COMPARING THE MATCHING PROPERTIES OF COARSENED EXACT MATCHING, PROPENSITY SCORE MATCHING, AND GENETIC MATCHING IN A NATIONWIDE DATA AND A SIMULATION EXPERIMENT

by

SHANSHAN QIN

(Under the Direction of Karen Samuelsen)

ABSTRACT

Sample matching is one statistical technique that can be applied to observational data to archive covariate balance and thus aid in estimating causal effects in studies lacking of randomization. This thesis (a) describes three types of sample matching methodologies-Propensity Score Matching (PSM), Coarsen Exact Matching (CEM), and Genetic Matching (GM), and (b) demonstrates and compares their application using empirical data from the Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS-K) and simulated data with seven scenarios differing by non-linear and/or non-additive associations between exposure and covariates. The study shows that CEM produces higher multivariate balance and consistently less biased effect estimate then the other two methods, although for data containing many categorical covariates curse of dimensionality is a noticeable concern in CEM. PSM and GM can result in more matched samples but carry higher extrapolation and model dependence in effect estimate.

INDEX WORDS: propensity score sample matching selection bias ECLS-K8 coarsened exact matching

COMPARING THE MATCHING PROPERTIES OF COARSENED EXACT MATCHING, PROPENSITY SCORE MATCHING, AND GENETIC MATCHING IN A NATIONWIDE OBSERVATIONAL DATA AND A MONTE CARLO EXPERIMENT

by

SHANSHAN QIN

BA, Renmin University of China, 2007

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF ART

ATHENS, GEORGIA

2011

COMPARING THE MATCHING PROPERTIES OF COARSENED EXACT MATCHING, PROPENSITY SCORE MATCHING, AND GENETIC MATCHING IN A NATIONWIDE OBSERVATIONAL DATA AND A MONTE CARLO EXPERIMENT

by

SHANSHAN QIN

Major Professor:	Karen Samuelsen
Committee:	Stephen Cramer
	Jonathan Templin

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2011

ACKNOWLEDGEMENTS

It is big pleasure to thank the many people who have made this thesis possible and because of whom my Master's study experience has been one that I will cherish forever.

It is difficult to overstate my gratitude to my Master's advisor Dr. Karen M. Samuelsen, who has stimulated and nourished my interest in scientific inquiry and methodology, and I would have been lost without her. Throughout my Master's study and thesis-writing period, she provided spring-like encouragement, sunny smile, sound advice, and good teaching, which were always there when I needed them.

It has been my good fortune to be a student of Dr. Stephen E. Cramer and Dr. Jonathan Templin who introduced me to the world of educational assessment and applied statistics. Their knowledge, experience and insights made me feel this career so fascinating. I am also indebted to all other professors and all graduate students in REMS for providing a stimulating and fun environment in which to learn and grow.

Dr. Wolfgang Lukowitz supported my Master's study and enlarged of my vision of academic system. My husband, Yujun is a source of my joy and a motion of my work. As natural scientists, they exemplified lively to me the objectivity, integrity, and persistence in academic research. Dedicating this thesis to them is a small token of my love and appreciation.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Making casual inferences based on observational data is a risky undertaking. Many educational and psychological researchers must resort to this strategy because randomized experiments, generally considered the gold standard in terms of assessing the causal effect of treatments, are not possible. There are a variety of reasons that a randomized experiment might not be possible. Among these are: the expense inherent in randomized designs; ethical concerns about withholding treatment from one sample; questions regarding the external generalizability of results due to the short term of the experiment; restrictions in the environment or settings, and; limited sample sizes in real practice (Rubin,2006).

The primary problem with observational data is that the probabilistic equivalence of pre-treatment covariates is likely not held; therefore the estimated treatment effect would be biased by confounders that unevenly distribute across groups. To remove the confounding bias in an observational study, sample matching has been developed in the statistical treatment effect literature (Cochran & Cox, 1957; Rubin, 1972), and the econometric policy evaluation literature (Dehejia & Wahba, 1999; J. J. Heckman, Urzua, & Vytlacil, 2006; J. Heckman, 1974; Quandt, 1958; Roy, 1951). The key goal of sample matching is to prune observations from the data so that the remaining data have more similar empirical distributions of the covariates across groups. Compared with the covariance adjustment directly on outcome variables, which is still the most common remedy for confounders in behavioral sciences, including education (Hahs-Vaughn & Onwuegbuzie, 2006), sampling matching has two appealing features (Cochran & Cox, 1957;

Rubin, 1972; Rubin, 1973; Rubin, 1997). First, if the observed values of confounders in different groups do not adequately overlap, sample matching would warn the investigator immediately, while covariance adjustment would still calculate the estimated mean effect adjusted for the influence of the mean covariates in the combined sample. In the extreme cases, the mean covariates in the combined samples can lie out of the range of covariates of any group, hence the adjusted mean effect, no matter how large the sample size is, cannot support any causal conclusions about treatment. Second, sampling matching reduces the model dependence in estimating the treatment effect because the estimated treatment effect in well-matched samples can be the mean difference on outcomes from difference groups, without subtracting out the part in outcomes that covariates account for, Covariate adjustment faces many choices of regression models of outcomes on covariates that could result in difference adjustments on estimated treatment effect.

The most widely applied sample matching is matching on unidimenstional balancing scores of observations. Observations with similar balancing scores are assumed to have similar covariates. As the most recognized balancing score, the propensity score is the conditional probability of being assigned to a treatment group given observed covariates. Propensity Score Matching (PSM) has been widely used in economic research (Cox-Edwards & Rodriguez-Oreggia, 2009; Czajka, Sharon M. Hirabayashi, Little, & Rubin, 1992; Lechner, 2002; Liu & Lynch, 2011; Mendola, 2007) and in medical research (Connors et al., 1996; Earle et al., 2001; Foody, Cole, Blackstone, & Lauer, 2001; Johnson et al., 2006; Lytle et al., 1999; Peterson et al., 2006). Despite the popularity of this approach its assumption either on ellipsoidal symmetry of covariates or on correct specification of the propensity score model limit the occasions where it could achieve bias reduction. When either of the two requirements is satisfied, PSM will reduce

bias in all linear combinations of the covariates, a property named "equal percent bias reduction" (EPBR) (Rubin, 1976; Rubin & Thomas, 1992a). If the second requirement is met, PSM will further reduce bias due to the covariates' nonlinear and interaction terms. If neither is met, PSM will improve balance on one covariate leading to a reduction in balance on others, or increase the bias of some functions of the covariates even if all univariate means are closer in the matched samples than the unmatched. Unfortunately, the latter scenario is often the reality.

To get by with weaker assumptions, two alternatives have been developed. One is the generalization of PSM, called Genetic Matching (GM), which matches samples on their weighted Mahalanobis distances calculated from the distance matrix including propensity scores and other functions of the original covariates (Diamond & Sekhon, In press.). GM adopts an iterative approach of automatically checking and improving covariate balance measured by univariate paired t-tests or univariate Kolmogorov-Smirnov (KS) tests. In every iteration, weights used in the distance calculation are adjusted to eliminate significant results from the univariate balance tests from the end of the last iteration. The iterative process ends when all univariate balance tests yield non-significant results. GM loosens the requirement on ellipsoidal distribution of covariates, however, unless nonlinear, and interaction terms are added into the distance matrix, it is still unclear whether GM can control the bias due to these terms. Another problem is the used univariate balance tests on covariates and on propensity scores, which do not assess joint balance of covariates.

The other alternative is multidimensional matching, or monotonic imbalance bounding (MIB), proposed by Iacus, King and Porro(Iacus, King, & Porro, In Press). A representative MIB technique is Coarsened Exact Matching (CEM). The essential idea of CEM is similar to blocking. CEM categorizes original covariates into user-defined intervals then matches treated units with

control units falling into the same the hyper-cuboids (Porro & Iacus, 2009) with all coarsened covariates as coordinates. Since how to coarsen each of the original covariates is user-defined, users would know how close the original values of covariates of units in the same hyper-cuboid are, without post-matching balance checking. In contrast, in PSM and GM the covariate difference among matched units are revealed after the matching meaning users have to alternate between modifying balancing score models and checking covariate balance. Moreover, changing the coarsening level of one covariate has no effect on the imbalance of any other covariate, but modifying balancing score models to reduce the imbalance of one covariate may worsen other covariates' imbalance. Also, like blocking, CEM directly makes the covariates' univariate distribution and multivariate distributions more similar across groups, so it should theoretically improve mean imbalance, variance imbalance, interaction imbalances and all other imbalance on covariates simultaneously. However, as the dimensions of covariates and the numbers of intervals of coarsened covariates increase, the curse of dimensionality may become the major concern with CEM and result in more hyper- cuboids and fewer matched samples in all hyper-cuboids. Iacus, King and Porro (In press) argued that it is better not regarded as a disadvantage of CEM in practice. First, if insufficient matches are found by CEM using even the coarsest categorization that substantive theory can tolerate, it actually indicates that the data quality is bad for making causal inference because the overlapping of multivariate distribution of covariates in different groups is too small. In such cases, the big overlapping in propensity scores may not reflect the poor multivariate overlapping hence using PSM can still results in a lot of matched samples, which, however, are not good matched. - Second, real data in social science studies tend to have highly correlated covariates. In this situation, units having the closer values on covariate

x_i will be likely to have closer value on covariate x_j, meaning that the increasing on dimensions does not reduce the number of matches heavily.

Comparisons between PSM and the two alternatives have been done by few researchers ((Diamond & Sekhon, In press; Iacus, King & Porro, In press). Both studies involved empirical and simulated examples. Diamond and Sekhon (In press) only assessed the univariate balance of covariates of matched samples and did not involve CEM in their study. They concluded that GM outperformed PSM with less unbalance of covariates and less variances in effect estimation. Iacus, King and Porro (In press) evaluated all three matching methods, however, their comparison between CEM and GM was only in one stimulation setting, which only considered the case where there was highly nonlinear relationship between the true propensity scores and the covariates. Their results indicated the superiority of CEM over the other two. The current study tries to address the limitations in those previous studies and to aid educational and psychological researchers in understanding and applying those matching methods. The empirical data used in this research came from Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K8) eighth-grade data files. The effects of on-time versus delayed kindergarten entry on children's reading performance in the spring of first grade are estimated using the different matching algorithms. Students who have delayed kindergarten entry are defined as the treatment group.

CHAPTER 2

LITERATURE REVIEW

Ignorability of Treatment Status

The theory of counterfactual causality in statistics (Sobel, 1996) and econometrics (J. J. Heckman & Vytlacil, 2007) claims that an experimental unit *i* has two theoretical outcomes — one that would be observed if the unit were in the treatment group ($\mathbf{Y_{i1}}$) and one that would be observed if the unit were in the control group ($\mathbf{Y_{i0}}$). **D** = 1 if $\mathbf{Y_{i1}}$is observed, and **D** = 0 corresponds to $\mathbf{Y_{i0}}$ being observed. The observed outcome is then

$$\boldsymbol{Y_i = DY_{i1} + (1 - D)Y_{i0}}$$

The difference between the two outcomes is the unbiased individual treatment effect for unit *i*, denoted as

$$\boldsymbol{TE_i \equiv Y_{i1} - Y_{i0}}$$

Two individual outcomes cannot be observed at one time, therefore any functions related to both of them, like TE_i, is not identifiable. In practice, researchers often have to focus on augmented treatment effects, such as Sample Average Treatment Effect (SATE)

$$\boldsymbol{SATE = E(Y_1) - E(Y_0)}$$

where E (.) denotes expectation in the sample combining all control units and treated units. There is also Population Average Treatment Effect, PATE (Imbens, 2004). SATE and PATE are equivalent when we have random samples from the population. Here, we only consider the situation when SATE=PATE. SATE can be identified if it equals the observable $\boldsymbol{E(Y_1|D = 1) - E(Y_0|D = 0)}$, the observed outcome difference between the treated group and the control group.

That happens only when the treatment is randomly assigned and samples comply with the assignment so that the units in the control group and the ones in the treated group are from the same population, then the conditional distribution of $\boldsymbol{Y_i}$ given $\boldsymbol{D}$ is the same as the marginal distribution of $\boldsymbol{Y_i}$, which is equivalent to saying that the potential outcomes are independent of treatment status, or that treatment status is ignorable. This can be denoted as

$$\boldsymbol{Y_0, Y_1 \perp\!\!\!\perp D}$$

$$\boldsymbol{SATE = E(Y_1) - E(Y_0) = E(Y_1|D = 1) - E(Y_0|D = 0)}$$

where $\perp\!\!\!\perp$ denotes independence.

In an observational study, where randomization is absent, matching supposes that if all confounders can be involved, denoted as $\boldsymbol{X}$, meaning there is no omitted confounders, and control units and treated units are both found to have similar distribution of $\boldsymbol{X}$, then

$$\boldsymbol{Y_0, Y_1 \perp\!\!\!\perp D \mid X}$$

$$\boldsymbol{E_X\{E(Y_1|x, D = 1) - E(Y_0|x, D = 0)\} = E_X\{E(Y_1|x) - E(Y_0|x)\}}$$

Where $\boldsymbol{E_X}$ denotes expectations with respect to the distribution of $\boldsymbol{X}$ in a sample combining control units and treated units. Matching assumes that after conditioning on $\boldsymbol{X}$, there is some randomness in the environment that switches units across treatment status (J. J. Heckman & Vytlacil, 2007). If all values of $\boldsymbol{X}$ are sampled from the entire population of units containing control and treated units together, we have

$$\boldsymbol{E_X\{E(Y_1|x) - E(Y_0|x)\} = E(Y_1) - E(Y_0)}$$

$$\boldsymbol{E_X\{E(Y_1|x, D = 1) - E(Y_0|x, D = 0)\} = SATE}$$

(Rosenbaum & Rubin, 1983). Therefore, the optimal purpose of sample matching is to balance the distribution of $\boldsymbol{X}$ between the treatment group and the control group, not to use $\boldsymbol{X}$ or any function of $\boldsymbol{X}$ to predict D.

Balancing Score

Different sample matching methods differ in the choice of function of $\boldsymbol{X}$ used to make the distribution of $\boldsymbol{X}$ in the control group and treatment group similar. Rosenbaum and Rubin named such functions balancing scores *b (x)* (1983). The similar *b (x)* should indicate the similar $\boldsymbol{X}$. When we substitute *b (x)* with *X*, maintaining the assumption about $\boldsymbol{X}$, we have

$$\boldsymbol{Y_0, Y_1} \perp\!\!\!\perp \boldsymbol{D} \mid \boldsymbol{b\,(x)}$$

$$\boldsymbol{E_{b\,(x)}\{E(\,Y_1|b\,(x), D = 1) - E(\,Y_0|b\,(x), D = 0)\} = E_{b\,(x)}\{E\big(\,Y_1\big|b(x)\big) - E\big(\,Y_0\big|b(x)\big)\},}$$

$$\boldsymbol{= E(Y_1) - E(Y_0)}$$

where *b (x)'s* are sampled from the entire combination of control and treated units, $E_{b\,(x)}$ denotes expectations with regards to the distribution of *b (x)* in this entire combination.

The finest balancing scores are first-order terms of $\boldsymbol{X}$, the coarsest ones are the many-to-one, or scalar functions of $\boldsymbol{X}$. The latter include propensity scores, Mahalanobis distances, and weighted Mahalanobis distances in GM. Between the two extreme cases of balancing scores are the coarsened or categorized $\boldsymbol{X}$.

The scalar function of $\boldsymbol{X}$, as a balancing score, has the advantage of avoiding what is known as the curse of dimensionality: the number of covariates is too large to allow a close match on every covariate simultaneously. It is like cutting a blueberry cheese cake. The more we cut, the smaller the cake pieces are, and the less likely two blue berries are found in one piece. For matching on a scalar function of $\boldsymbol{X}$ to maximize covariate balance, there are some requirements. The most practical one is the correct specification of each scalar function which should involve all first-order terms of $\boldsymbol{X}$ and their interaction or nonlinear terms if there is evidence that these terms influence the outcome. In reality, a scalar *b (x)* is usually regressed on

all observed first-order terms of *X*, mainly to remove mean imbalance on these covariates. Such models of *b (x)* can also control other types of imbalances on the distributions of confounders, when the following data restrictions are satisfied.

(1) All values of *X* are drawn randomly from specified populations (Rosenbaum and Rubin ,1985a);

(2) The multivariate distribution for $\boldsymbol{X}$ in every group is ellipsoidal (Rubin & Thomas, 1992a) —e.g., a normal distribution or t distribution—or the multivariate distribution of $\boldsymbol{X}$ in all groups is a discriminant mixture of proportional ellipsoidally symmetric densities (Rubin & Stuart, 2006).

According to Rubin and Stuart (Rubin & Stuart, 2006), a multivariate distribution of *X*, *F (X)* is a "discriminant mixture of proportional ellipsoidally symmetric" (DMPES) distribution if it possesses the following properties:

(i) *F (X)* is a mixture of *K* ellipsoidally symmetric distributions,

$$F(X) = \sum_{k=1}^{K} \alpha_k F_k\,(X)$$

where *K* is the number of groups that are involved in sample matching, F_k has a center $\boldsymbol{\mu}_k$ and an inner product $\boldsymbol{\Sigma_k}$, α_k is nonnegative for k=1, …,K and $\sum_{k=1}^{K} \alpha_k = 1$.

(ii) The *K* inner products are proportional,

$$\boldsymbol{\Sigma_i} \propto \boldsymbol{\Sigma_j},\ for\ i, j = 1, \ldots K.$$

(iii) The *K* centers are such that all best linear discriminants between any two groups are proportional,

$$(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\boldsymbol{\Sigma_k^{-1}} \quad \propto \quad (\boldsymbol{\mu}_{i'} - \boldsymbol{\mu}_{j'})\boldsymbol{\Sigma_{k'}^{-1}\Sigma_j}, \text{for all } i, j, k, i', j', k' = 1, \ldots, K.$$

In the case of matching a control group with a treat group, *K*=2.These restrictions seem too stringent to be real. But if they hold, Rubin and Stuart (2006) proved that if there is mean difference among the covariates of different groups, matching samples based on an arbitrary linear combination of $\boldsymbol{X}$, $\boldsymbol{Y} = \boldsymbol{\beta}'\boldsymbol{X}$, where $\boldsymbol{Y}$ have the correlation $\boldsymbol{\rho}$ with **Z**, the standardized best linear discriminant of *F (X)*, and the correlation $\sqrt{\mathbf{1}-\boldsymbol{\rho}^{\mathbf{2}}}$ with **W** which is orthogonal to **Z**, has the following property, named equal percent bias reducing (EPBR),

$$\boldsymbol{Y}=\boldsymbol{\rho Z}+\sqrt{\mathbf{1}-\boldsymbol{\rho}^{\mathbf{2}}}\ \boldsymbol{W}$$

$$\frac{E(\overline{Y}_{mT}(x)-\overline{Y}_{mC}(x))}{E(\overline{Y}_{rT}(x)-\overline{Y}_{rC}(x))}=\frac{E(\overline{Z}_{mT}(x)-\overline{Z}_{mC}(x))}{E(\overline{Z}_{rT}(x)-\overline{Z}_{rC}(x))}$$

Where *rT* refers to a random sample from N_T treated units, and *rC* means a random sample from N_C control units. EPBR says that after units are matched on **Z** the across-group difference on any linear combination of *X* is reduced by the same percentage because this percentage always equal to the percentage of difference reduction on **Z**, which . Matching methods which are not EPBR would increase the bias for some linear functions of $\boldsymbol{X}$ when reducing bias on others.

Propensity Score Matching

Propensity Score Matching (PSM) is an EPBR matching. A propensity score, e (x), is the conditional probability of units being assigned to the treatment group, that is *e(x)=Pr(D=1|X)*. In randomized experiments, units enter different groups by chance, so the distributions of $e\ (x)$ and $\boldsymbol{X}$ are very likely to similar across groups.

If $\boldsymbol{Y_0, Y_1} \perp\!\!\!\perp \boldsymbol{D} \mid \boldsymbol{X}$, then $\boldsymbol{Y_0, Y_1} \perp\!\!\!\perp \boldsymbol{D} \mid \boldsymbol{e\ (x)}$,

$$\boldsymbol{E_{e\ (x)}}\{\boldsymbol{E}(\boldsymbol{Y_1}|\boldsymbol{e}\ (\boldsymbol{x}), \boldsymbol{D}=\mathbf{1})-\boldsymbol{E}(\boldsymbol{Y_0}|\boldsymbol{e}\ (\boldsymbol{x}), \boldsymbol{D}=\mathbf{0})\}=\boldsymbol{E_{e\ (x)}}\{\boldsymbol{E}(\boldsymbol{Y_1}|\boldsymbol{e}(\boldsymbol{x}))-\boldsymbol{E}(\boldsymbol{Y_0}|\boldsymbol{e}(\boldsymbol{x}))\}$$

$$=\mathbf{E}(\boldsymbol{Y_1})-\mathbf{E}(\boldsymbol{Y_0})$$

Therefore *e(x)* is a balancing score of *X*, and *e (x)* should not equal to 1or 0, otherwise units cannot switch between groups (Heckman and Vytlacil, 2007).

In observational studies, true e (x) can be estimated by a logistic regression on $\boldsymbol{X}$. Matching based on *e (x),* as a linear combination of the first-ordered $\boldsymbol{X}$, is EPBR, when the distribution of $\boldsymbol{X}$ is DMPES or when estimated $e\ (x)$ is calculated from a correct specified model. In the first case, matching on $e\ (x)$ results in bias reduction in the discriminant Z and hence minimize the across-group differences of all linear combination of the first-ordered $\boldsymbol{X}$ (Rubin & Thomas, 1992b). It should be noticed that PSM obtains the mean balances of first-ordered $\boldsymbol{X}$ in expectation, or over repeated studies (Rosenbaum & Rubin, 1985). Just like in randomized experiments, chance imbalance on covariates could exist after well-performed matching. Additional adjustment on covariates (e.g covariance adjustment) might be still needed after PSM. Also, it is important to recognize that EPBR property mainly deals with the mean difference of first-order terms of $\boldsymbol{X}$ across groups, and it doesn't guarantee reducing difference on nonlinear or interaction terms of $\boldsymbol{X}$ across groups.When nonlinear or nonadditive relationships exist between treatment outcome and $\boldsymbol{X}$, PSM become less effectual.

There are numerous ways of matching samples on $e\ (x)$.The most desired and most difficult one is to match all treated units with control units having the exactly same values of $e(x)$'s—Exact Matching. In practice, researchers have to match samples within an acceptable neighborhood of $e(x)$—Inexact Matching. Because conditions for PSM to completely balance covariates across the entire samples are rarely available, Rosenbaum and Rubin (1984) recommended inexact matching followed by subclassifying samples into five strata based on propensity score. Five subclasses of many continuous distributions have been observed to remove at least 90% of the bias due to non-random selection effects (Cochran, 1968). Even

though the covariates have quite different distributions between the whole treated group and the whole control group, these differences can be reduced within each of the five strata. Then the treatment effect can be estimated for every stratum(Rosenbaum & Rubin, 1983). Also, matching with or without replacement of controls units, yields different results. Abadie and Imbens(Abadie & Imbens, 2006) found matching with replacement resulted in the higher degree of covariate balance and the lowest conditional bias.

Even after the application of subclassification on $e(x)$, within-stratum covariates may still have some imbalances. When this occurs, we must revise the model of $e(x)$, and then check the multivariate and univariate balances of $\boldsymbol{X}$ again. The two steps should replicate until the acceptable balances achieve or the unbalanced $\boldsymbol{X}$ are included in the later covariance adjustment on the outcomes.

Genetic Matching

Genetic Matching (GM), proposed by Diamond and Sekhon(Diamond & Sekhon, In Press), offers the benefit of combining the merits of traditional PSM and Mahalanobis Metric Matching (MMM) and the benefit of automatically checking balance and searching for best solutions, via software computational support and machine learning algorithms.

Just like PSM, MMM matches samples on a scalar function of $\boldsymbol{X}$, named Mahalanobis Distance (MD), the distance between a treated unit and a control unit in the high dimensional space of $\boldsymbol{X}$,

$$(\boldsymbol{X}_T - \boldsymbol{X}_C)' \boldsymbol{S}^{-1} (\boldsymbol{X}_T - \boldsymbol{X}_C)$$

Where $\boldsymbol{X}_T$ and $\boldsymbol{X}_C$ are vectors of covariates from different groups. One treated units will be matched with one or several control units having the smallest MDs with it. DMPES is also

required for MMM to be EPBR. Simulation studies (Rosenbaum & Rubin, 1985) showed that MMM was more successful than PSM in reducing the standardized mean difference of individual X, but it was far less successful in reducing the standardized mean difference on e (x), the scalar function of the covariates. The studies also found that and the combination of the two methods was superior to either of them alone in reducing covariate imbalance. Hence, it's argued that that e (x) should be included among the covariates matrix of MMM, alternatively, one may first match on the propensity score and then match based on MD within propensity score strata.

Rather than obtaining MD from one metrix as in MMM, the evolutionary algorithm of Genetic Matching (GM) searches amongst a range of metrics to find a generalized Mahalanobis distance, GMD. Candidate distance matrices differ in their assignment of weights for all confounding V's which include first-ordered $\boldsymbol{X}$, $e(x)$, and other functions of $\boldsymbol{X}$. The algorithm weights each variable according to its relative importance in achieving the best overall balance. As discussed below,

$$\text{GMD}\ (\boldsymbol{V_T}, \boldsymbol{V_C}, \boldsymbol{W}) = \sqrt{(\boldsymbol{V_T} - \boldsymbol{V_C})' \boldsymbol{W}\ (\boldsymbol{S}^{-\frac{1}{2}})\ (\boldsymbol{V_T} - \boldsymbol{V_C})}$$

where $\boldsymbol{V_T}$ and $\boldsymbol{V_C}$ are the vectors of covariates and their other functions from the treated group and the control group, and $\boldsymbol{S}^{\frac{1}{2}}$ is the Cholesky decomposition of S which is the variance-covariance matrix of $\boldsymbol{V}$. $\boldsymbol{W}$ denotes a weight component, which is a positive definite diagonal matrix. If both $\boldsymbol{Vs}$ only contain the first order term of $\boldsymbol{X}$ and $\boldsymbol{W}$ is an identity matrix, GMD=MD; if $e(x)$ is one element of $\boldsymbol{V}$ and its corresponding entry in W is 1 while other entries of $\boldsymbol{W}$ are 0, then GM is will be equivalent to propensity score matching. If neither matching on MD nor on $e(x)$ can achieve acceptable imbalance reduction, further modification on W will be conducted.

The general procedure for GM is

(1) Create an initial generation of ***Ws***, the population size of which might be n_w =1000. Calculate 1000 times of GMDs based on ***Ws***, then match treated units with replaceable control units based their GMD's, resulting in n_w ways of matching.

(2) Evaluate W via balance checking. Good Ws would be the ones that minimize any loss function specified by researchers, e.g the largest mean difference or distributional differences of individual elements of ***V***, which could be reflected in test statistics from t-tests or Kolmogorov–Smirnov tests. If all loss functions are minimized to acceptable levels, choose the matched samples based on the best W.

(3) If unacceptable confounder imbalances still exist after Step 2, a new generation of Ws evolved from the best sets of Ws of the initial generation, will be used and evaluated. These steps will be iterated until balance criteria are satisfied, e.g., none of P values from Kolmogorov–Smirnov test for e (x) is statistically significant. How the new Ws evolve from the old Ws involves a derivative-based, quasi-Newton parameter optimization method described by Mebane and Sekhon (Sekhon & Mebane, 1998) . The generation of W trials evolves towards those containing, on average, better Ws and asymptotically converges towards the optimal solution: the one which minimizes the loss function.

The package GenMatch in R was developed to conduct GM.

Simulations by Diamond and Sekhon(Diamond & Sekhon, In Press) considered seven matching scenarios which shared the outcome **Y**s as the linear combinations of treatment T and confounders but differed in the degree of linearity and additivity in the relationship between confounders and observed group membership. GM and PSM were compared in the seven scenarios, across which PSM always estimated e (x) as the linear combination of ***X*** that

influenced group assignments. Results showed that only when estimated e (x) in PSM was true chance of being assigned to treated group, estimated treatment effect after PSM had smaller absolute bias than the one after GM, and that in all seven situations GM led to smaller root mean square error (RMSE) of effect estimate and less significant P values in paired t-tests and KS-tests.

Coarsened Exact Matching

Both PSM and GM are matching on a scalar balancing score—unidimensional matching, While Coarsen Exact Matching (CEM) is matching on a vector of balancing scores, meaning a vector of coarsened $\boldsymbol{X}$. As discussed before, unidimensional matching tends to focus on obtaining univariate balance on the means of covariates. Mean balances of covariates might not remove bias in estimated effect due to imbalances on interactions, non-linear functions of $\boldsymbol{X}$ (Iacus et al., In Press).

Coarsened Exact Matching belongs to a multidimensional matching family, named Monotonic Imbalance Bounding (MIB). Besides dimensionality, the second biggest difference between MIB and matching on e(x) or on GDM is that MIB methods specify how covariates differ across groups before matching, instead of merely taking those difference as a post-matching result as in PSM or GM. MIB methods select units satisfying a series of conditions,

$$\begin{cases} D\left(f_1\left(x_{m_{T(\pi)}}\right), f_1\left(x_{m_{C(\pi)}}\right)\right) \leq \Upsilon_1\left(\pi_1\right) \\ \vdots \\ D\left(f_k\left(x_{m_{T(\pi)}}\right), f_k\left(x_{m_{C(\pi)}}\right)\right) \leq \Upsilon_k\left(\pi_k\right) \end{cases}$$

These inequalities state that, in every dimension of Xs, the distance D between the function f (•) of $\boldsymbol{X}$ from a treated and the function f (•) of $\boldsymbol{X}$ from its matched control unit should be smaller than a monotonically increasing function $\gamma\ (\pi)$, which means that if ϵ >0,

$$D\left(f_j\left(x_{m_{T(\pi)}}\right), f_j\left(x_{m_{C(\pi)}}\right)\right) \leq \gamma_j(\pi - \epsilon) < \gamma_j(\pi), \text{ j} = 1, \ldots, \text{k}.$$

$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots \pi_k)$ is a vector or matrix of tuning parameters that researchers can specify before matching to control the confounder difference in matched samples. For instance, let the first X be Age, then the original ages of units can be divided into classes with interval of 15 : (≤15), (16–30), (31–45), (46–60), … , where $\gamma_1(\pi_1) = \pi_1$=15, so when matching samples within the same age class, the maximum age difference between a treated unit and its matched a control unit cannot be bigger than 15.

The fact that one can choose to only change one element of π without altering other elements of π indicates that enlarging the tolerance for imbalance on one element of $\boldsymbol{X}$ would not affect the maximum imbalance on the rest elements (Iacus et al., In press). When the $\boldsymbol{X}$ of a treated unit and a control meet the above inequalities, the two units can be matched together.

In the case of CEM, $f(x) = x$, $D\left(f_1\left(x_{m_{T(\pi)}}\right), f_1\left(x_{m_{C(\pi)}}\right)\right) = |x_{m_{T(\pi)}} - x_{m_{C(\pi)}}|$, $\gamma(\pi) = \pi$. In every dimension, CEM divides the range of one element of $\boldsymbol{X}$, denoted as X_i, into V_i different classes whose widths can be equal or unequal, depending on substantive theory of researchers, so

$$\Upsilon_i(\pi_i) = (\Upsilon_{i1}(\pi_{i1}), \Upsilon_{i2}(\pi_{i2}) \ldots \Upsilon_{iV_i}(\pi_{iV_i})).$$

$\pi_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iV_i})$, its elements may or may not be equal to others. Using the previous example, this time the original ages of units, X_1 may be coarsened into classes of different intervals: (≤15), (16–20), (21–35), (36–45), (45-65), (≥66). Compared with the previous equal-interval classification, matching on age within the later kind of classes may be more theoretically meaningful for researchers who study labor forces. The following is the general procedure for CEM (Iacus, King, & Porro, 2009):

(1) Temporarily coarsen every X_i into several classes, then classes of X_i's form matching bins, strata, or hyper-cuboids.

(2) Sort units into hyper-cuboids according to their original X_i's.

(3) Keep the matched units, and use their original $\boldsymbol{X}$ in additional balance adjustment if necessary.

To control for the bias in estimated treatment effect due to the nonlinear or interaction terms of $\boldsymbol{X}$, unlike PSM and GM, CEM do not require identifying non-first-ordered terms of $\boldsymbol{X}$ which explain outcomes above and beyond the first-order terms. CEM is trying to make the marginal distributions of $\boldsymbol{X}$ in the treated group more similar to those in the control group. Such improvement can simultaneously happen in all dimensions of $\boldsymbol{X}$, and increasing distribution similarity of one X_i, will not influence the levels of similarity already archived on other X_i's, therefore the similarity on the multivariate distribution of $\boldsymbol{X}$ can be monotonically increasing, then covariate imbalance in the means, interaction, nonlinear functions of $\boldsymbol{X}$ across the treated and control groups can be reduced in CEM.

CEM and other MIB methods might face the curse of dimensionality due to sorting units in a high-dimensional space and find few matched observations even when using the most relaxed γ (π) that substantive theory can bear. However, it is the data quality that shall be blamed, not the MIB methods. With the same data, PSM and GM might generate more matched units than MIB, but inferences from these data will need more faith that all functions of *Xs* that would affect group assignment and outcome are already included in propensity score model and the distance matrix, and faith that balance on all these functions has been checked and satisfied.

Covariate Selection and Sensitivity Analysis

Despite the growing popularity of sample matching, relatively little has been written about covariate selection strategies. Most of those rare studies were done for PSM, but their suggestions can still provide some guidelines for GM and CEM. There is agreement that all confounding first-ordered $\boldsymbol{X}$ with their special functions should be controlled by matching. The controversy in practice is on how to deal with model parsimony and the variables only related to either outcome or treatment exposure.

In the applied literature of PSM, covariates selection was often based on a variable's predictive power on samples' observed group assignment, so many analyses chose their final PS model to be the ones that with the highest C statistics(Hong & Yu, 2008; Stürmer et al., 2006; Weitzen, Lapane, Toledano, Hume, & Mor, 2004) However, as discussed before, the final goal of matching is to obtain covariate balance, not model fit of PS model. Rosenbaum (2002) cautioned against reviewing the results of statistical significance tests as a way to select predictors included in the model of e (x). Westreich et al.(Westreich, Cole, Funk, Brookhart, & Stürmer, 2011)concluded that a high c-statistic in the propensity model is neither necessary nor sufficient for control of confounding and may result in less overlap in e (x)s between treated and untreated groups.

The analysis of Robins et al. (Robins, Mark, & Newey, 1992) showed that the asymptotic variance of an estimated effect based on a treatment exposure model is not increased and often decreased as the number of parameters in the exposure model is increased. However, this does not happen to all kinds of variables. Simulations (Brookhart et al., 2006; Robins et al., 1992; Rubin, 1997) suggested that all variables thought to be related to the outcome, whether or not they are related to exposure, should be included in a propensity score mode. Covariates that are

unrelated to the exposure but related to the outcome will increase the precision of the estimated treatment effect without increasing bias, and even if a covariate is theoretically unassociated with treatment exposure, there can be some slight chance relation between the covariate and the exposure for any given realization of a data set. In contrast, including variables that are related to the exposure but not the outcome will decrease the precision of the estimated exposure effect without decreasing bias. In small studies, the addition of these variables removes only a small amount of bias but can strongly decrease the precision of effect estimate.

No hidden bias in matching is a very strong assumption, requiring identifying all confounders. Therefore, post-matching sensitivity analysis should be conducted to assess the validness of this assumption in real data. An approach for such analysis is to assess how much hidden bias due to omitted confounders would need to be present to alter significance of effect estimate. Suppose two units, say, j and k, with the same observed covariate values Xs but different unobserved true propensity score e (x, u), where Us are some unobserved confounders. We define a sensitivity parameter Γ as

$$\frac{1}{\Gamma} \leq \frac{e_j(x, u_j)\{1 - e_j(x, u_j)\}}{e_k(x, u_k)\{1 - e_k(x, u_k)\}} \leq \Gamma$$

.When there is no hidden confounder U, $\Gamma = 1$, Unit j and Unit K would have the same conditional probability of treatment exposure. If $\Gamma = 2$, due to unequal Us, one unit might be twice as likely as another to receive the treatment. A sensitivity analysis could use increasing Γs to adjust the P values or confidence intervals in the significant test for treatment effect on matched samples, and make it more and more difficult to reject null hypothesis. Matched samples with lower risk of hidden bias are assumed to generate effect estimates more robust to enlarging Γs. For more details see Rosenbaum‘s demonstrations (Rosenbaum, 2002). Model-based Bayesian approaches for sensitivity analysis have also been developed (Gustafson,

McCandless, Levy, & Richardson, 2010; McCandless, Gustafson, & Levy, 2008), but the current study doesn't apply them due to their linearity and normality assumptions.

Imbalance Measure

Typically reported imbalance measurements in the matching literature are test statistics for univariate mean difference in covariate between the treated group and the control group. But univariate balance does not necessary mean multivariate balance, and mean balances do not necessarily indicate balance on other moments. The current study uses a multivariate imbalance measure $\mathcal{L}_1$recommended by Iacus, Gary King, and Porro (Iacus et al., In Press).

$\mathcal{L}_1$ represents the distance between the multivariate histograms of $\boldsymbol{X}$. The cell or hyper-cuboid frequency of the multidimensional histogram for one group is the within-group ratio of units with values of $(X_1, X_2, \cdots X_K)$ falling into a defined hyper-cuboid z, which is defined by H (X)= $(l_{1z}, l_{2z}, \cdots l_{kz})$, a vector of coarsened $\boldsymbol{X}$. Let $f_{l_{1z},l_{2z}\ldots l_{kz}}$ be the frequency for a multivariate cell z in the treated group, and $g_{l_{1z},l_{2z}\ldots l_{kz}}$ be its counterpart in the control group. Then

$$\mathcal{L}_1(\mathrm{H}) = \frac{1}{2}\sum_{\mathrm{z}=1}^{\mathrm{m}} \left| f_{l_{1z},l_{2z}\ldots l_{kz}} - g_{l_{1z},l_{2z}\ldots l_{kz}} \right|, \mathrm{z} \epsilon [1, \mathrm{m}].$$

$\mathcal{L}_1(\mathrm{H})$ stresses its dependence on the choice of coarsening levels of $\boldsymbol{X}$.

$\mathcal{L}_1(\mathrm{H})$ ranges from 0 to 1. For a given set of coarsen Xs, if the empirical multivariate distributions of Xs in different groups exactly coincide, $f_{l_{1z},l_{2z}\ldots l_{kz}} = \mathrm{g}_{\mathrm{l_{1z}},\mathrm{l_{2z}}\ldots\mathrm{l_{kz}}}$, $\mathcal{L}_1(H) = 0$; if they are completely separated, $\mathcal{L}_1(H) = 1$; In other cases, let say $\mathcal{L}_1(H) = .4$, then 60% of the area under the two histograms overlap. We expect better matched samples to have a smaller $\mathcal{L}_1(\mathrm{H})$.

CHAPTER 3

METHOD

Empirical Example: Delayed Kindergarten Entry

This study examined the performance of CEM, PSM, and GM in assessing the effects of age of entry into kindergarten on children's reading achievement in kindergarten. In the study, reading achievement of children who are delayed was compared to non-delayed entrants which included early and on-time students.

Data for the study came from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K8) sponsored by the U.S. Department of Education, National Center for Education Statistics (Tourangeau et al., 2009). The ECLS-K8 is following a nationally representative cohort of children from kindergarten through eighth grade. It includes information about children's family background, the nature and quality of the preschool and schools that they attend, and their developmental status in cognitive, physical, social, and emotional domains. The participants were mainly first time kindergarteners who were sampled in 1998-1999 school year and then followed up to the eighth grade. In the current study, children's scores from the first reading Item Response Theory (IRT) assessment at Fall kindergarten in ECLS-K8 were explored utilizing hierarchical linear models (HLM).

Covariate selection for Matching Samples

Since the mid 1900s, a considerable amount of research (Buntaine & Costenbader, 1997; Langer, Kalk, & Searls, 1984a; Stipek & Byler, 2001a) has examined the relationship between

the age of entry in kindergarten and children's school performance in the areas of reading. Also, a body of research has included other child, school, and family factors that relate to age of entry and its effects on children's school performance (Bickel, Zigmond, & Strayhorn, 1991; Langer, Kalk, & Searls, 1984b; May, Kundert, & Brent, 1995; Stipek & Byler, 2001b).

In the beginning of variable screening for matching samples, each available variable that was measured before kindergarten entry and showed to relate to reading achievements or kindergarten entrance in the literature were entered as the second predictor into a sample fixed-effects regression model of the reading theta of Fall kindergarten, where the entrance group membership (delayed/non-delayed) was the first predictor. This was done with Proc Mixed in SAS 9.1 (S.I. Incorporated, 2004). Meanwhile, this variable was used to predict entrance group membership in a simple logistic regression. This was done with Proc Logistic in SAS 9.1. For a continuous variable, its quadratic and interaction terms with categorical variables were also separately entered the above two regressions to assess their association with the dependent variables when the variability explained by their first-order term and entrance groups (in the first regression), or solely by their first-order tem (in the second regression), were partialled out. For a categorical variable, its two-way and three-way interactions with other categorical variables were also subsequently entered into the regression models.

Family background variables as important achievement predictors are the first groups of covariate candidates for sample matching, including socioeconomic status, ethnicity, parental education, primary language used in family communication, and family size (Lee, Burkam, & Economic, 2002; Tazouti et al., 2011), and studies indicated that holding children out of kindergarten is a much more common practice among middle and upper class families ((Bellisimo, Sacks, & Mergendoller, 1995; Graue & DiPerna, 2000).Children's individual

characteristics are the second groups of candidates, including gender, health status, cognitive traits, and temperament (Li-Grining, Votruba-Drzal, Maldonado-Carreño, & Haas, 2010). Bellisimo, et al. (1995) found that socioeconomic status has interacted with gender, in that high socioeconomic status of parents was significantly correlated with holding out boys, and not girls from school; while among low socioeconomic status families no gender differences were found in holding out (Stipek & Byler, 2001b). Children's initial cognitive abilities are correlated with later school performance {{84 Hindman,Annemarie H. 2011}}.

The third group of candidates included parental involvement, child care and preschool experiences (Fantuzzo et al., 2005; Hindman & Morrison, 2011; Stylianides & Stylianides, 2011). Children who attend childcare before kindergarten have more familiarity with cognitive skills and academic setting than those who do not attend. Parents' perception on children's characteristics impacts their decision on delaying kindergarten entry or not (Noel & Newman, 2008). Due to the fact that entry age restriction mostly applies to public kindergarten, School Type was also a covariate candidate that could impact parents' hold-out decision.

The original sample size from ECLS-K8 is 21409, after deleting students that were not first-time kindergarteners and did not change school in 1998 Fall Kindergarten, and students with missing ("REFUSED"," DON'T KNOW", or "NOT ASCERTAINED") values on the Reading IRT assessment of Fall Kindergarten assessment (C1R4RTHT), entry group membership (P1WHENEN_1=1 or 0), and all covariate candidates, the final sample size left for analysis is 14098. Most deleted cases was due to no record of the IRT Reading Theta (3739 cases), or no record of parents' rating on children's characteristics (3312 cases). In both the regression of reading IRT scores and the regression of entry group membership, alpha was set at 0.05 in candidates' regression coefficient tests. The initial variable screening revealed that all first-order

candidates along with some of interaction term and quadratic terms are significant predictors of reading theta. Given the large pre-matching sample size that was not a surprising result. Although matching samples on all those functions of *Xs* that significantly predicted the reading theta can reduce the risk of omitting important confounders, it also resulted in 0 matched samples in CEM and the multivariate imbalance measure $\mathcal{L}_1(\mathrm{H}) = 1$ for all three matching methods compared, indicating the complete discrepancy between the multivariate distribution of *Xs* in treated group and the one in the control group. At the end, only candidates that significantly predicted both reading performance and entry group membership (Table 1) were used to obtain final matched samples. If the interaction term of some variables were confounders in the two regressions, their first-order terms were also used in matching.

From Table 1, the scales of WKPARED, P1EXPECT, and P1HELPAR are ordinal, but they were treated as equal-interval in covariate screening, because that yielded significant regression coefficients and smaller Model BIC in the logistic regression of entry group membership. All statistics in Table 1 were calculated from the original values of candidates. In real matching, the original values of RACE and P1PRIMPK were lightly coarsened to increase the sample sizes in every class of the covariates. Specifically, Hispanic with race specified and Hispanic, no race specified were relabeled as one category, and Asian, Native Hawaiian or other Pacific Islander were grouped together. In P1PRIMPK, Relative care in child's home, "Relative care in another home", "Non-relative care in child's home", and "Non-relative care in another home" were renamed only as "Non-parental in-home care". Before matching, all continuous or equal-interval variables were standardized with mean=0 and variance=1, to eliminate the impact of different measurement scales on the calculation of *e (x)* and GMD. Without standardization, a *X* measured in larger scale will affect *e (x)* and GMD more than other *Xs*, increasing the balance

on this X could improve or worsen balance on other Xs more. All categorical covariates were dummy coded before matching to allow the latter process of softwares.

Only first-order covariates in Table 1 were used in CEM. Two ways of coarsening were tried. One was choosing the cut-points for all continuous covariates automatically according to Sturges' rule, the default rule in cem ( ) function of R 2.10.1. That means every continuous variables were coarsened into 15 equal-interval classes, so six continuous covariates (P1CONTRO, P1SADLON, WKPARED, WKSESL, P1EXPECT, P1HELPAR) and thirty six dummy variables resulted in 41006.25×10^4 hyper-cuboids. The other way of coarsening was selecting two continuous variables to specify cut-points while letting the others be chosen automatically. After exploring the effects of gradually relaxing any two variables' default coarsening on the matched sample size and on the imbalance measure $\mathcal{L}_1(\mathrm{H})$, which was done via relax.cem ( ) function in of R 2.10.1, P1SADLON and P1CONTRO were parent's rating on the child's sadness/loneliness level and self-control . P1CONTRO was only divided into six equal- interval classes, P1SADLON were cut by every 20 percentile, but because its 20 percentile equaled to its 40 percentile, there were only four intervals. The total number of hyper-cuboids in the second coarsening was 4274×10^4.

PSM and GM were conducted using the MatchIt ( ) function of R 2.10.1.. First, only first-order covariates entered matching to obtain matched samples, and then squared terms and interaction terms were introduced. For PSM, every treated unit child was randomly matched to five control units within 0.01 standard deviations of propensity scores. Different treated units can be matched to the same control unit. All units (treated and control) outside the common support or overlap of the propensity score distributions of the two groups (delayed and non-delayed) were discarded. In GM, the population size of Ws was set to be 500 for every generation.

Models for Assessing the effect of delayed kindergarten Entrance

To make the models for estimating treatment effect in the three matched samples much easier to compare, covariates were selected via PROC GLMSELECT in SAS 9.1, using BIC as criteria. All first-order candidates involved in covariate selection for matching were used in this screening. Both stepwise selection and backward selection were applied on the original unmatched 14,098 samples and the six sets of matched samples by the three methods. The entry or removal of a variable depended on whether or not this action reduced the model BIC. After covariates were selected, fixed effect and mixed effect models were tried in PROC MIXED to decide the final effect models, with model BIC and p-values of Covariance Parameter Estimates as criteria ($\alpha = 0.05$).

Simulation Study: Seven Matching Scenarios

This Monte-Carlo simulation setting was the same as the one used by Lee et al. {{154 Lee,Brian K. 2010}}, Diamond and Sekhon (in press) and Setoguchi et al. {{81 Setoguchi,Soko 2008}}. Setoguchi et al. used the binary treatment outcome, while Lee et al. followed by Diamond and Sekhon substituted that outcome with a continuous one. The current study used the continuous outcome.

The data were simulated for one hypothetical cohort studies (n=2000) with a binary exposure A and the effect of treatment exposure $B_A = -0.4$, the continuous outcome Y, and ten covariates ($W_i\ i = 1,2 \cdots 10$). N=2000 was chosen for the computation efficiency of the used computer servers, also for the fact that the treated sample sizes in the evaluation of many educational programs are no more than 1000. Four covariates ($W_1, W_{2,}\ W_3, W_4$) were associated with both A and Y (confounders), three ($W_{5,}\ W_6, W_7$) were associated with the exposure only

(exposure predictors), and three (W_8, W_9, W_{10}) were associated with the outcome only (outcome predictors). Six covariates ($W_1, W_3, W_5, W_6, W_8, W_9$) were binary whereas four (W_2, W_4, W_7, W_{10}) were standard normal random variables with mean=0 and variance =1 . Table 2 shows their intercorrelations.

Seven scenarios were explored. They shared the same 10 covariates for units with the same ID but differed in the degree of linearity and additivity in the true propensity score model—the chance of exposure:

Scenario A: additivity and linearity (first-order terms only)

Scenario B: mild non-linearity (one quadratic term)

Scenario C: moderate non-linearity (three quadratic terms)

Scenario D: mild non-additivity (three two-way interaction terms)

Scenario E: mild non-additivity and non-linearity (three two-way interaction terms and one quadraticterm)

Scenario F: moderate non-additivity (ten two-way interaction terms)

Scenario G:moderate non-additivity and non-linearity (ten two-way interaction terms and three quadratic terms)

The outcome Ys were generated as a linear combination of observed group membership, confounders, and outcome predictors with fixed coefficients across scenarios. Across scenarios, all three matching methods only matched samples on the first-order $W_1 \sim W_7$. Once matched samples were obtained, the outcomes were regressed on group membership only. To evaluate the three matching methods, the average bias in effect estimate was calculated as $\left|\frac{\overline{B_A\prime} - (-0.4)}{(-0.4)}\right|$, the average absolute percentage difference between the regression coefficients of exposure and the true exposure effect over 1000 replications. Appendix A contained R scriptfor this simulation.

As in the empirical study, automatic CEM was implemented, PSM was still a .01-SD clipper matching, and the population size of Ws in every generation of GM was 500.

CHAPTER 4

RESULTS AND DISCUSSION

Empirical Example: Delayed Kindergarten Entry

Figure 1 shows the profiles of L1, multidimensional balance measure of matched samples from CEM, PSM, and GM. Because L1 is subject to the chosen interval width of covariates, the x axis of Figure1 came from random intervals of covariates, a point on the L1 profile lines indicate a L1 measure given certain covariate intervals. The unidimensional balance measure was also reported (Table 3, Figure 2.1 to 2.6).

CEM yielded consistently lower L1s than the other two methods, no matter whether its cut points were automatically chosen (CEM1) or intentionally specified (CEM2). The independent t tests and chi-square tests on its matched samples (Table 3) were all statistically insignificant and had larger p-values then others matching methods across covariates . Also, all density plots of individual covariates (Figure 2.1) seemed to be very close between groups, expect for P1CONTRO in CEM2. The most important advantage of CEM is that every matched treated unit has at least control unit that is similar with it on all covariates that researchers want to control; thus the study's internal validity is enhanced and the problem of extrapolation was completely avoided. However, these came with the shrinkage of samples size from 14098 to 78 in CEM1 and to 189 in CEM2, although medium to high correlations (.25~.85) were found between family SES, parental education level, parents' expectation on the child's education, pre-kindergarten child care, kindergarten type, race and so on. Such smaller sample size may be alternative explanation for the larger p-values in univariate balances test. Meanwhile, the

external validity of the study on kindergarten entry was limited by missing characteristics of matched samples. For example, non-white and Hispanic groups, non-English speaking parents are not representing in the matched samples by CEM. There are two primary reasons for small sample size in CEM. One was that the large numbers of hype-cuboids that the original samples had to be sorted into. The other was that 36 out of 42 covariates were dummy variables. For example, one control unit and one treated unit already have very closer values on 41 covariates, if the 42th covariate is a dummy variable and the two units take on its different values, 0 and 1, they will never be matched together. But if the 41th covariate is continuous, like height, and the two units' difference on height is not extreme, researchers will have more flexibility in their matching decision.

Different from CEM, GM and PSM yielded 1962 to 5030 matched samples. The density plots of first-order covariates in GM and PSM still show high similarity across groups as in CEM. GM generated less matched samples than PSM but outperformed PSM with consistently smaller L1s. In the current case, including squared-terms and interaction terms did not help much in reducing L1 in GM and PSM. Although both PSM and GM used the same model to obtain the undimensional balancing score, GM actually gave up the intent of figuring out "Chance of exposure" because it chooses the model's coefficients directly according to covariate balance instead of success prediction of observed exposure, while the terminology in PSM, like "propensity score", still carry the hope of finding a true model for chance of exposure and tend to misdirect users to focus on the goodness of fit of PS models.

Notice that it is not guarantee that every matched treated unit in GM and PSM has at least one control unit processing similar values of all covariates controlled in matching. For instance, treated unit in GM and PSM may share with its matched control units the same values of most

covariates but not their prekindergarten child care experience, in the whole matched sample the prekindergarten child care experience still shows association with kindergarten entry group membership. Inferences based on such kind of matched units require either the strong belief that the pair of matched units just have different values of the covariate, child care experience in the case, by chance, or the belief that the impact of child care experience on reading performance, is just linear and additive, or the belief that the those interactions already achieve balance in the matched units, or the belief that those interaction are considered in the statistic model for estimating assessing treatment outcomes. Otherwise, the study's internal validity is still threatened by alternative explanation for the reading performance.

Based on L1 profile and sample sizes, samples from CEM2, PSM1, and GM2 were used for assessing the effect of delayed kindergarten entry. Table 4.1 to 4.3 display the coefficients and test statistics of the final selected models.

The effect Model of CEM2 is a fixed effect model, and had fewest covariates primarily due to the smallest samples (Treated: 84, Control: 105) and missing values of covariates. Based on the model, delayed kindergarten entry shows no statistically significant impact on the reading performance in Fall Kindergarten (Estimated group difference on reading= 0.1051, SE= 0.06965, DF=185, P-value=0.1329, Effect Size: Cohen's *d*= 0.214611941). GM2 also had a fixed effect model and statistically insignificant group difference estimate 0.04023 with SE= 0.02283, DF=1951, P-value=0.0781, Cohen's *d*= 0.079451). A 2-level random intercept model was applied to data of PSM1 using RACE as level 2 subjects with no significant level-2 predictor. The 2-level model produced statistically significant estimated group difference on reading= 0.04739 with SE= 0.01758, DF=5012, P-value=0.0071, and Cohen's *d*= 0.095512), but that

might be ascribed to the larger samples size of PSM1, given its estimated group difference was so close to that of GM2.

The absolute effect size and effect estimate of CEM2 departed from those of PSM1 and GM2. But it is not clear in this current analysis which method gives less biased results. First, the true effect is unknown in this empirical example. Second, many values of covariates were not represented in matched units of CEM1 as were those of PSM1 and GM2; therefore the population they can be generalized to might be different.

Simulation Study: Seven Matching Scenarios

Table 5 displays the average bias of estimated effect and the standard deviation of these biases across the 1000iterations. Biases were calculated as the absolute percentage differences from the true treatment effect of -0.4. Across all seven scenarios when sample size=2000, CEM consistently had the smallest average estimate bias between 6.46% (0.026) and 8.53% (0.0341), but the biases' standard deviations were between 0.038 and 0.041,the largest among the three methods. GM where Ws' population size=500, consistently had the second smallest average bias ranging from 8.47% (0.034) to 17.58% (0.070) and the second largest biases' standard deviations from 0.029 to 0.032 . PSM performed the worst with Bias between 21.40% (0.086) and 34.0% (0.136) , however, its biases were slightly more stable than those of the above two, ranging from 0.026and 0.030. In general, as non-additivity and non-linearity increased in the true propensity score model, all three methods generated larger and more varying bias.

CHAPTER 5

CONCLUSION

Compared with Propensity Score Matching and Genetic Matching, Coarsened Exact Matching shows better properties in achieving the multidimensional and unidimensional balance of covariates and in reduction in estimation bias. Therefore it is preferable when its matched sample is still representative of the population the study intends to be generalized to. Even for a nation-wide and carefully collected data like ECLSK, the curse of dimensionality is still quite a concern in CEM. But the concern could be less for data where continuous covariates take up larger share of covariates, especially in social science field.

In the cases where extrapolation level and modeling assumptions were still acceptable and CEM resulted in too small and limited samples, GM and PSM can be considered if collecting more data is not a practical option. On the algorithm, Genetic Matching could be a substitute for Propensity Score Matching. Matching samples with the same chance of exposure is to mimic the random selection process, where the chance of exposure is the same for every unit and the covariate balance is likely achieved automatically. The final goal of matching is covariate balance, while estimated chance of exposure can be just a byproduct in this process. It is not to say that since the fit of estimated chance of exposure is not sufficient for good matching, the considerations on the factors or mechanism that impact samples' observed group membership is not necessary. On the contrary, those considerations still provide distinctly important insight on finding confounding covariates.

As Diamond and Sekhon (Diamond & Sekhon, In Press) pointed out, matching is a case where computational power and machine-learning algorithms may help. There is little reason for a human to try the multitude of possible models or covariate cut-points to achieve balance when a computer can do this systematically and faster. Further study regarding the incorporation ofCoarsened Exact Matching with a genetic algorithm where multidimensional balance measure of covariates are cut-point selecting criteria, may be very useful.

REFERENCES

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica, 74*(1), 235-267. doi:http://www.econometricsociety.org

Bellisimo, Y., Sacks, C. H., & Mergendoller, J. R. (1995). Changes over time in kindergarten holding out: Parent and school contexts. *Early Childhood Research Quarterly, 10*(2), 205-22.

Bickel, D. D., Zigmond, N., & Strayhorn, J. (1991). Chronological age and entrance to first grade: Effects on elementary school success. *Early Childhood Research Quarterly, 6*(2), 105-17.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*(12), 1149-1156.

Buntaine, R. L., & Costenbader, V. K. (1997). The effectiveness of a transitional prekindergarten program on later academic achievement. *Psychology in the Schools, 34*(1), 41-50.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*(2), 295-313.

Cochran, W. G., & Cox, G. M. (1957). *Experimental designs* John Wiley & Sons Inc., New York & Chapman & Hall Ltd., London.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., . . . Knaus, W. A. (1996). The effectiveness of right heart catheterization in the initial care of critically III patients. *JAMA: The Journal of the American Medical Association, 276*(11), 889-897. doi:10.1001/jama.1996.03540110043030

Cox-Edwards, A., & Rodriguez-Oreggia, E. (2009). Remittances and labor force participation in mexico: An analysis using propensity score matching. *World Development, 37*(5), 1004-1014. doi:http://www.elsevier.com/wps/find/journaldescription.cws_home/386/description#description

Czajka, J. L., Sharon M. Hirabayashi, Little, R. J. A., & Rubin, D. B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business & Economic Statistics, 10*(2), pp. 117-131.

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

Diamond, A., & Sekhon, J. S. (In Press). *Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies.* Unpublished manuscript. Retrieved 03/10/2011, from http://sekhon.berkeley.edu/papers/GenMatch.pdf

Earle, C. C., Tsai, J. S., Gelber, R. D., Weinstein, M. C., Neumann, P. J., & Weeks, J. C. (2001). Effectiveness of chemotherapy for advanced lung cancer in the elderly: Instrumental variable and propensity analysis. *Journal of Clinical Oncology, 19*(4), 1064-1070.

Fantuzzo, J. W., Rouse, H. L., McDermott, P. A., Sekino, Y., Childs, S., & Weiss, A. (2005). Early childhood experiences and kindergarten success: A population-based study of a large urban setting. *School Psychology Review, 34*(4), 571-588.

Foody, J. M., Cole, C. R., Blackstone, E. H., & Lauer, M. S. (2001). A propensity analysis of cigarette smoking and mortality with consideration of the effects of alcohol. *The American Journal of Cardiology, 87*(6), 706-711. doi:DOI: 10.1016/S0002-9149(00)01487-9

Graue, M. E., & DiPerna, J. (2000). Redshirting and early retention: Who gets the "gift of time" and what are its outcomes? *American Educational Research Journal, 37*(2), 509-534.

Gustafson, P., McCandless, L. C., Levy, A. R., & Richardson, S. (2010). Simplified bayesian sensitivity analysis for mismeasured and unobserved confounders. *Biometrics, 66*(4), 1129-1137. doi:10.1111/j.1541-0420.2009.01377.x

Hahs-Vaughn, D., & Onwuegbuzie, A. J. (2006). Estimating and using propensity score analysis with complex samples. *Journal of Experimental Education, 75*(1), 31-65.

Heckman, J. (1974). Shadow prices, market wages, and labor supply. *Econometrica, 42*(4), pp. 679-694.

Heckman, J. J., Urzua, S., & Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics, 88*(3), 389-432. doi:10.1162/rest.88.3.389

Heckman, J. J., & Vytlacil, E. J. (2007). Chapter 71 econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to

evaluate social programs, and to forecast their effects in new environments. In James J. Heckman and Edward E. Leamer (Ed.), *Handbook of econometrics* (pp. 4875-5143) Elsevier. doi:DOI: 10.1016/S1573-4412(07)06071-0

Hindman, A. H., & Morrison, F. J. (2011). FAMILY INVOLVEMENT AND EDUCATOR OUTREACH IN HEAD START: Nature, extent, and contributions to early literacy skills. *Elementary School Journal, 111*(3), 359-386.

Hong, G., & Yu, B. (2008). Effects of kindergarten retention on children's social-emotional development: An application of propensity score method to multivariate, multilevel data. *Developmental Psychology, 44*(2), 407-421. doi:10.1037/0012-1649.44.2.407; 10.1037/0012-1649.44.2.407.supp (Supplemental)

Iacus, S. M., King, G., & Porro, G. (2009). CEM: Software for coarsened exact Matching .*30*(9)

Iacus, S. M., King, G., & Porro, G. (In Press). *Multivariate matching methods that are monotonic imbalance bounding.* Unpublished manuscript.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics, 86*(1), 4-29. doi:10.1162/003465304323023651

Johnson, M. L., Bush, R. L., Collins, T. C., Lin, P. H., Liles, D. R., Henderson, W. G., . . . Petersen, L. A. (2006). Propensity score analysis in observational studies: Outcomes after

abdominal aortic aneurysm repair. *The American Journal of Surgery, 192*(3), 336-343. doi:DOI: 10.1016/j.amjsurg.2006.03.009

Langer, P., Kalk, J. M., & Searls, D. T. (1984a). Age of admission and trends in achievement: A comparison of blacks and caucasians. *American Educational Research Journal, 21*(1), 61-78.

Langer, P., Kalk, J. M., & Searls, D. T. (1984b). Age of admission and trends in achievement: A comparison of blacks and caucasians. *American Educational Research Journal, 21*(1), 61-78.

Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Review of Economics and Statistics, 84*(2), 205-220. doi:10.1162/003465302317411488

Lee, V. E., Burkam, D. T., & Economic, P. I. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*

Li-Grining, C., Votruba-Drzal, E., Maldonado-Carreño, C., & Haas, K. (2010). Children's early approaches to learning and academic trajectories through fifth grade. *Developmental Psychology, 46*(5), 1062-1077. doi:10.1037/a0020066

Liu, X., & Lynch, L. (2011). Do agricultural land preservation programs reduce farmland loss? evidence from a propensity score matching estimator. *Land Economics, 87*(2), 188-201.

Lytle, B. W., Blackstone, E. H., Loop, F. D., Houghtaling, P. L., Arnold, J. H., Akhrass, R., . . . Cosgrove, D. M. (1999). Two internal thoracic artery grafts are better than one. *The Journal of Thoracic and Cardiovascular Surgery, 117*(5), 855-872. doi:DOI: 10.1016/S0022-5223(99)70365-X

May, D. C., Kundert, D. K., & Brent, D. (1995). Does delayed school entry reduce later grade retentions and use of special education services? *Remedial & Special Education, 16*(5), 288.

McCandless, L. C., Gustafson, P., & Levy, A. R. (2008). A sensitivity analysis using information about measured confounders yielded improved uncertainty assessments for unmeasured confounding. *Journal of Clinical Epidemiology, 61*(3), 247-255. doi:10.1016/j.jclinepi.2007.05.006

Mendola, M. (2007). Agricultural technology adoption and poverty reduction: A propensity-score matching analysis for rural bangladesh. *Food Policy, 32*(3), 372-393. doi:DOI: 10.1016/j.foodpol.2006.07.003

Noel, A. M., & Newman, J. (2008). Mothers' plans for children during the kindergarten hold-out year. *Early Child Development & Care, 178*(3), 289-303. doi:10.1080/03004430600800022

Peterson, E. D., Roe, M. T., Mulgund, J., DeLong, E. R., Lytle, B. L., Brindis, R. G., . . . Ohman, E. M. (2006). Association between hospital process performance and outcomes among patients with acute coronary syndromes. *JAMA: The Journal of the American Medical Association, 295*(16), 1912-1920. doi:10.1001/jama.295.16.1912

Porro, G., & Iacus, S. M. (2009). Random recursive partitioning: A matching method for the estimation of the average treatment effect. *Journal of Applied Econometrics, 24*(1), 163-185. doi:10.1002/jae.1026

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association, 53*(284), pp. 873-880.

Robins, J. M., Mark, S. D., & Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics, 48*(2), 479-495.

Rosenbaum, P. R. (2002). *Observational studies* Second edition; Series in Statistics; New York and Heidelberg:; Springer.

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

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling models that incorporate the propensity score. *American Statistician, 39*(1), 33.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers, 3*(2), pp. 135-146.

Rubin, D. B. (1972). *Estimating causal effects of treatments in experimental and observational studies*

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics, 29*(1), pp. 185-203.

Rubin, D. B. (1976). Multivariate matching methods that are equal percent bias reducing, II: Maximums on bias reduction for fixed sample sizes. *Biometrics, 32*(1), pp. 121-132.

Rubin, D. B. (1997). Estimating causal effects form large data sets using propensity scores. *Annals of Internal Medicine, 127*(8), 757-763.

Rubin, D. B., & Stuart, E. A. (2006). Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *The Annals of Statistics, 34*(4), 1814-1826.

Rubin, D. B., & Thomas, N. (1992a). Affinely invariant matching methods with ellipsoidal distributions. *The Annals of Statistics, 20*(2), pp. 1079-1093.

Rubin, D. B., & Thomas, N. (1992b). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika, 79*(4), pp. 797-809.

S.I. Incorporated, . (2004). *SAS 9.1. SAS/Stat user's guide, version 9.1*. Cary, NC, USA:

Sekhon, J. S., & Mebane, W. R. (1998). Genetic optimization using derivatives. *Political Analysis, 7*(1), 187-210. doi:10.1093/pan/7.1.187

Sobel, M. E. (1996). An introduction to causal inference. *Sociological Methods & Research, 24*(3), 353-379. doi:10.1177/0049124196024003004

Stipek, D., & Byler, P. (2001a). Academic achievement and social behaviors associated with age of entry into kindergarten. *Journal of Applied Developmental Psychology, 22*(2), 175-189. doi:10.1016/S0193-3973(01)00075-2

Stipek, D., & Byler, P. (2001b). Academic achievement and social behaviors associated with age of entry into kindergarten. *Journal of Applied Developmental Psychology, 22*(2), 175-189. doi:10.1016/S0193-3973(01)00075-2

Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., & Schneeweiss, S. (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*(5), 437-447.

Stylianides, A. J., & Stylianides, G. J. (2011). A type of parental involvement with an isomorphic effect on urban Children's mathematics, reading, science, and social studies achievement at kindergarten entry. *Urban Education, 46*(3), 408-425. doi:10.1177/0042085910377605

Tazouti, Y., Viriot-Goeldel, C., Matter, C., Geiger-Jaillet, A., Carol, R., & Deviterne, D. (2011). French nursery schools and german kindergartens: Effects of individual and contextual variables on early learning. *European Journal of Psychology of Education - EJPE (Springer Science & Business Media B.V.), 26*(2), 199-213. doi:10.1007/s10212-010-0043-4

Tourangeau, K., Nord, C., Le, T., Sorongon, A. G., Najarian, M., & National Center for, E. S. (2009). *Early childhood longitudinal study, kindergarten class of 1998-99 (ECLS-K): Combined user's manual for the ECLS-K eighth-grade and K-8 full sample data files and electronic codebooks. NCES 2009-004.*National Center for Education Statistics.

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., & Mor, V. (2004). Principles for modeling propensity scores in medical research: A systematic literature review. *Pharmacoepidemiology and Drug Safety, 13*(12), 841-853.

Westreich, D., Cole, S. R., Funk, M. J., Brookhart, M. A., & Stürmer, T. (2011). The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiology and Drug Safety, 20*(3), 317-320. doi:10.1002/pds.2074

Table 1

Selected Covariates for Matching

Variable Name in ECLS-K8 Dataset	Values Used in Covariate Screening	Applied in Method	Model BIC and Coefficient P value in Regression 1	Coefficient P value in Regression 2
GENDER	1 = Male 2 = Female	CEM PSM GM	24056.4 <.0001	<.0001
RACE @ (Child composite race)	1 = White, non-Hispanic 2 = Black or African American, non-Hispanic 3 = Hispanic, race specified 4 = Hispanic, no race specified 5 = Asian 6 = Native Hawaiian or other Pacific Islander 7 = American Indian or Alaskan Native 8 = More than one race, non-Hispanic	CEM PSM GM	23318.4 <.0001	<.0001
WKPARED ! (Highest level of education for the child's parents or nonparent guardians who reside in the household.)	1 = 8th grade or below 2 = 9th to 12th grade 3 = High school diploma/equivalent 4 = Voc/Tech program 5 = Some college 6 = Bachelor's degree 7 = Graduate/professional school/no degree 8 = Master's degree 9 = Doctorate or professional degree	CEM PSM GM	21703.3 <.0001	0.0008
WKSESL (Socioeconomic status (SES) scale)	Continuous (Higher values indicate higher SES)	CEM PSM CM	21453.2 <.0001	0.0053
WKSESL* RACE		PSM CM	21227.3 <.0001	0.0150
P1EXPECT ! (Parents' expectation for children's highest education level)	1 = To receive less than high school diploma 2 = To graduate from high school 3 = To attend two or more years of college	CEM PSM CM	23656.1 <.0001	0.0213

	4 = To finish a 4-or-5-year college degree 5 = To earn a master's degree or equivalent 6 = To get ph.d., md, or other higher degree			
P1LANGUG (Language (s) spoken most often at home by the parent (s)/guardian (s) in the household)	1 = Both only speak English language 2 = 1 (of 2) parents mainly speaks a non-English language 3 = Both only speak a non-English language	CEM PSM CM	24110.9 <.0001	0.0006
P1HFAMIL	1 = Two parents and sibling (s) 2 = Two parents, no siblings 3 = One parent and sibling (s) 4 = One parent, no siblings 5 = Other	CEM PSM CM	23604.5 <.0001	0.0010
P1PRIMPK@ (Primary, nonparental arrangement in which the child spent the most hours per week during the year before kindergarten)	0 = No non-parental care 1 = Relative care in child's home 2 = Relative care in another home 3 = Non-relative care in child's home 4 = Non-relative care in another home 5 = Head Start program 6 = Center-based program 7 = 2 or more programs 8 = Location of care varies	CEM PSM CM	22934.1 <.0001	0.0009
P1HSPREK (The Child attended Head Start before Kindergarten)	1 = Yes 2 or -1 = No	CEM PSM CM	23427.3 <.0001	<.0001
P1HSPREK* P1EXPECT !		PSM CM	23010.6 0.0023	0.0212
P1DISABL	1 = Yes 2 or -1= No	CEM PSM CM	24034.2 <.0001	<.0001
P1SADLON (Parent's rating on the child's Sadness/Loneliness level)	Continuous	CEM PSM CM	24138.1 0.0002	<.0001
P1SADLON*P1SADLON	Continuous	PSM CM	24076.0 <.0001	0.0284
P1CONTRO (Parents' rating on the child's self-control)	Continuous	CEM PSM CM	23804.5 <.0001	0.9442
P1CONTRO*P1DISABL		PSM CM	23728.6 0.0297	0.0008
P1HELPAR !	1 = Not at all 2 = Once or twice a week	CEM	24140.8	0.5866

(Parent's frequency of helping the child to do arts and crafts)	3 = 3 to 6 Times a week 4 = Everyday	PSM CM	0.0006	
P1HELPAR!*GENDER		PSM CM	24053.8 0.0006	0.0461
P1SINGSO (Parent's frequency of singing songs with the child)	1 = Not at all 2 = Once or twice a week 3 = 3 to 6 Times a week 4 = Everyday	CEM PSM CM	24087.6 <.0001	0.0023
S2KSCTYP (Type of School)	1-3 = Non Public 4 = Public	CEM PSM CM	23465.9 <.0001	<.0001
S2KSCTYP* P1EXPECT !		PSM CM	23071.1 <.0001	0.0033

Note:
! ------ The variable was used as a continuous variable in actual matching
@ ------ The original values of the variable was coarsened in actual matching

Table 2

Correlations of Simulated Covariates

		Confounders				Exposure Predictors			Outcome Predictor		
		w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}
Confounders	w_1	1	0	0	0						
	w_2	0	1	0	0						
	w_3	0	0	1	0						
	w_4	0	0	0	1						
Exposure	w_5	.2	0	0	0	1					
Predictors	w_6	0	.9	0	0	0	1				
	w_7	0	0	0	0	0	0	1			
Outcome Predictor	w_8	0	0	.2	0	0	0	0	1		
	w_9	0	0	0	.9	0	0	0	0	1	
	w_{10}	0	0	0	0	0	0	0	0	0	1

Table 3

Characteristics of Empirical data

		ORIGINAL		CEM 1		CEM 2		PSM 1	
		Control Units	Treated Units	Control Units	Treated Units	Control Units	Treated Units	Control Units	Treated Units
MALE (Gender=1)	Proportion	0.49	0.61	0.60	0.61	0.61	0.60	0.60	0.61
	Sample Size	6463	622	24	23	64	50	2420	615
	χ^2 test P	<0.0001**		0.96		0.8419		0.5643	
WHITE, NON-HISPANIC	Proportion	0.62	0.71	1.00	1.00	0.99	0.99	0.69	0.71
	Sample Size	8117	718	40	38	104	83	2790	713
	χ^2 test P	<0.0001**				0.8737		0.3418	
AFRICAN AMERICAN	Proportion	0.15	0.11	0	0	0	0	0.11	0.11
	Sample Size	1979	107					452	107
	χ^2 test P	<0.0001**						0.5903	
HISPANIC	Proportion	0.13	0.10	0	0	0	0	0.11	0.10
	Sample Size	1669	101					448	101
	χ^2 test P	0.01**						0.3198	
ASIAN OR PACIFIC ISLANDER	Proportion	0.05	0.04	0	0	0	0	0.03	0.03
	Sample Size	704	40					135	33
	χ^2 test P	0.0496						0.9063	
AMERICAN INDIAN OR ALASKA NATIVE	Proportion	0.02	0.03	0	0	0.01	0.01	0.03	0.03
	Sample Size	218	28			1	1	124	33
	χ^2 test P	0.0101				0.8737		0.7457	
PARENTS ONLY SPEAK ENGLISH	Proportion	0.90	0.93	1.00	1.00	1.00	1.00	0.92	0.93
	Sample Size	11741	947	40	38	105	84	3717	940
	χ^2 test P	0.0001**						0.2473	
BOTH ONLY SPEAK NON-ENGLISH	Proportion	0.07	0.04	0	0	0	0	0.06	0.04
	Sample Size	971	45					222	45
	χ^2 test P	0.0004						0.1866	
2 PARENTS PLUS	Proportion	0.67	0.72	0.98	0.97	0.93	0.91	0.71	0.72
	Sample Size	8717	726	39	37	98	77	2850	721

SIBLINGS FAMILY	χ^2 test P	0.0010**		0.97		0.6638		0.5974	
2 PARENTS NO SIBLING FAMILY	Proportion	0.10	0.09	0.025	0.026	0.04	0.05	0.09	0.09
	Sample Size	1287	93	1	1	4	4	359	93
	χ^2 test P	0.4991		0.97		0.7466		0.7486	
1 PARENT PLUS SIBLINGS FAMILY	Proportion	0.15	0.13	0	0	0.01	0.01	0.13	0.13
	Sample Size	1989	128			1	1	538	128
	χ^2 test P	0.0277*				0.8737		0.5886	
1 PARENT NO SIBLING FAMILY	Proportion	0.07	0.04	0	0	0.02	0.02	0.05	0.04
	Sample Size	886	44			2	2	204	44
	χ^2 test P	0.0027**				0.8212		0.3619	
PRIMARLY PARENTAL CARE	Proportion	0.16	0.19	0	0	0.03	0.04	0.19	0.19
	Sample Size	2132	194			3	3	761	193
	χ^2 test P	0.0182*				0.7808		0.8432	
PRIMARLY NON-PARENTAL IN HOME CARE	Proportion	0.24	0.26	3	3	0.15	0.17	0.25	0.26
	Sample Size	3124	262	0.075	0.0789	16	15	1025	261
	χ^2 test P	0.1534		1.00		0.6290		0.7588	
PRIMARLY HEAD START PROGRAM	Proportion	0.09	0.05	0	0	0	0	0.05	0.05
	Sample Size	1163	53					220	53
	χ^2 test P	<0.0001**						0.8034	
PRIMARLY CENTER-BASED PROGRAM	Proportion	0.46	0.46	0.9250	0.9211	0.82	0.79	0.46	0.45
	Sample Size	5973	461	37	35	86	66	1848	456
	χ^2 test P	0.9317		0.9479		0.5660		0.7342	
ATTEND 2 OR MORE PROGRAMS	Proportion	0.04	0.03	0	0	0	0	0.0338	0.008
	Sample Size	539	31					136	31
	χ^2 test P	0.0992						0.6368	
NOT AT ALL SING TO CHILD	Proportion	0.04	0.04	0	0	0	0	0.044	0.039
	Sample Size	575	40					180	39
	χ^2 test P	0.5035						0.4070	
ONCE OR TWICE A WEEK SING TO CHILD	Proportion	0.22	0.27	0.20	0.21	0.18	0.19	0.26	0.27
	Sample Size	2931	278	8	8	19	16	1043	274
	χ^2 test P	0.0002**		0.91		0.8670		0.3954	
3 TO 6 TIMES A WEEK SING TO CHILD	Proportion	0.28	0.28	0.425	0.4211	0.32	0.31	0.29	0.28
	Sample Size	3678	280	17	16	34	26	1148	280
	χ^2 test P	0.7495		0.97		0.8339		0.6615	
EVERYDAY SING TO CHILD	Proportion	0.45	0.41	0.3750	0.3684	0.50	0.50	0.41	0.41
	Sample Size	5901	415	15	14	54	42	1653	413
	χ^2 test P	0.0109*		0.95		0.9481		0.9886	
EVER ATTEND	Proportion	0.14	0.09	0	0	0	0	0.10	0.09
	Sample Size	1768	90					391	90

HEAD START PROGRAM	χ^2 test P	<.0001**						0.4574	
IN PUBILC KINDERGARTEN	Proportion	0.78	0.72	0.70	0.6842	0.72	0.73	0.73	0.72
	Sample Size	10202	1013	28	26	76	61	2952	726
	χ^2 test P	<.0001**		0.88		0.9709		0.4453	
P1DISABL	Proportion	0.13	0.20	0	0	0.04	0.05	0.17	0.20
	Sample Size	1706	204			4	4	683	197
	χ^2 test P	<.0001**				0.7466		0.0514	
WKPARED	Sample Size	13085	1013	40	38	105	84	4024	1006
	T-test P	0.0008**		0.9622		0.8987		0.4326	
WKSESL	Sample Size	13085	1013	40	38	105	84	4024	1006
	T-test P	0.0056**		0.9739		0.8567		0.4595	
P1EXPECT	Sample Size	13085	1013			105	84	4024	1006
	T-test P	0.0213*		0.9672		1.0000		0.4284	
P1SADLON	Sample Size	13085	1013	40	38	105	84	4024	1006
	T-test P	<.0001**		0.8691		0.1781		0.4871	
P1CONTRO	Sample Size	13085	1013	40	38	105	84	4024	1006
	T-test P	0.9420		0.9215		0.6685		0.9417	
P1HELPAR	Sample Size	13085	1013	40	38	105	84	4024	1006
	T-test P	0.5868		0.9276		0.6368		0.6860	

		PSM1		GM1		GM 2			
		Control Units	Treated Units	Control Units	Treated Units	Control Units	Treated Units		
MALE (Gender=1)	Proportion	0.59	0.61	0.60	0.61	0.60	0.61		
	Sample Size	2284	611	573	622	571	622		
	χ^2 test P	0.1684		0.6428		0.6210			
WHITE, NON-HISPANIC	Proportion	0.69	0.71	0.71	0.71	0.71	0.71		
	Sample Size	2703	709	674	718	683	718		
	χ^2 test P	0.3791		0.9442		0.8104			
AFRICAN AMERICAN	Proportion	0.12	0.11	0.11	0.11	0.11	0.11		
	Sample Size	449	107	104	107	102	107		
	χ^2 test P	0.4559		0.7771		0.9451			
HISPANIC	Proportion	0.11	0.10	0.10	0.10	0.10	0.10		
	Sample Size	421	101	92	101	97	101		
	χ^2 test P	0.5110		0.8375		0.9028			
ASIAN OR	Proportion	0.03	0.03	0.03	0.03	0.03	0.03		

PACIFIC ISLANDER	Sample Size	125	33	29	33	28	33		
	χ^2 test P	0.8889		0.7985		0.6709			
AMERICAN INDIAN OR ALASKA NATIVE	Proportion	0.03	0.03	0.04	0.03	0.04	0.03		
	Sample Size	103	32	35	35	36	35		
	χ^2 test P	0.3411		0.7810		0.7151			
PARENTS ONLY SPEAK ENGLISH	Proportion	0.93	0.93	0.94	0.93	0.95	0.93		
	Sample Size	3619	935	894	947	905	947		
	χ^2 test P	0.5855		0.5078		0.3119			
BOTH ONLY SPEAK NON-ENGLISH	Proportion	0.05	0.05	0.04	0.04	0.03	0.04		
	Sample Size	184	45	35	45	32	45		
	χ^2 test P	0.7601		0.3986		0.2086			
2 PARENTS PLUS SIBLINGS FAMILY	Proportion	0.70	0.72	0.72	0.72	0.73	0.72		
	Sample Size	2722	718	680	726	697	726		
	χ^2 test P	0.2550		0.9945		0.5644			
2 PARENTS NO SIBLING FAMILY	Proportion	0.10	0.09	0.09	0.09	0.09	0.09		
	Sample Size	394	93	86	93	87	93		
	χ^2 test P	0.4367		0.9274		0.9449			
1 PARENT PLUS SIBLINGS FAMILY	Proportion	0.14	0.13	0.13	0.13	0.12	0.13		
	Sample Size	531	127	127	128	119	128		
	χ^2 test P	0.4340		0.6230		0.8929			
1 PARENT NO SIBLING FAMILY	Proportion	0.05	0.04	0.04	0.04	0.04	0.04		
	Sample Size	185	43	42	44	40	44		
	χ^2 test P	0.5432		0.9292		0.8573			
PRIMARLY PARENTAL CARE	Proportion	0.19	0.19	0.19	0.19	0.19	0.19		
	Sample Size	721	192	181	194	181	194		
	χ^2 test P	0.6274		0.9648		0.8931			
PRIMARLY NON-PARENTAL IN HOME CARE	Proportion	0.26	0.26	0.26	0.26	0.26	0.26		
	Sample Size	1014	261	250	262	250	262		
	χ^2 test P	0.9792		0.8089		0.8956			
PRIMARLY HEAD START PROGRAM	Proportion	0.05	0.05	0.05	0.05	0.05	0.05		
	Sample Size	213	52	52	53	51	53		
	χ^2 test P	0.7328		0.8077		0.9232			
PRIMARLY CENTER-BASED PROGRAM	Proportion	0.46	0.45	0.46	0.46	0.46	0.46		
	Sample Size	1785	453	436	461	444	461		
	χ^2 test P	0.7454		0.8468		0.6931			
ATTEND 2 OR MORE PROGRAMS	Proportion	0.03	0.03	0.03	0.03	0.03	0.03		
	Sample Size	118	31	24	31	27	31		
	χ^2 test P	0.9119		0.4762		0.7539			
NOT AT ALL	Proportion	0.04	0.04	0.04	0.04	0.04	0.04		

SING TO CHILD	Sample Size	145	38	38	40	36	40		
	χ^2 test P	0.9129		0.9498		0.8295			
ONCE OR TWICE A WEEK SING TO CHILD	Proportion	0.26	0.27	0.27	0.27	0.26	0.27		
	Sample Size	1022	274	258	278	248	278		
	χ^2 test P	0.4683		0.8985		0.4433			
3 TO 6 TIMES A WEEK SING TO CHILD	Proportion	0.28	0.28	0.27	0.28	0.27	0.28		
	Sample Size	1102	277	251	280	263	280		
	χ^2 test P	0.6971		0.5527		0.9371			
EVERYDAY SING TO CHILD	Proportion	0.42	0.41	0.42	0.41	0.43	0.41		
	Sample Size	1626	412	402	415	410	415		
	χ^2 test P	0.7368		0.5317		0.3992			
EVER ATTEND HEAD START PROGRAM	Proportion	0.09	0.09	0.09	0.09	0.08	0.09		
	Sample Size	349	89	85	90	81	90		
	χ^2 test P	0.9455		0.9552		0.7404			
IN PUBILC KINDERGARTEN	Proportion	0.74	0.73	0.73	0.72	0.74	0.72		
	Sample Size	2900	726	695	729	711	729		
	χ^2 test P	0.2147		0.5284		0.2438			
P1DISABL	Proportion	0.17	0.19	0.19	0.20	0.19	0.20		
	Sample Size	648	194	183	204	184	204		
	χ^2 test P	0.0402*		0.6344		0.6112			
WKPARED	Sample Size	3895	1001	949	1013	957	1013		
	T-test P	0.4151		0.7435		0.9461			
WKSESL	Sample Size	3895	1001	949	1013	957	1013		
	T-test P	0.9957		0.7943		0.7870			
P1EXPECT	Sample Size	3895	1001	949	1013	957	1013		
	T-test P	0.9942		0.6902		0.8370			
P1SADLON	Sample Size	3895	1001	949	1013	957	1013		
	T-test P	0.1336		0.7812		0.1471			
P1CONTRO	Sample Size	3895	1001	949	1013	957	1013		
	T-test P	0.9560		0.8736		0.5431			
P1HELPAR	Sample Size	3895	1001	949	1013	957	1013		
	T-test P	0.6888		0.8487		0.7062			

Table 4.1

Assessing the Effect of Delayed Entry in Matched Data Of CEM2

Effect	Estimate	Error	*df*	F	P>F
Intercept	-0.8441	0.05731			
P1WHENEN_1	Delayed=0 Non-Delayed=-0.1007	0.06868	184	5.53	0.1443
P1CPREK_1	Yes=0 No=-0.2982	0.1268	184	11.2	0.0198
WKSESL	0.1764	0.0527	184	6.07	0.001

Table 4.2

Assessing the Effect of Delayed Entry in Matched Data Of GM2

Effect	Estimate	Error	*df*	F	P>F
Intercept	-1.1202	0.09797	1951		
P1WHENEN_1	Delayed=0 Non-Delayed= -0.04023	0.02283	1951	3.11	0.0781
P1CPREK_1	Yes=0 No=-0.1150	0.1268	1951	19.87	<.0001
S2KSCTYP_1	Yes=0 No=0.1006	0.02711	1951	13.77	0.0002
P1DISABL_1	Yes=0 No=0.1483	0.02889	1951	26.37	<.0001
WKSESL	0.1606	0.01371	1951	137.25	0.001
P1SINGSO_1	-0.0146	0.06194	1951	0.0576	0.8138
P1SINGSO_2	0.01722	0.02866	1951	0.36	0.548
P1SINGSO_3	0.08306	0.02803	1951	8.7616	0.0031
White, non-Hispanic	-0.00393	0.09332	1951	0.0016	0.9664
African American, non-Hispanic	-0.06578	0.09976	1951	0.4356	0.5097
Hispanic, race specified	-0.09663	0.1037	1951	0.8649	0.3517
Hispanic, no race specified	-0.127	0.1086	1951	1.3924	0.2396
Asian	0.09026	0.1128	1951	0.64	0.4236
Pacific Islander	0.1095	0.1542	1951	0.5041	0.4778

American Indian or Alaskan Native	-0.3652	0.1162	1951	9.8596	0.0017
P1READBO	0.044	0.01221	1951	12.97	0.0003
P1LEARN	0.04881	0.01182	1951	17.06	<.0001
P1IMPULS	-0.05453	0.01219	1951	20.02	<.0001

Table 4.3
Assessing the Effect of Delayed Entry in Matched Data Of PSM1

Fixed Effect	**Estimate**	**Error**	**DF**	**F**	**P>F**
Intercept	-1.1618	0.0472			
P1WHENEN_1	Delayed=0 Non-delayed=-0.04739	0.01758	5012	7.26	0.0071
P1CPREK_1	Yes=0 No=-0.1039	0.01584	5012	43	<.0001
S2KSCTYP_1	Yes=0 No= 0.1041	0.01682	5012	38.31	<.0001
GENDER_1	Yes=0 No=0.05086	0.01455	5012	12.22	0.0005
P1HSPREK_1	Yes=0 No=0.09662	0.02617	5012	13.63	0.0002
WKSESL	0.1472	0.008349	5012	310.81	<.0001
P1EXPECT	0.03931	0.007791	5012	25.46	<.0001
P1READBO	0.0424	0.007506	5012	31.91	<.0001
P1LEARN	0.05621	0.007303	5012	59.23	<.0001
P1IMPULS	-0.03459	0.007386	5012	21.94	<.0001
Random Effect	**Estimate**	**Error**	**DF**	**T**	**P>T**
White, non-Hispanic	0.07438	0.03763	5012	1.98	0.0481
African American, non-Hispanic	0.01348	0.04079	5012	0.33	0.7411
Hispanic, race specified	-0.03576	0.04346	5012	-0.82	0.4107
Hispanic, no race specified	-0.09371	0.04475	5012	-2.09	0.0363
Asian	0.176	0.04786	5012	3.68	0.0002
Pacific Islander	-0.01567	0.06259	5012	-0.25	0.8023
American Indian or Alaskan Native	-0.01567	0.05212	5012	-2.41	0.0161
More than one race, non-Hispanic	0.006815	0.05356	5012	0.13	0.8987

Table 5.
Performance of Matching Estimation Methods in the Simulation

		SCENARIO						
	METHOD	**A**	**B**	**C**	**D**	**E**	**F**	**G**
Average Absolute Bias (percent)	CEM	8.53246	7.05615	7.05867	8.43705	6.70642	8.39068	6.46776
	PSM	24.2892	21.5901	21.7248	24.3793	21.4006	25.2059	34.0266
	GM	17.5841	13.0349	9.92120	14.9690	8.57345	14.1142	8.47269
Standard Deviation	CEM	0.03904	0.03997	0.04102	0.03952	0.04136	0.03846	0.04055
	PSM	0.02715	0.02715	0.02655	0.02678	0.02600	0.02566	0.03007
	GM	0.03097	0.03109	0.03176	0.03050	0.03207	0.02923	0.02986

Figure 1. L1 Profiles of Six Matched Datasets

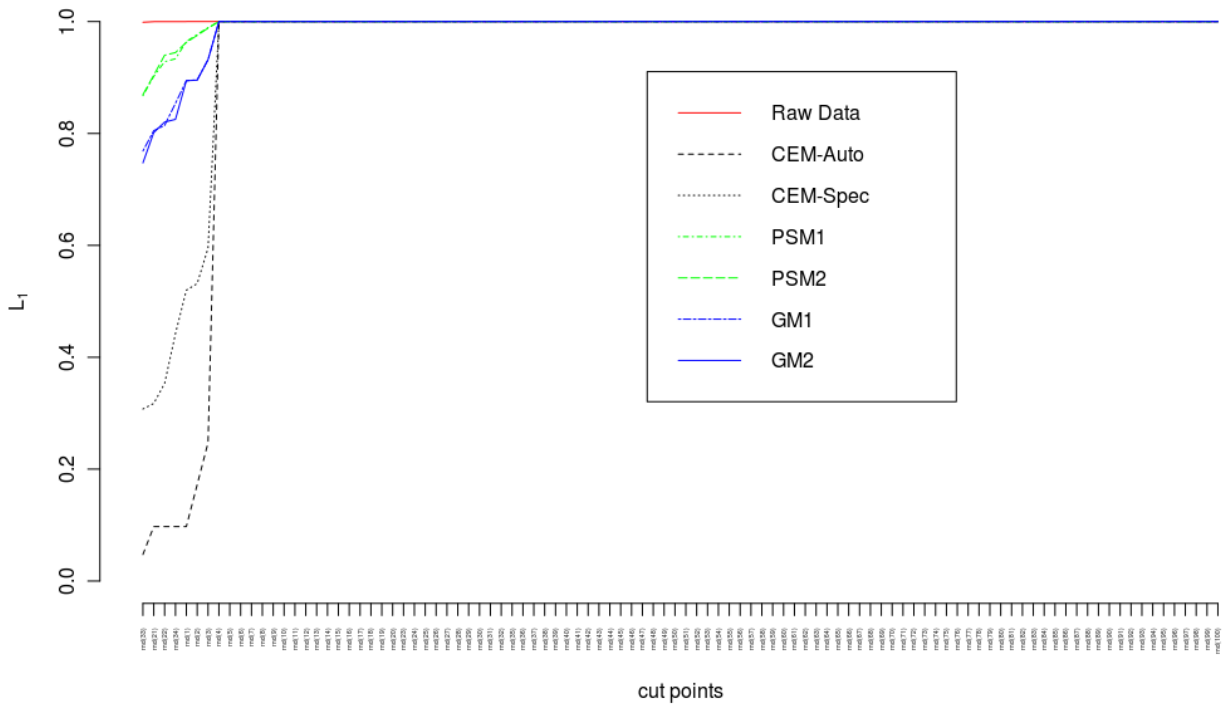


Figure 2.1 Covariate Density By Group-CEM1

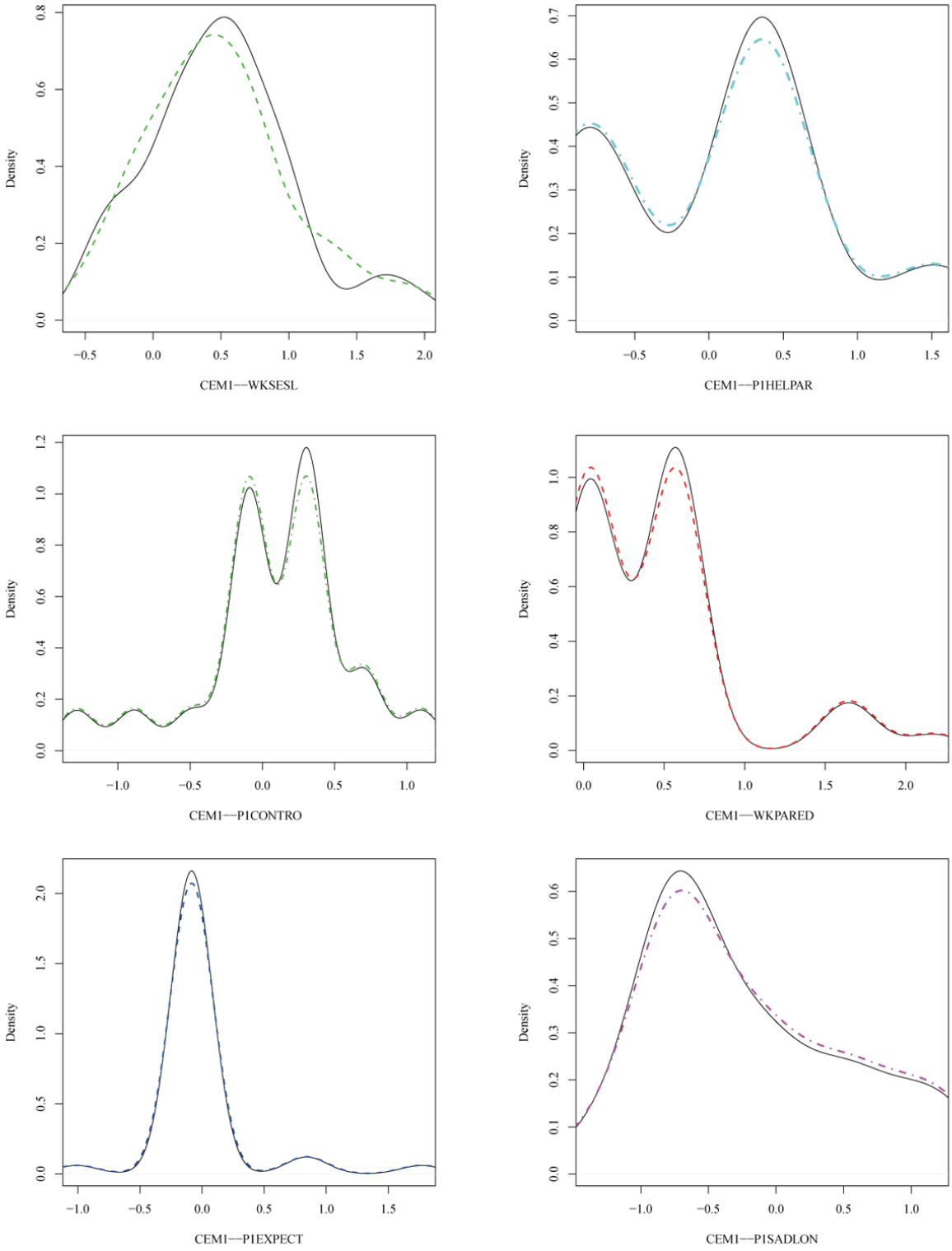


Figure 2.2 Covariate Density By Group-CEM2

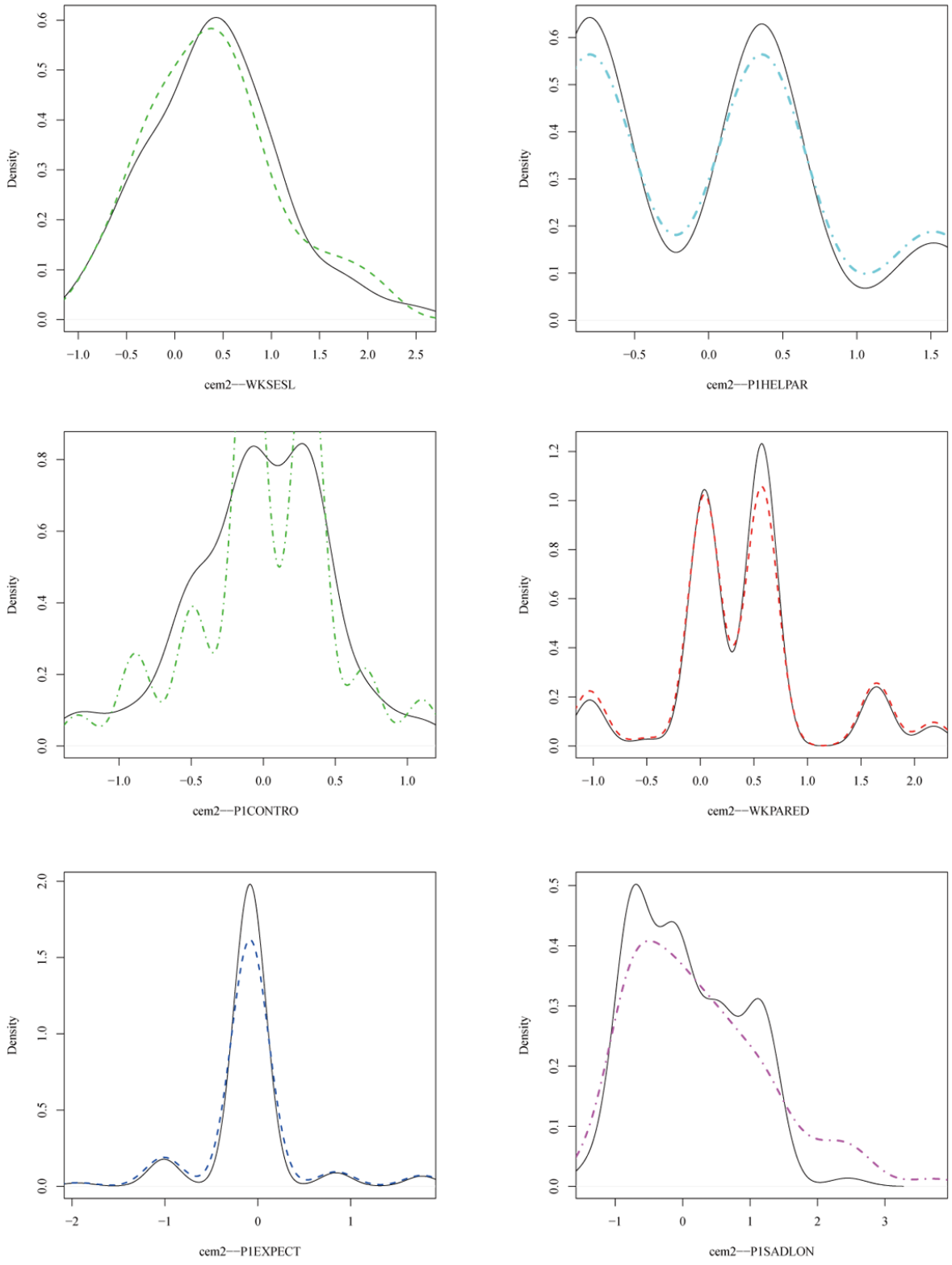


Figure 2.3 Covariate Density By Group-PSM1

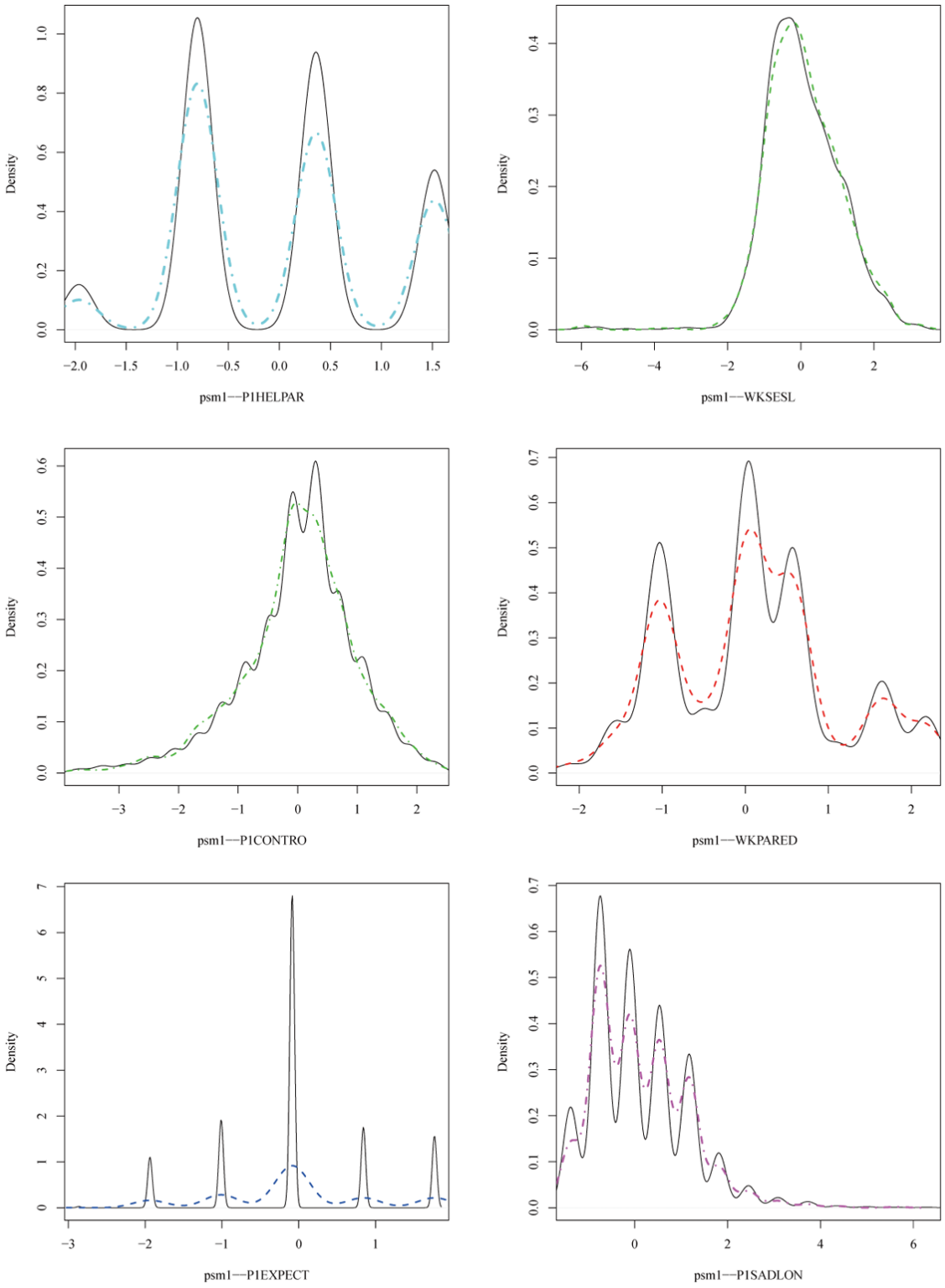


Figure 2.4 Covariate Density By Group-PSM2

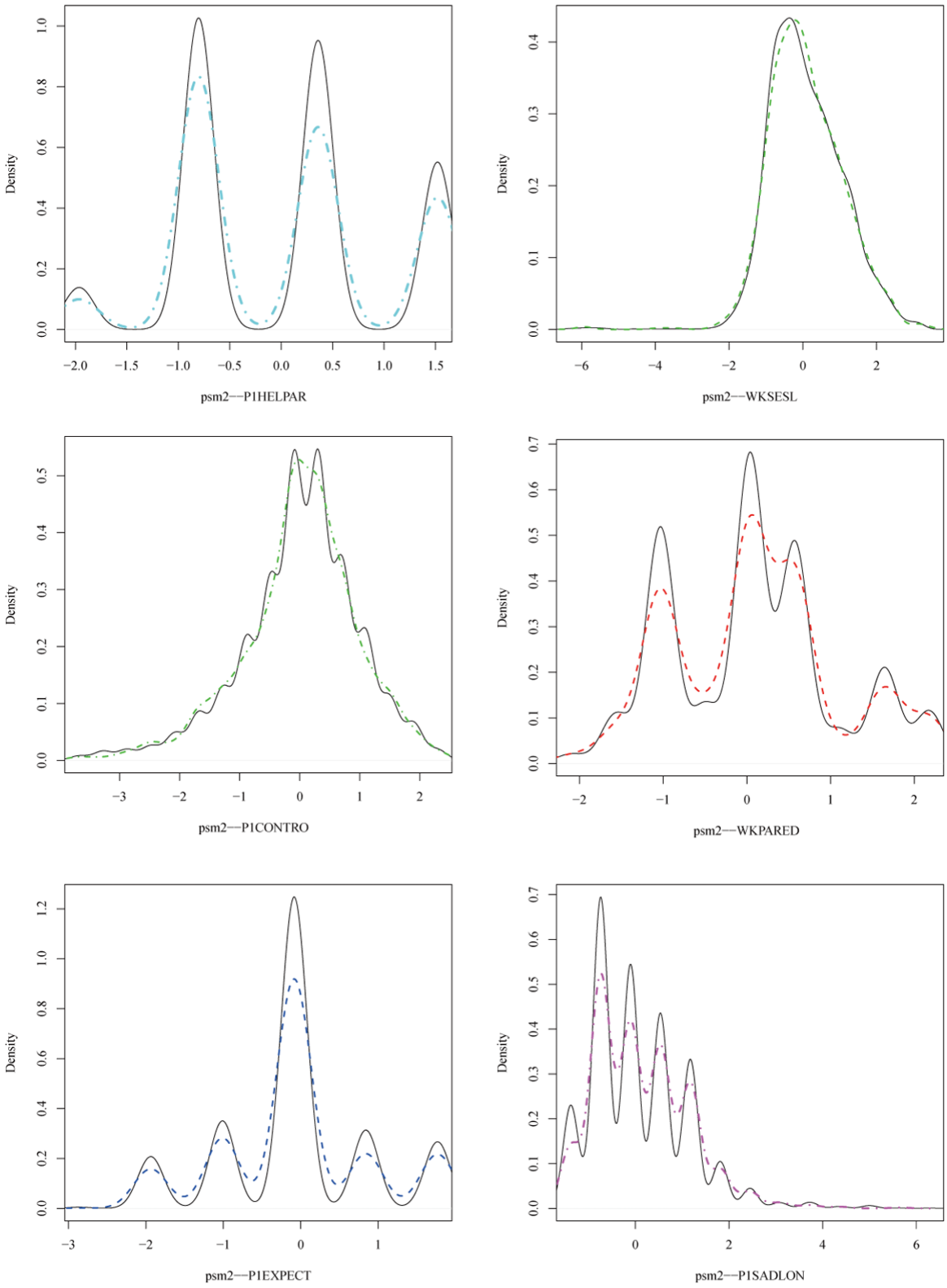


Figure 2.5 Covariate Density By Group-GM1

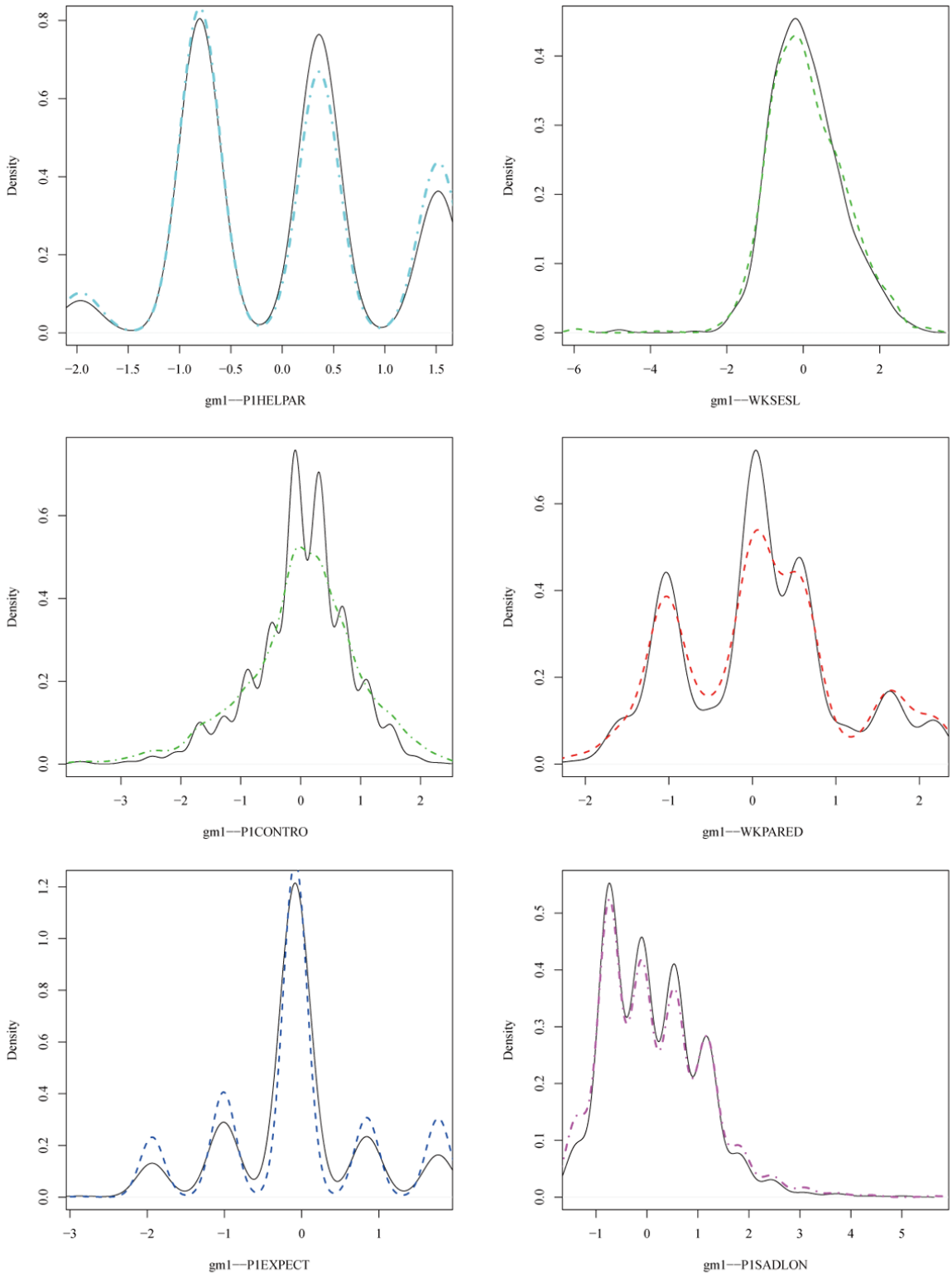


Figure 2.6 Covariate Density By Group-GM2

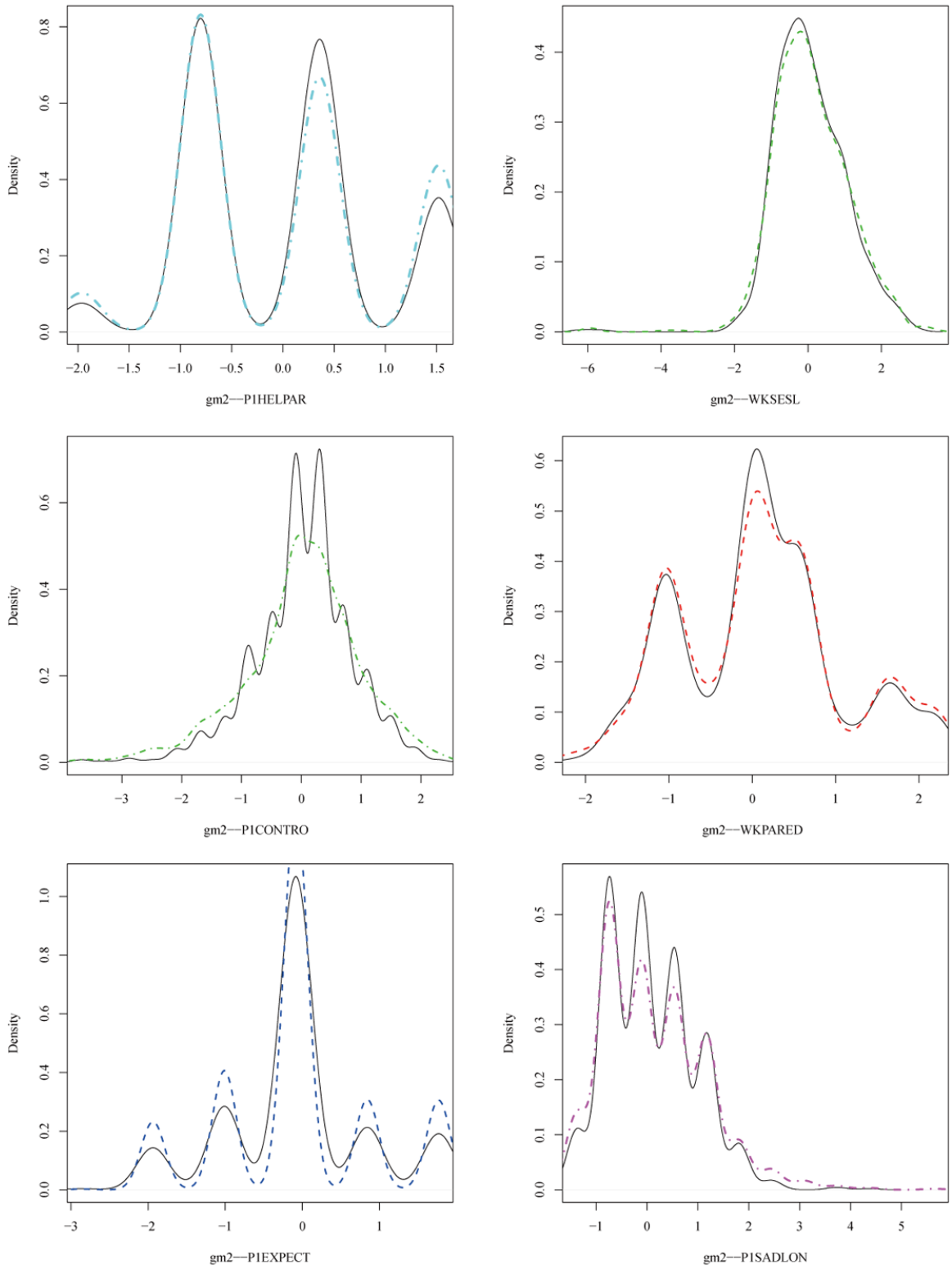


Figure 3.1 Residual Plot of CEM2

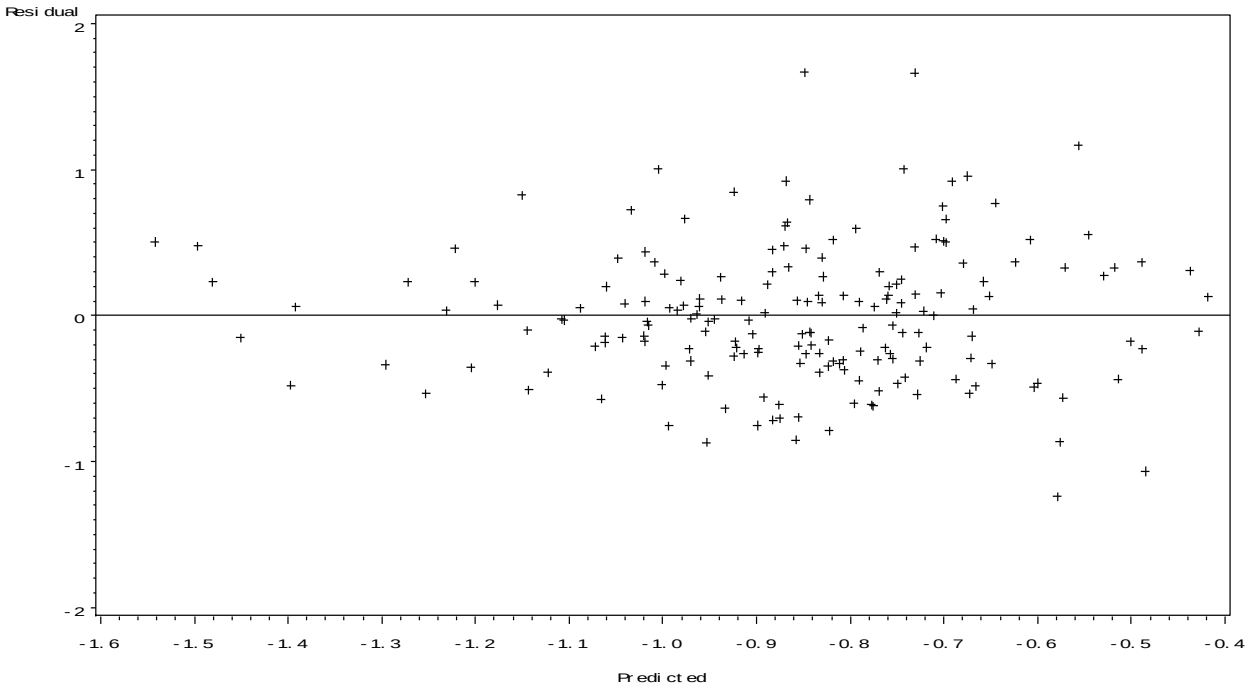


Figure 3.2 Residual Plot of GM2

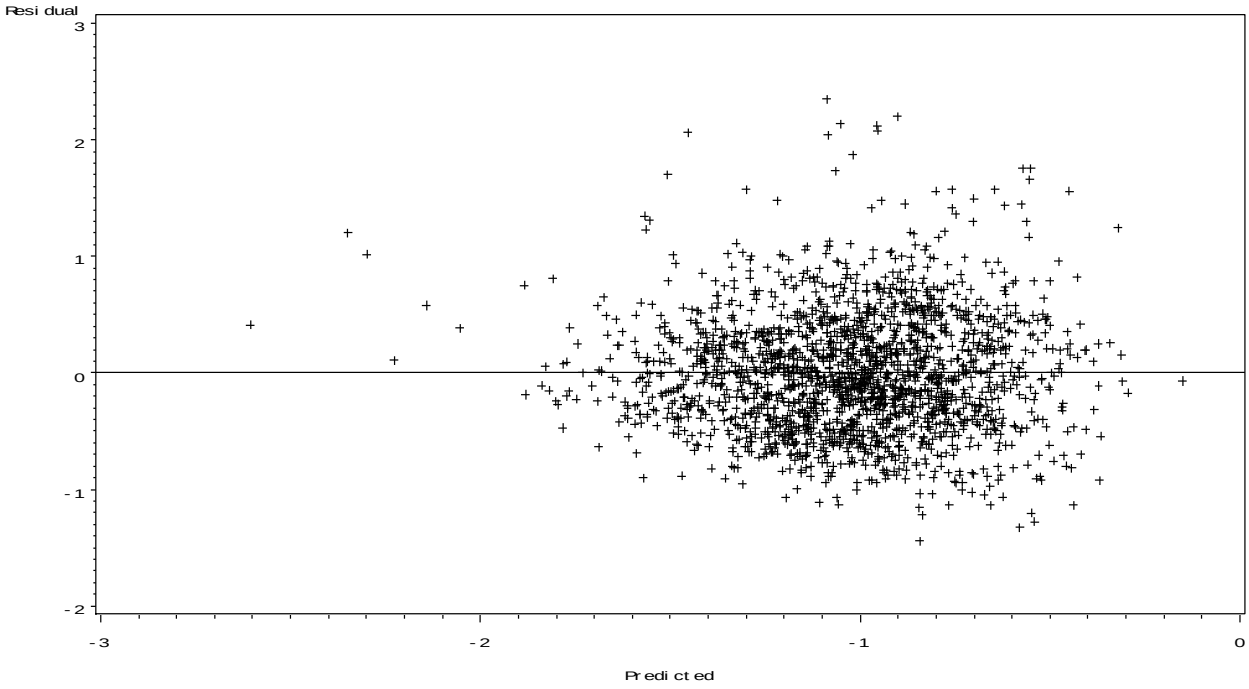
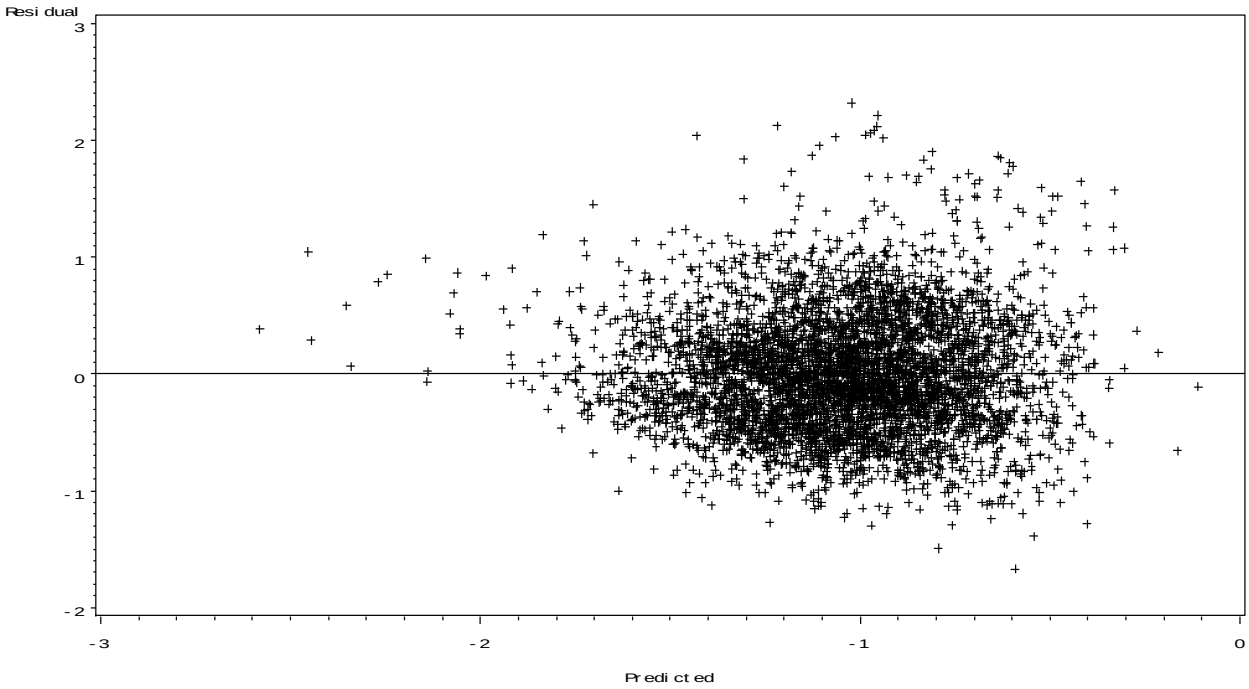


Figure 3.3 Residual Plot of PSM1



APPENDIX A

R SCRIPT FOR THE SIMULATION

```
library (cem)
library (MatchIt)
library (MASS)

matchample<-function (n){
se_ca<-rep (NA,n)
se_cb<-rep (NA,n)
se_cc<-rep (NA,n)
se_cd<-rep (NA,n)
se_ce<-rep (NA,n)
se_cf<-rep (NA,n)
se_cg<-rep (NA,n)
se_pa<-rep (NA,n)
se_pb<-rep (NA,n)
se_pc<-rep (NA,n)
se_pd<-rep (NA,n)
se_pe<-rep (NA,n)
se_pf<-rep (NA,n)
se_pg<-rep (NA,n)
se_ga<-rep (NA,n)
se_gb<-rep (NA,n)
se_gc<-rep (NA,n)
se_gd<-rep (NA,n)
se_ge<-rep (NA,n)
se_gf<-rep (NA,n)
se_gg<-rep (NA,n)
for (i in 1:i)
{set.seed (i)
e1<-rnorm (2000,0,1)
e2<-rnorm (2000,0,1)
e3<-rnorm (2000,0,1)
e4<-rnorm (2000,0,1)
v1<-rnorm (2000,0,1)
v3<-rnorm (2000,0,1)
v4<-rnorm (2000,0,1)
w2<-rnorm (2000,0,1)
w4<-rnorm (2000,0,1)
```

```
v5<-0.2*v1+e1
v6<-0.9*w2+e2
v8<-0.2*v3+e3
v9<-0.9*w4+e4
w7<-rnorm (2000,0,1)
w10<-rnorm (2000,0,1)
w1<-ifelse (v1<0.5,0,1)
w3<-ifelse (v3<0.5,0,1)
w5<-ifelse (v5<0.5,0,1)
w6<-ifelse (v6<0.5,0,1)
w8<-ifelse (v8<0.5,0,1)
w9<-ifelse (v9<0.5,0,1)
ta<-1/ (1+exp (- (0.8*w1-0.25*w2+0.6*w3-0.4*w4-0.8*w5-0.5*w6+0.7*w7)))
tb<-1/ (1+exp (- (0.8*w1-0.25*w2+0.6*w3-0.4*w4-0.8*w5-0.5*w6+0.7*w7-0.25*w2*w2)))
tc<-1/ (1+exp (- (0.8*w1-0.25*w2+0.6*w3-0.4*w4-0.8*w5-0.5*w6+0.7*w7-0.25*w2*w2-
0.4*w4*w4+0.7*w7*w7)))
td<-1/ (1+exp (- (0.8*w1-0.25*w2+0.6*w3-0.4*w4-0.8*w5-0.5*w6+0.7*w7+0.8*0.5*w1*w3-
0.25*0.7*w2*w4-0.4*0.5*w4*w5+
-0.8*0.5*w5*w6)))
te<-1/ (1+exp (- (0.8*w1-0.25*w2+0.6*w3-0.4*w4-0.8*w5-0.5*w6+0.7*w7+0.8*0.5*w1*w3-
0.25*0.7*w2*w4-0.4*0.5*w4*w5+
-0.8*0.5*w5*w6-0.25*w2*w2-0.4*w4*w4+0.7*w7*w7)))
tf<-1/ (1+exp (- (0.8*w1-0.25*w2+0.6*w3-0.4*w4-0.8*w5-0.5*w6+0.7*w7+0.8*0.5*w1*w3-
0.25*0.7*w2*w4+0.6*0.5*w3*w5-0.4*0.5*w4*w5+
-0.4*0.7*w4*w6-0.8*0.5*w5*w6-0.8*0.5*w5*w7+0.8*0.5*w1*w6-
0.25*0.7*w2*w3+0.6*0.5*w3*w4)))
tg<-1/ (1+exp (- (0.8*w1-0.25*w2+0.6*w3-0.4*w4-0.8*w5-0.5*w6+0.7*w7+0.8*0.5*w1*w3-
0.25*0.7*w2*w4+0.6*0.5*w3*w5-0.4*0.5*w4*w5+
-0.4*0.7*w4*w6-0.8*0.5*w5*w6-0.8*0.5*w5*w7+0.8*0.5*w1*w6-
0.25*0.7*w2*w3+0.6*0.5*w3*w4+
-0.25*w2*w2-0.4*w4*w4+0.7*w7*w7)))
g1<-runif (2000, 0, 1)
ga<-ifelse (g1<ta,1,0)
gb<-ifelse (g1<tb,1,0)
gc<-ifelse (g1<tc,1,0)
gd<-ifelse (g1<td,1,0)
ge<-ifelse (g1<te,1,0)
gf<-ifelse (g1<tf,1,0)
gg<-ifelse (g1<tg,1,0)
ya=-0.4*ga-3.85+0.3*w1-0.36*w2-0.73*w3-0.2*w4+0.71*w8-0.19*w9+0.26*w10
yb=-0.4*gb-3.85+0.3*w1-0.36*w2-0.73*w3-0.2*w4+0.71*w8-0.19*w9+0.26*w10
yc=-0.4*gc-3.85+0.3*w1-0.36*w2-0.73*w3-0.2*w4+0.71*w8-0.19*w9+0.26*w10
yd=-0.4*gd-3.85+0.3*w1-0.36*w2-0.73*w3-0.2*w4+0.71*w8-0.19*w9+0.26*w10
ye=-0.4*ge-3.85+0.3*w1-0.36*w2-0.73*w3-0.2*w4+0.71*w8-0.19*w9+0.26*w10
yf=-0.4*gf-3.85+0.3*w1-0.36*w2-0.73*w3-0.2*w4+0.71*w8-0.19*w9+0.26*w10
yg=-0.4*gg-3.85+0.3*w1-0.36*w2-0.73*w3-0.2*w4+0.71*w8-0.19*w9+0.26*w10
```

```
data1<-as.data.frame (cbind (ya,ta,ga,w1,w2,w3,w4,w5,w6,w7,w8,w9,w10))
data2<-as.data.frame (cbind (yb,tb,gb,w1,w2,w3,w4,w5,w6,w7,w8,w9,w10))
data3<-as.data.frame (cbind (yc,tc,gc,w1,w2,w3,w4,w5,w6,w7,w8,w9,w10))
data4<-as.data.frame (cbind (yd,td,gd,w1,w2,w3,w4,w5,w6,w7,w8,w9,w10))
data5<-as.data.frame (cbind (ye,te,ge,w1,w2,w3,w4,w5,w6,w7,w8,w9,w10))
data6<-as.data.frame (cbind (yf,tf,gf,w1,w2,w3,w4,w5,w6,w7,w8,w9,w10))
data7<-as.data.frame (cbind (yg,tg,gg,w1,w2,w3,w4,w5,w6,w7,w8,w9,w10))

autocem_a<-cem (treatment="ga",data=data1,drop=c ("ya","ta","w8","w9","w10"))
wa<-autocem_a$w
dcema<-cbind (wa,data1)
dcem_a<-dcema[which (wa>0.00000001),]
lgca<-lm (ya~ga,data=dcem_a)
coca<-coef (lgca)
b0ca<-coca[' (Intercept)']
b1ca<-coca['ga']
se_ca[i]<-b1ca+0.4

autocem_b<-cem (treatment="gb",data=data2,drop=c ("yb","tb","w8","w9","w10"))
wb<-autocem_b$w
dcemb<-cbind (wb,data2)
dcem_b<-dcemb[which (wb>0.00000001),]
lgcb<-lm (yb~gb,data=dcem_b)
cocb<-coef (lgcb)
b0cb<-cocb[' (Intercept)']
b1cb<-cocb['gb']
se_cb[i]<-b1cb+0.4

#dim (dcem_a) get dimension

autocem_c<-cem (treatment="gc",data=data3,drop=c ("yc","tc","w8","w9","w10"))
wc<-autocem_c$w
dcemc<-cbind (wc,data3)
dcem_c<-dcemc[which (wc>0.00000001),]
lgcc<-lm (yc~gc,data=dcem_c)
cocc<-coef (lgcc)
b0cc<-cocc[' (Intercept)']
b1cc<-cocc['gc']
se_cc[i]<-b1cc+0.4

autocem_d<-cem (treatment="gd",data=data4,drop=c ("yd","td","w8","w9","w10"))
wd<-autocem_d$w
dcemd<-cbind (wd,data4)
dcem_d<-dcemd[which (wd>0.00000001),]
lgcd<-lm (yd~gd,data=dcem_d)
cocd<-coef (lgcd)
```

```
b0cd<-cocd[' (Intercept)']
b1cd<-cocd['gd']
se_cd[i]<-b1cd+0.4

autocem_e<-cem (treatment="ge",data=data5,drop=c ("ye","te","w8","w9","w10"))
we<-autocem_e$w
dceme<-cbind (we,data5)
dcem_e<-dceme[which (we>0.00000001),]
lgce<-lm (ye~ge,data=dcem_e)
coce<-coef (lgce)
b0ce<-coce[' (Intercept)']
b1ce<-coce['ge']
se_ce[i]<-b1ce+0.4

autocem_f<-cem (treatment="gf",data=data6,drop=c ("yf","tf","w8","w9","w10"))
wf<-autocem_f$w
dcemf<-cbind (wf,data6)
dcem_f<-dcemf[which (wf>0.00000001),]
lgcf<-lm (yf~gf,data=dcem_f)
cocf<-coef (lgcf)
b0cf<-cocf[' (Intercept)']
b1cf<-cocf['gf']
se_cf[i]<-b1cf+0.4

autocem_g<-cem (treatment="gg",data=data7,drop=c ("yg","tg","w8","w9","w10"))
wg<-autocem_g$w
dcemg<-cbind (wg,data7)
dcem_g<-dcemg[which (wg>0.00000001),]
lgcg<-lm (yg~gg,data=dcem_g)
cocg<-coef (lgcg)
b0cg<-cocg[' (Intercept)']
b1cg<-cocg['gg']
se_cg[i]<-b1cg+0.4

psm_a<-matchit (ga~w1+w2+w3+w4+w5+w6+w7,data=data1,method = "nearest", discard =
"both", replace =  TRUE, ratio = 5,
caliper = 0.01, m.order = "random")
dpm_a<-match.data (psm_a)
lgpa<-lm (ya~ga,data=dpm_a)
copa<-coef (lgpa)
b0pa<-copa[' (Intercept)']
b1pa<-copa['ga']
se_pa[i]<-b1pa+0.4

psm_b<-matchit (gb~w1+w2+w3+w4+w5+w6+w7,data=data2,method = "nearest", discard =
"both", replace =  TRUE, ratio = 5,
```

```
caliper = 0.01, m.order = "random")
dpm_b<-match.data (psm_b)
lgpb<-lm (yb~gb,data=dpm_b)
copb<-coef (lgpb)
b0pb<-copb[' (Intercept)']
b1pb<-copb['gb']
se_pb[i]<-b1pb+0.4

psm_c<-matchit (gc~w1+w2+w3+w4+w5+w6+w7,data=data3,method = "nearest", discard =
"both", replace =  TRUE, ratio = 5,
caliper = 0.01, m.order = "random")
dpm_c<-match.data (psm_c)
lgpc<-lm (yc~gc,data=dpm_c)
copc<-coef (lgpc)
b0pc<-copc[' (Intercept)']
b1pc<-copc['gc']
se_pc[i]<-b1pc+0.4

psm_d<-matchit (gd~w1+w2+w3+w4+w5+w6+w7,data=data4,method = "nearest", discard =
"both", replace =  TRUE, ratio = 5,
caliper = 0.01, m.order = "random")
dpm_d<-match.data (psm_d)
lgpd<-lm (yd~gd,data=dpm_d)
copd<-coef (lgpd)
b0pd<-copd[' (Intercept)']
b1pd<-copd['gd']
se_pd[i]<-b1pd+0.4


psm_e<-matchit (ge~w1+w2+w3+w4+w5+w6+w7,data=data5,method = "nearest", discard =
"both", replace =  TRUE, ratio = 5,
caliper = 0.01, m.order = "random")
dpm_e<-match.data (psm_e)
lgpe<-lm (ye~ge,data=dpm_e)
cope<-coef (lgpe)
b0pe<-cope[' (Intercept)']
b1pe<-cope['ge']
se_pe[i]<-b1pe+0.4

psm_f<-matchit (gf~w1+w2+w3+w4+w5+w6+w7,data=data6,method = "nearest", discard =
"both", replace =  TRUE, ratio = 5,
caliper = 0.01, m.order = "random")
dpm_f<-match.data (psm_f)
lgpf<-lm (yf~gf,data=dpm_f)
copf<-coef (lgpf)
b0pf<-copf[' (Intercept)']
```

```
b1pf<-copf['gf']
se_pf[i]<-b1pf+0.4

psm_g<-matchit (gg~w1+w2+w3+w4+w5+w6+w7,data=data7,method = "nearest", discard =
"both", replace =  TRUE, ratio = 5,
caliper = 0.01, m.order = "random")
dpm_g<-match.data (psm_g)
lgpg<-lm (yg~gg,data=dpm_f)
copg<-coef (lgpg)
b0pg<-copg[' (Intercept)']
b1pg<-copg['gg']
se_pg[i]<-b1pg+0.4

gm_a<-matchit (ga~w1+w2+w3+w4+w5+w6+w7,data=data1,method = "genetic", pop.size=500)
dgm_a<-match.data (gm_a)
lgga<-lm (ya~ga,data=dgm_a)
coga<-coef (lgga)
b0ga<-coga[' (Intercept)']
b1ga<-coga['ga']
se_ga[i]<-b1ga+0.4

gm_b<-matchit (gb~w1+w2+w3+w4+w5+w6+w7,data=data2,method = "genetic", pop.size=500)
dgm_b<-match.data (gm_b)
lggb<-lm (yb~gb,data=dgm_b)
cogb<-coef (lggb)
b0gb<-cogb[' (Intercept)']
b1gb<-cogb['gb']
se_gb[i]<-b1gb+0.4

gm_c<-matchit (gc~w1+w2+w3+w4+w5+w6+w7,data=data3,method = "genetic",pop.size=500)
dgm_c<-match.data (gm_c)
lggc<-lm (yc~gc,data=dgm_c)
cogc<-coef (lggc)
b0gc<-cogc[' (Intercept)']
b1gc<-cogc['gc']
se_gc[i]<-b1gc+0.4

gm_d<-matchit (gd~w1+w2+w3+w4+w5+w6+w7,data=data4,method = "genetic",pop.size=500)
dgm_d<-match.data (gm_d)
lggd<-lm (yd~gd,data=dgm_d)
cogd<-coef (lggd)
b0gd<-cogd[' (Intercept)']
b1gd<-cogd['gd']
se_gd[i]<-b1gd+0.4

gm_e<-matchit (ge~w1+w2+w3+w4+w5+w6+w7,data=data5,method = "genetic",pop.size=500)
```

```
dgm_e<-match.data (gm_e)
lgge<-lm (ye~ge,data=dgm_e)
coge<-coef (lgge)
b0ge<-coge[' (Intercept)']
b1ge<-coge['ge']
se_ge[i]<-b1ge+0.4


gm_f<-matchit (gf~w1+w2+w3+w4+w5+w6+w7,data=data6,method = "genetic",pop.size=500)
dgm_f<-match.data (gm_f)
lggf<-lm (yf~gf,data=dgm_f)
cogf<-coef (lggf)
b0gf<-cogf[' (Intercept)']
b1gf<-cogf['gf']
se_gf[i]<-b1gf+0.4

gm_g<-matchit (gg~w1+w2+w3+w4+w5+w6+w7,data=data7,method = "genetic",pop.size=500)
dgm_g<-match.data (gm_g)
lggg<-lm (yg~gg,data=dgm_g)
cogg<-coef (lggg)
b0gg<-cogg[' (Intercept)']
b1gg<-cogg['gg']
se_gg[i]<-b1gg+0.4
}
mse_ca<-mean (se_ca,na.rm=TRUE)
mse_cb<-mean (se_cb,na.rm=TRUE)
mse_cc<-mean (se_cc,na.rm=TRUE)
mse_cd<-mean (se_cd,na.rm=TRUE)
mse_ce<-mean (se_ce,na.rm=TRUE)
mse_cf<-mean (se_cf,na.rm=TRUE)
mse_cg<-mean (se_cg,na.rm=TRUE)

mse_pa<-mean (se_pa,na.rm=TRUE)
mse_pb<-mean (se_pb,na.rm=TRUE)
mse_pc<-mean (se_pc,na.rm=TRUE)
mse_pd<-mean (se_pd,na.rm=TRUE)
mse_pe<-mean (se_pe,na.rm=TRUE)
mse_pf<-mean (se_pf,na.rm=TRUE)
mse_pg<-mean (se_pg,na.rm=TRUE)

mse_ga<-mean (se_ga,na.rm=TRUE)
mse_gb<-mean (se_gb,na.rm=TRUE)
mse_gc<-mean (se_gc,na.rm=TRUE)
mse_gd<-mean (se_gd,na.rm=TRUE)
mse_ge<-mean (se_ge,na.rm=TRUE)
mse_gf<-mean (se_gf,na.rm=TRUE)
```

```
mse_gg<-mean (se_gg,na.rm=TRUE)

mse<-rbind (mse_ca,mse_cb,mse_cc,mse_cd,mse_ce,mse_cf,mse_cg,mse_pa,mse_pb,
mse_pc,mse_pd,mse_pe,mse_pf,mse_pg,mse_ga,mse_gb,mse_gc,mse_gd,mse_ge,mse_gf,
mse_gg)
mse
}

SIM2000<-matchample (1000)
SIM2000


matchample2<-function (n){
se_ca<-rep (NA,n)
se_cb<-rep (NA,n)
se_cc<-rep (NA,n)
se_cd<-rep (NA,n)
se_ce<-rep (NA,n)
se_cf<-rep (NA,n)
se_cg<-rep (NA,n)
se_pa<-rep (NA,n)
se_pb<-rep (NA,n)
se_pc<-rep (NA,n)
se_pd<-rep (NA,n)
se_pe<-rep (NA,n)
se_pf<-rep (NA,n)
se_pg<-rep (NA,n)
se_ga<-rep (NA,n)
se_gb<-rep (NA,n)
se_gc<-rep (NA,n)
se_gd<-rep (NA,n)
se_ge<-rep (NA,n)
se_gf<-rep (NA,n)
se_gg<-rep (NA,n)
for (i in 1:i)
{set.seed (i)
e1<-rnorm (1000,0,1)
e2<-rnorm (1000,0,1)
e3<-rnorm (1000,0,1)
e4<-rnorm (1000,0,1)
v1<-rnorm (1000,0,1)
v3<-rnorm (1000,0,1)
v4<-rnorm (1000,0,1)
w2<-rnorm (1000,0,1)
w4<-rnorm (1000,0,1)
v5<-0.2*v1+e1
```

```
v6<-0.9*w2+e2
v8<-0.2*v3+e3
v9<-0.9*w4+e4
w7<-rnorm (1000,0,1)
w10<-rnorm (1000,0,1)
w1<-ifelse (v1<0.5,0,1)
w3<-ifelse (v3<0.5,0,1)
w5<-ifelse (v5<0.5,0,1)
w6<-ifelse (v6<0.5,0,1)
w8<-ifelse (v8<0.5,0,1)
w9<-ifelse (v9<0.5,0,1)
ta<-1/ (1+exp (- (0.8*w1-0.25*w2+0.6*w3-0.4*w4-0.8*w5-0.5*w6+0.7*w7)))
tb<-1/ (1+exp (- (0.8*w1-0.25*w2+0.6*w3-0.4*w4-0.8*w5-0.5*w6+0.7*w7-0.25*w2*w2)))
tc<-1/ (1+exp (- (0.8*w1-0.25*w2+0.6*w3-0.4*w4-0.8*w5-0.5*w6+0.7*w7-0.25*w2*w2-
0.4*w4*w4+0.7*w7*w7)))
td<-1/ (1+exp (- (0.8*w1-0.25*w2+0.6*w3-0.4*w4-0.8*w5-0.5*w6+0.7*w7+0.8*0.5*w1*w3-
0.25*0.7*w2*w4-0.4*0.5*w4*w5+
-0.8*0.5*w5*w6)))
te<-1/ (1+exp (- (0.8*w1-0.25*w2+0.6*w3-0.4*w4-0.8*w5-0.5*w6+0.7*w7+0.8*0.5*w1*w3-
0.25*0.7*w2*w4-0.4*0.5*w4*w5+
-0.8*0.5*w5*w6-0.25*w2*w2-0.4*w4*w4+0.7*w7*w7)))
tf<-1/ (1+exp (- (0.8*w1-0.25*w2+0.6*w3-0.4*w4-0.8*w5-0.5*w6+0.7*w7+0.8*0.5*w1*w3-
0.25*0.7*w2*w4+0.6*0.5*w3*w5-0.4*0.5*w4*w5+
-0.4*0.7*w4*w6-0.8*0.5*w5*w6-0.8*0.5*w5*w7+0.8*0.5*w1*w6-
0.25*0.7*w2*w3+0.6*0.5*w3*w4)))
tg<-1/ (1+exp (- (0.8*w1-0.25*w2+0.6*w3-0.4*w4-0.8*w5-0.5*w6+0.7*w7+0.8*0.5*w1*w3-
0.25*0.7*w2*w4+0.6*0.5*w3*w5-0.4*0.5*w4*w5+
-0.4*0.7*w4*w6-0.8*0.5*w5*w6-0.8*0.5*w5*w7+0.8*0.5*w1*w6-
0.25*0.7*w2*w3+0.6*0.5*w3*w4+
-0.25*w2*w2-0.4*w4*w4+0.7*w7*w7)))
g1<-runif (1000, 0, 1)
ga<-ifelse (g1<ta,1,0)
gb<-ifelse (g1<tb,1,0)
gc<-ifelse (g1<tc,1,0)
gd<-ifelse (g1<td,1,0)
ge<-ifelse (g1<te,1,0)
gf<-ifelse (g1<tf,1,0)
gg<-ifelse (g1<tg,1,0)
ya=-0.4*ga-3.85+0.3*w1-0.36*w2-0.73*w3-0.2*w4+0.71*w8-0.19*w9+0.26*w10
yb=-0.4*gb-3.85+0.3*w1-0.36*w2-0.73*w3-0.2*w4+0.71*w8-0.19*w9+0.26*w10
yc=-0.4*gc-3.85+0.3*w1-0.36*w2-0.73*w3-0.2*w4+0.71*w8-0.19*w9+0.26*w10
yd=-0.4*gd-3.85+0.3*w1-0.36*w2-0.73*w3-0.2*w4+0.71*w8-0.19*w9+0.26*w10
ye=-0.4*ge-3.85+0.3*w1-0.36*w2-0.73*w3-0.2*w4+0.71*w8-0.19*w9+0.26*w10
yf=-0.4*gf-3.85+0.3*w1-0.36*w2-0.73*w3-0.2*w4+0.71*w8-0.19*w9+0.26*w10
yg=-0.4*gg-3.85+0.3*w1-0.36*w2-0.73*w3-0.2*w4+0.71*w8-0.19*w9+0.26*w10
data1<-as.data.frame (cbind (ya,ta,ga,w1,w2,w3,w4,w5,w6,w7,w8,w9,w10))
```

```
data2<-as.data.frame (cbind (yb,tb,gb,w1,w2,w3,w4,w5,w6,w7,w8,w9,w10))
data3<-as.data.frame (cbind (yc,tc,gc,w1,w2,w3,w4,w5,w6,w7,w8,w9,w10))
data4<-as.data.frame (cbind (yd,td,gd,w1,w2,w3,w4,w5,w6,w7,w8,w9,w10))
data5<-as.data.frame (cbind (ye,te,ge,w1,w2,w3,w4,w5,w6,w7,w8,w9,w10))
data6<-as.data.frame (cbind (yf,tf,gf,w1,w2,w3,w4,w5,w6,w7,w8,w9,w10))
data7<-as.data.frame (cbind (yg,tg,gg,w1,w2,w3,w4,w5,w6,w7,w8,w9,w10))

autocem_a<-cem (treatment="ga",data=data1,drop=c ("ya","ta","w8","w9","w10"))
wa<-autocem_a$w
dcema<-cbind (wa,data1)
dcem_a<-dcema[which (wa>0.00000001),]
lgca<-lm (ya~ga,data=dcem_a)
coca<-coef (lgca)
b0ca<-coca[' (Intercept)']
b1ca<-coca['ga']
se_ca[i]<-b1ca+0.4

autocem_b<-cem (treatment="gb",data=data2,drop=c ("yb","tb","w8","w9","w10"))
wb<-autocem_b$w
dcemb<-cbind (wb,data2)
dcem_b<-dcemb[which (wb>0.00000001),]
lgcb<-lm (yb~gb,data=dcem_b)
cocb<-coef (lgcb)
b0cb<-cocb[' (Intercept)']
b1cb<-cocb['gb']
se_cb[i]<-b1cb+0.4

autocem_c<-cem (treatment="gc",data=data3,drop=c ("yc","tc","w8","w9","w10"))
wc<-autocem_c$w
dcemc<-cbind (wc,data3)
dcem_c<-dcemc[which (wc>0.00000001),]
lgcc<-lm (yc~gc,data=dcem_c)
cocc<-coef (lgcc)
b0cc<-cocc[' (Intercept)']
b1cc<-cocc['gc']
se_cc[i]<-b1cc+0.4

autocem_d<-cem (treatment="gd",data=data4,drop=c ("yd","td","w8","w9","w10"))
wd<-autocem_d$w
dcemd<-cbind (wd,data4)
dcem_d<-dcemd[which (wd>0.00000001),]
lgcd<-lm (yd~gd,data=dcem_d)
cocd<-coef (lgcd)
b0cd<-cocd[' (Intercept)']
b1cd<-cocd['gd']
se_cd[i]<-b1cd+0.4
```

```
autocem_e<-cem (treatment="ge",data=data5,drop=c ("ye","te","w8","w9","w10"))
we<-autocem_e$w
dceme<-cbind (we,data5)
dcem_e<-dceme[which (we>0.00000001),]
lgce<-lm (ye~ge,data=dcem_e)
coce<-coef (lgce)
b0ce<-coce[' (Intercept)']
b1ce<-coce['ge']
se_ce[i]<-b1ce+0.4

autocem_f<-cem (treatment="gf",data=data6,drop=c ("yf","tf","w8","w9","w10"))
wf<-autocem_f$w
dcemf<-cbind (wf,data6)
dcem_f<-dcemf[which (wf>0.00000001),]
lgcf<-lm (yf~gf,data=dcem_f)
cocf<-coef (lgcf)
b0cf<-cocf[' (Intercfpt)']
b1cf<-cocf['gf']
se_cf[i]<-b1cf+0.4

autocem_g<-cem (treatment="gg",data=data7,drop=c ("yg","tg","w8","w9","w10"))
wg<-autocem_g$w
dcemg<-cbind (wg,data7)
dcem_g<-dcemg[which (wg>0.00000001),]
lgcg<-lm (yg~gg,data=dcem_g)
cocg<-coef (lgcg)
b0cg<-cocg[' (Intercgpt)']
b1cg<-cocg['gg']
se_cg[i]<-b1cg+0.4

psm_a<-matchit (ga~w1+w2+w3+w4+w5+w6+w7,data=data1,method = "nearest", discard =
"both", replace =  TRUE, ratio = 5,
caliper = 0.01, m.order = "random")
dpm_a<-match.data (psm_a)
lgpa<-lm (ya~ga,data=dpm_a)
copa<-coef (lgpa)
b0pa<-copa[' (Intercept)']
b1pa<-copa['ga']
se_pa[i]<-b1pa+0.4

psm_b<-matchit (gb~w1+w2+w3+w4+w5+w6+w7,data=data2,method = "nearest", discard =
"both", replace =  TRUE, ratio = 5,
caliper = 0.01, m.order = "random")
dpm_b<-match.data (psm_b)
lgpb<-lm (yb~gb,data=dpm_b)
```

```
copb<-coef (lgpb)
b0pb<-copb[' (Intercept)']
b1pb<-copb['gb']
se_pb[i]<-b1pb+0.4

psm_c<-matchit (gc~w1+w2+w3+w4+w5+w6+w7,data=data3,method = "nearest", discard =
"both", replace =  TRUE, ratio = 5,
caliper = 0.01, m.order = "random")
dpm_c<-match.data (psm_c)
lgpc<-lm (yc~gc,data=dpm_c)
copc<-coef (lgpc)
b0pc<-copc[' (Intercept)']
b1pc<-copc['gc']
se_pc[i]<-b1pc+0.4

psm_d<-matchit (gd~w1+w2+w3+w4+w5+w6+w7,data=data4,method = "nearest", discard =
"both", replace =  TRUE, ratio = 5,
caliper = 0.01, m.order = "random")
dpm_d<-match.data (psm_d)
lgpd<-lm (yd~gd,data=dpm_d)
copd<-coef (lgpd)
b0pd<-copd[' (Intercept)']
b1pd<-copd['gd']
se_pd[i]<-b1pd+0.4


psm_e<-matchit (ge~w1+w2+w3+w4+w5+w6+w7,data=data5,method = "nearest", discard =
"both", replace =  TRUE, ratio = 5,
caliper = 0.01, m.order = "random")
dpm_e<-match.data (psm_e)
lgpe<-lm (ye~ge,data=dpm_e)
cope<-coef (lgpe)
b0pe<-cope[' (Intercept)']
b1pe<-cope['ge']
se_pe[i]<-b1pe+0.4

psm_f<-matchit (gf~w1+w2+w3+w4+w5+w6+w7,data=data6,method = "nearest", discard =
"both", replace =  TRUE, ratio = 5,
caliper = 0.01, m.order = "random")
dpm_f<-match.data (psm_f)
lgpf<-lm (yf~gf,data=dpm_f)
copf<-coef (lgpf)
b0pf<-copf[' (Intercept)']
b1pf<-copf['gf']
se_pf[i]<-b1pf+0.4
```

```
psm_g<-matchit (gg~w1+w2+w3+w4+w5+w6+w7,data=data7,method = "nearest", discard =
"both", replace =  TRUE, ratio = 5,
caliper = 0.01, m.order = "random")
dpm_g<-match.data (psm_g)
lgpg<-lm (yg~gg,data=dpm_f)
copg<-coef (lgpg)
b0pg<-copg[' (Intercept)']
b1pg<-copg['gg']
se_pg[i]<-b1pg+0.4

gm_a<-matchit (ga~w1+w2+w3+w4+w5+w6+w7,data=data1,method = "genetic", pop.size=500)
dgm_a<-match.data (gm_a)
lgga<-lm (ya~ga,data=dgm_a)
coga<-coef (lgga)
b0ga<-coga[' (Intercept)']
b1ga<-coga['ga']
se_ga[i]<-b1ga+0.4

gm_b<-matchit (gb~w1+w2+w3+w4+w5+w6+w7,data=data2,method = "genetic", pop.size=500)
dgm_b<-match.data (gm_b)
lggb<-lm (yb~gb,data=dgm_b)
cogb<-coef (lggb)
b0gb<-cogb[' (Intercept)']
b1gb<-cogb['gb']
se_gb[i]<-b1gb+0.4

gm_c<-matchit (gc~w1+w2+w3+w4+w5+w6+w7,data=data3,method = "genetic",pop.size=500)
dgm_c<-match.data (gm_c)
lggc<-lm (yc~gc,data=dgm_c)
cogc<-coef (lggc)
b0gc<-cogc[' (Intercept)']
b1gc<-cogc['gc']
se_gc[i]<-b1gc+0.4

gm_d<-matchit (gd~w1+w2+w3+w4+w5+w6+w7,data=data4,method = "genetic",pop.size=500)
dgm_d<-match.data (gm_d)
lggd<-lm (yd~gd,data=dgm_d)
cogd<-coef (lggd)
b0gd<-cogd[' (Intercept)']
b1gd<-cogd['gd']
se_gd[i]<-b1gd+0.4

gm_e<-matchit (ge~w1+w2+w3+w4+w5+w6+w7,data=data5,method = "genetic",pop.size=500)
dgm_e<-match.data (gm_e)
lgge<-lm (ye~ge,data=dgm_e)
coge<-coef (lgge)
```

```
b0ge<-coge[' (Intercept)']
b1ge<-coge['ge']
se_ge[i]<-b1ge+0.4


gm_f<-matchit (gf~w1+w2+w3+w4+w5+w6+w7,data=data6,method = "genetic",pop.size=500)
dgm_f<-match.data (gm_f)
lggf<-lm (yf~gf,data=dgm_f)
cogf<-coef (lggf)
b0gf<-cogf[' (Intercept)']
b1gf<-cogf['gf']
se_gf[i]<-b1gf+0.4

gm_g<-matchit (gg~w1+w2+w3+w4+w5+w6+w7,data=data7,method = "genetic",pop.size=500)
dgm_g<-match.data (gm_g)
lggg<-lm (yg~gg,data=dgm_g)
cogg<-coef (lggg)
b0gg<-cogg[' (Intercept)']
b1gg<-cogg['gg']
se_gg[i]<-b1gg+0.4
}
mse_ca<-mean (se_ca,na.rm=TRUE)
mse_cb<-mean (se_cb,na.rm=TRUE)
mse_cc<-mean (se_cc,na.rm=TRUE)
mse_cd<-mean (se_cd,na.rm=TRUE)
mse_ce<-mean (se_ce,na.rm=TRUE)
mse_cf<-mean (se_cf,na.rm=TRUE)
mse_cg<-mean (se_cg,na.rm=TRUE)

mse_pa<-mean (se_pa,na.rm=TRUE)
mse_pb<-mean (se_pb,na.rm=TRUE)
mse_pc<-mean (se_pc,na.rm=TRUE)
mse_pd<-mean (se_pd,na.rm=TRUE)
mse_pe<-mean (se_pe,na.rm=TRUE)
mse_pf<-mean (se_pf,na.rm=TRUE)
mse_pg<-mean (se_pg,na.rm=TRUE)

mse_ga<-mean (se_ga,na.rm=TRUE)
mse_gb<-mean (se_gb,na.rm=TRUE)
mse_gc<-mean (se_gc,na.rm=TRUE)
mse_gd<-mean (se_gd,na.rm=TRUE)
mse_ge<-mean (se_ge,na.rm=TRUE)
mse_gf<-mean (se_gf,na.rm=TRUE)
mse_gg<-mean (se_gg,na.rm=TRUE)

mse<-rbind (mse_ca,mse_cb,mse_cc,mse_cd,mse_ce,mse_cf,mse_cg,mse_pa,mse_pb,
```

```
mse_pc,mse_pd,mse_pe,mse_pf,mse_pg,mse_ga,mse_gb,mse_gc,mse_gd,mse_ge,mse_gf,
mse_gg)
mse
}
```