NEW DEVELOPMENTS IN CAUSAL INFERENCE USING BALANCE OPTIMIZATION SUBSET SELECTION

BY

HEE YOUN KWON

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Systems and Entrepreneurial Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Assistant Professor Niao He, Chair
Professor Sheldon H. Jacobson, Director of Research
Assistant Professor Karthekeyan Chandrasekaran
Professor Rakesh Nagi

ABSTRACT

Causal inference with observational data has drawn attention across various fields. These observational studies typically use matching methods which find matched pairs with similar covariate values. However, matching methods may not directly achieve covariate balance, a measure of matching effectiveness. As an alternative, the Balance Optimization Subset Selection (BOSS) framework, which seeks the optimal covariate balance directly, has been proposed. This dissertation extends the BOSS framework in various ways and is composed of the following five parts.

The first part of the dissertation investigates all the possible cases that may lead to bias in the context of BOSS and tries to mitigate the bias. Second, this dissertation then extends the BOSS by estimating and decomposing a treatment effect as a combination of heterogeneous treatment effects from a partitioned set using the BOSS. Third, the dissertation generalizes the BOSS framework from a binary treatment setting to a multi-treatment setting. A treatment effect estimate with multiple treatments can be computed by combining estimates obtained from BOSS with binary treatments. The fourth part discusses on how to handle missing data with BOSS. It includes a sensitivity analysis of BOSS studying how the estimated values are affected by violation of the conditional independence assumption and methods to apply BOSS after multiple imputation on missing covariates. In these discussions, the performances of BOSS estimators are compared to those of matching estimators. In the last part, BOSS is formulated as an LP by relaxing integer constraints in the original mixed integer programming formulation and properties of its dual problem are investigated.

To my family.

ACKNOWLEDGMENTS

The very first person that I would like to mention is my PhD advisor, Professor Sheldon H. Jacobson. I am very fortunate to have him as my advisor and really grateful for invaluable advice that he has provided me over the last several years. I am hugely indebted to his insightful guidance and thoughtful encouragement throughout my PhD study in Systems and Entrepreneurial Engineering at the University of Illinois at Urbana-Champaign.

I am also very grateful for my PhD Final Examination (Defense) Committee members – Professor Karthik Chandrasekaran, Professor Niao He, and Professor Rakesh Nagi – and Professor Jason J. Sauppe who gave me a lot of feedback when I was working on this research. I would like to express my gratitude to Professor Negar Kiyavash as well for being my Preliminary Examination Committee member and I thank Professor Alex Olshevsky and Professor Sewoong Oh for being my first year advisors at Illinois.

I would like to thank my MPhil advisor Professor Sujoy Mukerji and Professor Marcel Fafchamps for their help when I was transitioning from an MPhil student in economics at the University of Oxford to a PhD student in engineering at Illinois. I am grateful to Professor Gyo Taek Jin, Professor Sang-il Oum, Professor Yoon-Jae Whang, Professor Elias Sanidas and Professor Roland Herzog who provided support and guidance when I was starting an intellectual journey as a graduate student.

Additionally, my thanks should go to the Mavis Fellowship program from College of Engineering at Illinois for providing a great training opportunity in addition to the financial support. Financial support in the form of research and teaching assistantship from Department of Industrial and Enterprise Systems Engineering (ISE) · Computational Science and Engineering Program (CSE) · Department of Computer Science (CS) at Illinois is gratefully acknowledged. I also highly appreciate Samsung Scholarship Foundation which provided me a generous support during my master's study and the University of Oxford for its institutional support

when I was an MPhil student. I would have not been able to finish my graduate study without the support from these institutions and academic advisors.

I also would like to thank Ms. Holly Kizer, Ms. Aleta Lynch, and Ms. Elaine Wilson for their administrate support. Furthermore, I cannot list all of their names but I am really grateful to the current and former members of the Simulation and Optimization Laboratory and many other friends who helped me in various ways.

Lastly, I thank my family members – my parents, younger sister and younger brother. They have provided me an unceasing encouragement. This dissertation is dedicated to my family.

A part of this dissertation has been published in the Journal of the Operational Research Society.

TABLE OF CONTENTS

TABLES viii

FIGURES ix

ABBREVIATIONS x

NOTATION xii

CHAPTER 1 INTRODUCTION 1
1.1 Introduction 1
1.2 Background on Balance Optimization Subset Selection Framework 2
1.3 Overview 8

CHAPTER 2 BIAS IN BALANCE OPTIMIZATION SUBSET SELECTION 13
2.1 Introduction 13
2.2 Relationship between Bias and Imbalance Measure 14
2.3 Balance Hierarchy and Correct Imbalance Measure 18
2.4 Examples 24
2.5 Non-zero Optimum under Correct Imbalance Measure 33
2.6 Conditions for Zero Bias in BOSS 37
2.7 Concluding Remarks 38

CHAPTER 3 TREATMENT EFFECT DECOMPOSITION AND BOOTSTRAP HYPOTHESIS TESTING IN OBSERVATIONAL STUDIES . 40
3.1 Introduction 40
3.2 Balance Optimization Subset Selection (BOSS) 42
3.3 Decomposition of the Treatment Effect 43
3.4 Applying the Two-Sample Bootstrap Hypothesis Testing 48
3.5 Application: LaLonde Data 49
3.6 Concluding Remarks 56

CHAPTER 4 BALANCE OPTIMIZATION SUBSET SELECTION WITH MULTIPLE TREATMENT LEVELS 62
4.1 Introduction 62

4.2 Average Treatment Effect 63
4.3 Strong Ignorability and Weak Ignorability Assumptions 66
4.4 Matching with Multiple Treatment Levels 66
4.5 BOSS with Multiple Treatment Levels 68
4.6 Simulation Results 74
4.7 Concluding Remarks and Future Research Direction 78

CHAPTER 5 HANDLING MISSING DATA IN OBSERVATIONAL STUDIES WITH BALANCE OPTIMIZATION SUBSET SELECTION 80
5.1 Introduction 80
5.2 Simulating an Unobserved Covariate 82
5.3 Missing Values in Covariates and Multiple Imputation 92
5.4 Concluding Remarks 102

CHAPTER 6 DUALITY IN BALANCE OPTIMIZATION SUBSET SELECTION 111
6.1 Introduction 111
6.2 Basic Properties of the Primal and Dual Problems of BOSS 112
6.3 Relationship between Primal and Dual Solutions of BOSS 115
6.4 Concluding Remarks 126

CHAPTER 7 CONCLUSION 127

REFERENCES 130

TABLES

2.1 Relationship between objective function value and the bias 17

3.1 One-sided p-values of the Estimated Average Treatment Effects for the Treated 52
3.2 One-sided p-values when Bootstrap Hypothesis Testing is Conducted to Respective Treatment Effects 61

4.1 Comparison of Estimators (The First Experiment) 76
4.2 Comparison of Estimators (The Second Experiment) 78

5.1 Simulation Results under Three Scenarios 90
5.2 Matching Estimators (and Corresponding Standard Errors) 91
5.3 BOSS Estimators (and Corresponding Standard Errors) 91
5.4 Values of $(p_{11}, p_{10}, p_{01}, p_{00})$ under Each Scenario 104
5.5 Matching Estimators (and Corresponding Standard Errors) 106
5.6 BOSS Estimators (and Corresponding Standard Errors) 107
5.7 Values of $(p_{11}, p_{10}, p_{01}, p_{00})$ under Each Scenario 108
5.8 Propensity Score Matching Estimators after multiple imputation with $\mathfrak{L} = 5$ and 100 repetitions 110
5.9 BOSS Estimators after multiple imputation with $\mathfrak{L} = 5$ and 100 repetitions 110

6.1 Changing the RHS of the First Constraint in (6.18) 118
6.2 Changing the RHS of the Second Constraint in (6.18) 119
6.3 Changing the RHS of the Third Constraint in (6.18) 119
6.4 Changing the RHS of the Fourth Constraint in (6.18) 119

FIGURES

2.1 Balance Hierarchy Example 20
2.2 Another Balance Hierarchy Example 24
2.3 Range of Covariate Values 35

3.1 Histogram and Density of 1000 p-values of Bootstrap Hypothesis Testing with a Treatment Group from Experimental NSW Treatment Data and the Corresponding Control Group from Experimental NSW Control Data 58
3.2 Histogram and Density of 1000 p-values of Bootstrap Hypothesis Testing with a Treatment Group from Experimental NSW Treatment Data and the Corresponding Control Group from Non-experimental PSID Control Data 59
3.3 Empirical CDF of One-sided p-values 60

5.1 Graphical Representation of $(p_{11}, p_{10}, p_{01}, p_{00})$ Values under 36 Scenarios in Table 5.4 105
5.2 Graphical Representation of $(p_{11}, p_{10}, p_{01}, p_{00})$ Values under 34 Scenarios in Table 5.7 109

ABBREVIATIONS

BOSS	Balance Optimization Subset Selection
ATE	Average treatment effect
ATT	Average treatment effect for the treated
LP	Linear programming
MIP	Mixed integer programming
GMD	Generalized Mahalanobis distance
DOM	Difference of means
DOM+DOV	Difference of first and second moments
DOM2	Bivariate moment
KS	Kolmogorov-Smirnov test statistic
ecdf	Empirical cumulative distribution function
CvM	Cramer-von Mises test statistic
CPU	Central processing unit
GHz	Gigahertz
NSW	National Supported Work Demonstration program
PSID	Population Survey of Income Dynamics
RE74	Real earnings in year 1974
RE75	Real earnings in year 1975
RE78	Real earnings in year 1978
S1	Scenario 1

S2	Scenario 2
S3	Scenario 3
S4	Scenario 4
S5	Scenario 5
S6	Scenario 6
U74	Unemployment rate in year 1974
MSE	Mean squared error
RMSE	Root mean squared error
MCOV	Nearest neighbor covariate matching
GPSM	Generalized propensity score matching
GPSS	Sub-classification on the generalized propensity score
MCAR	Missing completely at random
MAR	Missing at random
MNAR	Missing not at random
Pri.DE	Pediatric Respiratory Infection in Deutschland
RSV	Respiratory Syncytial Virus
LRTI	Lower Respiratory Tract Infections
RHS	Right hand side

NOTATION

u	Unit
$\mathcal{U}$	Population; Set of all possible units (Note: $u \in \mathcal{U}$)
Z_u	Treatment level of unit $u \in U$
Z	Treatment level of a a unit selected randomly from population $\mathcal{U}$ where the random selection is uniform
L	Number of treatment levels $L = 2$ in a binary treatment setting and $L \geq 2$ in a multi-treatment setting
$\mathcal{L}$	Set of treatment levels $\mathcal{L} = \{0, 1, \cdots, L-1\}$ (Note: $\mathcal{U} = \{u : Z_u \in \mathcal{L}\}$)
U^i	Set of units in population whose treatment level is i $U^i = \{u \in U : Z_u = i\}$ for $i \in \mathcal{L}$ (Note: $U = U^0 \cup U^1 \cup \cdots \cup U^{L-1}$)
S^i	Samples drawn from U^i for $i \in \mathcal{L}$
t	Treated unit
c	Control unit; Untreated unit
T	Treatment group; Set of all the treated units $T = S^1$ under a binary treatment setting
C	Control pool; Set of all the control units $C = S^0$ under a binary treatment setting
C'	Control group; Set of control units selected from the control pool under a binary treatment setting (Note: $C' \subset C$)

N Number of observed units
$N = |T| + |C'|$ under a binary treatment setting
$N = |S^0| + |S^1| + \cdots + |S^{L-1}|$ under a multi-treatment setting

$\mathcal{N}$ Set of all the observed units
$\mathcal{N} = \{u_1, \cdots, u_N\}$
$\mathcal{N} = T \cup C$ under a binary treatment setting
$\mathcal{N} = S^0 \cup S^1 \cup \cdots \cup S^{L-1} = \cup_{i=0}^{L-1} S^i$ under a multi-treatment setting

S Set of some observed units (Note: $S \subset \mathcal{N}$)

$X_{u,k}$ The k-th covariate value of unit $u \in U$

K Number of covariate indices

$\mathcal{P}$ Set of all covariate indices
$\mathcal{P} = \{1, 2, \cdots, K\}$

$\mathcal{K}$ Set of some covariate indices (Note: $\mathcal{K} \subset \mathcal{P}$)

$\mathbf{D}$ Set of all possible clusters of covariate indices;
Power set of $\mathcal{P}$

$\mathbf{X}_u$ Vector of covariates of unit $u \in \mathcal{U}$
$\mathbf{X}_u = (X_{u,1}, X_{u,2}, \cdots, X_{u,K})$

$\mathbf{X}$ Vector of K covariate values for a unit selected randomly from population $\mathcal{U}$ where the random selection is uniform

$\mathcal{X}$ Set of possible values for $\mathbf{X}$ (Support of $\mathbf{X}$)

$\mathcal{X}_k(S)$ Set of possible values for the k-th covariate value for all units $u \in S$
$\mathcal{X}_k(S) = \{X_{u,k} : u \in S\}$

$\mathcal{I}$ Imbalance measure
(Note: Under a binary treatment setting, it denotes an imbalance measure between a treatment group T and a control group C' for all covariates whose incides are in $\mathcal{P}$ if not noted otherwise.)

$\mathcal{I}_{\text{DOM}}$ Imbalance measure balancing the difference of means
Under a binary treatment setting, the imbalance measure between sets G_1 and G_2 is given by
$\mathcal{I}_{\text{DOM}}(G_1, G_2) = \sum_{k=1}^{K} \left| \frac{1}{|G_1|} \sum_{u \in G_1} X_{u,k} - \frac{1}{|G_2|} \sum_{u \in G_2} X_{u,k} \right|$
Under a multi-treatment setting,
$\mathcal{I}_{\text{DOM}}\left(S^l, (S^l)', (S^l)'', S^m, (S^m)', (S^m)''\right) = \mathcal{I}_{\text{DOM}}\left(S^l, (S^m)'\right) +$
$\mathcal{I}_{\text{DOM}}\left(S^m, (S^l)'\right) + \mathcal{I}_{\text{DOM}}\left(S^-, (S^m)''\right) + \mathcal{I}_{\text{DOM}}\left(S^-, (S^l)''\right)$

$\mathscr{I}_{\text{DOM:}\mathcal{K}}$ Imbalance measure balancing the difference of means for covariates whose indices are in $\mathcal{K}$
Under a binary treatment setting, the imbalance measure between sets G_1 and G_2 is given by
$\mathscr{I}_{\text{DOM:}\mathcal{K}}(G_1, G_2) = \sum_{k\in\mathcal{K}} \left|\frac{1}{|G_1|}\sum_{u\in G_1} X_{u,k} - \frac{1}{|G_2|}\sum_{u\in G_2} X_{u,G_2}\right|$

$\mathscr{I}_{\text{DOM+DOV}}$ Imbalance measure balancing the difference of first and second moments
Under a binary treatment setting, the imbalance measure between sets G_1 and G_2 is given by
$\mathscr{I}_{\text{DOM+DOV}}(G_1, G_2) = \mathscr{I}_{\text{DOM}} + \sum_{k=1}^{K} \left|\frac{1}{|G_1|}\sum_{u\in G_1} (X_{u,k})^2 - \frac{1}{|G_2|}\sum_{u\in G_2} (X_{u,k})^2\right|$
Under a multi-treatment setting,
$\mathscr{I}_{\text{DOM+DOV}}\left(S^l, (S^l)', (S^l)'', S^m, (S^m)', (S^m)''\right) = \mathscr{I}_{\text{DOM+DOV}}\left(S^l, (S^m)'\right) +$
$\mathscr{I}_{\text{DOM+DOV}}\left(S^m, (S^l)'\right) + \mathscr{I}_{\text{DOM+DOV}}\left(S^-, (S^m)''\right) + \mathscr{I}_{\text{DOM+DOV}}\left(S^-, (S^l)''\right)$

$\binom{\mathcal{P}}{2}$ Set of all the possible covariate index pairs
$\binom{\mathcal{P}}{2} = \{(k_1, k_2) \,|\, k_1, k_2 \in \mathcal{P}\}$

$\mathscr{I}_{\text{DOM2}}$ Imbalance measure balancing the bivariate moment
Under a binary treatment setting, the imbalance measure between sets G_1 and G_2 is given by
$\mathscr{I}_{\text{DOM2}}(G_1, G_2) = \mathscr{I}_{\text{DOM+DOV}}(G_1, G_2) +$
$\sum_{(k_1,k_2)\in\binom{\mathcal{P}}{2}} \left|\frac{1}{|G_1|}\sum_{u\in G_1} X_{u,k_1}X_{u,k_2} - \frac{1}{|G_2|}\sum_{u\in G_2} X_{u,k_1}X_{u,k_2}\right|$
Under a multi-treatment setting,
$\mathscr{I}_{\text{DOM2}}\left(S^l, (S^l)', (S^l)'', S^m, (S^m)', (S^m)''\right) = \mathscr{I}_{\text{DOM2}}\left(S^l, (S^m)'\right) +$
$\mathscr{I}_{\text{DOM2}}\left(S^m, (S^l)'\right) + \mathscr{I}_{\text{DOM2}}\left(S^-, (S^m)''\right) + \mathscr{I}_{\text{DOM2}}\left(S^-, (S^l)''\right)$

$\mathscr{I}_{\text{Corr:}\mathcal{K}}$ Imbalance measure balancing the correlation terms of the form $\prod_{k\in\mathcal{K}}(X_{u,k})^{p_k}$ for covariates whose indices are in $\mathcal{K}$ and $p_k \in \mathbb{R}$
Under a binary treatment setting, the imbalance measure between sets G_1 and G_2 is given by
$\mathscr{I}_{\text{Corr:}\mathcal{K}}(G_1, G_2) = \sum_{k\in\mathcal{K}} \left|\frac{1}{|G_1|}\sum_{u\in G_1} \prod_{k\in\mathcal{K}}(X_{u,k})^{p_k} - \frac{1}{|G_2|}\sum_{u\in G_2} \prod_{k\in\mathcal{K}}(X_{u,k})^{p_k}\right|$

$\widehat{F_k}(T, x)$ Empirical distribution function of the treatment group T
$\widehat{F_k}(T, x) = \left|\{u \in T : X_{u,k} \le x\}\right| / |T|$

$\widehat{F_k}(C', x)$ Empirical distribution function of the control group C'
$\widehat{F_k}(C', x) = \left|\{u \in C' : X_{u,k} \le x\}\right| / |C'|$

$\mathscr{I}_{\text{KS}}$ Imbalance measure balancing the Kolmogorov-Smirnov test statistic
Under a binary treatment setting, the imbalance measure between sets G_1 and G_2 is given by

$\mathscr{I}_{\mathrm{KS}}(G_1, G_2) = \sum_{k=1}^{K} \max_{x \in \mathscr{X}_k(G_1 \cup G_2)} \left| \widehat{F}_k(G_1, x) - \widehat{F}_k(G_2, x) \right|$
Under a multi-treatment setting,
$\mathscr{I}_{\mathrm{KS}}\left(S^l, (S^l)', (S^l)'', S^m, (S^m)', (S^m)''\right) = \mathscr{I}_{\mathrm{KS}}\left(S^l, (S^m)'\right) +$
$\mathscr{I}_{\mathrm{KS}}\left(S^m, (S^l)'\right) + \mathscr{I}_{\mathrm{KS}}\left(S^-, (S^m)''\right) + \mathscr{I}_{\mathrm{KS}}\left(S^-, (S^l)''\right)$

$\mathscr{I}_{\mathrm{ecdf:D}}$ Imbalance measure balancing the difference of joint empirical cumulative distribution functions of clusters in $\mathbf{D}$ using the Kolmogorov-Smirnov test statistic
Under a binary treatment setting, the imbalance measure between sets G_1 and G_2 is given by
$\mathscr{I}_{\mathrm{ecdf:D}}(G_1, G_2) = \sum_{D \in \mathbf{D}} \max_{x \in \mathscr{X}_D(G_1 \cup G_2)} \left| \widehat{F}_D(G_1, x) - \widehat{F}_D(G_2, x) \right|$
Under a multi-treatment setting,
$\mathscr{I}_{\mathrm{ecdf:D}}\left(S^l, (S^l)', (S^l)'', S^m, (S^m)', (S^m)''\right) = \mathscr{I}_{\mathrm{ecdf:D}}\left(S^l, (S^m)'\right) +$
$\mathscr{I}_{\mathrm{ecdf:D}}\left(S^m, (S^l)'\right) + \mathscr{I}_{\mathrm{ecdf:D}}\left(S^-, (S^m)''\right) + \mathscr{I}_{\mathrm{ecdf:D}}\left(S^-, (S^l)''\right)$

$\mathscr{I}_{\mathrm{CvM}}$ Imbalance measure balancing the Cramer-von Mises test statistic
Under a binary treatment setting,the imbalance measure between sets G_1 and G_2 is given by
$\mathscr{I}_{\mathrm{ecdf:D}}(G_1, G_2) = \frac{|G_1| \cdot |G_2|}{(|G_1| + |G_2|)^2} \sum_{k=1}^{K} \left(\sum_{x \in \mathscr{X}_k(G_1 \cup G_2)} (\widehat{F}_k(G_1, x) - \widehat{F}_k(G_2, x))^2 \right)$

$\mathscr{I}_{\mathrm{CvM:D}}$ Imbalance measure balancing the difference of joint empirical cumulative distribution functions of clusters in $\mathbf{D}$ using the Cramer-von Mises test statistic
Under a binary treatment setting, the imbalance measure between sets G_1 and G_2 is given by
$\mathscr{I}_{\mathrm{CvM:D}}(G_1, G_2) = \frac{|G_1| \cdot |G_2|}{(|G_1| + |G_2|)^2} \sum_{D \in \mathbf{D}} \left(\sum_{x \in \mathscr{X}_D(G_1 \cup G_2)} (\widehat{F}_D(G_1, x) - \widehat{F}_D(G_1, x))^2 \right)$

N_k Set of values that the k-th covariate can have

$E_{k,j}$ Set of units in $\mathcal{N}$ whose k-th covariate is equal to j

$\eta_{k,j}(G)$ Number of units in G whose k-th covariate is equal to j
$\eta_{k,j}(G) = |E_{k,j} \cap G|$ for $G \in \{T, C'\}$

$\mathscr{I}_{\mathcal{K}}$ Imbalance measure for $\mathcal{K}$ using binning
Under a binary treatment setting,
$\mathscr{I}_{\mathcal{K}}(T, C') = \sum_{k \in \mathcal{K}} \sum_{j \in N_k} |\eta_{k,j}(T) - \eta_{k,j}(C')|$

M_k Set of histogram bin indices for covariate k

$B_{k,b}$ Set of units whose k-th covariate value is in the b-th bin for $b \in M_k$

$\mathscr{I}_{\mathrm{hist}}$ Imbalance measure for $\mathcal{P}$ using histogram binning
Under a binary treatment setting,
$\mathscr{I}_{\mathrm{hist}}(T, C') = \sum_{k=1}^{K} \sum_{b \in M_k} \left| \frac{|T \cap B_{k,b}|}{|T|} - \frac{|C' \cap B_{k,b}|}{|C'|} \right|$

$h^z(\mathbf{X}_u)$ Response function for u in treatment level z
Under a binary treatment setting, $h^1(\mathbf{X}_u)$ denotes a treatment response function $h^0(\mathbf{X}_u)$ denote a control response function

ϵ_u^z Error term in a responses for treatment level z of a unit u

Y_u^z Response of unit u for treatment level z
$Y_u^z = h^z(\mathbf{X}_u) + \epsilon_u^z$ for $z \in \mathcal{L}$
(Note: Under a binary treatment setting, Y_u^1 is a treated response value which is observable for $u \in T$ and unobservable for $u \in C$ while Y_u^0 is a control response value which is observable for $u \in C$ and unobservable for $u \in T$.)

$\mathbb{E}(Y^z)$ Average population response for treatment level z

τ^1 Population average treatment effect for the treated (PATT) under a binary treatment setting
$\tau^1 = \mathbb{E}[Y^1 - Y^0|Z = 1]$

τ_T^1 Sample average treatment effect for the treated (SATT) under a binary treatment setting
$\tau_T^1 = \frac{1}{|T|} \sum_{t \in T} \left(Y_t^1 - Y_t^0\right)$

$\widetilde{\tau}_T^1(C')$ BOSS Estimator of SATT τ_T^1 in a binary treatment setting
$\widetilde{\tau}_T^1(C') = \frac{1}{|T|} \sum_{t \in T} Y_t^1 - \frac{1}{|C'|} \sum_{c \in C'} Y_c^0$

$\tau(m, l)$ Population average treatment effect (PATE) between treatment level l and treatment level m
$\tau(m, l) = \mathbb{E}[Y^m] - \mathbb{E}[Y^l]$

$\tau_{m,l}$ Sample average treatment effect (SATE) between treatment level l and treatment level m
$\tau_{m,l} = \frac{1}{|S|} \sum_{u \in S} \left(Y_u^m - Y_u^l\right) = \tau_{m,l} = \frac{1}{|S^0|+|S^1|+\cdots+|S^{L-1}|} \sum_{i=0}^{L-1} \sum_{u \in S^i} \left(Y_u^m - Y_u^l\right)$

$\nu(m, l)$ Population average treatment effect which only considers a pair of treatment levels under a multi-treatment setting
$\nu(l, m) = \mathbb{E}\left[Y^m \mid Z \in \{l, m\}\right] - \mathbb{E}\left[Y^l \mid Z \in \{l, m\}\right]$

$\hat{\nu}(m, l)$ Estimator of $\nu(l, m)$

$\widetilde{\tau}_{BOSS}(m, l)$ BOSS estimator of $\mathbb{E}[\tau_{m,l}]$
$\widetilde{\tau}_{BOSS}(m, l) = \frac{1}{|\mathcal{N}|}\Big\{ \sum_{u \in (S^m)'} Y_u^m - \sum_{u \in S^l} Y_u^l + \sum_{u \in S^m} Y_u^m - \sum_{u \in (S^l)'} Y_u^l + \sum_{u \in (S^m)''} Y_u^m - \sum_{u \in (S^l)''} Y_u^l \Big\}$

$f(u, j)$ $f(u, j) = \arg\min_{v: Z_v = j} \|X_v - X_u\|$

$\widetilde{\tau}_{GNN}(m, l)$ Generalized Nearest Neighbor covariate matching estimator of $\tau(m, l)$
$\widetilde{\tau}_{GNN}(m, l) = \frac{1}{|\mathcal{N}|} \sum_{u \in \mathcal{N}} \left(Y_{f(u,m)}^m - Y_{f(u,l)}^l\right)$

$p(z \mid \mathbf{X}_u)$	Generalized propensity score for a unit u whose covaraite vector is given by $\mathbf{X}_u$ $p(j \mid x) = \text{Prob}(Z_v = j \mid \mathbf{X}_v = x)$
$g(u, j)$	$g(u, j) = \arg\min_{v:Z_v=j} \lvert p(j\lvert\mathbf{X}_v) - p(j\lvert\mathbf{X}_u)\rvert$
$\widetilde{\tau}_{GPS}(m, l)$	Generalized Propensity Score (GPS) matching estimator of $\tau(m, l)$ $\widetilde{\tau}_{GPS}(m, l)] = \frac{1}{\lvert\mathcal{N}\rvert} \sum_{u\in\mathcal{N}} \left(Y^m_{g(u,m)} - Y^l_{g(u,l)}\right)$
$\overline{Y^1_T}$	Average treated response of units in a treatment group T
$\overline{Y^0_{C'}}$	Average untreated (control) response of units in a control group C'
$\mathscr{P}$	Partition of T $\mathscr{P} = \{T_1, T_2, \cdots, T_P\}$
$\mathcal{B}(T, C')$	Bias term in the estimator $\widetilde{\tau}^1_T(C')$ for τ^1_T $\mathcal{B}(T, C') = \frac{1}{\lvert T\rvert} \sum_{t\in T} h^0(\mathbf{X}_t) - \frac{1}{\lvert C'\rvert} \sum_{c\in C'} h^0(\mathbf{X}_c)$
$\mathcal{E}(T, C')$	Error term in the estimator $\widetilde{\tau}^1_T(C')$ for τ^1_T $\mathcal{E}(T, C') = \frac{1}{\lvert T\rvert} \sum_{t\in T} \epsilon^0_t - \frac{1}{\lvert C'\rvert} \sum_{c\in C'} \epsilon^0_c$
$\mathcal{B}$	$\mathcal{B}(T, C') - \frac{\lvert T_1\rvert}{\lvert T\rvert}\mathcal{B}(T, C'_1) - \frac{\lvert T_2\rvert}{\lvert T\rvert}\mathcal{B}(T_2, C'_2)$
$\mathcal{E}$	$\mathcal{E}(T, C') - \frac{\lvert T_1\rvert}{\lvert T\rvert}\mathcal{E}(T_1, C'_1) - \frac{\lvert T_2\rvert}{\lvert T\rvert}\mathcal{E}(T_2, C'_2)$
$\mathscr{A}$	Set of treated responses for units in T $\mathscr{A} = \{Y^1_t \mid t \in T\}$
$\mathscr{B}$	Set of untreated responses for units in C' $\mathscr{B} = \{Y^0_c \mid c \in C'\}$
$\mu_{\mathscr{A}}$	Mean of observed responses in $\mathscr{A}$ $\mu_{\mathscr{A}} = \overline{Y^1_T} = \frac{1}{\lvert T\rvert} \sum_{t\in T} Y^1_t$
$\mu_{\mathscr{B}}$	Mean of observed responses in $\mathscr{B}$ $\mu_{\mathscr{B}} = \overline{Y^0_{C'}} = \frac{1}{\lvert C'\rvert} \sum_{c\in C'} Y^0_c$
H_0	Null hypothesis
H_1	Alternative hypothesis
$\mathfrak{M}$	Number of iterations in the bootstrap procedure
δ	Difference in means of elements in A and B $\delta = \overline{Y^1_T} - \overline{Y^0_{C'}}$
Γ	Set of observed responses of units in T and C' $\Gamma = \mathscr{A} \cup \mathscr{B} = \{Y^1_t \mid t \in T\} \cup \{Y^0_c \mid c \in C'\}$

$Y^1_{T,\mathfrak{m}}$	Set of $\lvert T\rvert$ observations drawn from Γ in the $\mathfrak{m}$-th iteration
$Y^0_{C',\mathfrak{m}}$	Set of $\lvert C'\rvert$ observations drawn from Γ in the $\mathfrak{m}$-th iteration
$\overline{Y^1_{T,\mathfrak{m}}}$	Mean of observed values in $Y^1_{T,\mathfrak{m}}$
$\overline{Y^0_{C',\mathfrak{m}}}$	Mean of observed values in $Y^0_{C',\mathfrak{m}}$
$\delta_{\mathfrak{m}}$	$\delta_{\mathfrak{m}} = \overline{Y^1_{T,\mathfrak{m}}} - \overline{Y^0_{C',\mathfrak{m}}}$ in the $\mathfrak{m}$-th iteration
p	p-value for the one-sided hypothesis test H_0 versus H_1 $p = \frac{\sum_{\mathfrak{m}=1}^{\mathfrak{M}} \mathbb{1}[\delta_{\mathfrak{m}} \geq \delta]}{\mathfrak{M}}$ in one-sided hypothesis testing $p = \frac{\sum_{l=1}^{L} \mathbb{1}[\lvert\delta_{\mathfrak{m}}\rvert \geq \lvert\delta\rvert]}{\mathfrak{M}}$ in two-sided hypothesis testing
U	Previously unobserved binary covariate variable
$\mathscr{U}$	Set of the new binary values U that is attached to the given data to form an augmented set of covariates
p_{ij}	$\text{Prob}\{U = 1 \mid Z = z, B = j\}$ for $i, j \in \{0, 1\}$ for a treatment indicator Z and a binary outcome variable B
d	$p_{01} - p_{00}$
$p_{i\cdot}$	$\text{Prob}(U = 1 \mid Z = i)$
s	$p_{1\cdot} - p_{0\cdot}$
Y_u	Observed response for unit $u \in \mathcal{N}$ $Y_u = Y^1_u$ if $u \in T$ $Y_u = Y^0_u$ if $u \in C'$
$\overline{Y}$	Sample mean of the observed response values $\overline{Y} = \sum_{u \in \mathcal{N}} Y_u / \lvert\mathcal{N}\rvert$
B_u	Binary outcome $B_u = \mathbb{1}\{Y_u > \overline{Y}\}$
M	Number of repetitions in the sensitivity analysis
$e(\mathbf{X}_u)$	Propensity score for u under a binary treatment setting Probability of u receiving a treatment $e(\mathbf{X}_u) = \text{Prob}(Z_u = 1 \mid \mathbf{X}_u)$
$C_k(t)$	$C_k(t) = \arg\min_{c \in C_k} \lvert e(\mathbf{X}_t) - e(\mathbf{X}_c)\rvert$
$\hat{\tau}^1_{T,k}$	Matching estimator value obtained from the k-th estimation out of M repetitions $\hat{\tau}^1_{T,k} = \frac{1}{\lvert T_k\rvert} \sum_{t \in T_k} \left(Y^1_t - \sum_{c \in C_k(t)} \frac{1}{\lvert C_k(t)\rvert} Y^0_c \right)$

C_k Control pool with newly added $\mathscr{U}$ to the original control pool C in the k-th repetition of sensitivity analysis

C'_k Control group (selected from control pool C_k) that minimizes an imbalance $\mathscr{I}_{\text{DOM}}(T_k, C'_k)$

$\widetilde{\tau}^1_{T,k}(C'_k)$ BOSS estimator value obtained from the k-th estimation out of M repetitions
$\widetilde{\tau}^1_T(C'_k) = \frac{1}{|T_k|}\sum_{t\in T_k} Y^1_t - \frac{1}{|C'_k|}\sum_{c\in C'_k} Y^0_c$

$\overline{\hat{\tau}^1_T}$ Matching estimator of SATT after M repetitions
$\overline{\hat{\tau}^1_T} = \frac{1}{M}\sum_{k=1}^{M} \hat{\tau}^1_{T,k}$

$\overline{\widetilde{\tau}^1_T(C')}$ BOSS estimator of SATT after M repetitions
$\overline{\widetilde{\tau}^1_T(C')} = \frac{1}{M}\sum_{k=1}^{M} \widetilde{\tau}^1_{T,k}(C'_k)$

se^2_k Variance of the k-th estimator ($\hat{\tau}^1_{T,k}$ and $\widetilde{\tau}^1_{T,k}(C')$) among the M repetitions
For BOSS, $\text{se}^2_k = \frac{1}{|T|^2}\cdot|T|\cdot\text{Var}_{t\in T}(Y^1_t) + \frac{1}{|C'_k|^2}\cdot|C'_k|\cdot\text{Var}_{c\in C'_k}(Y^0_c) = \frac{\text{Var}_{t\in T}(Y^1_t)}{|T|} + \frac{\text{Var}_{c\in C'_k}(Y^0_c)}{|C'_k|}$

se^2_W Within-imputation variance $\text{se}^2_W = \frac{1}{M}\sum_{k=1}^{M}\text{se}^2_k$

se^2_B Between-imputation variance
For matching, $\text{se}^2_B = \frac{1}{M-1}\sum_{k=1}^{M}\left(\hat{\tau}^1_{T,k} - \overline{\hat{\tau}^1_T}\right)^2$.
For BOSS, $\text{se}^2_B = \frac{1}{M-1}\sum_{k=1}^{M}\left(\widetilde{\tau}^1_{T,k}(C'_k) - \overline{\widetilde{\tau}^1_T(C')}\right)^2$.

se^2_T Total variance for the estimator after M repetitions (e.g., the matching estimator $\overline{\hat{\tau}^1_T}$ and the BOSS estimator $\overline{\widetilde{\tau}^1_T(C')}$)
$\text{se}^2_T = \text{se}^2_W + \left(\frac{M+1}{M}\right)\text{se}^2_B$

$M_{u,k}$ Indicator for missing entries
$M_{u,k} = \begin{cases} 1 \text{ if } X_{u,k} \text{ is missing} \\ 0 \text{ otherwise} \end{cases}$

M_u Vector of K indicator values for unit u
$M_u = (M_{u,1}, M_{u,2}, \cdots, M_{u,K})$

$\underline{\mathbf{X}}$ Matrix of covariate values of N units in $\mathcal{N}$
$\underline{\mathbf{X}} = [\mathbf{X}_{u_1}\ \mathbf{X}_{u_2}\ \cdots\ \mathbf{X}_{u_N}]' \in \mathbb{R}^{n\times K}$

$\mathbf{M}$ Missing indicator matrix of $\underline{\mathbf{X}}$
$\mathbf{M} = [M_{u_1}\ M_{u_2}\ \cdots\ M_{u_N}]' \in \mathbb{R}^{n\times K}$

$\mathscr{M}$ Set of missing covariate values
$\mathscr{M} = \{X_{u,k} | M_{u,k} = 1 \text{ for } u \in \mathcal{N}, k \in \mathcal{P}\}$

$\mathscr{O}$ Set of observed covariate values
$\mathscr{O} = \{X_{u,k} | M_{u,k} = 0 \text{ for } u \in \mathcal{N}, k \in \mathcal{P}\}$.

$\mathfrak{Q}$ Number of imputations in multiple imputation

$\mathbf{X}^{<i>}$ The i-th complete dataset obtained by multiple imputation on $\underline{\mathbf{X}}$ for $i \in \{1, 2, \cdots, \mathfrak{Q}\}$

$C^{',<i>}$ Subset of C that minimizes an imbalance measure $\mathscr{I}(T, C^{',<i>})$ using covariate values in $\mathbf{X}^{<i>}$.

$\widetilde{\tau}_T^1(C^{',<i>})$ $\frac{1}{|T|}\sum_{t\in T} Y_t^1 - \frac{1}{|C^{',<i>}|}\sum_{c\in C^{',<i>}} Y_c^0$

$\widetilde{\tau}_T^{1,W}$ Within approach BOSS estimator of SATT after multiple imputation on $\underline{\mathbf{X}}$
$\widetilde{\tau}_T^{1,W} = \frac{\sum_{i=1,2,\cdots,\mathfrak{Q}} \widetilde{\tau}_T^1(C^{',<i>})}{\mathfrak{Q}}$

$\underline{\mathbf{X}}^A$ The average of $\mathfrak{Q}$ complete datasets from multiple imputation on $\underline{\mathbf{X}}$
$\underline{\mathbf{X}}^A = \frac{\mathbf{X}^{<1>}+\mathbf{X}^{<2>}+\cdots+\mathbf{X}^{<\mathfrak{Q}>}}{\mathfrak{Q}}$

$\widetilde{\tau}_T^{1,A}$ Across approach BOSS estimator of SATT after multiple imputation on $\underline{\mathbf{X}}$
$\widetilde{\tau}_T^{1,A} = \widetilde{\tau}_T^1(C^{',A}) \equiv \frac{1}{|T|}\sum_{t\in T} Y_t^1 - \frac{1}{|C^{',A}|}\sum_{c\in C^{',A}} Y_c^0$

$\mathbf{e}(\mathbf{X}^{<i>})$ Vector of estimated propensity scores for each complete dataset for
$\mathbf{X}^{<i>} = [\mathbf{X}_{u_1}^{<i>}\mathbf{X}_{u_2}^{<i>}\cdots\mathbf{X}_{u_N}^{<i>}]$
$\mathbf{e}(\mathbf{X}^{<i>}) = [e(\mathbf{X}_{u_1}^{<i>})\ e(\mathbf{X}_{u_2}^{<i>})\ \cdots\ e(\mathbf{X}_{u_N}^{<i>})]' \in \mathbb{R}^N$

$\hat{C}^{<i>}$ Set of matched control units thorough propensity score matching using $\mathbf{e}(\mathbf{X}^{<i>})$
$\hat{C}^{<i>} = \{c \mid c \in \arg\min_{c\in C} \|e(\mathbf{X}_c^{<i>}) - e(\mathbf{X}_t^{<i>})\| \text{ for some } t \in T\}$

$\hat{\tau}_T^1(\hat{C}^{<i>})$ $\frac{1}{|T|}\sum_{t\in T} Y_t^1 - \frac{1}{|\hat{C}^{<i>}|}\sum_{c\in\hat{C}^{<i>}} Y_c^0$

$\hat{\tau}_T^1(\hat{C}^{<i>})$ Within approach matching estimator of SATT after multiple imputation on $\underline{\mathbf{X}}$
$\hat{\tau}_T^{1,W} = \frac{\sum_{i=1,2,\cdots,m} \hat{\tau}_T^1(\hat{C}^{<i>})}{m}$

$\mathbf{e}^A$ Average vector of $< \mathfrak{Q} >$ propensity score vectors
$$\mathbf{e}^A(\mathbf{X}^{<1>}, \mathbf{X}^{<2>}, \cdots, \mathbf{X}^{<\mathfrak{Q}>}) = \frac{1}{\mathfrak{Q}}\sum_{i=1,2,\cdots,\mathfrak{Q}} \begin{bmatrix} e(\mathbf{X}_{u_1}^{<i>}) \\ e(\mathbf{X}_{u_2}^{<i>}) \\ \cdots \\ e(\mathbf{X}_{u_N}^{<i>}) \end{bmatrix}$$

$\hat{C}^A$ $\{c \mid c \in \arg\min_{c \in C} \|e(\mathbf{X}_c^A) - e(\mathbf{X}_t^A)\| \text{ for some } t \in T\}$

$\hat{\tau}_T^1(\hat{C}^A)$ Across approach matching estimator of SATT after multiple imputation on $\underline{\mathbf{X}}$
$\hat{\tau}_T^1(\hat{C}^A) \equiv \frac{1}{|T|} \sum_{t \in T} Y_t^1 - \frac{1}{|\hat{C}^A|} \sum_{c \in \hat{C}^A} Y_c^0$

$\underline{Z}$ $(Z_{u_1}, Z_{u_2}, \cdots, Z_{u_N})$

$\underline{B}$ Basis matrix in LP

v_c Indicator of $c \in C$ being in the C'
$v_c = \begin{cases} 1 \text{ if } c \in C' \\ 0 \text{ if } c \notin C' \end{cases}$

w_k Imbalance variable for the k-th covariate

$(v_{c_1}, v_{c_2}, \cdots, v_{c_{|C|}}, w_1, w_2, \cdots, w_K)$ Primal variables in the LP formulation of BOSS

$(y_{1^+}, y_{2^+}, \cdots, y_{K^+}, y_{1^-}, y_{2^-}, \cdots, y_{K^-}, y_s)$ Dual variable in the LP formulation of BOSS

$(v_{c_1}^*, v_{c_2}^*, \cdots, v_{c_{|C|}}^*, w_1^*, w_2^*, \cdots, w_K^*)$ Optimal primal solution in the LP formulation of BOSS

$(y_{1^+}^*, y_{2^+}^*, \cdots, y_{K^+}^*, y_{1^-}^*, y_{2^-}^*, \cdots, y_{K^-}^*, y_s^*)$ Optimal dual solution in the LP formulation of BOSS

V^* Optimal objective value in the LP formulation of BOSS

I_m $m \times m$ identity matrix

γ_k Perturbation factor for the k-th constraint

e_k The k-th unit vector

CHAPTER 1

INTRODUCTION

1.1 Introduction

Identifying causal relationships is an important task in many scientific research. In randomized experiments, units to be treated are selected at random from a pool of subjects. As a result, characteristics of treated units and those of control units are balanced stochastically. Hence, an estimator of treatment effects (the difference in average treated response and average control response) that are unbiased can be obtained. However, randomized experiments are often not available because of reasons such as high cost, impracticality, and ethical issues.

Rubin (1973) took a study on the effect of seat-belt usage on a human subject's damage after car crash as an example where using the observational data is inevitable. One cannot conduct experiment after randomly assigning a human subject to a group with seat-belt and another group without seat-belt. Using a human subject when conducting an experiment to study the effect of exposure to a harmful material like toxin or radiation is also unethical. In such cases, analyzing an observational data is essential.

Historically, matching methods have been used extensively to analyze observational data. Matching finds a control unit that has similar covariate values with a treated unit for all the treated units. By doing so, a bias that arises from systematic differences in covariate distribution between the treated and the control can be reduced. Depending on how to compare the difference between the treated unit and the control unit, there are several types of matching. Propensity score matching uses the probability of being treated given covariates as a single dimensional summary of each unit's covariates. Mahalanobis matching uses Mahalanobis metric when comparing the units.

One of the reasons matching methods have been popular is that matching minimizes the total distances between the matched units given a metric and it is solv-

able in polynomial time (See Section 1.2.2). However, matching methods have some drawbacks. One significant drawback is that they do not directly optimize the imbalance measure while they aim to reduce the imbalance between covariate distributions of a treatment group and a control group. Hence, researchers should repeat the steps of finding matches and checking balance iteratively.

To overcome the drawback, Balance Optimization Subset Selection (BOSS) which directly minimizes a given imbalance measure using mixed integer programming (MIP) was suggested by Nikolaev et al. (2013). Previously, Zubizarreta (2012) also suggested to solve an MIP containing a term directly minimizing the imbalance for optimal matching. However, BOSS is different from matching in that it focuses on balancing and estimating at the group level while all the matching methods focus on unit level responses.

The purpose of this dissertation research is to analyze and extend the BOSS framework so that it can be applicable to a more broad area. The first chapter of this dissertation is organized as follows. In Section 1.2, the BOSS framework will be reviewed. Section 1.3 provides an overview of topics to be covered in the dissertation. In the overview in Section 1.3, each subsection will briefly review the related research that has been done in the field and introduce the analysis and experimentation methods used in each Capter of this dissertation.

1.2 Background on Balance Optimization Subset Selection Framework

1.2.1 Randomized Experiments and Observational Data

Scientists have used randomized experiments to identify causal relationships. Randomization allows the treatment effect to be isolated from potential confounding factors. However, there are circumstances where one cannot conduct randomized experiments because of ethical issues or impracticality as mentioned in the previous section. In many cases where randomized experiments are not available, researchers should rely on observational data. For comparison of randomized experiments and observational studies in medical settings, see Hannan (2008), Hartz et al. (2005), Jepsen et al. (2004), and Port (2000).

Optimization can be used as a pre-processing step in experiments, as noted in

Bertsimas et al. (2015), to overcome limitations of randomization: When only a small number of samples can be used because of subjects' rarity or high cost, the imbalance between the treatment and control groups constructed by randomization may be severe. The groups obtained by solving an optimization problem may be more balanced.

When analyzing observational data, it is important to post-process the data so that one can find treatment and control groups offering an estimate of the treatment effect with minimal bias. The value of interest is the "average treatment effect for the treated" (ATT) and the objective is to the selection bias from inherent differences in covariates. Both matching and BOSS can serve as a post-processing step for the analysis of observational data.

1.2.2 Matching

One approach that has been used for analyzing observational data is the matching method which tries to match each unit in the treatment group with a unit from the control pool with the same or similar covariates in order to reduce differences in the covariate distribution (Rubin, 2006). There are several types of matching such as propensity score matching and Mahalanobis matching.

Matching appeals to many researchers because of its attractive features. It is a combinatorial optimization problem which has been extensively studied. The traditional matching problem, which is also known as an optimal assignment problem minimizing the total distances between the matched units, is solvable in polynomial time since its constraint matrix is totally unimodular (Zubizarreta, 2012).

The treatment effect can be estimated without bias if exact matches between treatment and control groups are obtained. However, it is difficult to have exact matches for all units in the treatment group when the number of covariates is not small. In such cases, researchers have to use inexact or incomplete matching. Inexact matching adopts a notion of distance measure to identify good matches having small distances. Incomplete matching is not desirable since it may leave some important treatment units unmatched. In addition, with inexact matching, the quality of the resulting matches is difficult to evaluate because it is unknown whether another metric would result in a control group with more balance.

1.2.3 Balance Optimization Subset Selection (BOSS)

Nikolaev et al. (2013) introduces a new approach: BOSS. This approach is motivated from the hypothesis that bias in the estimate of the treatment effect can be minimized by optimizing covariate balance directly. In this process, it is not required to find individually matched samples and it can be guaranteed that optimal balance is achieved.

BOSS was introduced as a way of estimating a sample average treatment effect for the treated (SATT) under a binary treatment setting whether each unit is either treated or not treated. Denote a population of units by $\mathcal{U}$. Let T be a set of treated units (i.e., treatment group) and C be a set of control units (i.e., control pool). Let $\mathcal{N}$ be a set of observed samples and $u \in \mathcal{N}$ be an observed unit. Let Z_u be a treatment indicator for a unit u such that

$$Z_u = \begin{cases} 1 \text{ if } u \text{ is treated} \\ 0 \text{ otherwise} \end{cases} \tag{1.1}$$

under a binary treatment setting where there are 2 treatment levels. Later, in Chapter 4, the notion of the treatment indicator will be extended so that it can have a value in $\{0, 1, \cdots, L-1\}$ under a multi-treatment setting where there are $L \geq 2$ treatment levels. The treatment group T can be written as $\{u \,|\, Z_u = 1, u \in \mathcal{N}\}$, the control pool C can be written as $\{u \,|\, Z_u = 0, u \in \mathcal{N}\}$, and $\mathcal{N} = T \cup C$ holds under a binary treatment setting.

Let Z be a treatment indicator (which is equal to 1 if treated and 0 if not) of a unit that is selected at random from the population $\mathcal{U}$ with a uniform random selection. Denote a unit u's treated response by Y_u^1 and its control response by Y_u^0. Let Y^1 be a treated response of a unit selected at random from the population $\mathcal{U}$ and Y^0 be a control response of such a unit.

The population average treatment effect for the treated (PATT) is defined as

$$\tau^1 \equiv \mathbb{E}[Y^1 - Y^0 | Z = 1] \tag{1.2}$$

and SATT is defined as

$$\tau_T^1 \equiv \frac{1}{|T|} \sum_{t \in T} \left(Y_t^1 - Y_t^0 \right). \tag{1.3}$$

Note that one cannot observe the control response of a treated unit Y_t^0 for $t \in T$ as Y_u^0 is not observable for $u \notin C$. Hence Y_t^0 should be estimated using the observed

values.

BOSS estimates the SATT with

$$\widetilde{\tau}^1_T(C') \equiv \frac{1}{|T|}\sum_{t\in T} Y^1_t - \frac{1}{|C'|}\sum_{c\in C'} Y^0_c. \tag{1.4}$$

using a control group C' obtained by solving an imbalance minimization problem to balance covariate distributions of T and C'. Various forms of imbalance measures can be used for the imbalance minimization problem. A difference of means (DOM) imbalance measure, $\mathscr{I}_{\text{DOM}}$ in (1.5), is one example. Let $\mathcal{P} = \{1, 2, \cdots, K\}$ be a set of covariate indices where K is a total number of covariates in each unit. Let $\mathbf{X}_u = (X_{u,1}, X_{u,2}, \cdots, X_{u,K})$ be a set of unit u's covariates where $X_{u,k}$ denotes the k-th covariate of the unit u. The imbalance measure $\mathscr{I}_{\text{DOM}}$ is defined as

$$\mathscr{I}_{\text{DOM}}(T, C') \equiv \sum_{k=1}^{K}\left|\frac{1}{|T|}\sum_{t\in T} X_{t,k} - \frac{1}{|C'|}\sum_{u\in C'} X_{c,k}\right|. \tag{1.5}$$

Many more imbalance measures and their relationship will be introduced in Chapter 2.

To obtain a control group C' that is balanced with a treatment group T using the DOM imbalance measure, BOSS method computationally solve for the following optimization problem:

$$C' = \arg\min_{C'\subset C, |C'|=s} \mathscr{I}_{\text{DOM}}(T, C'). \tag{1.6}$$

for a positive integer $s \in \mathbb{N}$ where the control group is composed of discrete (full) control units. Equivalently, BOSS method solves the following mixed integer programming in (1.7) to get $C' = \{c \in C : v_c = 1\}$ which minimizes the $\mathscr{I}_{\text{DOM}}$ between T and C' (Sauppe, 2015).

$$\begin{aligned}
\min \quad & \sum_{k\in\mathcal{P}} w_k & & \\
\text{s.t.} \quad & \tfrac{1}{s}\sum_{c\in C} v_c X_{c,k} - \tfrac{1}{|T|}\sum_{t\in T} X_{t,k} \le w_k & & \forall k \in \mathcal{P} \\
& \tfrac{1}{|T|}\sum_{t\in T} X_{t,k} - \tfrac{1}{s}\sum_{c\in C} v_c X_{c,k} \le w_k & & \forall k \in \mathcal{P} \\
& \sum_{c\in C} v_c = s & & \\
& v_c \in \{0,1\} & & \forall c \in C \\
& w_k \ge 0 & & \forall k \in \mathcal{P}.
\end{aligned} \tag{1.7}$$

Nikolaev et al. (2013) showed that, under a relaxed version of strong ignorability assumption which will be discussed below in Assumption 1' (and in more detail in Chapter 5), the BOSS estimator is unbiased if full covariate balance, $\{\mathbf{X}_u\}_{u\in T}=\{\mathbf{X}_u\}_{u\in C'}$, is given. This condition can be relaxed when the functional form for the response function is known.

Responses can be written as $Y_u^z = h^z(\mathbf{X}_u) + \epsilon_u^z$ for $z \in \{0, 1\}$ with a response function $h^z(\cdot)$ and error term ϵ_u^z. When the response functions are linear in covariates (i.e., $h^0(\mathbf{X}_u) = \beta^T\mathbf{X}_u + \alpha = \sum_{k\in\mathcal{P}}\beta_k X_{u,k} + \alpha$ for all $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^K$ where the set of covariate is $\mathcal{P} = \{1, 2, \cdots, K\}$), then the BOSS estimator is unbiased if $\mathscr{I}_{\text{DOM}}(T, C') = 0$ (Sauppe and Jacobson, 2017). This can be extended to higher order functional forms of the response functions given that appropriate imbalance measures are zero. It will be discussed again when defining the balance hierarchy and correct imbalance measure in Section 1.3.1 and Chapter 2.

In Nikolaev et al. (2013), the authors assume *Strong Ignorability*. As Sekhon (2009) mentioned, this assumption is composed of two conditions known as *unconfoundedness* and *common overlap*. The first condition states that, given the covariate values, potential outcomes and assignment to treatment are independent. The second condition is that each unit, given its covariate values, has a positive probability of belonging to either the treatment pool or the control pool. The Strong Ignorability assumption is also used for matching methods. It is a common assumption made by most researchers working with observational data.

Assumption 1. (Strong Ignorability)

$$Y_u^1, Y_u^0 \perp\!\!\!\perp Z_u \;\Big|\; \mathbf{X}_u \text{ and } 0 < P(Z_u = 1|\mathbf{X}_u) < 1 \tag{1.8}$$

Assumption 1'. (relaxed version of Assumption 1 for ATT)

$$Y_u^0 \perp\!\!\!\perp Z_u \;\Big|\; \mathbf{X}_u \text{ and } P(Z_u = 1|\mathbf{X}_u) < 1 \tag{1.9}$$

In this dissertation, the notation for the treatment group, T, is the same as the treatment pool, T since all treatment units are used for matching and BOSS. The terminologies, a treatement pool and a treatment group, are used interchangeably. On the other hand, the control group C' is not the same as the control pool C in general as the control group is chosen from the control pool by solving an imbal-

ance minimization problem. The objective of the BOSS problem is to minimize the difference of the control and treatment groups' covariate distributions as measured by some imbalance measure $\mathscr{I}$. The benefits of having closely balanced groups with a small imbalance measure are discussed in Zubizarreta (2012): a closely balanced group is more robust to misspecifications of the model and it offers more precise estimates when a model of covariance adjustment is used.

1.2.4 Comparison between Genetic Matching and BOSS

Diamond and Sekhon (2013) proposes Genetic Matching, a multivariate matching method which generalizes existing matching methods such as the ones using propensity score and Mahalanobis Distance. Diamond and Sekhon (2013) argues that while there is no dispute that covariate imbalance should be minimized, many researchers who use matching method for empirical studies fail to report whether covariate balance is achieved or not. This is because manually checking and modifying the specification of the matching method is tedious and prone to error. To overcome this challenge, the Genetic Matching method utilizes a genetic algorithm to find the matching metric that minimizes covariate imbalance after the matching. Sekhon and Grieve (2008) shows that the resulting covariate balance with the Genetic Matching method is better than that of propensity score matching, and leads to less bias in the treatment effect estimate.

This key idea of moving to the computational domain to overcome the challenges and human bias that may arise during sequential modifications to the matching metric is what is common between Genetic Matching and the BOSS framework. The main difference is that Genetic Matching still focuses on unit-level matching in order to minimize the following distance metric.

The generalized Mahalanobis distance from p.934 of Diamond and Sekhon (2013):

$$\mathrm{GMD}(\mathbf{X}_{u_1}, \mathbf{X}_{u_2}, W) = \sqrt{(\mathbf{X}_{u_1} - \mathbf{X}_{u_2})^T (S^{-1/2})^T W S^{-1/2} (\mathbf{X}_{u_1} - \mathbf{X}_{u_2})} \qquad (1.10)$$

Note that $\mathbf{X}_{u_i}$ $(i = 1, 2)$ in the above equation can contain the propensity score in addition to covariate values. In the Genetic Matching, the diagonal matrix W

is determined with the genetic algorithm adjusting the weights on each covariate so that it maximizes balance. On the other hand, BOSS method finds a subset that directly minimizes the imbalance by computationally solving an optimization problem without requiring units to be matched. Determining which imbalance measure to use is one of the issues that both of these methods are confronting.

1.3 Overview

This dissertation is composed of five parts. In the first part of my dissertation, cases that may result in bias in BOSS will be investigated and the way that the bias can be reduced will be discussed. While analyzing the cases that may lead to bias, notions of balance hierarchy and correct imbalance measure will be introduced. In the second part, treatment effect of the entire set is decomposed into a combination of heterogeneous treatment effects from its partition. Additionally, how researchers can conduct a bootstrap hypothesis testing to check the statistical significance of the treatment effect values obtained by BOSS will be explained. In the third part, BOSS framework which was originally introduced and discussed under a binary treatment setting will be extended so that it can be applied in a multi-treatment setting where there are more than two treatment levels. In the fourth part, how to handle missing data with BOSS is discussed. It includes a sensitivity analysis of BOSS studying how the estimated values are affected by violation of the conditional independence assumption and methods to apply BOSS after multiple imputation on missing covariates. Last part of the dissertation will discuss about duality (the relationship between a primal and its dual in the BOSS framework) after formulating the BOSS as an LP.

1.3.1 Bias in BOSS

Researchers often confront with bias resulted from differences in treatment units' and control units' covariate distributions when dealing with observational data. Observational studies using BOSS is not an exception. The bias issues have been extensively dealt with many researchers in studies using matching (e.g., Rubin (1973), Rosenbaum and Rubin (1985), Heckman et al. (1998), Abadie and Imbens (2012)). Investigating the bias and reducing it are also necessary for BOSS.

First, understanding how bias and imbalance measures are related is needed. Note that the difference between the estimated SATT and the true SATT can be expressed as a sum of selection bias and error terms.

To understand the relationship between bias and imbalance measures, define a notion of "ranking" between imbalance measures. An imbalance measure $\mathscr{I}_1$ has a *higher rank in balance hierarchy* than an imbalance measure $\mathscr{I}_2$ if $\mathscr{I}_2 = 0$ is implied by $\mathscr{I}_1 = 0$. For example, $\mathscr{I}_{\text{DOM+DOV}}(T, C') = 0$ implies that $\mathscr{I}_{\text{DOM}}(T, C') = 0$ where $\mathscr{I}_{\text{DOM+DOV}}(T, C')$ is defined as

$$\mathscr{I}_{\text{DOM+DOV}}(T, C') = \mathscr{I}_{\text{DOM}}(T, C') + \sum_{k=1}^{K} \left| \frac{1}{|T|} \sum_{t \in T} (X_{t,k})^2 - \frac{1}{|C'|} \sum_{c \in C'} (X_{c,k})^2 \right| \quad (1.11)$$

with $\mathscr{I}_{\text{DOM}}$ given in (1.5).

The balance hierarchy among the imbalance measures are given in Chapter 2. In the graphical representation of a balance hierarchy, there is a directed path from an imbalance measure ranked lower to an imbalance measure ranked higher.

As discussed earlier, there exists an imbalance measure which is sufficient to guarantee an unbiasedness of a BOSS estimator given a functional form of the response functions (e.g., $\mathscr{I}_{\text{DOM}}$ for linear response functions). If an imbalance measure is ranked higher than what is required by the response functions' functional form, then the imbalance measure is said to be *correct*. In the dissertation, it is be shown that $\mathscr{I}_{\text{ecdf:}\mathbf{D}}$ for a full joint distributional balance is correct for any functional form of response functions. Two additional imbalance measures ware introduced using a Cramer-von Mises test statistic.

Even with a correct imbalance measure, there can be a non-zero bias if the optimization problem for BOSS has non-zero optimum. For the optimization problem to have its optimal value zero, sufficient data with enough overlap are needed. In the dissertation, notions of "more overlap" and "enough overlap" are defined based on homogeneity of T and C.

Lastly, sub-optimality, which may occur from ineffective algorithm or time constraints, can also lead to bias in BOSS. If all the three issues (1.Use of incorrect imbalance measure; 2.Insufficient data; 3.Sub-optimality) are resolved, then there will not be any bias. In the dissertation, it is illustrated with numerical examples.

1.3.2 Treatment Effect Decomposition and Bootstrap Hypothesis Testing in BOSS

Heterogeneous treatment effects refer different treatment effects from different subgroups (Imai et al., 2013; Xie et al., 2012). Treatment effect estimate can be decomposed as a combination of heterogeneous treatment effects from its subsets, in particular, sets in a partition of the entire treated unit set and corresponding sets of control units. The decomposition technique that is proposed is my dissertation are different from the sub-classification when using observational data while the two methods coincide when using experimental data. Finding treatment effects of specific subgroups and understanding how those heterogeneous treatment effects are related with the treatment effects of an entire set are of interest.

Propensity score sub-classification, which is also known as stratification, uses propensity score to divide units into several groups so that units in the same strata can have similar propensity scores. Then the treatment effect estimate can be found by using a weighted average of the estimates from the sub-groups. On the other hand, the new method of decomposition with BOSS estimators for observational data use BOSS method to find control groups that is balanced with partitioned sets of the treatment group.

As a next step, a two-sample bootstrap hypothesis test is conducted to check the statistical significance of the BOSS estimators from the subsets. Denote the set of treated responses of units in T and the set of control responses of units in C' by $A = \{Y_t^1 \mid t \in T\}$ and $B = \{Y_c^0 \mid c \in C'\}$, respectively. Recall that there will be no bias if C' is from BOSS with zero $\mathscr{I}_{\text{DOM}}$ where the response functions are linear and the BOSS estimator of SATT is given by (1.4). If SATT estimate is zero, then $\mu_A = \mu_B$ should hold where $\mu_A = \frac{1}{|T|}\sum_{t\in T} Y_t^1$ and $\mu_B = \frac{1}{|C'|}\sum_{c\in C'} Y_c^0$. Hence, the following hypothesis H_0 and H_1 can be constructed to test for zero/non-zero treatment effect estimate given that C' is from BOSS with $\mathscr{I}_{\text{DOM}} = 0$ and no bias.

$$H_0 : \mu_A - \mu_B = 0 \text{ and } H_1 : \mu_A - \mu_B \neq 0. \tag{1.12}$$

In the dissertation, the testing procedure is then applied to LaLonde (1986) data. LaLonde (1986) analyzed a dataset from National Supported Work (NSW) Demonstration program. It is a labor training program conducted in 1970s and the dataset has been used by numerous researchers. In addition to the entire sample analysis described above, a sub-sample analysis is done. The p-values obtained by entire sample analysis and sub-sample analysis obtained by Bootstrap hypothesis

tests will be provided.

The technique developed here enables to identify whether a specific subgroup has a significant treatment effect. It can be also applied to any program evaluation data not only restricted to this labor training program which was taken for a demonstration in my dissertation.

1.3.3 BOSS under a Multi-Treament Setting

Matching method was introduced in a binary treatment setting (Cochran and Rubin, 1973; Rubin, 1973) and later it was extended in a multi-treatment setting (Imbens, 2000; Lechner, 2001; Yang et al., 2016). On the other hand, so far, BOSS estimators were only applied to datasets having binary treatment with treatment indicator 0 for control units and 1 for treated units. In this section, BOSS estimator that is applicable to datasets having multiple treatments is proposed. As in the binary treatment setting, BOSS under a multi-treatment setting is a non-matching technique which directly minimizes an extended imbalance measure and it does not require to find a matched pair for each treated unit.

Before introducing a new BOSS estimator for multiple treatments, a matching method for the multi-treatment setting by Yang et al. (2016) is reviewed. Then a new BOSS estimator for multiple treatments is proposed and it is shown that the proposed estimator is unbiased for the expected value of the sample average treatment effect under certain conditions.

In the dissertation, a computational result demonstrating that the proposed BOSS estimator outperforms the other matching estimators in terms of the size of bias on a simulated dataset is provided.

1.3.4 Handling Missing Data in Observational Studies with BOSS

In this part of the dissertation, methods that are applicable when there is a missing covariate vector or some missing entries in the set of covariates are discussed. Both matching and BOSS relies on strong ignorability assumption as mentioned earlier. In this dissertation, a sensitivity analysis of BOSS estimators is conducted to investigate how the estimated values change when the conditional independence (unconfoundedness) assumption is violated because of a missing covariate vector. To conduct a sensitivity analysis, a number of hidden/unmeasured covariate vec-

tors were generated based on some parameter values using the method proposed and implemented by Ichino et al. (2008) and Nannicini (2007). How this addition affects the BOSS estimates and its standard error compared to matching estimates are investigated.

In addition, two different BOSS methods that are applicable after multiple imputation on a dataset with missing entries in covariates are compared. Mitra and Reiter (2016) studied two propensity score matching methods (namely, Across and Within approaches) and showed that Across approach leads to a smaller bias than Within approach. Across approach estimates the propensity score, averages the propensity scores from the multiple datasets whose missing values are imputed, and finds a single treatment effect estimate. On the other hand, Within approach conducts the propensity score matching to find a treatment effect estimate on each imputed dataset and finds a mean of the estimates. Similar approaches in estimation using BOSS are defined and examined. The performance of two BOSS methods are compared to each other as well as to the corresponding methods in matching.

1.3.5 Duality in BOSS

Last part of the dissertation studies a dual problem of the BOSS formulation. As stated earlier, BOSS can be formulated as an MIP. The integer constraints in the MIP can be relaxed and the BOSS can be formulated as an LP if a fractional contribution of control units in the optimal control group is allowed. Accordingly, a dual problem of the LP can be found. Since understanding the meaning of the variables and constraints in the dual problem is important, it is investigated as a last part of the dissertation.

1.3.6 Outline

This dissertation is organized in the order that is mentioned above: Chapter 2 discusses on bias in Balance Optimization Subset Selection, Chapter 3 discusses on treatment effect decomposition and bootstrap hypothesis testing in observational studies, Chapter 4 discusses on BOSS under a multi-treatment setting, Chapter 5 discusses on handling missing data with BOSS, Chapter 6 discusses on duality in BOSS, and Chapter 7 concludes.

CHAPTER 2

BIAS IN BALANCE OPTIMIZATION SUBSET SELECTION

2.1 Introduction

The Balance Optimization Subset Selection (BOSS) framework was proposed by Nikolaev et al. (2013) as an alternative to matching methods for causal inference using observational data. Due to advances in computing technology, this optimization approach using Mixed Integer Programming (MIP) was enabled. This approach directly minimizes an imbalance measure while traditional matching methods do so indirectly.

Recently, Zubizarreta (2012) formulated a matching method as an MIP with a term directly minimizing imbalance which differs from traditional matching methods. The formulation for BOSS is similar to this matching method as an MIP but it differs in that BOSS does not require each treatment unit to be matched with a control unit.

BOSS finds a control group that is more balanced (i.e., a control group that is more similar to the treatment group in its covariate value distribution) than those identified by traditional matching methods because BOSS directly minimizes a given imbalance measure. In this chapter, selection bias is defined as the difference in mean control response of the treatment units and that of the selected control units. Within the BOSS framework, eliminating this bias as much as possible is one of the aims of this chapter.

There have been numerous papers that have studied bias in causal analysis. Unlike randomized experiments which select treated and control units at random, bias from systematic differences in these two different groups of units is inevitable in observational studies. Rubin (1973) viewed matching as a way of removing and controlling the bias. Rosenbaum and Rubin (1985) explored the bias which arises from incomplete matching. Heckman et al. (1998) characterized the selection bias by decomposing it. They tested assumptions for matching methods and showed

which component of the bias is associated with matching. Abadie and Imbens (2012) introduced matching estimators that are bias-corrected.

As stated, bias associated with the matching estimator has been studied by many researchers. However, a study of the bias in the treatment effect estimator based on the BOSS framework is also necessary. What causes bias in BOSS and how to reduce it are considered in this chapter.

When using Balance Optimization Subset Selection, bias in the treatment effect estimator occurs as a result of the following cases. First, it can occur when an incorrect imbalance measure that does not comply with a functional form of the response function is used. Second, even with a correct imbalance measure, bias can be generated when there are insufficient data leading to residual imbalance between the outcome groups. Non-zero bias can also result when an optimal value of zero cannot be found because of technical issues like ineffective algorithms and time constraints. It is called suboptimality in this chapter. Examples corresponding to each of these cases generating bias will be provided.

The chapter is organized as follows. Section 2.2 investigates the relationship between bias and the imbalance measure used by BOSS. Section 2.3 defines the concept of the Balance Hierarchy and discusses how to identify the correct imbalance measure. Section 2.4 provides examples illustrating the relationship between bias and imbalance measures. Section 2.5 further studies the cases where residual balance remains after optimization due to insufficient data and/or suboptimality from technical issues. Section 2.6 combines the requirements to guarantee zero bias from the BOSS estimator. Section 2.7 provides concluding comments.

2.2 Relationship between Bias and Imbalance Measure

2.2.1 Role of Covariate Balance for Reducing Bias in the Treatment Effect Estimator

This section studies the relationship between bias and an imbalance measure. Examples illustrating this relationship is provided in this section and Section 2.4. Before going on to the examples, first recall that the sample average treatment effect for the treated (SATT) is defined as (1.3).

Further recall that the treated and untreated responses for unit u are of the forms $Y_u^1 = h^1(\mathbf{X}_u) + \epsilon_u^1$ and $Y_u^1 = h^0(\mathbf{X}_u) + \epsilon_u^0$ for a treatment response function $h^1(\cdot)$ and a control response function $h^0(\cdot)$. Then the difference between the estimated treatment effect, $\widetilde{\tau}_T^1(C')$, and the SATT, τ_T^1, as the sum of selection bias and the error terms can be written as follows.

$$\begin{aligned}
\widetilde{\tau}_T^1(C') - \tau_T^1 &= \left(\frac{1}{|T|}\sum_{t\in T} Y_t^1 - \frac{1}{|C'|}\sum_{c\in C'} Y_c^0\right) - \left(\frac{1}{|T|}\sum_{t\in T}\left(Y_t^1 - Y_t^0\right)\right) \\
&= \left(\frac{1}{|T|}\sum_{t\in T} Y_t^0 - \frac{1}{|C'|}\sum_{c\in C'} Y_c^0\right) \\
&= \underbrace{\left(\frac{1}{|T|}\sum_{t\in T} h^0(\mathbf{X}_t) - \frac{1}{|C'|}\sum_{c\in C'} h^0(\mathbf{X}_c)\right)}_{\mathcal{B}(T,C')\text{ : selection bias}} + \underbrace{\left(\frac{1}{|T|}\sum_{t\in|T|} \epsilon_t^0 - \frac{1}{|C'|}\sum_{c\in C'} \epsilon_c^0\right)}_{\mathcal{E}(T,C')\text{: error terms}}
\end{aligned} \tag{2.1}$$

Here it is assumed that the error term ϵ_u^0 for any unit u is zero in expectation. Hence, in expectation, the difference of the estimated treatment effect and the SATT, $\widetilde{\tau}_T^1(C') - \tau_T^1$, is reduced to the selection bias, $\mathcal{B}(T, C')$, which is defined as the control response function mean of the treatment units minus that of the control units.

On the response functions $h^0(\mathbf{X}_u)$ and $h^1(\mathbf{X}_u)$, the BOSS framework does not have any specific requirements other than those required for unbiasedness based on imbalance measures. These two response functions need not be the same. Heterogeneity in the effects for treatment and control units will be identified after conducting BOSS from a non-zero treatment effect estimate.

In the special case where the treatment response function is the same as the control response function, the following relationship between the estimated ATT and selection bias holds. Under this case that $h^0(\mathbf{X}_u) = h^1(\mathbf{X}_u)$, the SATT, τ_T^1, is zero in expectation. Accordingly, the expected value of the estimated treatment effect is the same as the expected value of selection bias. That is,

$$\mathbb{E}\left[\widetilde{\tau}_T^1(C')\right] = \mathbb{E}\left[\frac{1}{|T|}\sum_{t\in T} Y_t^1 - \frac{1}{|C'|}\sum_{c\in C'} Y_c^0\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\frac{1}{|T|}\sum_{t\in T} h^1(\mathbf{X}_t) - \frac{1}{|C'|}\sum_{c\in C'} h^0(\mathbf{X}_c)\right] + \mathbb{E}\left[\frac{1}{|T|}\sum_{t\in T}\epsilon_t^1 - \frac{1}{|C'|}\sum_{c\in C'}\epsilon_c^0\right] \\
&= \mathbb{E}\left[\frac{1}{|T|}\sum_{t\in T} h^0(\mathbf{X}_t) - \frac{1}{|C'|}\sum_{c\in C'} h^0(\mathbf{X}_c)\right] \\
&= \mathbb{E}\left[\mathcal{B}(T, C')\right]
\end{aligned}
\tag{2.2}
$$

In BOSS, the objective function of the optimization problem is defined as some measure of covariate imbalance. A non-zero objective value, in general, leads to non-zero bias. However, zero bias may be obtained even though there is residual imbalance in the full joint distribution. Examples 1 and 2 illustrate those possibilities.

Example 1. Consider the three covariate case with the untreated response function $h^0(\mathbf{X}_u) = 1.4X_{u,1} + 1.3X_{u,2} + 0.9X_{u,3}$ for both treatment and control individuals. Then the bias is defined by

$$
\mathcal{B}(T, C') \equiv \frac{\sum_{t\in T} h^0(\mathbf{X}_t)}{|T|} - \frac{\sum_{c\in C'} h^0(\mathbf{X}_c)}{|C'|}. \tag{2.3}
$$

Given the linear response function,

$$
\begin{aligned}
\mathcal{B}(T, C') =& \frac{\sum_{t\in T}(1.4X_{t,1} + 1.3X_{t,2} + 0.9X_{t,3})}{|T|} - \frac{\sum_{c\in C'}(1.4X_{c,1} + 1.3X_{c,2} + 0.9X_{c,3})}{|C'|} \\
=& 1.4(\overline{X_{T,1}} - \overline{X_{C',2}}) + 1.3(\overline{X_{T,2}} - \overline{X_{C',1}}) + 0.9(\overline{X_{T,3}} - \overline{X_{C',3}}),
\end{aligned}
\tag{2.4}
$$

where $\overline{X_{G,k}} = \sum_{u\in G} X_{u,k}/|G|$ for $G \in \{T, C'\}$.

Under some generic form for the response function, it is common to have both non-zero objective value and non-zero bias. To illustrate this case of non-zero objective with a particular objective function as defined in (2.6), it is necessary to introduce some additional notation.

Let $\mathcal{P} = \{1, 2, 3\}$ be the set of covariate indices. Denote the set of values that the k-th covariate can have by N_k and suppose that the possible values for each covariate are 1 and 2: $N_k = \{1, 2\}$ for $k \in \mathcal{P}$. Let $E_{k,j}$ denote the bin of units whose k-th covariate is equal to j, i.e., $E_{k,j}$ is the set of units that have the value j for their k-th covariate where $k \in \mathcal{P}$, and let $\eta_{k,j}(G) \equiv |E_{k,j} \cap G|$ for $G \in \{T, C'\}$. For each covariate, the first bin with $j = 1$ is for covariate value 1 and the second bin

with $j = 2$ is for covariate value 2. Assume that the distributions of the treatment group T and the selected control group C' satisfy the following:

$$\begin{aligned}
&\eta_{1,1}(T) = 100, \quad \eta_{1,2}(T) = 50,\\
&\eta_{2,1}(T) = 50, \quad \eta_{2,2}(T) = 100,\\
&\eta_{3,1}(T) = 75, \quad \eta_{3,2}(T) = 75,\\
&\eta_{1,1}(C') = 100, \quad \eta_{1,2}(C') = 50,\\
&\eta_{2,1}(C') = 48, \quad \eta_{2,2}(C') = 102,\\
&\eta_{3,1}(C') = 76, \quad \eta_{3,2}(C') = 74.
\end{aligned} \tag{2.5}$$

For a subset of covariate indices, $\mathcal{K} \subset \mathcal{P}$, use the objective function from Sauppe et al. (2014) p.551:

$$\mathscr{I}_{\mathcal{K}}(T, C') = \sum_{k \in \mathcal{K}} \sum_{j \in N_k} |\eta_{k,j}(T) - \eta_{k,j}(C')| \tag{2.6}$$

Here $\mathcal{K} = \mathcal{P} = \{1, 2, 3\}$. The objective value that is computed based on (2.6) is given by $0+0+2+2+1+1 = 6$ and the bias is given by $\mathcal{B}(T, C') = -17/1500$ since $\overline{X_{T,1}} = \overline{X_{C',1}} = (100+2\cdot 50)/150$, $\overline{X_{T,2}} = (50+2\cdot 100)/150$, $\overline{X_{C',2}} = (48+2\cdot 102)/150$, $\overline{X_{T,3}} = (75 + 2 \cdot 75)/150$, and $\overline{X_{C',3}} = (76 + 2 \cdot 74)/150$.

Example 2. Note that sometimes zero bias may be obtained even in the case when the objective value is not zero. To illustrate, consider an example with the same $\eta_{k,j}$ as above but with a different response function $h^0(\mathbf{X}_u) = 1.4X_{u,1} + 1.3X_{u,2} + 2.6X_{u,3}$. In this case, a zero-bias is obtained although the objective function is non-zero.

The relationship between the bias and the objective function value after optimization is summarized in Table 2.1. All four scenarios can happen as indicated. Example 3 and 4 corresponding to zero-objective value will be introduced later in Section 2.4.1.

Table 2.1: Relationship between objective function value and the bias

	zero objective value	non-zero objective value
zero bias	With a correct imbalance measure: Example 3 on page 25	Example 2 on page 17
non-zero bias	With an incorrect imbalance measure: Example 4 on page 27	Example 1 on page 16

Sauppe and Jacobson (2017) shows that covariate balance can be used to guarantee unbiasedness of treatment effect estimates under certain assumptions on the control response function which relates the covariates to the control responses. The required imbalance measures for several different functional forms of the control response function are summarized below.

Let $\mathcal{K}$ denote a subset of the set of all the covariate indices $\mathcal{P} = \{1, 2, \cdots, K\}$. If the response function has a term of the form $\sum_{k\in\mathcal{K}} \beta_k X_{u,k}$ for a unit u, then an imbalance measure that includes the following term is needed:

$$\mathscr{I}_{\text{DOM:}\mathcal{K}}(T, C') = \sum_{k\in\mathcal{K}} \left| \frac{1}{|T|} \sum_{t\in T} X_{t,k} - \frac{1}{|C'|} \sum_{c\in C'} X_{c,k} \right|. \tag{2.7}$$

If the response function has a term of the form $\gamma_{\mathcal{K}} \prod_{k\in\mathcal{K}} (X_{u,k})^{p_k}$ for some constants $p_k \in \mathbb{R}$ for $k \in \mathcal{K}$ and $\gamma_{\mathcal{K}} \in \mathbb{R}$, then an imbalance measure that includes the difference in corresponding correlation terms as in the following equation is needed:

$$\mathscr{I}_{\text{Corr:}\mathcal{K}}(T, C') = \sum_{k\in\mathcal{K}} \left| \frac{1}{|T|} \sum_{t\in T} \prod_{k\in\mathcal{K}} (X_{t,k})^{p_k} - \frac{1}{|C'|} \sum_{c\in C'} \prod_{k\in\mathcal{K}} (X_{c,k})^{p_k} \right|. \tag{2.8}$$

When a response function is *separable* for each covariate, an imbalance measure that balances the marginal distribution of those covariates such as $\mathscr{I}_{\text{KS}}$ or $\mathscr{I}_{\text{CvM}}$, which will be discussed in Section 2.3.1, will be needed. (Definition of a separable response function and a non-separable response function will be explained in page 21.)

2.3 Balance Hierarchy and Correct Imbalance Measure

In this section, correct and incorrect imbalance measures will be defined. First, a concept of balance hierarchy is introduced.

Definition 1. An imbalance measure $\mathscr{I}_1$ is said to have a *higher rank in the balance hierarchy* than an imbalance measure $\mathscr{I}_2$ if $\mathscr{I}_1 = 0$ implies $\mathscr{I}_2 = 0$. In other words, $\mathscr{I}_1$ is more highly ranked than $\mathscr{I}_2$ if $\mathscr{I}_1$ requires more balance than $\mathscr{I}_2$.

Consider the following imbalance measures including the ones used in the previous examples.

For covariates $k = 1, 2, \cdots, K$,

- Difference of Means:

$$\mathscr{I}_{\text{DOM}}(T, C') = \sum_{k=1}^{K} \left| \frac{1}{|T|} \sum_{t \in T} X_{t,k} - \frac{1}{|C'|} \sum_{c \in C'} X_{c,k} \right| \tag{2.9}$$

- Difference of First and Second Moments:

$$\mathscr{I}_{\text{DOM+DOV}}(T, C') = \mathscr{I}_{\text{DOM}} + \sum_{k=1}^{K} \left| \frac{1}{|T|} \sum_{t \in T} (X_{t,k})^2 - \frac{1}{|C'|} \sum_{c \in C'} (X_{c,k})^2 \right| \tag{2.10}$$

- Bivariate Moment:

$$\mathscr{I}_{\text{DOM2}}(T, C') = \mathscr{I}_{\text{DOM+DOV}} + \sum_{(k_1,k_2) \in \binom{\mathcal{P}}{2}} \left| \frac{1}{|T|} \sum_{t \in T} X_{t,k_1} X_{t,k_2} - \frac{1}{|C'|} \sum_{c \in C'} X_{c,k_1} X_{c,k_2} \right| \tag{2.11}$$

  where $\binom{\mathcal{P}}{2}$ denotes a set of all the possible covariate index pairs.

- Kolmogorov-Smirnov Test Statistic:

$$\mathscr{I}_{\text{KS}}(T, C') = \sum_{k=1}^{K} \max_{x \in \mathscr{X}_k(T \cup C')} \left| \widehat{F}_k(T, x) - \widehat{F}_k(C', x) \right| \tag{2.12}$$

  where $\mathscr{X}_k(S) \equiv \{X_{u,k} : u \in S\}$ denotes the set of possible values for the k-th covariate value for all units $u \in S$, and $\widehat{F}_k(T, x)$ is the empirical distribution function of the treatment group T while $\widehat{F}_k(C', x)$ is that of the control group C'. That is, $\widehat{F}_k(T, x) = \left|\{u \in T : X_{u,k} \leq x\}\right|/|T|$ and $\widehat{F}_k(C', x) = \left|\{u \in C' : X_{u,k} \leq x\}\right|/|C'|$.

- Joint Distribution:

$$\mathscr{I}_{\text{ecdf:}\mathbf{D}}(T, C') = \sum_{D \in \mathbf{D}} \max_{x \in \mathscr{X}_D(T \cup C')} \left| \widehat{F}_D(T, x) - \widehat{F}_D(C', x) \right| \tag{2.13}$$

  where $\mathbf{D}$ denotes the set of all possible covariate clusters. A covariate cluster denotes a set of covariate indices and thus $\mathbf{D}$ is the power set of $\mathcal{P} = \{1, 2, \cdots, K\}$.

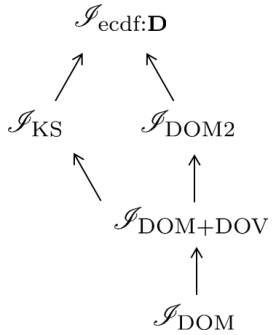
According to Definition 1, the imbalance measure $\mathscr{I}_{\text{DOM+DOV}}$ is more highly ranked in the balance hierarchy than the imbalance measure $\mathscr{I}_{\text{DOM}}$ because it requires more balance between the matched groups to achieve zero objective value when using this imbalance measure as an objective. Similarly, $\mathscr{I}_{\text{DOM2}}$ is more highly ranked than both $\mathscr{I}_{\text{DOM}}$ and $\mathscr{I}_{\text{DOM+DOV}}$. As additional terms to be balanced are included, the corresponding imbalance measure will be ranked higher and higher.

In addition, $\mathscr{I}_{\text{KS}}$ is more highly ranked in the balance hierarchy than both imbalance measures $\mathscr{I}_{\text{DOM+DOV}}$ and $\mathscr{I}_{\text{DOM}}$ since $\mathscr{I}_{\text{KS}}$ requires the entire marginal distribution of the covariates to be balanced while the others only require a balance in first and second moments.

In that sense, the imbalance measure $\mathscr{I}_{\text{ecdf:}\mathbf{D}}$ in (2.13) would be among the most highly ranked imbalance measures since it requires a balance in joint distributions while the other measures such as $\mathscr{I}_{\text{DOM}}$, $\mathscr{I}_{\text{DOM+DOV}}$, and $\mathscr{I}_{\text{KS}}$ require balance on marginal distributions only and $\mathscr{I}_{\text{DOM2}}$ balances bivariate moments only among the many other possible combinations of clusters for balancing.

Note that some ranks of imbalance measures are not comparable. For example, there does not exist a strict ranking between a middle ranked imbalance measure and an imbalance measure created by combining a low ranked imbalance measure with a high ranked imbalance measure. Additionally, among the above examples, the rank of $\mathscr{I}_{\text{DOM2}}$ and $\mathscr{I}_{\text{KS}}$ cannot be compared since $\mathscr{I}_{\text{DOM2}}$ has a correlation term that $\mathscr{I}_{\text{KS}}$ does not balance while the former does not balance the entire marginal distribution as the latter does.

Figure 2.1: Balance Hierarchy Example



The ranks of the above imbalance measures are summarized in Figure 2.1. The arrow indicates a direction from a low ranked imbalance measure to a high ranked imbalance measure. If there does not exist a directed path from one imbalance

measure to another imbalance measure, then the ranks of the two imbalance measures are not comparable.

From the way that each balance measure is defined, it can be easily shown that the following relation holds:

$$\mathscr{I}_{\text{ecdf:}\mathbf{D}} = 0 \Longrightarrow \mathscr{I}_{\text{KS}} = 0 \tag{2.14a}$$

$$\mathscr{I}_{\text{ecdf:}\mathbf{D}} = 0 \Longrightarrow \mathscr{I}_{\text{DOM2}} = 0 \tag{2.14b}$$

$$\mathscr{I}_{\text{KS}} = 0 \Longrightarrow \mathscr{I}_{\text{DOM+DOV}} = 0 \tag{2.15}$$

$$\mathscr{I}_{\text{DOM2}} = 0 \Longrightarrow \mathscr{I}_{\text{DOM+DOV}} = 0 \tag{2.16}$$

$$\mathscr{I}_{\text{DOM+DOV}} = 0 \Longrightarrow \mathscr{I}_{\text{DOM}} = 0 \tag{2.17}$$

Equivalently, the following holds and it motivates a definition of *correct* balance measure.

$$\mathscr{I}_{\text{DOM}} \neq 0 \Longrightarrow \mathscr{I}_{\text{DOM+DOV}} \neq 0 \Longrightarrow \mathscr{I}_{\text{DOM2}} \neq 0 \Longrightarrow \mathscr{I}_{\text{ecdf:}\mathbf{D}} \neq 0 \tag{2.18}$$

$$\mathscr{I}_{\text{KS}} \neq 0 \Longrightarrow \mathscr{I}_{\text{ecdf:}\mathbf{D}} \neq 0 \tag{2.19}$$

Definition 2. An imbalance measure is said to be *correct* for a given problem if it is equally or more highly ranked in the balance hierarchy than the imbalance measure that is required by the functional form of the response function to ensure that the bias equals zero.

In addition, consider a case with a non-separable response function. A response function is said to be separable if it can be written in the form $f(X_{u,k_1}) + f(X_{u,k_2}) + \cdots + f(X_{u,k_m})$ for individual covariates in a set $\mathcal{K} = \{k_1, \cdots k_m\} \subset \mathcal{P} = \{1, 2, \cdots, K\}$ where $\mathcal{K}$ is the set of covariate indices that constitutes the response function. The response function is non-separable if it is not separable.

If the response function is non-separable and is a function of covariates that belong to a certain covariate cluster $\mathcal{K}$, it is necessary to balance the joint distribution of those covariates using the more generalized version of $\mathscr{I}_{\text{KS}}$ like (2.13) or that of $\mathscr{I}_{\text{CvM}}$ like (2.22) defined with a power set of $\mathcal{K}$ in the next section, Section 2.3.1.

Theorem 1. *(Corollary 13 on p.332 of Sauppe and Jacobson, 2017) If the imbalance measure for the entire joint distribution of covariates is* 0*, then zero bias is guaranteed for any functional form of the response function. That is,*

$$\mathscr{I}_{ecdf:\boldsymbol{D}}(T, C') = \sum_{D \in \boldsymbol{D}} max_{x \in \mathscr{X}_D(T \cup C')} \left| \widehat{F}_D(T, x) - \widehat{F}_D(C', x) \right| = 0 \tag{2.20}$$

will imply that $\mathcal{B}(T, C') = 0$ *regardless of the response function's form, where* $\mathscr{X}_D(S)$ *is the set of possible values of the covariate combination D for all units u in S,* $\widehat{F}_D(T, x)$ *is the empirical cumulative density function of the treatment group T, and* $\widehat{F}_D(C', x)$ *is that of the control group* C'.

Theorem 2. *Assuming that there are no unobserved covariates, the imbalance measure* $\mathscr{I}_{ecdf:\boldsymbol{D}}$*, which is the most highly ranked in the balance hierarchy, is correct for any functional form of the response function where* $\boldsymbol{D}$ *includes all possible covariate clusters.*

Proof. This follows from Definition 2 and Theorem 1. □

The above theorems are discussed with the full joint distribution because balance on the joint distribution ensures balance on any other sub-joint distributions. For example, balance on the joint distribution of covariates $\mathcal{P} = \{1, 2, 3\}$ ensures balance on the entire set $\mathcal{K}_1 = \{1, 2, 3\}$, and balance on the pairwise-joints $\mathcal{K}_2 = \{1, 2\}$, $\mathcal{K}_3 = \{1, 3\}$, and $\mathcal{K}_4 = \{2, 3\}$ in addition to balance on the marginals $\mathcal{K}_5 = \{1\}$, $\mathcal{K}_6 = \{2\}$, and $\mathcal{K}_7 = \{3\}$.

2.3.1 Another Imbalance Measure for Balancing the Distribution

This section introduces a new imbalance measure using the Cramer-von Mises criterion, to compare the resulting covariate distributions of the treatment group and the control group in addition to various imbalance measures used in Nikolaev et al. (2013), Sauppe et al. (2014), and Sauppe and Jacobson (2017).

The following test statistic for two sample comparison is taken from Anderson (1962). If this value is greater than a certain tabulated value, then reject the hypothesis that the empirical distributions of the two samples are the same. Hence it can be a criterion testing whether covariate balance is achieved between the two groups.

- Cramer-von Mises Test Statistic

$$\mathscr{I}_{\text{CvM}}(T, C') = \frac{|T| \cdot |C'|}{(|T| + |C'|)^2} \sum_{k=1}^{K} \left(\sum_{x \in \mathscr{X}_k(T \cup C')} (\widehat{F}_k(T, x) - \widehat{F}_k(C', x))^2 \right) \tag{2.21}$$

  where $\widehat{F}_k$ denotes the empirical distribution function as in (2.12).

In addition, consider the joint distribution version of this Cramer-von Mises Test Statistic using the empirical cumulative density functions for all possible covariate clusters $D \in \mathbf{D}$:

$$\mathscr{I}_{\text{CvM:}\mathbf{D}}(T, C') = \frac{|T| \cdot |C'|}{(|T| + |C'|)^2} \sum_{D \in \mathbf{D}} \left(\sum_{x \in \mathscr{X}_D(T \cup C')} (\widehat{F}_D(T, x) - \widehat{F}_D(C', x))^2 \right) \tag{2.22}$$

The following theorem shows the relationship between $\mathscr{I}_{\text{KS}}$ and $\mathscr{I}_{\text{CvM}}$: they are equally ranked in the balance hierarchy.

Theorem 3. *The imbalance measure $\mathscr{I}_{KS}$ is equal to zero if and only if $\mathscr{I}_{CvM}$ is equal to zero.*

Proof.

$$\mathscr{I}_{\text{KS}}(T, C') = \sum_{k=1}^{K} \max_{x \in \mathscr{X}_k(T \cup C')} \left| \widehat{F}_k(T, x) - \widehat{F}_k(C', x) \right| = 0 \tag{2.23}$$

$$\iff \widehat{F}_k(T, x) - \widehat{F}_k(C', x) = 0 \quad \forall x \in \mathscr{X}_k(T \cup C') \tag{2.24}$$

$$\iff (\widehat{F}_k(T, x) - \widehat{F}_k(C', x))^2 = 0 \quad \forall x \in \mathscr{X}_k(T \cup C') \tag{2.25}$$

$$\iff \mathscr{I}_{\text{CvM}}(T, C') = \frac{|T| \cdot |C'|}{(|T| + |C'|)^2} \sum_{k=1}^{K} \left(\sum_{x \in \mathscr{X}_k(T \cup C')} (\widehat{F}_k(T, x) - \widehat{F}_k(C', x))^2 \right) = 0. \tag{2.26}$$

□

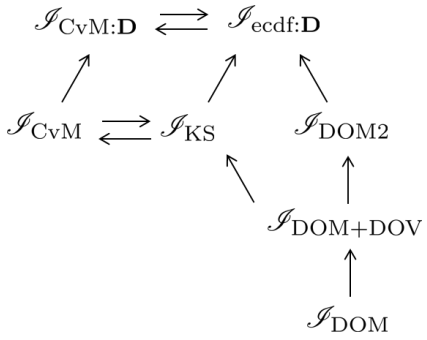
Similarly, Theorem 4 demonstrates that $\mathscr{I}_{\text{ecdf:}\mathbf{D}}$ and $\mathscr{I}_{\text{CvM:}\mathbf{D}}$ are equally ranked in the balance hierarchy.

Theorem 4. *The imbalance measure $\mathscr{I}_{ecdf:\mathbf{D}}$ is equal to zero if and only if the imbalance measure $\mathscr{I}_{CvM:\mathbf{D}}$ is equal to zero.*

Proof. The proof of Theorem 4 is the same as the proof of Theorem 3 when $x \in \mathscr{X}_k(T \cup C')$ is replaced by $x \in \mathscr{X}_D(T \cup C')$. □

Theorem 3 tells that $\mathscr{I}_{\mathrm{CvM}}$ has the same rank in the balance hierarchy as $\mathscr{I}_{\mathrm{KS}}$. Similarly, from Theorem 4, $\mathscr{I}_{\mathrm{ecdf:}\mathbf{D}}$ is ranked the same as $\mathscr{I}_{\mathrm{CvM:}\mathbf{D}}$. Accordingly, the balance hierarchy in Figure 2.1 can be extended to the directed graph as in Figure 2.2.

Figure 2.2: Another Balance Hierarchy Example



As in Theorem 4, the imbalance measure $\mathscr{I}_{\mathrm{CvM:}\mathbf{D}}$ is also correct for any functional form of the response function. However, the imbalance measure $\mathscr{I}_{\mathrm{CvM:}\mathbf{D}}$ as well as $\mathscr{I}_{\mathrm{ecdf:}\mathbf{D}}$ is not often used in practice since it is extremely difficult to achieve an objective value of zero with these measures. In particular, as the number of covariates increases, the number of possible covariate clusters increases exponentially and thus it becomes more difficult to achieve an optimal value of zero. Such cases yielding a non-zero optimal value although the imbalance measure is correct will be discussed in Section 2.5. Before that, examples with incorrect imbalance measures are provided in the following section.

2.4 Examples

2.4.1 Examples with Incorrect Imbalance Measures

Use of DOM Imbalance Measure When The Response Function Is Not Linear

The first example provided in this section corresponds to a case with a correct imbalance measure where zero objective value implies zero bias.

Example 3. Consider the Difference of Means (DOM) imbalance measure in (2.9):

$$\mathscr{I}_{\text{DOM}}(T, C') = \sum_{k=1}^{K} \left| \frac{1}{|T|} \sum_{t \in T} X_{t,k} - \frac{1}{|C'|} \sum_{c \in C'} X_{c,k} \right| \tag{2.27}$$

Suppose that $K = 1$, and both the treatment group and the control group are composed of 100 units respectively. Say, $T = \{t_1, \cdots, t_{100}\}$ and $C' = \{c_1, \cdots, c_{100}\}$. Moreover, suppose that

$$X_{t,1} = \begin{cases} 50 \text{ for } 1 \leq i \leq 90, t_i \in T \\ 25 \text{ for } 91 \leq i \leq 95, t_i \in T \\ 75 \text{ for } 96 \leq i \leq 100, t_i \in T \end{cases} \tag{2.28}$$

and

$$X_{c,1} = \begin{cases} 50 \text{ for } 1 \leq i \leq 10, c_i \in C' \\ 25 \text{ for } 11 \leq i \leq 55, c_i \in C' \\ 75 \text{ for } 56 \leq i \leq 100, c_i \in C' \end{cases} \tag{2.29}$$

Then

$$\begin{aligned} \mathscr{I}_{\text{DOM}}(T, C') &= \left| \frac{1}{|T|} \sum_{t \in T} X_{t,1} - \frac{1}{|C'|} \sum_{c \in C} X_{c,1} \right| \\ &= \left| \left(\frac{50 \cdot 90 + 25 \cdot 5 + 75 \cdot 5}{100} - \frac{50 \cdot 10 + 25 \cdot 45 + 75 \cdot 45}{100} \right) \right| \\ &= 0 \end{aligned} \tag{2.30}$$

When the response function is of the form $h(\mathbf{X}) = \beta^T \mathbf{X} + \alpha$ then an objective value of zero will imply zero bias. This is a direct consequence of Theorem 5.

Theorem 5. *(Theorem 7 on p.330 of Sauppe and Jacobson, 2017) Having the DOM imbalance measure equal zero is necessary and sufficient for the bias to equal to zero for all possible linear response functions (i.e., $h^0(\mathbf{X}_u) = \beta^T \mathbf{X}_u + \alpha = \sum_{k \in \mathcal{P}} \beta_k X_{u,k} + \alpha$ for all $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^K$ where the set of all the covariate indices is $\mathcal{P} = \{1, 2, \cdots, K\}$).*

Note that the condition $\mathscr{I}_{\text{DOM}} = 0$ is sufficient but not necessary for $\mathcal{B}(T, C') = 0$ when the response is linear with fixed α and β. As an example, consider the following case with two covariates ($K = 2$). Suppose that there are 100 units respectively in the treatment group T and the selected control group C'. Denote

the treatment units as $t_1, t_2, \cdots, t_{100}$ and the control units as $c_1, c_2, \cdots, c_{100}$. Let the covariate values be given by

$$\mathbf{X}_{t_i} = \begin{cases} (1,0) & \text{for } 1 \le i \le 45 \\ (0,-1) & \text{for } 46 \le i \le 90 \\ (2,3) & \text{for } 91 \le i \le 95 \\ (-3,-2) & \text{for } 96 \le i \le 100 \end{cases} \tag{2.31}$$

for the treatment units and

$$\mathbf{X}_{c_i} = \begin{cases} (1,0) & \text{for } 1 \le i \le 5 \\ (0,-1) & \text{for } 6 \le i \le 10 \\ (2,3) & \text{for } 11 \le i \le 55 \\ (-3,-2) & \text{for } 56 \le i \le 100 \end{cases} \tag{2.32}$$

for the control units.

Suppose that $h^0(\mathbf{X}_u) = X_{u,1} + X_{u,2}$ is the response function of both groups (i.e., $\alpha = 0$ and $\beta = (1,1) \in \mathbb{R}^2$). Then

$$\begin{aligned} &\frac{1}{|T|}\sum_{t\in T} X_{t,1} - \frac{1}{|C'|}\sum_{c\in C'} X_{c,1} \\ &= \frac{1\cdot 45 + 0\cdot 45 + 2\cdot 5 - 3\cdot 5}{100} - \frac{1\cdot 5 + 0\cdot 5 + 2\cdot 45 - 3\cdot 45}{100} = 0.8 \neq 0 \end{aligned} \tag{2.33}$$

and

$$\begin{aligned} &\frac{1}{|T|}\sum_{t\in T} X_{t,2} - \frac{1}{|C'|}\sum_{c\in C'} X_{c,2} \\ &= \frac{0\cdot 45 - 1\cdot 45 + 3\cdot 5 - 2\cdot 5}{100} - \frac{0\cdot 5 - 1\cdot 5 + 3\cdot 45 - 2\cdot 45}{100} = -0.8 \neq 0 \end{aligned} \tag{2.34}$$

hence

$$\mathscr{I}_{\text{DOM}} = 1.6 \neq 0. \tag{2.35}$$

However,

$$\begin{aligned}\mathcal{B}(T, C') &= \frac{1}{|T|}\sum_{t\in T}(X_{t,1} + X_{t,2}) - \frac{1}{|C'|}\sum_{c\in C'}(X_{c,1} + X_{c,2}) \\ &= \frac{1\cdot 45 - 1\cdot 45 + 5\cdot 5 - 5\cdot 5}{100} - \frac{1\cdot 5 - 1\cdot 5 + 5\cdot 45 - 5\cdot 45}{100} = 0.\end{aligned} \tag{2.36}$$

Now consider the case when the response function is not linear (e.g., $h^0(\mathbf{X}) = \sum_{k\in\mathcal{K}}\beta_k X_{u,k} + \sum_{k\in\mathcal{K}}\gamma_k(X_{u,k})^2 + \sum_{(k_1,k_2)\in\binom{\mathcal{K}}{2}}\gamma_{k_1,k_2}X_{u,k_1}X_{u,k_2} + \alpha$). In that case, a value of zero from the DOM imbalance measure does not imply zero bias.

Example 4. Consider the response function of the quadratic form:

$$h^0(\mathbf{X}_u) = \sum_{k\in\mathcal{K}} X_{u,k} + \sum_{k\in\mathcal{K}} X_{u,k}^2 = X_{u,1} + X_{u,1}^2 \tag{2.37}$$

where the covariates are given by (2.28) and (2.29) as above. Then the bias is given by

$$\begin{aligned}&\mathcal{B}(T, C') \\ &= \frac{1}{|T|}\sum_{t\in T} h^0(\mathbf{X}_t) - \frac{1}{|C'|}\sum_{c\in C'} h^0(\mathbf{X}_c) \\ &= \frac{2550\cdot 90 + 650\cdot 5 + 5700\cdot 5}{100} - \frac{2550\cdot 10 + 650\cdot 45 + 5700\cdot 45}{100} \\ &= -2285\end{aligned} \tag{2.38}$$

When the response function is not linear, an appropriate imbalance measure is needed to be set. The following theorem for the case with the second moment terms in the response function has been stated on p.34 of Sauppe (2015) without proof and the detailed proof is provided here.

Theorem 6. *Let $\mathcal{P} = \{1, 2, \cdots, K\}$ and $\binom{\mathcal{P}}{2}$ denote a set of all the possible covariate index pairs. Then the bias $\mathcal{B}(T, C')$ equals zero for all response functions of the second-order form $h^0(\mathbf{X}_u) = \sum_{k\in\mathcal{P}}\beta_k X_{u,k} + \sum_{k\in\mathcal{P}}\gamma_k(X_{u,k})^2 + \sum_{(k_1,k_2)\in\binom{\mathcal{P}}{2}}\gamma_{k_1,k_2}X_{u,k_1}X_{u,k_2} +$*

α if and only if $\mathscr{I}_{DOM2} = 0$ where $\mathscr{I}_{DOM2}$ is the imbalance measure given in (2.11)*:*

$$\mathscr{I}_{DOM2}(T, C') = \sum_{k\in\mathcal{P}} \left| \frac{1}{|T|} \sum_{t\in T} X_{t,k} - \frac{1}{|C'|} \sum_{c\in C'} X_{c,k} \right| + \sum_{k\in\mathcal{P}} \left| \frac{1}{|T|} \sum_{t\in T} X_{t,k}^2 - \frac{1}{|C'|} \sum_{c\in C'} X_{c,k}^2 \right| + \sum_{(k_1,k_2)\in\binom{\mathcal{P}}{2}} \left| \frac{1}{|T|} \sum_{t\in T} X_{t,k_1} X_{t,k_2} - \frac{1}{|C'|} \sum_{c\in C'} X_{c,k_1} X_{c,k_2} \right|. \tag{2.39}$$

Proof. The sufficiency part can be proved as follows. Note that $\mathscr{I}_{\text{DOM2}} = 0$ is equivalent to the following three conditions in (2.40) to (2.42):

$$\frac{1}{|T|} \sum_{t\in T} X_{t,k} - \frac{1}{|C'|} \sum_{c\in C'} X_{c,k} = 0 \quad \forall k \in \mathcal{P}, \tag{2.40}$$

$$\frac{1}{|T|} \sum_{t\in T} X_{t,k}^2 - \frac{1}{|C'|} \sum_{c\in C'} X_{c,k}^2 = 0 \quad \forall k \in \mathcal{P}, \tag{2.41}$$

and

$$\frac{1}{|T|} \sum_{t\in T} X_{t,k_1} X_{t,k_2} - \frac{1}{|C'|} \sum_{c\in C'} X_{c,k_1} X_{c,k_2} = 0 \quad \forall (k_1, k_2) \in \binom{\mathcal{P}}{2}. \tag{2.42}$$

It implies that

$$\begin{aligned}
\mathcal{B}(T, C') =& \frac{1}{|T|} \sum_{t\in T} h^0(\mathbf{X}_t) - \frac{1}{|C'|} \sum_{c\in C'} h^0(\mathbf{X}_c) \\
=& \frac{1}{|T|} \sum_{t\in T} \left(\sum_{k\in\mathcal{P}} \beta_k X_{t,k} + \sum_{k\in\mathcal{P}} \gamma_k (X_{t,k})^2 + \sum_{(k_1,k_2)\in\binom{\mathcal{P}}{2}} \gamma_{k_1,k_2} X_{t,k_1} X_{t,k_2} + \alpha \right) \\
& - \frac{1}{|C'|} \sum_{c\in C'} \left(\sum_{k\in\mathcal{P}} \beta_k X_{c,k} + \sum_{k\in\mathcal{P}} \gamma_k (X_{c,k})^2 + \sum_{(k_1,k_2)\in\binom{\mathcal{P}}{2}} \gamma_{k_1,k_2} X_{c,k_1} X_{c,k_2} + \alpha \right) \\
=& \sum_{k\in\mathcal{P}} \beta_k \left(\frac{1}{|T|} \sum_{t\in T} X_{t,k} - \frac{1}{|C'|} \sum_{c\in C} X_{c,k} \right) + \sum_{k\in\mathcal{P}} \gamma_k \left(\frac{1}{|T|} \sum_{t\in T} X_{t,k}^2 - \frac{1}{|C'|} \sum_{c\in C'} X_{c,k}^2 \right) \\
& + \sum_{(k_1,k_2)\in\binom{\mathcal{P}}{2}} \gamma_{k_1,k_2} \left(\frac{1}{|T|} \sum_{t\in T} X_{t,k_1} X_{t,k_2} - \frac{1}{|C'|} \sum_{c\in C'} X_{c,k_1} X_{c,k_2} \right) = 0
\end{aligned}$$

To prove necessity, suppose to the contrary that at least one of the three conditions (2.40), (2.41), and (2.42) is not satisfied. Then

$$\frac{1}{|T|}\sum_{t\in T} X_{u,k^*} - \frac{1}{|C'|}\sum_{c\in C'} X_{c,k^*} \neq 0 \quad \text{for some } k^* \in \mathcal{P}, \tag{2.43}$$

$$\frac{1}{|T|}\sum_{t\in T} X^2_{t,k^{**}} - \frac{1}{|C'|}\sum_{c\in C'} X^2_{c,k^{**}} \neq 0 \quad \text{for some } k^{**} \in \mathcal{P}, \tag{2.44}$$

or

$$\frac{1}{|T|}\sum_{t\in T} X_{t,\widetilde{k_1}} X_{t,\widetilde{k_2}} - \frac{1}{|C'|}\sum_{c\in C'} X_{c,\widetilde{k_1}} X_{c,\widetilde{k_2}} \neq 0 \quad \text{for some } (\widetilde{k_1},\widetilde{k_2}) \in \binom{\mathcal{P}}{2}. \tag{2.45}$$

As in the proof for Theorem 5, take one of the cases among (2.43), (2.44), and (2.45) that do not satisfy (2.40), (2.41), and (2.42). Let $\beta_{k^*} = 1$ and all the other coefficients be 0 in case of (2.43) is selected. Moreover, let $\gamma_{k^{**}} = 1$ and all the other coefficients be 0 in case of (2.44), and let $\gamma_{\widetilde{k_1},\widetilde{k_2}} = 1$ and all the other coefficients be 0 in case of (2.45).

Then it becomes one of the following cases.

$$\mathcal{B}(T,C') = \frac{1}{|T|}\sum_{t\in T} X_{t,k^*} - \frac{1}{|C'|}\sum_{c\in C'} X_{c,k^*} \neq 0, \tag{2.46}$$

$$\mathcal{B}(T,C') = \frac{1}{|T|}\sum_{t\in T} X^2_{t,k^{**}} - \frac{1}{|C'|}\sum_{c\in C'} X^2_{c,k^{**}} \neq 0, \tag{2.47}$$

or

$$\mathcal{B}(T,C') = \frac{1}{|T|}\sum_{t\in T} X_{t,\widetilde{k_1}} X_{t,\widetilde{k_2}} - \frac{1}{|C'|}\sum_{c\in C'} X_{c,\widetilde{k_1}} X_{c,\widetilde{k_2}} \neq 0. \tag{2.48}$$

Hence, the bias is not equal to zero for such choice of coefficients. □

Note that for fixed α and β, the condition $\mathscr{I}_{\text{DOM2}} = 0$ is sufficient but not necessary for $\mathcal{B}(T,C') = 0$. The following is an example where $\mathcal{B}(T,C') = 0$ while $\mathscr{I}_{\text{DOM2}} \neq 0$.

Suppose that there are two covariates, $\mathcal{P} = \{1,2\}$, and the response function is given by $h^0(\mathbf{X}_u) = X_{u,1} + X_{u,2} + X^2_{u,1} + X^2_{u,2} + X_{u,1}X_{u,2}$ for both the treatment group and the control group. That is, $\alpha = 0$ and

$$\beta_k = \gamma_k = \gamma_{k_1,k_2} = 1 \quad \forall k \in \mathcal{P}, (k_1,k_2) \in \binom{\mathcal{P}}{2} \tag{2.49}$$

in the second-order response function.

Let the covariate values for units in $T = \{t_1, t_2, \cdots, t_{100}\}$ and $C' = \{c_1, c_2, \cdots, c_{100}\}$

be given by

$$\mathbf{X}_{t_i} = \begin{cases} (1, -\frac{5}{2}) & \text{for } 1 \le i \le 45 \\ (0,0) & \text{for } 46 \le i \le 95 \\ (\frac{-1+\sqrt{57}}{4}, \frac{-1-\sqrt{57}}{4}) & \text{for } 96 \le i \le 100 \end{cases} \tag{2.50}$$

and

$$\mathbf{X}_{c_i} = \begin{cases} (1, -\frac{5}{2}) & \text{for } 1 \le i \le 5 \\ (0,0) & \text{for } 6 \le i \le 55 \\ (\frac{-1+\sqrt{57}}{4}, \frac{-1-\sqrt{57}}{4}) & \text{for } 56 \le i \le 100 \end{cases} \tag{2.51}$$

for $i = 1, 2, \cdots, 100$. Then

$$\frac{1}{|T|}\sum_{t\in T} X_{t,1} - \frac{1}{|C'|}\sum_{c\in C'} X_{c,1} = \frac{5-\sqrt{57}}{10} \neq 0, \tag{2.52}$$

$$\frac{1}{|T|}\sum_{t\in T} X_{t,2} - \frac{1}{|C'|}\sum_{c\in C'} X_{c,2} = \frac{-9+\sqrt{57}}{10} \neq 0, \tag{2.53}$$

$$\frac{1}{|T|}\sum_{t\in T} X_{t,1}^2 - \frac{1}{|C'|}\sum_{c\in C'} X_{c,1}^2 = \frac{-21+\sqrt{57}}{20} \neq 0, \tag{2.54}$$

$$\frac{1}{|T|}\sum_{t\in T} X_{t,2}^2 - \frac{1}{|C'|}\sum_{c\in C} X_{c,2}^2 = \frac{21-\sqrt{57}}{20} \neq 0, \tag{2.55}$$

$$\frac{1}{|T|}\sum_{t\in T} X_{t,1}X_{t,2} - \frac{1}{|C'|}\sum_{c\in C'} X_{c,1}X_{c,2} = \frac{2}{5} \neq 0. \tag{2.56}$$

Hence, $\mathscr{I}_{\text{DOM2}} \neq 0$ but

$$\begin{aligned} \mathcal{B}(T, C') = &\sum_{k=1}^{2}\left(\frac{1}{|T|}\sum_{t\in T} X_{t,k} - \frac{1}{|C'|}\sum_{c\in C'} X_{c,k}\right) + \sum_{k=1}^{2}\left(\frac{1}{|T|}\sum_{t\in T} X_{t,k}^2 - \frac{1}{|C'|}\sum_{c\in C'} X_{c,k}^2\right) \\ &+ \left(\frac{1}{|T|}\sum_{t\in T} X_{t,1}X_{t,2} - \frac{1}{|C'|}\sum_{c\in C'} X_{c,1}X_{c,2}\right) \\ =&\frac{5-\sqrt{57}}{10} + \frac{-9+\sqrt{57}}{10} + \frac{-21+\sqrt{57}}{20} + \frac{21-\sqrt{57}}{20} + \frac{2}{5} = 0. \end{aligned} \tag{2.57}$$

Use of Imbalance Measure Balancing Only the Marginal Distribution When Covariates Are Correlated

Consider the case that the two groups have an identical marginal distribution but not identical joint distributions. Suppose that $K = 2$ and the covariate values for the treatment units are given by

$$\mathbf{X}_{t_i} = \begin{cases} (i, 101 - i) & \text{for } 1 \le i \le 100 \\ (i - 100, i - 100) & \text{for } 101 \le i \le 200 \\ (i - 200, i - 200) & \text{for } 201 \le i \le 300 \end{cases} \tag{2.58}$$

while those of the control units are given by

$$\mathbf{X}_{c_i} = \begin{cases} (i, 101 - i) & \text{for } 1 \le i \le 100 \\ (i - 100, 201 - i) & \text{for } 101 \le i \le 200 \\ (i - 200, i - 200) & \text{for } 201 \le i \le 300 \end{cases} \tag{2.59}$$

for $T = \{t_1, t_2, \cdots, t_{300}\}$ and $C' = \{c_1, c_2, \cdots, c_{300}\}$

Then the imbalance measure derived from the Kolmogorov-Smirnov Test Statistic in (2.60) equals zero,

$$\mathscr{I}_{\text{KS}}(T, C') = \sum_{k=1}^{K} \max_{x \in \mathscr{X}_k(T \cup C')} \left| \widehat{F}_k(T, x) - \widehat{F}_k(C', x) \right| \tag{2.60}$$

where $\mathscr{X}_k(S)$, $\widehat{F}_k(T, x)$, and $\widehat{F}_k(C', x)$ are defined as in (2.12).

On the other hand, if the response function is not separable so that it contains an interaction term, then having the same marginal distribution is insufficient to ensure zero bias. For example, suppose that the response function is given by $h^0(\mathbf{X}_u) = h^1(\mathbf{X}_u) = X_{u,1}X_{u,2}$. It implies

$$\begin{aligned} \mathscr{B}(T, C') &= \frac{1}{|T|} \sum_{t \in T} h^0(\mathbf{X}_t) - \frac{1}{|C'|} \sum_{c \in C'} h^0(\mathbf{X}_c) \\ &= \frac{\sum_{i=1}^{100} \left(i(101 - i) + 2i^2 \right)}{300} - \frac{\sum_{i=1}^{100} \left(2i(101 - i) + i^2 \right)}{300} = \frac{1111}{2}. \end{aligned} \tag{2.61}$$

2.4.2 Example with An Improper Bin Size

Insufficient Granularity in Coarsened Binning Method

Consider the following histogram binning imbalance measure (Nikolaev et al., 2013):

$$\mathscr{I}_{\text{hist}}(T, C') = \sum_{k=1}^{K} \sum_{b \in M_k} \left| \frac{|T \cap B_{k,b}|}{|T|} - \frac{|C' \cap B_{k,b}|}{|C'|} \right| \tag{2.62}$$

where $B_{k,b}$ denotes the set of units whose value for the k-th covariate is in the b-th bin, for $b \in M_k$, the set of indices for covariate k's bins.

Let K, the number of covariates, be 1 for the following example. Suppose further that there are 300 treatment units $\{t_1, \cdots, t_{300}\}$ in the treatment group and 300 control units $\{c_1, \cdots, c_{300}\}$ in the control group where the treatment units have covariate values given by

$$X_{t_i,1} = \begin{cases} 4i + 1 & \text{for } 1 \leq i \leq 100 \\ 4(i - 100) + 1 & \text{for } 101 \leq i \leq 200 \\ 4(i - 200) + 3 & \text{for } 201 \leq i \leq 300 \end{cases} \tag{2.63}$$

and the control units have their covariate values given by

$$X_{c_i,1} = \begin{cases} 4i + 3 & \text{for } 1 \leq i \leq 100 \\ 4(i - 100) + 3 & \text{for } 101 \leq i \leq 200 \\ 4(i - 200) + 1 & \text{for } 201 \leq i \leq 300. \end{cases} \tag{2.64}$$

Form the coarsened bins so that the b-th bin of the single covariate represents an interval $[4b, 4b + 4]$ for $b \in M_k = \{1, 2, \cdots, 100\}$. For the single covariate, there are 100 bins in total and each bin contains exactly three treatment units and three control units.

Then

$$\mathscr{I}_{\text{hist}}(T, C') = \sum_{k=1}^{1} \sum_{b=1}^{100} \left| \frac{|T \cap B_{k,b}|}{|T|} - \frac{|C' \cap B_{k,b}|}{|C'|} \right| = 0 \tag{2.65}$$

since $|T \cap B_{1,b}| = |C' \cap B_{1,b}| = 3$ for each b and $|T| = |C'| = 300$. In other words, with this particular coarsened binning for the imbalance measure, an objective value of zero is obtained. On the other hand, the bias need not be zero.

Suppose that the response function is defined as follows:

$$h^0(\mathbf{X}_u) = 1.5X_{u,1}. \tag{2.66}$$

Then the bias is given by

$$\begin{aligned}\mathcal{B}(T, C') &= \frac{1}{300}\sum_{i=1}^{100}[1.5 \cdot (4i+1) \cdot 2 + 1.5 \cdot (4i+3)] \\ &\quad - \frac{1}{300}\sum_{i=1}^{100}[1.5 \cdot (4i+3) \cdot 2 + 1.5 \cdot (4i+1)] \\ &= \frac{1}{300}\sum_{t=1}^{100}(-3) = -1.\end{aligned} \tag{2.67}$$

Note that this case with a single covariate corresponds to a coarsened matching method – matching the units that lie in the same bins.

2.5 Non-zero Optimum under Correct Imbalance Measure

The optimal value from BOSS may not be zero if there is either insufficient data and/or optimization issues, regardless of whether the chosen imbalance measure is correct for the problem. These cases in general lead to non-zero bias.

2.5.1 Insufficient Data

This section provides an example of insufficient data. Let $K = 1$. Suppose that the covariate values of treatment units $\{t_1, t_2, \cdots, t_{300}\}$ are distributed as

$$X_{t_i,1} = \begin{cases} i & \text{for } 1 \le i \le 200 \\ i - 200 + 50 & \text{for } 201 \le i \le 300 \end{cases}. \tag{2.68}$$

Suppose that the pool of control units $C = \{c_1, \cdots, c_{400}\}$ have covariates given by

$$X_{c_i,1} = \begin{cases} i & \text{for } 1 \leq i \leq 100 \\ i - 100 & \text{for } 101 \leq i \leq 200 \\ i - 200 + 50 & \text{for } 201 \leq i \leq 300 \\ i - 300 + 50 & \text{for } 301 \leq i \leq 400 \end{cases}. \tag{2.69}$$

Suppose that $h^1(\mathbf{X}_u) = h^0(\mathbf{X}_u) = \mathbf{X}_u$ and thus $\mathscr{I}_{\text{DOM}}$ is a correct imbalance measure. However, an imbalance measure of zero cannot be achieved in this case. When restricting to the case that the size of the selected control group is the same as that of the treatment group (i.e., $|C'| = 300$), the minimum value in the imbalance measure is obtained by choosing

$$C' = \{c_{51}, c_{52}, \cdots, c_{100}, c_{151}, c_{152}, \cdots c_{400}\}. \tag{2.70}$$

(Note that there are many alternate optima in this problem.) In this case, the bias is equal to the following:

$$\mathscr{B}(T, C') = \frac{1}{|T|} \sum_{t \in T} X_{t,k} - \frac{1}{|C'|} \sum_{c \in C'} X_{c,k} = \frac{25}{3}. \tag{2.71}$$

This example corresponds to the case of "non-zero objective – non-zero bias" in Table 2.1. In this case, the bias can be corrected by adding appropriate control units to the data so that there exists sufficient overlap between the treatment group and control pool and thus balanced groups can be chosen for estimation.

How Much Overlap Is Needed to Be Sufficient?

In this section, how much overlap is needed to achieve a zero objective value in the imbalance minimization problem of BOSS is investigated. To achieve a zero objective value, there does not need to be complete overlap between the two intervals that the treatment and control groups are laid on. For example, an objective value of zero for the DOM imbalance measure can be achieved in the following case where treatment units are in $[1, 100]$ and units in the control pool are in $[2, 101]$.

Let $K = 1$. Suppose that the covariate values of treatment units $t \in T$ are

distributed as

$$X_{t_i,1} = i \quad \forall t_i \in T = \{t_1, \cdots, t_{100}\} \tag{2.72}$$

while the control units $c \in C = \{c_1, \cdots, c_{200}\}$ have

$$\begin{aligned} X_{c_i,1} &= i + 1 \text{ for } c_i \in \{c_1, \cdots, c_{100}\} \\ X_{c_i,1} &= i - 99 \text{ for } c_i \in \{c_{101}, \cdots, c_{200}\} \end{aligned} \tag{2.73}$$

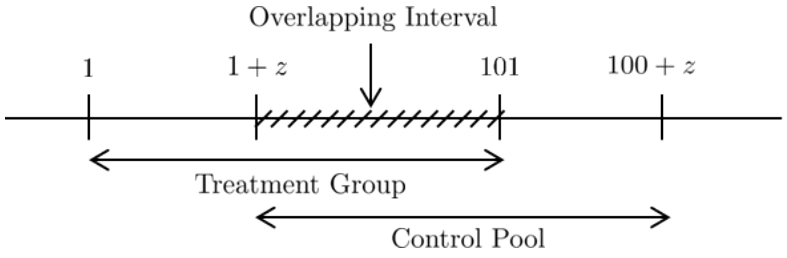
and the response function is given as above: $h^0(\mathbf{X}_u) = h^1(\mathbf{X}_u) = \mathbf{X}_u$. Then the value of 0 in the imbalance measure can be achieved by choosing $C' = \{c_1, \cdots, c_{98}, c_{101}, c_{198}\}$ so that $\frac{1}{|T|}\sum_{t\in T} X_{t,k} = \frac{1}{|C'|}\sum_{c\in C'} X_{c,k} = 50.5$ and thus $\frac{1}{|T|}\sum_{t\in T} X_{t,k} - \frac{1}{|C'|}\sum_{c\in C'} X_{c,k} = 0$. There are many alternative optima to this problem. For example, all of the following sets for C' satisfy the condition for zero optimum:

$$\{c_1, \cdots c_{98}, c_{102}, c_{197}\}, \{c_1, \cdots c_{98}, c_{103}, c_{196}\}, \cdots, \{c_1, \cdots c_{98}, c_{149}, c_{150}\},$$
$$\{c_2, \cdots, c_{97}, c_{148}, c_{149}, c_{150}, c_{151}\}, \cdots, \{c_{25}, \cdots c_{74}, c_{125}, \cdots c_{174}\}.$$

Given the information on the response function, how much they should be overlapped to achieve zero imbalance measure can also be computed. As an illustration, consider the same setting as above except that the control pool's covariate value distribution is given by

$$\begin{aligned} X_{c_i,1} &= i + z \text{ for } c_i \in \{c_1, \cdots, c_{100}\} \\ X_{c_i,1} &= i + z - 100 \text{ for } c_i \in \{c_{101}, \cdots, c_{200}\}. \end{aligned} \tag{2.74}$$

Figure 2.3: Range of Covariate Values



Here the control pool size is twice as large as the treatment group. If the control group is chosen without replacement and with a condition that $|C'| = |T|$ as done so far, then any integer $z \in [-25, 25]$ would be sufficient to have optimal value 0 when minimizing the DOM imbalance measure. Any z values outside of this

interval (i.e., $z > 25$ or $z < -25$) will result in insufficient overlap. Extreme cases with an insufficient overlap would be the cases without any overlapping interval between the range of the treatment group covariate values and the range of the control pool covariate values (here in this example, that is when $z > 100$ or $z < -99$).

However, zero objective value may not be achieved even though the z is in the interval $[-25, 25]$. If $z = 23.002$, then the sum of control units' covariate values, $\sum_{c \in C'} X_{c,k}$, is fractional for any selection of a control group C' from the pool C while the the sum of the treatment units' covariate values, $\sum_{t \in T} X_{t,k}$, is integral. Accordingly, with such a value for z, a zero bias cannot be achieved since any fractional number cannot be equal to an integer while $|T|$ and $|C'|$ are the same.

This suggests that, although it is not necessary to have a complete overlap between the treatment and control values, merely having a large overlapping region between covariate value range of the treatment group and that of the control pool may not be very informative about whether an objective value of zero is achievable or not. Sufficient overlap between the treatment pool and the control pool should be defined based on homogeneity of the two sets.

Definition 3. Given an imbalance measure $\mathscr{I}$, a treatment group T is has *more overlap* with C than with $\widetilde{C}$ if

$$\min_{C' \subseteq C} \mathscr{I}(T, C') < \min_{\widetilde{C}' \subseteq \widetilde{C}} \mathscr{I}(T, \widetilde{C}') \tag{2.75}$$

In addition, there is *enough overlap* between the treatment pool T and the control pool C if a control group $C' \subseteq C$ such that $\mathscr{I}(T, C') = 0$ can be found.

Note that a stricter overlap condition is required when using an imbalance measure with a higher ranking in the balance hierarchy. Suppose that there are imbalance measures $\mathscr{I}_1$ and $\mathscr{I}_2$ where $\mathscr{I}_1$ has a higher rank than $\mathscr{I}_2$. Then from Definition 1, $\mathscr{I}_1 = 0$ implies that $\mathscr{I}_2 = 0$. As such, the amount of overlap that is necessary to ensure that $\mathscr{I}_1 = 0$ is at least as large as (and almost certainly larger than) the amount of overlap necessary to ensure that $\mathscr{I}_2 = 0$.

From Theorem 1 and Definition 3, the best data to add to the control pool is data that provides enough overlap for $\mathscr{I}_{\text{ecdf:}\mathbf{D}}$, so that it is possible to find a $C' \subseteq C$ that satisfies $\mathscr{I}_{\text{ecdf:}\mathbf{D}}(T, C') = 0$. Such a control group would then provide a treatment effect estimate with zero bias regardless of the functional form of the response function. In order for the optimal value of $\mathscr{I}_{\text{ecdf:}\mathbf{D}}$ to equal zero, it is necessary to

have a unit c to C that is identical to a unit t for each $t \in T$.

2.5.2 Suboptimality from Technical Issues in Optimization

In many simulations of the BOSS method using a large number of covariates and large treatment and control pools, an optimal value of zero could not be obtained due to time constraints. See Sauppe et al. (2014) for examples. This issue could potentially be resolved by adopting a more effective algorithm to solve the problem, by assigning more time for computation, or by using a more powerful computer. However, this is not always possible because the general BOSS problems are computationally intractable. Sauppe et al. (2014) shows that BOSS is NP-hard except in some special cases.

2.6 Conditions for Zero Bias in BOSS

Putting all the discussions from earlier sections together, the earlier discussions can be summarized with the following theorem.

Theorem 7. *Suppose that the BOSS problem satisfies all of the following three conditions:*

1. *Given the functional form of the response functions, a correct imbalance measure that corresponds to it is used.*
2. *There are sufficient data to ensure that an optimal value of* 0 *in the correct imbalance measure can be obtained.*
3. *There is an effective algorithm as well as a fast enough computer that can solve the problem to optimality without hitting the time constraints.*

Then there is no bias in the estimate derived from the solution identified by BOSS.

Proof. Condition 2 implies that there are enough data to achieve an optimal value of zero in theory. And this zero imbalance measure is achievable in practice since there is not any optimization issues from Condition 3. Lastly, having an optimal value equal to zero implies that there is zero bias since a correct imbalance measure is used from Condition 1. □

In other words, the three cases 1, 2, and 3 listed below exhaust all the possible cases that may lead to bias. When these bias issues are resolved, then there is not any bias when estimating the average treatment effect for the treated.

1. Incorrect imbalance measure
2. Insufficient data (data without enough overlap)
3. Suboptimality

While finding an exact functional form is a huge challenge, case 1 is easy to resolve since it is known which imbalance measure will be correct from Definitions 1 and 2, and Theorems 5 and 6 after identifying the functional form of the response function. However, when there are limited observational data, it is very likely to confront case 2. Additionally case 3, suboptimality, may arise when there are a large number of covariates to be balanced. When the number of covariates increases, it is not only difficult to have an optimal value of zero but also it is difficult to actually find an optimal solution given the computational intractability of the problem.

2.7 Concluding Remarks

The three conditions guaranteeing zero bias, which is the exact opposite of the three cases for generating bias as discussed in the previous section, can be elaborated as follows.

First, a correct imbalance measure is needed. Researchers should understand the concept of the Balance Hierarchy and use an imbalance measure that is correct. As an example, a linear response function requires a DOM imbalance measure, $\mathscr{I}_{\text{DOM}}$, as in (2.9). As another example, when there is a correlation term in the response function, a corresponding $\mathscr{I}_{\text{Corr:}\mathcal{K}}$ term like (2.8) in the imbalance measure for the set of covariate indices, $\mathcal{K}$, is needed.

The imbalance measures that are highest in the balance hierarchy for separable response functions, $\mathscr{I}_{\text{KS}}$ in (2.12) and $\mathscr{I}_{\text{CvM}}$ in (2.21), and those for non-separable case, $\mathscr{I}_{\text{ecdf:}\mathbf{D}}$ in (2.13) and $\mathscr{I}_{\text{CvM:}\mathbf{D}}$ in (2.22) were also discussed. Note that, because the general non-separable case includes the separable case, $\mathscr{I}_{\text{ecdf:}\mathbf{D}}$ and $\mathscr{I}_{\text{CvM:}\mathbf{D}}$ are more highly ranked than $\mathscr{I}_{\text{KS}}$ and $\mathscr{I}_{\text{CvM}}$. Further recall that, although the imbalance measures $\mathscr{I}_{\text{ecdf:}\mathbf{D}}$ and $\mathscr{I}_{\text{CvM:}\mathbf{D}}$ are ranked the highest and are correct for

any functional form of the response function, they are not often used in practice as it is difficult to obtain sufficient data to ensure that an objective value of zero can be achieved.

Second, there should be sufficient data. The control pool data that is available for solving the optimization problem should have enough overlap with the treatment group data.

Lastly, there should not be any issues in optimization. That is, an effective algorithm is needed to solve the problem and a computer that is able to solve the problem to optimality within a reasonable time limit.

Preliminary work on the balance hierarchy comparing the various imbalance measures and identifying which imbalance measure should be used for various cases had been done by Sauppe and Jacobson (2017). The hierarchy presented in this chapter provides an alternate perspective. The hierarchy defined in this chapter aggregates the terms in the control response function and uses it to rank the various imbalance measures while Sauppe and Jacobson (2017) focuses on the individual terms and the relationship between them.

In addition, imbalance measures like $\mathscr{I}_{\text{CvM}}$ and $\mathscr{I}_{\text{CvM:}\mathbf{D}}$ are newly adopted using the concept of the Cramer-von Mises test statistic. Furthermore, the discussion in this chapter provides additional value since it exhausts all possible cases for bias and provides many examples.

In this chapter, when bias can occur in the BOSS method was considered. As noted, it may be difficult to eliminate all bias. Even though zero bias may not be achieved, it is always important to select a control group whose distribution is as similar to that of the treatment group because closely balanced groups offer many benefits.

CHAPTER 3

TREATMENT EFFECT DECOMPOSITION AND BOOTSTRAP HYPOTHESIS TESTING IN OBSERVATIONAL STUDIES

3.1 Introduction

Determining whether a treatment causes a certain effect is essential in many fields. In randomized experiments, researchers can easily estimate the treatment effect. Units under study are randomly assigned to either a treatment pool or a control pool in an experimental setting. This ensures that there will be no systematic difference in their characteristics on average. However, in many cases where randomized experiments cannot be implemented, researchers have to rely on observational data. For example, when examining whether radiation exposure causes cancer, researchers cannot conduct a randomized experiment because assigning individuals to be exposed to (possibly harmful) radiation would be unethical. Similarly, verifying a causal relation between smoking and cancer must rely on observational data as one cannot force some randomly selected individuals to smoke.

Matching is a common method for causal inference with observational data. Each unit has a set of attributes called covariates. Recall that a treatment pool (or a treatment group) is a set of units who were treated and a control pool is the set of units who were not treated. The objective of matching methods is to find exact or close matches from the control pool having similar covariate values for each unit in the treatment pool. The chosen units from the control pool constitutes a control group. Covariate balance, a function of the treated units' and control units' covariate vectors, measures the similarity between the covariate distributions of the treatment group and the control group and thus the effectiveness of the matching. However, matching methods have limitations in that they aim to achieve covariate balance indirectly. To overcome this drawback, the Balance Optimization Subset Selection (BOSS) framework was proposed by Nikolaev et al. (2013).

This chapter uses BOSS to estimate the treatment effects with a new perspective. First, the treatment group T and control pool C are partitioned into several

subsets depending on specific covariate values of interest. Partitioning the set based on specific covariate values is not a new idea. *Subclassification* is a term that has been used in the literature to subdivide units into smaller groups based on certain covariates or propensity score (Cochran, 1968; Rosenbaum and Rubin, 1985). The propensity score subclassificiation method is also known as *Stratification* as it forms strata where units in each stratum have approximately the same propensity score. The method introduced in this chapter coincides with subclassification in the case of experimental data but it differs in the case of observational data in that within each subclass BOSS method is used to find a subset of control units that is most balanced with the treatment group.

After partitioning these sets, the size of a treatment effect by each of the components in the partition of the treatment group is computed. Treatment effects that differ across diverse sub-populations are referred as *Heterogeneous Treatment Effects* (Xie et al., 2012; Imai et al., 2013). Since "effects vary across individuals, between groups, over time, and across space" as noted in Elwert and Winship (2010), it is important to understand the treatment effect heterogeneity that arises from different subsets. The purpose of this chapter is to identify heterogeneous treatment effects in a certain partition of the entire set of units. How the treatment effects from the subsets are related to the treatment effect of the entire set will be investigated and their statistical significance will be checked.

Consider investigating whether a power toothbrush is effective in removing plaque (Grender et al., 2013). Possible covariate values for the study include gender, age, and dietary pattern. The set of units for the study may be divided into two groups (by gender) or several groups (by the age of the participants). Would the treatment effect of the power toothbrush differ by gender? Would it be different for different age groups? Would the identified treatment effects be significant?

To answer such questions, the treatment effect of the entire treatment and control units is decomposed first. In particular, Balance Optimization Subset Selection (BOSS) is used to compute the treatment responses. Denote a treatment group by T and a control pool by C. In BOSS, a control group $C' \subset C$, as set of units that minimizes covariate imbalance is chosen. Then the treatment effect can be estimated with a difference in average treated response value of the treated units in T and the average untreated response value of the control units in C. (See Nikolaev et al. (2013) and Sauppe et al. (2014) for a detailed explanation of BOSS.) As a next step, how a bootstrap hypothesis test can be conducted on the results obtained by BOSS is introduced to determine statistical significance of the effects. These

methods are applied to the LaLonde (1986) dataset for a labor program evaluation. Lastly, the dataset is re-analyzed by creating sub-samples composed of a subset of units from a treatment group and a corresponding control pool.

The chapter is organized as follows. Section 3.2 reviews the BOSS framework. Section 3.3 introduces decomposition of the treatment effect. Section 3.4 discusses how to apply the two-sample bootstrap hypothesis test to check statistical significance of the heterogeneous treatment effects. Section 3.5 applies the theory to the LaLonde data. Section 3.6 provides concluding remarks.

3.2 Balance Optimization Subset Selection (BOSS)

In the following sections, the methods of treatment effect decomposition and bootstrap hypothesis testing which can be applied on the estimates obtained by BOSS are introduced.

Matching methods find pairs for each treated unit with a control unit and try to achieve balance in the distribution of the covariates indirectly. BOSS finds the best control group by directly minimizing a covariate imbalance measure chosen by researchers. That is, BOSS selects a control group C' from the control pool C that has similar covariate values in aggregate (i.e., that maximize balance) with the treatment group T through BOSS.

Recall that Y_t^1 and Y_t^0 denote the treated and untreated responses of a treated unit $t \in T$ respectively and the SATT is defined as (1.3). Since Y_t^0 is unavailable, the SATT is estimated through an estimator in (1.4) using a control group C' obtained by BOSS.

BOSS solves a minimization problem that has an imbalance measure as an objective function. There are many possible choices for imbalance measures. If some assumptions are made on the functional form of the response functions, then a full joint distribution balance is not needed to construct an unbiased treatment effect estimates. In this chapter, a particular form of the imbalance measure, $\mathscr{I}_{\text{DOM}}$, defined in (2.9) using difference of means will be used. Note that the decomposition and bootstrap hypothesis testing methods which will be explained in the following sections can be applied to BOSS estimators obtained with any imbalance measure.

3.3 Decomposition of the Treatment Effect

This section considers decomposing the treatment effect as a weighted average of treatment effects from sets in a partition of T and corresponding subsets of C. First, consider a base case that the treated group T is partitioned into two sets T_1 and T_2. Define the sample average treatment effect for the treated in T_1 and that in T_2 by $\tau^1_{T_1}$ in (3.1) and $\tau^1_{T_2}$ in (3.2) respectively:

$$\tau^1_{T_1} = \frac{1}{|T_1|} \sum_{t \in T_1} (Y^1_t - Y^0_t) \tag{3.1}$$

and

$$\tau^1_{T_2} = \frac{1}{|T_2|} \sum_{t \in T_2} (Y^1_t - Y^0_t). \tag{3.2}$$

Then the sample average treatment effect for the treated in the entire treatment pool, τ^1_T, can be decomposed as follows.

Theorem 8. *Suppose that T denotes the set of treated units. Then*

$$\tau^1_T = \frac{|T_1|}{|T|} \tau^1_{T_1} + \frac{|T_2|}{|T|} \tau^1_{T_2} \tag{3.3}$$

where $\{T_1, T_2\}$ is a partition of T.

Proof. Since $T_1 \cup T_2 = T$ and $T_1 \cap T_2 = \varnothing$, then

$$\begin{aligned}
\tau^1_T &= \frac{1}{|T|} \sum_{t \in T} (Y^1_t - Y^0_t) \\
&= \frac{1}{|T|} \left\{ \sum_{t \in T_1} (Y^1_t - Y^0_t) + \sum_{t \in T_2} (Y^1_t - Y^0_t) \right\} \\
&= \frac{|T_1|}{|T|} \tau^1_{T_1} + \frac{|T_2|}{|T|} \tau^1_{T_2}
\end{aligned}$$

□

Corollary. *Let $\mathscr{P} = \{T_1, T_2, \cdots, T_P\}$ be a partition of the set of treated units, T. Then the following relationship holds:*

$$\tau^1_T = \sum_{j=1}^{P} \frac{|T_j|}{|T|} \tau^1_{T_j} \tag{3.4}$$

Theorem 8 states the treatment effect of an entire sample T can be written as a weighted average of the treatment effects of its partition T_1 and T_2 where the weights are proportional to the cardinality of those sets in the partition. Note that, while discussion in this chapter will focus on two subset case, it can be generalized into a case with multiple subsets in the partition.

Suppose that the control pool C is also partitioned into two sets C_1 and C_2 using the same criteria with covariates used for partitioning the treated set T. Recall that the responses of each unit u are given by $Y_u^z = h^z(\mathbf{X}_u) + \epsilon_u^z$ for $z \in \{0, 1\}$ with response function $h^z(\cdot)$ and error term ϵ_u^z from Strong Ignorability assumption. Define $\mathcal{B}(T, C')$ and $\mathcal{E}(T, C')$ as

$$\mathcal{B}(T, C') = \frac{1}{|T|} \sum_{t \in T} h^0(\mathbf{X}_t) - \frac{1}{|C'|} \sum_{c \in C'} h^0(\mathbf{X}_c) \tag{3.5}$$

$$\mathcal{E}(T, C') = \frac{1}{|T|} \sum_{t \in T} \epsilon_t^0 - \frac{1}{|C'|} \sum_{c \in C'} \epsilon_c^0 \tag{3.6}$$

Then the SATT, τ_T^1, can be written as

$$\tau_T^1 = \widetilde{\tau}_T^1(C') - \mathcal{B}(T, C') - \mathcal{E}(T, C'). \tag{3.7}$$

as noted in (2.1). The term $\mathcal{B}(T, C')$ is the selection bias that arises from the difference in distributions of the covariate values between the treatment group and the control group. Recall that C' is a control group obtained by

$$C' = \arg\min_{C' \subset C} \mathscr{I}(T, C'). \tag{3.8}$$

Similarly, for $i = 1, 2$, the control groups C_i' can be found from

$$C_i' = \arg\min_{C_i' \subset C_i} \mathscr{I}(T_i, C_i'). \tag{3.9}$$

Define $\mathcal{B}(T_i, C_i')$ and $\mathcal{E}(T_i, C_i')$ by replacing T and C' to T_i and C' respectively in (3.5) and (3.6).

Using these terms, how and why the estimates from a single control group C' differ from the estimates from the partitioned sets C_1' and C_2' can be analyzed. From Theorem 8:

$$\widetilde{\tau}_T^1(C') = \frac{|T_1|}{|T|} \widetilde{\tau}_{T_1}^1(C_1') + \frac{|T_2|}{|T|} \widetilde{\tau}_{T_2}^1(C_2') + \mathcal{B} + \mathcal{E} \tag{3.10}$$

where $\mathcal{B} = \mathcal{B}(T, C') - \frac{|T_1|}{|T|}\mathcal{B}(T_1, C'_1) - \frac{|T_2|}{|T|}\mathcal{B}(T_2, C'_2)$ and $\mathcal{E} = \mathcal{E}(T, C') - \frac{|T_1|}{|T|}\mathcal{E}(T_1, C'_1) - \frac{|T_2|}{|T|}\mathcal{E}(T_2, C'_2)$.

Alternatively, the the sum of bias and the error terms in (3.10), $\mathcal{B} + \mathcal{E}$, which is the ATT estimator for the entire group T minus the weighted average of the estimators from sets in its partition $\{T_1, T_2\}$, can be written as a function of T, T_1, T_2, C, C_1, C_2, Y_c^0 using the definition of the estimators: See (3.11). Let D_1 be the portion of the difference that arises from discrepancy in $|T_i|/|T|$ and $|C_i|/|C|$ for $i = 1, 2$ and D_2 be a term that arises from the difference in $C'_1 \cup C'_2$ and C'.

$$
\begin{aligned}
&\widetilde{\tau}^1_T(C') - \frac{|T_1|}{|T|}\widetilde{\tau}^1_{T_1}(C'_1) - \frac{|T_2|}{|T|}\widetilde{\tau}^1_{T_2}(C'_2) \\
&= \left(\frac{1}{|T|}\sum_{t\in T} Y_t^1 - \frac{1}{|C'|}\sum_{c\in C'} Y_c^0\right) - \frac{|T_1|}{|T|}\left(\frac{1}{|T_1|}\sum_{t\in T_1} Y_t^1 - \frac{1}{|C'_1|}\sum_{c\in C'_1} Y_c^0\right) \\
&\quad - \frac{|T_2|}{|T|}\left(\frac{1}{|T_2|}\sum_{t\in T_2} Y_t^1 - \frac{1}{|C'_2|}\sum_{c\in C'_2} Y_c^0\right) \\
&= \underbrace{\frac{1}{|C'_1|}\left(\frac{|T_1|}{|T|} - \frac{|C'_1|}{|C'|}\right)\sum_{c\in C'_1} Y_c^0 + \frac{1}{|C'_2|}\left(\frac{|T_2|}{|T|} - \frac{|C'_2|}{|C'|}\right)\sum_{c\in C'_2} Y_c^0}_{\triangleq D_1} \\
&\quad + \underbrace{\frac{1}{|C'|}\left(\sum_{c\in C'_1} Y_c^0 + \sum_{c\in C'_2} Y_c^0 - \sum_{c\in C'} Y_c^0\right)}_{\triangleq D_2}
\end{aligned}
\tag{3.11}
$$

In the next two sub-sections, it will be shown that $D_2 = 0$ and D_1 is small in case of experimental data while $D_1 = 0$ and $D_2 \neq 0$ in observational data.

3.3.1 Experimental Data

In the experimental setting, units were randomly assigned to either a treated pool (treatment group) or a control pool. Hence, it was possible to assume that the covariate distribution of the treated units and that of the control units are stochastically balanced, given that the size of the groups are large enough. Randomization provides an unbiased estimate of the treatment effect as it gives stochastic balance.

Hence, in the experimental setting, the entire control pool C can be taken as the control group C' without using BOSS to find a control group from the pool of control units having the same distribution with the treated units. Similarly, C_1 in-

stead of C_1' and C_2 instead of C_2' can be used in the analysis with the experimental data. Note that randomization ensures stochastic balance while BOSS is designed to achieve empirical balance in a particular sample.

Further recall that $\{C_1, C_2\}$ is a partition of the control pool C from definition. As a result, in (3.11),

$$D_2 = 0 \tag{3.12}$$

when using $C' = C$, $C_1' = C_1$, and $C_2' = C_2$.

Hence (3.11) can be simplified as

$$\widetilde{\tau}_T^1(C) - \frac{|T_1|}{|T|}\widetilde{\tau}_{T_1}^1(C_1) - \frac{|T_2|}{|T|}\widetilde{\tau}_{T_2}^1(C_2) = D_1 \tag{3.13}$$

Therefore, the difference between the treatment effect estimate using the entire control pool and the estimate obtained as the weighted average of estimates from the partition in (3.14)

$$\widetilde{\tau}_T^1(C) - \left(\frac{|T_1|}{|T|}\widetilde{\tau}_{T_1}^1(C_1) + \frac{|T_2|}{|T|}\widetilde{\tau}_{T_2}^1(C_2)\right) \tag{3.14}$$

can be computed from the following expression:

$$\frac{1}{|C_1|}\left(\frac{|T_1|}{|T|} - \frac{|C_1|}{|C|}\right)\sum_{c\in C_1} Y_c^0 + \frac{1}{|C_2|}\left(\frac{|T_2|}{|T|} - \frac{|C_2|}{|C|}\right)\sum_{c\in C_2} Y_c^0. \tag{3.15}$$

Note that the estimate from the entire control pool and the weighted estimate will be identical when T_1 and C_1' are equally proportioned with respect to the entire pool (i.e., $|T_1|/|T| = |C_1|/|C|$). As a result, if these ratios are different, the procedure of partitioning the treatment group, computing estimates for each part, and then averaging would give a different estimate than just computing the estimate for the entire treatment group at once. However this difference would be small since it is likely that $|T_1|/|T|$ is similar to $|C_1|/|C|$ and at the same time $|T_2|/|T|$ is similar to $|C_2|/|C|$ from randomization when the sample sizes are large enough.

Note that the approach that is taken here with experimental data is simply the subclassification method as partitioning the set based on covariate values of interest was used. However, in case of observational data, the approach is no longer coincides with the traditional subclassification method.

3.3.2 Non-experimental Data

When analyzing observational data with a partition of size two (that is, $\{T_1, T_2\}$ of T), the two control groups C_1' and C_2' are selected from C_1 and C_2 respectively and C' is selected from C by solving imbalance minimization problems. While C_1 and C_2 form a partition of the control pool C, the set $\{C_1', C_2'\}$ does not necessarily constitute a partition of C'. Hence, in general the term D_2 in (3.11) is not equal to zero unlike the case with experimental data.

Assume that the control group is found from the control pool with additional condition that the control group size is equal to the treatment group size. Since under this assumption, the following holds:

$$|T| = |C'|,\ |T_1| = |C_1'|,\ \text{and}\ |T_2| = |C_2'|. \tag{3.16}$$

It implies that

$$\frac{|T_1|}{|T|} - \frac{|C_1'|}{|C'|} = 0 \text{ and } \frac{|T_2|}{|T|} - \frac{|C_2'|}{|C'|} = 0. \tag{3.17}$$

Note that general BOSS method does not require the control group and the treatment group to be of the same size and even in the case (3.17) will hold since in BOSS the control groups are chosen in a way that $|C'| = \gamma|T|$ and $|C_i'| = \gamma|T_i|$ for $i = 1, 2$ hold for some fixed γ. Hence $D_1 = 0$ and thus (3.11) becomes

$$\widetilde{\tau}_T^1(C') - \frac{|T_1|}{|T|}\widetilde{\tau}_{T_1}^1(C_1') - \frac{|T_2|}{|T|}\widetilde{\tau}_{T_2}^1(C_2') = D_2. \tag{3.18}$$

Unlike the case with experimental data, with observational data the new method is different from subclassification in that the BOSS approach is used to find a subset of control units from each subclass that is formed by simple partitioning. As mentioned, the subgroups C_1' and C_2' in the new method are not necessarily a partition of the set C'. The subgroups are guaranteed to have the best possible balance with respective treatment groups since they are found by solving imbalance minimization problems. In this sense, the new method overcomes the drawback of the propensity score subclassification method which requires units in the same stratum to have (approximately) the same propensity score.

In both cases with experimental and non-experimental data, the decomposition method adds value in that it enables to check the composition of treatment effects that arises from the partition of the entire treated set. In the next section, whether these estimated treatment effect values are significantly greater than zero will be

investigated by applying bootstrapping.

3.4 Applying the Two-Sample Bootstrap Hypothesis Testing

In this section, the two-sample bootstrap hypothesis testing is applied to results from BOSS. See MacKinnon (2009) for a reference on Bootstrap hypothesis testing. Recall the two sets of units that are considered: a treatment group, T, and a control group, C'. As assumed in the previous section, it will be assumed that $|T| = |C'|$ in case of non-experimental data when C' is a set of units selected by solving BOSS. Let $\mathscr{A} = \{Y_t^1 \mid t \in T\}$ be a set of treated responses for units in T, and $\mathscr{B} = \{Y_c^0 \mid c \in C'\}$ be a set of untreated responses for units in C'.

Consider the means of response values in the two sets $\mathscr{A}$ and $\mathscr{B}$:

$$\mu_{\mathscr{A}}\left(= \overline{Y_T^1}\right) = \frac{1}{|T|}\sum_{t\in T} Y_t^1 \tag{3.19}$$

and

$$\mu_{\mathscr{B}}\left(= \overline{Y_{C'}^0}\right) = \frac{1}{|C'|}\sum_{c\in C'} Y_c^0 \tag{3.20}$$

Note that the estimated treatment effect is given by

$$\widetilde{\tau}_T^1(C') = \frac{1}{|T|}\sum_{t\in T} Y_t^1 - \frac{1}{|C'|}\sum_{c\in C'} Y_c^0 = \mu_{\mathscr{A}} - \mu_{\mathscr{B}}. \tag{3.21}$$

Suppose that the control group C' is obtained by solving the optimization problem from BOSS with zero DOM imbalance measure where the response functions are linear. If there is no treatment effect, then the responses of T and the responses of C' should have the same mean. Hence, the following one-sided hypothesis test on whether there is a treatment effect can be constructed given that the control group is obtained by optimization with a zero objective value with zero bias.

H_0 : Elements of $\mathscr{A}$ and those of $\mathcal{B}$ have the same mean. (i.e., There is no treatment effect: $\mu_{\mathscr{A}} - \mu_{\mathcal{B}} = 0$).

H_1 : Elements of $\mathscr{A}$ and those of $\mathcal{B}$ do not have the same mean (i.e., $\mu_{\mathscr{A}} > \mu_{\mathscr{B}}$).

Bootstrap Procedure

1. Set $\mathfrak{M}$ to a large number. Set the iteration number $\mathfrak{m} = 1$.
2. Compute $\delta = \overline{Y_T^1} - \overline{Y_{C'}^0}$, difference in means of $\{Y_t^1\}_{t\in T}$ and $\{Y_c^0\}_{c\in C'}$.
3. Combine the two sets of observed response values into one set:

$$\Gamma = \mathscr{A} \cup \mathscr{B} = \{Y_t^1 \,|\, t \in T\} \cup \{Y_c^0 \,|\, c \in C'\} \tag{3.22}$$

4. Draw a sample of $|\mathscr{A}| = |T|$ observations with replacement from Γ and denote the mean of the sample by $\overline{Y_{T,m}^1}$. Similarly, draw a second sample of $|\mathscr{B}| = |C'|$ observations with replacement from the set Γ and denote the mean of this sample by $\overline{Y_{C',m}^0}$.
5. Compute $\delta_{\mathfrak{m}} = \overline{Y_{T,\mathfrak{m}}^1} - \overline{Y_{C',\mathfrak{m}}^0}$.
6. Increase the iteration number $\mathfrak{m}$ by 1. Repeat Step 4 and Step 5 for $\mathfrak{m} = 1, 2, \cdots, \mathfrak{M}$. Obtain $\mathfrak{M}$ values of $\delta_{\mathfrak{m}}$: $\delta_1, \delta_2, \cdots, \delta_{\mathfrak{M}}$.
7. Compute the p-value for the one-sided hypothesis test H_0 versus H_1

$$p = \frac{\sum_{\mathfrak{m}=1}^{\mathfrak{M}} \mathbb{1}[\delta_{\mathfrak{m}} \geq \delta]}{\mathfrak{M}} \tag{3.23}$$

Note that the Bootstrap procedure can also be applied for the two-sided hypothesis test H_0: $\mu_{\mathscr{A}} = \mu_{\mathscr{B}}$ vs $H_1 : \mu_{\mathscr{A}} \neq \mu_{\mathscr{B}}$ by replacing δ and δ_l in Step 7 by their absolute values $|\delta|$ and $|\delta_l|$. The p-value is then computed as

$$p = \frac{\sum_{\mathfrak{m}=1}^{\mathfrak{M}} \mathbb{1}[|\delta_{\mathfrak{m}}| \geq |\delta|]}{\mathfrak{M}} \tag{3.24}$$

3.5 Application: LaLonde Data

LaLonde used a dataset from National Supported Work Demonstration (NSW), an employment program conducted in the United States during the mid 1970s (LaLonde, 1986). Since then, this dataset has been analyzed by many researchers

(e.g., Heckman and Hotz (1989), Dehejia and Wahba (1999, 2002), Imbens (2003), Smith and Todd (2005), Abadie and Imbens (2012), and Colson et al. (2016)). In LaLonde (1986), LaLonde argues that the non-experimental estimates have specification errors as they fail to replicate the estimates from experimental data. He critiqued the use of non-experimental data by noting that the estimator from the non-experimental data differs greatly when compared to an estimator from the experimental data. The argument was later disputed by other researchers using his data. Heckman and Hotz (1989) discusses how to select a proper estimator from a wide range of non-experimental estimators. To overcome LaLonde's critique on the non-experimental results, Dehejia and Wahba (1999, 2002) used a propensity score matching method. Using this matching method for estimating the SATT with non-experimental data, the SATT value was comparable to what is obtained from the benchmark case with experimental data. The effectiveness of the matching method for program evaluation is further discussed in Smith and Todd (2005). Colson et al. (2016) compares various matching methods using simulated data and the LaLonde data.

An analysis of the LaLonde (1986) data using BOSS was first conducted by Cho et al. (2013). The analysis in this chapter differs from theirs in that this chapter consider a treatment effect estimator from decomposed parts of the data and conduct bootstrap hypothesis tests and a sub-sample analysis.

3.5.1 Entire Sample Analysis

The decomposition in Theorem 8 is applied to the LaLonde (1986) data. Specifically, the Dehejia and Wahba (1999) sample of the NSW Data is used as an experimental benchmark. The dataset includes information on earnings in 1974 (RE74) in addition to those in 1975 (RE75) and 1978 (RE78). This dataset is composed of 185 treated units and 260 control units. The observational dataset from Population Survey of Income Dynamics (PSID) by LaLonde (1986) is also obtained from the webpage of Rajeev Dehejia. The PSID dataset contains information on 2490 control units.

There are eight covariates in each dataset – age, education level, Black, Hispanic, marital status, no-degree, RE74, RE75 – and 1 response variable, RE78. In these datasets, RE74 and RE75 are pre-intervention earnings and RE78 are post-intervention earnings. BOSS balances these covariates. Then the response values

of the treatment group and that of the control group are compared to estimate the treatment effect. The dataset also has a treatment indicator showing whether a unit is treated.

Among the set of 185 treated units in the Dehejia-Wahba Sample, there are 52 units who were employed in both of the years investigated – 1974 and 1975. The remaining 133 units were unemployed in one or both of the years. The former group of 52 treated people is composed of those who were employed at some point as indicated by their nonzero total annual income. Denote the set of units who were employed in both years as the *Nonzero Income* dataset and the remaining as the *Zero Income* dataset. The latter group, Zero Income dataset, can also be noted as those who had a gap year since those units had zero income in 1974 or 1975 (or both). Using the same definition, all the other datasets are also partitioned into two sets – a Nonzero Income dataset and a Zero Income dataset. Among the units in the experimental control pool, there are 58 units in the Nonzero Income dataset and 202 units in the Zero Income dataset. In the PSID control pool, there are 2188 Nonzero Income units and 302 Zero Income units.

It is of interest to see how the treatment effects from the subsets of T and the corresponding subsets of C are related to the treatment effect of the entire set. The treatment effect of the entire NSW data can be decomposed using the treatment effects from the Nonzero Income dataset and the Zero Income dataset.

Suppose that T denotes the set of treated units in the NSW dataset. The sample average treatment effect for the treated in the NSW program, τ_T^1, can be decomposed as

$$\tau_T^1 = \frac{|T_1|}{|T|}\tau_{T_1}^1 + \frac{|T_2|}{|T|}\tau_{T_2}^1 \tag{3.25}$$

where $T_1 \subset T$ is a set of Nonzero Income units who were employed in both years (1974 and 1975) and $T_2 \subset T$ is a set of Zero Income units who were unemployed in 1974 or 1975.

While τ_T^1 is the value of interest, it cannot be observed directly from the data since the untreated responses of the treated units are unavailable. The outcome of the treated units in 1978 after treatment, not the untreated outcome of those units (i.e., the income level they would have gotten in 1978 if they were not treated), can be observed.

Throughout this chapter, DOM imbalance measure is used for the Balance Optimization Subset Selection. Time limits for solving the optimization problem are set to 300 seconds for each run of the BOSS method in a dual-core Windows lap-

top with Intel Core i7-4500U at 1.80GHz and a quad-core Windows desktop with Intel Core i5 at 2.67GHz. The estimated values of the SATT is rounded to the nearest integer.

In this section, p-values for one-sided hypothesis testing are reported to see whether the treatment effects are greater than zero. The p-values listed in Table 3.1 are obtained by taking the average of p-values after applying bootstrapping with $\mathfrak{M} = 3000$ for 100 replications for each specification. (In each simulation, 3000 samples of $|T|$ and $|C'|$ observations are drawn by repeating the Step 4 and Step 5 in section 3.4 for 3000 times. The same process is repeated for 100 times to find the p-values reported in the tables.)

Table 3.1: One-sided p-values of the Estimated Average Treatment Effects for the Treated

Using C from Experimental NSW Control Data			Using C' from Non-experimental PSID Control Data		
Estimated ATT		p-values	Estimated ATT		p-values
$\widehat{\tau}^1_T(C)$	1794	0.003	$\widehat{\tau}^1_T(C')$	1190	0.063
$\widehat{\tau}^1_{T_1}(C_1)$	-572	0.665	$\widehat{\tau}^1_T(C'_1)$	-2711	0.977
$\widehat{\tau}^1_{T_2}(C_2)$	2563	<0.001	$\widehat{\tau}^1_T(C'_2)$	3493	<0.001

For comparison, the SATT is also estimated using a control group C' that was obtained by propensity score matching with an R package `Matching`. The estimated values that are computed from PSID control data by using `Match` function without replacement are $\widehat{\tau}^1_T(C') = -85$ (one-sided p-value = 0.543), $\widehat{\tau}^1_T(C'_1) = -4529$ (0.999) and $\widehat{\tau}^1_T(C'_2) = 2899$ (0.001). The estimates from BOSS $\widehat{\tau}^1_T(C')$ and $\widehat{\tau}^1_{T_1}(C'_1)$ were closer to the benchmark values from experimental data than those of matching while the estimate $\widehat{\tau}^1_{T_2}(C_2)$ from matching was closer to the benchmark that that of BOSS. Note that $|1190 - 1794| < |-85 - 1794|$, $|-2711 - (-572)| < |-4529 - (-572)|$, and $|3493 - 2563| > |2899 - 2563|$. In addition, the values of matching estimates were not consistent in that they varied a lot when using different types of software implemented for matching.

The null hypothesis is rejected if the p-value is less than or equal to an α for some significance level α (e.g., $\alpha = 0.05$). In Table 3.1, the p-values from Nonzero Income datasets, 0.67 and 0.98, are greater than 0.1. Hence the null hypothesis that there is no treatment effect in the case of the Nonzero Income datasets

is not rejected under the significance level $\alpha = 0.05$ or $\alpha = 0.1$. In other cases, there was enough evidence supporting the alternative hypothesis since p-values are smaller than the given significance level $\alpha = 0.1$ or $\alpha = 0.05$.

The NSW program was effective when seeing the entire units that were the subjects of the experiment. However, that significant effect comes from the Zero Income dataset, not the Nonzero Income dataset. The program was not effective in increasing the salary earned in 1978 for those who were already employed in 1974 and 1975 when compared with what they would have earned if they were not participating in the NSW program. The significant effect observed is from those who were unemployed in at least one year among the investigated years – 1974 or 1975. A larger number of units in the Zero Income treated dataset who were unemployed in 1974 or 1975 became employed in 1978 and earned more when compared to those in Zero Income control dataset with similar covariates who did not receive the NSW program's training.

3.5.2 Sub-sample Analysis

In this section, the analyses in Section 3.5.1 are redone each with a thousand sub-samples. A sub-sample is constructed by combining 30 treated units that are randomly selected from the (all / Nonzero Income / Zero Income) NSW treatment data and the corresponding (all / Nonzero Income / Zero Income) controls from either the NSW control data or PSID control data.

The control group of size 30 which minimizes the DOM imbalance measure is selected through BOSS from those sub-samples with the results reported in the Figures 3.1a to 3.2c. The estimator values for SATT are different from the estimated SATT obtained in earlier sections. There is such a difference because the same control units can be chosen several times for different sub-samples. In the previous discussion, one control unit can only be chosen once since there was a single control group for each estimate. However, this sub-sample analysis provides valuable information on what can be expected as SATT values when BOSS is conducted with a treated sample of size 30.

Results with Experimental (NSW) Control Data

An estimated SATT is computed for each of the 1000 sub-samples and the significance of those values are examined using the Bootstrap Hypothesis testing method explained in the previous section. The treatment groups, T, used to compute p-values in Figure 3.1a are sets of 30 treated units randomly taken from the entire treatment group of the experimental NSW treated data. The treatment groups, T_1, used in Figure 3.1b are chosen in a similar manner as in Figure 3.1a. However, the treatment groups, T_1, are from the Nonzero Income treated dataset defined in Section 3.3, not the entire treated dataset. Similarly, treatment groups, T_2, used in Figure 3.1c are drawn from the Zero Income dataset.

The entire units in the experimental NSW control data are used as a control pool for the first set of experiments with BOSS whose results is depicted in Figure 3.1a. When computing p-values the figure, selected control groups respectively of the same size with sub-sample's treatment group size (i.e., 30 units in the experiment) are used. For Figures 3.1b and 3.1c, the units in the Nonzero Income dataset and the Zero Income dataset respectively are used among the experimental NSW control data as their corresponding control pools.

In the first set of experiments depicted in Figure 3.1a, the average value of the $\widetilde{\tau}^1_T(C')$ is \$1564. There are 177 observations with p-value smaller than a significance level 0.05 and 298 observations with p-value smaller than a significance level 0.1 out of 1000 sub-samples investigated. The significance of the treatment effect of the entire treated dataset is mainly from some sub-population which has large treatment effect values.

In the experiments depicted in Figure 3.1b, the average of the estimated treatment effects $\widetilde{\tau}^1_{T_1}(C'_1)$ from 1000 different choices of T_1 and the corresponding C'_1 is $-$\$579. There were only 11 observations whose p-values were smaller than the significance level 0.1 (none for the significance level 0.05). This small number of significant cases is likely to be what might have seen just because of repeated tests since some will be significant by chance when repeating. These results illustrate that there was no significant effect for most choices of T_1 when selecting T_1 of size 30 from the units who were employed in both 1974 and 1975.

In the experiments of Figure 3.1c, the average $\widetilde{\tau}^1_{T_2}(C'_2)$ is \$2428 and 356 out of 1000 observations had their p-values smaller than 0.05 and 570 out of 1000 had their p-values smaller than 0.1.

The three figures demonstrate that the units that showed significant increase in

their response values are from the Zero Income dataset, a group of people who were not employed in at least one of the two years, not the Nonzero Income dataset. This result with sub-samples is consistent with the result from the previous section. Also note that the distribution of p-values are skewed to the right in case of the first experiment with entire treatment units and the third experiment with Zero Income units while that of the second experiment with Nonzero Income units are skewed to the left.

Results with Non-experimental (PSID) Control Data

Now the treatment effect estimate is computed using BOSS on the NSW treated units and the PSID control units. In Figure 3.2a, results of p-values obtained after conducting BOSS with 1000 sub-samples composed of the treatment groups T of size 30 randomly drawn from the entire NSW treated units and all the PSID control units are depicted. Similarly, Figures 3.2b and 3.2c contain results from 1000 sub-samples composed of T_1's of size 30 randomly drawn from the Nonzero Income / Zero Income NSW treated units and the Nonzero Income / Zero Income PSID control units respectively. These results from Figures 3.2a to 3.2c and earlier results in Figures 3.1a to 3.1c are also summarized in Table 3.2.

These results in Figures 3.2a, 3.2b, and 3.2c are consistent with the results in Figures 3.1a, 3.1b, and 3.1c in that the significant positive effect was derived from sub-samples which contain Zero Income units mainly. There were zero treatment groups having a significant positive treatment effect among the 1000 treatment groups T_1 selected from the Nonzero Income dataset with the one-sided p-value less than the significance level 0.1. In the Nonzero Income dataset, all the others had no significant positive effect from the training of the NSW program.

Furthermore, general skewness pattern of the density function depicted in Figures 3.2a, 3.2b, and 3.2c coincides with those in Figures 3.1a, 3.1b, and 3.1c: Figures 3.2a and 3.2c are right-skewed while Figure 3.2b is left-skewed. Left-skewness indicates that there were many observations with insignificant (i.e., large) p-values.

The average estimated SATT is \$1441 in the fourth set of experiments depicted in Figure 3.2a. In the set, there were 148(285) observations whose p-values were smaller than 0.05(0.1). The average estimated SATT is −\$4127 in the fifth set of experiments in Figure 3.2b. Among the values in the fifth set, zero observations had their p-value smaller than 0.1. In the last set of experiments used to generate

Figure 3.2c, the average estimated SATT is \$3030. In the last case, 432 observations had their p-values smaller than 0.05 and 658 observations had their p-values smaller than 0.1.

3.6 Concluding Remarks

Firstly, in this chapter, a method for decomposition of the treatment effect as a weighted average of effects from its partition was introduced. The difference between the entire set's estimate and the weighted average of estimates from the partitioned sets is explained. The general expression for the difference is given by $D_1 + D_2$ as in (3.11). This is equal to D_1 in case of experimental data and to D_2 in case of non-experimental data. Secondly, a bootstrap hypothesis testing procedure which can be applied to get the p-values giving information on the significance of the treatment effect values was provided.

As an application of the theoretical results, the LaLonde (1986) data was analyzed. In the partition $\{T_1, T_2\}$ of T that was considered, T_1 is the Nonzero Income dataset defined as the set of units who were employed in both years (1974 and 1975) and T_2 is the Zero Income units defined as those who are not. Though partitioning of the units under consideration into Nonzero Income and Zero Income units was used, the methods discussed in this chapter can be applied with the same logic to any other partition and the effects from those subsets can be checked.

Using two-sample bootstrap hypothesis testing, the significance of the treatment effect values was tested. From the literature, it is well known that the sample average treatment effect of the treated value is about \$1800. On the other hand, none of the previous studies focused on checking the significance of the heterogeneous treatment effect from decomposed parts. The new method proposed in this chapter suggests that the treatment effect from those who were employed in both years investigated was not significantly greater than zero and the significant effect is from those who were not employed in at least one of those years.

Additionally, the above results were confirmed by taking 6 sets of 1000 subsamples with 30 treated units and the corresponding control units. The results obtained by analyzing the sub-samples are consistent with those with the entire sample.

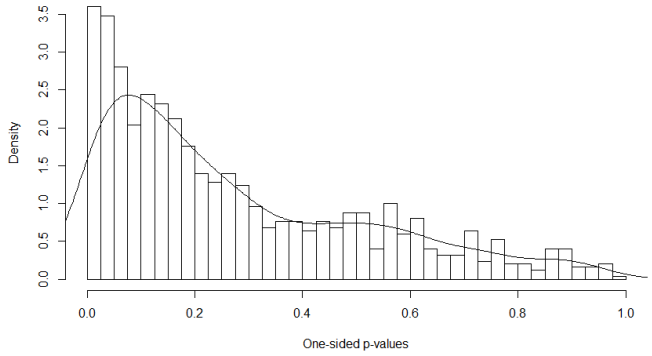
When applying BOSS in this chapter, one specific imbalance measure, namely $\mathscr{I}_{\text{DOM}}$, was used. Using other imbalance measures that are more highly ranked in

the balance hierarchy as defined in Chapter 2 is also possible. However using other imbalance measures will increase the computational cost although more balance in the covariate distributions may be obtained.

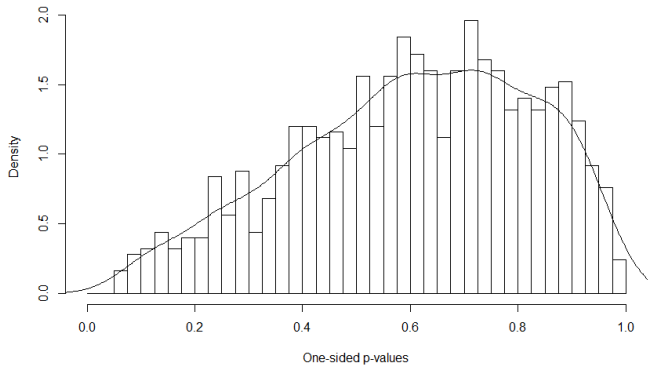
This way of evaluating the labor program's effectiveness would be able to shed light on related future research – not only for the evaluation of the labor training program but also for other program evaluations as well. In particular, it is possible to identify whether a specific subset of the groups under consideration has a significant treatment effect by using the approach proposed in this chapter.

Figure 3.1: Histogram and Density of 1000 p-values of Bootstrap Hypothesis Testing with a Treatment Group from Experimental NSW Treatment Data and the Corresponding Control Group from Experimental NSW Control Data

(a) Simulation 1: p-values with T of size 30 and Corresponding C'



(b) Simulation 2: p-values with Nonzero Income T_1 of size 30 and Corresponding Nonzero Income C_1'



(c) Simulation 3: p-values with Zero Income T_2 of size 30 and Corresponding Zero Income C_2'

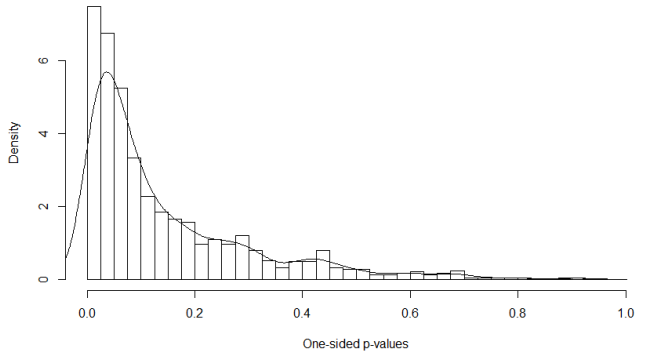
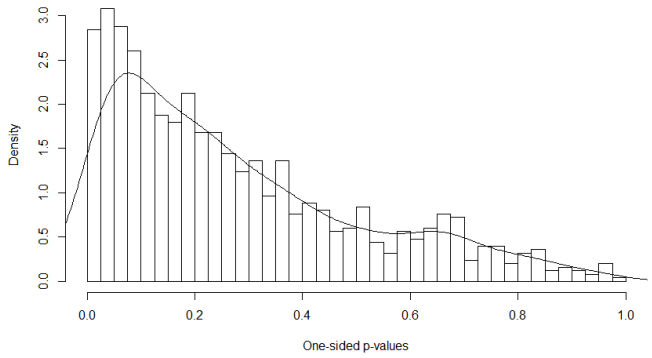
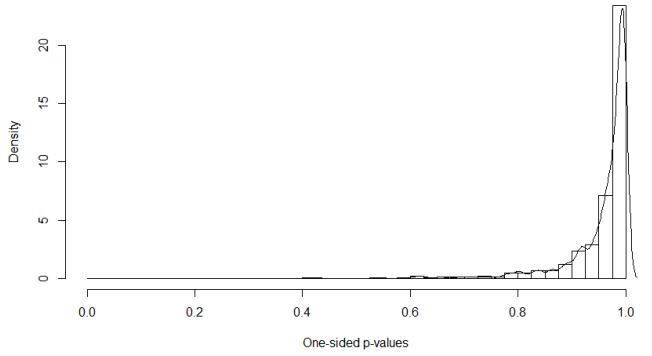


Figure 3.2: Histogram and Density of 1000 p-values of Bootstrap Hypothesis Testing with a Treatment Group from Experimental NSW Treatment Data and the Corresponding Control Group from Non-experimental PSID Control Data

(a) Simulation 4: p-values with T of size 30 and Corresponding C'



(b) Simulation 5: p-values with Nonzero Income T_1 of size 30 and Corresponding Nonzero Income C'_1



(c) Simulation 6: p-values with Zero Income T_2 of size 30 and Corresponding Zero Income C'_2

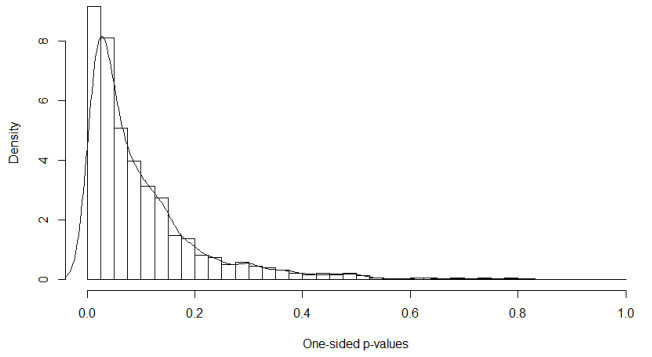
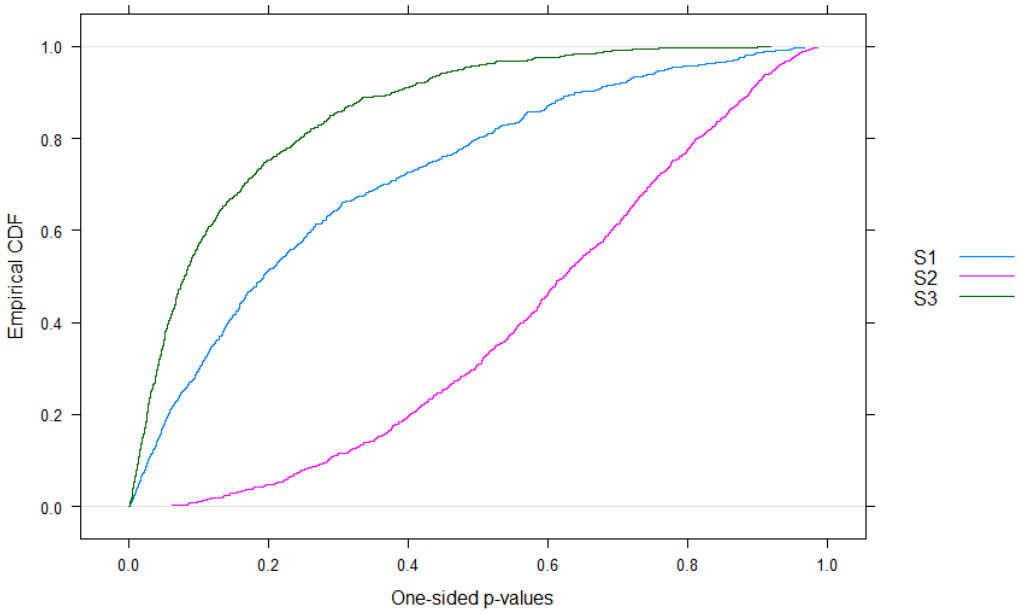


Figure 3.3: Empirical CDF of One-sided p-values

(a) Simulations 1, 2, and 3 (respectively S1, S2, and S3)



(b) Simulations 4, 5, and 6 (respectively S4, S5, and S6)

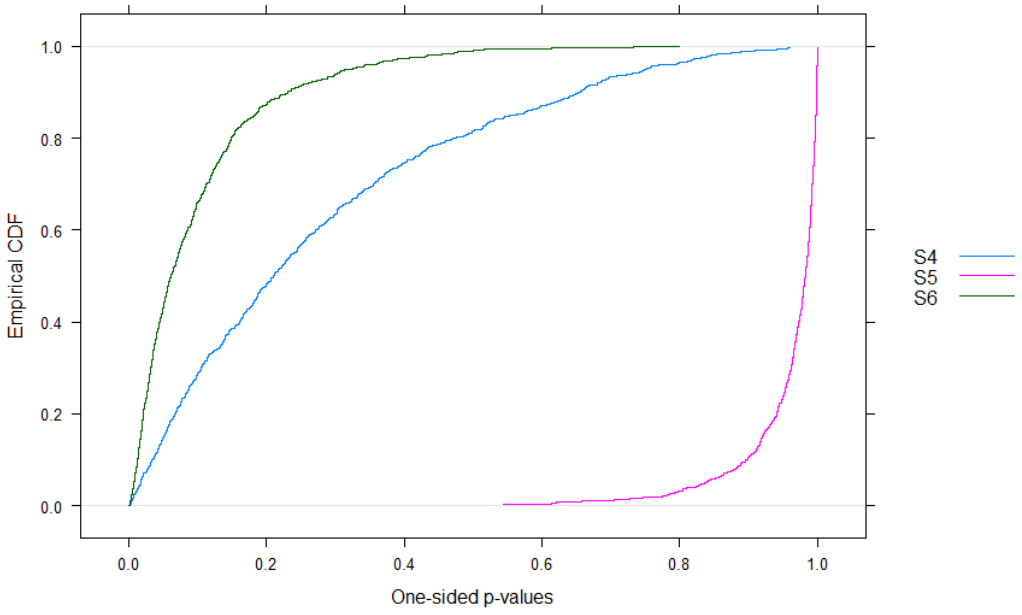


Table 3.2: One-sided p-values when Bootstrap Hypothesis Testing is Conducted to Respective Treatment Effects

Interval	Number of p-values in the Specified Interval											
	Experimental (NSW) Control Data						Non-experimental (PSID) Control Data					
	T with C'		T_1 with C'_1		T_2 with C'_2		T with C'		T_1 with C'_1		T_2 with C'_2	
	Freq.	Cumulative Density	Freq.	Cumulative Density	Freq.	Cumulative Density	Freq.	Cumulative Density	Freq.	Cumulative Density	Freq.	Cumulative Density
0.00-0.05	177	0.177	0	0.000	356	0.356	148	0.148	0	0.000	432	0.432
0.05-0.10	121	0.298	11	0.011	214	0.570	137	0.285	0	0.000	226	0.658
0.10-0.15	119	0.417	19	0.030	103	0.673	100	0.385	0	0.000	146	0.804
0.15-0.20	97	0.514	18	0.048	80	0.753	98	0.483	0	0.000	71	0.875
0.20-0.25	67	0.581	31	0.079	51	0.804	84	0.567	0	0.000	38	0.913
0.25-0.30	66	0.647	36	0.115	54	0.858	67	0.634	0	0.000	26	0.939
0.30-0.35	41	0.688	28	0.143	33	0.891	58	0.692	0	0.000	21	0.960
0.35-0.40	38	0.726	53	0.196	20	0.911	53	0.745	0	0.000	13	0.973
0.40-0.45	35	0.761	58	0.254	32	0.943	42	0.787	1	0.001	9	0.982
0.45-0.50	39	0.800	55	0.309	15	0.958	29	0.816	0	0.001	9	0.991
0.50-0.55	32	0.832	69	0.378	10	0.968	32	0.848	1	0.002	4	0.995
0.55-0.60	40	0.872	85	0.463	7	0.975	22	0.870	1	0.003	0	0.995
0.60-0.65	30	0.902	83	0.546	8	0.983	27	0.897	5	0.008	2	0.997
0.65-0.70	16	0.918	68	0.614	10	0.993	37	0.934	4	0.012	1	0.998
0.70-0.75	22	0.940	91	0.705	2	0.995	16	0.950	6	0.018	1	0.999
0.75-0.80	18	0.958	73	0.778	2	0.997	15	0.965	14	0.032	1	1.000
0.80-0.85	8	0.966	68	0.846	1	0.998	17	0.982	27	0.059	0	1.000
0.85-0.90	20	0.986	75	0.921	1	0.999	7	0.989	46	0.105	0	1.000
0.90-0.95	8	0.994	54	0.975	1	1.000	5	0.994	131	0.236	0	1.000
0.95-1.00	6	1.000	25	1.000	0	1.000	6	1.000	764	1.000	0	1.000
Total	1000		1000		1000		1000		1000		1000	

CHAPTER 4

BALANCE OPTIMIZATION SUBSET SELECTION WITH MULTIPLE TREATMENT LEVELS

4.1 Introduction

Health-care professionals often encounter data having multiple treatments. For example, pharmaceutical scientists may be interested in testing the effect of different doses of pills and comparing the treatment effects between different dosages (e.g., control units, treated units with one dose of pills, and treated units with two doses of pills). Furthermore, the treatment levels may not be hierarchical: pharmaceutical scientists may be interested in the comparative effectiveness of two different types of pills on patients.

Matching methods for causal inference were introduced under a binary treatment setting where units are either treated or non-treated (Cochran and Rubin, 1973; Rubin, 1973). Extensions of matching methods on multi-treatment data where there are more than two treatments have been made (Imbens, 2000; Lechner, 2001; Yang et al., 2016). Evaluation of the European labor market program was provided as an example having multiple treatments by Lechner (2001). In the European labor market program, there are a number of training programs and individuals can participate in one of them. Additional applications of the earlier theoretical results on dose responses can be found in Wang et al. (2001) and Foster (2003).

The purpose of this chapter, Chapter 4, is to develop a non-matching technique that is applicable to multi-treatment data using the Balance Optimization Subset Selection framework, which was introduced in a binary treatment setting by Nikolaev et al. (2013). To date, BOSS has been discussed only under the binary treatment settings (e.g., level 0 for control units and level 1 for treated units). In this chapter, the earlier result is extended to multiple treatments. In particular, after reviewing the theoretical results discussed in Yang et al. (2016), comparable results on average treatment effects with multi-treatments are developed using

BOSS.

Chapter 4 is organized as follows. Section 4.2 defines population average treatment effect and sample average treatment effect under the multi-treatment setting. Section 4.3 discusses the strong and weak ignorability assumptions. Section 4.4 reviews the theory for matching under multiple treatments. Section 4.5 provides some theoretical results for BOSS under multiple treatments. Section 4.6 reproduces the simulation study that was conducted by Yang et al. (2016) to compare their estimator with a multi-treatment BOSS estimator and provides additional simulation results. Section 4.7 provides concluding remarks and directions for future research.

4.2 Average Treatment Effect

In this section, definitions of Population Average Treatment Effect (PATE) and Sample Average Treatment Effect (SATE) and their relationship are provided. Let $\mathcal{U}$ be a population, or set of all possible units. Suppose that there are L treatment levels, $0, 1, \cdots, L-1$. Let random variable Z denote a treatment level of a randomly selected unit where the random selection is uniform and let Z_u denote a treatment level for unit $u \in \mathcal{U}$. Note that, while Z and $newTI$ could have only a value that is either 0 or 1 in the previous chapters, now they can have any value in $\{0, 1, \cdots, L-1\}$. Let $\mathcal{L} = \{0, 1, \cdots, L-1\}$ be a set of possible treatment levels. Then, $Z_u \in \mathcal{L}$. Denote the set of units with treatment level i as

$$U^i \equiv \{u \in \mathcal{U} : Z_u = i\} \text{ for } i \in \mathcal{L} \tag{4.1}$$

with $\mathcal{U} = U^0 \cup U^1 \cup \cdots \cup U^{L-1}$. By design, $U^m \cap U^l = \varnothing$ for any $m, l \in \mathcal{L}$, $m \neq l$. Suppose that sets $S^0, S^1, \cdots, S^{L-1}$ are samples drawn from $U^0, U^1, \cdots, U^{L-1}$ respectively, with the observed unit set $\mathcal{N}$ given by $S^0 \cup S^1 \cup \cdots \cup S^{L-1}$.

As defined under a binary treatment setting, let Y_u^i denote the response of unit u for treatment level i. Note that hypothetically each unit u has a response Y_u^i for each treatment levels $i \in \mathcal{L}$. However, the response Y_u^i of unit u can be observed only if $Z_u = i$. Let Y^i denote the response for treatment level i for a unit that is selected uniformly from population $\mathcal{U}$. Then $\mathbb{E}[Y^m]$ denotes the average population response for treatment level m. The value of interest is the difference of the average population responses – PATE between treatment level l and treatment

level m, which is defined as

$$\tau(m, l) \equiv \mathbb{E}[Y^m - Y^l] = \mathbb{E}[Y^m] - \mathbb{E}[Y^l]. \tag{4.2}$$

The SATE between treatment level l and treatment level m is defined as

$$\tau_{m,l} \equiv \frac{1}{|\mathcal{N}|} \sum_{u \in \mathcal{N}} \left(Y_u^m - Y_u^l \right). \tag{4.3}$$

Since $S^0, S^1, \cdots, S^{L-1}$ are mutually exclusive and exhaustive sets, (4.3) can be written as

$$\tau_{m,l} = \frac{1}{|S^0| + |S^1| + \cdots + |S^{L-1}|} \sum_{i=0}^{L-1} \sum_{u \in S^i} \left(Y_u^m - Y_u^l \right). \tag{4.4}$$

Note that $\tau_{m,l}$ cannot be directly computed since the values Y_u^l for $u \notin S^l$ and Y_u^m for $u \notin S^m$ cannot be observed. The objective is to find an unbiased estimator for $\tau_{m,l}$ using the observed values.

Assumption 2. Assume that the sets S^i, $i \in \mathcal{L}$ are random samples drawn uniformly without replacement from the corresponding populations U^i, $i \in \mathcal{L}$, respectively.

Assumption 3. Suppose that the probability of a unit being at treatment level i is given by

$$\text{Prob}(Z = i) = \frac{|S^i|}{|S^0| + |S^1| + \cdots + |S^{L-1}|} \text{ for all } i \in \{0, 1, \cdots, L-1\}. \tag{4.5}$$

Under Assumptions 2 and 3, Theorem 9 shows that $\mathbb{E}[\tau_{m,l}]$, the expected value of the sample average treatment effect, is equal to $\tau(m, l)$, the population average treatment effect.

Theorem 9. *Under Assumptions 2 and 3,*

$$\mathbb{E}[\tau_{m,l}] = \tau(m, l). \tag{4.6}$$

Proof.

$$
\begin{aligned}
\mathbb{E}[\tau_{m,l}] &= \mathbb{E}\left[\frac{1}{|S^0| + |S^1| + \cdots + |S^{L-1}|}\sum_{i=0}^{L-1}\sum_{u\in S^i}\left(Y_u^m - Y_u^l\right)\right] \\
&= \frac{1}{|S^0| + |S^1| + \cdots + |S^{L-1}|}\sum_{i=0}^{L-1}\sum_{u\in S^i}\mathbb{E}\left[Y_u^m - Y_u^l\right] \\
&= \frac{1}{|S^0| + |S^1| + \cdots + |S^{L-1}|}\sum_{i=0}^{L-1}|S^i|\cdot\mathbb{E}\left[Y^m - Y^l|Z=i\right] \\
&= \sum_{i=0}^{L-1}\left(\frac{|S^i|}{|S^0| + |S^1| + \cdots + |S^{L-1}|}\right)\mathbb{E}\left[Y^m - Y^l|Z=i\right] \\
&= \sum_{i=0}^{L-1}\text{Prob}(Z=i)\cdot\mathbb{E}\left[Y^m - Y^l|Z=i\right] \\
&= \mathbb{E}[Y^m - Y^l] = \tau(m,l).
\end{aligned}
\tag{4.7}
$$

From the second line to the third line, equation (4.8) is used:

$$
\mathbb{E}[Y_u^i] = \mathbb{E}[Y^i|Z=j] \text{ for all } u \in S^j. \tag{4.8}
$$

which holds by Assumption 2 for uniform random sampling. □

In the binary treatment case with $L = 2$ (i.e., in the observational data there are two sets of samples, S^0 and S^1, with treatment level 0 and treatment level 1 respectively), S^0 is referred to as C, a set of control units, while S^1 is referred as T, a set of treated units. Then SATE in (4.3) reduces to

$$
\begin{aligned}
\tau_{1,0} &= \frac{1}{|S^0| + |S^1|}\left(\sum_{u\in S^0}(Y_u^1 - Y_u^0) + \sum_{u\in S^1}(Y_u^1 - Y_u^0)\right) \\
&= \frac{1}{|C| + |T|}\left(\sum_{c\in C}(Y_c^1 - Y_c^0) + \sum_{t\in T}(Y_t^1 - Y_t^0)\right).
\end{aligned}
\tag{4.9}
$$

It coincides with the definition of SATE for binary treatment case stated in Sauppe (2015).

4.3 Strong Ignorability and Weak Ignorability Assumptions

Strong ignorability and weak ignorability assumptions are reviewed in this section. Denote a vector of K covariates for unit $u \in \mathcal{U}$ by $\mathbf{X}_u = (X_{u,1}, X_{u,2}, \cdots, X_{u,K})$. Let $\mathbf{X}$ be a vector of K covariate values for a unit selected randomly from population $\mathcal{U}$ where the random selection is uniform. Let $\mathcal{X}$ denote the set of possible values for $\mathbf{X}$.

A *strong ignorability* condition and a *weak ignonrability* condition are said to be satisfied if the following statements in Assumptions 4 and 5 hold, respectively (Yang et al., 2016). Assumption 4, which is discussed in Imbens (2000) and Lechner (2001), is a generalized version of the strong ignorability assumption for binary treatment case introduced in Rosenbaum and Rubin (1983b).

Assumption 4. *Strong Ignorability*: The random variables $Y^0, Y^1, \cdots Y^{L-1}, Z$ and $\mathbf{X}$ satisfy

$$(Y^0, Y^1, \cdots Y^{L-1}) \perp\!\!\!\perp Z \mid \mathbf{X} \tag{4.10}$$

and

$$0 < P(Z = i \mid \mathbf{X} = x) < 1 \text{ for all } i \in \mathcal{L} \text{ and } x \in \mathcal{X}. \tag{4.11}$$

where the notation $\perp\!\!\!\perp$ denotes conditional independence.

Assumption 5. *Weak Ignorability* (Yang et al., 2016)
The random variables Y^i, Z and $\mathbf{X}$ satisfy

$$Y^i \perp\!\!\!\perp Z \mid \mathbf{X} \text{ for all } i \in \mathcal{L} \tag{4.12}$$

and

$$0 < P(Z = i \mid \mathbf{X} = x) < 1 \text{ for all } i \in \mathcal{L} \text{ and } x \in \mathcal{X}. \tag{4.13}$$

Strong ignorability (Assumption 4) is stronger than the weak ignorability (Assumption 5) since mutual independence implies pairwise independence and not vice versa.

4.4 Matching with Multiple Treatment Levels

This section reviews the matching method with more than two treatment levels (Yang et al., 2016). The estimator of PATE in Yang et al. (2016) was constructed

by matching under weak ignorability (Assumption 5).

Consider a conventional method that compares a pair of treatment levels in a setting with multiple treatment levels. The method leads to an estimator $\hat{\nu}(m, l)$ of $\nu(m, l)$ in (4.14) which focuses on a specific sub-population composed of units whose treatment level is either m or l.

$$\nu(m, l) = \mathbb{E}\left[Y^m - Y^l \mid Z \in \{l, m\}\right] = \mathbb{E}\left[Y^m \mid Z \in \{l, m\}\right] - \mathbb{E}\left[Y^l \mid Z \in \{l, m\}\right]. \quad (4.14)$$

However, $\nu(m, l) \neq \tau(m, l)$ since the sub-population, $U^l \cup U^m$, is different from the set of all units, $\mathcal{U} = \{u : Z_u \in \mathcal{L}\}$. Furthermore, the estimators $\hat{\nu}(m, l)$ and $\hat{\nu}(m', l')$ are not comparable for $(m', l') \neq (m, l)$ since they use different sub-populations for estimation (Yang et al., 2016).

To avoid the issue that arises from using a different sub-population, Yang et al. (2016) proposes a method that estimates $\mathbb{E}[Y^l]$ for $l \in \mathcal{L}$ based on the entire population. As a result, they can estimate PATE between a treatment level l and a treatment level m, defined as $\mathbb{E}[Y^m] - \mathbb{E}[Y^l]$. Yang et al. (2016) proposes the following methods of estimating PATE with multi-treatment levels.

- Generalized Nearest Neighbor Covariate Matching
  - The estimator of $\tau(m, l)$ is given by

$$\widetilde{\tau}_{GNN}(m, l) = \frac{1}{|\mathcal{N}|} \sum_{u \in \mathcal{N}} \left(Y^m_{f(u,m)} - Y^l_{f(u,l)}\right) \quad (4.15)$$

    where the function $f(\cdot, \cdot)$, which takes a unit and a treatment level as its arguments, is given by $f(u, j) \equiv \arg\min_{v:Z_v=j} \|X_v - X_u\|$. The notation $\|\cdot\|$ denotes a metric used (e.g., Mahalanobis metric).

- Generalized Propensity Score (GPS) Matching
  - The estimator of $\tau(m, l)$ is given by

$$\widetilde{\tau}_{GPS}(m, l) = \frac{1}{|\mathcal{N}|} \sum_{u \in \mathcal{N}} \left(Y^m_{g(u,m)} - Y^l_{g(u,l)}\right) \quad (4.16)$$

    where the function $g(\cdot, \cdot)$ is given by $g(u, j) \equiv \arg\min_{v:Z_v=j} |p(j|\mathbf{X}_v) - p(j|\mathbf{X}_u)|$. The notation $p(\cdot \mid \cdot)$ denotes a generalized propensity score, defined as

$$p(j \mid x) = \text{Prob}(Z_v = j \mid \mathbf{X}_v = x). \quad (4.17)$$

Under Assumptions 2 and 5, both $\widetilde{\tau}_{GNN}(m, l)$ and $\widetilde{\tau}_{GPS}(m, l)$ are consistent estimators of $\tau(m, l)$ as both estimators are asymptotically normally distributed (Yang et al., 2016).

This chapter uses the BOSS method to define an estimator for $\tau(m, l)$. Furthermore, the resulting generalized BOSS method subsumes the estimator $\widetilde{\tau}_{GNN}(m, l)$ from generalized matching method in that the BOSS method with imbalance measure that balances the full joint distribution is equivalent to matching.

4.5 BOSS with Multiple Treatment Levels

This section extends the BOSS framework to a multiple treatment setting. The BOSS framework, which is used to estimate the sample average treatment effect for the treated (SATT) for the entire population in a binary treatment setting, is reviewed in Section 4.5.1. The BOSS method is applied to estimate PATE and expected value of SATE under multiple treatment setting in Section 4.5.2.

4.5.1 BOSS Method on Estimating ATT in a Binary Treatment Setting

First, the conventional BOSS method under a binary treatment setting is reviewed in this subsection. As described in Section 4.2 and earlier chapters, the control pool is denoted by $C\left(= S^0\right)$ and the treatment group is denoted by $T\left(= S^1\right)$. In the BOSS method under a binary treatment setting, the objective is to find a control group C' from the control pool C that minimizes an imbalance measure. Then the estimator for the sample average treatment effect for the treated (SATT) is given by

$$\widetilde{\tau}_T^1(C') \equiv \frac{1}{|T|} \sum_{t \in T} Y_t^1 - \frac{1}{|C'|} \sum_{c \in C'} Y_c^0. \tag{4.18}$$

Note that SATT is defined as

$$\tau_T^1 \equiv \frac{1}{|T|} \sum_{t \in T} \left(Y_t^1 - Y_t^0\right) \tag{4.19}$$

while PATT is defined as $\tau^1 \equiv \mathbb{E}[Y^1 - Y^0 | Z = 1]$.

Note that, under Assumption 2, the expected value of SATT is equal to PATT (Sauppe, 2015). Furthermore, suppose that a subset of control units $C' \subset C$ is

selected in a way that the covariate distributions of the treatment group and the control group are identical (i.e., $\{\mathbf{X}_u\}_{u\in C'} = \{\mathbf{X}_u\}_{u\in T}$). Then, under the strong ignorability assumption (Assumption 4), $\mathbb{E}[\widetilde{\tau}^1_T(C')] = \tau^1$ (Nikolaev et al., 2013).

4.5.2 BOSS Method to Estimate ATE in a Multi-Treatment Setting

In this section, the new BOSS method under a multi-treatment setting is stated and a proof on the unbiasedness of the new estimator will be provided. Recall that $\mathcal{N} \equiv \cup_{i=0}^{L-1} S^i$. The set of observed units, $\mathcal{N} = \{u_1, \cdots, u_N\}$, can be written as a union of S^l, S^m, and $\mathcal{N} \setminus (S^l \cup S^m)$. Hence, the matching estimator $\widetilde{\tau}_{GNN}(m, l)$ in (4.15) can be written as

$$\begin{aligned}
&\widetilde{\tau}_{GNN}(m, l) \\
&= \frac{1}{|\mathcal{N}|} \sum_{u\in\mathcal{N}} \left(Y^m_{f(u,m)} - Y^l_{f(u,l)}\right) \\
&= \frac{1}{|\mathcal{N}|} \left\{ \sum_{u\in S^l} \left(Y^m_{f(u,m)} - Y^l_u\right) + \sum_{u\in S^m} \left(Y^m_u - Y^l_{f(u,l)}\right) + \sum_{u\in(\mathcal{N}\setminus(S^l\cup S^m))} \left(Y^m_{f(u,m)} - Y^l_{f(u,l)}\right) \right\}
\end{aligned} \tag{4.20}$$

where $f(u, j) \equiv \arg\min_{v:Z_v=j} ||X_v - X_u||$ returns a unit v in S^j whose covariate value is the closest to u. Following facts are used when going from the first line to the second line in (4.20): $f(u, l) = u$ for $u \in S^l$ and $f(u, m) = u$ for $u \in S^m$.

Estimating $\tau(m, l)$ requires Y^m to be estimated for units in S^l, Y^l for units in S^m, and both Y^m and Y^l for units outside of S^l and S^m. In the binary treatment setting, BOSS shifts from estimating unit-level responses (as done in matching) to estimating group-level responses. A similar approach can be done here. Rather than estimating the unit-level responses Y^m_u for each unit $u \in S^l$ by matching, the average response $\sum_{u\in S^l} Y^m_u/|S^l|$ is estimated. To do this, a group of units $(S^m)'$ are obtained from S^m that is balanced with respect to S^l. A similar process is used to estimate the other parts as well.

Equation (4.21) is the BOSS estimator of $\mathbb{E}[\tau_{m,l}]$,

$$\widetilde{\tau}_{BOSS}(m, l) = \frac{1}{|\mathcal{N}|} \left\{ \sum_{u\in(S^m)'} Y^m_u - \sum_{u\in S^l} Y^l_u + \sum_{u\in S^m} Y^m_u - \sum_{u\in(S^l)'} Y^l_u + \sum_{u\in(S^m)''} Y^m_u - \sum_{u\in(S^l)''} Y^l_u \right\} \tag{4.21}$$

where $(S^l)', (S^l)'', (S^m)'$, and $(S^m)''$ are optimal groups that minimizes the BOSS imbalance measures (e.g., imbalance measures given in (4.22) to (4.25)). These groups $(S^l)', (S^l)'', (S^m)'$, and $(S^m)''$ are selected from S^l and S^m, respectively, while allowing multiple instances of each unit in each set so that $(S^l)', (S^l)'', (S^m)'$, and $(S^m)''$ become multisets.

Denote the union of sampled units other than those belonging to S^m and S^l by $S^- \equiv \mathcal{N} \setminus (S^l \cup S^m)$. Consider various forms of imbalance measures that compare the following pairs:

- S^l and $(S^m)'$ where $|S^l| = |(S^m)'|$
- S^m and $(S^l)'$ where $|S^m| = |(S^l)'|$
- S^- and $(S^m)''$ where $|S^-| = |(S^m)''|$
- S^- and $(S^l)''$ where $|S^-| = |(S^l)''|$

Solve for a set of units $(S^m)' \subset S^m$ that is balanced with S^l and a set of units $(S^l)' \subset S^l$ that is balanced with S^m by minimizing the imbalance measures for binary treatment case, $\mathscr{I}\left(S^l, (S^m)'\right)$ and $\mathscr{I}\left(S^m, (S^l)'\right)$, respectively. In addition, solve for two sets of units $(S^m)'' \subset S^m$ and $(S^l)'' \subset S^l$ that are balanced with S^- by minimizing $\mathscr{I}\left(S^-, (S^m)''\right)$ and $\mathscr{I}\left(S^-, (S^l)''\right)$.

Recall that Chapter 2 provided several imbalance measures for the binary treatment case. The two sets (T, C') in (2.9) to (2.13) can be replaced with an appropriate pair of sets such as $(S^l, (S^m)')$, $(S^m, (S^l)')$, $(S^-, (S^m)'')$, and $(S^-, (S^l)')$ to minimize imbalance between them.

The imbalance measures given in (4.22) to (4.25) for the multi-treatment case can be obtained by combining the imbalance measures, $\mathscr{I}_{\text{DOM}}$, $\mathscr{I}_{\text{DOM+DOV}}$, $\mathscr{I}_{\text{DOM2}}$, and $\mathscr{I}_{\text{KS}}$ for the binary treatment case.

$$\begin{aligned}
&\bullet \mathscr{I}_{\text{DOM}}\left(S^l, (S^l)', (S^l)'', S^m, (S^m)', (S^m)''\right) \qquad (4.22)\\
&\quad = \mathscr{I}_{\text{DOM}}\left(S^l, (S^m)'\right) + \mathscr{I}_{\text{DOM}}\left(S^m, (S^l)'\right)\\
&\qquad + \mathscr{I}_{\text{DOM}}\left(S^-, (S^m)''\right) + \mathscr{I}_{\text{DOM}}\left(S^-, (S^l)''\right)\\
&\bullet \mathscr{I}_{\text{DOM+DOV}}\left(S^l, (S^l)', (S^l)'', S^m, (S^m)', (S^m)''\right) \qquad (4.23)\\
&\quad = \mathscr{I}_{\text{DOM+DOV}}\left(S^l, (S^m)'\right) + \mathscr{I}_{\text{DOM+DOV}}\left(S^m, (S^l)'\right)\\
&\qquad + \mathscr{I}_{\text{DOM+DOV}}\left(S^-, (S^m)''\right) + \mathscr{I}_{\text{DOM+DOV}}\left(S^-, (S^l)''\right)
\end{aligned}$$

$$\bullet \mathscr{I}_{\mathrm{DOM2}}\Big(S^l, (S^l)', (S^l)'', S^m, (S^m)', (S^m)''\Big) \tag{4.24}$$
$$= \mathscr{I}_{\mathrm{DOM2}}\Big(S^l, (S^m)'\Big) + \mathscr{I}_{\mathrm{DOM2}}\Big(S^m, (S^l)'\Big)$$
$$+ \mathscr{I}_{\mathrm{DOM2}}\Big(S^-, (S^m)''\Big) + \mathscr{I}_{\mathrm{DOM2}}\Big(S^-, (S^l)''\Big)$$
$$\bullet \mathscr{I}_{\mathrm{KS}}\Big(S^l, (S^l)', (S^l)'', S^m, (S^m)', (S^m)''\Big) \tag{4.25}$$
$$= \mathscr{I}_{\mathrm{KS}}\Big(S^l, (S^m)'\Big) + \mathscr{I}_{\mathrm{KS}}\Big(S^m, (S^l)'\Big)$$
$$+ \mathscr{I}_{\mathrm{KS}}\Big(S^-, (S^m)''\Big) + \mathscr{I}_{\mathrm{KS}}\Big(S^-, (S^l)''\Big)$$

When responses have certain functional forms, an unbiased estimator exists when an appropriate imbalance measure is zero. Note that, from the weak ignorability assumption, the response values Y_u^i are of the form $h^i(\mathbf{X}_u)+\epsilon_u^i$ where $h^i(\cdot)$ is a response function and ϵ_u^i is an error in the response value for unit $u \in \mathcal{N}$ with level $i \in \mathcal{L}$. The error term ϵ_u^i is assumed to have mean 0 and variance 1 for any unit u. Theorem 10 shows that $\widetilde{\tau}_{BOSS}(m, l)$ in (4.21) is unbiased for linear response values when the DOM imbalance measure under a multi-treatment setting in (4.22) is zero.

Theorem 10. *Under Assumptions 2 and 5, the BOSS estimator $\widetilde{\tau}_{BOSS}(m, l)$ is an unbiased estimator for $\mathbb{E}[\tau_{m,l}]$ between treatment level l and treatment level m, if the DOM imbalance measure, $\mathscr{I}_{DOM}$, is zero and the response functions are linear.*

In other words, if $\mathscr{I}_{DOM}\Big(S^l, (S^l)', (S^l)'', S^m, (S^m)', (S^m)''\Big) = 0$, then

$$\mathbb{E}[\widetilde{\tau}_{BOSS}(m, l)] = \mathbb{E}[\tau_{m,l}] \tag{4.26}$$

holds when level-l and level-m response functions are linear, namely,

$$h^i(\mathbf{X}_u) = \beta_i^T \mathbf{X}_u + \alpha_i = \sum_{k=1}^{K} \beta_{i,k} X_{u,k} + \alpha_i \tag{4.27}$$

for $i \in \{l, m\}$, $\beta_i = (\beta_{i,1}, \beta_{i,2}, \cdots, \beta_{i,K}) \in \mathbb{R}^K$, $\mathbf{X}_u = (X_{u,1}, X_{u,2}, \cdots, X_{u,K}) \in \mathbb{R}^K$, and $\alpha_i \in \mathbb{R}$.

Proof. It will be shown that $\mathbb{E}[\widetilde{\tau}_{BOSS}(m, l) - \tau_{m,l}] = 0$. (See (4.3) and (4.21).)

First, note that

$$
\begin{aligned}
&\widetilde{\tau}_{BOSS}(m,l) - \tau_{m,l} \\
&= \frac{1}{|\mathcal{N}|}\left\{\sum_{u\in(S^m)'} Y_u^m - \sum_{u\in S^l} Y_u^l + \sum_{u\in S^m} Y_u^m - \sum_{u\in(S^l)'} Y_u^l + \sum_{u\in(S^m)''} Y_u^m - \sum_{u\in(S^l)''} Y_u^l\right\} \\
&\quad - \frac{1}{|\mathcal{N}|}\left\{\sum_{i=0}^{L-1}\sum_{u\in S^i}\left(Y_u^m - Y_u^l\right)\right\} \\
&= \frac{1}{|\mathcal{N}|}\left\{\sum_{u\in(S^m)'} Y_u^m - \sum_{u\in S^l} Y_u^l + \sum_{u\in S^m} Y_u^m - \sum_{u\in(S^l)'} Y_u^l + \sum_{u\in(S^m)''} Y_u^m - \sum_{u\in(S^l)''} Y_u^l\right\} \\
&\quad - \frac{1}{|\mathcal{N}|}\left\{\sum_{u\in S^l}\left(Y_u^m - Y_u^l\right) + \sum_{u\in S^m}\left(Y_u^m - Y_u^l\right) + \sum_{u\in S^-}\left(Y_u^m - Y_u^l\right)\right\}. \\
&= \frac{1}{|\mathcal{N}|}\left\{\left(\sum_{u\in(S^m)'} Y_u^m - \sum_{u\in S^l} Y_u^m\right) - \left(\sum_{u\in(S^l)'} Y_u^l - \sum_{u\in S^m} Y_u^l\right)\right. \\
&\quad \left. + \left(\sum_{u\in(S^m)''} Y_u^m - \sum_{u\in S^-} Y_u^m\right) - \left(\sum_{u\in(S^l)''} Y_u^l - \sum_{u\in S^-} Y_u^l\right)\right\}.
\end{aligned}
\tag{4.28}
$$

By substituting $Y_u^i = h^i(\mathbf{X}_u) + \epsilon_u^i$ to (4.28), and taking expectation,

$$
\begin{aligned}
&\mathbb{E}[\widetilde{\tau}_{BOSS}(m,l) - \tau_{m,l}] \\
&= \frac{1}{|\mathcal{N}|}\left\{\left(\sum_{u\in(S^m)'} h^m(\mathbf{X}_u) - \sum_{u\in S^l} h^m(\mathbf{X}_u)\right) - \left(\sum_{u\in(S^l)'} h^l(\mathbf{X}_u) - \sum_{u\in S^m} h^l(\mathbf{X}_u)\right)\right. \\
&\quad \left. + \left(\sum_{u\in(S^m)''} h^m(\mathbf{X}_u) - \sum_{u\in S^-} h^m(\mathbf{X}_u)\right) - \left(\sum_{u\in(S^l)''} h^l(\mathbf{X}_u) - \sum_{u\in S^-} h^l(\mathbf{X}_u)\right)\right\}.
\end{aligned}
\tag{4.29}
$$

since the error terms have zero-mean.

By the linearity of the response functions stated in (4.27),

$$
\begin{aligned}
&\mathbb{E}[\widetilde{\tau}_{BOSS}(m,l) - \tau_{m,l}] \\
&= \frac{1}{|\mathcal{N}|}\left\{\beta_m^T\Big(\sum_{u\in(S^m)'} \mathbf{X}_u - \sum_{u\in S^l} \mathbf{X}_u\Big) - \beta_l^T\Big(\sum_{u\in(S^l)'} \mathbf{X}_u - \sum_{u\in S^m} \mathbf{X}_u\Big)\right. \\
&\quad \left. + \beta_m^T\Big(\sum_{u\in(S^m)''} \mathbf{X}_u - \sum_{u\in\mathcal{N}\setminus(S^l\cup S^m)} \mathbf{X}_u\Big) - \beta_l^T\Big(\sum_{u\in(S^l)''} \mathbf{X}_u - \sum_{u\in\mathcal{N}\setminus(S^l\cup S^m)} \mathbf{X}_u\Big)\right\}.
\end{aligned}
\tag{4.30}
$$

From (4.22) and the condition that $\mathscr{I}_{\mathrm{DOM}}\Big(S^l, (S^l)', (S^l)'', S^m, (S^m)', (S^m)''\Big) = 0$,

$$\mathscr{I}_{\mathrm{DOM}}\Big(S^l, (S^m)'\Big) = \sum_{k=1}^{K} \left| \frac{1}{|S^l|} \sum_{u \in S^l} X_{u,k} - \frac{1}{|(S^m)'|} \sum_{u \in (S^m)'} X_{u,k} \right| = 0, \tag{4.31}$$

$$\mathscr{I}_{\mathrm{DOM}}\Big(S^m, (S^l)'\Big) = \sum_{k=1}^{K} \left| \frac{1}{|S^m|} \sum_{u \in S^m} X_{u,k} - \frac{1}{|(S^l)'|} \sum_{u \in (S^l)'} X_{u,k} \right| = 0, \tag{4.32}$$

$$\mathscr{I}_{\mathrm{DOM}}\Big(S^-, (S^m)''\Big) = \sum_{k=1}^{K} \left| \frac{1}{|S^-|} \sum_{u \in S^-} X_{u,k} - \frac{1}{|(S^m)''|} \sum_{u \in (S^m)''} X_{u,k} \right| = 0, \tag{4.33}$$

and

$$\mathscr{I}_{\mathrm{DOM}}\Big(S^-, (S^l)''\Big) = \sum_{k=1}^{K} \left| \frac{1}{|S^-|} \sum_{u \in S^-} X_{u,k} - \frac{1}{|(S^l)''|} \sum_{u \in (S^l)''} X_{u,k} \right| = 0. \tag{4.34}$$

Moreover, as the multisets $(S^l)', (S^l)'', (S^m)'$, and $(S^m)''$ are chosen in a way such that $|S^l| = |(S^m)'|$, $|S^m| = |(S^l)|$, and $|S^-| = |(S^m)''| = |(S^l)''|$,

$$\sum_{u \in S^l} X_{u,k} - \sum_{u \in (S^m)'} X_{u,k} = 0 \quad \forall k \in \mathcal{P}, \tag{4.35}$$

$$\sum_{u \in S^m} X_{u,k} - \sum_{u \in (S^l)'} X_{u,k} = 0 \quad \forall k \in \mathcal{P}, \tag{4.36}$$

$$\sum_{u \in S^-} X_{u,k} - \sum_{u \in (S^m)''} X_{u,k} = 0 \quad \forall k \in \mathcal{P}, \tag{4.37}$$

and

$$\sum_{u \in S^-} X_{u,k} - \sum_{u \in (S^l)''} X_{u,k} = 0 \quad \forall k \in \mathcal{P}. \tag{4.38}$$

Substituting (4.35), (4.36), (4.37), and (4.38) into (4.30) leads to the desired result. □

Corollary. *Under Assumption 3, the BOSS estimator $\widetilde{\tau}_{BOSS}(m, l)$ is an unbiased estimator of $\tau(m, l)$ between treatment level l and treatment level m if the conditions in Theorem 10 are satisfied.*

Proof. From Theorem 9, $\mathbb{E}[\tau_{m,l}] = \tau(m, l)$. In addition, from Theorem 10,

$$\mathbb{E}[\widetilde{\tau}_{BOSS}(m, l)] - \mathbb{E}[\tau_{m,l}] = 0. \tag{4.39}$$

Hence $\mathbb{E}[\widetilde{\tau}_{BOSS}(m,l)] = \tau(m,l)$ holds. □

Note that these results can be extended to higher-order functional forms of the response functions provided that the appropriate imbalance measures are used. For example, the multi-level BOSS estimator for ATE will be unbiased if level-l and level-m response functions are of the form

$$h^i(\mathbf{X}_u) = \sum_{k\in\mathcal{P}} \beta_{i,k} X_{u,k} + \sum_{k\in\mathcal{P}} \gamma_{i,k}(X_{u,k})^2 + \sum_{(k_1,k_2)\in\binom{\mathcal{P}}{2}} \gamma_{i,k_1,k_2} X_{u,k_1} X_{u_1,k_2} + \alpha_i \text{ for } i \in \{l,m\} \tag{4.40}$$

and $\mathscr{I}_{\text{DOM2}}\Big(S^l, (S^l)', (S^l)'', S^m, (S^m)', (S^m)''\Big) = 0$.

4.6 Simulation Results

In this section, BOSS estimator $\widetilde{\tau}_{BOSS}(m,l)$ is computed using simulated data. Its bias and Root Mean Squared Error (RMSE) values are compared to those of matching estimators. For comparison, in Section 4.6.1, one of the simulations that Yang et al. (2016) conducted in their paper is reproduced first. In Section 4.6.2 additional simulation results under a different setting are presented.

4.6.1 First Experiment

In this reproduction of a simulation study that was proposed by Yang et al. (2016), there are seven covariates and three levels for each unit. Response functions are linear for all three levels $i = 0, 1,$ and 2, where $\beta_i = (\beta_{i,1}, \cdots, \beta_{i,K}) \in \mathbb{R}^K$ are given by $\beta_0 = (-1.5, 1, 1, 1, 1, 1, 1), \beta_1 = (-3, 2, 3, 1, 2, 2, 2)$, and $\beta_2 = (1.5, 3, 1, 2, -1, -1, -1)$ and $\alpha_i = 0 \in \mathbb{R}$ for all $i \in \{0, 1, 2\}$. The responses Y_u^i are of the form $h^i(\mathbf{X}_u) + \epsilon_u^j$ where ϵ_u^i is the normally distributed error term with mean 0 and variance 1.

Covariate values for each unit are constructed using a constant, multivariate normal distribution, uniform distribution, Chi-square distribution, and a Bernoulli random variable. The first covariate $X_{u,1}$ is equal to 1 for every unit u while $(X_{u,2}, X_{u,3}, X_{u,4})$ follows the multivariate normal distribution with the following

mean μ and covariance matrix Σ:

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} 2 & 1 & -1 \\ 1 & 1 & -0.5 \\ -1 & -0.5 & 1 \end{bmatrix} \tag{4.41}$$

In addition, $X_{u,5}$ is uniformly distributed in the interval $[-3, 3]$, $X_{u,6}$ follows a Chi-square distribution with 1 degree of freedom, $X_{u,7}$ is a Bernoulli random variable with $p = 0.5$.

Assignment of the treatment levels is done using a multinomial distribution with $p_i = p(i \mid \mathbf{X}_u)$ for $i \in \{0, 1, 2\}$ where $p(i \mid \mathbf{X}_u)$ denotes a generalized propensity score defined in (4.17) and is computed by

$$p(i \mid \mathbf{X}_u) = \frac{\exp(\gamma_i^T \mathbf{X}_u)}{\sum_{j=0}^{2} \exp(\gamma_j^T \mathbf{X}_u)} \tag{4.42}$$

with values of γ_0, γ_1, and γ_2 given in Yang et al. (2016):

$$\begin{aligned} \gamma_0 &= (0, 0, 0, 0, 0, 0, 0), \\ \gamma_1 &= (0, 0.7, 0.7, 0.7, -0.7, 0.7, 0.7), \\ \gamma_2 &= (0, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4). \end{aligned} \tag{4.43}$$

Sample sizes were set to $|S^0| = |S^1| = |S^2| = 500$.

One hundred datasets were generated using the method mentioned above. BOSS was applied to each dataset with 300 second time limit. This time limit was chosen because most objective values of the BOSS instances in this experiment with the DOM imbalance measure become lower than 0.1 within 300 seconds. Simulation results are reported in Table 4.1. In the table, MCOV denotes the nearest neighbor covariate matching with multiple treatment levels described in Section 4.4 and GPSM denotes a matching estimator under a multi-treatment level using a generalized propensity score. GPSS is an estimator under multi-treatment levels that is obtained using sub-classification on the generalized propensity score.

To calculate the MCOV, GPSM, and GPSS estimator values, an R code that is made available by Yang et al. (2016) is used. To compute the BOSS estimator, Python is used to do post-processing for multi-treatment levels using the control groups obtained by C++ program by Sauppe et al. (2014) for BOSS method with two treatment levels.

As a pre-processing step, auxiliary data is first constructed by combining each combination of two treatment levels. Then, for each pair, the DOM imbalance measure with two treatment levels is minimized. For each run, time limit was set as 300 seconds. For the simulation, a desktop equipped with a quad-core Intel i7-6700 at 3.40GHz is used.

True SATE values for each treatment level pairs for the simulated data are given by $\tau_{1,0} = -0.292$, $\tau_{2,0} = -0.397$, and $\tau_{2,1} = -0.104$. These values are computed using what each unit in a certain level would have gotten for its outcome from the data construction. It is different from the true PATE values which are given by $\tau(m, l) = 0$ for any $l, m \in \{0, 1, 2\}$ since Assumption 3 does not hold. Bias values which are computed by subtracting the true SATE values from estimated SATE are reported in Table 4.1. In the table, RMSE values are also reported.

Table 4.1: Comparison of Estimators (The First Experiment)

Estimation Method	Bias			RMSE		
	$\tau_{1,0}$	$\tau_{2,0}$	$\tau_{2,1}$	$\tau_{1,0}$	$\tau_{2,0}$	$\tau_{2,1}$
MCOV	0.220	0.191	-0.030	0.258	0.227	0.155
GPSM	0.173	0.032	-0.140	0.533	0.363	0.586
GPSS	0.174	0.003	-0.171	0.463	0.261	0.500
BOSS	<0.001	-0.014	-0.015	0.294	0.268	0.313
Random Selection	1.379	0.556	-0.823	1.428	0.593	0.904

Note that the BOSS estimator outperforms the other estimation methods in terms of the bias in all the cases except for the estimator of $\mathbb{E}[\tau_{2,0}]$ in GPSS. The BOSS estimator has similar or smaller RMSE value than GPSM and GPSS estimators while it has slightly larger RMSE than covariate matching. The covariate matching method shows the smallest RMSE value. However, it is known that the covariate matching estimator leads to a lower coverage rate when compared to GPSM and GPSS method (Yang et al., 2016). For the entire simulation of 100 datasets, BOSS took approximately 30000 seconds while the matching methods took 577 seconds. Estimates generated by randomly selecting the control groups instead of optimally selecting them showed the worst values in the size of the bias and RMSE.

4.6.2 Second Experirment

This second experiment was conducted using the same program written from the previous simulation with different input data under the same computing environment as before. The data generation process is described below.

In this setup for the generation of the input data, there are three treatment levels denoted by i and three covariates for each unit: $i = 0, 1,$ and 2 and $K = 3$. In the response functions $h^i(\mathbf{X}_u)$ for $i \in \{0, 1, 2\}$ used in the simulation, $\beta_i = (\beta_{i,1}, \cdots, \beta_{i,K}) \in \mathbb{R}^K$ are given by $\beta_0 = (1, 2, 3), \beta_1 = (4, 5, 6)$ and $\beta_2 = (7, 8, 9)$ and $\alpha_i = 0 \in \mathbb{R}$ for $i \in \{0, 1, 2\}$. As before, the responses Y_u^i are of the form $h^i(\mathbf{X}_u) + \epsilon_u^i$ where the error term ϵ_u^i follows a normal distribution with mean 0 and variance 1.

The covariates are constructed using a normal distribution. $X_{u,1}$, $X_{u,2}$, $X_{u,3}$ are independent random variables that follows a normal distribution with mean 0 and variance 5, 2, and 7 respectively. That is, the covariate vector $(X_{u,1}, X_{u,2}, X_{u,3})$ follows the multivariate normal distribution with the following mean μ and covariance matrix Σ:

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 7 \end{bmatrix} \tag{4.44}$$

Sample sizes were set to $|S^0| = 400$, $|S^1| = 600$, $|S^2| = 500$. Assignment of the treatment levels is done using a multinomial distribution with $p_0 = 4/15$, $p_1 = 6/15$, and $p_2 = 5/15 = 1/3$ so that Assumption 2 holds in this case.

Hence, for each treatment level, the expected value of the true SATE values is equal to 0 which is the true PATE value. For the simulation, 1000 datasets are generated using the method described above and the time limit for each BOSS simulation with DOM imbalance measure was set to 30 seconds as the objective value falls below 0.02 within 30 seconds.

The true SATE values obtained in the generated data were

$$\tau_{1,0} = -0.003, \tau_{2,0} = -0.010, \text{ and } \tau_{2,1} = -0.006 \tag{4.45}$$

Bias and RMSE values from this experiment are reported in Table 4.2.

In this second experiment, BOSS with DOM imbalance measure performed similarly compared to that of matching estimators. Note that the estimates by BOSS can be improved by imposing an imbalance measure that are ranked higher with a cost of longer computational time. As seen in the first experiment, both

Table 4.2: Comparison of Estimators (The Second Experiment)

Estimation Method	Bias			RMSE		
	$\tau_{1,0}$	$\tau_{2,0}$	$\tau_{2,1}$	$\tau_{1,0}$	$\tau_{2,0}$	$\tau_{2,1}$
MCOV	-0.002	0.006	0.008	0.125	0.186	0.204
GPSM	-0.013	-0.004	0.009	0.483	0.796	0.864
GPSS	-0.013	>-0.001	0.012	0.140	0.210	0.220
BOSS	-0.016	-0.008	0.007	0.395	0.416	0.363
Random Selection	1.379	0.556	-0.823	1.428	0.593	0.904

matching and BOSS performed better than the random selection in terms of bias and RMSE value sizes.

4.7 Concluding Remarks and Future Research Direction

In this chapter, the BOSS framework was extended to a multi-treatment setting. The simulation results in this chapter demonstrate that the BOSS estimator provides comparable results when compared to matching estimators when considering the size of the bias. Estimation using BOSS is an alternative way to get an unbiased estimate of the expected value of the SATE which can be used under a multi-treatment setting. When using the real-world data where the true vales of the estimates are not known, computing the BOSS estimates in addition to matching estimates will allow to have more reliable results.

When applying the BOSS estimator to a dataset having multiple treatments, checking that the optimal objective value is small enough is necessary since the Theorem 10 of unbiasedness holds only when the BOSS instances are fully optimized to zero objective value. If there does not exist close enough samples between the treatment level groups, the optimal value 0 may not be achieved. Hence, one should be aware when dealing with datasets with small and dissimilar treatment level groups.

A possible future research direction is extending this BOSS method so that it can be applicable to estimating the average treatment effect of a dataset having a continuous treatment. One way of applying the multi-treatment BOSS to continuous treatment setting is by dividing the continuous treatments into multiple treatments based on their treatment values. For example, when the continuous

treatment is years of education, units can be divided into multiple levels based on the discrete years that are rounded to the nearest integers or their education level (e.g., elementary school, middle school, high school, college) depending on what is needed in the analysis.

The proposed BOSS estimator has a broad application area. It will be applicable to any study involving the assessment of comparative effectiveness of many different types of treatments such as a various medicine or policy effectiveness test as BOSS estimator can now be applied to a data having multiple treatments.

CHAPTER 5

HANDLING MISSING DATA IN OBSERVATIONAL STUDIES WITH BALANCE OPTIMIZATION SUBSET SELECTION

5.1 Introduction

In observational studies, researchers often encounter incomplete data which contains missing values. Missing data can occur from non-response of respondents or loss of data in the data collection or storage process. In longitudinal studies, failure to collect responses from the same set of participants because of unavailability of some participants after a certain amount of time can also result in missing data (Graham, 2012). This chapter considers two possible cases of missing data: [Case 1] some covariates are not included at all and [Case 2] all covariates that affect the responses are included but some entries are missing.

In Case 1, some covariate vectors that affect responses for the units are not observed while all the other covariates are fully observed. Such a situation may occur from a failure to record all the covariates. The conditional independence assumption in strong ignorability does not hold since the assumption states that, given the observed covariates, response values are conditionally independent of whether the units are treated or not. Sensitivity analysis on violations of the conditional independence assumption is necessary for all estimators based on strong ignorability. Suppose instead that the conditional independence assumption holds once a previously unobserved covariate is included. The sensitivity of the Balance Optimization Subset Selection (BOSS) estimators is examined by simulating such binary unobserved confounders. This result is then compared to that of matching estimators on two datasets – the LaLonde (1986) data and the Pri.DE (Stampf, 2014) data. For the LaLonde data, it will be shown that BOSS estimators are less sensitive to the addition of a covariate and the BOSS estimators have much more stable variances compared to matching estimators. With the Pri.DE data, comparable sensitivity results are obtained for BOSS and matching.

In Case 2, datasets with all covariates known but with missing entries for some

units are considered. Imputation of missing values is a method that is frequently used to apply causal inference methods such as BOSS or matching, whose theories were developed with complete data in mind. In particular, multiple imputation generates multiple complete datasets by imputing the missing values using the observed values. With these datasets after multiple imputation, two approaches of applying BOSS - Within and Across - are considered as studied by Mitra and Reiter (2016) for matching. The Within approach applies inference methods to each dataset first and then combines the estimates afterwards. The Across approach first aggregates information from the multiple datasets and then applies the method to a single dataset with aggregated information. Through an example, it is illustrated that the method using the Across approach has smaller bias than the method using the Within approach in BOSS. The better performance of the Across approach over the Within approach is consistent with results by Mitra and Reiter (2016) for matching. Furthermore, in the example, the BOSS method with each approach gives smaller bias compared to the corresponding matching method.

This chapter is organized as follows. The first type of missing data considered is Case 1. Section 5.2 discusses sensitivity analysis that can be done by simulating a missing covariate in the first case of missing data. Section 5.2.1 reviews the concept of strong ignorability. Section 5.2.2 summarizes the sensitivity analysis method of matching estimators using simulation proposed by Ichino et al. (2008) and implemented by Nannicini (2007) in Stata. Section 5.2.3 introduces how to compute standard errors for each estimator and applies the method of sensitivity analysis to BOSS estimators. Sensitivity analysis results for matching and BOSS estimators are discussed in Section 5.2.3 (with the LaLonde data) and Section 5.2.3 (with the Pri.DE data). Section 5.3 focuses on Case 2 with missing entries in some covariate vectors where all the covariate vectors are included in the dataset. Section 5.3.1 discusses the Across and the Within approaches in BOSS on multiply-imputed data and reviews those approaches in propensity score matching. Section 5.3.2 describes the data generation method for an experiment with these approaches and Section 5.3.2 presents simulation results. Section 5.4 provides concluding comments and future research directions.

5.2 Simulating an Unobserved Covariate

When evaluating estimation methods, researchers need to check whether the underlying assumptions hold. As stated in the standard matching literature (e.g., Rosenbaum and Rubin (1983b)), the validity of matching estimators depends on a *strong ignorability* assumption. Similarly, as Nikolaev et al. (2013) showed, the validity of Balance Optimization Subset Selection (BOSS) estimators for causal inference relies on the same assumption.

As stated in Chapter 1, the strong ignorability assumption is composed of two parts – *conditional independence* and *common overlap* (Sekhon, 2009). The conditional independence assumption, also known as *unconfoundedness*, states that, given the measured covariate values, the treatment assignment is conditionally independent with the distribution of the potential outcomes. This assumption means that all the covariates affecting the response values are measured or controlled in order not to affect the treatment effect estimates and thus there is no confounding factor that is uncontrolled.

In practice, the strong ignorability assumption may not always hold. The first part of this chapter focuses on a sensitivity analysis of how the estimator value changes when deviating from the conditional independence assumption. There have been several parametric approaches for sensitivity analyses of matching estimators (Rosenbaum and Rubin, 1983a; Imbens, 2003; Brumback et al., 2004; Greenland, 2005; Li et al., 2011). In addition, Ichino et al. (2008) and Nannicini (2007) provide simulation-based, non-parametric, sensitivity analysis approaches.

Cosistent with what Nannicini (2007) has done on matching estimators, a sensitivity analysis of BOSS estimators will be conducted by simulating possible unmeasured covariates. By comparing the sensitivity of BOSS estimators to that of matching, how the estimators produced by matching and BOSS are affected when deviating from the conditional independence assumption will be examined.

5.2.1 Strong Ignorability Assumption

Consider a binary treatment setting again. Suppose that Z is a treatment indicator which is equal to 1(0) if (not) treated. Let Y^1 and Y^0 be treated and untreated outcome values and let $\mathbf{X}$ (with support $\mathcal{X}$) be a vector of K covariates for a unit

that is selected uniformly at random. Then the strong ignorability assumption is

$$Y^1, Y^0 \perp\!\!\!\perp Z \mid \mathbf{X} \tag{5.1a}$$

and

$$0 < P(Z = 1|\mathbf{X} = x) < 1 \text{ for all } x \in \mathcal{X}. \tag{5.1b}$$

By definition, $Y^1 \perp\!\!\!\perp Z \mid \mathbf{X}$ $\left(Y^0 \perp\!\!\!\perp Z \mid \mathbf{X}\right)$ implies that the distribution of Y^1 (Y^0), given the covariate values $\mathbf{X}$, is the same regardless of the treatment indicator Z. Hence,

$$\mathbb{E}\left[Y^1 \mid \mathbf{X}, Z = 0\right] = \mathbb{E}\left[Y^1 \mid \mathbf{X}, Z = 1\right] = \mathbb{E}\left[Y^1 \mid \mathbf{X}\right] \tag{5.2a}$$

and

$$\mathbb{E}\left[Y^0 \mid \mathbf{X}, Z = 0\right] = \mathbb{E}\left[Y^0 \mid \mathbf{X}, Z = 1\right] = \mathbb{E}\left[Y^0 \mid \mathbf{X}\right]. \tag{5.2b}$$

under this assumption.

Recall that the treated and untreated outcome values of a specific unit u were denoted by Y_u^1 and Y_u^0, respectively. As done in Chapters 1 to 3, the value of interest is the sample average treatment effect for the treated (SATT) defined in (1.3).

Propensity score matching (Rosenbaum and Rubin, 1983b) and Balance Optimization Subset Selection (Nikolaev et al., 2013, BOSS) are two methods used to estimate SATT. Let $e(\mathbf{X}_u)$ denote the propensity score of unit u under a binary treatment setting, the probability that unit u receives treatment given its covariate values. Note that $e(\mathbf{X}_u)$ can be expressed as

$$e(\mathbf{X}_u) = p(1|\mathbf{X}_u) \tag{5.3}$$

using the notation $p(j|\mathbf{X}_u)$ from previous chapter, which is the generalized propensity score for treatment level j, since the "treatment" in a binary treatment setting corresponds to a treatment level 1. Rosenbaum and Rubin (1983b) show that, under the strong ignorability assumption, the following relationship of conditional independence given the propensity score values holds:

$$Y^1, Y^0 \perp\!\!\!\perp Z \mid e(\mathbf{X}). \tag{5.4}$$

By matching each treated unit $t \in T$ with a control unit $c \in C$ that has the same propensity score, an unbiased estimate of τ_T^1 can be obtained under the strong

ignorability assumption. If there does not exist an exact matching for each $t \in T$, then *inexact matching* is used. The inexact matching method considered in this chapter is *Nearest-Neighbor Matching* (Rubin, 1973), which uses the following propensity score matching estimator for τ_T^1,

$$\hat{\tau}_T^1 = \frac{1}{|T|} \sum_{t \in T} \left(Y_t^1 - \sum_{c \in C(t)} \frac{1}{|C(t)|} Y_c^0 \right), \tag{5.5}$$

where $C(t) \equiv \arg\min_{c \in C} |e(\mathbf{X}_t) - e(\mathbf{X}_c)|$.

BOSS estimates τ_T^1 using the estimator given in (1.4). It has been shown that $\widetilde{\tau}_T^1(C')$ is an unbiased estimator of τ_T^1 under the strong ignorability assumption when appropriate covariate balance is achieved given some functional form of the response function (Sauppe and Jacobson, 2017). BOSS can incorporate various imbalance measures as studied in Chapter 2. For the simulations reported in Sections 5.2 and 5.3, the DOM imbalance measure under a binary treatment setting given in (2.9) is used. As stated in Theorem 5, BOSS yields an unbiased estimator of SATT if the response functions are linear and the DOM imbalance measure is minimized to zero.

5.2.2 Sensitivity Analysis of Matching Estimators using Simulation

This section summarizes the simulation-based sensitivity analysis method proposed by Ichino et al. (2008) and Nannicini (2007). Consider the scenario that the assumption stated in (5.1a) does not hold because there exists a missing covariate (for all units in the dataset). Suppose that the conditional independence assumption holds given the observed covariate values $\mathbf{X}$ and an additional (previously unobserved) binary variable U. Suppose that the Assumption 6 holds hereafter.

Assumption 6. Treatment assignments are conditionally independent with the outcome values, given the new set of covariates including both the observed covariates $\mathbf{X}$ and the unobserved binary variable U:

$$Y^1, Y^0 \perp\!\!\!\perp Z \,\Big|\, (\mathbf{X}, U). \tag{5.6}$$

Note that (5.2a) and (5.2b) does not hold. Instead,

$$\mathbb{E}\left[Y^1 \mid \mathbf{X}, U, Z = 0\right] = \mathbb{E}\left[Y^1 \mid \mathbf{X}, U, Z = 1\right] \tag{5.7a}$$

and

$$\mathbb{E}\left[Y^0 \mid \mathbf{X}, U, Z = 0\right] = \mathbb{E}\left[Y^0 \mid \mathbf{X}, U, Z = 1\right]. \tag{5.7b}$$

hold under Assumption 6.

Define B to be a binary outcome variable. Define the following four probabilities

$$p_{ij} = \text{Prob}\{U = 1 \mid Z = i, B = j\} \text{ for } i, j \in \{0, 1\}. \tag{5.8}$$

Either 0 or 1 is assigned to a binary variable U using the p_{ij} values. Then, in addition to existing covariates which were observed, a set $\mathscr{U}$ composed of the new binary values U is attached to the given data to form an augmented set of covariates. Note that, when generating one covariate vector for all the units with $\mathscr{U}$, the number of elements in the set $\mathscr{U}$ is equal to the number of observed units (i.e., $|\mathscr{U}| = \|$). Then the standard Matching and BOSS methods can applied to this augmented dataset since the strong ignorability assumption holds with this new (augmented) dataset.

Sensitivity analysis of a propensity score matching estimator is conducted using the Stata code (Nannicini, 2007), by simulating many different $\mathscr{U}$ with different sets of p_{ij} values. In Section 5.2.3, the sensitivity of matching and BOSS estimators are examined with the National Supported Work Demonstration (NSW) program data (LaLonde, 1986). In Section 5.2.3, wthe sensitivity of the estimators are examined with the Pediatric Respiratory Infection in Deutschland (Pri.DE) data from Stampf (2014).

The Stata code (Nannicini, 2007) uses the propensity score matching algorithm `attnd` (Becker and Ichino, 2002). The simulation results obtained using matching are presented in Sections 5.2.3 and 5.2.3 and these results are compared to the BOSS results. In addition to replicating the three scenarios considered in Nannicini (2007), additional 36 scenarios are simulatd using p_{ij} values listed in Table 5.4 and Figure 5.1 for the LaLonde data following the method in Ichino et al. (2008). In a similar manner, additional simulations of 34 scenarios were conducted with various p_{ij} values reported in Table 5.7 and Figure 5.2 for the Pri.DE data.

The p_{ij} values in Table 5.4 and Table 5.7 are obtained by varying $d = p_{01} - p_{00}$

and $s = p_{1\cdot} - p_{0\cdot}$ where

$$\begin{aligned} p_{i\cdot} &= \text{Prob}(U = 1 \mid Z = i) \\ &= \frac{\text{Prob}(U = 1, Z = i)}{\text{Prob}(Z = i)} \\ &= \frac{\text{Prob}(U = 1, Z = i, B = 0)}{\text{Prob}(Z = i)} + \frac{\text{Prob}(U = 1, Z = i, B = 1)}{\text{Prob}(Z = i)} \\ &= \text{Prob}\{U = 1 \mid Z = i, B = 0\} \cdot \text{Prob}(B = 0 \mid Z = i) \\ &\quad + \text{Prob}\{U = 1 \mid Z = i, B = 1\} \cdot \text{Prob}(B = 1 \mid Z = i) \\ &= p_{i0} \cdot \text{Prob}(B = 0 \mid Z = i) + p_{i1} \cdot \text{Prob}(B = 1 \mid Z = i) \end{aligned} \tag{5.9}$$

for $i \in \{0, 1\}$. To generate the values of p_{ij} in the Table 5.4 and Table 5.7, it was assumed that $p_{11} - p_{00} = 0$ and $\text{Prob}(U = 1) = 0.4$ (Nannicini, 2007). Fixing α and β results in the following system of four equations with four variables p_{ij}, $i, j \in \{0, 1\}$:

$$0 = p_{11} - p_{00} \tag{5.10}$$

$$0.4 = \text{Prob}(U = 1) = \sum_{i=0}^{1} \sum_{j=0}^{1} p_{ij} \cdot \text{Prob}(B = j \mid Z = i) \cdot \text{Prob}(Z = i) \tag{5.11}$$

$$\alpha = d = p_{01} - p_{00} \tag{5.12}$$

$$\beta = s = p_{1\cdot} - p_{0\cdot} = \sum_{j=0}^{1} p_{1j} \cdot \text{Prob}(B = j \mid Z = 1) - \sum_{j=0}^{1} p_{0j} \cdot \text{Prob}(B = j \mid Z = 0). \tag{5.13}$$

From (5.10) to (5.13), the values of p_{11}, p_{10}, p_{01}, and p_{00} are determined and used to simulate the binary unobserved confounder.

5.2.3 Sensitivity Analysis and Comparison of Estimators Using Simulation

Sensitivity analysis of BOSS estimators is conducted using the same approach used for sensitivity analysis of matching estimators discussed in Section 5.2.2. The hidden/unmeasured covariate were generated to see how this addition affects the treatment effect value (the estimator of the SATT) and its standard error.

Let $\hat{\tau}^1_{T,k}$ and $\widetilde{\tau}^1_{T,k}(C'_k)$, respectively, denotes the matching estimator and the BOSS estimator obtained from the k-th estimation out of M repetitions. These estimates

are computed by

$$\hat{\tau}^1_{T,k} = \frac{1}{|T_k|} \sum_{t \in T_k} \left(Y^1_t - \sum_{c \in C_k(t)} \frac{1}{|C_k(t)|} Y^0_c \right), \tag{5.14}$$

and

$$\widetilde{\tau}^1_{T,k}(C'_k) = \frac{1}{|T_k|} \sum_{t \in T_k} Y^1_t - \frac{1}{|C'_k|} \sum_{c \in C'_k} Y^0_c \tag{5.15}$$

where $C_k(t) \equiv \arg\min_{c \in C_k} |e(\mathbf{X}_t) - e(\mathbf{X}_c)|$ and $C'_k \subset C_k$ is the control group that minimizes an imbalance $\mathscr{I}_{\text{DOM}}(T_k, C'_k)$ between the treatment group T_k and the control group C'_k.

Note that units composing the k-th treatment group T_k and the k'-th treatment group $T_{k'}$ ($k \neq k'$) are the same but the augmented set of covariates possessed by the units are different. Similarly, units composing the k-th control pool C_k and the k'-th control pool $C_{k'}$ ($k \neq k'$) are the same but the augmented set of covariates are different. However, note that, in matching, $C_k(t) \neq C_{k'}(t)$ thus $\cup_{t \in T_k} C_k(t) \neq \cup_{t \in T_{k'}} C_{k'}(t)$ and, in BOSS, $C'_k \neq C'_{k'}$ since the fact that the augmented set of covariates for the control pool and the treated group are different for k and k' ($k \neq k'$) results in different control groups for k and k' ($k \neq k'$). At each time the augmented data is different because of randomness in the data generation process for the unobserved confounder U although the same data generation method is used with the same p_{ij} values for a given scenario. Hence, the estimator values for the k-th estimation is different from the k'-th estimation for $k \neq k'$.

In the Stata program `sensatt`, the matching estimator of SATT is found using the average estimated values across several repetitions (say, M). Likewise, the average value of the BOSS estimators are found from the M repetitions for each scenario:

$$\overline{\hat{\tau}^1_T} = \frac{1}{M} \sum_{k=1}^{M} \hat{\tau}^1_{T,k} \tag{5.16}$$

and

$$\overline{\widetilde{\tau}^1_T(C')} = \frac{1}{M} \sum_{k=1}^{M} \widetilde{\tau}^1_{T,k}(C'_k). \tag{5.17}$$

For BOSS estimation, Python is used to generate $\mathscr{U}$ to construct an augmented dataset together with the original set of covariates.

For the simulation results reported in Section 5.2.3 and 5.2.3, the number of repetitions in each scenario was set to $M = 100$ for both matching and BOSS. For the simulation of BOSS, each individual run was given a time limit of 300 seconds

since most of the decreases in objective values occurred within 300 seconds. All the simulations were done using a quad-core Windows desktop with an Intel Core i7-6700 CPU at 3.40GHz.

Additionally, to compute standard errors for BOSS estimators, the method which used to compute standard errors for matching estimators in the `sensatt` package was adopted as follows. The notion of *within-imputation variance* se_W^2 and *between-imputation variance* se_B^2 (Nannicini, 2007) for matching methods are given by

$$\text{se}_W^2 = \frac{1}{M}\sum_{k=1}^{M} \text{se}_k^2, \tag{5.18}$$

and

$$\text{se}_B^2 = \frac{1}{M-1}\sum_{k=1}^{M}\left(\hat{\tau}_{T,k}^1 - \overline{\hat{\tau}_T^1}\right)^2, \tag{5.19}$$

where se_k^2 is the variance estimate for the k-th matching estimator $\hat{\tau}_{T,k}^1$ in (5.14). For BOSS, (5.18) was also used for se_W^2 and the following in (5.20) is used for se_B^2

$$\text{se}_B^2 = \frac{1}{M-1}\sum_{k=1}^{M}\left(\widetilde{\tau}_{T,k}^1(C'_k) - \overline{\widetilde{\tau}_T^1(C')}\right)^2, \tag{5.20}$$

where se_k^2 is the variance estimate for the k-th BOSS estimator $\widetilde{\tau}_{T,k}^1(C')$. The variance estimate se_k^2 for the k-th BOSS estimator $\widetilde{\tau}_{T,k}^1(C')$ in (5.15) is computed by

$$\text{se}_k^2 = \frac{1}{|T|^2}\cdot|T|\cdot\text{Var}_{t\in T}(Y_t^1) + \frac{1}{|C'_k|^2}\cdot|C'_k|\cdot\text{Var}_{c\in C'_k}(Y_c^0) = \frac{\text{Var}_{t\in T}(Y_t^1)}{|T|} + \frac{\text{Var}_{c\in C'_k}(Y_c^0)}{|C'_k|} \tag{5.21}$$

while assuming independent outcomes over units (Becker and Ichino, 2002). Using the within-imputation variance se_W^2 and the between-imputation variance se_B^2, the total variances se_T^2 for the matching estimator $\overline{\hat{\tau}_T^1}$ and the BOSS estimator $\overline{\widetilde{\tau}_T^1(C')}$ are computed with

$$\text{se}_T^2 = \text{se}_W^2 + \left(\frac{M+1}{M}\right)\text{se}_B^2. \tag{5.22}$$

Simulation Result – LaLonde Data

This section discusses a sensitivity analysis of matching and BOSS estimators for the LaLonde (1986) data. In the LaLonde dataset, there is a set of eight observed

covariates and one outcome variable. As described in earliery chapter, real earnings in year 1978, $RE78$, is the outcome (response) variable used in the dataset and denoted as Y. Furthermore, it is assumed that there is a binary, unmeasured confounder U that should have been included in the set of covariates, as it satisfies the Assumption 6 when included in the set of covariates. The three possible scenarios that were considered in Nannicini (2007) was followed as a first step of the analysis.

To generate the binary covariate vector $\mathscr{U}$, take the following four probabilities into account: p_{11}, p_{10}, p_{01}, and p_{00} defined in (5.8) after transforming the continuous outcome variable Y to a binary outcome variable B using the indicator function:

$$B_u = \mathbb{1}\{Y_u > \overline{Y}\}, \tag{5.23}$$

where $\mathcal{N}$ is the set of all units and $\overline{Y} = \sum_{u\in\mathcal{N}} Y_u/|\mathcal{N}|$ is the sample mean of the outcome values Y.

Note that in the LaLonde data

$$\text{Prob}(B = 0 \mid Z = 1) = \frac{176}{185}, \qquad \text{Prob}(B = 1 \mid Z = 1) = \frac{9}{185},$$
$$\text{Prob}(B = 0 \mid Z = 0) = \frac{1198}{2490}, \qquad \text{Prob}(B = 1 \mid Z = 0) = \frac{1292}{2490}.$$

The values of the original propensity score matching estimator and BOSS estimator before including any additional binary confounder are as follows. The estimated SATT value using matching was 2126 with a standard error of 1542 and that of BOSS was 1117 with standard error 784. These values are rounded to the nearest integers.

The three scenarios examined are described below. For the analysis in this chapter, the following covariates are used: [1] age in years, [2] education in years, [3] whether the individual is black as a binary variable, [4] whether he/she is Hispanic as a binary variable, [5] whether the individual is married as a binary variable, [6] whether the individual has no high school degree as a binary variable, [7] earnings in 1974, and [8] earnings in 1975. Note that, in Nannicini (2007), they used age, age2(squared age), educ, educ2(squared education years), marriage, black, Hispanic, RE74, RE75, RE742(squared 1974 earnings) RE752(squared 1975 earnings) as their set of existing covariates.

In the first scenario, the new variable U was set so that it has the same p_{ij} values obtained from the unemployment rate in year 1974, namely U74. The four

probabilities p_{11}, p_{10}, p_{01}, and p_{00} from the dataset are

$$\begin{aligned} p_{11} &= \text{Prob}\{U74 = 1 \mid Z = 1, B = 1\} = 0.78, \\ p_{10} &= \text{Prob}\{U74 = 1 \mid Z = 1, B = 0\} = 0.70, \\ p_{01} &= \text{Prob}\{U74 = 1 \mid Z = 0, B = 1\} = 0.02, \\ p_{00} &= \text{Prob}\{U74 = 1 \mid Z = 0, B = 0\} = 0.15. \end{aligned} \tag{5.24}$$

In the second scenario, the following p_{ij} values were considered:

$$p_{11} = 0.8,\ p_{10} = 0.8,\ p_{01} = 0.6,\ p_{00} = 0.3 \tag{5.25}$$

In the third scenario, the following values were used to simulate the unmeasured covariate U:

$$p_{11} = 0.8,\ p_{10} = 0.8,\ p_{01} = 0.6,\ p_{00} = 0.1 \tag{5.26}$$

Note that the only difference in generating the unmeasured covariate values in Scenarios 2 and 3 is the value of p_{00}, while all the other probabilities are the same. Under the three scenarios, the estimated value for SATT has been changed to those that are reported in Table 5.1. The values reported in Table 5.1 are rounded to the nearest integers. Note that the resulting values are different from those reported in Nannicini (2007) for the corresponding scenarios because a different set of covariates is used.

Table 5.1: Simulation Results under Three Scenarios

	Scenario 1		Scenario 2		Scenario 3	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
Matching	1175	3103	996	2746	-1408	6306
BOSS	1065	811	935	840	929	836

Now the same dataset is tested under 36 different scenarios with p_{ij} $(i, j \in \{0, 1\})$ reported in Table 5.4 obtained by varying the values of d and s. Again in Tables 5.2 and 5.3, the values are rounded. For both matching and BOSS estimators, increasing both d and s made the added (previously unobserved) binary variable work as a confounder reducing the estimated values. In the BOSS estimator table, there were a few deviations from this tendency. It might have occurred since the computationally optimization problem with a given time limit did not return fully optimized solutions because of the time limit for each run while the

added covariate with high d and s indeed worked as a variable reducing the estimates compared to the one with smaller values for d and s.

The results in Tables 5.2 and 5.3 are different. Matching estimators have shown a decrease in their estimator values along with an increase in their variance. On the other hand, in the case of BOSS estimators, stable variances are observed across various values of d and s. Furthermore, the variances of the BOSS estimators are smaller than those of the matching estimators.

Table 5.2: Matching Estimators (and Corresponding Standard Errors)

	$s = 0.1$	$s = 0.2.$	$s = 0.3$	$s = 0.4$	$s = 0.5$	$s = 0.6$
$d = 0.1$	1403 (1859)	1392 (1916)	1287 (2216)	1201 (2591)	1162 (2705)	1051 (3283)
$d = 0.2$	1447 (1840)	1269 (2147)	1270 (2409)	1116 (2580)	934 (3400)	879 (3723)
$d = 0.3$	1284 (2224)	933 (2681)	1010 (2654)	787 (3453)	655 (3798)	344 (4496)
$d = 0.4$	1228 (2319)	880 (2979)	884 (3000)	529 (4181)	−323 (5480)	−182 (5739)
$d = 0.5$	481 (3338)	661 (3205)	−327 (4712)	−300 (4963)	−616 (5457)	−888 (6988)
$d = 0.6$	499 (3904)	−1241 (5409)	−1087 (5965)	−2810 (7833)	−3515 (9938)	−3373 (11264)

Table 5.3: BOSS Estimators (and Corresponding Standard Errors)

	$s = 0.1$	$s = 0.2$	$s = 0.3$	$s = 0.4$	$s = 0.5$	$s = 0.6$
$d = 0.1$	979 (834)	998 (825)	972 (832)	960 (836)	988 (830)	975 (831)
$d = 0.2$	972 (833)	939 (837)	987 (830)	943 (837)	976 (832)	986 (831)
$d = 0.3$	929 (845)	959 (836)	933 (840)	918 (843)	944 (837)	953 (834)
$d = 0.4$	910 (842)	909 (843)	903 (840)	936 (838)	897 (843)	907 (844)
$d = 0.5$	890 (843)	880 (845)	903 (845)	894 (845)	884 (841)	878 (842)
$d = 0.6$	901 (842)	847 (846)	878 (842)	855 (844)	830 (847)	866 (845)

Simulation Result – Pri.DE Data

In this section, a sensitivity analysis of matching and BOSS estimators was conducted using Pri.DE data (Stampf, 2014). Pri.DE data is used to investigate whether or not current respiratory syncytial virus (RSV) infection in infants and young children causes severe lower respiratory tract infections (LRTI).

In this dataset, in addition to the information on RSV infection and severe LRTI (respectively, the treatment T and the outcome B) as binary variables, there are 13 covariates: gender, ethnic group, preterm delivery, former RSV infection, congenital heart defect, region in Germany, age, breast feeding, siblings, passive exposure to smoking at home, external care, parental atopy, and number of diagnosed

LRTI. Among the 3078 infants and young children considered, 1031 of them had a current RSV infection and 2047 of them did not.

From the data, the following probabilities can be computed:

$$\text{Prob}(B = 0 \mid Z = 1) = \frac{343}{1031}, \qquad \text{Prob}(B = 1 \mid Z = 1) = \frac{688}{1031},$$
$$\text{Prob}(B = 0 \mid Z = 0) = \frac{932}{2047}, \qquad \text{Prob}(B = 1 \mid Z = 0) = \frac{1115}{2047}.$$

By using the above values together with various d and s in the system of four equations in (5.10) to (5.13), the values for the four unknowns p_{11}, p_{10}, p_{01}, and p_{00}, can also be computed and these values are reported in Table 5.7.

In the case of matching, the default estimator values for SATT before adding any unobserved confounder is 0.092 with standard error 0.027. In the case of BOSS, the default estimated value is 0.179 with standard error 0.021. By varying d and s, the sensitivity analysis results as reported in Tables 5.5 and 5.6 are obtained for matching and BOSS, respectively .

Unlike the previous case with the LaLonde dataset, the sensitivities to the violation of the conditional independence assumption were comparable in the matching and BOSS estimators for the Pri.DE data. The sizes of the standard errors for both matching and BOSS estimators were similar. In most simulations for BOSS with this dataset, the optimal solutions were found within a few seconds which is much earlier than the time limit, 300 seconds. Specifically, 99.1% of the 3400 simulations were finished within 3 seconds and 99.6% of them were finished within 4 seconds.

5.3 Missing Values in Covariates and Multiple Imputation

In this section, two methods that are applicable to multiply-imputed data sets are compared. The multiple imputation is used to impute values where there are missing entries in the data while all the relevant covariate vectors are observed (This is Case 2 mentioned in the introduction of this chapter).

There are several methods to handle the missing entries. Appropriate methods should be used depending on the missing data mechanism. According to Little and Rubin (2002), the missing data mechanisms can be classified into three categories

based on how the missing values are related to the other data values. If missing values are independent of both observed and unobserved values, then the missing values are *missing completely at random* (MCAR). *Missing at random* (MAR) is when the missing data mechanism does not depend on unobserved values given the observed data while *missing not at random* (MNAR) is when the missing data mechanism does depend on unobserved values and thus the missing values cannot be accounted by only using observed values.

A possible method one can consider when handling missing data is to abandon all the units which have missing values. However, if the missing data mechanism is not MCAR, then this approach will lead to bias. When the number of units that are ignored is large, other types of information loss such as a reduction in precision may occur. Another method for handling the missing data is *imputation*(i.e, filling in missing values). There are many types of imputation methods – mean imputation using the average values, regression imputation using predicted values through regression, and hot deck imputation using other unit's data with similar values (Gelman and Hill, 2006).

In this section, a specific multiple imputation method, which uses Bayesian linear regression to fill in missing data values, is used. Recently, Mitra and Reiter (2016) compared two different propensity score matching methods for causal inference with datasets after multiple imputation – namely, the Across approach and the Within approach. Similar approaches are developed in the second half of Chapter 5 using a non-matching technique called BOSS and the BOSS estimates' performance measures on simulated datasets are provided. Two different BOSS methods will be applied to multiply-imputed datasets, and the performance measure of estimates from the two methods will be compared.

5.3.1 Within Approach and Across Approach

Multiple imputation generates multiple complete datasets by imputing missing values. Using these datasets, two ways to apply the BOSS method - Within approach and Across approach - are explained here. For each dataset containing missing values, the multiple imputation method outputs m datasets with different sets of imputed values for entries that were missing. Note that, given a complete dataset, the BOSS method finds a control group that minimizes an imbalance measure such that the control group is balanced with respect to the treatment group,

so bias from the differences of the two groups can be reduced when estimating the treatment effect.

Suppose that there are N units in the set of samples $\mathcal{N}$, namely $u_1, u_2, \cdots, u_N$. Denote a vector of K covariates for unit $u \in \mathcal{N}$ by $\mathbf{X}_u = (X_{u,1}, X_{u,2}, \cdots, X_{u,K}) \in \mathbb{R}^K$. For each unit $u \in U$, let $M_u = (M_{u,1}, M_{u,2}, \cdots, M_{u,K}) \in \mathbb{R}^K$ be a vector of K indicator variables for missing entries, where

$$M_{u,k} = \begin{cases} 1 \text{ if } X_{u,k} \text{ is missing} \\ 0 \text{ otherwise} \end{cases} \tag{5.27}$$

Let $\underline{\mathbf{X}} = [\mathbf{X}_{u_1}\ \mathbf{X}_{u_2}\ \cdots\ \mathbf{X}_{u_N}]' \in \mathbb{R}^{n \times K}$ be the matrix of covariate values of N units in $\mathcal{N}$ and $\mathbf{M} = [M_{u_1}\ M_{u_2}\ \cdots\ M_{u_N}]' \in \mathbb{R}^{n \times K}$ be the corresponding missing indicator matrix. Then $\mathscr{M} = \{X_{u,k} | M_{u,k} = 1 \text{ for } u \in \mathcal{N}, k \in \mathcal{P}\}$ is a set of missing covariate values and $\mathscr{O} = \{X_{u,k} | M_{u,k} = 0 \text{ for } u \in \mathcal{N}, k \in \mathcal{P}\}$ is a set of observed covariate values. Assume that missing entries only occur among the covariates (i.e., the treatment indicator and responses are fully observed for all the units).

Multiple imputation fills in missing covariate values multiple times (say, $\mathfrak{Q}$) using a predictive distribution of the missing covariate values given observed covariates and the treatment indicators. Denote the complete datasets obtained by using multiple imputation with $\underline{\mathbf{X}}$ by $\mathbf{X}^{<1>}, \mathbf{X}^{<2>}, \cdots, \mathbf{X}^{<\mathfrak{Q}>}$. Two approaches – the Within approach and the Across approach – that are applicable with the multiply imputed datasets will be introduced.

The Within approach finds $\mathfrak{Q}$ different control groups for $\mathfrak{Q}$ different complete datasets, estimates $\mathfrak{Q}$ treatment effects using those control groups, one by one, and computes the average of the $\mathfrak{Q}$ treatment effect estimates that are obtained to get a single value for the estimated treatment effect given the original dataset with missing values. A detailed implementation of the Within approach is given below.

Divide the set of observed units, $\mathcal{N}$, into a treatment group $T = \{u \in \mathcal{N} | Z_u = 1\}$ and a control pool $C = \{u \in \mathcal{N} | Z_u = 0\}$. For the i-th complete dataset $\mathbf{X}^{<i>}$, to reduce the bias from differences in covariate distributions of the treatment group and the control pool, BOSS first finds a control group $C^{',<i>} \subset C$ that minimizes an imbalance measure $\mathscr{I}(T, C^{',<i>})$ as a function of covariate values in $\mathbf{X}^{<i>}$. Then BOSS estimates the sample average treatment effect for the treated (SATT) using

the i-th complete dataset, with

$$\widetilde{\tau}_T^1(C^{'<i>}) \equiv \frac{1}{|T|}\sum_{t\in T} Y_t^1 - \frac{1}{|C^{'<i>}|}\sum_{c\in C^{'<i>}} Y_c^0. \tag{5.28}$$

As a next step, the information from these estimated values for all the complete datasets is aggregated by taking the average:

$$\widetilde{\tau}_T^{1,W} = \frac{\sum_{i=1,2,\cdots,\mathfrak{Q}} \widetilde{\tau}_T^1(C^{'<i>})}{\mathfrak{Q}}, \tag{5.29}$$

Equation (5.29) is the SATT estimate for $\underline{\mathbf{X}}$ after multiple imputation using the Within approach.

The Across approach first aggregates the information of the $\mathfrak{Q}$ different complete datasets by taking an average of those datasets, finds a single control group using the dataset obtained by averaging, and then obtains an estimated value for the treatment effect by applying BOSS to the averaged dataset.

Denote the average of $\mathfrak{Q}$ complete datasets by $\underline{\mathbf{X}}^A$:

$$\underline{\mathbf{X}}^A = \frac{\mathbf{X}^{<1>} + \mathbf{X}^{<2>} + \cdots + \mathbf{X}^{<\mathfrak{Q}>}}{\mathfrak{Q}}. \tag{5.30}$$

Then SATT can be estimated with

$$\widetilde{\tau}_T^{1,A} = \widetilde{\tau}_T^1(C^{',A}) \equiv \frac{1}{|T|}\sum_{t\in T} Y_t^1 - \frac{1}{|C^{',A}|}\sum_{c\in C^{',A}} Y_c^0 \tag{5.31}$$

using a control group $C^{',A} \subset C$ minimizing an imbalance measure $\mathscr{I}(T, C^{',A})$, which is a function of covariate values in $\underline{\mathbf{X}}^A$. The value $\widetilde{\tau}_T^{1,A}$ obtained from (5.31) is the SATT estimate for $\underline{\mathbf{X}}$ after multiple imputation using the Across approach.

The propensity score matching method can also be applied to the $\mathfrak{Q}$ complete datasets using both the Within approach and the Across approach. The Within approach in the context of the propensity score matching does the following: it finds $\mathfrak{Q}$ vectors of propensity scores respectively for the $\mathfrak{Q}$ complete datasets and computes $\mathfrak{Q}$ treatment effect estimates using the propensity score matching method for each dataset and the average of the $\mathfrak{Q}$ treatment effects to obtain a single treatment effect estimate for the original dataset. The Across approach in the context of the propensity score matching denotes a method which estimates a treatment effect using a single vector of propensity scores obtained by taking the average of the

$\mathfrak{Q}$ propensity score vectors. The Across and Within approaches were investigated by Mitra and Reiter (2016) and it was shown that the Across approach results in smaller bias compared to the Within approach when applied with propensity score matching. The details for implementation of these two approaches are as follows.

For the i-th complete dataset $\mathbf{X}^{<i>} = [\mathbf{X}_{u_1}^{<i>}\ \mathbf{X}_{u_2}^{<i>}\ \cdots\ \mathbf{X}_{u_N}^{<i>}]' \in \mathbb{R}^{n\times K}$, denote the estimated propensity score for unit $u \in \mathcal{N}$ by $e(\mathbf{X}_u^{<i>}) = \text{Prob}(Z_u = 1 \mid \mathbf{X}_u^{<i>}) \in \mathbb{R}^1$ which is a function of $\mathbf{X}_u^{<i>} \in \mathbb{R}^K$. Construct a vector of estimated propensity scores for each complete dataset $\mathbf{X}^{<i>}$: $\mathbf{e}(\mathbf{X}^{<i>}) = [e(\mathbf{X}_{u_1}^{<i>})\ e(\mathbf{X}_{u_2}^{<i>})\ \cdots\ e(\mathbf{X}_{u_N}^{<i>})]' \in \mathbb{R}^N$

Obtain a set of matched control units through propensity score matching using $\mathbf{e}(\mathbf{X}^{<i>})$ for the i-th complete dataset, denoted by $\hat{C}^{<i>}$. In this chapter, a nearest neighbor matching (Rubin, 1973) is used with $\hat{C}^{<i>} = \{c \mid c \in \arg\min_{c\in C} \|e(\mathbf{X}_c^{<i>}) - e(\mathbf{X}_t^{<i>})\| \text{ for some } t \in T\}$. Then SATT can be estimated using $\hat{C}^{<i>}$ from the i-th complete dataset with

$$\hat{\tau}_T^1(\hat{C}^{<i>}) \equiv \frac{1}{|T|}\sum_{t\in T} Y_t^1 - \frac{1}{|\hat{C}^{<i>}|}\sum_{c\in\hat{C}^{<i>}} Y_c^0. \tag{5.32}$$

Matching method is applied for each complete dataset $\hat{C}^{<i>}$ ($i = 1, 2, \cdots, \mathfrak{Q}$). Then, as for the Within approach for BOSS, the Within approach for matching also aggregates the information by taking the average.

$$\hat{\tau}_T^{1,W} = \frac{\sum_{i=1,2,\cdots,m}\hat{\tau}_T^1(\hat{C}^{<i>})}{m}, \tag{5.33}$$

Equation (5.33) is the Within approach matching estimator for $\underline{\mathbf{X}}$ after multiple imputation.

While the Within approach performs separate matching procedures for each of the $\mathfrak{Q}$ complete datasets, the Across approach first combines the information from the $\mathfrak{Q}$ propensity score vectors into one propensity score by averaging them:

$$\mathbf{e}^A(\mathbf{X}^{<1>}, \mathbf{X}^{<2>}, \cdots, \mathbf{X}^{<\mathfrak{Q}>}) = \frac{1}{\mathfrak{Q}}\sum_{i=1,2,\cdots,\mathfrak{Q}}\begin{bmatrix} e(\mathbf{X}_{u_1}^{<i>}) \\ e(\mathbf{X}_{u_2}^{<i>}) \\ \cdots \\ e(\mathbf{X}_{u_N}^{<i>}) \end{bmatrix} = \begin{bmatrix} \frac{1}{\mathfrak{Q}}\sum_{i=1,2,\cdots,\mathfrak{Q}} e(\mathbf{X}_{u_1}^{<i>}) \\ \frac{1}{\mathfrak{Q}}\sum_{i=1,2,\cdots,\mathfrak{Q}} e(\mathbf{X}_{u_2}^{<i>}) \\ \cdots \\ \frac{1}{\mathfrak{Q}}\sum_{i=1,2,\cdots,\mathfrak{Q}} e(\mathbf{X}_{u_m}^{<i>}) \end{bmatrix} \in \mathbb{R}^N \tag{5.34}$$

Using the aggregated estimated propensity score vector, $\mathbf{e}^A(\mathbf{X}^{<1>}, \mathbf{X}^{<2>}, \cdots, \mathbf{X}^{<\mathfrak{Q}>})$,

a nearest neighbor matching is performed to obtain a set of control units,

$$\hat{C}^A = \{c \mid c \in \arg\min_{c \in C} \|e(\mathbf{X}_c^A) - e(\mathbf{X}_t^A)\| \text{ for some } t \in T\}. \tag{5.35}$$

which are used to estimate estimate SATT,

$$\hat{\tau}_T^1(\hat{C}^A) \equiv \frac{1}{|T|} \sum_{t \in T} Y_t^1 - \frac{1}{|\hat{C}^A|} \sum_{c \in \hat{C}^A} Y_c^0 \tag{5.36}$$

given the the Across approach matching estimator of the given dataset after multiple imputation. Equation (5.36) is the Across approach matching estimator for $\underline{\mathbf{X}}$.

One of the objectives of this chapter is to compare these two approaches in the context of BOSS framework. It is of interest to see whether the better performance of the Across approach is still observed (as observed with matching method) when the two approaches were applied using BOSS method. In addition, a comparison of BOSS and matching methods to estimate the treatment effect with multiply-imputed datasets will be provided. The study in this chapter will be useful in that it helps to better understand and assess which method(s) should be used in the presence of missing data.

5.3.2 Simulation

This section provides data generation process and simulation results to compare the Within and the Across approaches in matching and BOSS. The two approaches will be compared within the BOSS framework and the difference between the BOSS estimates and the matching estimates will be also be investigated.

Data Generation

Simulated datasets are generated using the method described in Mitra and Reiter (2016). Each unit has two covariates (i.e., $K = 2$) which are normally distributed with mean and variance,

$$\mathbf{X}_u = \begin{bmatrix} X_{u,1} \\ X_{u,2} \end{bmatrix} \sim N\left(\begin{bmatrix} 10 \\ 10 \end{bmatrix}, \begin{bmatrix} 5 & 2.5 \\ 2.5 & 5 \end{bmatrix} \right) \tag{5.37}$$

Repeat the process of generating the covariates for $N = 1100$ units to generate the set of covariates for each unit. The value of N was chosen as 1000 in order to make the approximate number of the treated units and that of the control units which will be determined later be 100 and 1000 respectively. Suppose that both treatment response functions and control response functions are of the form

$$Y_u^1 = Y_u^0 = Y_u = \sum_{k=1,2} X_{u,k} + \epsilon_u \quad \forall u \in \mathcal{N}, \tag{5.38}$$

where ϵ_u are error terms in the response values for each unit $u \in \mathcal{N}$ that are normally distributed with mean 0 and variance 1. Since $Y_u^1 = Y_u^0$ for all $u \in \mathcal{N}$, then he true value of SATT is 0 (see (1.3)).

Assume that all the values of $X_{u,1}$, Z_u, and Y_u are fully observed for all $u \in \mathcal{N}$ and missing values are in the second covariates $X_{u,2}$ for some $u \in \mathcal{N}$. Depending on the treatment assignment mechanism, three scenarios can be considered. In the first scenario (Scenario 1), whether Z_u is equal to 0 or 1 is determined by $X_{u,1}$ only. In the second scenario (Scenario 2), whether Z_u is equal to 0 or 1 is determined by $X_{u,2}$ only. In the third scenario (Scenario 3), the contributions of $X_{u,1}$ and $X_{u,2}$ in determining assignment to treatment are identical.

Scenario 1

The assignment to treatment for a unit u depends only on the first covariate of the unit, $X_{u,1}$. Suppose that the treatment indicator value is assigned using a logistic function for the propensity score:

$$e(\mathbf{X}_u) = \text{Prob}(Z_u = 1 \mid \mathbf{X}_u) = \frac{e^{\alpha_e + \beta_e X_{u,1}}}{1 + e^{\alpha_e + \beta_e X_{u,1}}}, \tag{5.39}$$

where $\alpha_e = -7.8$ and $\beta_e = 0.5$.

Two sub-scenarios, Scenario 1-1 and Scenario 1-2, can be constructed based on the assignment of the missing data. In Scenario 1-1, the values for the second covariate are missing only for some control units $u \in C$. The probability of $X_{u,2}$ being missing is again given by a logistic function:

$$\text{Prob}(M_{u,2} = 1 \mid Z_u = 0, \mathbf{X}_u) = \frac{e^{\alpha_m + \beta_m X_{u,1}}}{1 + e^{\alpha_m + \beta_m X_{u,1}}} \tag{5.40}$$

where $\alpha_m = -10.1$ and $\beta_m = 0.9$. Note that in Scenario 1-1,

$$\mathbb{E}[\text{Prob}(M_{u,2} = 1 \mid Z_u = 0, \mathbf{X}_u)] = \frac{e^{\alpha_m + \beta_m \mathbb{E}[X_{u,1}]}}{1 + e^{\alpha_m + \beta_m \mathbb{E}[X_{u,1}]}} = \frac{e^{-1.1}}{1 + e^{-1.1}} \approx 0.3 \tag{5.41}$$

since $\mathbb{E}[X_{u,1}] = 10$ from (5.37).

In Scenario 1-2, the values for the second covariate are missing for both treated units and control units. The control units' covariate values are missing using the same equation (5.40) as in Scenario 1-1 and treated units' covariate values are missing completely at random with

$$\text{Prob}(M_{u,2} = 1 \mid Z_u = 1, \mathbf{X}_u) = 0.3, \tag{5.42}$$

so that approximately 30% of the control units' second covariate values and exactly 30% of the treated units' second covariate values are missing in this scenario.

The covariate values that were originally assigned to those covariates, which happened to be missing, are discarded and the missing covariate values are imputed multiple times ($\mathfrak{L} = 5$ times) using a Bayesian linear regression of $X_{\mathcal{N},2} \equiv (X_{u_1,2}, X_{u_2,2}, \cdots, X_{u_N,2})$ on $X_{\mathcal{N},1} \equiv (X_{u_1,1}, X_{u_2,1}, \cdots, X_{u_N,1})$ and $Z_{\mathcal{N}} \equiv (Z_{u_1}, Z_{u_2}, \cdots, Z_{u_N})$. For the multiple imputation, `mice` (Multiple Imputation by Chained Equations; an R package) is used. After obtaining m complete datasets through multiple imputation on missing covariates, the Within and Across approaches of both BOSS and propensity score matching are applied as described in Section 5.3.1. The outputs of those methods are $\widetilde{\tau}_T^{1,W}$ and $\widetilde{\tau}_T^{1,A}$ for BOSS and $\hat{\tau}_T^{1,W}$ and $\hat{\tau}_T^{1,A}$ for propensity score matching.

For all the scenarios, the same process is repeated 100 times while the values of $\underline{\mathbf{X}} = [\mathbf{X}_{u_1} \ \mathbf{X}_{u_2} \ \cdots \ \mathbf{X}_{u_{1100}}]' \in \mathbb{R}^{1100\times 2}$, $\underline{Z} = (Z_{u_1}, Z_{u_2}, \cdots, Z_{u_{1100}}) \in \mathbb{R}^{1100}$, $Y = (Y_{u_1}, Y_{u_2}, \cdots, Y_{u_{1100}}) \in \mathbb{R}^{1100}$ and $\mathbf{M} = [M_{u_1} \ M_{u_2} \ \cdots \ M_{u_{1100}}]' \in \mathbb{R}^{1100\times 2}$ are newly generated each time.

Scenario 2

Assignment to treatment for a unit u depends only on the second covariate of the unit, $X_{u,2}$, following

$$e(\mathbf{X}_u) = \text{Prob}(Z_u = 1 \mid \mathbf{X}_u) = \frac{e^{\alpha_e + \beta_e X_{u,2}}}{1 + e^{\alpha_e + \beta_e X_{u,2}}} \tag{5.43}$$

where $\alpha_e = -7.8$ and $\beta_e = 0.5$.

Consider the two sub-scenarios, Scenario 2-1 and Scenario 2-2, based on the missing entry assignment. The values for the second covariate are missing only for some control units $u \in C$ in Scenario 2-1, while they are missing for some control and treated units in Scenario 2-2. The probability of the control units having a missing covariate in Scenario 2-1 and Scenario 2-2 is given by (5.40). In Scenario

2-1, There are no missing covariates among treated units and the probability of $X_{u,2}$ of the treated units being a missing covariate is 0. In Scenario, the probability of $X_{u,2}$ of the treated units being a missing covariate is given by (5.42) as before.

After multiple imputations of the missing covariates using the Bayesian linear regression implemented in `mice`, compute $\widetilde{\tau}_T^{1,W}$, $\widetilde{\tau}_T^{1,A}$, $\hat{\tau}_T^{1,W}$, and $\hat{\tau}_T^{1,A}$. Repeat the above process for 100 times each with newly generated values for $\underline{\mathbf{X}}$, $\underline{Z}$, Y, and $\mathbf{M}$.

Scenario 3

The following equation is used for determining assignment to treatment so that the first covariate and the second covariate can have the same impact on the propensity score:

$$e(\mathbf{X}_u) = \text{Prob}(Z_u = 1 \mid \mathbf{X}_u) = \frac{e^{\alpha_e + \beta_{e,1} X_{u,1} + \beta_{e,2} X_{u,2}}}{1 + e^{\alpha_e + \beta_{e,1} X_{u,1} + \beta_{e,2} X_{u,2}}} \tag{5.44}$$

where $\alpha_e = -7.8$ and $\beta_{e,1} = \beta_{e,2} = 0.255$.

The missing data mechanism for the two sub-scenarios, Scenarios 3-1 and 3-2, is the same as for the Scenarios 1-1 and 2-1, and Scenario 1-2 and 2-2, respectively as described above. Multiple imputation of the missing covariates is conducted using the Bayesian linear regression and the four estimates for SATT ($\widetilde{\tau}_T^{1,W}$, $\widetilde{\tau}_T^{1,A}$, $\hat{\tau}_T^{1,W}$, and $\hat{\tau}_T^{1,A}$) are computed for 100 times each with newly generated values of $\underline{\mathbf{X}}$, $\underline{Z}$, Y, and $\mathbf{M}$.

Simulation Results

The simulation results of the four estimates for the six scenarios (Scenarios 1-1, 1-2, 2-1, 2-2, 3-1, and 3-2) are given in Tables 5.8 and 5.9. All the values reported in Table 5.9 for BOSS are those that are obtained after the 300 second time limit. This time limit was chosen because after this amount of time there was not a significant decrease in the objective value when solving BOSS instances.

Note that the point estimate values equal to bias values because the true SATT value is equal to 0 for all scenarios while bias is defined as the point estimate minus the true value. The point estimates for the matching and BOSS estimators reported in Tables 5.8 and 5.9 are obtained by taking an average of the matching estimates $\hat{\tau}_{T,k}^{1,W}$, and the BOSS estimates $\hat{\tau}_{T,k}^{1,A}$, $\widetilde{\tau}_{T,k}^{1,W}$, and $\widetilde{\tau}_{T,k}^{1,A}$ from the k-th repetition

for $k = 1, 2, \cdots, 100$ (See Section 5.3.1). That is,

$$\overline{\hat{\tau}_T^{1,W}} = \frac{1}{100}\sum_{k=1}^{100}\hat{\tau}_{T,k}^{1,W}, \quad \overline{\hat{\tau}_T^{1,A}} = \frac{1}{100}\sum_{k=1}^{100}\hat{\tau}_{T,k}^{1,A}, \tag{5.45}$$

and

$$\overline{\widetilde{\tau}_{T,k}^{1,W}} = \frac{1}{100}\sum_{k=1}^{100}\widetilde{\tau}_{T,k}^{1,W}, \quad \overline{\widetilde{\tau}_{T,k}^{1,A}} = \frac{1}{100}\sum_{k=1}^{100}\widetilde{\tau}_{T,k}^{1,A}. \tag{5.46}$$

The variance of the estimates reported in Table 5.8 and 5.9 are computed using

$$\frac{1}{100-1}\sum_{k=1}^{100}\left(\hat{\tau}_{T,k}^{1,W} - \overline{\hat{\tau}_T^{1,W}}\right)^2, \quad \frac{1}{100-1}\sum_{k=1}^{100}\left(\hat{\tau}_{T,k}^{1,A} - \overline{\hat{\tau}_T^{1,A}}\right)^2, \tag{5.47}$$

and

$$\frac{1}{100-1}\sum_{k=1}^{100}\left(\widetilde{\tau}_{T,k}^{1,W} - \overline{\hat{\tau}_T^{1,W}}\right)^2, \quad \frac{1}{100-1}\sum_{k=1}^{100}\left(\widetilde{\tau}_{T,k}^{1,A} - \overline{\hat{\tau}_T^{1,A}}\right)^2, \tag{5.48}$$

respectively for the Within/Across approach matching estimators and the Within/Across approach BOSS estimators. Note that the above form of variance corresponds to the between-imputation variance in Section 5.2.3. The mean squared error (MSE) is computed using

$$\frac{1}{100}\sum_{k=1}^{100}\left(\hat{\tau}_{T,k}^{1,W} - 0\right)^2, \quad \frac{1}{100}\sum_{k=1}^{100}\left(\hat{\tau}_{T,k}^{1,A} - 0\right)^2, \tag{5.49}$$

and

$$\frac{1}{100}\sum_{k=1}^{100}\left(\widetilde{\tau}_{T,k}^{1,W} - 0\right)^2, \quad \frac{1}{100}\sum_{k=1}^{100}\left(\widetilde{\tau}_{T,k}^{1,A} - 0\right)^2. \tag{5.50}$$

since the true value for these estimates is 0.

As one can see from the Tables 5.8 and 5.9, the Across approach outperforms the Within approach for both BOSS and matching methods: the magnitude of the point estimate for SATT (i.e., the magnitude of the bias) is smaller in the Across approach when compared to the Within approach. Simulation results reported in this chapter confirm the results of Mitra and Reiter (2016) for matching and additionally one can see that the simulation for BOSS exhibit the same pattern.

Furthermore, the performance of BOSS estimators and corresponding matching estimators can be compared. In all six scenarios, the BOSS does better than matching for both approaches in that the Within approach for BOSS shows smaller

magnitude of bias compared to the Within approach for matching and the Across approach for BOSS shows smaller magnitude of bias compared to the Across approach for matching. However, BOSS takes a longer computational time than matching since matching problems are poly-time solvable while the general BOSS problems are computationally intractable (Sauppe et al., 2014). The time limit that was chosen for solving BOSS instances can be reduced but that will result in less accurate point estimate values. There is a trade-off between getting a better result with smaller bias and using less time in computation.

5.4 Concluding Remarks

This chapter applied sensitivity analysis method using additional simulated confounder to BOSS estimators. The chapter also discussed two methods after multiple imputation of missing values in observed covariates. Advantages and disadvantages of BOSS estimators over matching estimators in these analyses were discussed. Note that, when strong ignorability assumption is involved for the estimation of causal effects, conducting the sensitivity analysis provides valuable information on sensitivity of the estimators. Further note that BOSS methods have potential for greater performance over matching when dealing with missing entries through multiple imputation. With advances in computing techniques for faster computation, longer computational time for BOSS will become less of an issue if a smaller bias can be obtained with the method.

When comparing the matching and the BOSS estimators using simulation of unobserved confounders, BOSS estimators are less sensitive compared to matching estimates to the violation of the conditional independence assumption in the sensitivity analysis using LaLonde data. In addition, BOSS estimators provide stable standard errors when adding a binary confounder with large d and s values while the standard errors of matching estimators increased dramatically. A precise information on estimated values from matching for large d and s could not be obtained because the large standard errors generated a very wide confidence interval for SATT.

BOSS estimators are better than matching estimators as BOSS estimators show more robustness to failure of the conditional independence assumption when estimating SATT with this dataset. However, BOSS estimators also have some drawbacks – computing them takes much longer than matching and the estimated val-

ues from BOSS can be biased if the objective is not optimized to zero because of the time limit or insufficient data. These shortcomings were also observed in BOSS simulations reported in this chapter. Note that matching estimators can also be biased because of their use of inexact matches when exact matches are unavailable. Therefore, researchers should decide which estimator to use considering both their strengths and weaknesses depending on their situation.

For the Pri.DE data, the sensitivity results for matching and BOSS are comparable. In case of BOSS, while the computational time required was greater than that of matching, it was much shorter for Pri.DE dataset than for LaLonde dataset because most optimal solutions were found within the time limit that was set.

The examples discussed in this chapter may not be appropriate to be generalized to all the settings. However, the simulation study that is conducted in this chapter provides a sensitivity analysis results of BOSS estimators and suggests that the Across approach outperforms the Within approach for BOSS as well as for matching under the cases that are examined.

Table 5.4: Values of $(p_{11}, p_{10}, p_{01}, p_{00})$ under Each Scenario

	$s = 0.1$	$s = 0.2$	$s = 0.3$	$s = 0.4$	$s = 0.5$	$s = 0.6$
$d = 0.1$	$(0.49, 0.49, 0.44, 0.34)$	$(0.59, 0.59, 0.43, 0.33)$	$(0.68, 0.68, 0.43, 0.33)$	$(0.77, 0.77, 0.42, 0.32)$	$0.87, 0.87, 0.41, 0.31)$	$(0.96, 0.96, 0.41, 0.31)$
$d = 0.2$	$(0.49, 0.49, 0.49, 0.29)$	$(0.59, 0.59, 0.48, 0.28)$	$(0.68, 0.68, 0.48, 0.28)$	$(0.77, 0.77, 0.47, 0.27)$	$0.87, 0.87, 0.46, 0.26)$	$(0.96, 0.96, 0.45, 0.25)$
$d = 0.3$	$(0.49, 0.49, 0.54, 0.24)$	$(0.59, 0.59, 0.53, 0.23)$	$(0.68, 0.68, 0.52, 0.22)$	$(0.77, 0.77, 0.52, 0.22)$	$0.87, 0.87, 0.51, 0.21)$	$(0.96, 0.96, 0.50, 0.20)$
$d = 0.4$	$(0.49, 0.49, 0.59, 0.19)$	$(0.59, 0.59, 0.58, 0.18)$	$(0.68, 0.68, 0.57, 0.17)$	$(0.77, 0.77, 0.56, 0.16)$	$0.87, 0.87, 0.56, 0.16)$	$(0.96, 0.96, 0.55, 0.15)$
$d = 0.5$	$(0.49, 0.49, 0.63, 0.13)$	$(0.59, 0.59, 0.63, 0.13)$	$(0.68, 0.68, 0.62, 0.12)$	$(0.77, 0.77, 0.61, 0.11)$	$0.87, 0.87, 0.61, 0.11)$	$(0.96, 0.96, 0.60, 0.10)$
$d = 0.6$	$(0.49, 0.49, 0.68, 0.08)$	$(0.59, 0.59, 0.67, 0.07)$	$(0.68, 0.68, 0.67, 0.07)$	$(0.77, 0.77, 0.66, 0.06)$	$0.87, 0.87, 0.65, 0.05)$	$(0.96, 0.96, 0.65, 0.05)$

Figure 5.1: Graphical Representation of $(p_{11}, p_{10}, p_{01}, p_{00})$ Values under 36 Scenarios in Table 5.4

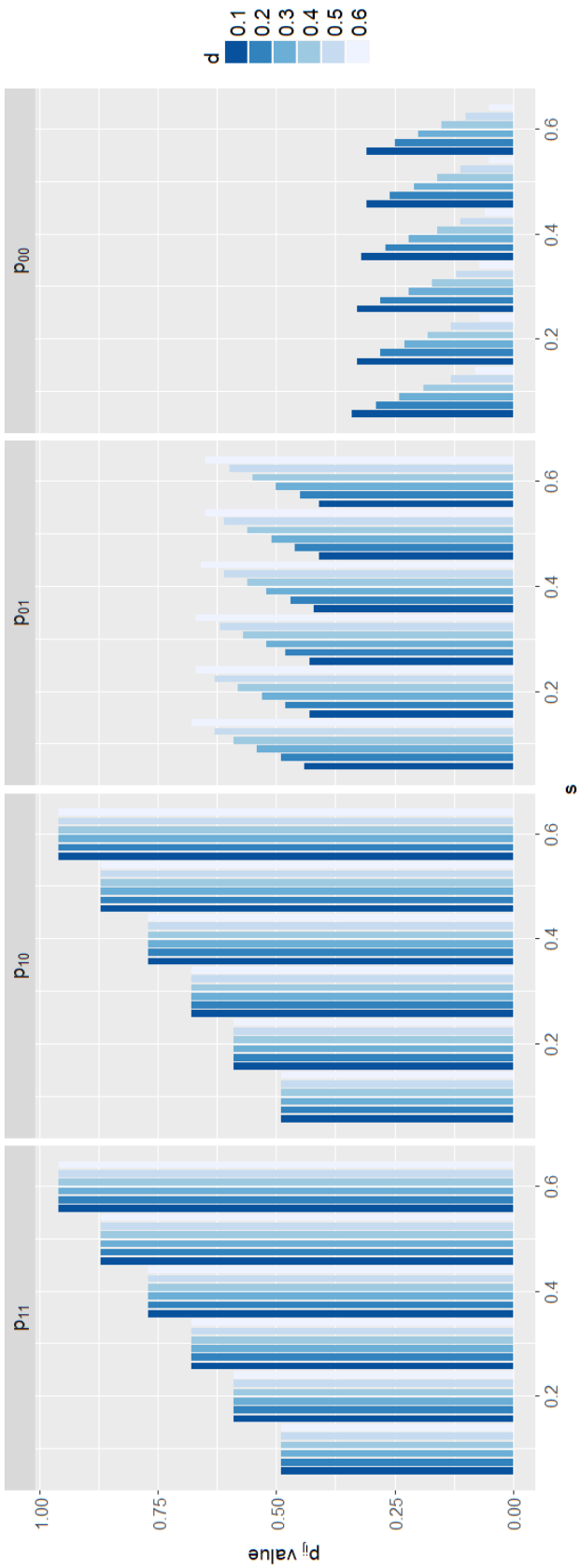


Table 5.5: Matching Estimators (and Corresponding Standard Errors)

	$s = 0.1$	$s = 0.2$	$s = 0.3$	$s = 0.4$	$s = 0.5$	$s = 0.6$	$s = 0.7$	$s = 0.8$	$s = 0.9$
$d = 0.1$	0.068 (0.032)	0.061 (0.032)	0.048 (0.034)	0.029 (0.037)	0.017 (0.039)	−0.007 (0.043)	−0.033 (0.048)	−0.069 (0.056)	−0.145 (0.062)
$d = 0.2$	0.061 (0.032)	0.044 (0.031)	0.017 (0.035)	−0.009 (0.036)	−0.044 (0.037)	−0.089 (0.040)	−0.137 (0.042)	−0.232 (0.042)	N/A
$d = 0.3$	0.049 (0.033)	0.021 (0.033)	−0.012 (0.033)	−0.060 (0.033)	−0.102 (0.034)	−0.165 (0.036)	−0.265 (0.032)	N/A	N/A
$d = 0.4$	0.037 (0.032)	0.003 (0.031)	−0.045 (0.033)	−0.095 (0.031)	−0.170 (0.032)	N/A	N/A	N/A	N/A
$d = 0.5$	0.033 (0.032)	−0.013 (0.031)	−0.071 (0.031)	N/A	N/A	N/A	N/A	N/A	N/A
$d = 0.6$	0.027 (0.032)	−0.028 (0.031)	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 5.6: BOSS Estimators (and Corresponding Standard Errors)

	$s = 0.1$	$s = 0.2$	$s = 0.3$	$s = 0.4$	$s = 0.5$	$s = 0.6$	$s = 0.7$	$s = 0.8$	$s = 0.9$
$d = 0.1$	0.12 (0.028)	0.11 (0.032)	0.052 (0.025)	0.053 (0.024)	0.068 (0.026)	0.082 (0.027)	0.095 (0.025)	0.10 (0.027)	0.11 (0.027)
$d = 0.2$	0, 11 (0.031)	0.092 (0.029)	0.020 (0.025)	0.027 (0.025)	0.044 (0.025)	0.060 (0.027)	0.073 (0.027)	0.088 (0.026)	N/A
$d = 0.3$	0.097 (0.029)	0.067 (0.028)	−0.002 (0.027)	0.003 (0.027)	0.019 (0.026)	0.036 (0.026)	0.056 (0.026)	N/A	N/A
$d = 0.4$	0.083 (0.029)	0.054 (0.029)	−0.032 (0.028)	−0.028 (0.025)	−0.003 (0.026)	N/A	N/A	N/A	N/A
$d = 0.5$	0.072 (0.028)	0.035 (0.029)	−0.056 (0.028)	N/A	N/A	N/A	N/A	N/A	N/A
$d = 0.6$	0.075 (0.028)	0.017 (0.029)	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 5.7: Values of $(p_{11}, p_{10}, p_{01}, p_{00})$ under Each Scenario

	$s = 0.1$	$s = 0.2$	$s = 0.3$	$s = 0.4$	$s = 0.5$	$s = 0.6$	$s = 0.7$	$s = 0.8$	$s = 0.9$
$d = 0.1$	(0.47, 0.47, 0.41, 0.31)	(0.53, 0.53, 0.38, 0.28)	(0.60, 0.60, 0.35, 0.25)	(0.67, 0.67, 0.31, 0.21)	(0.73, 0.73, 0.28, 0.18)	(0.80, 0.80, 0.24, 0.14)	(0.87, 0.87, 0.21, 0.11)	(0.93, 0.93, 0.18, 0.08)	(1.00, 1.00, 0.14, 0.04)
$d = 0.2$	(0.47, 0.47, 0.46, 0.26)	(0.53, 0.53, 0.42, 0.22)	(0.60, 0.60, 0.39, 0.19)	(0.67, 0.67, 0.36, 0.16)	(0.73, 0.73, 0.32, 0.12)	(0.80, 0.80, 0.29, 0.09)	(0.87, 0.87, 0.26, 0.06)	(0.93, 0.93, 0.22, 0.02)	N/A
$d = 0.3$	(0.47, 0.47, 0.50, 0.20)	(0.53, 0.53, 0.47, 0.17)	(0.60, 0.60, 0.44, 0.14)	(0.67, 0.67, 0.40, 0.10)	(0.73, 0.73, 0.37, 0.07)	(0.80, 0.80, 0.34, 0.04)	(0.87, 0.87, 0.30, 0.00)	N/A	N/A
$d = 0.4$	(0.47, 0.47, 0.55, 0.15)	(0.53, 0.53, 0.52, 0.12)	(0.60, 0.60, 0.48, 0.08)	(0.67, 0.67, 0.45, 0.05)	(0.73, 0.73, 0.41, 0.01)	N/A	N/A	N/A	N/A
$d = 0.5$	(0.47, 0.47, 0.59, 0.09)	(0.53, 0.53, 0.56, 0.06)	(0.60, 0.60, 0.53, 0.03)	N/A	N/A	N/A	N/A	N/A	N/A
$d = 0.6$	(0.47, 0.47, 0.64, 0.04)	(0.53, 0.53, 0.61, 0.01)	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Figure 5.2: Graphical Representation of $(p_{11}, p_{10}, p_{01}, p_{00})$ Values under 34 Scenarios in Table 5.7

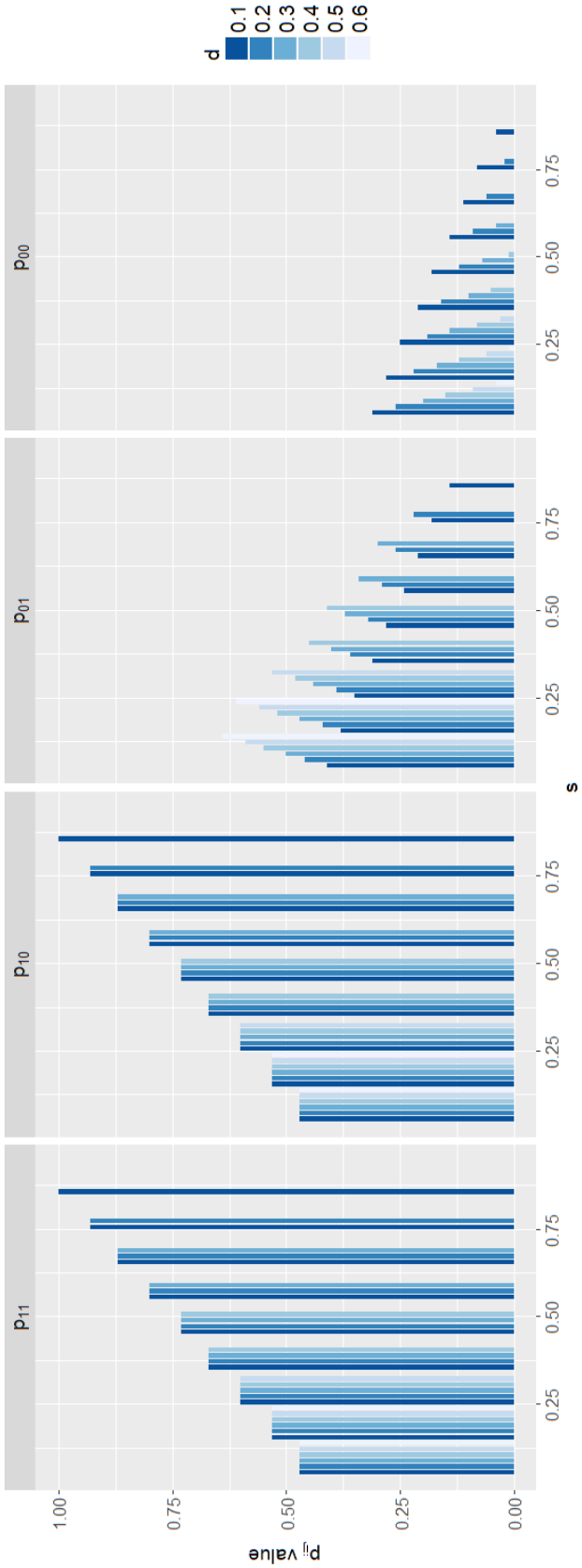


Table 5.8: Propensity Score Matching Estimators after multiple imputation with $\mathfrak{L} = 5$ and 100 repetitions

Scenario	Within approach			Across approach		
	Estimate ($\hat{\tau}_T^{1,W}$)	Variance	MSE	Estimate ($\hat{\tau}_T^{1,A}$)	Variance	MSE
	Only control units have the missing covariates $\mathbf{X}_{u,2}$ such that $M_{u,2} = 1$					
1-1	0.060	0.059	0.062	0.034	0.075	0.075
2-1	0.811	0.042	0.699	0.540	0.075	0.367
3-1	0.520	0.043	0.313	0.377	0.056	0.197
	Both treated and control units have the missing covariates $\mathbf{X}_{u,2}$ such that $M_{u,2} = 1$					
1-2	0.068	0.060	0.064	0.038	0.092	0.093
2-2	0.850	0.053	0.774	0.287	0.111	0.192
3-2	0.525	0.046	0.321	0.314	0.073	0.171

Table 5.9: BOSS Estimators after multiple imputation with $\mathfrak{L} = 5$ and 100 repetitions

Scenario	Within approach			Across approach		
	Estimate ($\widetilde{\tau}_T^{1,W}$)	Variance	MSE	Estimate ($\widetilde{\tau}_T^{1,A}$)	Variance	MSE
	Only control units have the missing covariates $\mathbf{X}_{u,2}$ such that $M_{u,2} = 1$					
1-1	−0.037	0.049	0.050	0.022	0.058	0.058
2-1	0.728	0.035	0.565	0.227	0.039	0.090
3-1	0.414	0.044	0.214	0.137	0.046	0.065
	Both treated and control units have the missing covariates $\mathbf{X}_{u,2}$ such that $M_{u,2} = 1$					
1-2	−0.031	0.056	0.056	0.022	0.064	0.064
2-2	0.734	0.035	0.574	0.228	0.064	0.115
3-2	0.411	0.044	0.213	0.142	0.066	0.085

CHAPTER 6

DUALITY IN BALANCE OPTIMIZATION SUBSET SELECTION

6.1 Introduction

The Balance Optimization Subset Selection (BOSS) framework by Nikolaev et al. (2013) is a causal inference method to estimate treatment effects through a modern optimization lens. It is known that certain instances of BOSS (depending on the imbalance measures used) can be formulated as a mixed-integer linear programming problem.

The mixed integer programming (MIP) formulation of the problem (1.6) with $\mathscr{I}_{\text{DOM}}$ is provided in (6.1) of Section 6.2. As mentioned earlier, with the difference of imbalance measure, the treatment effect estimator is guaranteed to be unbiased for linear response functions given that the strong ignorability assumption is satisfied and that all imbalance is removed (Sauppe and Jacobson, 2017).

BOSS has been discussed with integral control groups where each unit from the control pool is either included in the control group or not. However, the earlier results developed with control groups that are integral can be extend to fractional control groups where the control units are weighted in their contributions to both the balance constraints and the average control response of the control group. With such a change of allowing a fractional contribution of the control units (i.e., through a relaxation of the integrality constraints), the mixed integer linear programming problem can be transformed into a linear programming (LP) problem. See Section 6.2 for details.

With an LP, duality theory can be applied to this problem. In Section 6.2, how the LP can be formulated for the BOSS problem with $\mathscr{I}_{\text{DOM}}$ will be shown and its dual problem will be constructed. After having a look at basic properties of the primal and dual problems, additional properties will be studied in Section 6.3. Section 6.4 concludes the chapter with thoughts on future research direction.

6.2 Basic Properties of the Primal and Dual Problems of BOSS

In this section, the BOSS is formulated as an LP with the DOM imbalance measure, $\mathscr{I}_{\text{DOM}}$. Before having a look at the LP problem, consider the mixed integer programming formulation of the BOSS problem. The following formulation as an MIP is applicable to some specific (e.g., moment-based) forms of imbalance measures.

Recall that the vector of covariates for unit u is denoted by $\mathbf{X}_u = \{X_{u,1}, X_{u,2}, \cdots, X_{u,K}\}$ for covariate indices $1, 2, \cdots, K$. As stated in Chapter 1, BOSS with the DOM imbalance measure $\mathscr{I}_{\text{DOM}}$ can be formulated as the mixed integer program in (6.1) to get $C' = \{c \in C : v_c = 1\}$ which minimizes $\mathscr{I}_{\text{DOM}}$ between T and C' (Sauppe, 2015). Since C is a pool of discrete control units, the units in C can be denoted by $c_1, c_2, \cdots, c_{|C|}$ without loss of generality.

$$
\begin{aligned}
\min \quad & \sum_{k \in \mathcal{P}} w_k & & \\
\text{s.t.} \quad & \tfrac{1}{s} \sum_{c \in C} v_c X_{c,k} - \tfrac{1}{|T|} \sum_{t \in T} X_{t,k} \leq w_k & & \forall k \in \mathcal{P} \\
& \tfrac{1}{|T|} \sum_{t \in T} X_{t,k} - \tfrac{1}{s} \sum_{c \in C} v_c X_{c,k} \leq w_k & & \forall k \in \mathcal{P} \\
& \sum_{c \in C} v_c = s & & \\
& v_c \in \{0, 1\} & & \forall c \in C \\
& w_k \geq 0 & & \forall k \in \mathcal{P}.
\end{aligned}
\tag{6.1}
$$

The integer constraints in the MIP can be relaxed and BOSS can be formulated as an LP if a fractional contribution of control units in the optimal control group is permitted.

$$
\begin{aligned}
\min \quad & \sum_{k \in \mathcal{P}} w_k & & \\
\text{s.t.} \quad & \tfrac{1}{s} \sum_{c \in C} v_c X_{c,k} - \tfrac{1}{|T|} \sum_{t \in T} X_{t,k} \leq w_k & & \forall k \in \mathcal{P} \\
& \tfrac{1}{|T|} \sum_{t \in T} X_{t,k} - \tfrac{1}{s} \sum_{c \in C} v_c X_{c,k} \leq w_k & & \forall k \in \mathcal{P} \\
& \sum_{c \in C} v_c = s & & \\
& v_c \geq 0, & & \forall c \in C \\
& w_k \geq 0 & & \forall k \in \mathcal{P}.
\end{aligned}
\tag{6.2}
$$

In this LP, v_c (for $c \in C$) and w_k (for $k \in \mathcal{P}$) are primal variables. The values s, $X_{c,k}$, $X_{t,k}$ for $c \in C, t \in T, k \in \mathcal{P}$ are constants given before solving the problem.

Note that the formulation after relaxation of the integer constraints allows repetition in inclusion of units in the control group selection as there is no upper bound on v_c (and thus v_c can be greater than zero).

To simplify the notation, replace the average values of each covariate in the treatment group $\frac{1}{|T|}\sum_{t\in T} X_{t,k}$ by $\overline{\mu}_{T,k}$. In addition, by rearranging the order so that primal variables appear in the left-hand side of the constraints and constants appear in the right-hand side of the constraints, the primal LP can be written as follows.

Primal Problem

$$\begin{aligned}
\min \quad & \sum_{k\in\mathcal{P}} w_k & & \\
\text{s.t.} \quad & \tfrac{1}{s}\sum_{c\in C} v_c X_{c,k} - w_k & \le \overline{\mu}_{T,k} \quad & \forall k \in \{1,2,\cdots,K\} \\
& -\tfrac{1}{s}\sum_{c\in C} v_c X_{c,k} - w_k & \le -\overline{\mu}_{T,k} \quad & \forall k \in \{1,2,\cdots,K\} \\
& \sum_{c\in C} v_c & = s & \\
& v_c & \ge 0 \quad & \forall c \in C \\
& w_k & \ge 0 \quad & \forall k \in \{1,2,\cdots,K\}.
\end{aligned} \tag{6.3}$$

Primal Problem (in Canonical Form)

$$\begin{aligned}
\min \quad & \sum_{k\in\mathcal{P}} w_k & & \\
\text{s.t.} \quad & -\tfrac{1}{s}\sum_{c\in C} v_c X_{c,k} + w_k & \ge -\overline{\mu}_{T,k} \quad & \forall k \in \{1,2,\cdots,K\} \\
& \tfrac{1}{s}\sum_{c\in C} v_c X_{c,k} + w_k & \ge \overline{\mu}_{T,k} \quad & \forall k \in \{1,2,\cdots,K\} \\
& \sum_{c\in C} v_c & \ge s & \\
& -\sum_{c\in C} v_c & \ge -s & \\
& v_c & \ge 0 \quad & \forall c \in C \\
& w_k & \ge 0 \quad & \forall k \in \{1,2,\cdots,K\}
\end{aligned} \tag{6.4}$$

Denote the dual variables corresponding to the first K constraints (for $k = 1,2,\cdots,K$) in (6.3) by $y_{1^+}, y_{2^+}, \cdots, y_{K^+}$, the dual variables corresponding to the next K constraints by $y_{1^-}, y_{2^-}, \cdots, y_{K^-}$, and the dual variable corresponding to the equality constraint $v_c = s$ by y_s. The the dual problem of the LP in (6.3) can be written as

Dual Problem

$$\begin{aligned}
\max \quad & \sum_{k=1}^{K} \overline{\mu}_{T,k} y_{k^+} + \sum_{k=1}^{K} \overline{\mu}_{T,k} y_{k^-} && + s \cdot y_s \\
\text{s.t.} \quad & \sum_{k=1}^{K} \tfrac{1}{s} X_{c,k} y_{k^+} - \sum_{k=1}^{K} \tfrac{1}{s} X_{c,k} y_{k^-} + y_s && \leq 0 \quad \forall c \in C \\
& -y_{k^+} - y_{k^-} && \leq 1 \quad \forall k \in \{1, 2, \cdots, K\} \\
& y_{k^+}, y_{k^-} && \leq 0 \quad \forall k \in \{1, 2, \cdots, K\}.
\end{aligned} \tag{6.5}$$

Note that, in the primal problem, there are $|C| + K$ primal variables: $v_{c_1}, v_{c_2}, \cdots, v_{c_{|C|}}, w_1, w_2, \cdots, w_K$. In the dual problem, there are $2K + 1$ dual variables: $y_{1^+}, y_{2^+}, \cdots, y_{K^+}, y_{1^-}, y_{2^-}, \cdots, y_{K^-}, y_s$.

Suppose that a primal optimal solution with optimal objective value 0. Then, from strong duality, the following holds:

$$\sum_{k=1}^{K} \overline{\mu}_{T,k} y^*_{k^+} + \sum_{k=1}^{K} \overline{\mu}_{T,k} y^*_{k^-} + s \cdot y^*_s = 0 \tag{6.6}$$

Dual variables with values 0, i.e., $(y^*_{1^+}, y^*_{2^+}, \cdots, y^*_{K^+}, y^*_{1^-}, y^*_{2^-}, \cdots, y^*_{K^-}, y^*_s) = (0, 0, \cdots, 0)$, satisfy all the dual constraints and have optimal value 0. Hence, if the primal optimal solution has the objective value 0, then an optimal dual solution is given by $(y^*_{1^+}, y^*_{2^+}, \cdots, y^*_{K^+}, y^*_{1^-}, y^*_{2^-}, \cdots, y^*_{K^-}, y^*_s) = (0, 0, \cdots, 0)$ with objective value 0.

Note that there are "corner cases" where the average covariate values for the treatment group are zero (i.e, $\overline{\mu}_{T,k} = 0$ for some $k \in \mathcal{P}$). In these corner cases, the corresponding dual variables y_{k^+} and y_{k^-} need not be zero in an optimal dual solution with objective value zero.

In such a case when the primal and dual optimal solutions have objective optimal value 0, the w_k values in the primal optimal solution are 0. That is, $(w^*_1, w^*_2, \cdots, w^*_K) = (0, 0, \cdots, 0)$. From the first and second type constraints in the primal problem, the following holds:

$$\frac{1}{s} \sum_{c \in C} v^*_c X_{c,k} - \overline{\mu}_{T,k} \leq 0 \quad \forall k \in \mathcal{P} \tag{6.7}$$

and

$$\overline{\mu}_{T,k} - \frac{1}{s} \sum_{c \in C} v^*_c X_{c,k} \leq 0 \quad \forall k \in \mathcal{P}. \tag{6.8}$$

Equivalently,

$$0 \leq \frac{1}{s} \sum_{c \in C} v^*_c X_{c,k} - \overline{\mu}_{T,k} \leq 0 \quad \forall k \in \mathcal{P} \tag{6.9}$$

implies

$$\frac{1}{s}\sum_{c\in C} v_c^* X_{c,k} - \overline{\mu}_{T,k} = 0 \quad \forall k \in \mathcal{P}. \tag{6.10}$$

In the following section, more properties of the primal and dual solutions of the BOSS problem will be discussed.

6.3 Relationship between Primal and Dual Solutions of BOSS

This section investigates how the optimal solutions of the primal and dual problems for BOSS are related. Both general results and examples are provided. The Theorem 11 shows the properties of the dual solutions that correspond to a primal solution satisfying a certain condition.

Theorem 11. *In the dual problem of the BOSS with $\mathscr{I}_{DOM}$, the dual solutions should have $y_{k^+}^* = 0$ or $y_{k^-}^* = 0$ for each $k \in \mathcal{P}$ for k such that $w_k^* > 0$. That is, $y_{k^+}^* \cdot y_{k^-}^* = 0$ for $k \in \{1, 2, \cdots, K\}$ such that $w_k^* > 0$.*

Proof. Suppose that the optimal objective value of the primal problem is 0. Then an optimal dual solution is given by $y_{1^+}^* = y_{2^+}^* = \cdots = y_{K^+}^* = y_{1^-}^* = y_{2^-}^* = \cdots = y_{k^-}^* = 0$ and thus it satisfy the statement that $y_{k^+}^* = 0$ or $y_{k^-}^* = 0$ for each $k \in \mathcal{P}$. In fact in this case there is no $k \in \mathcal{P}$ such that $w_k^* > 0$ since $w_k^* = 0\ \forall k \in \mathcal{P}$.

Now assume that the optimal objective value of the primal problem is positive. Suppose that the first and second types of constraints hold with strict inequalities for a given $k \in \mathcal{P}$: That is,

$$\frac{1}{s}\sum_{c\in C} v_c^* X_{c,k} - w_k^* < \overline{\mu}_{T,k} \tag{6.11}$$

and

$$-\frac{1}{s}\sum_{c\in C} v_c^* X_{c,k} - w_k^* < -\overline{\mu}_{T,k} \tag{6.12}$$

This contradicts that $(v_{c_1}^*, v_{c_2}^*, \cdots, v_{c_{|C|}}^*, w_1^*, w_2^*, \cdots, w_K^*)$ are optimal solution of the primal problem because one can decrease the optimal value by replacing w_k^* which is greater than $\left|\frac{1}{s}\sum_{c\in C} v_c^* X_{c,k} - \overline{\mu}_{T,k}\right|$ by $\widetilde{w}_k^* = \left|\frac{1}{s}\sum_{c\in C} v_c^* X_{c,k} - \overline{\mu}_{T,k}\right|$ since the set of values $(v_{c_1}^*, v_{c_2}^*, \cdots, y_{c_{|C|}}^*, w_1^*, w_2^*, \cdots, w_{k-1}^*, \widetilde{w}_k^*, w_{k+1}^*, \cdots w_K^*)$ will still satisfy all the primal

constraints while giving smaller objective value:

$$w_1 + w_2 + \cdots + w_K \geq w_1^* + w_2^* + \cdots + w_{k-1}^* + \widetilde{w}_k^* + w_{k+1}^* + \cdots w_K^*. \tag{6.13}$$

Hence, at least one of the first and second types of constraints should hold with equality. The above argument holds for all $k \in \mathcal{P}$. As a result,

$$\frac{1}{s}\sum_{c\in C} v_c^* X_{c,k} - w_k^* = \overline{\mu}_{T,k} \tag{6.14}$$

or

$$-\frac{1}{s}\sum_{c\in C} v_c^* X_{c,k} - w_k^* = -\overline{\mu}_{T,k} \tag{6.15}$$

for all $k \in \mathcal{P}$.

Consider k such that $w_k^* > 0$. Since it was assumed that the optimal objective value of the primal problem is positive, there exist k such that $w_k^* > 0$ among $k \in \mathcal{P}$.

If both (6.14) and (6.15) hold, then

$$w_k^* = \frac{1}{s}\sum_{c\in C} v_c^* X_{c,k} - \overline{\mu}_{T,k} = -\frac{1}{s}\sum_{c\in C} v_c^* X_{c,k} + \overline{\mu}_{T,k} = \left|\frac{1}{s}\sum_{c\in C} v_c^* X_{c,k} - \overline{\mu}_{T,k}\right| = 0 \tag{6.16}$$

Hence it cannot be the case for k such that $w_k^* > 0$.

In other words, only one of (6.14) and (6.15) holds with equality and the other one holds with strict inequality. If (6.14) is the constraint with strict inequality, then by complementary slackness condition, $y_{k^+}^* = 0$. If (6.15) is the constraint with strict inequality, then again by complementary slackness condition, $y_{k^-}^* = 0$. □

The previous result can be demonstrated with an example. Consider a primal problem stated in (6.3) with one unit in the treatment group (namely, $T = \{t_1\}$; $|T| = 1$) two units in the control pool ($|C| = 2$; $C = \{c_1, c_2\}$) and two covariates for each unit ($K = 2$). Then the primal problem becomes (6.17).

$$
\begin{aligned}
\min \quad & \sum_{k\in\mathcal{P}} w_k \\
\text{s.t.} \quad & \tfrac{1}{s}\left(v_{c_1}X_{c_1,1} + v_{c_2}X_{c_2,1}\right) - w_1 \leq X_{t_1,1} \\
& \tfrac{1}{s}\left(v_{c_1}X_{c_1,2} + v_{c_2}X_{c_2,2}\right) - w_2 \leq X_{t_1,2} \\
& -\tfrac{1}{s}\left(v_{c_1}X_{c_1,1} + v_{c_2}X_{c_2,1}\right) - w_1 \leq -X_{t_1,1} \\
& -\tfrac{1}{s}\left(v_{c_1}X_{c_1,2} + v_{c_2}X_{c_2,2}\right) - w_2 \leq -X_{t_1,2} \\
& v_{c_1} + v_{c_2} = s \\
& v_{c_1}, v_{c_2} \geq 0 \\
& w_1, w_2 \geq 0.
\end{aligned}
\tag{6.17}
$$

Let the covariate values for each unit be $\mathbf{X}_{t_1} = (1,1), \mathbf{X}_{c_1} = (1,2), \mathbf{X}_{c_2} = (1,3)$ and the constant s be given by $s = 1$. Then the primal LP can be written as (6.18).

$$
\begin{aligned}
\min \quad & w_1 + w_2 \\
\text{s.t.} \quad & v_{c_1} + v_{c_2} - w_1 \leq 1 \\
& 2v_{c_1} + 3v_{c_2} - w_2 \leq 1 \\
& -v_{c_1} - v_{c_2} - w_1 \leq -1 \\
& -2v_{c_1} - 3v_{c_2} - w_2 \leq -1 \\
& v_{c_1} + v_{c_2} = 1 \\
& v_{c_1}, v_{c_2}, w_1, w_2 \geq 0.
\end{aligned}
\tag{6.18}
$$

Note that the optimal solution of the primal problem in (6.18) is $\left(v^*_{c_1}, v^*_{c_2}, w^*_1, w^*_2\right) = (1,0,0,1)$ with objective value $V^* = 1$. The dual of the above primal problem in (6.18) is given by

$$
\begin{aligned}
\min \quad & y_{1^+} + y_{2^+} - y_{1^-} - y_{2^-} + y_s \\
\text{s.t.} \quad & y_{1^+} + 2y_{2^-} - y_{1^-} - 2y_{2^-} + y_s \leq 0 \\
& y_{1^+} + 3y_{2^+} - y_{1^-} - 3y_{2^-} + y_s \leq 1 \\
& -y_{1^+} - y_{1^-} \leq 1 \\
& -y_{2^+} - y_{2^-} \leq 1 \\
& y_{1^+}, y_{2^+}, y_{1^-}, y_{2^-} \leq 0 \\
& y_s \quad \text{free.}
\end{aligned}
\tag{6.19}
$$

Solving the dual problem gives $\left(y^*_{1^+}, y^*_{2^+}, y^*_{1^-}, y^*_{2^-}, y^*_s\right) = (0,-1,-1,0,1)$. One can check the result of Theorem 11 that $y^*_{k^+} \cdot y^*_{k^-} = 0$ holds for $k = 2$ $\left(\text{i.e., } y^*_{2^+} \cdot y^*_{2^-} = 0\right)$ where $w^*_2 > 0$. Here, $y^*_{1^+} \cdot y^*_{1^-} = 0$ also holds but only k values such that $w_k > 0$ are of interest since, for $\widetilde{k}$ such that $w_{\widetilde{k}} = 0$, the $\widetilde{k}$-th covariate is already balanced.

How the optimal value is affected by the change of the covariates that are not

Table 6.1: Changing the RHS of the First Constraint in (6.18)

$\gamma_1 = \text{RHS} - \text{RHS}_0$	RHS	$v_{c_1}^*$	$v_{c_2}^*$	w_1^*	w_2^*	V^*	$\Delta V^* = V^* - V_0^*$
-3	-2	1	0	3	1	4	3
-2	-1	1	0	2	1	3	2
-1	0	1	0	1	1	2	1
0	$\text{RHS}_0 = 1$	1	0	0	1	$V_0^* = 1$	$0 = \gamma_1 \cdot 0$
1	2	1	0	0	1	1	$0 = \gamma_1 \cdot 0$
2	3	1	0	0	1	1	$0 = \gamma_1 \cdot 0$
3	4	1	0	0	1	1	$0 = \gamma_1 \cdot 0$

balanced will be assessed. Specifically, how the optimal value changes as the right-hand side of the first $2K$ constraints which is equal to $\overline{\mu}_{T,k}$ for the first K constraints and $-\overline{\mu}_{T,k}$ for the following K constraints in the general primal problem given in (6.3) change will be investigated. These values can be explained by using the dual variables for an interval of perturbation values that are sufficiently small. Such an interval will be explained precisely in Theorem 13.

Before stating the theorem, revisit the example given above. With the primal LP in (6.18), the sensitivity analysis of the optimal values can be conducted on the change of the right-hand side values. The following four tables show how the objective value changes as γ_i values added to the i-th constraint changes while keeping the other coefficients and the right-hand side values.

To compute the values in Table 6.1, consider the modified LP given in (6.20):

$$\begin{aligned}
\min \quad & w_1 + w_2 \\
\text{s.t.} \quad & v_{c_1} + v_{c_2} - w_1 \leq 1 + \gamma_1 \\
& 2v_{c_1} + 3v_{c_2} - w_2 \leq 1 \\
& -v_{c_1} - v_{c_2} - w_1 \leq -1 \\
& -2v_{c_1} - 3v_{c_2} - w_2 \leq -1 \\
& v_{c_1} + v_{c_2} = 1 \\
& v_{c_1}, v_{c_2}, w_1, w_2 \geq 0.
\end{aligned} \tag{6.20}$$

Similarly, one can construct the modified problem when changing the RHS value of the second, third, and fourth constraints respectively and get the following Tables 6.2 to 6.4.

Recall that $\left(y_{1^+}^*, y_{2^+}^*, y_{1^-}^*, y_{2^-}^*, y_s^*\right)$ is given by $(0, -1, -1, 0, 1)$. Additionally note that in Table 6.1,

$$\Delta V^* = \gamma_1 \cdot y_{1^+}^* \text{ for } \gamma_1 \geq 0. \tag{6.21}$$

Table 6.2: Changing the RHS of the Second Constraint in (6.18)

$\gamma_2 = \text{RHS} - \text{RHS}_0$	RHS	$v_{c_1}^*$	$v_{c_2}^*$	w_1^*	w_2^*	V^*	$\Delta V^* = V^* - V_0^*$
-3	-2	1	0	0	4	4	$3 = \gamma_2 \cdot (-1)$
-2	-1	1	0	0	3	3	$2 = \gamma_2 \cdot (-1)$
-1	0	1	0	0	2	2	$1 = \gamma_2 \cdot (-1)$
0	$\text{RHS}_0 = 1$	1	0	0	1	$V_0^* = 1$	$0 = \gamma_2 \cdot (-1)$
1	2	1	0	0	0	0	$-1 = \gamma_2 \cdot (-1)$
2	3	1	0	0	0	0	-1
3	4	1	0	0	0	0	-1

Table 6.3: Changing the RHS of the Third Constraint in (6.18)

$\gamma_3 = \text{RHS} - \text{RHS}_0$	RHS	$v_{c_1}^*$	$v_{c_2}^*$	w_1^*	w_2^*	V^*	$\Delta V^* = V^* - V_0^*$
-3	-4	1	0	3	1	4	$3 = \gamma_3 \cdot (-1)$
-2	-3	1	0	2	1	3	$2 = \gamma_3 \cdot (-1)$
-1	-2	1	0	1	1	2	$1 = \gamma_3 \cdot (-1)$
0	$\text{RHS}_0 = -1$	1	0	0	1	$V_0^* = 1$	$0 = \gamma_3 \cdot (-1)$
1	0	1	0	0	1	0	0
2	1	1	0	0	1	0	0
3	2	1	0	0	1	0	0

In Table 6.2,

$$\Delta V^* = \gamma_2 \cdot y_{2^+}^* \text{ for } \gamma_2 \leq 1. \tag{6.22}$$

In Table 6.3,

$$\Delta V^* = \gamma_3 \cdot y_{1^-}^* \text{ for } \gamma_3 \leq 0. \tag{6.23}$$

In Table 6.4,

$$\Delta V^* = \gamma_4 \cdot y_{2^-}^* \text{ for } \gamma_4 \geq -2. \tag{6.24}$$

These range of values can be computed as follows. First convert the LP in

Table 6.4: Changing the RHS of the Fourth Constraint in (6.18)

$\gamma_3 = \text{RHS} - \text{RHS}_0$	RHS	$v_{c_1}^*$	$v_{c_2}^*$	w_1^*	w_2^*	V^*	$\Delta V^* = V^* - V_0^*$
-3	-4	0.5	0.5	0	1.5	1.5	0.5
-2	-3	1	0	0	1	1	$0 = \gamma_4 \cdot 0$
-1	-2	1	0	0	1	1	$0 = \gamma_4 \cdot 0$
0	$\text{RHS}_0 = -1$	1	0	0	1	$V_0^* = 1$	$0 = \gamma_4 \cdot 0$
1	0	1	0	0	1	1	$0 = \gamma_4 \cdot 0$
2	1	1	0	0	1	1	$0 = \gamma_4 \cdot 0$
3	2	1	0	0	1	1	$0 = \gamma_4 \cdot 0$

(6.18) into a standard form as in (6.25).

$$\begin{array}{lrl} \min & w_1 + w_2 & \\ \text{s.t.} & v_{c_1} + v_{c_2} - w_1 + z_1 & = 1 \\ & 2v_{c_1} + 3v_{c_2} - w_2 + z_2 & = 1 \\ & -v_{c_1} - v_{c_2} - w_1 + z_3 & = -1 \\ & -2v_{c_1} - 3v_{c_2} - w_2 + z_4 & = -1 \\ & v_{c_1} + v_{c_2} & = 1 \\ & v_{c_1}, v_{c_2}, w_1, w_2, z_1, x_2, x_3, x_4 & \geq 0. \end{array} \tag{6.25}$$

Two-phase simplex method will be applied to find the optimal solution and basis in the optimal tableau of the problem. To find the initial basic feasible solution, construct the following auxiliary problem in (6.26).

$$\begin{array}{lrl} \min & z_5 + z_6 + z_7 & \\ \text{s.t.} & v_{c_1} + v_{c_2} - w_1 + z_1 & = 1 \\ & 2v_{c_1} + 3v_{c_2} - w_2 + z_2 & = 1 \\ & v_{c_1} + v_{c_2} + w_1 - z_3 + z_7 & = 1 \\ & 2v_{c_1} + 3v_{c_2} + w_2 - z_4 + z_6 & = 1 \\ & v_{c_1} + v_{c_2} + z_5 & = 1 \\ & z_1, z_2, \cdots, z_7 & \geq 0. \end{array} \tag{6.26}$$

This auxiliary problem has the optimal solution with $z_5 = z_6 = z_7 = 0$ with optimal value 0 and hence one can start the Phase 2 using the feasible solution obtained at the end of Phase 1. The simplex method in Phase 2 gives the final tableau with $(z_1, z_4, w_1, v_{c_1}, w_2) = (0, 2, 0, 1, 1)$ as basic variable and $(v_{c_2}, z_2, z_3) = (0, 0, 0)$ as non-basic variable. The basis matrix B is given by

$$B = \begin{bmatrix} 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 2 & -1 \\ 0 & 0 & -1 & -1 & 0 \\ 0 & 1 & 0 & -2 & -1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{6.27}$$

Theorem 12. *(p.207 of Bertsimas and Tsitsiklis (1997))Consider an LP of the*

form

$$\begin{aligned} max \quad & \boldsymbol{c}^T\boldsymbol{x} + \boldsymbol{0}^T\boldsymbol{x}_s \\ s.t. \quad & \begin{bmatrix} A & I_m \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{x}_s \end{bmatrix} = \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0} \\ & \boldsymbol{x}_s \geq \boldsymbol{0}. \end{aligned} \tag{6.28}$$

with an m×n matrix A, an m×m identity matrix I_m, an n×1 vector $\boldsymbol{x}$, an m×1 vector $\boldsymbol{x}_s$ and an $m \times 1$ vector $\boldsymbol{b}$. Then, given that the optimal basis B doesn't change for a small enough perturbation, any constraint has its shadow price which is equal to the optimal value of the dual variable that corresponds to the constraint. Furthermore, the tolerance interval of perturbation factor γ_k for the k-th constraint can be computed using the inequalities from the optimal tableau:

$$B^{-1}\boldsymbol{b} + \gamma_k B^{-1} e_k \geq \boldsymbol{0} \tag{6.29}$$

where B is an optimal basis and $e_k \in \mathbb{R}^m$ is the k-th unit vector.

Consider applying the perturbation to first constraint in the standard form which was found in (6.25) as in (6.30).

$$\begin{aligned} \min \quad & w_1 + w_2 \\ \text{s.t.} \quad & v_{c_1} + v_{c_2} - w_1 + z_1 = 1 + \gamma_1 \\ & 2v_{c_1} + 3v_{c_2} - w_2 + z_2 = 1 \\ & -v_{c_1} - v_{c_2} - w_1 + z_3 = -1 \\ & -2v_{c_1} - 3v_{c_2} - w_2 + z_4 = -1 \\ & v_{c_1} + v_{c_2} = 1 \\ & v_{c_1}, v_{c_2}, w_1, w_2, z_1, x_2, x_3, x_4 \geq 0. \end{aligned} \tag{6.30}$$

Then for γ satisfying the

$$B^{-1}\mathbf{b} + \gamma_1 B^{-1} e_1 = \begin{bmatrix} 1 & 0 & -1 & 0 & -2 \\ 0 & -1 & 0 & 1 & 4 \\ 0 & 0 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 1 \end{bmatrix} + \gamma_1 \begin{bmatrix} 1 & 0 & -1 & 0 & -2 \\ 0 & -1 & 0 & 1 & 4 \\ 0 & 0 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$
$$= \begin{bmatrix} 0 + \gamma_1 \cdot 1 \\ 2 + \gamma_1 \cdot 0 \\ 0 + \gamma_1 \cdot 0 \\ 1 + \gamma_1 \cdot 0 \\ 1 + \gamma_1 \cdot 0 \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{6.31}$$

yields the tolerance interval $\gamma_1 \geq 0$ coinciding with (6.21) reported in the table. Similarly, the following inequalities hold:

$$B^{-1}\mathbf{b} + \gamma_2 B^{-1} e_2 = \begin{bmatrix} 0 + \gamma_2 \cdot 0 \\ 2 + \gamma_2 \cdot (-1) \\ 0 + \gamma_2 \cdot 0 \\ 1 + \gamma_2 \cdot 0 \\ 1 + \gamma_2 \cdot (-1) \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{6.32}$$

$$B^{-1}\mathbf{b} + \gamma_3 B^{-1} e_3 = \begin{bmatrix} 0 + \gamma_3 \cdot (-1) \\ 2 + \gamma_3 \cdot 0 \\ 0 + \gamma_3 \cdot (-1) \\ 1 + \gamma_3 \cdot 0 \\ 1 + \gamma_3 \cdot 0 \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{6.33}$$

$$B^{-1}\mathbf{b} + \gamma_4 B^{-1} e_4 = \begin{bmatrix} 0 + \gamma_4 \cdot 0 \\ 2 + \gamma_4 \cdot 1 \\ 0 + \gamma_4 \cdot 0 \\ 1 + \gamma_4 \cdot 0 \\ 1 + \gamma_4 \cdot 0 \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{6.34}$$

For the perturbation in the second, the third, and the fourth constraint, the tolerance interval that the optimal basis doesn't change is given by $\gamma_2 \leq 1$, $\gamma_3 \leq 0$, and $\gamma_4 \geq -2$ respectively as in (6.22), (6.23), and (6.24).

Theorem 13. *The following relationship holds between the means of treated units' k-th covariate values the corresponding dual variables* (y_{k^+}, y_{k^-}) *for* $k \in \mathcal{P}$ *such that* $w_k > 0$*: When* $\overline{\mu}_{T,k}$ *increases by* $\widetilde{\gamma}_k$*, for* $\widetilde{\gamma}_k$ *that lies in the tolerance interval, the optimal objective value of the primal problem increases by* $\widetilde{\gamma}_k(y^*_{k^+} - y^*_{k^-})$*. In addition, the tolerance interval for* γ_k *is given by set of inequalities given in* (6.35) *and* (6.36)*.*

$$B^{-1}b + \widetilde{\gamma}_k B^{-1}e_k \geq \boldsymbol{0} \tag{6.35}$$

$$B^{-1}b - \widetilde{\gamma}_k B^{-1}e_{k+K} \geq \boldsymbol{0} \tag{6.36}$$

Proof. From Theorem 12, note that y_{k^+} for $k \in \mathcal{P}$ are shadow prices for the k-th constraint where $\gamma_k \cdot y_k$ is an increment in the objective function value given a relaxation of a corresponding primal constraint by γ_k within the tolerance interval. Likewise, y_{k^-} for $k \in \mathcal{P}$ are shadow prices corresponding to $(k + K)$-th constraint of the primal problem and hence the objective value increases by $\gamma_{k+K} \cdot y_{k^-}$ when increasing the right-hand side of the $(k + K)$-th constraint of the primal problem by γ_{k+K} where γ_{k+K} likes within the tolerance interval. Additionally, the tolerance interval is given by the following inequalities:

$$B^{-1}b + \gamma_k B^{-1}e_k \geq \mathbf{0} \tag{6.37}$$

and

$$B^{-1}b + \gamma_{k+K} B^{-1}e_{k+K} \geq \mathbf{0}. \tag{6.38}$$

Note that increasing $\overline{\mu}_{T,k}$ by $\widetilde{\gamma}_k$ is equivalent to increasing the RHS of the k-th constraint by $\widetilde{\gamma}_k$ and decreasing the RHS of the $(k + K)$-th constraint by $\widetilde{\gamma}_k$ (i.e., $\gamma_k = \widetilde{\gamma}_k$ and $\gamma_{k+K} = -\widetilde{\gamma}_k$). Hence the tolerance interval is given by combining (6.35) and (6.36) together.

Furthermore, note that while the increment of the RHS of the k-th constraint and the decrement of the RHS of the $(k + K)$-th constraint by the same amount occurs simultaneously, one can consider them sequentially. Furthermore, recall that from Theorem 11, at least one of $y^*_{k^+}$ and $y^*_{k^-}$ for $k \in \mathcal{P}$ such that $w_k > 0$ is zero.

Suppose that $y^*_{k^+} = 0$. Then apply the increment by γ_k in RHS of the k-th constraint first. Since the optimal basis doesn't change with the change by amount of $\widetilde{\gamma}$ within the tolerance interval, the value of the dual variable doesn't change and the change in objective value is $0 = \widetilde{\gamma}_k y^*_{k^+}$. Hence, in the second step, the increment in the objective value by the decrement by $\widetilde{\gamma}_k$ in RHS of the $(k + K)$-th

constraint is still given by $-\widetilde{\gamma}_k y^*_{k^-}$ where $y^*_{k^-}$ is the optimal value of the $(k+K)$-th dual variable of the original problem.

Similarly, if $y^*_{k^-} = 0$, then apply change in RHS to the $(k+K)$-th constraint first. Within the tolerance level, the optimal basis remains the same and the optimal dual solution doesn't change. Hence, in the second step when applying the increment by $\widetilde{\gamma}_k$ to RHS to the k-th constraint the objective value increases by $\widetilde{\gamma}_k y^*_k$.

As a result, in both cases, the optimal objective value increases by $\widetilde{\gamma}_k(y^*_{k^+} - y^*_{k^-})$ when increasing the value of $\overline{\mu}_{T,k}$ by $\widetilde{\gamma}_k$ for $\widetilde{\gamma}_k$ within the tolerance interval given above. □

In the above example in (6.17), if the $X_{t_1,2}$ changes from 1 to $1 + \widetilde{\gamma}_2$, then the optimal objective value will change by $\widetilde{\gamma}_2 y^*_2 - \widetilde{\gamma}_2 y^*_4 = \widetilde{\gamma}_2 \cdot (-1) - \widetilde{\gamma}_2 \cdot = 0$ for $\widetilde{\gamma}_2$ satisfying $\widetilde{\gamma}_2 \leq 1$ since

$$\left.\begin{aligned} \gamma_2 = \widetilde{\gamma}_2 \leq 1 \\ \gamma_4 = -\widetilde{\gamma}_2 \geq -2 \end{aligned}\right\} \Rightarrow \widetilde{\gamma}_2 \leq 1. \tag{6.39}$$

Indeed one can check that the optimal objective value decreases by 1 from 1 to 0 when $X_{t_1,2}$ increases by 1 from 1 to 2 (and the optimal objective value increases by 1 from 1 to 2 when decreasing $X_{t_1,2}$ by 1 from 1 to 0.)

Furthermore, more things can be said about dual solution of those variables y^*_k and y^*_{k+K} for $k \in \mathcal{P}$ such that $w^*_k > 0$.

Theorem 14. *In dual problems of the BOSS, if $k \in \mathcal{P}$ is a covariate index such that $w^*_k > 0$, then the dual solutions should satisfy $(y^*_{k^+}, y^*_{k^-}) = (-1, 0)$ or $(y^*_{k^+}, y^*_{k^-}) = (0, -1)$.*

Proof. Suppose that $k \in \mathcal{P}$ is a covariate index such that $w^*_k > 0$. Then from the complementary slackness condition,

$$-y_{k^+} - y_{k^-} = 1 \tag{6.40}$$

for such k.

Recall that $y^*_{k^+} = 0$ or $y^*_{k^-} = 0$ for each $k \in \mathcal{P}$ for k such that $w^*_k > 0$ from Theorem 11. If $y^*_{k^+} = 0$, then $y^*_{k+^-} = -1$ from (6.40). Similarly, if $y^*_{k^-} = 0$, then $y^*_{k^+} = -1$ from (6.40). Hence, the dual solution should satisfy $(y^*_{k^+}, y^*_{k^-}) = (-1, 0)$ or $(y^*_{k^+}, y^*_{k^-}) = (0, -1)$. □

Again, recall the example in (6.18). It was discussed that the dual solution is given by $\left(y^*_{1^+}, y^*_{2^+}, y^*_{1^-}, y^*_{2^-}, y^*_s\right) = (0, -1, -1, 0, 1)$ where the primal solution of the problem is given by $\left(v^*_{c_1}, v^*_{c_2}, w^*_1, w^*_2\right) = (1, 0, 0, 1)$. Note that $w^*_2 > 0$ and the corresponding dual variables $(y^*_{2^+}, y^*_{2^-})$ have the values $(-1, 0)$ as stated in Theorem 14.

Note that the condition of w^*_k being greater than zero is necessary in the above theorem. While $(y^*_{1^+}, y^*_{1^-}) = (0, -1)$ with $k = 1, K = 2$, and $w^*_{1^+} = 0$ holds in the example of (6.18), this is a special case that is happened by coincidence. In general, $(y^*_{k^+}, y^*_{k^-})$ is neither $(-1, 0)$ nor $(0, -1)$ for $w^*_k = 0$. See the following primal problem (6.41) and its dual problem (6.42) for an example:

Primal Problem

$$\begin{aligned}
\min \quad & w_1 + w_2 \\
\text{s.t.} \quad & 2v_{c_1} + 0v_{c_2} - w_1 \leq 1 \\
& 0v_{c_1} + 1v_{c_2} - w_2 \leq 1 \\
& -2v_{c_1} - 0v_{c_2} - w_1 \leq -1 \\
& -0v_{c_1} - 1v_{c_2} - w_2 \leq -1 \\
& v_{c_1} + v_{c_2} = 1 \\
& v_{c_1}, v_{c_2}, w_1, w_2 \geq 0.
\end{aligned} \tag{6.41}$$

Dual Problem

$$\begin{aligned}
\min \quad & y_{1^+} + y_{2^+} - y_{1^-} - y_{2^-} + y_s \\
\text{s.t.} \quad & 2y_{1^+} + 0y_{2^+} - 2y_{1^-} - 0y_{2^-} + y_s \leq 0 \\
& 0y_{1^+} + 1y_{2^+} - 0y_{1^-} - 1y_{2^-} + y_s \leq 1 \\
& -y_{1^+} - y_{1^-} \leq 1 \\
& -y_{2^+} - y_{2^-} \leq 1 \\
& y_{1^+}, y_{2^+}, y_{1^-}, y_{2^-} \leq 0 \\
& y_s \quad \text{free.}
\end{aligned} \tag{6.42}$$

In this example which is obtained by plugging $\mathbf{X}_{t_1} = (1, 1), \mathbf{X}_{c_1} = (2, 0), \mathbf{X}_{c_2} = (0, 1)$ in problem stated in (6.17), the primal solution is given by $\left(v^*_{c_1}, v^*_{c_2}, w^*_1, w^*_2\right) = (0.5, 0.5, 0, 0.5)$ and the dual solution is given by $\left(y^*_{1^+}, y^*_{2^+}, y^*_{1^-}, y^*_{2^-}, y^*_s\right) = (0, 0, -0.5, -1, -1)$. Note that, while $(y^*_{2^+}, y^*_{2^-}) = (0, -1)$ as $w^*_2 = 0.5 > 0$, the value of w^*_1 is equal to 0 and the first and third dual variables are given by $(y^*_{1^+}, y^*_{1^-}) = (0, -0.5)$ which is not $(-1, 0)$ nor $(0, -1)$.

Additionally, the following result immediately follows from the previous result.

Theorem 15. *If the dual optimal solution satisfy $y^*_{k^+} \neq -1$ and $y^*_{k^-} \neq -1$, then the*

corresponding primal optimal solution satisfy $w_k^* = 0$ *meaning that k-th covariate is balanced with the current choice of control group.*

Proof. This follows from previous theorem (Theorem (14)) as k-th and $(k + K)$-th constraints are binding with $w_k^* = 0$. □

6.4 Concluding Remarks

In this chapter, the LP formulation of a particular BOSS problem, its dual problems, and the properties of the primal and dual solutions are studied. Note that the number of dual variables $(2K + 1)$ is is typically much smaller than the number of primal variables $(|C| + K)$ and the dual variables corresponding to $w_k^* > 0$ would be either $(y_{k^+}, y_{k^-}) = (-1, 0)$ or $(y_{k^+}, y_{k^-}) = (0, -1)$. Furthermore, values of the dual solutions give insights on which covariates are balanced with the current optimal control group and which are not. In addition, the optimal objective value will change as the covariate values of the given units change and dual solution gives information on how much change will be made through the adjustment .

The discussion of the chapter was made based on BOSS with the difference of means imbalance measure $\mathscr{I}_{\mathrm{DOM}}$, which only involves the first order terms of the covariates. The discussion can be extended into other LP formulation of BOSS having higher order terms in covariates. This extension can be made since, regardless of which form the polynomials of covariates in imbalance measure take, they are regarded as a constant once the covariate values of the control units and the treatment units are given.

CHAPTER 7

CONCLUSION

This dissertation has made five major extensions to the Balance Optimization Subset Selection (BOSS) framework. Main objective of BOSS is to eliminate or reduce selection bias that can arise from imbalance between the treated and the control by directly solving a computational optimization problem and finding a set of control units which minimizes the imbalance defined by a researcher. In the first part (Chapter 2), cases that may lead to bias and examples for those cases were provided. In the chapter, balance hierarchy and a correct imbalance measure which corresponds to the form of the response functions are defined. Additionally, new imbalance measures drawn from the Cramer-von Mises test statistic were introduced. The cases of insufficient data and suboptimality that can happen in causal analysis with BOSS were also presented.

The second part (Chapter 3) of this dissertation discussed how we can decompose a treatment effect estimate as a combination of heterogeneous treatment effects from a partitioned set. The method introduced in the chapter is different from the traditional propensity score subclassification method in that a subset is found in each subclass of the control pool using BOSS instead of using the stratum determined by the propensity score. Then, by conducting a bootstrap hypothesis test on each component, a statistical significance of these heterogeneous treatment effects are examined. These methods were applied to a dataset from the National Supported Work Demonstration (NSW) program which was conducted in the 1970s. By examining the statistical significance, it was shown that the program was not significantly effective to a specific subgroup composed of those who were already employed.

In the third part (Chapter 4), the BOSS framework was extended to a non-binary treatment (i.e., multi-treatment) setting. A treatment effect estimator under a multi-treatment setting was proposed and it was shown that this estimator is unbiased when there is no residual imbalance under a weak ignorability assumption. How the estimate can be computed by combining estimates obtained from BOSS

with binary treatments were explained and the BOSS estimates are computationally compared to those obtained by matching. In the example with simulated dataset, the BOSS estimator showed comparable results to matching estimators in terms of the size of bias which is smaller than the bias of estimates obtained through random selection of control groups.

The fourth part (Chapter 5) handled cases where there are missing covariates in dataset. Sensitivity of BOSS estimators were examined by generating a previously unobserved covariates with various parameter values and it was compared with matching estimators' sensitivity. Furthermore, two methods that can be applied after imputing the missing entries were discussed. In the examples that were discussed in the chapter, the BOSS methods on multiply imputed data had smaller bias than the corresponding methods in matching.

Lastly, in the fifth part (Chapter 6), a dual problem of BOSS and its solution are investigated. Most BOSS problems can be formulated as mixed integer linear programs. In the chapter, BOSS was formulated as a linear program by relaxing the integrality condition on contribution of control units in the optimal control group and a dual problem of the LP was found. After investigating the relationship between the primal and dual solutions of BOSS, it was discussed how the dual solution provides information on changes in the objective value under perturbations of the covariate values by building on top of standard duality results.

One thing to note is that a large portion of discussions in this dissertation were made by using the difference of means imbalance measure, $\mathscr{I}_{\text{DOM}}$. As mentioned, many discussions such as the treatment effect decomposition technique, the unbiasedness theorems in a multi-treatment setting, the methods of handling missing data, and the duality results can be easily extended to other imbalance measures containing higher order terms. However, one should be aware when extending the earlier results to imbalance measures such as $\mathscr{I}_{\text{KS}}$ and $\mathscr{I}_{\text{CvM}}$ that are not moment-based forms since these results may not be directly applicable. For example, the discussions on duality results are limited to the moment-based imbalance measures having the constraints of the same form as the difference of means imbalance measure.

There are many possible directions for future research. One of the routes is a further investigation on how regression and BOSS can be used together in a synergistic manner. The relationship between regression and BOSS is similar to the relationship between regression and matching: Two methods are compensating each other rather than competing. It has been shown that a better performance can

be expected when matching and regression are used together than the case when the either method is used alone (Rubin, 1979; Rubin and Thomas, 2000). It is also well discussed in a survey paper of matching method by Stuart (2010).

BOSS method can be used first to find optimally balanced groups of units with an imbalance measure that uses approximation of the response functions and then regression can be conducted using those two groups to find a more accurate polynomial approximation of those response functions. To improve the result, these methods can be used one after the other in succession. Actually implementing these ideas to use regression and BOSS together on simulated or real-world data is a possible future research direction.

Another possible direction is to extend the BOSS framework to a continuous treatment setting as mentioned at the end of Chapter 4. Matching methods were extended so that it can handle continuous treatment (Hirano and Imbens, 2004). A similar development is needed in BOSS framework as well so that the BOSS method can be applied directly to such a dataset with continuous treatment without using the provisional way of making the continuous treatment into a discrete multiple treatment levels.

As seen in the dissertation, the BOSS methods has a potential to be performed better than the matching method that it often has smaller size of bias in many applications while it does not require individual unit matches. Guaranteed unbiasedness of BOSS estimators given less strict conditions than exact individual unit matches is a huge advantage. However, BOSS also has a disadvantage compared to matching that it generally takes a longer computational time. In addition, sometimes there may be circumstances that having exact unit matches with the same or close covariates is needed as it can convey important qualitative information. Note that matching can be incorporated into BOSS method as discussed in Sauppe et al. (2014). One method may provide more benefits over the other under specific requirement and they may be also used together for various purposes. Hence, understanding the comparative advantage of these methods under different situations is important and one should decide appropriate estimation method for treatment effects that meets one's needs.

REFERENCES

Abadie, A. and Imbens, G. W. (2012). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11.

Anderson, T. W. (1962). On the distribution of the two-sample cramer-von mises criterion. *The Annals of Mathematical Statistics*, 33(3):1148–1159.

Becker, S. O. and Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal*, 2(4):358–377.

Bertsimas, D., Johnson, M., and Kallus, N. (2015). The power of optimization over randomization in designing experiments involving small samples. *Operations Research*, 63(4):868–876.

Bertsimas, D. and Tsitsiklis, J. N. (1997). *Introduction to linear optimization*. Athena Scientific, Belmont, MA, U.S.A.

Brumback, B. A., Hernán, M. A., Haneuse, S. J., and Robins, J. M. (2004). Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine*, 23(5):749–767.

Cho, W. K. T., Sauppe, J. J., Nikolaev, A. G., Jacobson, S. H., and Sewell, E. C. (2013). An optimization approach for making causal inferences. *Statistica Neerlandica*, 67(2):211–226.

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

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446.

Colson, K. E., Rudolph, K. E., Zimmerman, S. C., Goin, D. E., Stuart, E. A., van der Laan, M., and Ahern, J. (2016). Optimizing matching and analysis combinations for estimating causal effects. *Scientific Reports*, 6(23222).

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

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

Diamond, A. and Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945.

Elwert, F. and Winship, C. (2010). Effect heterogeneity and bias in main-effects-only regression models. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, pages 327–36.

Foster, E. M. (2003). Propensity score matching: An illustrative analysis of dose response. *Medical care*, 41(10):1183–1192.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge, U.K.

Graham, J. W. (2012). *Missing Data : Analysis and Design*. Springer Science & Business Media, New York, NY, U.S.A.

Greenland, S. (2005). Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2):267–306.

Grender, J., Williams, K., Walters, P., Klukowska, M., and Reick, H. (2013). Plaque removal efficacy of oscillating-rotating power toothbrushes: Review of six comparative clinical trials. *American Journal of Dentistry*, 26(2):68–74.

Hannan, E. L. (2008). Randomized clinical trials and observational studies: Guidelines for assessing respective strengths and limitations. *JACC: Cardiovascular Interventions*, 1(3):211–217.

Hartz, A., Bentler, S., Charlton, M., Lanska, D., Butani, Y., Soomro, G. M., and Benson, K. (2005). Assessing observational studies of medical treatments. *Emerging themes in epidemiology*, 2(1):8.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66(5):1017–1098.

Heckman, J. J. and Hotz, V. J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association*, 84(408):862–874.

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. In Gelman, A. and Meng, X.-L., editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, chapter 7, pages 73–84. John Wiley & Sons Ltd, Chichester, West Sussex, England.

Ichino, A., Mealli, F., and Nannicini, T. (2008). From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *Journal of Applied Econometrics*, 23(3):305–327.

Imai, K., Ratkovic, M., et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, pages 126–132.

Jepsen, P., Johnsen, S. P., Gillman, M., and Sørensen, H. T. (2004). Interpretation of observational studies. *Heart*, 90(8):956–960.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, pages 604–620.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies*, pages 43–58. Springer, Berlin · Heidelberg, Germany.

Li, L., Shen, C., Wu, A. C., and Li, X. (2011). Propensity score-based sensitivity analysis method for uncontrolled confounding. *American Journal of Epidemiology*, page kwr096.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John John Wiley & Sons, Ltd., Hoboken, NJ, U.S.A., 2nd edition.

MacKinnon, J. G. (2009). Bootstrap hypothesis testing. In *Handbook of Computational Econometrics*, pages 183–213. John Wiley & Sons, Ltd., Chichester, United Kingdom.

Mitra, R. and Reiter, J. P. (2016). A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical Methods in Medical Research*, 25(1):188–204.

Nannicini, T. (2007). Simulation-based sensitivity analysis for matching estimators. *The Stata Journal*, 7(3):334.

Nikolaev, A. G., Jacobson, S. H., Cho, W. K. T., Sauppe, J. J., and Sewell, E. C. (2013). Balance optimization subset selection (BOSS): An alternative approach for causal inference with observational data. *Operations Research*, 61(2):398–412.

Port, F. K. (2000). Role of observational studies versus clinical trials in esrd research. *Kidney International*, 57:S3–S6.

Rosenbaum, P. R. and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 212–218.

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

Rosenbaum, P. R. and Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics*, 41(1):103–116.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 29(1):159–183.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, pages 318–328.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press, Cambridge, U.K.

Rubin, D. B. and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450):573–585.

Sauppe, J. J. (2015). *Balance Optimization Subset Selection: A Framework for Causal Inference with Observational Data*. PhD thesis, University of Illinois at Urbana-Champaign.

Sauppe, J. J. and Jacobson, S. H. (2017). The role of covariate balance in observational studies. *Naval Research Logistics (NRL)*, 64(4):323–344.

Sauppe, J. J., Jacobson, S. H., and Sewell, E. C. (2014). Complexity and approximation results for the balance optimization subset selection model for causal inference in observational studies. *INFORMS Journal on Computing*, 26(3):547–566.

Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science*, 12:487–508.

Sekhon, J. S. and Grieve, R. (2008). A new non-parametric matching method for bias adjustment with applications to economic evaluations. In *iHEA 2007 6th World Congress: Explorations in Health Economics Paper*.

Smith, J. A. and Todd, P. E. (2005). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1):305–353.

Stampf, S. (2014). *Propensity Score Based Data Analysis*. R package vignette, R package version 1.42.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25(1):1.

Wang, J., Donnan, P. T., Steinke, D., and MacDonald, T. M. (2001). The multiple propensity score for analysis of dose-response relationships in drug safety studies. *Pharmacoepidemiology and Drug safety*, 10(2):105–111.

Xie, Y., Brand, J. E., and Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological Methodology*, 42(1):314–347.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., and Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, 72(4):1055–1065.

Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500):1360–1371.