term for subject i	
y_i	the outcome of subject i	
$e(\mathbf{x}_i)$	propensity scores for subject i	§ 3.2 PS model
(c)	superscript c used to denote (within) the subclass c for most of the notations above in Section 3.1	§ 3.3 EV subclassification
	different notations for δ are introduced in the following:	Figure A3.1, A3.2
c	subscript c used to denote treatment effect in subclass c	
^	hat is used to denote the treatment effect estimate	
reg.	subscript reg. is used to indicate the regression application	
$\hat{V}_{reg.c}$	the estimated variance of $\hat{\delta}_{reg.c}$	

$\overline{V_{reg.}^{-1}}$	the average of $\hat{V}_{reg.c}$ for c = 1, …, 5	
k	for this section, subscript k is introduced to denote the control group, k = 0, or the treatment group, k = 1; for k = 0, i = 1, …, n_0, for k = 1, i = 1, …, n_1	§ 3.4 Estimation of δ
–	bar is used to denote the mean (of the outcome)	
$\hat{V}_c$	the estimated variance of $\hat{\delta}_c$	
w_c	subclass-specific weights within subclass c, $\sum_{c=1}^{5} w_c = 1$; $w_c = w_c^{EW}$ for using equal weights, $w_c = w_c^{IV}$ for using inverse variance weights	
$\hat{\delta}^{(j)}$	$\hat{\delta}$ of the j^{th} simulated sample, for j = 1, 2, …, 1000, $\underset{\sim}{\hat{\delta}} = (\hat{\delta}^{(1)}, \hat{\delta}^{(2)}, ..., \hat{\delta}^{(1000)})$	

Figure A3.1 illustrates how each step in the simulation procedure flows in the diagram, Figure A3.2, A3.3 and Table A3.4 provide extra information for Figure A3.1.

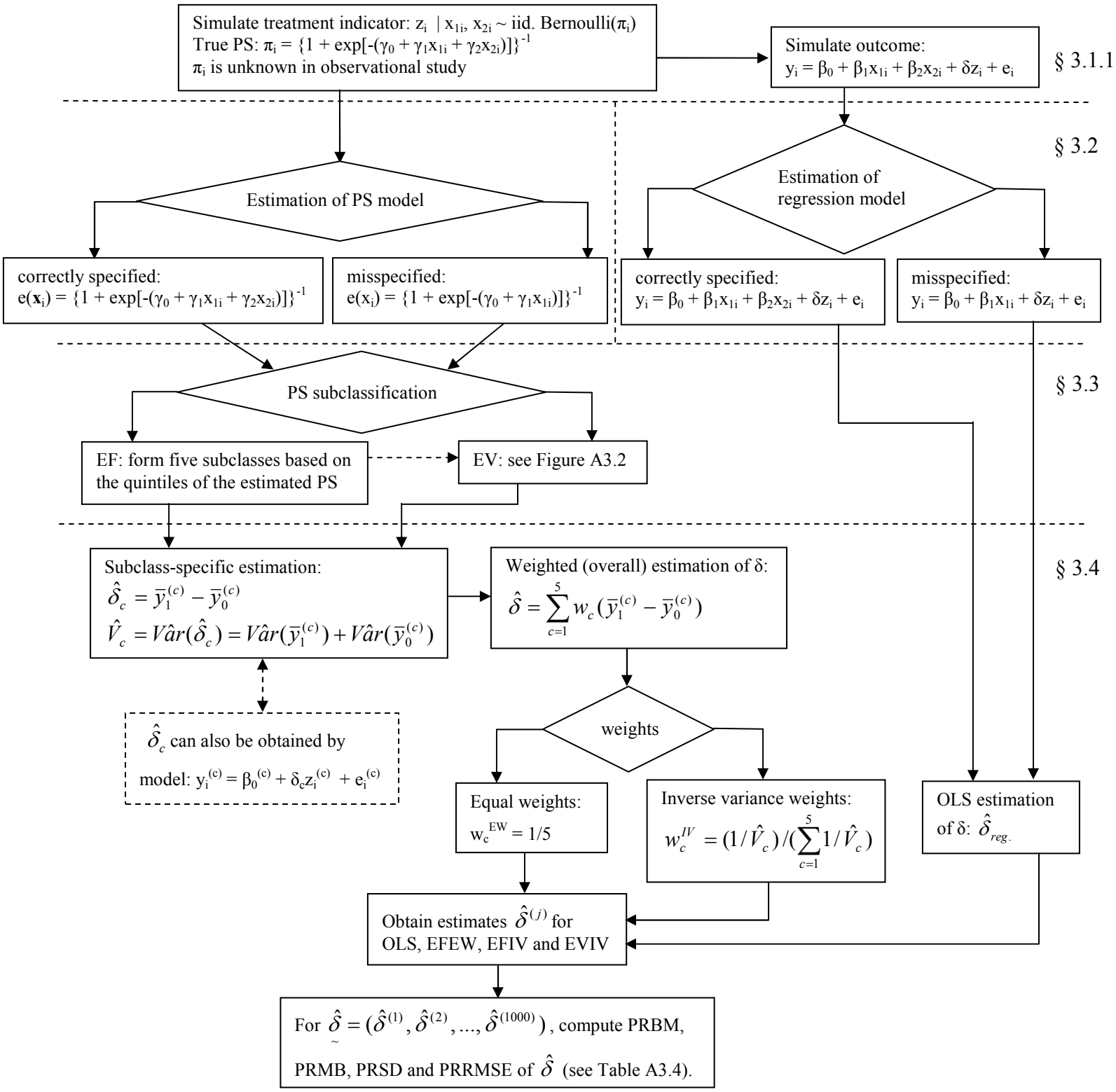

Figure A3. 1 Diagram under scenario involving two independent covariates, (x_1, x_2)

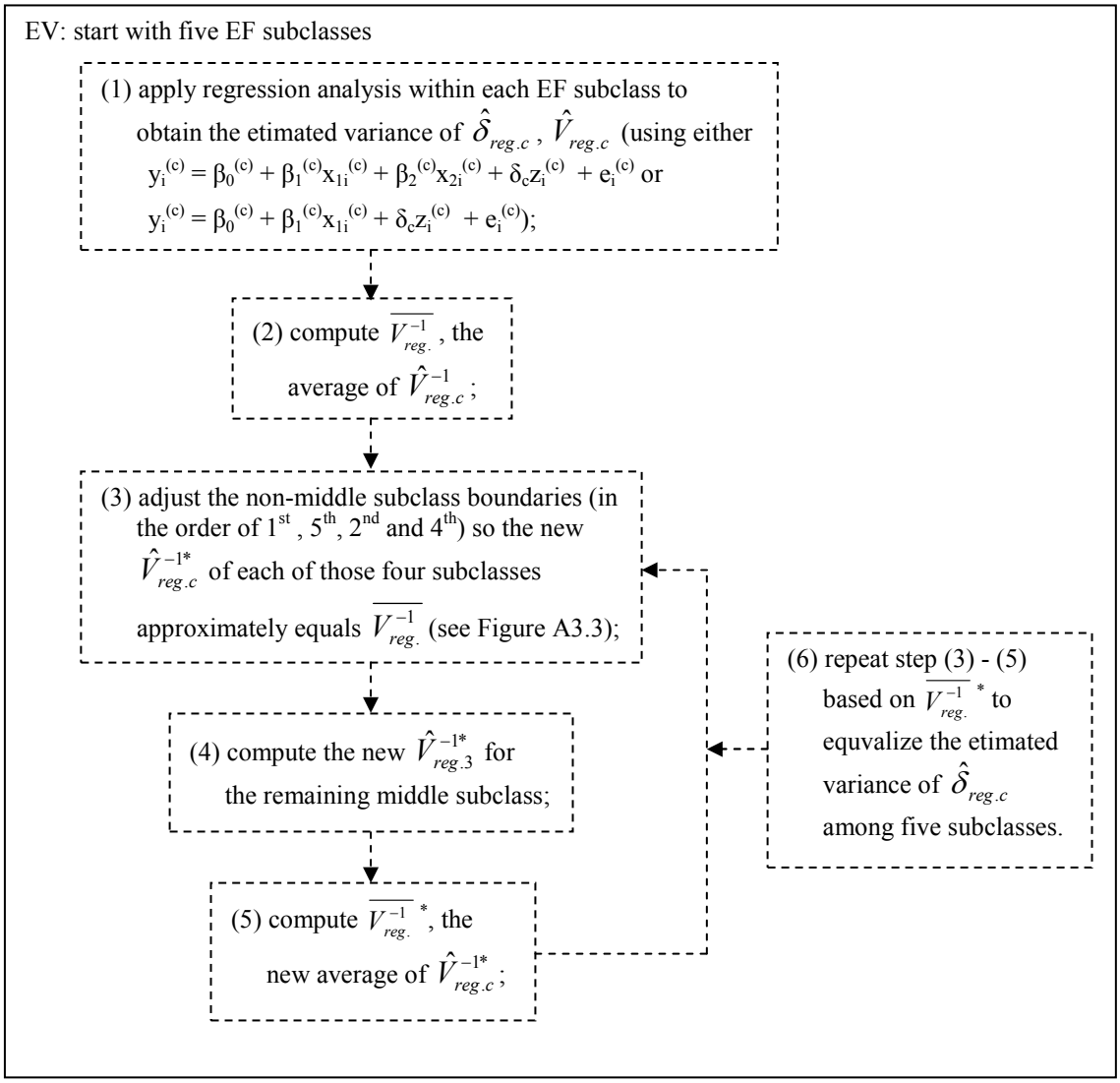

Figure A3. 2 Diagram of EV subclassification under Scenario involving (x_1, x_2)

Note for Figure A3.2: A predetermined number of iterations can be set to achieve approximately equal inverse variances among subclasses. Since $\hat{V}_{reg.c}^{-1*}$ may not be the same among subclasses, one can also set up a tolerance level to determine whether the differences among $\hat{V}_{reg.c}^{-1*}$ are acceptable as approximately equal.

Note for Figure A3.1 and Figure A3.2: for the diagram under the scenario involving a single covariate and its squared term, (x, x^2), in Figure A3.1, x_{1i} will be replaced by x_i, and x_{2i} will be replaced by x_i^2; the correctly specified regression model and propensity scores model will be excluded following Drake's (1993) omission. In Figure A3.2, $x_{1i}^{(c)}$ will be replaced by $x_i^{(c)}$, and $x_{2i}^{(c)}$ will be replaced by $(x_i^{(c)})^2$.

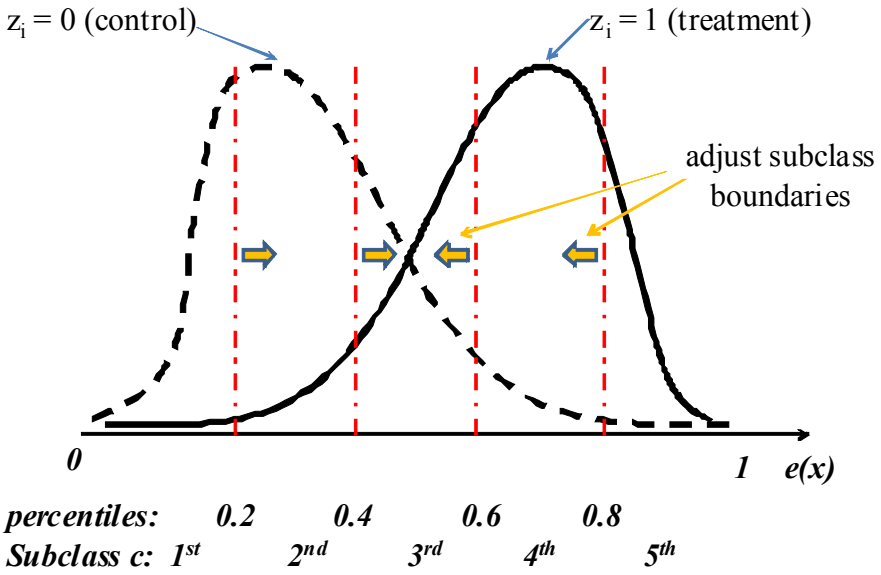

Figure A3. 3 Adjusting boundaries in the order of 1st, 5th, 2nd and 4th subclass

Table A3. 3 Percentage relative measure formulas of $\hat{\delta}$

Measure	Formula
$\bar{\delta}$	mean($\underset{\sim}{\hat{\delta}}$)
$\tilde{\delta}$	median($\underset{\sim}{\hat{\delta}}$)
PRBM	$100\times[(\tilde{\delta} - \delta)/\delta]\%$
PRMB	$100\times[(\bar{\delta} - \delta)/\delta]\%$
PRSD	$100\times[SD(\underset{\sim}{\hat{\delta}})/\delta]\%$
PRRMSE	$\{[PRMB(\hat{\delta})]^2 + [PRSD(\hat{\delta})]^2\}^{1/2}$

Table A3.5 displays the true parameter values.

Table A3. 4 Values of parameters

δ	γ_0	γ_1	γ_2	β_0	β_1	β_2
1	0	0.4	(0.4, 0.7, 1.1)	1	1	(1, 2, 3)
3	0	0.4	(0.4, 0.7, 1.1)	1	1	(1, 2, 3)

A3.3.2 Example of EV only produces less than five subclasses

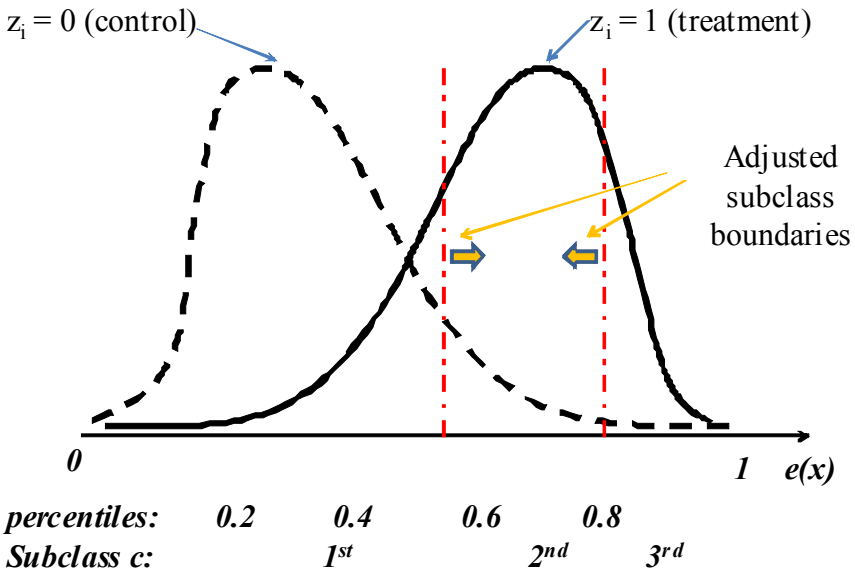

Figure A3. 4 EV subclassification procedure does not produce a result for five subclasses

A3.5.1 Correctly specified model involving covariates (x_1, x_2)

Table A3.5.1. 1 PRBM for the OLS and three propensity score estimators using correctly specified propensity score models and correctly specified regression models

Parameters			PRBM of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	0.2584	7.9828	6.5922	7.2794
		0.7	-0.3078	10.0068	7.9962	9.3018
		1.1	-0.0657	12.6586	9.7177	12.2720
	2	0.4	0.1384	12.1111	9.3967	9.9853
		0.7	0.1951	16.5513	12.3161	13.7956
		1.1	0.2007	21.3605	14.9981	17.9560
	3	0.4	-0.3334	15.1411	11.7769	13.1588
		0.7	-0.1133	22.9553	15.6461	16.9937
		1.1	0.2547	31.4995	19.9846	22.1701
3	1	0.4	0.0655	2.6509	2.2054	2.3596
		0.7	0.0674	3.3501	2.7700	3.1725
		1.1	0.1155	4.3770	3.3814	4.1709
	2	0.4	0.0857	3.8933	3.2217	3.4362
		0.7	0.0483	5.6407	4.0954	4.4872
		1.1	0.0745	7.5077	5.2929	6.2697
	3	0.4	-0.0174	5.1354	4.2482	4.5834
		0.7	-0.0098	7.6276	5.2931	5.5716
		1.1	0.0769	10.3472	6.5117	7.4084

The results in Table A3.5.1.1 indicate that when using correctly specified propensity scores, the PRBM of all propensity score estimators (EF-EW, EF-IV and EV-

IV) increase as γ_2 (i.e. the influence of x_2 on the propensity scores) increases. This pattern is consistent for each level of β_2 (i.e. the influence of x_2 on the outcome). All propensity score estimators produce positive biases. The PRBM of all propensity score estimators at $\delta=3$ are smaller than they are at $\delta=1$. This is because the PRBM is computed by $100\times[(\tilde{\delta} - \delta)/\delta]\%$, so at $\delta=3$, the denominator increases by a factor of three as compared to when $\delta=1$. Among propensity score estimators, EF-IV has a lower PRBM than the other two estimators, and EV-IV has a lower PRBM than EF-EW. OLS has the lowest PRBM of all estimators in this investigation, which is as we expect, since the estimating model is correctly specified.

Table A3.5.1. 2 PRSD for the OLS and three propensity score estimators using correctly specified propensity score models and correctly specified regression models

Parameters			PRSD of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	6.79	7.34	7.26	7.31
		0.7	6.90	7.49	7.38	7.43
		1.1	7.14	8.09	7.65	7.84
	2	0.4	6.79	8.09	7.87	7.89
		0.7	6.66	7.91	7.73	7.70
		1.1	7.25	8.78	8.12	8.52
	3	0.4	6.54	8.93	8.49	8.83
		0.7	6.98	9.68	9.01	9.44
		1.1	7.34	10.07	8.95	9.23
3	1	0.4	2.27	2.48	2.43	2.41
		0.7	2.23	2.41	2.40	2.44
		1.1	2.45	2.70	2.59	2.63
	2	0.4	2.19	2.73	2.59	2.61
		0.7	2.27	2.82	2.67	2.68
		1.1	2.42	2.99	2.79	2.88
	3	0.4	2.18	2.87	2.75	2.83
		0.7	2.21	3.17	2.88	2.96
		1.1	2.25	3.21	2.78	2.94

The results in Table A3.5.1.2 indicate that when using correctly specified propensity scores, the PRSD of all propensity score estimators increase as γ_2 increases at $\delta=1$, $\beta_2=1$. At $\delta=1$, $\beta_2=2$, the PRSD of all propensity score estimators slightly decrease as γ_2 increases from 0.4 to 0.7, and they increase as γ_2 increases from 0.7 to 1.1. At $\delta=1$,

β_2=3, the PRSD of all propensity score estimators increase as γ_2 increases from 0.4 to 0.7, and they slightly decrease as γ_2 increases from 0.7 to 1.1. At δ=3, the PRSD of all propensity score estimators do not show the similar pattern when δ=1. These results also reveal that as β_2 increases, the PRSD of all propensity score estimators increase. The PRSD of all propensity score estimators at δ=3 are smaller than they are at δ=1. This is because the PRSD is computed by $100\times[SD(\underset{\sim}{\hat{\delta}})/\delta]\%$, so at δ=3, the denominator increases by a factor of three as compared to when δ=1. Among propensity score estimators, in 16 out of 18 parameter combinations, EF-IV has the lowest PRSD and EV-IV has a lower PRSD than EF-EW. OLS has the lowest PRSD of all estimators in this simulation, which is as we expect.

Table A3.5.1. 3 PRRMSE (percentage relative RMSE) for the OLS and three propensity score estimators using correctly specified propensity score models and correctly specified regression models

Parameters			PRRMSE of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	6.7903	10.8127	9.8156	10.2883
		0.7	6.8961	12.4953	10.9645	11.9653
		1.1	7.1455	14.7873	12.1948	14.4460
	2	0.4	6.7943	14.2792	12.2970	12.8363
		0.7	6.6647	18.5281	14.4931	15.7898
		1.1	7.2576	23.4215	17.3140	20.1358
	3	0.4	6.5398	17.6182	14.7415	15.6954
		0.7	6.9752	24.9592	18.1341	19.7593
		1.1	7.3461	32.7267	21.5123	24.1578
3	1	0.4	2.2760	3.6169	3.2764	3.4063
		0.7	2.2261	4.1749	3.6690	4.0212
		1.1	2.4509	5.0969	4.2513	4.9833
	2	0.4	2.1861	4.7090	4.0261	4.2088
		0.7	2.2732	6.1935	4.8391	5.2158
		1.1	2.4217	7.9695	5.9526	6.8997
	3	0.4	2.1795	5.9609	5.0528	5.4015
		0.7	2.2117	8.2908	6.0022	6.4622
		1.1	2.2549	10.8837	7.1386	7.9919

The results in Table A3.5.1.3 indicate that when using correctly specified propensity scores, the PRRMSE of all propensity score estimators increases as γ_2 increases. We observe the same pattern for all propensity score estimators at each level of β_2. At δ=1, β_2=1, 3 and at δ=3, β_2=2, 3, the PRRMSE of the correctly specified OLS

estimator increases slightly as γ_2 increases. Table A3.5.1.3 also shows that as β_2 increases, the PRRMSE of all propensity score estimators increase. The PRRMSE of all estimators at δ=3 are smaller than they are at δ=1, as obtained by $\{[\text{PRMB}(\hat{\delta})]^2 + [\text{PRSD}(\hat{\delta})]^2\}^{1/2}$. Among propensity score estimators, EF-IV has the lowest PRRMSE, and EV-IV has a lower PRRMSE than EF-EW. OLS has the lowest PRRMSE of all estimators in this study which is as we expect.

A3.5.2 Misspecified model with a covariate omission (excluding x_2)

Table A3.5.2. 1 PRBM for the OLS and three propensity score estimators using misspecified propensity score models and misspecified regression models

Parameters			PRBM of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	38.8467	42.4454	42.4563	42.6973
		0.7	63.6339	66.8608	66.6900	66.9474
		1.1	88.2792	91.2719	90.9991	90.9619
	2	0.4	77.3830	80.9922	80.7373	80.9318
		0.7	126.6983	130.2715	130.2548	130.4198
		1.1	176.1955	178.8781	178.8499	178.9682
	3	0.4	116.1411	119.8310	119.8196	119.8072
		0.7	190.6813	193.9786	194.2208	193.7174
		1.1	266.0288	268.3200	268.0300	268.5865
3	1	0.4	12.9568	14.1038	14.0620	14.0938
		0.7	21.1244	22.2965	22.1812	22.2519
		1.1	29.7682	30.8179	30.8117	30.7692
	2	0.4	25.9439	27.0475	27.0711	27.0987
		0.7	42.0446	43.2246	43.0975	43.1800
		1.1	58.9364	59.8422	59.8501	59.9135
	3	0.4	38.9153	40.0536	40.0188	40.1469
		0.7	62.9571	64.0750	64.0156	64.0850
		1.1	88.5816	89.3763	89.4871	89.5133

The results in Table A3.5.2.1 indicate that for misspecified propensity scores with a covariate omission, the PRBM of all propensity score estimators increase dramatically as γ_2 increases. This pattern is consistent with all propensity score estimators at each

level of β_2. The misspecified OLS estimator that omits covariate x_2 follows the same pattern. All estimators produce positive biases. As γ_2 and β_2 increase, the PRBM of all estimators increase substantially.

The PRBM of all estimators at $\delta=3$ are smaller than they are at $\delta=1$. This is because the PRBM is computed by $100\times[(\tilde{\delta} - \delta)/\delta]\%$, so at $\delta=3$, the denominator increases by a factor of three as compared to when $\delta=1$. Overall, none of the estimators perform well in terms of PRBM when an independent covariate is omitted from the propensity score models and from the regression models.

Table A3.5.2. 2 PRSD for the OLS and three propensity score estimators using misspecified propensity score models and misspecified regression models

Parameters			PRSD of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	8.98	9.18	9.19	9.23
		0.7	9.05	9.26	9.28	9.34
		1.1	8.65	8.84	8.86	8.88
	2	0.4	14.15	14.13	14.20	14.24
		0.7	13.33	13.49	13.57	13.55
		1.1	13.39	13.45	13.51	13.50
	3	0.4	20.12	20.22	20.35	20.36
		0.7	19.88	19.93	20.08	20.11
		1.1	18.52	18.73	18.66	18.66
3	1	0.4	3.05	3.12	3.134	3.11
		0.7	2.85	2.89	2.89	2.92
		1.1	2.89	2.95	2.96	2.97
	2	0.4	4.62	4.71	4.73	4.73
		0.7	4.53	4.52	4.58	4.57
		1.1	4.31	4.35	4.36	4.36
	3	0.4	6.77	6.80	6.80	6.79
		0.7	6.19	6.24	6.26	6.31
		1.1	6.10	6.13	6.16	6.15

The results in Table A3.5.2.2 indicate that when using misspecified propensity scores with a covariate omission, at δ=1, 3 and β_2=2, 3, the PRSD of all propensity score estimators decrease slightly as γ_2 increases. At δ=1, β_2=1, the PRSD of all propensity score estimators increase slightly as γ_2 increases from 0.4 to 0.7, and they decrease as γ_2 increases from 0.7 to 1.1. At δ=3, β_2=1, the PRSD of all propensity score estimators

decrease slightly as γ_2 increases from 0.4 to 0.7, then they increase as γ_2 increases from 0.7 to 1.1. The misspecified OLS estimator shows a similar pattern except at $\delta=1$, $\beta_2=2$. These results also show that as β_2 increases, the PRSD of all estimators increase. The PRSD of all estimators at $\delta=3$ are smaller than they are at $\delta=1$. This is because the PRSD is computed by $100\times[SD(\underset{\sim}{\hat{\delta}})/\delta]\%$, so at $\delta=3$, the denominator increases three times as compared to $\delta=1$. Overall, none of the estimators performs well in terms of PRSD when an independent covariate is omitted from the propensity score model and from the regression model.

Table A3.5.2. 3 PRRMSE for the OLS and three propensity score estimators using misspecified propensity score models and misspecified regression models

Parameters			PRRMSE of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	39.8482	43.5270	43.2714	43.5305
		0.7	63.8598	67.3843	67.1912	67.3387
		1.1	88.7667	91.7266	91.4893	91.6179
	2	0.4	78.6804	82.2270	82.0772	82.2708
		0.7	127.3754	130.8780	130.7482	130.8603
		1.1	176.9333	179.7323	179.6653	179.7998
	3	0.4	117.6982	121.0903	121.1026	121.2980
		0.7	190.7942	193.7795	193.7847	193.9205
		1.1	266.3356	268.9625	268.8701	269.0876
3	1	0.4	13.2953	14.5263	14.4491	14.5167
		0.7	21.3121	22.4762	22.4002	22.4646
		1.1	29.8063	30.8157	30.7541	30.7895
	2	0.4	26.2576	27.4124	27.3979	27.4574
		0.7	42.3822	43.4465	43.3895	43.4451
		1.1	59.1745	60.1203	60.0728	60.1136
	3	0.4	39.3786	40.5684	40.5059	40.5480
		0.7	63.3176	64.3530	64.3128	64.3619
		1.1	88.7543	89.6156	89.5828	89.6089

The results in Table A3.5.2.3 indicate that when using misspecified propensity scores with a covariate omission, the PRRMSE of all propensity score estimators increases drastically as γ_2 increases. This pattern is consistent with all propensity score estimators at each level of β_2. The misspecified OLS estimator with a covariate omission

follows the same pattern. As γ_2 and β_2 increase, the PRRMSE of all estimators increase substantially.

The PRRMSE of all estimators at δ=3 are smaller than they are at δ=1, as obtained by $\{[\text{PRMB}(\hat{\delta})]^2 + [\text{PRSD}(\hat{\delta})]^2\}^{1/2}$. Overall, none of the estimators perform well in terms of PRRMSE when an independent covariate is omitted from the propensity score models and from the regression models.

A3.5.3 Model with a quadratic term misspecification (omitting x^2)

Table A3.5.3. 1 PRBM for the OLS and three propensity score estimators using misspecified propensity scores model and misspecified regression models

Parameters			PRBM of $\hat{\delta}$				Proportion
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV	EV-N/A
1	1	0.4	55.0540	30.9273	13.5805	16.7561	
		0.7	76.3387	39.0552	18.1776	24.6651	
		1.1	89.8479	44.4803	19.7065	31.2844	
	2	0.4	111.5960	59.9801	12.5918	18.1536	
		0.7	151.4055	76.6458	19.0281	29.5986	
		1.1	179.8477	88.6031	24.1015	42.9639	
	3	0.4	166.4412	89.1653	11.8878	20.7355	0.177
		0.7	228.8786	115.7972	18.2056	34.6258	0.136
		1.1	268.6373	132.2227	23.5426	51.0012	0.064
3	1	0.4	18.4694	10.3999	4.5970	5.7084	
		0.7	25.4915	12.8725	6.1505	8.2263	
		1.1	29.8830	14.9346	6.6572	10.6868	
	2	0.4	36.5291	19.8024	4.3274	6.2102	
		0.7	51.1085	25.9451	6.3914	10.0749	
		1.1	59.6798	29.5189	8.1019	14.1791	
	3	0.4	55.4942	29.6058	3.7125	6.9259	0.189
		0.7	76.7960	38.6405	5.8580	11.6378	0.119
		1.1	89.5648	43.8916	7.7963	17.1402	0.062

Note: EV-N/A indicates that the proportion of the EV subclassification approach does not produce five subclasses out of 1000 random samples (Section 3.3). The results of EV-N/A are the same for other tables in this subsection.

The results in Table A3.5.3.1 indicate that under quadratic term misspecification, the PRBM of all propensity score estimators increase as γ_2 increases. This pattern is consistent with all propensity score estimators at each level of β_2. The misspecified OLS estimator, which omits a quadratic term, follows the same pattern. The EF-IV has the lowest PRBM of all estimators in this simulation, while the OLS estimator has the highest PRBM. The results show that as β_2 increases, the PRBM of the OLS, EF-EW and EV-IV estimators increase. However, the results obtained from the EF-IV estimator do not reflect that pattern. The PRBM of all estimators at $\delta=3$ are smaller than they are at $\delta=1$. This is because the PRBM is computed by $100\times[(\tilde{\delta} - \delta)/\delta]\%$, so at $\delta=3$, the denominator increases by a factor of three as compared to when $\delta=1$.

Table A3.5.3. 2 PRSD for the OLS and three propensity score estimators using misspecified propensity score models and misspecified regression models

Parameters			PRSD of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	10.29	8.81	7.56	7.89
		0.7	10.52	9.07	7.76	8.22
		1.1	10.14	9.50	8.22	8.48
	2	0.4	17.79	13.28	8.46	9.11
		0.7	16.07	12.08	8.58	9.82
		1.1	16.07	12.67	8.59	9.88
	3	0.4	26.05	19.13	8.80	10.97
		0.7	24.08	18.12	9.26	11.88
		1.1	22.29	17.22	9.27	12.67
3	1	0.4	3.54	3.08	2.60	2.74
		0.7	3.37	2.91	2.53	2.69
		1.1	3.31	3.29	2.77	2.82
	2	0.4	5.82	4.65	2.81	3.13
		0.7	5.61	4.28	2.88	3.23
		1.1	5.31	4.29	2.97	3.36
	3	0.4	8.71	6.29	2.99	3.68
		0.7	8.10	5.905	3.09	3.96
		1.1	7.37	5.83	3.16	4.20

The results in Table A3.5.3.2 indicate that under quadratic term misspecification, at each level of β_2, there is no clear pattern for all estimators as γ_2 increases. These results also show that as β_2 increases, the PRSD of all estimators increase. The PRSD of all estimators at δ=3 are smaller than they are at δ=1. This is because the PRSD is computed by $100\times[SD(\hat{\delta})/\delta]\%$, so at δ=3, the denominator increases by a factor of three

as compared to when δ=1. Overall, the OLS estimator has the highest percentage relative SD, and EF-IV has the lowest PRSD of all estimators.

Table A3.5.3. 3 PRRMSE for the OLS and three propensity score estimators using misspecified propensity score models and misspecified regression models

Parameters			PRRMSE of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	56.0204	32.1665	15.5849	18.3650
		0.7	76.9019	40.1091	19.5790	26.0662
		1.1	90.3418	45.4292	21.2294	32.5615
	2	0.4	112.3836	61.0151	15.2208	20.3182
		0.7	152.4889	77.5881	20.6333	31.1008
		1.1	180.6401	89.6126	25.4063	43.9807
	3	0.4	168.7110	91.3191	14.7846	23.7789
		0.7	231.2271	117.2974	20.4009	36.9556
		1.1	270.2290	132.6870	25.5963	52.9490
3	1	0.4	18.8745	10.8929	5.3348	6.2569
		0.7	25.6272	13.3113	6.6116	8.6937
		1.1	30.0473	15.2473	7.2418	10.9874
	2	0.4	37.2443	20.2725	5.1292	6.9690
		0.7	51.3175	26.2450	6.9663	10.5315
		1.1	60.0477	29.8170	8.5925	14.6687
	3	0.4	56.3381	30.2764	4.8511	7.9727
		0.7	77.2127	39.1856	6.6446	12.4333
		1.1	89.9871	44.3666	8.4588	17.6084

The results in Table A3.5.3.3 indicate that under quadratic term misspecification, the PRRMSE of all propensity score estimators increases as γ_2 increases. We observe the same pattern for all propensity score estimators at each level of β_2. The misspecified

OLS estimator follows a similar pattern. These results show that as β_2 increases, the PRRMSE of the OLS, EF-EW and EV-IV estimator increases. However, a similar pattern has not been observed in the results obtained from the EF-IV estimator. The PRRMSE of all estimators at $\delta=3$ are smaller than they are at $\delta=1$, as obtained by $\{[\text{PRMB}(\hat{\delta})]^2 + [\text{PRSD}(\hat{\delta})]^2\}^{1/2}$. Overall, the OLS estimator has the highest PRRMSE, and EF-IV has the lowest PRRMSE of all estimators. We also provide the results for using true propensity scores and the correctly specified regression model in the following.

A3.5.4 Using true propensity scores

True propensity scores are generated under two scenarios. In one scenario, true propensity scores are generated from two independent covariates. In another scenario, true propensity scores are generated from one covariate and its square. Ideally, the true propensity scores provide the "best" possible situation with respect to propensity scores estimation. We can see the effect of estimating the propensity scores from Section 3.5.1 and A3.5.1, which shows that results using estimated propensity scores are very similar to results using true propensity scores.

A3.5.4.1 Two independent covariates (x_1, x_2)

Under this scenario, the results of the OLS estimator are the same as they are in Section 3.5.1.

Table A3.5.4.1. 1 PRBM for the OLS and three propensity score estimators using true propensity scores and correctly specified regression models for (x_1, x_2)

Parameters			PRBM of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	0.2584	7.9935	6.4839	7.1960
		0.7	-0.3078	10.2172	8.0307	9.3445
		1.1	-0.0657	12.4214	9.5089	12.0687
	2	0.4	0.1384	11.9425	9.4926	10.0540
		0.7	0.1951	16.5535	12.1748	13.8818
		1.1	0.2007	21.6156	15.2071	18.0531
	3	0.4	-0.3334	15.1627	12.1306	12.9605
		0.7	-0.1133	22.7805	15.2906	17.2285
		1.1	0.2547	31.1220	19.8846	22.0174
3	1	0.4	0.0655	2.6894	2.3015	2.3691
		0.7	0.0674	3.3819	2.8045	3.2839
		1.1	0.1155	4.3855	3.4471	4.1359
	2	0.4	0.0857	3.9068	3.0665	3.3242
		0.7	0.0483	5.5606	4.1123	4.4783
		1.1	0.0745	7.4166	5.2643	6.2034
	3	0.4	-0.0174	5.1016	4.3000	4.6936
		0.7	-0.0098	7.7401	5.2908	5.6467
		1.1	0.0769	10.4917	6.5388	7.2311

The results in Table A3.5.4.1.1 indicate that when using true propensity scores, the PRBM of all propensity score estimators increase as γ_2 (i.e. the influence of x_2 on the propensity scores) increases. This pattern is consistent with all propensity score estimators (EF-EW, EF-IV and EV-IV) at each level of β_2 (i.e. the influence of x_2 on the outcome). All propensity score estimators produce positive biases. These results also

show that as β_2 increases, the PRBM of all propensity score estimators increase. The PRBM of all propensity scores estimators at $\delta=3$ are smaller than they are at $\delta=1$. This is because the PRBM is computed by $100\times[(\tilde{\delta} - \delta)/\delta]\%$, so at $\delta=3$, the denominator increases by a factor of three as compared to when $\delta=1$. Among propensity score estimators, EF-IV has the lowest PRBM, while EV-IV has a lower PRBM than EF-EW. OLS has the lowest PRBM of all estimators in this investigation, which is as we expect, since the estimating model is correctly specified.

Table A3.5.4.1. 2 PRMB for the OLS and three propensity score estimators using true propensity scores and correctly specified regression models

Parameters			PRMB of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	0.1461	7.9108	6.5464	7.1295
		0.7	-0.0872	10.0884	8.1664	9.4786
		1.1	-0.2448	12.2969	9.4200	12.0602
	2	0.4	0.2306	11.8418	9.5693	10.2478
		0.7	0.3096	16.8163	12.2690	13.6785
		1.1	0.2194	21.7645	15.3156	18.0929
	3	0.4	-0.0923	15.2042	12.1506	13.1490
		0.7	-0.0476	22.8755	15.6814	17.2711
		1.1	0.1624	31.4081	19.6356	21.9471
3	1	0.4	0.0887	2.6495	2.2170	2.3832
		0.7	0.0633	3.4684	2.8501	3.2342
		1.1	0.0989	4.2983	3.3640	4.2618
	2	0.4	0.0028	3.8697	3.1307	3.3252
		0.7	0.0386	5.5211	4.0124	4.4209
		1.1	0.1338	7.3758	5.2576	6.2471
	3	0.4	0.0682	5.3333	4.3708	4.7575
		0.7	0.0190	7.6596	5.2598	5.6940
		1.1	0.0108	10.4015	6.5322	7.2886

The results in Table A3.5.4.1.2 indicate that when using true propensity scores, the PRMB of all propensity score estimators increase as γ_2 increases. All propensity score estimators show the same pattern at each level of β_2. These results reveal that as β_2 increases, the PRMB of all propensity score estimators increase. The PRMB of all propensity scores estimators at δ=3 are smaller than they are at δ=1. This is because the

PRMB is computed by $100\times[(\bar{\delta} - \delta)/\delta]\%$, so at $\delta=3$, the denominator increases by a factor of three as compared to when $\delta=1$. Among propensity score estimators, EF-IV has the lowest PRMB, while EV-IV has a lower PRMB than EF-EW. OLS has the lowest PRMB, which is as we expect.

Table A3.5.4.1. 3 PRSD for the OLS and three propensity score estimators using true propensity scores and correctly specified regression models

Parameters			PRSD of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	6.79	7.30	7.17	7.31
		0.7	6.90	7.79	7.72	7.73
		1.1	7.14	8.72	8.38	8.63
	2	0.4	6.79	9.28	8.98	9.13
		0.7	6.66	7.90	7.73	7.80
		1.1	7.25	9.05	8.40	8.71
	3	0.4	6.54	12.86	12.64	12.79
		0.7	6.98	10.65	9.93	10.22
		1.1	7.34	10.05	8.98	9.41
3	1	0.4	2.27	2.48	2.44	2.44
		0.7	2.23	2.57	2.55	2.59
		1.1	2.45	3.02	2.93	2.98
	2	0.4	2.19	3.10	2.98	3.04
		0.7	2.27	2.80	2.66	2.70
		1.1	2.42	3.04	2.86	2.93
	3	0.4	2.18	4.30	4.22	4.33
		0.7	2.21	3.41	3.18	3.26
		1.1	2.25	3.30	2.83	2.97

The results in Table A3.5.4.1.3 indicate that when using true propensity scores, the PRSD of all propensity score estimators increase as γ_2 increases at β_2=1, but decrease at β_2=3. At β_2=2, the PRSD of all propensity score estimators decrease as γ_2 increases from 0.4 to 0.7, then they slightly increase as γ_2 increases from 0.7 to 1.1. At δ=1, β_2=1, 3 and at δ=3, β_2=2, 3, the PRSD of the correctly specified OLS estimator increases slightly as γ_2 increases. These results reveal that as β_2 increases, the PRSD of all propensity score estimators increase. The PRSD of all propensity score estimators at δ=3 are smaller than they are at δ=1. This is because the PRSD is computed by $100\times[SD(\underset{\sim}{\hat{\delta}})/\delta]\%$, so at δ=3, the denominator increases by a factor of three times as compared to when δ=1. Among propensity score estimators, EF-IV has the lowest PRSD, while EV-IV has a lower PRSD than EF-EW. OLS has the lowest PRSD of all estimators in this simulation, which is as we expect.

Table A3.5.4.1. 4 PRRMSE for the OLS and three propensity score estimators using true propensity scores and correctly specified regression models

Parameters			PRRMSE of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	6.7903	10.7622	9.7107	10.2124
		0.7	6.8961	12.7486	11.2400	12.2335
		1.1	7.1455	15.0774	12.6085	14.8285
	2	0.4	6.7943	15.0478	13.1252	13.7245
		0.7	6.6647	18.5788	14.4984	15.7479
		1.1	7.2576	23.5720	17.4692	20.0783
	3	0.4	6.5398	19.9143	17.5324	18.3457
		0.7	6.9752	25.2336	18.5597	20.0677
		1.1	7.3461	32.9765	21.5902	23.8779
3	1	0.4	2.2760	3.6288	3.2931	3.4125
		0.7	2.2261	4.3184	3.8237	4.1457
		1.1	2.4509	5.2521	4.4641	5.2019
	2	0.4	2.1861	4.9587	4.3205	4.5055
		0.7	2.2732	6.1898	4.8118	5.1779
		1.1	2.4217	7.9789	5.9833	6.8993
	3	0.4	2.1795	6.8504	6.0781	6.4325
		0.7	2.2117	8.3847	6.1452	6.5608
		1.1	2.2549	10.9117	7.1192	7.8707

The results in Table A3.5.4.1.4 indicate that when using true propensity scores, the PRRMSE of all propensity score estimators increase as γ_2 increases. We observe the same pattern for all propensity score estimators at each level of β_2. At δ=1, β_2=1, 3 and at δ=3, β_2=2, 3, the PRRMSE of the correctly specified OLS estimator increases slightly as γ_2 increases. These results show that as β_2 increases, the PRRMSE of all propensity score

estimators increase. The PRRMSE of all estimators at δ=3 are smaller than they are at δ=1, as obtained by $\{[\text{PRMB}(\hat{\delta})]^2 + [\text{PRSD}(\hat{\delta})]^2\}^{1/2}$. Among propensity score estimators, EF-IV has the lowest PRRMSE, while EV-IV has a lower PRRMSE than EF-EW. OLS has the lowest PRRMSE of all estimators in this simulation, which is as we expect.

A3.5.4.2 A single covariate and its quadratic term (x, x^2)

Table A3.5.4.2. 1 PRBM for the OLS and three propensity score estimators using true propensity scores and correctly specified regression models for (x, x^2)

Parameters			PRBM of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	-0.1139	21.3826	6.6092	8.4888
		0.7	-0.3982	24.8033	6.7840	11.3165
		1.1	-0.4850	26.6910	5.2646	13.2725
	2	0.4	0.1686	36.8611	6.3982	7.7744
		0.7	0.3043	43.5561	7.8635	10.9106
		1.1	0.2857	50.8252	9.0525	16.3767
	3	0.4	0.1439	52.0723	6.0684	7.3383
		0.7	0.2464	64.5238	7.6980	10.2169
		1.1	-0.3978	73.6806	10.0725	14.1167
3	1	0.4	0.0230	7.0168	2.1918	2.9273
		0.7	0.0072	8.2801	2.3460	3.9381
		1.1	-0.0429	9.2416	1.8194	4.4918
	2	0.4	0.0120	11.9193	2.0412	2.4615
		0.7	0.0210	15.1123	2.7147	3.8347
		1.1	-0.0063	17.0066	2.9754	5.4170
	3	0.4	-0.0943	17.3109	1.8690	2.3036
		0.7	-0.0669	21.5020	2.4994	3.3268
		1.1	-0.0049	24.6366	3.1991	4.7650

The results in Table A3.5.4.2.1 are very similar to the results in Table A3.5.4.1.1. However, at δ=1,3 and β_2=1, the PRBM of the EF-IV estimator slightly decreases as γ_2

(i.e. the influence of the quadratic term x^2 on the propensity scores) increases from 0.7 to 1.1.

Table A3.5.4.2. 2 PRMB for the OLS and three propensity score estimators using true propensity scores and correctly specified regression models

Parameters			PRMB of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	-0.1139	21.0598	6.6573	8.6520
		0.7	-0.3982	24.6755	6.6363	11.2855
		1.1	-0.4850	26.8094	5.1720	13.2672
	2	0.4	0.1686	36.3735	6.2855	7.7631
		0.7	0.3043	43.5656	7.8587	11.0499
		1.1	0.2857	50.6402	8.8744	15.9418
	3	0.4	0.1439	51.8913	6.0753	7.5356
		0.7	0.2464	64.1325	7.7809	10.1618
		1.1	-0.3978	73.2957	9.8899	14.2632
3	1	0.4	0.0230	7.1217	2.2970	2.9467
		0.7	0.0072	8.3032	2.3545	3.8807
		1.1	-0.0429	9.1431	1.8255	4.5805
	2	0.4	0.0120	11.9427	2.1193	2.6005
		0.7	0.0210	15.0204	2.7424	3.8269
		1.1	-0.0063	17.0273	3.0354	5.4318
	3	0.4	-0.0943	17.2241	1.9552	2.4051
		0.7	-0.0669	21.5834	2.5487	3.3352
		1.1	-0.0049	24.5617	3.2754	4.7408

The results in Table A3.5.4.2.2 are similar to the results in Table A3.5.4.1.2. However, at δ=1, 3 and β_2=1, the PRMB of the EF-IV estimator decreases as γ_2 increases from 0.7 to 1.1.

Table A3.5.4.2. 3 PRSD for the OLS and three propensity score estimators using true propensity scores and correctly specified regression models

Parameters			PRSD of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	6.64	8.60	7.19	7.30
		0.7	6.81	9.57	7.63	7.79
		1.1	7.10	12.19	8.11	8.39
	2	0.4	6.59	12.64	8.22	8.33
		0.7	6.71	11.18	7.79	8.04
		1.1	6.92	13.77	7.87	8.51
	3	0.4	6.44	18.19	10.05	10.24
		0.7	7.02	16.37	8.43	8.68
		1.1	7.19	17.28	8.43	8.75
3	1	0.4	2.22	2.82	2.42	2.45
		0.7	2.22	3.12	2.49	2.59
		1.1	2.45	4.24	2.72	2.76
	2	0.4	2.19	4.32	2.80	2.87
		0.7	2.26	3.83	2.50	2.66
		1.1	2.43	4.51	2.78	2.84
	3	0.4	2.25	6.02	3.43	3.43
		0.7	2.21	5.18	2.75	2.87
		1.1	2.37	5.60	2.74	2.85

The results in Table A3.5.4.2.3 indicate that when using true propensity scores, at δ=1, 3 and β_2=1, the PRSD of the propensity score estimators increase as γ_2 increases. At β_2=2, 3, the PRSD of propensity score estimators decrease as γ_2 increases from 0.4 to 0.7, then the majority of them slightly increase as γ_2 increases from 0.7 to 1.1. At each level of β_2, the PRSD of the correctly specified OLS estimator slightly increases as γ_2

increases, except at $\delta=3$ and $\beta_2=3$. These results show that as β_2 (i.e. the influence of the quadratic term x^2 on the outcome) increases, the PRSD of all propensity score estimators increase. The PRSD of all estimators at $\delta=3$ are smaller than they are at $\delta=1$. This is because the PRSD is computed by $100\times[SD(\underset{\sim}{\hat{\delta}})/\delta]\%$, so at $\delta=3$, the denominator increases by a factor of three as compared to when $\delta=1$. Among propensity score estimators, EF-IV has the lowest PRSD, while EV-IV has a lower PRSD than EF-EW. OLS has the lowest PRSD of all estimators in this simulation, which is as we expect.

Table A3.5.4.2. 4 PRRMSE for the OLS and three propensity score estimators using true propensity scores and correctly specified regression models

Parameters			PRRMSE of $\hat{\delta}$			
δ	β_2	γ_2	OLS	EF-EW	EF-IV	EV-IV
1	1	0.4	6.6404	22.7491	9.8006	11.3193
		0.7	6.8227	26.4651	10.1156	13.7121
		1.1	7.1087	29.4489	9.6179	15.6978
	2	0.4	6.5904	38.5077	10.3447	11.3865
		0.7	6.7148	44.9770	11.0647	13.6656
		1.1	6.9232	52.4787	11.8622	18.0704
	3	0.4	6.4361	54.9864	11.7476	12.7149
		0.7	7.0208	66.1875	11.4699	13.3634
		1.1	7.1924	75.3046	12.9947	16.7313
3	1	0.4	2.2167	7.6587	3.3383	3.8318
		0.7	2.2180	8.8715	3.4275	4.6652
		1.1	2.4521	10.0788	3.2796	5.3464
	2	0.4	2.1945	12.7001	3.5141	3.8708
		0.7	2.2591	15.4998	3.7111	4.6601
		1.1	2.4315	17.6135	4.1140	6.1315
	3	0.4	2.2485	18.2458	3.9440	4.1906
		0.7	2.2069	22.1964	3.7517	4.3996
		1.1	2.3694	25.1922	4.2730	5.5309

The results in Table A3.5.4.2.4 are very similar to the results in Table A3.5.4.1.4. However, at δ=1, 3 and β_2=1, the PRRMSE of the EF-IV estimator increases as γ_2 increases from 0.4 to 0.7, then it slightly decreases as γ_2 increases from 0.7 to 1.1. At δ=1, β_2=2, 3 and δ=3, β_2=3, the PRRMSE of the EF-IV estimator decreases as γ_2 increases from 0.4 to 0.7, then it slightly increases as γ_2 increases from 0.7 to 1.1. At each level of

β_2, the PRRMSE of the correctly specified OLS estimator increase slightly as γ_2 increases except at $\delta=3$, $\beta_2=3$. These results also show that as β_2 increases, PRRMSE of all propensity score estimators increase. The PRRMSE of all estimators at $\delta=3$ are smaller than they are at $\delta=1$, as obtained by $\{[\text{PRMB}(\hat{\delta})]^2 + [\text{PRSD}(\hat{\delta})]^2\}^{1/2}$. Among propensity score estimators, EF-IV has the lowest PRRMSE, while EV-IV has a lower PRRMSE than EF-EW. OLS has the lowest PRRMSE of all estimators in this simulation, which is as we expect.

A4.1 Notations, theory development diagram and Lemmas

Table A4.1 summarizes the notations used in this chapter.

Table A4. 1 Notations

Notation	Description	Section
(a_c, b_c)	lower bound and upper bound of subclass c, where c = 1, 2, …, C	§ 4.1
s_i^c	subclass indicator for subject i, where $s_i^c = 1$ if and only if $x_i \in (a_c, b_c)$ and $s_i^c = 0$ otherwise; by definition, s_i^c is a function of x_i such that, $s_i^c = s^c(x_i)$; let $\mathbf{s}_i = (s_i^1, \ldots, s_i^c, \ldots, s_i^C)_{1xC} \Rightarrow \mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_i, \ldots, s_n)_{1xnC}$; set $\mathbf{s}^c = (s_1^c, \ldots, s_i^c, \ldots, s_n^c)_{1xn}$; $n_1^{(c)} = \sum_{i=1}^{n} z_i s_i^c$, $n_0^{(c)} = \sum_{i=1}^{n} (1 - z_i) s_i^c$	
$\sigma^2_{x\|z,s^c}$	conditional variance of covariate x_i i = 1, … , n	
σ_e^2	variance of error term e_i, i = 1, … , n	
B_c	bias of $\hat{\delta}_c = \bar{y}_1^{(c)} - \bar{y}_0^{(c)}$, $E(\hat{\delta}_c - \delta)$	
V_c	variance of $\hat{\delta}_c$, $Var(\hat{\delta}_c)$	
$\underset{\sim}{w}$	vector of the weights, $\underset{\sim}{w} = (w_1, \ldots, w_c, \ldots, w_C)$, superscript ^{EW} is used to indicate equal weights, superscript ^{IV} is used to indicate inverse variance weights; superscript $^{EF\text{-}EW}$, $^{EF\text{-}IV}$ and $^{EV\text{-}IV}$ are used to denote different propensity scores estimators	§ 4.2, § 4.3

Figure A4.1 illustrates how theoretical development flows for each step in the diagram.

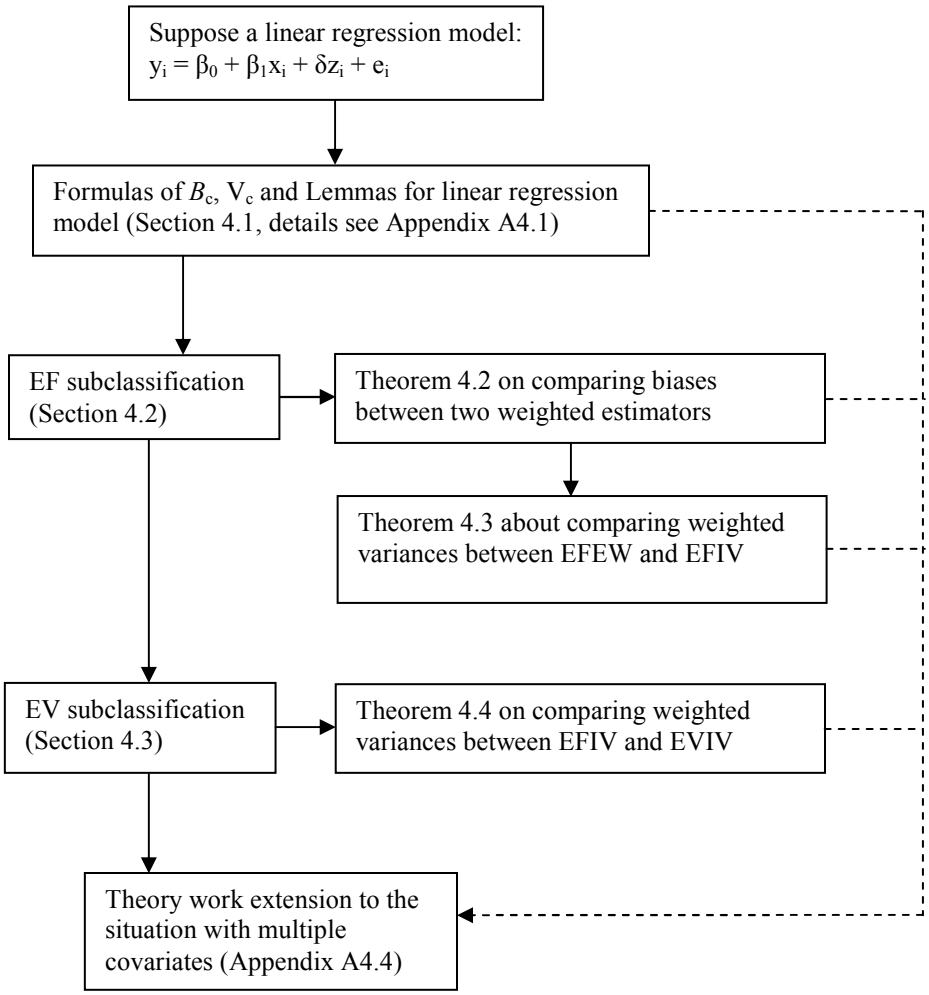

Figure A4. 1 Diagram of theory development

The following subsections provide a detailed development of B_c, V_c and Lemmas under a linear regression model.

A4.1.1 Preparing Lemmas to develop B_c and V_c under a linear regression model

In order to develop the expression of B_c and V_c, we provide Lemmas in the following.

Lemma A4.1 $u_1 \perp\!\!\!\perp u_2 \Rightarrow v_1=g_1(u_1) \perp\!\!\!\perp v_2=g_2(u_2) \mid w_1=h_1(u_1), w_2=h_2(u_2)$, where v_1, w_1 are functions of u_1 and v_2, w_2 are functions of u_2.

$$\text{Proof}: f(v_1, v_2 \mid w_1, w_2) = \frac{f(v_1, v_2, w_1, w_2)}{f(w_1, w_2)} = \frac{f(v_1, w_1) f(v_2, w_2)}{f(w_1) f(w_2)} = f(v_1 \mid w_1) f(v_2 \mid w_2)$$
$$= f(v_1 \mid w_1, w_2) f(v_2 \mid w_1, w_2) \quad //$$

The functions in Lemma A4.1 can be vector valued.

Lemma A4.2 For $i \neq j$, $x_i \perp\!\!\!\perp x_j \mid \mathbf{z}, \mathbf{s}$; $y_i \perp\!\!\!\perp y_j \mid \mathbf{z}, \mathbf{s}$.

Proof:

$(y_i, x_i, z_i) \perp\!\!\!\perp (y_j, x_j, z_j)$ for all $i \neq j$.

i. $(x_i, x_j) \mid \mathbf{z}, \mathbf{s}$ has the same distribution as $(x_i, x_j) \mid z_i, z_j, \mathbf{s}_i, \mathbf{s}_j$

   where $\mathbf{s}_i = (s_i^1, \ldots, s_i^c, \ldots, s_i^C)$ for all i

   and $s_i^c = 1$ if and only if $x_i \in (a_c, b_c) \Rightarrow s_i^c = s^c(x_i)$, $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_i, \ldots, \mathbf{s}_n)$.

   Let $u_1 = (x_i, z_i)$, $u_2 = (x_j, z_j)$, $v_1 = x_i$, $v_2 = x_j$, $w_1 = (z_i, \mathbf{s}_i)$, $w_2 = (z_j, \mathbf{s}_j)$ in Lemma A4.1, and note that $(y_i, x_i, z_i) \perp\!\!\!\perp (y_j, x_j, z_j)$ for all $i \neq j \Rightarrow (x_i, z_i) \perp\!\!\!\perp (x_j, z_j)$.

   Therefore, by Lemma A4.1, we have $x_i \perp\!\!\!\perp x_j \mid z_i, z_j, \mathbf{s}_i, \mathbf{s}_j$, which is equivalent to $x_i \perp\!\!\!\perp x_j \mid \mathbf{z}, \mathbf{s}$ //

ii. Similarly, $y_i \perp\!\!\!\perp y_j \mid \mathbf{z}, \mathbf{s}$ //

Lemma A4.3 $(\mathbf{x}, \mathbf{z}) \perp\!\!\!\perp \mathbf{e} \Rightarrow \mathbf{x} \perp\!\!\!\perp \mathbf{e} \mid \mathbf{z}$.

Proof:

$(\mathbf{x}, \mathbf{z}) \perp\!\!\!\perp \mathbf{e} \Rightarrow \mathbf{x} \perp\!\!\!\perp \mathbf{e}, \mathbf{z} \perp\!\!\!\perp \mathbf{e} \Rightarrow f(\mathbf{e}) \perp\!\!\!\perp f(\mathbf{e} \mid \mathbf{z})$

$(\mathbf{x}, \mathbf{z}) \perp\!\!\!\perp \mathbf{e} \Rightarrow f(\mathbf{x}, \mathbf{z}, \mathbf{e}) = f(\mathbf{x}, \mathbf{z}) f(\mathbf{e})$

$$f(\underset{\sim}{x}, \underset{\sim}{e} \mid \underset{\sim}{z}) = \frac{f(\underset{\sim}{x}, \underset{\sim}{e}, \underset{\sim}{z})}{f(\underset{\sim}{z})} = \frac{f(\underset{\sim}{x}, \underset{\sim}{z}, \underset{\sim}{e})}{f(\underset{\sim}{z})} = \frac{f(\underset{\sim}{x}, \underset{\sim}{z}) f(\underset{\sim}{e})}{f(\underset{\sim}{z})} = f(\underset{\sim}{x} \mid \underset{\sim}{z}) f(\underset{\sim}{e}) = f(\underset{\sim}{x} \mid \underset{\sim}{z}) f(\underset{\sim}{e} \mid \underset{\sim}{x})$$

$\Rightarrow \mathbf{x} \perp\!\!\!\perp \mathbf{e} \mid \mathbf{z}$ //

Lemma A4.4 $(\mathbf{x}, \mathbf{z}) \perp\!\!\!\perp \mathbf{e} \Rightarrow \mathbf{x} \perp\!\!\!\perp \mathbf{e} \mid \mathbf{z}, \mathbf{s}$ where $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_i, \ldots, \mathbf{s}_n)$, $\mathbf{s}_i = (s_i^1, \ldots, s_i^c, \ldots, s_i^C)$.

Proof:

$(\mathbf{x}, \mathbf{z}) \perp\!\!\!\perp \mathbf{e} \Rightarrow (\mathbf{x}, \mathbf{z}, \mathbf{s}) \perp\!\!\!\perp \mathbf{e} \Rightarrow (\mathbf{z}, \mathbf{s}) \perp\!\!\!\perp \mathbf{e} \Rightarrow f(\mathbf{e}) = f(\mathbf{e} \mid \mathbf{z}, \mathbf{s})$

$(\mathbf{x}, \mathbf{z}, \mathbf{s}) \perp\!\!\!\perp \mathbf{e} \Rightarrow f(\mathbf{x}, \mathbf{z}, \mathbf{s}, \mathbf{e}) = f(\mathbf{x}, \mathbf{z}, \mathbf{s}) f(\mathbf{e})$

$$f(\underset{\sim}{x}, \underset{\sim}{e} \mid \underset{\sim}{z}, \underset{\sim}{s}) = \frac{f(\underset{\sim}{x}, \underset{\sim}{e}, \underset{\sim}{z}, \underset{\sim}{s})}{f(\underset{\sim}{z}, \underset{\sim}{s})} = \frac{f(\underset{\sim}{x}, \underset{\sim}{z}, \underset{\sim}{s}, \underset{\sim}{e})}{f(\underset{\sim}{z}, \underset{\sim}{s})} = \frac{f(\underset{\sim}{x}, \underset{\sim}{z}, \underset{\sim}{s}) f(\underset{\sim}{e})}{f(\underset{\sim}{z}, \underset{\sim}{s})}$$
$$= f(\underset{\sim}{x} \mid \underset{\sim}{z}, \underset{\sim}{s}) f(\underset{\sim}{e}) = f(\underset{\sim}{x} \mid \underset{\sim}{z}, \underset{\sim}{s}) f(\underset{\sim}{e} \mid \underset{\sim}{z}, \underset{\sim}{s})$$

$\Rightarrow \mathbf{x} \perp\!\!\!\perp \mathbf{e} \mid \mathbf{z}, \mathbf{s}$ //

Suppose the propensity scores are estimated by the logistic regresion model:

$$e(x_i) = \{1 + \exp[-(\gamma_0 + \gamma x_i)]\}^{-1} \qquad \text{(A4.1)}$$

If we ignore subscript i for convenience, the propensity scores e(x) is a one-to-one monotone function of x if $\gamma \neq 0$. Hence, condition on propensity scores e(x) is equivalent to condition on x.

A4.1.2 Expectation of the within subclass treatment effect estimator

Recall that in Section 4.1, we assumed that there are at least two observations in both the treatment group and the control group within each subclass. This condition can be expressed mathematically as saying the observed data are in $A^* = \{(\underset{\sim}{z}, \underset{\sim}{s}) : n_1^{(c)} \geq 2 \text{ and } n_0^{(c)} \geq 2 \text{ for all } c = 1, ..., C\}$. If we explicitly take this condition into account in the expectation of $\bar{x}_1^{(c)}$ in equation (4.4), we write $E(\bar{x}_1^{(c)})$ in the conditional form $E(\bar{x}_1^{(c)} \mid A^*)$. In order to obtain this expectation, we need z_i and s_i^c to become tractable by conditioning. That is, by the law of iterated expectation, we express $E(\bar{x}_1^{(c)} \mid A^*)$ as iterated expectation $E[E(\bar{x}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) \mid A^*]$. Here, conditional on A^*, i.e. given $(\underset{\sim}{z}, \underset{\sim}{s}) \in A^*$, we have $E(\bar{x}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) = E(\bar{x}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})$.

Next, we develop

$$E(\bar{x}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = E(\frac{\sum_{i=1}^{n} z_i s_i^c x_i}{\sum_{i=1}^{n} z_i s_i^c} \mid \underset{\sim}{z}, \underset{\sim}{s})$$

$$= \frac{\sum_{i=1}^{n} z_i s_i^c E(x_i \mid z_i, s_i^c)}{\sum_{i=1}^{n} z_i s_i^c}.$$

Notice that, if either $z_i = 0$ or $s_i^c = 0$, then $z_i s_i^c = 0$, so for all i, $z_i s_i^c E(x_i \mid z_i, s_i^c) = z_i s_i^c E(x_i \mid z_i = 1, s_i^c = 1)$ and

$$E(\bar{x}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = \frac{\sum_{\{i: z_i=1, s_i^c=1\}} z_i s_i^c \underbrace{E(x_i \mid z_i = 1, s_i^c = 1)}_{\mu_{x|z=1, s^c=1}}}{\sum_{\{i: z_i=1, s_i^c=1\}} z_i s_i^c}.$$

Since (x_i, z_i) have the same distribution as (x, z) for all i, we have

$E(x_i \mid z_i = 1, s_i^c = 1) = E(x \mid z = 1, s^c = 1)$ for all i, so

$$E(\bar{x}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = \frac{\sum_{i=1}^{n} z_i s_i^c E(x \mid z = 1, s^c = 1)}{\sum_{i=1}^{n} z_i s_i^c}$$

$$= E(x \mid z = 1, s^c = 1)$$

Therefore, by the law of iterative expectation, we have

$$E(\bar{x}_1^{(c)} \mid A^*) = E[E(\bar{x}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) \mid A^*]$$

$$= E[E(\bar{x}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) \mid A^*]$$

$$= E[E(x \mid z = 1, s^c = 1) \mid A^*]$$

$$= E(x \mid z = 1, s^c = 1).$$

Similarly, in the notation of Section 4.1, we have $E(\bar{x}_0^{(c)}) = E(x \mid z = 0, s^c = 1)$.

Therefore, $E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid A^*) = \beta_1[E(x \mid z = 1, s^c = 1) - E(x \mid z = 0, s^c = 1)] + \delta$.

If we let $B_c = \beta_1[E(x \mid z = 1, s^c = 1) - E(x \mid z = 0, s^c = 1)]$, then

$$Bias[\hat{\delta}(\underset{\sim}{w}) \mid A^*] = \sum_{c=1}^{C} w_c B_c .$$

A4.1.3 Variance of the within subclass treatment effect estimator

If we explicitly take A^* into account in the variance of $\bar{y}_1^{(c)} - \bar{y}_0^{(c)}$, we write $Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)})$ in the conditional form $Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid A^*)$. In order to obtain this variance, we need z_i and s_i^c to become tractable by conditioning. That is, by the law of iterated variance, we express $Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid A^*)$ as

$E[Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) \mid A^*] + Var[E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) \mid A^*]$. Here, conditional on A^*, i.e. given $(\underset{\sim}{z}, \underset{\sim}{s}) \in A^*$, we have $Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) = Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})$ and $E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) = E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})$. Hence, the variance of $\bar{y}_1^{(c)} - \bar{y}_0^{(c)}$ becomes

$$Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid A^*) = E[Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) \mid A^*] + Var[E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) \mid A^*].$$

From Section 4.1 and A4.1.2, we have

$E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = \beta_1[E(x \mid z = 1, s^c = 1) - E(x \mid z = 0, s^c = 1)] + \delta$. Thus,

$E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})$ is a non-random constant, so $Var[E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) \mid A^*] = 0$.

Next, we obtain $Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})$ as the following,

$$Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = Var(\bar{y}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) + Var(\bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) - 2Cov(\bar{y}_1^{(c)}, \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})$$

Since $\underset{\sim}{x} \perp \underset{\sim}{e} \mid \underset{\sim}{z}, \underset{\sim}{s}$ by Lemma A4.4

$$Var(\bar{y}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = Var(\beta_0 + \beta_1 \bar{x}_1^{(c)} + \delta + \bar{e}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = Var(\beta_1 \bar{x}_1^{(c)} + \bar{e}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})$$

$$= \beta_1^2 Var(\bar{x}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) + Var(\bar{e}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) + 2\beta_1 \underbrace{Cov(\bar{x}_1^{(c)}, \bar{e}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})}_{0}$$

$$= \beta_1^2 Var(\bar{x}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) + Var(\bar{e}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})$$

$$Var(\bar{e}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = Var(\frac{\sum_{i=1}^{n} z_i s_i^c e_i}{\sum_{i=1}^{n} z_i s_i^c} \mid \underset{\sim}{z}, \underset{\sim}{s}) = \frac{\sum_{i=1}^{n} (z_i s_i^c)^2 Var(e_i \mid \underset{\sim}{z}, \underset{\sim}{s})}{(\sum_{i=1}^{n} z_i s_i^c)^2} = \frac{\sum_{i=1}^{n} (z_i s_i^c)^2 Var(e_i)}{(\sum_{i=1}^{n} z_i s_i^c)^2} = \frac{\sum_{i=1}^{n} (z_i s_i^c)^2 \sigma_e^2}{(\sum_{i=1}^{n} z_i s_i^c)^2}$$

$$\left[z_i = 1, 0,\ s_i^c = 1, 0 \Rightarrow (z_i s_i^c)^2 = z_i s_i^c \Rightarrow \sum_{i=1}^{n} (z_i s_i^c)^2 = \sum_{i=1}^{n} z_i s_i^c \right]$$

$$= \frac{\sum_{i=1}^{n} z_i s_i^c \sigma_e^2}{(\sum_{i=1}^{n} z_i s_i^c)^2} = \frac{\sigma_e^2}{\sum_{i=1}^{n} z_i s_i^c} = \frac{\sigma_e^2}{n_1^{(c)}}$$

Similarly, $Var(\bar{e}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = Var(\frac{\sum_{i=1}^{n} (1 - z_i) s_i^c e_i}{\sum_{i=1}^{n} (1 - z_i) s_i^c} \mid \underset{\sim}{z}, \underset{\sim}{s}) = \frac{\sigma_e^2}{\sum_{i=1}^{n} (1 - z_i) s_i^c} = \frac{\sigma_e^2}{n_0^{(c)}}$

$$Var(\bar{x}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = Var(\frac{\sum_{i=1}^{n} z_i s_i^c x_i}{\sum_{i=1}^{n} z_i s_i^c} \mid \underset{\sim}{z}, \underset{\sim}{s}) = \frac{\sum_{i=1}^{n} (z_i s_i^c)^2 Var(x_i \mid \underset{\sim}{z}, \underset{\sim}{s}) + 2\sum_{i<}\sum_{j} z_i s_i^c z_j s_j^c Cov(x_i, x_j \mid \underset{\sim}{z}, \underset{\sim}{s})}{(\sum_{i=1}^{n} z_i s_i^c)^2}$$

$$\left[\text{by Lemma A4.2 } Cov(x_i, x_j \mid \underset{\sim}{z}, \underset{\sim}{s}) = 0 \text{ for } i \neq j\right]$$

$$= \frac{\sum_{i=1}^{n} z_i s_i^c Var(x_i \mid \underset{\sim}{z}, \underset{\sim}{s})}{(\sum_{i=1}^{n} z_i s_i^c)^2}$$

Notice that, if either $z_i = 0$ or $s_i^c = 0$, then $z_i s_i^c = 0$, so for all i, $z_i s_i^c Var(x_i \mid z_i, s_i^c) = z_i s_i^c Var(x_i \mid z_i = 1, s_i^c = 1)$ and

$$Var(\bar{x}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = \frac{\sum_{i=1}^{n} z_i s_i^c Var(x_i \mid z_i = 1, s_i^c = 1)}{(\sum_{i=1}^{n} z_i s_i^c)^2}$$

$$= \frac{\sum_{i=1}^{n} z_i s_i^c \sigma^2_{x|z=1, s^c=1}}{(\sum_{i=1}^{n} z_i s_i^c)^2} = \frac{\sigma^2_{x|z=1, s^c=1}}{\sum_{i=1}^{n} z_i s_i^c} = \frac{\sigma^2_{x|z=1, s^c=1}}{n_1^{(c)}}$$

Similarly, $Var(\bar{x}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = Var(\frac{\sum_{i=1}^{n} (1-z_i) s_i^c x_i}{\sum_{i=1}^{n} (1-z_i) s_i^c} \mid \underset{\sim}{z}, \underset{\sim}{s}) = \frac{\sigma^2_{x|z=0, s^c=1}}{\sum_{i=1}^{n} (1-z_i) s_i^c} = \frac{\sigma^2_{x|z=0, s^c=1}}{n_0^{(c)}}$.

Now, $Var(\bar{y}_1 \mid \underset{\sim}{z}, \underset{\sim}{s}) = \beta_1^2 \frac{\sigma^2_{x|z=1, s^c=1}}{n_1^{(c)}} + \frac{\sigma_e^2}{n_1^{(c)}} = \frac{1}{n_1^{(c)}}(\beta_1^2 \sigma^2_{x|z=1, s^c=1} + \sigma_e^2)$, and

$$Var(\bar{y}_0 \mid \underset{\sim}{z}, \underset{\sim}{s}) = \beta_1^2 \frac{\sigma^2_{x|z=0, s^c=1}}{n_0^{(c)}} + \frac{\sigma_e^2}{n_0^{(c)}} = \frac{1}{n_0^{(c)}}(\beta_1^2 \sigma^2_{x|z=0, s^c=1} + \sigma_e^2).$$

Next, for $i \neq j$, $x_i \perp\!\!\!\perp x_j \mid \mathbf{z}, \mathbf{s}$; $z_i(1 - z_i) = 0$ and $\mathbf{x} \perp\!\!\!\perp \mathbf{e} \mid \mathbf{z}, \mathbf{s}$. We have

$$Cov(\bar{y}_1^{(c)}, \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = Cov(\beta_0 + \beta_1 \bar{x}_1^{(c)} + \delta + \bar{e}_1^{(c)}, \beta_0 + \beta_1 \bar{x}_0^{(c)} + \bar{e}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})$$

$$= Cov(\beta_1 \bar{x}_1^{(c)} + \bar{e}_1^{(c)}, \beta_1 \bar{x}_0^{(c)} + \bar{e}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})$$

$$= \beta_1^2 \underbrace{Cov(\bar{x}_1^{(c)}, \bar{x}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})}_{0} + \beta_1 \underbrace{Cov(\bar{x}_1^{(c)}, \bar{e}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})}_{0} + \beta_1 \underbrace{Cov(\bar{e}_1^{(c)}, \bar{x}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})}_{0}$$

$$+ Cov(\bar{e}_1, \bar{e}_0 \mid \underset{\sim}{z}, \underset{\sim}{s})$$

$$Cov(\bar{e}_1, \bar{e}_0 \mid \underset{\sim}{z}, \underset{\sim}{s}) = Cov(\frac{\sum_{i=1}^{n} z_i s_i^c e_i}{\sum_{i=1}^{n} z_i s_i^c}, \frac{\sum_{i=1}^{n} (1-z_i) s_i^c e_i}{\sum_{i=1}^{n} (1-z_i) s_i^c} \mid \underset{\sim}{z}, \underset{\sim}{s})$$

$$= \frac{1}{(\sum_{i=1}^{n} z_i s_i^c)[\sum_{i=1}^{n} (1-z_i) s_i^c]} \sum_i \sum_j z_i s_i^c (1-z_j) s_j^c Cov(e_i, e_j \mid \underset{\sim}{z}, \underset{\sim}{s})$$

$$\left[z_i s_i^c (1-z_i) s_i^c = 0 \text{ and } Cov(e_i, e_j \mid \underset{\sim}{z}, \underset{\sim}{s}) = 0 \text{ for } i \neq j \right]$$

$$= 0$$

$$\Rightarrow Cov(\bar{y}_1^{(c)}, \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = 0$$

Therefore,

$Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = \frac{1}{n_1^{(c)}} (\beta_1^2 \sigma^2_{x|z=1, s^c=1} + \sigma_e^2) + \frac{1}{n_0^{(c)}} (\beta_1^2 \sigma^2_{x|z=0, s^c=1} + \sigma_e^2)$, and we have the variance of $\hat{\delta}_c$ as the following

$$Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid A^*) = E[Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) \mid A^*]$$

$$= E[\frac{1}{n_1^{(c)}} (\beta_1^2 \sigma^2_{x|z=1, s^c=1} + \sigma_e^2) + \frac{1}{n_0^{(c)}} (\beta_1^2 \sigma^2_{x|z=0, s^c=1} + \sigma_e^2) \mid A^*] + 0$$

$$= (\beta_1^2 \sigma^2_{x|z=1, s^c=1} + \sigma_e^2) E(\frac{1}{n_1^{(c)}} \mid A^*) + (\beta_1^2 \sigma^2_{x|z=0, s^c=1} + \sigma_e^2) E(\frac{1}{n_0^{(c)}} \mid A^*).$$

A4.1.4 Covariance of two within subclass treatment effect estimators

For two different subclasses, e.g. subclass c and d, if we explicitly take A^* into account in the covariance of $\bar{y}_1^{(c)} - \bar{y}_0^{(c)}$ and $\bar{y}_1^{(d)} - \bar{y}_0^{(d)}$, we write $Cov(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}, \bar{y}_1^{(d)} - \bar{y}_0^{(d)})$ in the conditional form $Cov(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}, \bar{y}_1^{(d)} - \bar{y}_0^{(d)} \mid A^*)$. In order to obtain this covariance, we need z_i and $\mathbf{s}_i$ to become tractable by conditioning. That is, by the law of iterated covariance, we express $Cov(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}, \bar{y}_1^{(d)} - \bar{y}_0^{(d)} \mid A^*)$ as

$$E[Cov(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}, \bar{y}_1^{(d)} - \bar{y}_0^{(d)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) \mid A^*] +$$
$$Cov[E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*), E(\bar{y}_1^{(d)} - \bar{y}_0^{(d)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) \mid A^*].$$

Here, conditional on A^*, i.e. given $(\underset{\sim}{z}, \underset{\sim}{s}) \in A^*$, we have

$$Cov(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}, \bar{y}_1^{(d)} - \bar{y}_0^{(d)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) = Cov(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}, \bar{y}_1^{(d)} - \bar{y}_0^{(d)} \mid \underset{\sim}{z}, \underset{\sim}{s}) \text{ and}$$

$$E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) = E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}).$$

From Section A4.1.3, we have $E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})$ and $E(\bar{y}_1^{(d)} - \bar{y}_0^{(d)} \mid \underset{\sim}{z}, \underset{\sim}{s})$ are non-random constants, so $Cov[E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*), E(\bar{y}_1^{(d)} - \bar{y}_0^{(d)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) \mid A^*] = 0$.

Hence, the covariance $Cov(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}, \bar{y}_1^{(d)} - \bar{y}_0^{(d)} \mid A^*)$ becomes

$$E[Cov(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}, \bar{y}_1^{(d)} - \bar{y}_0^{(d)} \mid \underset{\sim}{z}, \underset{\sim}{s}) \mid A^*].$$

We have $\bar{y}_1^{(c)} - \bar{y}_0^{(c)} = \sum_{i=1}^{n} (\frac{z_i}{n_1^{(c)}} - \frac{1-z_i}{n_0^{(c)}}) s_i^c y_i$, then

$$Cov(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}, \bar{y}_1^{(d)} - \bar{y}_0^{(d)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = \sum_{i=1}^{n}\sum_{j=1}^{n} (\frac{z_i}{n_1^{(c)}} - \frac{1-z_i}{n_0^{(c)}})(\frac{z_j}{n_1^{(c)}} - \frac{1-z_j}{n_0^{(c)}}) s_i^c s_j^d Cov(y_i, y_j \mid \underset{\sim}{z}, \underset{\sim}{s})$$

.

By Lemma A4.2, $Cov(y_i, y_j \mid \mathbf{z}, \mathbf{s}) = 0$ for $i \neq j$. For $i = j$, since $c \neq d$, then $s_i^c s_j^d = 0$ because either $s_i^c = 0$ or $s_j^d = 0$. Hence, $Cov(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}, \bar{y}_1^{(d)} - \bar{y}_0^{(d)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = 0$. Therefore, $Cov(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}, \bar{y}_1^{(d)} - \bar{y}_0^{(d)} \mid A^*) = 0$.

If we let $V_c = Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid A^*)$, then we have $Var[\hat{\delta}(\underset{\sim}{w}) \mid A^*] = \sum_{c=1}^{C} w_c^2 V_c$. If we also let $RMSE_c$ denote the RMSE of $\hat{\delta}_c$ conditional on A^*, then $RMSE_c = \sqrt{(B_c)^2 + V_c}$.

A4.4 Theory development extension to multiple covariates

We assume the outcome (y_i) variable is generated from a linear regression model that includes a treatment indicator (z_i) and multiple covariates:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \delta z_i + e_i \qquad \text{(A4.1)}$$

where $\mathbf{x}_i$ is the covariates vector of subject i, and $(\mathbf{x}, \mathbf{z}) \perp\!\!\!\perp \mathbf{e}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ is the regression parameter vector. We define the propensity scores of subject i as $P\{z_i = 1 \mid \mathbf{x}_i\} = e(\mathbf{x}_i)$.

A4.4.1 Expectation of the within subclass treatment effect estimator

Following the development in Section A4.1.2, if we explicitly take A^* into account in the expectation of $\bar{y}_1^{(c)} - \bar{y}_0^{(c)}$, we write $E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)})$ in the conditional form $E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid A^*)$. In order to obtain this expectation, we need z_i and s_i^c to become tractable by conditioning. That is, by the law of iterated expectation, we express $E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid A^*)$ as $E[E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) \mid A^*]$. Here, conditional on A^*, i.e. given $(\underset{\sim}{z}, \underset{\sim}{s}) \in A^*$, we have $E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) = E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})$. Hence, the expectation of $\bar{y}_1^{(c)} - \bar{y}_0^{(c)}$ becomes $E[E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) \mid A^*]$.

Next, we develop

$$E(\bar{y}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = E\left(\frac{\sum_{i=1}^{n} z_i s_i^c y_i}{\sum_{i=1}^{n} z_i s_i^c} \mid \underset{\sim}{z}, \underset{\sim}{s}\right)$$

$$= \frac{\sum_{i=1}^{n} z_i s_i^c E(y_i \mid z_i, s_i^c)}{\sum_{i=1}^{n} z_i s_i^c}.$$

Notice that, if either $z_i = 0$ or $s_i^c = 0$, $z_i s_i^c = 0$, so for all i, $z_i s_i^c E(x_i \mid z_i, s_i^c) = z_i s_i^c E(x_i \mid z_i = 1, s_i^c = 1)$ and

$$E(\bar{y}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = \frac{\sum_{\{i: z_i=1, s_i^c=1\}} z_i s_i^c \underbrace{E(y_i \mid z_i = 1, s_i^c = 1)}_{\mu_{y|z=1, s^c=1}}}{\sum_{\{i: z_i=1, s_i^c=1\}} z_i s_i^c}.$$

Since (y_i, z_i) have the same distribution as (y, z) for all i, we have

$$E(y_i \mid z_i = 1, s_i^c = 1) = E(y \mid z = 1, s^c = 1) \text{ for all } i, \text{ so}$$

$$E(\bar{y}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = \frac{\sum_{i=1}^{n} z_i s_i^c E(y \mid z = 1, s^c = 1)}{\sum_{i=1}^{n} z_i s_i^c} = E(y \mid z = 1, s^c = 1) = \mu_1^{(c)}$$

Therefore, by the law of iterative expectation, we have

$$\begin{aligned} E(\bar{y}_1^{(c)} \mid A^*) &= E[E(\bar{y}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) \mid A^*] \\ &= E[E(\bar{y}_1^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) \mid A^*] \\ &= E[E(y \mid z = 1, s^c = 1) \mid A^*] \\ &= E(y \mid z = 1, s^c = 1) \\ &= [E(\underset{\sim}{x} \mid z = 1, s^c = 1)]\underset{\sim}{\beta} + \delta \\ &= \underset{\sim}{\mu}_{x|z=1, s^c=1}\underset{\sim}{\beta} + \delta. \end{aligned}$$

Similarly, we have $E(\bar{y}_0^{(c)} \mid A^*) = E(y \mid z = 0, s^c = 1) = \underset{\sim}{\mu}_{x|z=0, s^c=1}\underset{\sim}{\beta}$. Therefore,

$$\begin{aligned} E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid A^*) &= E(y \mid z = 1, s^c = 1) - E(y \mid z = 0, s^c = 1)] \\ &= (\underset{\sim}{\mu}_{x|z=1, s^c=1} - \underset{\sim}{\mu}_{x|z=0, s^c=1})\underset{\sim}{\beta} + \delta \end{aligned}.$$

The weighted treatment effect estimator is

$$\hat{\delta}(\underset{\sim}{w}) = \sum_{c=1}^{C} w_c(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}),\ w_c \geq 0,\ \sum_{c=1}^{C} w_c = 1,$$

$$\Rightarrow E[\hat{\delta}(\underset{\sim}{w}) \mid A^*] = \sum_{c=1}^{C} w_c E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid A^*) = [\sum_{c=1}^{C} w_c(\underset{\sim}{\mu}_{x|z=1,\, s^c=1} - \underset{\sim}{\mu}_{x|z=0,\, s^c=1})]\underset{\sim}{\beta}$$

$$\Rightarrow Bias[\hat{\delta}(\underset{\sim}{w}) \mid A^*] = [\sum_{c=1}^{C} w_c(\underset{\sim}{\mu}_{x|z=1,\, s^c=1} - \underset{\sim}{\mu}_{x|z=0,\, s^c=1})]\underset{\sim}{\beta} - \delta$$

$$= \sum_{c=1}^{C} w_c B_c^m$$

where $B_c^m = (\underset{\sim}{\mu}_{x|z=1,\, s^c=1} - \underset{\sim}{\mu}_{x|z=0,\, s^c=1})\underset{\sim}{\beta} - \delta$.

A4.4.2 Variance of the within subclass treatment effect estimator

Similarly as in Section A4.1.3, if we explicitly take A^* into account in the variance of $\bar{y}_1^{(c)} - \bar{y}_0^{(c)}$, we write $Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)})$ in the conditional form $Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid A^*)$. In order to obtain this variance, we need z_i and s_i^c to become tractable by conditioning. That is, by the law of iterated variance, we express $Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid A^*)$ as

$E[Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) \mid A^*] + Var[E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) \mid A^*]$. Here, conditional on A^*, i.e. given $(\underset{\sim}{z}, \underset{\sim}{s}) \in A^*$, we have $Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) = Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})$ and $E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}, A^*) = E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})$. Hence, the variance of $\bar{y}_1^{(c)} - \bar{y}_0^{(c)}$ becomes

$$Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid A^*) = E[Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) \mid A^*] + Var[E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) \mid A^*].$$

From A4.4.1, we have

$E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) = E(y \mid z = 1, s^c = 1) - E(y \mid z = 0, s^c = 1)] = (\underset{\sim}{\mu}_{x|z=1, s^c=1} - \underset{\sim}{\mu}_{x|z=0, s^c=1})\underset{\sim}{\beta} + \delta$. Thus, $E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s})$ is a non-random constant, so $Var[E(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) \mid A^*] = 0$.

Therefore, we have the variance of $\hat{\delta}_c$ as

$Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid A^*) = E[Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid \underset{\sim}{z}, \underset{\sim}{s}) \mid A^*]$.

A4.4.3 Covariance of two within subclass treatment effect estimators

For two different subclasses, e.g. subclass c and d, if we explicitly take A^* into account in the covariance of $\bar{y}_1^{(c)} - \bar{y}_0^{(c)}$ and $\bar{y}_1^{(d)} - \bar{y}_0^{(d)}$, we write $Cov(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}, \bar{y}_1^{(d)} - \bar{y}_0^{(d)})$ in the conditional form $Cov(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}, \bar{y}_1^{(d)} - \bar{y}_0^{(d)} \mid A^*)$. In order to obtain this covariance, we need z_i and $\mathbf{s}_i$ to become tractable by conditioning. Following the same development as in Section A4.1.3, we have

$Cov(\bar{y}_1^{(c)} - \bar{y}_0^{(c)}, \bar{y}_1^{(d)} - \bar{y}_0^{(d)} \mid A^*) = 0$.

We let $V_c^{'} = Var(\bar{y}_1^{(c)} - \bar{y}_0^{(c)} \mid A^*)$, then we have $Var[\hat{\delta}(\underset{\sim}{w} \mid A^*)] = \sum_{c=1}^{C} w_c^2 V_c^{'}$. If we also let $RMSE_c$ denote the RMSE of $\hat{\delta}_c$ conditional on A^*, then $RMSE_c = \sqrt{(B_c^{'})^2 + V_c^{'}}$.

A4.4.4 Theory development extension under the same subclassification

Ignoring i for convenience, we provide the lemma below that if the propensity score is an increasing function of covariates, then the difference between the mean of the covariates for the treatment group and the control group is nonnegative.

Lemma A4.5: Suppose P{z = 1 | **x**} = e(**x**) is an increasing function of **x**, then E[**x** | z = 1, e(**x**) $\in$ (a, b)] $\geq$ E[**x** | x = 0, e(**x**) $\in$ (a, b)].

Proof :

$$f(\underset{\sim}{x} \mid z=1) = \frac{f(\underset{\sim}{x}, z=1)}{P(z=1)} = \frac{P(z=1 \mid \underset{\sim}{x}) f(\underset{\sim}{x})}{P(z=1)} = \frac{e(\underset{\sim}{x}) f(\underset{\sim}{x})}{P(z=1)},$$

$$f(\underset{\sim}{x} \mid z=0) = \frac{f(\underset{\sim}{x}, z=0)}{P(z=0)} = \frac{P(z=0 \mid \underset{\sim}{x}) f(\underset{\sim}{x})}{P(z=0)} = \frac{[1-P(z=1 \mid \underset{\sim}{x})] f(\underset{\sim}{x})}{P(z=0)} = \frac{[1-e(\underset{\sim}{x})] f(\underset{\sim}{x})}{P(z=0)},$$

$$E[\underset{\sim}{X} \mid z=1, e(\underset{\sim}{x}) \in (a,b)] = \frac{\int_{e(\underset{\sim}{x}) \in (a,b)} \underset{\sim}{x} f(\underset{\sim}{x} \mid z=1) d\underset{\sim}{x}}{\int_{e(\underset{\sim}{x}) \in (a,b)} f(\underset{\sim}{x} \mid z=1) d\underset{\sim}{x}} = \frac{\int_{e(\underset{\sim}{x}) \in (a,b)} \underset{\sim}{x} \frac{e(\underset{\sim}{x}) f(\underset{\sim}{x})}{P(z=1)} d\underset{\sim}{x}}{\int_{e(\underset{\sim}{x}) \in (a,b)} \frac{e(\underset{\sim}{x}) f(\underset{\sim}{x})}{P(z=1)} d\underset{\sim}{x}}$$

$$= \frac{\int_{e(\underset{\sim}{x}) \in (a,b)} \underset{\sim}{x} e(\underset{\sim}{x}) f(\underset{\sim}{x}) d\underset{\sim}{x}}{\int_{e(\underset{\sim}{x}) \in (a,b)} e(\underset{\sim}{x}) f(\underset{\sim}{x}) d\underset{\sim}{x}}$$

$$E[\underset{\sim}{X} \mid z=0, e(\underset{\sim}{x}) \in (a,b)] = \frac{\int_{e(\underset{\sim}{x}) \in (a,b)} \underset{\sim}{x} f(\underset{\sim}{x} \mid z=0) d\underset{\sim}{x}}{\int_{e(\underset{\sim}{x}) \in (a,b)} f(\underset{\sim}{x} \mid z=0) d\underset{\sim}{x}} = \frac{\int_{e(\underset{\sim}{x}) \in (a,b)} \underset{\sim}{x} \frac{[1-e(\underset{\sim}{x})] f(\underset{\sim}{x})}{P(z=0)} d\underset{\sim}{x}}{\int_{e(\underset{\sim}{x}) \in (a,b)} \frac{[1-e(\underset{\sim}{x})] f(\underset{\sim}{x})}{P(z=0)} d\underset{\sim}{x}}$$

$$= \frac{\int_{e(\underset{\sim}{x}) \in (a,b)} \underset{\sim}{x} [1-e(\underset{\sim}{x})] f(\underset{\sim}{x}) d\underset{\sim}{x}}{\int_{e(\underset{\sim}{x}) \in (a,b)} [1-e(\underset{\sim}{x})] f(\underset{\sim}{x}) d\underset{\sim}{x}}$$

let $c = \int_{e(\underset{\sim}{x})\in(a,b)} f(\underset{\sim}{x})d\underset{\sim}{x}$, so $\underset{\sim}{x} \mid e(\underset{\sim}{x}) \in (a,b) \sim pdf.\frac{1}{c} f(\underset{\sim}{x})$

$$E[\underset{\sim}{x} \mid z=1, e(\underset{\sim}{x}) \in (a,b)] = \frac{E[\underset{\sim}{x}\, e(\underset{\sim}{x}) \mid e(\underset{\sim}{x}) \in (a,b)]}{E[e(\underset{\sim}{x}) \mid e(\underset{\sim}{x}) \in (a,b)]}$$

$$E[\underset{\sim}{x} \mid z=0, e(\underset{\sim}{x}) \in (a,b)] = \frac{E[\underset{\sim}{x} \mid e(\underset{\sim}{x}) \in (a,b)] - E[\underset{\sim}{x}\, e(\underset{\sim}{x}) \mid e(\underset{\sim}{x}) \in (a,b)]}{1 - E[e(\underset{\sim}{x}) \mid e(\underset{\sim}{x}) \in (a,b)]}$$

for convenience, drop the conditioning notation, then

$$\frac{E[\underset{\sim}{x}\, e(\underset{\sim}{x})]}{E[e(\underset{\sim}{x})]} \geq \frac{E[\underset{\sim}{x}] - E[\underset{\sim}{x}\, e(\underset{\sim}{x})]}{1 - E[e(\underset{\sim}{x})]}$$

$$\Leftrightarrow E[\underset{\sim}{x}\, e(\underset{\sim}{x})] - E[\underset{\sim}{x}\, e(\underset{\sim}{x})]E[e(\underset{\sim}{x})] \geq E[e(\underset{\sim}{x})]E[\underset{\sim}{x}] - E[e(\underset{\sim}{x})]E[\underset{\sim}{x}\, e(\underset{\sim}{x})]$$

$$\Leftrightarrow E[\underset{\sim}{x}\, e(\underset{\sim}{x})] \geq E[e(\underset{\sim}{x})]E[\underset{\sim}{x}] \Leftrightarrow E[\underset{\sim}{x}\, e(\underset{\sim}{x})] - E[\underset{\sim}{x}]E[e(\underset{\sim}{x})] = Cov[\underset{\sim}{x}, e(\underset{\sim}{x})] \geq 0,$$

which is true by the Covariance Inequality Theorem in Casella and Berger (2001), because e(**x**) is assumed to be an increasing function of **x**. //

Lemma A4.5 indicates that if the propensity score is an increasing function of covariates, then the difference between the mean of the covariates from the treatment group and the control group is nonnegative.

Under the same subclassification, assume $B_c \geq 0$ for all subclasses, Theorem 4.4, Theorem 4.5 and their corollaries will apply under the situation with multiple covariates. We can also provide the following corollary from Theorem 4.4.

Corollary A4.4.3 For any subclassification, $\mathbf{w}^{EW}$ = (1/C, …, 1/C) be equal weights. Then $|Bias[\hat{\delta}(\underset{\sim}{w}^*)]| \geq |Bias[\hat{\delta}(\underset{\sim}{w}^{EW})]|$ if $\mathbf{u}^*$ and $\boldsymbol{B}$ are concordant.

An example of when $\mathbf{u}^*/\mathbf{u}$ and $\boldsymbol{B}$ are discordant: consider subclass c and d are two arbitrary but distinct subclasses, where $1 \le c, d \le C$, then discordance indicates $(\frac{u_c^*}{u_c} - \frac{u_d^*}{u_d})(B_c - B_d) \le 0$ for all c, d. Using Lemma 4.3, we develop the following Lemma to compare the variances of two propensity score estimators under the same subclassification but using different weights.

Lemma A4.6 If $u_c > 0,\ w_c = \frac{u_c}{\sum_{c=1}^{C} u_c},\ u_c^* \ge 0,\ w_c^* = \frac{u_c^*}{\sum_{c=1}^{C} u_c^*}$ for all c and

$(\frac{u_c^*}{u_c} - \frac{u_d^*}{u_d})(V_c - V_d) \le 0$ for all c, d, then $Var[\hat{\delta}(\underset{\sim}{w}^*)] \le aVar[\hat{\delta}(\underset{\sim}{w})]$ where

$a = \sum_{c=1}^{C} w_c^{*2} / \sum_{c=1}^{C} w_c^2$.

Proof : In Lemma 4.3, let $i = c,\ a_i = \frac{w_c^{*2}}{w_c^2},\ b_i = V_c,\ t_i = w_c^2$ to conclude

$$(\sum_{i=1}^{n} t_i a_i)(\sum_{i=1}^{n} t_i b_i) - (\sum_{i=1}^{n} t_i)(\sum_{i=1}^{n} t_i a_i b_i) = (\sum_{c=1}^{C} w_c^{*2})(\sum_{c=1}^{C} w_c^2 V_c) - (\sum_{c=1}^{C} w_c^2)(\sum_{c=1}^{C} w_c^{*2} V_c)$$

$$= (\sum_{c=1}^{C} w_c^2)\{aVar[\hat{\delta}(\underset{\sim}{w})] - Var[\hat{\delta}(\underset{\sim}{w}^*)]\} \ge 0 \quad //$$

We develop the following corollary of Lemma A4.6.

Corollary A4.6.1 If $u_c > 0,\ w_c = \frac{u_c}{\sum_{c=1}^{C} u_c},\ u_c^* \geq 0,\ w_c^* = \frac{u_c^*}{\sum_{c=1}^{C} u_c^*}$ for all c, if

$(\frac{u_c^*}{u_c} - \frac{u_d^*}{u_d})(V_c - V_d) \leq 0$ for all c, d, and if $\sum_{c=1}^{C} w_c^{*2} \leq \sum_{c=1}^{C} w_c^2$ then $Var[\hat{\delta}(\underset{\sim}{w}^*)] \leq Var[\hat{\delta}(\underset{\sim}{w})]$.

Corollary A4.6.2 If $u_c > 0$ and u_c is a non-increasing function of V_c for all c, then $Var[\hat{\delta}(\underset{\sim}{w}^{EW})] \leq Var[\hat{\delta}(\underset{\sim}{w})]$.

We also provide Lemma A4.7 to show that for any EV subclassification, the IV estimator has the smallest variance.

Lemma A4.7 For EV subclassification, let $u_c^* = 1/V_c$ be the inverse variance weight, where $V_1 = \ldots V_c = \ldots V_C$. Let u_c be an arbitrary weight and $\underset{\sim}{w}^* = \left(\frac{1}{C}, \ldots, \frac{1}{C}\right)$ be equal variance weights, then $Var[\hat{\delta}(\underset{\sim}{w}^*)] \leq Var[\hat{\delta}(\underset{\sim}{w})]$ always holds.

Proof :

$$Var[\hat{\delta}(\underset{\sim}{w})] = \sum_{c=1}^{C} w_c^2 V_c = \sum_{c=1}^{C} w_c^2 V_1 = V_1 \sum_{c=1}^{C} w_c^2$$

$$Var[\hat{\delta}(\underset{\sim}{w}^*)] = \sum_{c=1}^{C} (w_c^*)^2 V_c = \sum_{c=1}^{C} \frac{1}{C^2} V_c = \frac{1}{C^2} \sum_{c=1}^{C} V_1 = \frac{CV_1}{C^2} = \frac{V_1}{C}$$

$$Var[\hat{\delta}(\underset{\sim}{w}^*)] \leq Var[\hat{\delta}(\underset{\sim}{w})] \Leftrightarrow \frac{V_1}{C} \leq V_1 \sum_{c=1}^{C} w_c^2 \Leftrightarrow \sum_{c=1}^{C} w_c^2 = \sum_{c=1}^{C} \frac{u_c^2}{(\sum_{c=1}^{C} u_c)^2} = \frac{\sum_{c=1}^{C} u_c^2}{(\sum_{c=1}^{C} u_c)^2} \geq \frac{1}{C}$$

we know the sample variance $s_x^2 = \text{var}(x) = \dfrac{\sum x_i^2 - n\bar{x}^2}{n-1} \geq 0$

$$\Leftrightarrow (n-1)s_x^2 = \sum x_i^2 - n\bar{x}^2 \geq 0 \Leftrightarrow \sum x_i^2 \geq n\bar{x}^2 = \frac{(n\bar{x})^2}{n} = \frac{(\sum x_i)^2}{n}$$

$$Var[\hat{\delta}(\underset{\sim}{w}^*)] \leq Var[\hat{\delta}(\underset{\sim}{w})] \Leftrightarrow \frac{V_1}{C} \leq V_1 \sum_{c=1}^{C} w_c^2 \Leftrightarrow \sum_{c=1}^{C} w_c^2 = \sum_{c=1}^{C} \frac{u_c^2}{(\sum_{c=1}^{C} u_c)^2} = \frac{\sum_{c=1}^{C} u_c^2}{(\sum_{c=1}^{C} u_c)^2} \geq \frac{1}{C}$$

we know the sample variance $s_x^2 = \text{var}(x) = \dfrac{\sum x_i^2 - n\bar{x}^2}{n-1} \geq 0$

$$\Leftrightarrow (n-1)s_x^2 = \sum x_i^2 - n\bar{x}^2 \geq 0 \Leftrightarrow \sum x_i^2 \geq n\bar{x}^2 = \frac{n^2\bar{x}^2}{n} = \frac{(n\bar{x})^2}{n} = \frac{(\sum x_i)^2}{n}$$

it always true that $s_{u_c}^2 \geq 0$, so

$$(C-1)s_{u_c}^2 = \sum_{c=1}^{C} u_c^2 - C\bar{u}^2 \geq 0 \Leftrightarrow \sum_{c=1}^{C} u_c^2 - \frac{(C\bar{u})^2}{C} = \sum_{c=1}^{C} u_c^2 - \frac{(\sum_{c=1}^{C} u_c)^2}{C} \geq 0 \Leftrightarrow \sum_{c=1}^{C} u_c^2 \geq \frac{(\sum_{c=1}^{C} u_c)^2}{C} \quad //$$

We develop the following corollary of Lemma 4.7.

Corollary A4.7.1 Suppose $B_c \geq 0 \ \forall$ for all subclasses. For EV subclassification, let $u_c^* = 1/V_c$ be the inverse variance weight, and let u_c be an arbitrary weight. If $\mathbf{u}^*$ and $\boldsymbol{B}$ are concordant, then $|Bias[\hat{\delta}(\underset{\sim}{w})]| \geq |Bias[\hat{\delta}(\underset{\sim}{w}^*)]|$ and $Var[\hat{\delta}(\underset{\sim}{w}^*)] \leq Var[\hat{\delta}(\underset{\sim}{w})]$, so

$RMSE[\hat{\delta}(\underset{\sim}{w}^*)] \leq RMSE[\hat{\delta}(\underset{\sim}{w})]$.

Figure A5. 1 PSB Subclassification procedure

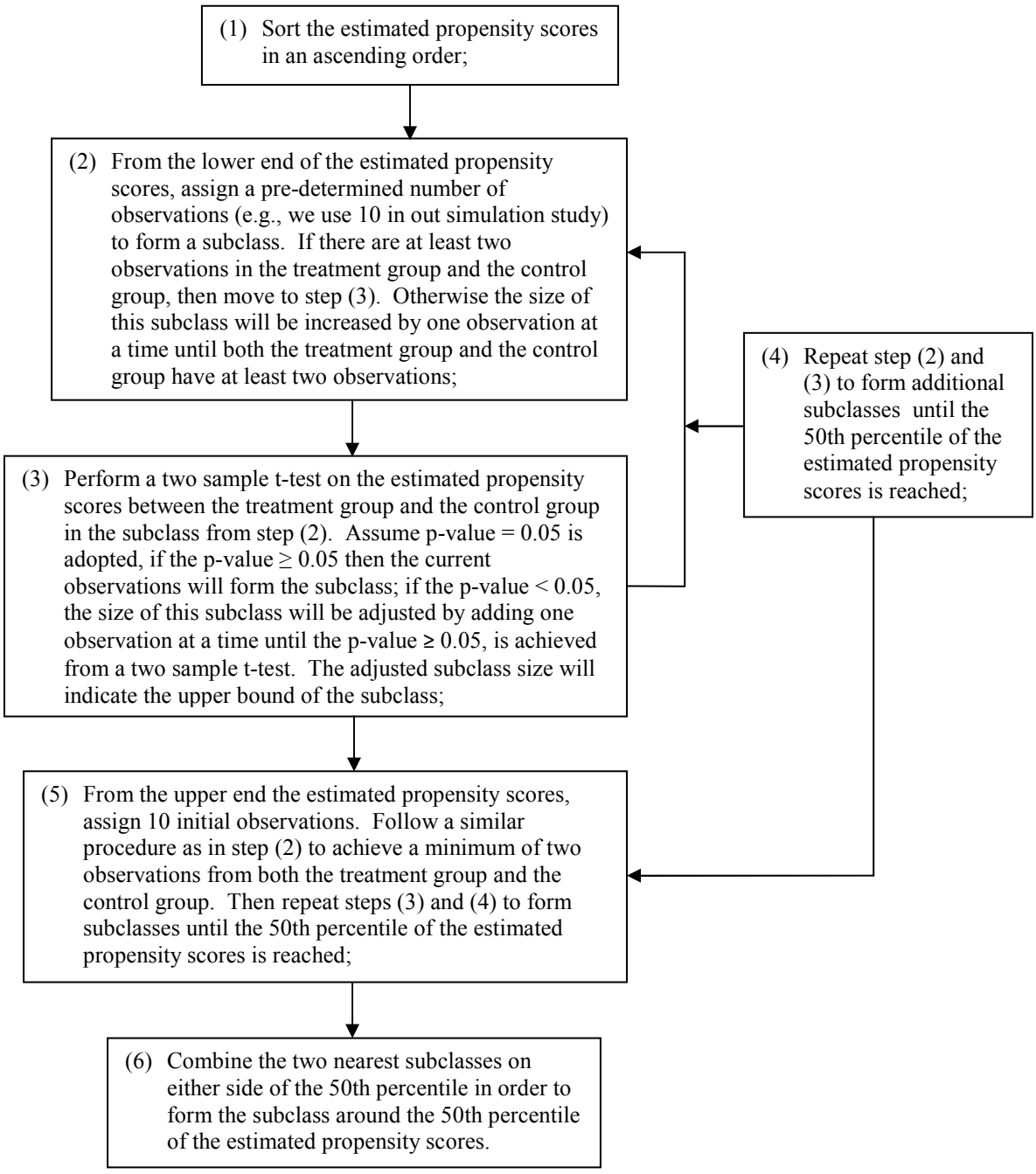

Figure A5. 2 PSB Subclassification procedure by restricting the size of the lowest subclass

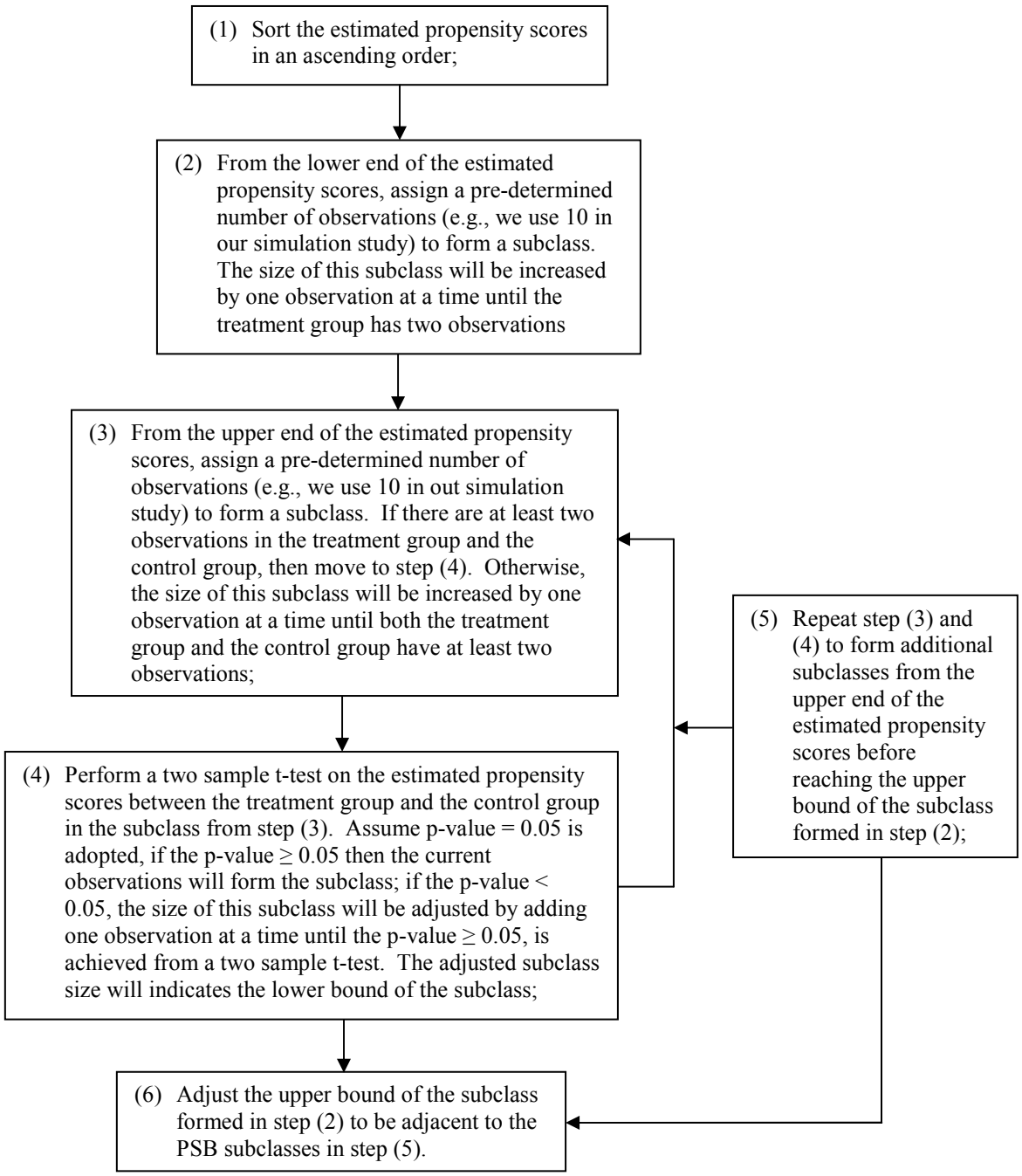

Figure A5. 3 Simulation study diagram for two independent covariates, (x_1, x_2)