APPROVAL SHEET

Title of Dissertation: A SWAPPING METHOD AND EXPLORATORY ANALYSIS FOR AVERAGE TREATMENT EFFECT ESTIMATION BASED ON PARTIAL BALANCING AND SIMULTANEOUS INFERENCE OF REGRESSION MODELS

Name of Candidate: Rowena F. Bastero
Doctor of Philosophy, 2017

Dissertation and Abstract Approved: _____

Dr. Bimal K. Sinha
Professor
Department of Mathematics & Statistics

Date Approved: _____

ABSTRACT

Title of dissertation: A SWAPPING METHOD AND EXPLORATORY ANALYSIS FOR AVERAGE TREATMENT EFFECT ESTIMATION BASED ON PARTIAL BALANCING AND SIMULTANEOUS INFERENCE OF REGRESSION MODELS

Rowena F. Bastero, Doctor of Philosophy, 2017

Dissertation directed by: Dr. Bimal K. Sinha
Department of Mathematics and Statistics

To provide significant outcomes, it is imperative that health care professionals, medical practitioners and policy-makers acquire evidence of the effectiveness of different treatments or interventions. This is most commonly done by looking into treatment and control groups and determining if the treatment has a causal effect on the outcome. Ideally, treatment assignment is performed through randomization so that the groups formed are comparable with respect to their features; thus, randomized controlled trials are the gold standard. However, some factors, such as cost, time, and ethical issues behind the treatment, may make it difficult to assign treatments at random. This leads to the use of observational studies instead in assessing the treatment or intervention effect.

The lack of randomization can be an issue for both observational studies and clinical trial studies so that systematic differences in the covariates of the treatment and control groups may exist, which poses an inherent problem in estimating

average treatment effect. Current trends in data analysis utilize propensity score matching as a remedy to the imbalance among covariates between the treatment and control groups under comparison. However, assumed matched pairs or groups formed through propensity scores continue to reflect imbalance in the covariates between the two groups. Hence, an improved method is proposed, based on the direct classification of categorical covariates and balance between groups on continuous covariates, which consequently provides a more stable estimate of the average treatment effect. The proposed method begins with forming homogeneous subgroups in terms of qualitative features and then generates the estimates of average treatment effect using a method that infuses "swapping" of models based on classical regression and eventual combination of such estimates over all subgroups based on meta-analyses procedures. The "swapping" procedure allows for the imputation of the missing potential outcome for units in one of the control and treatment groups while meta-analysis provides some means of combining the effect sizes calculated from each matched group while addressing the issue of homogeneity. Simulation studies show that the proposed method is able to capture the true treatment effect and provide more stable estimates in comparison to standard propensity score measures.

Exploratory analysis via simultaneous regression inference is likewise presented to provide information on the magnitude of difference between the treatment and control regression models. The confidence bands generated through this analysis provide graphical representations of the average treatment effect.

A SWAPPING METHOD AND EXPLORATORY ANALYSIS FOR AVERAGE TREATMENT EFFECT ESTIMATION BASED ON PARTIAL BALANCING AND SIMULTANEOUS INFERENCE OF REGRESSION MODELS

by

Rowena F. Bastero

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland Baltimore County in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Dr. Bimal K. Sinha, Chair/Advisor
Dr. Chuanhua Xing, Co-Chair/Co-Advisor
Dr. Thomas Mathew
Dr. DoHwan Park
Dr. Yi Huang
Dr. Martin Klein

Acknowledgments

I would like to take this opportunity to express my utmost gratitude to the following people who have been instrumental in the completion of this thesis:

To Dr. Bimal K. Sinha, my adviser, for generously sharing his ideas and time during the entire thesis process and for continuously encouraging me to pursue research and teaching. You are the embodiment of a true mentor. Throughout my entire graduate school experience, you have shown genuine concern for me and for that, I will forever be grateful.

To Dr. Chuanhua Xing, my co-adviser, for her patience, motivation and immense generosity in sharing her time and knowledge. I could not imagine having accomplished this dissertation without your guidance.

To the committee members for my dissertation, Dr. Thomas Mathew, Dr. Yi Huang, Dr. DoHwan Park and Dr. Martin Klein, for imparting their knowledge for the improvement of this work. Your insights and comments are very much appreciated.

To the faculty and staff of the Department of Mathematics and Statistics for their encouragement. Your gestures of support, however small, have sustained me all throughout my graduate studies.

To my classmates and friends in the program, for their help and prayers. My graduate school experience was memorable because of the friendships we have built over the years.

To my extended families in Maryland, for always encouraging me to move

forward amidst the struggles of being away from my immediate family. I will always be thankful to God for giving me these families who have adopted me like their own these past years.

To my friends in the Philippines, for always reaching out and keeping our friendship strong despite the distance. I have been motivated throughout this academic journey because of your encouragement.

To my siblings, for always helping me out and for taking good care of me. Thank you for being understanding and supportive. The two of you have been good role models in my pursuit for academic excellence.

To my Mama and Papa, for your undying support and affection. I am lucky to have you as my parents.

To my husband, Arthur, for being patient with me through all these years. Thank you for the sacrifices you have made for me. This victory is just as much yours as it is mine.

And most especially, to God Almighty, for without You, nothing is possible. You have been forever faithful to Your people and all praises and glory belong to You. Thank you for the wisdom, the strength and the provisions You have given me. May You alone, Dear God, be glorified in this work.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ATE	Average Treatment Effect
BMI	Body Mass Index
IPTW	Inverse Probability of Treatment Weighting
MLE	Maximum Likelihood Estimator
NSW	National Supported Work
PS	Propensity Scores
RE78	Real Earnings in 1978
RMLE	Restricted Maximum Likelihood Estimator
SE	Standard Error

Chapter 1

Introduction

1.1 Causal Inference

In causal inference, the potential outcome framework is founded on the idea that every unit has a pair of potential outcomes $\left(Y^{(0)}, Y^{(1)}\right)$ which refers to the unit's response had it been assigned to the control and treatment group, respectively. However, in observational studies or non-randomized trials, having both outcomes at once is impossible and consequently, one is always missing. The observed value for each subject is then defined as

$$Y_i = D_i Y_i^{(1)} + (1 - D_i)\, Y_i^{(0)}, \tag{1.1}$$

where $D_i = 0$ if the ith unit is in the control group and $D_i = 1$ if the ith unit is in the treatment group. Under this setup, the treatment effect for each subject is $Y_i^{(1)} - Y_i^{(0)}$ and the average treatment effect (ATE)is defined by $E\left[Y_i^{(1)} - Y_i^{(0)}\right]$, which is the target estimand for inference.

In randomized controlled trials, the treatment allocation is not confounded by the covariates. This allows for the direct comparison of outcomes between the treatment and control groups. However, in observational studies, there may exist systematic differences in the covariates of the two groups; hence, estimates derived by direct comparison are biased because differences in the response between the two

groups may be attributed to the disparity in covariates and not the treatment effect itself. Under such situations, methods have been proposed to remove bias caused by confounding. With this, proper estimation of the average treatment effect is conducted so that the difference in treatment outcomes realized is truly attributed to the treatment and not on the said systematic differences in covariates between groups.

Propensity scores analysis is a commonly used method that addresses this issue. This matching method posits that less biased estimates are realized when the comparison of outcomes is made on groups that are as similar as possible in terms of the covariates. Such balance between the two groups is achieved by matching based on a single score, $e(x)$, known as propensity scores (PS), that summarizes the n-dimensional vector of pre-treatment characteristics [24]. This score is defined as the conditional probability of having been assigned with the treatment given a vector $\mathbf{x}$ of observed covariates; that is $e(x) = P(D_i = 1|\mathbf{x})$. A vast literature on the matching and estimation methods based on the propensity scores are available such as Stratification Matching [25, 1], Propensity Score Matching [24, 26, 1], and Inverse Probability of Treatment Weighting (IPTW) using propensity scores [27, 1] which are briefly discussed in the following sections.

1.2 Stratification Matching

A common procedure in controlling the systematic differences existing between covariates is creating subgroups based on the computed propensity scores and di-

rectly comparing the units in treatment and control groups from each subgroup formed. To achieve this stratification of subjects into mutually exclusive subgroups, subjects are ranked based on their propensity scores and are then stratified into subsets based on previously defined threshold based on it. In the application of this method, it has been established that 90% of the biases due to the covariates are removed by using five subgroups [25, 1]. This is done by checking the balance achieved by comparing the covariates of the treatments across the different strata. The key point of this method lies on the idea that when the propensity score model has been correctly specified, the distribution of measured covariates will be approximately similar between treated and untreated subjects within the same stratum [4], which can be then be likened to a randomized study. Hence, direct comparisons between the two groups can be performed.

An estimate for the average treatment effect under this setup is obtained using a direct adjustment of the subgroup-specific estimates generated. This is simply formulated as the average of the K subgroup differences of the mean responses between the treatment groups [25]. This expression is summarized [35] as

$$\hat{\tau} = \sum_{k=1}^{K} \frac{n_k}{N} (\bar{y}_{1k} - \bar{y}_{0k}) \tag{1.2}$$

with a corresponding variance of

$$Var(\hat{\tau}) = \sum_{k=1}^{K} \left(\frac{n_k}{N}\right)^2 Var(\bar{y}_{1k} - \bar{y}_{0k}). \tag{1.3}$$

In theory, all strata must be balanced with respect to the different covariates to ensure that an unbiased average treatment effect is calculated. However, in practice, this balance might not be easily achieved. In such a case, it is vital that the clinical

significance be justified or the amount of reduction in imbalance after stratification be investigated.

1.3 Propensity Score Matching

Propensity score matching forms matched sets of units in the treatment and control groups whose propensity scores are almost similar. The most common of which is the one-to-one matching or pair matching. In this method, a treated unit i with an estimated propensity score $\hat{e}_i$ is matched to a control unit (or set of control units $C(i)$) where $C(i) = min_j||p_i - p_j||$. For pair matching, this is a singleton which is most often the case. Multiple nearest neighbors, nonetheless, may be realized but is very rare especially if the set of covariates include continuous variables [5].

With these matched samples, the treatment effect is calculated by directly comparing the outcomes, as in randomized studies. For continuous responses, this refers to the difference between the mean of the response variable of the treated group and untreated group in the matched sample. On one hand, if the outcome is binary, the treatment effect is estimated as the difference in the proportion of units that are experiencing the event of interest between the treated and untreated groups [24, 26]. Other measures that gauge average treatment effect used in many applications are relative risk and odds ratio [2, 3, 4].

Generally, the average treatment effect estimator for continuous response is

given by

$$\begin{aligned}\hat{\tau} &= \frac{1}{N_T}\sum_{i\in T}\left(Y_i^T - \sum_{j\in C(i)} w_{ij}Y_j^C\right) \\ &= \frac{1}{N^T}\sum_{i\in T} Y_i^T - \frac{1}{N^T}\sum_{j\in C} w_{ij}Y_j^C, \end{aligned} \tag{1.4}$$

where Y_i^T is the response of the treated unit, Y_j^C is the response of the matched control unit, N^T is the number of units in the treatment group and $w_{ij} = \frac{1}{N_i^C}$ are its weights. In the case of a pair matching, this formulation simplifies to a direct computation of the average treatment effect as described. The corresponding variance of the estimator, assuming fixed weights and independent outcomes across units, is given by

$$\begin{aligned} Var(\hat{\tau}) &= \frac{1}{(N_T)^2}\left[\sum_{i\in T} Var\left(Y_i^T\right) + \sum_{j\in C}(w_{ij})^2 Var\left(Y_j^C\right)\right] \\ &= \frac{1}{N^T}Var\left(Y_i^T\right) + \frac{1}{(N^C)^2}\sum_{j\in C}(w_{ij})^2 Var\left(Y_j^C\right). \end{aligned} \tag{1.5}$$

It is suggested that within the matched sample, the treatment and control units should be regarded as independent [29]. In contrast, some researchers argue that the estimate and its corresponding variance should be derived under the paired setup [12]. Hence, a paired t-test and Mcnemar Test [20] should be used for statistical inference in continuous and dichotomous responses, respectively.

1.4 Inverse Probability Treatment Weighting

Another common propensity score approach is the inverse probability of treatment weighting (IPTW). This method is particularly attractive because a treatment effect estimate can be calculated even if the response is a rare binary outcome and

adjustments on the covariates are necessary [36]. In this procedure, once the propensity scores are estimated, an estimator for the average treatment effect is given by

$$\hat{\tau} = \left(\sum_{i=1}^{n} \frac{Y_i D_i}{\hat{e}_i}\right)\left(\sum_{i=1}^{n} \frac{D_i}{\hat{e}_i}\right)^{-1} - \left(\sum_{i=1}^{n} \frac{Y_i(1-D_i)}{(1-\hat{e}_i)}\right)\left(\sum_{i=1}^{n} \frac{(1-D_i)}{(1-\hat{e}_i)}\right)^{-1} \tag{1.6}$$

which is essentially the difference in marginal means of the treatment and control groups, $\mu_T - \mu_C$. The corresponding estimate of the variance of this treatment effect estimator is derived [36] as

$$n\hat{Var}(\hat{\tau}) = \hat{V}_{un} - \hat{\mathbf{v}}^T \left(2\hat{\mathbf{M}}_1 - \hat{\mathbf{M}}_2\right)\hat{\mathbf{v}}, \tag{1.7}$$

where supposing $\hat{w}_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i}{\hat{e}_i}$ and $\hat{w}_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{1-D_i}{1-\hat{e}_i}$,

$$\begin{aligned}
\hat{V}_{un} &= \frac{1}{\hat{w}_1^2}\frac{1}{n}\sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_1)^2 D_i}{\hat{e}_i^2} + \frac{1}{\hat{w}_0^2}\frac{1}{n}\sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_0)^2(1-D_i)}{(1-\hat{e}_i)^2} \\
\hat{\mathbf{v}} &= \frac{1}{\hat{w}_1}\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\mathbf{x}}_i(Y_i - \hat{\mu}_1)D_i(1-\hat{e}_i)}{\hat{e}_i} + \frac{1}{\hat{w}_0}\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\mathbf{x}}_i(Y_i - \hat{\mu}_0)(1-D_i)\hat{e}_i}{(1-\hat{e}_i)} \\
\hat{\mathbf{M}}_1 &= \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{x}}_i\hat{\mathbf{x}}_i^T\hat{e}_i(1-\hat{e}_i)\right)^{-1} \\
\hat{\mathbf{M}}_2 &= \hat{\mathbf{M}}_1\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{x}}_i\hat{\mathbf{x}}_i^T(D_i - \hat{e}_i)^2\right)\hat{\mathbf{M}}_1.
\end{aligned} \tag{1.8}$$

This estimator has been shown to have comparable statistical properties to stratification and matching methods and, in certain circumstances, it is a more preferable analysis procedure.

Looking into the idea of balance in propensity scores, it is defined based on the closeness of the propensity score calculated for each i. However, careful assessment on the matches based on the first two methods continue to show imbalanced covariates; thus, a modified method is proposed. In this procedure, matching is conducted based on some of the covariate information where each group made is

assured to be balanced with respect to the categorical variables and an alternative test in checking the balance of continuous variables within each matched group is presented.

In the implementation of these different methods, careful consideration must be taken to ensure that the assumption of strong ignorability imposed by the propensity score is achieved. This means that the researchers must include all variables related to the treatment and the response, while excluding the response variable, in the construction of the propensity score model. Also, irrelevant covariates will inflate variance estimates as it adds up to the noise in the model, while deleting important variables can result in serious bias [22]. However, it is preferred to include unimportant covariates and consequently lose efficiency rather than increase the bias by deleting a variable [28]. It is for this reason, that most propensity score models are built on as many variables as possible given the data. In the proposed method, however, the mechanism works when the data set contains manageable number of variables and the goal is to estimate the average treatment effect.

Another advantage in comparison to propensity score methods is the use of meta-analysis which addresses homogeneity in the combined treatment effect. Meanwhile, the "swapping" method allows the incorporation of covariate information in average treatment effect estimation. All of these analyses are geared towards providing appealing features in estimating average treatment effect in comparison to propensity score methods.

1.5 Motivation for the Proposed Procedure

In analyzing observational studies, the main issue lies in the presence of systematic differences in the features of the treatment and the control groups. The solution provided by propensity score analysis is to find ample matches for the treatment and control unit which are close to each other based on the balancing score, $e(x)$. However, this definition of balance, when closely assessed, do not absolutely provide matches whose characteristics or baseline features are the same.

For instance, propensity score matching methods are applied on Lalonde data set from a 1999 seminal study on the comparison of treatment and control groups to determine causal effects of a job training program [15]. The Lalonde data combines the treated units from a randomized evaluation of the National Supported Work (NSW) demonstration with the control units drawn from survey data. This public data is widely used to illustrate propensity scores and matching methods. The outcome of interest is RE78 (real earnings in 1978) with the treatment defined as the participation in the NSW job training program. The continuous covariates considered in the study are age and education quantified by the number of years in school while the categorical variables are Black (1 if black, 0 otherwise), Hispanic (1 if Hispanic, 0 otherwise), married (1 if married, 0 otherwise) and nodegree (1 if no degree, 0 otherwise). The total number of observations used in the analysis is $n = 445$, $n_C = 260$ of which are under the control group while the remaining $n_T = 185$ observations are treatments.

A propensity score logistic model is estimated so that

$$\begin{aligned} logit(P(D_i|x_i)) &= 1.1092 + 0.0101age - 0.00009age^2 - 0.2245black - 0.0719ed \\ &\quad -0.8515hisp + 0.1608married - 0.9026nodeg. \end{aligned} \tag{1.9}$$

With this estimated model, a balancing score is assigned to each of the unit in the treatment and control groups. Under the pair matching method, a treatment unit is matched to a control unit where its corresponding absolute difference in propensity score is smallest in the pool of control units. However, assessing the features of the matched pair, it can be shown that disparity in characteristics of some matches formed exist. A subset of the matches made using this method is shown in Table 1.1.

Matches	**Group**	**Age**	**Educ**	**Black**	**Hisp**	**Married**	**Nodeg**	$e(x)$
1	T	27	10	0	1	0	1	0.2393
	C	17	10	0	1	0	1	0.2391
2	T	17	9	0	1	0	1	0.2394
	C	17	10	0	1	0	1	0.2392
3	T	20	11	1	0	0	1	0.2559
	C	18	8	0	1	0	1	0.2558
4	T	21	7	0	1	0	1	0.2786
	C	19	11	0	1	1	1	0.2738
5	T	26	10	1	0	0	1	0.2813
	C	19	6	0	1	0	1	0.2856

Table 1.1: Subset of Matches Formed on NSW Data Based on Pair Matching

Evidently, in some cases such as Matches 1 and 2, exact matches are formed with respect to the categorical variable. However, some matches, like Matches 3, 4, 5, do not share the same categorical characteristics. For instance, Matches 3

and 4 involve comparing a black, non-hispanic program participant to a non-black, hispanic respondent. Similarly, in Match 4, a married respondent is being matched with a participant who is single by civil status. This illustrates the inability of the propensity score method to come up with matches that reflect "balance" with respect to at least the categorical features. It can also be observed that for Match 2, the pair is composed of two 17-year old respondents while for Match 1, there is a 10-year age gap between the pair derived. Meanwhile, with respect to the number of years spent in school, Match 1 is comparable, where both respondents have spent 10 years in school. However, for Match 5, the control unit has spent fours years in school less than the treatment unit. These observations suggest that while the propensity scores calculated are quite similar, the covariates of the matches made are not completely "balanced".

Looking into the quality of the matched groups formed via the stratification method, the same propensity score model is used and the balancing scores are ranked to create the five strata. The results of the strata formed are shown in Table 1.2. It is shown that the composition of the categorical variables in some of the strata formed vary. For Strata 1 and 2, very minimal disparity is established in the composition of the treatment and control groups. However, Stratum 3 is composed of 9% married respondents for the treatment group and only 3% are married for the control group. Strata 4 and 5 have likewise varied percentages of respondents who are married and has no degree. Such results indicate a sense of imbalance in the groups formed using propensity scores.

From this illustration, it can be shown that the features of the matches made

Stratum	**Group**	n	**Black**	**Hisp**	**Married**	**Nodeg**
1	T	29	20 (69%)	9 (31%)	2 (7%)	29 (100%)
	C	54	38 (70%)	16 (30%)	3 (6%)	54 (100%)
2	T	32	32 (100%)	0 (0%)	0 (0%)	32 (100%)
	C	60	60 (100%)	0 (0%)	0 (0%)	60 (100%)
3	T	34	32 (94%)	1 (3%)	3 (9%)	33 (97%)
	C	51	48 (94%)	1 (2%)	2 (4%)	50 (98%)
4	T	41	32 (78%)	1 (2%)	20 (49%)	35 (85%)
	C	45	35 (78%)	1 (2%)	26 (58%)	43 (96%)
5	T	46	40 (87%)	0 (0%)	8 (17%)	2 (4%)
	C	40	34 (85%)	0 (0%)	9 (23%)	0 (0%)

Table 1.2: Matched Groups of NSW Data Based on Stratification Matching

may be different; hence, it is not completely balanced. Although it is balanced by the closeness of its propensity scores, the features of the covariates of the matches are not alike which may indicate that systematic differences in the treatment and control groups continue to exist. This may consequently lead to biased estimates of the average treatment effect. In the proposed method, this imbalance is being addressed by forming subgroups that are perfectly balanced with respect to the categorical features of the data set. An alternative way of checking the balance with respect to the continuous variables are also assessed and a test of homogeneity is introduced in the estimation of the average treatment effect. All of these components are geared towards providing a better estimate of the average treatment effect.

Chapter 2

Proposed Balancing and Estimation Method

2.1 Overview

As an alternative to the methods of propensity score analysis, a new procedure of analyzing observational data to establish average treatment effect is proposed. In this method, subgroups are formed based on the nature of the relevant covariates in predicting the outcome of interest. Classical regression and the "swapping" idea are then infused and implemented to find appropriate matches that will estimate missing potential outcomes. Also, meta-analysis procedures are used to combine subgroup estimates and consequently to generate the average treatment effect. In the succeeding sections, the data setup and general approach for the proposed method are more clearly described.

2.2 Data Setup

For a subgroup i, let $\underset{\sim}{Y}$ be an $N \times 1$ vector of observed outcomes on the response variable and $\mathbf{X}$ be an $N \times q$ matrix of observed covariates, where N is the total number of observations and q be the number of selected independent variables

available in the data set. Thus, the data matrix is given by

$$\underset{\sim}{\mathbf{Y}} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \\ Y_{n+1} \\ \vdots \\ Y_N \end{pmatrix}, \mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1q} \\ \vdots & & & \\ X_{n1} & X_{n2} & \dots & X_{nq} \\ X_{(n+1)1} & X_{(n+1)2} & \dots & X_{(n+1)q} \\ \vdots & & & \\ X_{N1} & X_{N2} & \dots & X_{Nq} \end{pmatrix} \tag{2.1}$$

Assume that the first n observations belong to the treatment group and the remaining $m = N - n$ observations are members of the control group so that the data matrix can be partitioned into the treatment group and control group as

$$\underset{\sim}{\mathbf{Y}}^{(1)} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X^{(1)}} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1q} \\ \vdots & & & \\ X_{n1} & X_{n2} & \dots & X_{nq} \end{pmatrix} \tag{2.2}$$

and

$$\underset{\sim}{\mathbf{Y}}^{(0)} = \begin{pmatrix} Y_{n+1} \\ \vdots \\ Y_N \end{pmatrix}, \mathbf{X^{(0)}} = \begin{pmatrix} X_{(n+1)1} & X_{(n+1)2} & \dots & X_{(n+1)q} \\ \vdots & & & \\ X_{N1} & X_{N2} & \dots & X_{Nq} \end{pmatrix}, \tag{2.3}$$

respectively, with $min\{n, m\} > q$. These submatrices can then be redefined such that each unit $j = 1, 2, \dots, n_i$ in the treatment group will have $\left(Y_j^{(1)}, \mathbf{X}_j^{(1)}\right) = (U_j, \mathbf{v}_j)$ where U_j is the observed response value and $\mathbf{v}_j$ is the realized vector of covariates. Similarly, the observed quantities for the control units are redefined such that $\left(Y_j^{(0)}, \mathbf{X}_j^{(0)}\right) = (Z_j, \mathbf{w}_j)$, where Z_j and $\mathbf{w}_j$ are the observed outcome value and vector of independent variable values, respectively, for every $j = 1, 2, \dots, m_i$. It

is emphasized that based on the nature of observational studies, the matrices $\mathbf{v}$ and $\mathbf{w}$ share the same characteristics but balance is not guaranteed to occur between the two groups. This balance is desired so that any realized effect can be attributed solely to the treatment and not on the systematic differences between the groups.

2.3 General Approach

In general, the proposed method involves three major steps: creation of balanced subgroups, estimation of treatment effect for each subgroup and combination of these estimates for the average treatment effect. Through these steps, issues on balance and homogeneity of subgroups are addressed. It must be noted, however, that this procedure is most applicable for data sets with manageable number of confounders. This proposed method is performed through the following mechanism:

(1) Create subgroups based on the $k(k \leq q)$ categorical variables, forming a total of $G = \prod_{l=1}^{k} b_l$ subgroups, where b_l is the number of levels in the lth categorical variable. This step assures that each of the G subgroup are balanced with respect to the categorical variables. Hence, no balance checking procedure is necessary between treatment and control groups in each of the G subgroup.

In the case that a large number of subgroups formed have sparse data, the levels of some categorical variables are pooled if there is a natural way of doing it. If subgroups remain to be sparse even after pooling, these sparse subgroups are then discarded. This results to n^* and m^* remaining observations classified into G^* valid subgroups. Each of these $i = 1, 2, \ldots, G^*$ have n_i observations for

the treatment group and m_i for the control group such that $\sum_{i=1}^{G^*} n_i = n^*$ and $\sum_{i=1}^{G^*} m_i = m^*$ with $N^* = n^* + m^*$ total number of observations remaining.

Note that in each of these subgroups, the data setup is as described in section 2.2.

(2) Subgroups with non-sparse observations are further validated for balance with respect to the remaining $p = q - k$ continuous variables. A multivariate test is applied to check whether the balance is achieved on the continuous variables. The test is for the equality of two mean vectors under the assumption of unequal dispersion matrices Σ. Several tests have been proposed in the literature for solving this problem, also known as multivariate Behrens-Fisher problem. Available tests are Yao's Test [37], Johansen's Test [13], Nel and Van der Merve's Test [21], Krishnamoorthy and Yu's test [14], and the generalized p-value test [34]. See subsection 2.4 for the details of these different tests.

(3) After checking the balance of the covariates between the treatment and control groups, estimates of the treatment effect, θ_i, are derived for each subgroup. That is,

   (a) If balance in a specific subgroup is achieved, proceed to the direct computation of effect sizes. Since balance is ensured, no systematic difference exist as in a randomized study. Thus, direct difference in the averages of treatment and control groups can be applied to calculate the subgroup estimate.

   (b) If balance is not achieved, proceed to the "swapping" method from which

the missing potential outcome is imputed. The details are described in Chapter 3.

(4) Upon deriving the treatment effect θ_i in each subgroup, a test for homogeneity is conducted. That is, test if $\theta_1 = \theta_2 = \ldots = \theta_{G^*} = \theta$, where θ is a common treatment effect value.

   (a) If the assumption of homogeneity is satisfied, the G^* effect sizes are combined using meta-analysis methods.

   (b) If the assumption of homogeneity is not satisfied, a random effects model is postulated and a homogeneity parameter is estimated.

The combined effect sizes is the estimated ATE, $E[Y^{(1)} - Y^{(0)}]$. The estimators and standard errors of this proposed procedure are derived in details in Chapter 3 for continuous response variable.

2.4 Test for Equality of Mean Vector of Covariates

Balance checking among covariates is an essential step prior to the computation of treatment effect. Each subgroup is assured to have balance with respect to the categorical variables but not on the continuous covariates. To check this balance, a multivariate test for the equality of two mean vectors under the assumption of unequal dispersion matrices is applied. Several progresses have been done in solving this problem. Among these tests are Yao's invariant test [37], Johansen's invariant test [13], Nel and Van der Merwe invariant test [21] and Krishnamoorthy and Yu

modified invariant test [14]. Also, new methods are being proposed that provide more reasonable size and power of the test [34].

The general formulation of the test suggests that for a subgroup, suppose that the treatment and control groups are from a p-variate normal populations given by $N(\underset{\sim}{\mu}_T, \Sigma_T)$ and $N(\underset{\sim}{\mu}_C, \Sigma_C)$, respectively, where $\underset{\sim}{\mu}_T$ and $\underset{\sim}{\mu}_C$ are unknown $p \times 1$ vectors and Σ_T and Σ_C are unknown $p \times p$ positive definite matrices. For balance checking, the test H_o: $\underset{\sim}{\mu}_T = \underset{\sim}{\mu}_C$ vs $H_1 : \underset{\sim}{\mu}_T \neq \underset{\sim}{\mu}_C$ must be conducted. For $n > p$, consider the estimators $\bar{\mathbf{X}}_T = \dfrac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{X}_{jT}$ and $\bar{\mathbf{X}}_C = \dfrac{1}{m_i}\sum_{j=1}^{m_i} \mathbf{X}_{jC}$ which are sufficient for the mean vectors of the two groups. Also, let $\mathbf{A}_T = \sum_{j=1}^{n_i}(\mathbf{X}_{jT} - \bar{\mathbf{X}})(\mathbf{X}_{jT} - \bar{\mathbf{X}})'$ and $\mathbf{A}_C = \sum_{j=1}^{m_i}(\mathbf{X}_{jC} - \bar{\mathbf{X}})(\mathbf{X}_{jC} - \bar{\mathbf{X}})'$ be sufficient estimators of the dispersion matrices. It is noted that for $k = T, C$, these quantities are independent variables with distributions

$$\bar{\mathbf{X}}_k \sim N\left(\mu_k, \frac{\Sigma_k}{n_{ik}}\right) \tag{2.4}$$

$$\mathbf{A}_k \sim W_p(n_{ik} - 1, \Sigma_k) \tag{2.5}$$

where $W_p(df = r, \Sigma)$ is a $p-$dimensional Wishart distribution. Define

$$\mathbf{S}_k = \frac{\mathbf{A}_k}{n_{ik} - 1} \tag{2.6}$$

$$\tilde{\mathbf{S}}_k = \frac{\mathbf{S}_k}{n_{ik}} \tag{2.7}$$

$$\tilde{\mathbf{S}} = \tilde{\mathbf{S}}_T + \tilde{\mathbf{S}}_C \tag{2.8}$$

$$T^2 = (\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)'\tilde{\mathbf{S}}^{-1}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C) \tag{2.9}$$

where the test based on T^2 is a natural invariant test.

Among the available multivariate tests are the following:

(a) Yao's test [37] suggests a multivariate test based on $T^2 \sim \left(\frac{vp}{v-p+1}\right) F_{p,v-p+1}$ where

$$v = \left[\frac{1}{n_{iT}}\left(\frac{\bar{\mathbf{X}}_d'\tilde{\mathbf{S}}^{-1}\tilde{\mathbf{S}}_T\tilde{\mathbf{S}}^{-1}\bar{\mathbf{X}}_d}{\bar{\mathbf{X}}_d'\tilde{\mathbf{S}}^{-1}\bar{\mathbf{X}}_d}\right) + \frac{1}{n_{iC}}\left(\frac{\bar{\mathbf{X}}_d'\tilde{\mathbf{S}}^{-1}\tilde{\mathbf{S}}_C\tilde{\mathbf{S}}^{-1}\bar{\mathbf{X}}_d}{\bar{\mathbf{X}}_d'\tilde{\mathbf{S}}^{-1}\bar{\mathbf{X}}_d}\right)\right]^{-1} \tag{2.10}$$

and

$$\bar{\mathbf{X}}_d = \bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C. \tag{2.11}$$

(b) Johansen's test [13] suggests that $T^2 \sim q^* F_{p,v}$ where

$$\begin{aligned} q^* &= \frac{p+2D-6D}{p(p-1)+2} \\ v &= \frac{p(p+2)}{3D} \end{aligned}$$

and

$$D = \frac{1}{2}\sum_{k=T,C}\frac{trace[(I-(\tilde{\mathbf{S}}_T^{-1}+\tilde{\mathbf{S}}_C)^{-1}\tilde{\mathbf{S}}_k^{-1})^2] + trace[(I-(\tilde{\mathbf{S}}_T^{-1}+\tilde{\mathbf{S}}_C)^{-1}\tilde{\mathbf{S}}_k^{-1})]^2}{n_{ik}} \tag{2.12}$$

Studies conducted on this test show that it performs better than the usual Hotelling's T where the covariance matrices are assumed to be equal.

(c) Nel and Van der Merve's test [21] is an approximate solution to the multivariate Behrens-Fisher problem which is a non-variant test based on $T^2 \sim \left(\frac{vp}{v-p+1}\right) F_{p,v-p+1}$ where v is redefined as

$$v = \frac{trace(\tilde{\mathbf{S}}^2) + [trace(\tilde{\mathbf{S}})]^2}{\frac{1}{n_{iT}}[trace(\tilde{\mathbf{S}}_T^2) + [trace(\tilde{\mathbf{S}}_T)]^2] + \frac{1}{n_{iC}}[trace(\tilde{\mathbf{S}}_C^2) + [trace(\tilde{\mathbf{S}}_C)]^2]} \tag{2.13}$$

(d) Krishnamoorthy and Yu's test [14] modifies the approximation provided by Nel and Van der Merve's invariant test. This modification simplifies to the

approximate solution to the univariate case and is similarly based on $T^2 \sim \left(\frac{vp}{v-p+1}\right) F_{p,v-p+1}$ but the degrees of freedom is given by

$$v = \frac{p + p^2}{\frac{1}{n_{iT}}[trace(\tilde{\mathbf{S}}_T\tilde{\mathbf{S}}^{-1})^2 + [trace(\tilde{\mathbf{S}}_T\tilde{\mathbf{S}}^{-1})]^2] + \frac{1}{n_{iC}}[trace(\tilde{\mathbf{S}}_C\tilde{\mathbf{S}}^{-1})^2 + [trace(\tilde{\mathbf{S}}_C)\tilde{\mathbf{S}}^{-1}]^2]} \tag{2.14}$$

The simulations based on the test demonstrated that it is as powerful as the other methods explored while controlling for the sample sizes.

At present, the multivariate Behrens-Fisher problem remains to be an interesting research area in statistics. Equally more powerful tests are being developed such as the test based on Roy's union-intersection principle with generalized $P-$value [34] which might be useful in establishing the equality of continuous covariates between the treatment and control groups given a specific subgroup.

For subgroups where the equality of mean vectors are established, direct calculation of the effect size is done as there exists no systematic differences between groups that could affect the average treatment effect. This treatment effect is calculated as the difference in mean response values of the treatment and control groups. If the null hypothesis is rejected, then the proposed swapping procedure is applied.

Chapter 3

Average Treatment Effect Estimation

3.1 Overview

Upon the creation of subgroups and balance checking with respect to the continuous variables, the "swapping" method is applied to subgroups that are found to be unbalanced while direct computation is done for balanced ones. This will allow generation of subgroup estimates for the treatment effect.

3.2 Swapping Method

The rationale of the "swapping" method is based on the idea that in the presence of a good model for the distribution of the response under the treatment group, the missing potential outcome $Y^{(1)}$ for units in the control group can be predicted. Similarly, the missing $Y^{(0)}$ of the treatment units could be estimated by looking into its predictive values under the model established using the control group. As an initial step to the swapping method, models are built using classical regression methods. For a fixed subgroup i, the response variable in the treatment group is

$$U_j = \beta_{0,T} + \beta_1 v_{j1} + \ldots + \beta_p v_{jp} + \epsilon_j \tag{3.1}$$

where $\epsilon_j \sim N(0, \sigma_T^2)$. Hence, $U_j|\underset{\sim}{v}_j \sim N(\underset{\sim}{v}_j'\underset{\sim}{\beta}_T, \sigma_T^2)$ for $j = 1, 2, \ldots, n_i$. In classical regression, a centered form of the model can be considered so that

$$\begin{aligned} U_j &= \beta_{0,T}' + \beta_1 v_{j1} + \ldots + \beta_p v_{jp} + \epsilon_j \\ &= \beta_{0,T}' + \beta_1(v_{j1} - \bar{v}_1) + \ldots + \beta_p(v_{jp} - \bar{v}_p) + \epsilon_j \end{aligned} \tag{3.2}$$

where $\bar{v}_l = \sum_{j=1}^{n_i} v_{jl}$ and

$$\beta_{0,T}' = \beta_{0,T} + \beta_1\bar{v}_1 + \ldots + \beta_p\bar{v}_p \tag{3.3}$$

for $l = 1, 2, \ldots, p$.

In matrix form, vectors $\underset{\sim}{U}, \underset{\sim}{\beta}_T, \underset{\sim}{\epsilon}$ and matrix $\mathbf{V}^*$ can be defined as

$$\underset{\sim}{U} = \begin{pmatrix} U_1 \\ \vdots \\ U_{n_i} \end{pmatrix}_{n_i \times 1}, \underset{\sim}{\beta}_T = \begin{pmatrix} \beta_{0,T} \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1)\times 1}, \underset{\sim}{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{n_i} \end{pmatrix}_{(n_i \times 1)}$$

where $\underset{\sim}{\epsilon} \sim N_{n_i}(\underset{\sim}{0}, \sigma_T^2\mathbf{I})$ and

$$\mathbf{V}^* = \begin{pmatrix} 1 & v_{11} - \bar{v}_1 & \ldots & v_{1p} - \bar{v}_p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & v_{n_i 1} - \bar{v}_1 & \ldots & v_{n_i p} - \bar{v}_p \end{pmatrix}_{(n_i \times (p+1))} \tag{3.4}$$

which provides the least squares estimates of

$$\underset{\sim}{\hat{\beta}}_T^* = \begin{pmatrix} \hat{\beta}_{0,T}' \\ \hat{\beta}_{1,T} \\ \vdots \\ \hat{\beta}_{p,T} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_{0,T}' \\ \underset{\sim}{\hat{\beta}}_T^{**} \end{pmatrix} = \left(\mathbf{V}^{*'}\mathbf{V}^*\right)^{-1}\mathbf{V}^{*'}\underset{\sim}{U} = \begin{pmatrix} n_i & 0 & 0 & \ldots & 0 \\ 0 & & & & \\ \vdots & & \mathbf{S}_{vv} & & \\ 0 & & & & \end{pmatrix}^{-1} \mathbf{V}^{*'}\underset{\sim}{U}. \tag{3.5}$$

where $\mathbf{S}_{vv}$ is a $p \times p$ matrix of sum of squares and cross-products of deviations with the $ij-$elements $a_{ij} = \sum_{j=1}^{n_i}(v_{jl} - \bar{v}_l)(v_{jq} - \bar{v}_q)$ for $l, q = 1, 2, \ldots, p$.

This formulation simplifies the distributional properties of $\hat{\underset{\sim}{\beta}}^*_T$ into

$$E\left(\hat{\underset{\sim}{\beta}}^*_T\right) = \begin{pmatrix} \beta'_{0,T} \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \beta'_{0,T} \\ \underset{\sim}{\beta}^* \end{pmatrix} \tag{3.6}$$

$$Cov\left(\hat{\underset{\sim}{\beta}}^*_T\right) = \sigma^2_T \begin{pmatrix} \frac{1}{n_i} & 0 & 0 & \ldots & 0 \\ 0 & & & & \\ \vdots & & \mathbf{S}^{-1}_{vv} & & \\ 0 & & & & \end{pmatrix}. \tag{3.7}$$

Focusing on the control group of subgroup i, the same least squares estimation on response variable of the control

$$Z_j = \beta_{0,C} + \beta_1 w_{j1} + \ldots + \beta_p w_{jp} + \epsilon_j \tag{3.8}$$

is performed where $\epsilon_j \sim N(0, \sigma^2_C)$. Thus, $Z_j|\underset{\sim}{w}_j \sim N(\underset{\sim}{w}'_j\underset{\sim}{\beta}_C, \sigma^2_C)$ for $j = 1, 2, \ldots, m_i$. For a more parsimonious form for the dispersion matrix of the estimates, a centered form is likewise considered in its estimation where

$$\begin{aligned} Z_j &= \beta'_{0,C} + \beta_1 w_{j1} + \ldots + \beta_p w_{jp} + \epsilon_j \\ &= \beta'_{0,C} + \beta_1(w_{j1} - \bar{w}_1) + \ldots + \beta_p(w_{jp} - \bar{w}_p) + \epsilon_j \end{aligned} \tag{3.9}$$

with $\bar{w}_l = \sum_{j=1}^{m_i} w_{jl}$ for $l = 1, 2, \ldots, p$ and

$$\beta'_{0,C} = \beta_{0,C} + \beta_1\bar{w}_1 + \beta_2\bar{w}_2 + \ldots + \beta_p\bar{w}_p. \tag{3.10}$$

Similarly, the vector of response of the control group can be expressed in matrix form as $\underset{\sim}{Z} = \mathbf{W}^{*'}\underset{\sim}{\beta}_C + \underset{\sim}{\epsilon}$ where

$$\underset{\sim}{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_{m_i} \end{pmatrix}_{m_i \times 1}, \underset{\sim}{\beta}_C = \begin{pmatrix} \beta_{0,C} \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1)\times 1},$$

$$\mathbf{W}^* = \begin{pmatrix} 1 & w_{11} - \bar{w}_1 & \dots & w_{1p} - \bar{w}_p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & w_{m_i 1} - \bar{w}_1 & \dots & w_{m_i p} - \bar{w}_p \end{pmatrix}_{(m_i \times (p+1))}, \underset{\sim}{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{m_i} \end{pmatrix}_{(m_i \times 1)}$$

with $\underset{\sim}{\epsilon} \sim N_{m_i}(\underset{\sim}{0}, \sigma_C^2 \mathbf{I})$. Under this specification, the least squares estimates of the response in the control group is given by

$$\underset{\sim}{\hat{\beta}}^*_C = \begin{pmatrix} \hat{\beta}'_{0,C} \\ \hat{\beta}_{1,C} \\ \vdots \\ \hat{\beta}_{p,C} \end{pmatrix} = \begin{pmatrix} \hat{\beta}'_{0,C} \\ \underset{\sim}{\hat{\beta}}^{**}_C \end{pmatrix} = \left(\mathbf{W}^{*'}\mathbf{W}^*\right)^{-1}\mathbf{W}^{*'}\underset{\sim}{Z} = \begin{pmatrix} m_i & 0 & 0 & \dots & 0 \\ 0 & & & & \\ \vdots & & \mathbf{S}_{ww} & & \\ 0 & & & & \end{pmatrix}^{-1} \mathbf{W}^{*'}\underset{\sim}{Z} \tag{3.11}$$

where $\mathbf{S}_{ww}$ is a $p \times p$ matrix of sum of squares and cross-products of deviations with $ij-$element $b_{ij} = \sum_{j=1}^{m_i}(w_{jl} - \bar{w}_l)(w_{jq} - \bar{w}_q)$ for $l, q = 1, 2, \dots, p$ and

$$\underset{\sim}{\hat{\beta}}^*_C = \begin{pmatrix} \hat{\beta}'_{0,C} \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}_{(p+1)\times 1}$$

Under this centered form, the expectation of $\hat{\underset{\sim}{\beta}}^*_C$ remains the same as the treatment group except for its intercept $\beta'_{0,C}$ while its variance is simplified into the form

$$E\left(\hat{\underset{\sim}{\beta}}^*_C\right) = \begin{pmatrix} \beta'_{0,C} \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \beta'_{0,C} \\ \underset{\sim}{\beta}^* \end{pmatrix} \tag{3.12}$$

$$Cov\left(\hat{\underset{\sim}{\beta}}^*_C\right) = \sigma^2_C \begin{pmatrix} \frac{1}{m_i} & 0 & 0 & \dots & 0 \\ 0 & & & & \\ \vdots & & \mathbf{S}^{-1}_{ww} & & \\ 0 & & & & \end{pmatrix}. \tag{3.13}$$

A crucial assumption in the model specifications above is that there exists a common $\underset{\sim}{\beta}^*$ in both groups which implies that the effect of the covariates on the response is the same regardless of treatment assignment. Furthermore, any effect of the treatment on a unit is reflected only through the intercepts of the models. To mitigate the estimation of the common parameter $\underset{\sim}{\beta}^*$, the two independent estimates $\hat{\underset{\sim}{\beta}}^{**}_T$ and $\hat{\underset{\sim}{\beta}}^{**}_C$ with known variance-covariance matrices, $\Sigma_T = \sigma^2_T S^{-1}_{vv}$ and $\Sigma_C = \sigma^2_T S^{-1}_{ww}$, respectively, are combined via multivariate meta-analysis by

$$\begin{aligned} \hat{\underset{\sim}{\beta}}^* &= \left(\sum_{k=T,C} \Sigma^{-1}_k\right)^{-1} \left(\sum_{k=T,C} \Sigma^{-1}_k \hat{\underset{\sim}{\beta}}^{**}_k\right) \\ &= \left(\frac{S_{vv}}{\sigma^2_T} + \frac{S_{ww}}{\sigma^2_C}\right)^{-1} \left(\sum_{k=T,C} \Sigma^{-1}_k \hat{\underset{\sim}{\beta}}^{**}_k\right), \end{aligned} \tag{3.14}$$

where

$$\hat{\underset{\sim}{\beta}}^* \sim N\left(\underset{\sim}{\beta}^*, \left(\sum_{k=T,C} \Sigma^{-1}_k\right)^{-1}\right)$$

and σ_T^2 and σ_C^2 are estimated by $\hat{\sigma}_T^2 = \dfrac{SSE_T}{n_i - (p+1)}$ and $\hat{\sigma}_C^2 = \dfrac{SSE_C}{m_i - (p+1)}$, respectively. Marginally, $\dfrac{(n_i - (p+1))\hat{\sigma}_T^2}{\sigma_T^2} \sim \chi^2_{(n_i-(p+1))}$ and $\dfrac{(m_i - (p+1))\hat{\sigma}_C^2}{\sigma_C^2} \sim \chi^2_{(m_i-(p+1))}$ so that the hypothesis $Ho : \sigma_T^2 = \sigma_C^2$ can be tested by the usual F-test. The non-rejection of the null allows simplification of $\left(\sum_{k=T,C} \Sigma_k^{-1}\right)^{-1}$ into $\dfrac{1}{\sigma^2}(S_{vv} + S_{ww})$. Another assumption made on this procedure is that homogeneous effect sizes exist between treatment and control groups since the possible values of the covariates belong to the same space with some overlap. Hence, the two estimates may be combined accordingly.

After the model-building and estimation of the common parameter $\underset{\sim}{\beta}^*$, the "swapping" mechanism is applied. For a subgroup i, the swapped quantities

$$\tilde{U}_j = \hat{\beta}'_{0,C} + \hat{\beta}_1(v_{11} - \bar{w}_1) + \ldots + \hat{\beta}_p(v_{n_i p} - \bar{w}_p) \tag{3.15}$$

$$\tilde{Z}_j = \hat{\beta}'_{0,T} + \hat{\beta}_1(w_{11} - \bar{v}_1) + \ldots + \hat{\beta}_p(w_{m_i p} - \bar{v}_p) \tag{3.16}$$

are defined which estimates the missing potential outcome in the framework described below:

(a) For the units in the treatment group where the potential outcome observed is only $Y_j^{(1)}$, estimate its corresponding potential outcome $Y_j^{(0)}$ by $\tilde{U}_j$. Therefore, for $j = 1, 2, \ldots, n_i$ in the treatment group, potential outcomes $(Y_j^{(1)}, Y_j^{(0)})$ are estimated by $(U_j, \tilde{U}_j)$. The rationale is that $\tilde{U}_j$ is the estimated response of unit i had it taken the placebo or the control.

(b) For the control group where only the potential outcome $Y_j^{(0)}$ is observed, estimate the treatment potential outcome $Y_j^{(1)}$ using $\tilde{Z}_j$ for each unit $j = 1, 2, \ldots, m_i$.

This will provide an estimate for potential outcome $(Y_j^{(1)}, Y_j^{(0)})$ of $(\tilde{Z}_j, Z_j)$. This structure posits that $\tilde{Z}_j$ is the response value of unit j had it undergone the treatment.

These model-building procedures for both groups are done for subgroups $1, 2, \ldots, G^*$ with unbalanced covariates. Upon model-fitting, a certain criteria is defined to aid in determining a good model to be used for "swapping". Measure such as the coefficient of determination, R^2, can be used to facilitate the choice of model for cases assuming normality. Another *adhoc* method that can be used in choosing the model used for swapping is the absolute difference of the quantities $\underset{\sim}{v}_j\hat{\underset{\sim}{\beta}}_T^{**} - \underset{\sim}{v}_j\hat{\underset{\sim}{\beta}}^*$ and $\underset{\sim}{w}_j\hat{\underset{\sim}{\beta}}_C^{**} - \underset{\sim}{w}_j\hat{\underset{\sim}{\beta}}^*$. The smaller of the quantities $s = \sum_{j=1}^{n_i} \frac{|\underset{\sim}{v}_j\hat{\underset{\sim}{\beta}}_T^{**} - \underset{\sim}{v}_j\hat{\underset{\sim}{\beta}}^*|}{n_i}$ and $t = \sum_{j=1}^{m_i} \frac{|\underset{\sim}{w}_j\hat{\underset{\sim}{\beta}}_C^{**} - \underset{\sim}{w}_j\hat{\underset{\sim}{\beta}}^*|}{m_i}$ will determine which model has a better fit relative to the combined estimate $\hat{\underset{\sim}{\beta}}^*$. If $s < t$, then the treatment model has a better fit and hence, it will be used to predict the potential outcome $Y^{(1)}$ of the control units through the estimated $\tilde{Z}_j$. On one hand, if $s > t$, then the control model may be used to estimate the missing $Y^{(0)}$ of the treatment units through $\tilde{U}_j$. The rationale in this procedure is that the model whose predicted value from the combined coefficients is closest to the ones in the classical model would provide a better prediction for the missing potential outcome. This matching and "swapping" method allows estimation of $(Y^{(1)}, Y^{(0)})$ from which the parameter of interest $E(Y^{(1)} - Y^{(0)})$ is derived.

3.3 Estimation

To estimate the parameter or the *effect size* of interest, ATE, defined by $E(Y^{(1)}-Y^{(0)})$, we consider the fitted models under the treatment and control groups in each G^* subgroups. For a subgroup i, one of the four cases could occur:

(1) There is a good fit, as defined in Section 3.2, in the control group but not the treatment group. In this instance, for each treatment unit $j = 1, 2, ..., n_i$, $\delta_j = U_j - \tilde{U}_j$ is computed as the treatment effect. To derive its distribution, note that $\hat{\underset{\sim}{\beta}}^*$ is computed using standard meta-analysis formula so that

$$\begin{aligned}
\hat{\underset{\sim}{\beta}}^* &= \left(\sum_{k=T,C} \Sigma_k^{-1}\right)^{-1} \left(\sum_{k=T,C} (\Sigma_k^{-1} \hat{\underset{\sim}{\beta}}_k^{**})\right) \\
&= \left(\Sigma_T^{-1} + \Sigma_C^{-1}\right)^{-1} \left(\Sigma_T^{-1} \hat{\underset{\sim}{\beta}}_T^{**} + \Sigma_C^{-1} \hat{\underset{\sim}{\beta}}_C^{**}\right) \\
&= \left(\frac{S_{vv}}{\sigma_T^2} + \frac{S_{ww}}{\sigma_C^2}\right)^{-1} \left(\frac{S_{vv}}{\sigma_T^2}\mathbf{D}_T \underset{\sim}{U} + \frac{S_{ww}}{\sigma_C^2}\mathbf{D}_C \underset{\sim}{Z}\right) \\
&= \left(\frac{S_{vv}}{\sigma_T^2} + \frac{S_{ww}}{\sigma_C^2}\right)^{-1} \frac{S_{vv}}{\sigma_T^2}\mathbf{D}_T \underset{\sim}{U} + \left(\frac{S_{vv}}{\sigma_T^2} + \frac{S_{ww}}{\sigma_C^2}\right)^{-1} \frac{S_{ww}}{\sigma_C^2}\mathbf{D}_C \underset{\sim}{Z} \quad (3.17)
\end{aligned}$$

where $\mathbf{D}_T$ is the $p \times n_i$ matrix of the $(\mathbf{V}^{*'}\mathbf{V}^*)^{-\mathbf{1}}\mathbf{V}^{*'}$ with the first row removed from the calculation of the treatment group data while $\mathbf{D}_C$ is the $p \times m_i$ matrix of the $(\mathbf{W}^{*'}\mathbf{W}^*)^{-\mathbf{1}}\mathbf{W}^{*'}$ from the control group data with the first row removed. To simplify the derivation, quantities

$$\mathbf{D}_T^* = \left(\frac{S_{vv}}{\sigma_T^2} + \frac{S_{ww}}{\sigma_C^2}\right)^{-1} S_{vv}\mathbf{D}_T = \begin{pmatrix} d_{T11} & \dots & d_{Tn_i1} \\ \vdots & \vdots & \vdots \\ d_{T1p} & \dots & d_{Tn_ip} \end{pmatrix} = \begin{pmatrix} \underset{\sim}{d}_{T1} \\ \vdots \\ \underset{\sim}{d}_{Tp} \end{pmatrix} \quad (3.18)$$

and

$$\mathbf{D}_C^* = \left(\frac{S_{vv}}{\sigma_T^2} + \frac{S_{ww}}{\sigma_C^2}\right)^{-1} S_{ww}\mathbf{D}_C = \begin{pmatrix} d_{C11} & \dots & d_{Cm_i1} \\ \vdots & \vdots & \vdots \\ d_{C1p} & \dots & d_{Cm_ip} \end{pmatrix} = \begin{pmatrix} \underset{\sim}{d}_{C1} \\ \vdots \\ \underset{\sim}{d}_{Cp} \end{pmatrix} \quad (3.19)$$

of dimensions $p \times n_i$ and $p \times m_i$, respectively, are defined where row vectors $\underset{\sim}{d}_{Tl} = \begin{pmatrix} d_{T1l} & \dots & d_{Tn_il} \end{pmatrix}$ and $\underset{\sim}{d}_{Cl} = \begin{pmatrix} d_{C1l} & \dots & d_{Cn_il} \end{pmatrix}$ for $l = 1, 2, \dots, p$. This means that

$$\underset{\sim}{\hat{\beta}}^* = \frac{1}{\sigma_T^2}\mathbf{D}_T^*\underset{\sim}{U} + \frac{1}{\sigma_C^2}\mathbf{D}_C^*\underset{\sim}{Z} \quad (3.20)$$

with dimension $p \times 1$. Hence,

$$\begin{aligned} \tilde{U}_j &= \begin{pmatrix} 1 & (v_{j1} - \bar{w}_1) & \dots & (v_{jp} - \bar{w}_p) \end{pmatrix} \begin{pmatrix} \bar{Z} \\ \frac{1}{\sigma_T^2}\mathbf{D}_T^*\underset{\sim}{U} + \frac{1}{\sigma_C^2}\mathbf{D}_C^*\underset{\sim}{Z} \end{pmatrix} \\ \tilde{U}_j &= \begin{pmatrix} 1 & (v_{j1} - \bar{w}_1) & \dots & (v_{jp} - \bar{w}_p) \end{pmatrix} \begin{pmatrix} \bar{Z} \\ \frac{1}{\sigma_T^2}\underset{\sim}{d}_{T1}\underset{\sim}{U} + \frac{1}{\sigma_C^2}\underset{\sim}{d}_{C1}\underset{\sim}{Z} \\ \vdots \\ \frac{1}{\sigma_T^2}\underset{\sim}{d}_{Tp}\underset{\sim}{U} + \frac{1}{\sigma_C^2}\underset{\sim}{d}_{Cp}\underset{\sim}{Z} \end{pmatrix}. \end{aligned}$$

Therefore, $\tilde{U}_j$ can similarly be defined by the following three quantities:

$$\tilde{U}_j = \hat{\beta}'_{0,C} + \hat{\beta}_1(v_{j1} - \bar{w}_1) + \ldots + \hat{\beta}_p(v_{jp} - \bar{w}_p) \tag{3.21}$$

$$\tilde{U}_j = \begin{pmatrix} 1 & (v_{j1} - \bar{w}_1) & \ldots & (v_{jp} - \bar{w}_p) \end{pmatrix} \begin{pmatrix} \bar{Z} \\ \underset{\sim}{\hat{\beta}^*} \end{pmatrix} \tag{3.22}$$

$$\begin{aligned} \tilde{U}_j = {} & \bar{Z} + \frac{1}{\sigma_T^2}(v_{j1} - \bar{w}_1)\underset{\sim}{d}_{T1}\underset{\sim}{U} + \frac{1}{\sigma_C^2}(v_{j1} - \bar{w}_1)\underset{\sim}{d}_{C1}\underset{\sim}{Z} + \ldots + \\ & \frac{1}{\sigma_T^2}(v_{jp} - \bar{w}_p)\underset{\sim}{d}_{Tp}\underset{\sim}{U} + \frac{1}{\sigma_C^2}(v_{jp} - \bar{w}_p)\underset{\sim}{d}_{Cp}\underset{\sim}{Z}. \end{aligned} \tag{3.23}$$

For the distributional properties of $\tilde{U}_j$, it can be easily established that

$$E(\tilde{U}_j) = \begin{pmatrix} 1 & (v_{j1} - \bar{w}_1) & \ldots & (v_{jp} - \bar{w}_p) \end{pmatrix} \begin{pmatrix} \beta'_{0,C} \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \tag{3.24}$$

$$\begin{aligned} Var(\tilde{U}_j) &= v_j^{*'} \begin{pmatrix} \frac{\sigma_C^2}{m_i} & \frac{\sum_{q=1}^{m_i} d_{Cq1}}{m_i} & \frac{\sum_{q=1}^{m_i} d_{Cq2}}{m_i} & \ldots & \frac{\sum_{q=1}^{m_i} d_{Cqp}}{m_i} \\ \frac{\sum_{q=1}^{m_i} d_{Cq1}}{m_i} & & & & \\ \frac{\sum_{q=1}^{m_i} d_{Cq2}}{m_i} & & \left(\sum_{i=T,C} \Sigma_i^{-1}\right)^{-1} & & \\ \vdots & & & & \\ \frac{\sum_{q=1}^{m_i} d_{Cqp}}{m_i} & & & & \end{pmatrix} v_j^* \\ &= v_j^{*'} \Lambda_C v_j^* \end{aligned} \tag{3.25}$$

since the covariance of $\bar{Z}$ and any parameter $\hat{\beta}_l, l = 1, 2, \ldots, p$ is given by

$$\begin{aligned} Cov\left(\bar{Z}, \hat{\beta}_l\right) &= Cov\left(\bar{Z}, \frac{1}{\sigma_T^2}\underset{\sim}{d}_{Tl}\underset{\sim}{U} + \frac{1}{\sigma_C^2}\underset{\sim}{d}_{Cl}\underset{\sim}{Z}\right) \\ &= Cov\left(\frac{Z_1 + \ldots + Z_{m_i}}{m_i}, \frac{1}{\sigma_C^2}\left(d_{C1l}Z_1 + \ldots + d_{Cm_il}Z_{m_i}\right)\right) \\ &= \frac{1}{m_i}\frac{1}{\sigma_C^2}\left[d_{C1l}Var(Z_1) + \ldots + d_{Cm_il}Var(Z_{m_i})\right] \\ &= \frac{1}{m_i}\sum_{q=1}^{m_i} d_{Cql}. \end{aligned} \tag{3.26}$$

Under this distributional properties of $\tilde{U}_j$, the distribution of the n_i differences between observed treatment response and the estimated control response, $\underset{\sim}{\delta}$, is derived as

$$\underset{\sim}{\delta} = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_{n_i} \end{pmatrix} \sim N_{n_i}\left(\alpha\underset{\sim}{1}, \Psi\right)$$

with $\alpha = \beta_{0,T} - \beta_{0,C}$ and the elements of the Ψ matrix given by

$$Var(\delta_j) = \sigma_T^2 + \underset{\sim}{v}_j^{*'}\Lambda_c\underset{\sim}{v}_j^* - 2\underset{\sim}{v}_j^{'}\underset{\sim}{d}_{Tj}^* \tag{3.27}$$

$$Cov(\delta_j, \delta_{j'}) = -\underset{\sim}{v}_{j'}^{'}\underset{\sim}{d}_{Tj}^* - \underset{\sim}{v}_j^{'}\underset{\sim}{d}_{Tj'}^* + \underset{\sim}{v}_j^{*'}\Lambda_c^*\underset{\sim}{v}_{j'}^* + \frac{1}{m_i}\sum_{q=1}^{m_i}\underset{\sim}{v}_j^{'}d_{Cq}^* + \frac{1}{m_i}\sum_{q=1}^{m_i}\underset{\sim}{v}_{j'}^{'}d_{Cq}^*. \tag{3.28}$$

where

$$\underset{\sim}{v}_j = \begin{pmatrix} (v_{j1} - \bar{w}_1) \\ \vdots \\ (v_{jp} - \bar{w}_p) \end{pmatrix}, \underset{\sim}{d}_{Tj}^* = \begin{pmatrix} d_{Tj1} \\ \vdots \\ d_{Tjp} \end{pmatrix}, \underset{\sim}{d}_{Cj}^* = \begin{pmatrix} d_{Cj1} \\ \vdots \\ d_{Cjp} \end{pmatrix},$$

$$\Lambda_C^* = \begin{pmatrix} \frac{\sigma_C^2}{m_i} & 0 & 0 & \dots & 0 \\ 0 & & & & \\ \vdots & & \left(\sum_{i=T,C} \Sigma_i^{-1}\right)^{-1} & & \\ 0 & & & & \end{pmatrix}$$

and $\underset{\sim}{v}_j^*$ and Λ_C are as previously defined. Details of the derivation are shown in Appendix A.

With the n_i differences in response for each j, $\delta_j's$, calculated for a specific subgroup, an optimal estimate of the effect size for a subgroup i, Δ_i in the treatment group can be provided by

$$\hat{\Delta}_i = \frac{\underset{\sim}{1}'\Psi^{-1}\underset{\sim}{\delta}}{\underset{\sim}{1}'\Psi^{-1}\underset{\sim}{1}} \sim N\left(\alpha, \sigma_i^{*2}\right) \tag{3.29}$$

where $\sigma_i^{*2} = \dfrac{1}{\underset{\sim}{1}'\Psi^{-1}\underset{\sim}{1}}$.

(2) There is a good fit in the treatment group but not the control group. In this case, good estimates will be derived using the predictive model under the treatment group over the control. For each control unit $j = 1, 2, ..., m_i$, $\delta_j^* = \tilde{Z}_j - Z_j$ can be calculated as the treatment effect under the same rationale as Case 1. Ultimately, a corresponding effect size for subgroup i in the control group may be computed as

$$\hat{\boldsymbol{\Delta}}_i = \frac{\underset{\sim}{1}'\Psi^{*-1}\underset{\sim}{\delta}^*}{\underset{\sim}{1}'\Psi^{*-1}\underset{\sim}{1}} \sim N\left(\alpha, \sigma_i^{*2}\right) \tag{3.30}$$

where $\sigma_i^{*2} = \dfrac{1}{\underset{\sim}{1}'\Psi^{*-1}\underset{\sim}{1}}$ and Ψ^* have diagonal elements $\sigma_C^2 + \underset{\sim}{w}_j^{*'}\Lambda_T\underset{\sim}{w}_j^* - 2\underset{\sim}{w}_j'\underset{\sim}{d}_{Cj}^*$ and off-diagonal elements $-\underset{\sim}{w}_{j'}'\underset{\sim}{d}_{Cj}^* - \underset{\sim}{w}_j'\underset{\sim}{d}_{Cj'}^* + \underset{\sim}{w}_j^{*'}\Lambda_T\underset{\sim}{w}_{j'}^* + \dfrac{1}{n}\sum_{q=1}^{n}\underset{\sim}{w}_j'\underset{\sim}{d}_{Tq}^* + \dfrac{1}{n}\sum_{q=1}^{n}\underset{\sim}{w}_{j'}'\underset{\sim}{d}_{Tq}^*$

where

$$\Lambda_T = \begin{pmatrix} \frac{\sigma_T^2}{n_i} & \frac{\sum_{q=1}^{n_i} d_{Tq1}}{n_i} & \cdots & \frac{\sum_{q=1}^{n_i} d_{Tqp}}{n_i} \\ \frac{\sum_{q=1}^{n_i} d_{Tq1}}{n_i} & & & \\ \vdots & \left(\sum_{i=T,C} \Sigma_i^{-1}\right)^{-1} & & \\ \frac{\sum_{q=1}^{n_i} d_{Tqp}}{n_i} & & & \end{pmatrix}_{(p+1)\times(p+1)}$$

$$\Lambda_T^* = \begin{pmatrix} \frac{\sigma_T^2}{m_i} & 0 & 0 & \cdots & 0 \\ 0 & & & & \\ \vdots & \left(\sum_{i=T,C} \Sigma_i^{-1}\right)^{-1} & & & \\ 0 & & & & \end{pmatrix}_{(p+1)\times(p+1)}$$

$$\underset{\sim}{w}_j^* = \begin{pmatrix} 1 \\ (w_{j1} - \bar{v}_1) \\ \vdots \\ (w_{jp} - \bar{v}_p) \end{pmatrix}_{((p+1)\times 1)}, \underset{\sim}{w}_j = \begin{pmatrix} (w_{j1} - \bar{v}_1) \\ \vdots \\ (w_{jp} - \bar{v}_p) \end{pmatrix}_{(p\times 1)}, \underset{\sim}{d}_{Cj}^* = \begin{pmatrix} d_{Cj1} \\ \vdots \\ d_{Cjp} \end{pmatrix}_{(p\times 1)}$$

The derivations follow from the same formulation as Appendix A of Case 1.

(3) There is a good fit in both groups. If the model under the control group is deemed better, the method and estimates presented in Case 1 will be applied. Otherwise, those derived in Case 2 will be used to calculate for the subgroup' respective effect size, δ_j.

(4) There is no good fit in neither groups; that is, the available covariates are not good predictors of the response in the said subgroup. This implies that there

is no systematic difference in the covariates that will affect the difference in treatments realized. This is further justified by the existing balance in the subgroup with respect to the covariates. Thus, the effect size for a subgroup i under this circumstance is given by

$$\hat{\Delta}_i = \bar{U} - \bar{Z} \tag{3.31}$$

where $\bar{U} = \frac{1}{n_i}\sum_{i=1}^{n_i} U_i$ and $\bar{Z} = \frac{1}{m_i}\sum_{j=1}^{m_i} Z_i$ with a corresponding estimated variance of

$$\hat{\sigma}_i^{*2} = Var\left(\bar{U} - \bar{Z}\right) = \frac{\hat{\sigma}_T^2}{n_i} + \frac{\hat{\sigma}_C^2}{m_i}. \tag{3.32}$$

Given these four cases, the effect size $\hat{\Delta}_i$ and standard error for each subgroup may be calculated. To estimate the ATE, $E[Y^{(1)} - Y^{(0)}]$, meta-analysis procedures are proposed to combine the G^* effect sizes [9].

Consider the G^* independent subgroups with the ith subgroup having an estimated effect size of $\hat{\Delta}_i$ and estimated variance of $\hat{\Delta}_i$ of $\hat{\sigma}_i^{*2}, i = 1, 2, \ldots, G^*$ as computed above. Assuming homogeneous effect sizes across groups, that is, $\Delta_1 = \ldots = \Delta_{G^*} = \Delta$ where Δ is the common population effect size, then the combined estimate of the average treatment effect is given by the weighted combination of the $\hat{\Delta}_i$'s defined as

$$\hat{\Delta} = \frac{\sum_{i=1}^{G^*} w_i \hat{\Delta}_i}{\sum_{i=1}^{G^*} w_i},$$

where w_i is a non-negative weight assigned to the ith group. The choice of w_i is one that makes the $Var(\hat{\Delta})$ as smallest as possible and this is given by $w_i = \frac{1}{\sigma_i^{2*}}$ for

$i = 1, 2, \ldots, G^*$. However, this value is unknown and hence must be estimated by $\hat{\sigma}_i^{*2}$ resulting to a modified weighted combination of average treatment effect size of

$$\tilde{\Delta} = \frac{\sum_{i=1}^{G^*} \hat{\Delta}_i / \hat{\sigma}_i^{*2}}{\sum_{i=1}^{G^*} 1/\hat{\sigma}_i^{*2}}, \tag{3.33}$$

with an estimated variance of

$$\hat{\sigma}^2(\tilde{\Delta}) = \frac{1}{\sum_{i=1}^{G^*} 1/\hat{\sigma}_i^{*2}}. \tag{3.34}$$

It must be noted that these pooled estimates are based on the assumption of homogeneous effect sizes. To test the validity of this assumption, a chi-square test on the data from the G^* subgroups is applied. Using $\tilde{\Delta}$, the test statistic χ^2 is calculated where

$$\chi^2 = Q_c = \sum_{i=1}^{G^*} \frac{(\hat{\Delta}_i - \tilde{\Delta})^2}{\hat{\sigma}_i^{*2}} = \sum_{i=1}^{G^*} \frac{\hat{\Delta}_i^2}{\hat{\sigma}_i^{*2}} - \frac{\left(\sum_{i=1}^{G^*} \hat{\Delta}_i / \hat{\sigma}_i^{*2}\right)^2}{\sum_{i=1}^{G^*} 1/\hat{\sigma}_i^{*2}} \tag{3.35}$$

where the null hypothesis is rejected if H_o if $\chi^2 > \chi^2_{G^*-1,\alpha}$. The test statistic Q_c is also called the Cochran's test statistic [9]. Upon the non-rejection of the null, one is able to pool the G^* effect sizes as discussed above. However, when there is lack of homogeneity, it may be an indicator that some of the covariates have behaved differently in each of the G^* subgroups which implies that combining effect sizes as described above may not be suitable. In such a case, a random effects model is considered.

In the one-way random effects model, $\tilde{\Delta}_i \sim N(\Delta, \psi^2 + \sigma_i^{*2})$, where $\psi^2 \geq 0$ refers to the variability between the subgroups constructed or the heterogeneity

parameter [10]. For known ψ^2 and $\sigma_i^{*2}, i = 1, 2, \ldots, G^*$, the maximum likelihood estimator of $\tilde{\Delta}$ is given by

$$\tilde{\Delta} = \frac{\sum_{i=1}^{G^*} \hat{\Delta}_i/(\psi^2 + \hat{\sigma}_i^{*2})}{\sum_{i=1}^{G^*} 1/(\psi^2 + \hat{\sigma}_i^{*2})}, \tag{3.36}$$

with a corresponding variance of $\tilde{\Delta}$ given by

$$\hat{\sigma}^2(\tilde{\Delta}) = \frac{1}{\sum_{i=1}^{G^*} 1/(\psi^2 + \hat{\sigma}_i^{*2})}. \tag{3.37}$$

A wide literature is available for the estimation of ψ^2 [10]. Among these include the maximum likelihood estimator (MLE), restricted maximum likelihood estimator (RMLE), ANOVA-type estimator by Rao, Kaplan and Cochran [23], Mandel-Paule estimator [19] and Sidik and Jonkman estimator [33]. In this paper, a commonly used estimator for ψ^2 called the Dersimonian and Laird estimator [6] is used. The estimator is defined by

$$\hat{\psi}^2_{DSL} = \frac{Q_c - (G^* - 1)}{\sum_{i=1}^{G^*} \hat{r}_i - \frac{\sum_{i=1}^{G^*} \hat{r}_i^2}{\sum_{i=1}^{G^*} \hat{r}_i}}, \tag{3.38}$$

where $\hat{r}_i = 1/\hat{\sigma}_i^{*2}$ and Q_c is the test Cochran's test statistic defined and used above for the test of homogeneity of effect sizes. The pooled estimates derived under these meta-analysis methods represent the estimate for the ATE.

Chapter 4

Applications

4.1 Overview

The proposed procedure presented in the preceding chapter is applied on different simulated data sets. In these simulations, the response variable is generated from covariates with the same increased effect. Also. scenarios where the response variable is based on covariates having mixed, either increasing or decreasing, effect on the outcome. It is likewise implemented on real data sets; one involving a study on the social sciences and another on the medical field. The analyses on these simulated and real data sets using the proposed methods are compared with known propensity score methods.

4.2 Simulation with Covariates Having Increased Effects on Y_i

For each simulation run, we generate two continuous covariates for $n = 2,200$ records such that $x_i = (x_{i1}, x_{i2})^{'}$ where $\mu = (10, 10)^{'}$ and Σ has variances equal to 5 with correlation 0.5. Three categorical covariates, each with two levels, were also generated such that $x_{i3} \sim Be(0.7), x_{i4} \sim Be(0.3)$ and $x_{i5} \sim Be(0.6)$. The response Y is generated for all i as $Y_i = x_{i1} + x_{i2} + 0.1x_{i3} + 0.1x_{i4} + 0.1x_{i5} + \tau D_i + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$, where τ is the population average treatment effect. Three mechanisms

are considered for assigning treatment, including:

(i) assignment depends only on x_1 and x_3

(ii) assignment depends only on x_2 and x_4

(iii) assignment depends equally on x_1 and x_2, as well as equally on x_3 , x_4 and x_5.

Without loss of generality for additive treatment effects, $\tau = 0$ for all simulations. Scenarios 1 and 2 reflect the instances when only a subset of the relevant predictors of the outcome affect the treatment assignment, while the last scenario looks into the circumstance when all variables affect both the response and the treatment assignment.

4.2.1 Scenario 1: assignment depends only on x_1 and x_3

In this simulation, treatments are assigned from Bernoulli distributions where

$$logit(P(D_i = 1|x_i)) = -7.8 + 0.5x_{i1} + 0.5x_{i3}. \tag{4.1}$$

Thus, treatment assignment depends only on x_1 and x_3 . The simulation setup generates $n_C = 1,737$ control units and $n_T = 463$ treatment units.

A total of $G = 2^3 = 8$ groups are created based on the levels of the categorical variables x_3, x_4 and x_5. The multivariate test on the equality of the mean vector of continuous variables with unequal Σ also suggests that there is an imbalance in all groups; hence, the swapping mechanism is applied. To perform this, regression models of the response Y on x_1 and x_2 are built using treatment units only as well as control units only for each of the eight groups. To determine which model to

use for the swapping mechanism in order to predict the missing potential outcome, the *adhoc* method presented in Section 3.2 is implemented for each subgroup. In the 8 groups, the treatment model is used twice so that $Y^{(1)}$ is predicted for the control units; thus,case 2 of treatment effect estimation is implemented. Also, case 1 is implemented for the rest of the 6 subgroups. Upon testing the homogeneity of the 8 treatment effects, non-rejection of the null is concluded and therefore, effect sizes can be combined via meta-analysis formula.

To compare the performance of the generated average treatment effect estimate through the proposed method, propensity score-based estimates are likewise calculated. Results of these are shown in Table 4.1.

Method	**ATE**	**SE(ATE)**	**95% C.I.**
Swapping	-0.0129	0.0884	(-0.1858 , 0.1624)
Pair Matching	1.8734	0.5782	(0.7401, 3.0067)
Stratification	1.4999	0.5081	(0.5040, 2.4957)
IPTW	1.7369	0.9778	(-0.1796 , 3.6534)

Table 4.1: Average Treatment Effect Estimates of Simulation 1, Scenario 1

The results of the different ATE estimation procedures indicate the superiority of the proposed method over the propensity score-based methods. First, the estimate derived from "swapping" is the most stable as evident in its standard error (SE), which is smallest among the four estimates. Second, the absolute value of the estimate generated from the proposed method is closest to the true ATE, $\tau = 0$. This suggests that the proposed method provides the most reliable average treatment

effect estimate. Third, the confidence interval formed through these estimates does not include $\tau = 0$ except for the proposed model and IPTW estimate. Hence, both procedures are able to include the true ATE; however, the former is still preferred over latter taking into account their standard errors and absolute differences from $\tau = 0$.

4.2.2 Scenario 2: assignment depends only on x_2 and x_4

In this scenario, treatment units are assigned from Bernoulli distributions where

$$logit(P(D_i = 1|x_i)) = -7.8 + 0.5x_{i2} + 0.5x_{i4}. \tag{4.2}$$

This setup results to a total of $n_C = 1,752$ control units and $n_T = 448$ treatment units. Eight subgroups are then formed that are assured to be balanced with respect to categorical variables but have been established to be unbalanced with respect to the continuous variables x_1 and x_2. This results to the implementation of the "swapping" mechanism. Based on the *adhoc* method suggested in choosing the groups for swapping, the control group model is selected seven times for imputing $Y^{(0)}$ for the treatment units. This means that treatment effect estimate presented in case 1 has been implemented seven times. The null hypothesis of the test of homogeneity is not rejected; therefore, the eight treatment effect estimates calculated can be combined using meta-analysis. Table 4.2 presents the result of the proposed method along with the other propensity score-based estimates for comparison under this scenario.

Method	ATE	SE(ATE)	95% C.I.
Swapping	0.0742	0.0999	(-0.1217, 0.2725)
Pair Matching	1.5832	0.6220	(0.3641, 2.8023)
Stratification	1.3596	0.5154	(0.3494,2.3698)
IPTW	-0.1589	0.9565	(-2.0336, 1.7158)

Table 4.2: Average Treatment Effect Estimates of Simulation 1, Scenario 2

The results of the methods applied on the simulated data for the second scenario show that the proposed method provides an estimate with the smallest standard error and is closest in magnitude to the true average treatment effect, $\tau = 0$. While the IPTW estimate is comparable to the derived estimate from "swapping" in its ability to capture $\tau = 0$, the proposed estimate is still preferred since its standard error is smallest among all estimates.

4.2.3 Scenario 3: treatment assignment depends equally on x_1 and x_2 , as well as equally on x_3 , x_4 , and x_5

In this scenario, treatment assignment is made from Bernoulli distributions where

$$logit(P(D_i = 1x_i)) = -7.8 + 0.25x_{i1} + 0.25x_{i2} + 0.15x_{i3} + 0.15x_{i4} + 0.15x_{i5}. \quad (4.3)$$

This formulation indicates that the treatment assignment depends on all the available covariates. This generates $n_C = 1,810$ control units and $n_T = 390$ treatment units. The balance with respect to the categorical variable is guaranteed when distinct subgroups are created based on the different levels of the categorical variables.

The multivariate test on the vector of continuous variables for these subgroups suggest the need for "swapping" models for the estimation of missing potential outcome. The estimated regression model suggests that the control model is used seven times; thus, case 1 was implemented seven times for the calculation of treatment effect estimates. It is further established that the treatment effects are homogeneous and can be combined accordingly. The results of the proposed method is shown in Table 4.3.

Method	**ATE**	**SE(ATE)**	**95% C.I.**
Swapping	-0.1101	0.1596	(-0.4519, 0.2079)
Pair Matching	1.9155	0.6316	(0.6776, 3.1534)
Stratification	1.7183	0.4485	(0.8392, 2.5974)
IPTW	0.8607	1.3399	(-1.7655, 3.4869)

Table 4.3: Average Treatment Effect Estimates of Simulation 1, Scenario 3

Table 4.3 shows similar conclusions as the previous scenarios. An estimate with a smaller standard error and an absolute value closest to $\tau = 0$ is realized from the proposed method, which is as desired. The calculated confidence interval based on the "swapping" method also includes the true ATE.

In general, the use of covariate information, as in the "swapping" method, is preferred over the propensity score-based procedures due to the stability of estimates derived and its closeness to the true average treatment effect. More specifically, estimates derived from the nearest neighbor matching are based on matches that have varied covariate feature upon inspection. Also, in the stratification method,

some matched groups remain unbalanced with respect to some covariate despite redefinition of the propensity scores. This may be an indicator that the calculated within-stratum estimates, and consequently, the average treatment effect estimate, are biased. Lastly, for the IPTW, although it is able to reflect the true treatment effect, relatively larger standard errors are realized. This results to wider confidence intervals and larger p-values that provide unreliable conclusions.

However, it is noted that the three treatment assignment mechanism induce low prevalence of the treatment units. As a result, one-to-one matching and stratification have the least desirable results due to the elimination of several units in the estimation of ATE. This also suggests that the data set realized after propensity score matching via stratification and one-to-one pairing may have minimal overlap with the data set used in the proposed procedure and IPTW. Hence, comparable results are generated for the "swapping" method and IPTW but not for stratification and one-to-one matching.

Running the procedure on $k = 1,500$ simulated data, the type I error at $\alpha = 0.05$ and power of the test at $\tau = 0.25$ and $\tau = 0.5$ are derived. The results suggests that the Type I error is a little inflated; however, the test has a good power of identifying the true average treatment effect. Table 4.4 shows the results of this procedure.

Scenario	**Type I Error**	**Power** ($\tau = 0.25$)	**Power** ($\tau = 0.5$)
Scenario 1	0.0604	0.7327	0.9121
Scenario 2	0.0574	0.5625	0.8793
Scenario 3	0.0554	0.6938	0.8920

Table 4.4: Type I Error and Power of the Test

Given the simulated data sets for all three scenarios, an approximation of the standard error and confidence interval of the ATE estimator is derived via bootstrapping. Through this procedure, we are able to obtain estimates of characteristics of ATE by generating re-samples of size $1,200$ and repeating this procedure sufficiently large number of times, B. In this case, $B = 2,000$. The results suggest that low standard errors are generated through the proposed method. Also, the confidence interval is able to capture the true ATE, $\tau = 0$ as shown in Table 4.5.

Scenario	**Bootstrap Estimate**	**Standard Error**	**Confidence interval**
Scenario 1	0.0014	0.0027	(-0.0039, 0.0067)
Scenario 2	-0.0039	0.0028	(-0.0094, 0.0017)
Scenario 3	-0.0044	0.0029	(-0.0101, 0.0014)

Table 4.5: Bootstrap Estimates of ATE under Simulation 1

To further look into the performance of the swapping method, coverage probabilities were calculated for each scenario considered. Given the data set, re-samples of size $1,200$ were generated and an estimate of the ATE along with its corresponding CI are derived. This procedure is repeatedly performed 500 times. The coverage probability is then calculated by determining the proportion of confidence inter-

vals containing the true ATE $\tau = 0$ among the 500 confidence intervals derived via bootstrapping. Results are shown in Table 4.6.

Scenario	**Coverage Probability**
Scenario 1	86.4
Scenario 2	87.2
Scenario 3	86.8

Table 4.6: Coverage Probability of Swapping Method under Simulation 1

Based on the scenarios from which the simulated data are generated, it is shown that the proposed procedure is able to capture the true average treatment effect as well as provide a good power of the test. In these circumstances, the "swapping" method performs better than the known propensity score methods as shown in the small absolute differences between its estimated value and true value. However, the coverage probabilities also suggest that there is a need for some mechanism to correct the inaccuracies of these procedure.

4.3 Simulation with Covariates Having Mixed Effects on Y_i

The same specifications on the covariates are generated for the second simulation. The response variable, however, is defined by covariates having increased and decreased effect on its value. In this simulation, the response Y is generated for all i as $Y_i = 0.9x_{i1} - 0.5x_{i2} + 0.1x_{i3} - 0.3x_{i4} + 0.15x_{i5} + \tau D_i + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$ and $\tau = 0$ without loss of generality. The same three scenarios on treatment assignment as Section 4.2 are likewise considered.

4.3.1 Scenario 1: assignment depends only on x_1 and x_3

Under the scenario that the treatment assignment is based on a Bernoulli distribution with $logit(P(D_i = 1|x_i)) = -7.8 + 0.5x_{i1} + 0.5x_{i3}$, $n_C = 1,926$ units are assigned to the control group while the remaining $n_T = 274$ units are given the treatment. With the eight subgroups formed, "swapping" method is applied since balance is not established with respect to its continuous covariates. Separate regression models are built and its estimated parameters are combined through multivariate meta-analysis under the common parameter $\underset{\sim}{\beta}^*$ assumption. Based on the *adhoc* method, smaller average absolute differences in the control units are realized for six groups; hence, the potential outcome $Y^{(0)}$ is imputed for the treatment units. Meta-analysis methods are implemented to combine these subgroup estimates since the null hypothesis for the test of homogeneity is not rejected. Table 4.7 shows the performance of the proposed method on the simulated data in comparison to the known propensity score methods.

Method	**ATE**	**SE(ATE)**	**95% C.I.**
Swapping	0.0169	0.0882	(-0.1599, 0.1978)
Pair Matching	0.3705	0.1944	(-0.0105, 0.7515)
Stratification	0.2752	0.6932	(-1.0845, 1.6339)
IPTW	0.0538	0.4763	(-0.8800, 0.9873)

Table 4.7: Average Treatment Effect Estimates of Simulation 2, Scenario 1

Based on the results above, it could be deduced that the proposed method has the least absolute value difference from the true average treatment effect $\tau = 0$. It

also has the smallest standard error, as desired, as this indicates that the method produces the most stable estimate. However, it can be deduced from the confidence intervals that the "swapping" method and the propensity score methods are able to capture the true ATE. While this suggests that the common propensity scores will generate valid conclusions, the proposed method is preferred due to its close estimate to the population average treatment effect.

4.3.2 Scenario 2: assignment depends only on x_2 and x_4

Another scenario considered is when the treatment assignment follows a Bernoulli distribution where its probability of assigning a unit to the treatment group is given by $logit(P(D_i = 1|x_i)) = -7.8 + 0.5x_{i2} + 0.5x_{i4}$. This treatment assignment generates $n_C = 1,958$ control units and $n_T = 242$ treatment units. "Swapping" is implemented on each subgroup created by performing regression analysis, multivariate meta-analysis and the adhoc method. The subgroups are combined using meta-analysis to generate an average treatment effect estimate, as shown in Table 4.8, along with the results of the common propensity score methods implemented on the simulated data.

Method	**ATE**	**SE(ATE)**	**95% C.I.**
Swapping	0.0120	0.0843	(-0.1627, 0.1788)
Pair Matching	-0.0998	0.1885	(-0.4623, 0.2697)
Stratification	0.1584	0.5839	(-0.9861, 1.3028)
IPTW	0.1371	0.1501	(-0.1501, 0.4313)

Table 4.8: Average Treatment Effect Estimates of Simulation 2, Scenario 2

In all four methods, the true ATE is captured by the confidence intervals. The "swapping" method is preferred though since it has the closest absolute difference from the true effect and it reflects the smallest standard error for the estimates.

4.3.3 Scenario 3: treatment assignment depends equally on x_1 and x_2 , as well as equally on x_3 , x_4 and x_5

The last simulation scenario lies on the idea that the treatment assignment mechanism depends on drawing from a Bernoulli distribution with the probability of assigning a unit to the treatment group depending on all available covariates. One data set generated under this scenario contains $n_C = 1,982$ observations for the control group and $n_T = 218$ for the treatment group. The same procedures as the first two scenarios were implemented on the eight balanced subgroups formed based on the categorical variables and the results are reflected in Table 4.9 along with the estimates calculated via propensity score methods.

Method	**ATE**	**SE(ATE)**	**95% C.I.**
Swapping	-0.0651	0.0879	(-0.2460, 0.1181)
Pair Matching	0.2637	0.2111	(-0.1501, 0.6775)
Stratification	0.0840	0.5271	(-0.9491, 1.1171)
IPTW	-0.0120	0.1623	(-0.3301, 0.3061)

Table 4.9: Average Treatment Effect Estimates of Simulation 2, Scenario 3

Table 4.9 shows that the proposed method and the IPTW have comparable estimates for the average treatment effect and their standard error. These estimates

are also the closest values to $\tau = 0$. However, it can be deduced from the confidence intervals that the all methods are able to capture the true ATE.

As a summary, the estimates generated from the "swapping" are generally preferred under these scenarios since it provides the smallest absolute value difference from $\tau = 0$ and standard errors. However, it should be mentioned that the same treatment assignment mechanisms have been imposed on Model 2 as Model 1; thus, much smaller number of treatment units are generated. Under the stratification and one-to-one matching procedures, several of the control units will be discarded and consequently, distinct data sets will be used for the estimation of the ATE for the "swapping", stratification and one-to-one matching procedure. Since IPTW estimates the ATE only as a function of the $e(x)$ and does not discard units in the estimation procedure, it reasonably compares well to the "swapping" method.

To calculate the type I error of the procedure $\alpha = 0.05$ and its power with $\tau = 0.25$ and $\tau = 0.5$, $k = 1,500$ simulated data are derived. Table 4.10 reflects the results of the said power analysis. The results suggests that the Type I error is close to the set $\alpha = 0.05$ and the power substantially increases to roughly 87% when the true value of the parameter is $\tau = 0.5$.

Scenario	**Type I Error**	**Power** ($\tau = 0.25$)	**Power** ($\tau = 0.5$)
Scenario 1	0.0514	0.6413	0.8766
Scenario 2	0.0509	0.6297	0.8786
Scenario 3	0.0489	0.6255	0.8651

Table 4.10: Type I Error and Power of the Test

Similar to simulation 1, bootstrap estimates are likewise calculated to depict the characteristics of the ATE. As shown in Table 4.11, lower standard errors are realized for the proposed method. However, the results also suggest that when the treatment assignment is based only on x_1 and x_3, the "swapping" method fails to capture the true ATE under the specified model where covaariates have mixed effects on the response.

Scenario	**Bootstrap Estimate**	**Standard Error**	**Confidence interval**
Scenario 1	0.0210	0.0013	(0.0185, 0.0233)
Scenario 2	-0.0038	0.0026	(-0.0089, 0.0012)
Scenario 3	-0.0026	0.0026	(-0.0077, 0.0023)

Table 4.11: Bootstrap Estimates of ATE under Simulation 2

The accuracy of the proposed method under this simulation setup is likewise investigated by calculating the coverage probabilities of the different scenarios. Some level of inaccuracy was also detected given the low coverage probabilities calculated, as shown in Table 4.12. This inaccuracy could potentially be attributed to the minimal treatment units possible drawn in the re-samples which may affect the estimation of the ATE.

Scenario	**Coverage Probability**
Scenario 1	84.2
Scenario 2	85.2
Scenario 3	85.8

Table 4.12: Coverage Probability of Swapping Method under Simulation 2

In general, the simulations conducted show that the proposed method is able to capture the true average treatment effect and provide estimates that are close to the true value and have small standard errors. This is particularly true for cases when the treatment assignment depends on the subset of covariates x_2 and x_4 or when it depends on all of the available confounders on the data set. Hence, scenarios have been detected where the proposed method performs well.

4.4 IPTW vs. Proposed Method

In the simulations discussed in Sections 4.2 and 4.3, the inverse probability treatment weighting method of the propensity score analysis is generally shown to be comparable to that estimates of the "swapping" method. To compare the performance of the proposed method and IPTW, a simulation scenario where the latter is proven to be successful in estimating the average treatment effect is applied on the proposed method.

A small simulation study has been conducted by Williamson, et al. [36], to illustrate the statistical properties of IPTW and its estimates provided in the section 1.4. In this simulation, treatment assignment is generated with $D \sim Bernoulli(0.5)$ and four independent baseline confounders X_1, X_2, X_3 and X_4 are considered, each coming from a $Normal(0, 9)$. The continuous response variable Y is simulated from $Normal(\mu, 0.25)$ where $\mu = 0.8X_1 + 0.5X_2 + 0.15X_3 + 2D$. Under these specifications, the marginal correlations within the treatment and control groups between the response variable and each of the four confounders X_1, X_2, X_3 and X_4 are 0.8,

0.5, 0.15 and 0, respectively. Furthermore, the true treatment effect as measured by the mean difference, is 2. The data generation procedure suggests drawing 100, 200 and 1,000 sample sizes, resulting to approximately 50, 100 and 500 units for the treatment and control groups. For each sample size, 5,000 data sets were simulated. With this scenario, it has been established that there is no large bias in the IPTW estimate for all the sample sizes considered. To investigate the performance of the proposed method in this scenario where the IPTW provides good estimate, the "swapping" method is implemented on a simulated data with the same nature.

Table 4.13 reflects the results of the IPTW estimators [36] and the proposed model estimates implemented on the simulated data described above. It can be concluded that the proposed method provided comparable estimates and that it is able to manifest the true average treatment effect for all sample sizes considered. The estimated variance, calculated by the mean of the variance estimates, for the IPTW is higher than that of the proposed method which suggests superiority of the estimates generated from the "swapping" method. In all cases, the empirical variance, defined by the empirical variance across all 5,000 simulations, for the proposed model are very close to the estimated variance which reflects that more stable estimates of the average treatment effect are derived from the proposed method.

Sample	**Estimate**		**Est Var**		**Emp Var**		**Type I error**	
	IPTW	Proposed	IPTW	Proposed	IPTW	Proposed	IPTW	Proposed
100	2.01	1.998	0.342	0.010	0.327	0.011	0.042	0.056
200	2.00	1.999	0.170	0.005	0.170	0.005	0.046	0.047
1000	2.00	1.999	0.034	0.001	0.034	0.001	0.052	0.043

Table 4.13: Comparison of IPTW and Proposed Method (True Value = 2.0)

As a summary, the proposed method performs better than the IPTW estimates as shown by its comparable estimate and lower variances. For the proposed model, the type I error is likewise not significantly deflated and is approximately equal to $\alpha = 0.05$. It can be noticed that the type I error lowers as the sample size increases, which is as expected.

In terms of the power, it can be shown that the IPTW and "swapping" method have comparable power except when the sample size is small (n=100) for $\tau = 0.25$. In this scenario, it is shown that the IPTW is able to reject the null hypothesis more often and depict the true nature of the average treatment effect. Thus, IPTW is a preferable analysis in determining the average treatment effect. However, when $\tau = 0.5$, comparable power between the two procedures are detected regardless of sample size. Results are shown in Table 4.14.

Sample	$\tau = 0.25$		$\tau = 0.50$	
	IPTW	Proposed	IPTW	Proposed
100	0.8374	0.5954	0.8074	0.8174
200	0.7622	0.8654	0.9996	0.9916
1000	0.9999	0.9999	0.9999	0.9999

Table 4.14: Comparison of Power of IPTW and Proposed Method

Also, in a study by Austin, stratification on propensity score result in great bias when estimating average treatment effect [4]. It also tends to perform well when the covariate distributions have common support and/or the covariate distributions have substantial overlap between the treatment and the control group. On matching, it has been shown that it does not perform well for simulations where the covariate distribution for the treatment group is not contained within that of the control group [8]. Thus, the proposed method is deemed to be the most superior estimation procedure in comparison to the common propensity score methods.

4.5 Data Analysis

4.5.1 National Supported Work Data

To illustrate the proposed procedure, we apply the proposed matching method and estimation procedure of the average treatment effect to Lalonde data set on the comparison of treatment and control groups to determine causal effects of a job training program, described on Section 1.5. Revisiting the data set, the outcome of

interest is the real earnings in 1978 with the participation in the NSW job training program as the treatment. The covariates considered are age, education, black, hispanic, married and no degree. The total number of observations is $n = 445$; $n_C = 260$ are under the control group and $n_T = 185$ observations are under the treatment group.

At the initial stage, subgroups are created based on the four categorical variables mentioned above. With each variable having two levels, $G = 2^4 = 16$ total subgroups that are balanced with respect to the categorical variables are constructed. The frequencies of the groups are shown in Table 4.15.

Group	Characteristic	Control	Treat	Total
1	non-black, non-hispanic, single, with degree	6	7	13
2	black, non-hispanic, single, with degree	26	37	63
3	non-black, hispanic, single, with degree	1	1	2
4	black, hispanic, single, with degree	0	0	0
5	non-black, non-hispanic, married, with degree	0	2	2
6	black, non-hispanic, married, with degree	9	6	15
7	non-black, hispanic, married, with degree	1	1	2
8	black, hispanic, married, with degree	0	0	0
9	non-black, non-hispanic, single, no degree	10	8	18
10	black, non-hispanic, single, no degree	154	90	244
11	non-black, hispanic, single, no degree	23	7	30
12	black, hispanic, single, no degree	0	0	0
13	non-black, non-hispanic, married, no degree	1	1	2
14	black, non-hispanic, married, no degree	26	23	49
15	non-black, hispanic, married, no degree	3	2	5
16	black, hispanic, married, no degree	0	0	0

Table 4.15: Frequency of Subgroups of NSW Data

Since Groups 3, 4, 5, 7, 8, 12, 13, 15 and 16 are sparse groups, 13 observations from these groups are discarded for further analysis. This results to a total of $n^* = 432$ experimental units with $n_C^* = 254$ control units and $n_T^* = 178$ treatment units. For the 7 remaining subgroups, we test for equality of mean vector with unknown Σ to verify the balance of the groups with respect to the continuous variables. The results of the multivariate testing based on Yao's multivariate test are reflected in Table 4.16.

Group	p-value	Conclusion
1	0.9452	Do not reject
2	0.8387	Do not reject
6	0.7156	Do not reject
9	0.5501	Do not reject
10	0.9572	Do not reject
11	0.8878	Do not reject
14	0.6634	Do not reject

Table 4.16: Multivariate Test on NSW Data

The results above reflect that there is balance in the subgroups with respect to the continuous variables. This implies that there is no systematic difference in the covariates between the treatment and the control group in all the subgroups formed. Hence, it mimics a randomized setup and thus, direct estimation can be made in each group as in Case 4, for the average treatment effect estimation. Table 4.17 below reflects the estimates of the average treatment difference in each subgroup, where the direct difference above is applied.

Group	**Frequency**		**Mean (SD)**		**Mean**	**Standard**
	Control	Treatment	Control	Treatment	**Difference**	**Error**
1	6	7	9960.1 (6311.6)	7337.2 (3926.3)	-2622.9	2864.9
2	26	37	3821.7 (5587.5)	7756.1 (8567.0)	3934.5	1916.9
6	9	6	3978.3 (4398.8)	12183.9 (13543.7)	8205.6	4785.9
9	10	8	5866.2 (4257.8)	8870.8 (6850.6)	3004.6	2629.5
10	154	90	4189.6 (5519.6)	4833.5 (7645.0)	643.9	847.1
11	23	7	6615.7 (5651.8)	8302.7 (8711.6)	1687.0	2776.1
14	26	23	3952.7 (5485.8)	7050.8 (6840.6)	3098.1	1762.5

Table 4.17: Subgroup Estimates of NSW Data

Prior to using the weighted combination formula for $\tilde{\Delta}$, the data is tested for homogeneity. With the quantities given in Table 4.17, the test statistic of χ^2 is calculated as

$$\chi^2 = \sum_{i=1}^{G^*} \frac{(\hat{\mathbf{\Delta}}_i - \tilde{\Delta})^2}{\hat{\sigma}_i^{*2}} = \sum_{i=1}^{G^*} \frac{\hat{\mathbf{\Delta}}_i^2}{\hat{\sigma}_i^{*2}} - \frac{\left(\sum_{i=1}^{G^*} \hat{\mathbf{\Delta}}_i/\hat{\sigma}_i^{*2}\right)^2}{\sum_{i=1}^{G^*} 1/\hat{\sigma}_i^{*2}} = 13.33 - \frac{0.0037^2}{0.00000242} = 7.8225 \tag{4.4}$$

The rejection region states that the null hypothesis is rejected if $\chi^2 > \chi^2_{0.05,6}$. With the test statistic value of 7.8225 and a critical value of $\chi^2_{0.05,6} = 12.529$, the null is accepted and it can be concluded that the assumption of homogeneity is satisfied. Thus, the weighted combination formula provides a suitable estimate of the average treatment effect. The average treatment effect for this example is given by

$$\tilde{\Delta} = 1506.7$$

with a corresponding standard error of $\hat{\sigma}(\tilde{\Delta}) = 641.8281$.

To investigate the performance of this method, the estimated average treatment effect is compared with those derived from propensity scores methods. In stratification, the estimated treatment effect is computed by summing the within-stratum difference in means between the treatment and control groups. For the matching, each control group is matched with a treatment group and the estimate is derived by taking the mean of these paired differences. Also, a randomized-experiment benchmark estimate has been presented [7] to facilitate comparison of the generated estimates. The results are shown in Table 4.18.

Propensity Scores		**Proposed Method**	**Randomization**
Stratification	Matching		
\$1,608 (1,571)	\$1,691 (2,209)	\$1,506 (642)	\$1,794 (633)

Table 4.18: NSW Data Analysis based on Propensity Score and Proposed Methods

Based on the results above, it can be observed that both propensity score estimates are closer to the experimental benchmark as compared to the proposed method. However, it was shown that the standard errors of estimates from stratification and matching are substantially higher than that of the "swapping" estimate.

4.5.2 *Lilly* Clinical Trial Data

We also applied the proposed matching method to estimate the ATE of a randomized, open-label clinical trial performed by *Lilly*. The primary objective of the study was to demonstrate that *dulaglutide* (treatment A) was noninferior to *liraglutide* (treatment B) in changing the control in glycosylated hemoglobin A1C in patients with type 2 diabetes. The outcome of interest is the hemoglobin levels of patients and the treatments considered are *dulaglutide* and *liraglutide*. The covariates explored are age, BMI, ethnicity (Hispanic/Latino or otherwise), region of residence (North America, South America and Europe), and gender (Male and Female). The total number of observations is $n = 551$, with $n_A = 275$ subjects treated with *dulaglutide* and $n_B = 276$ subjects with *luraglutide*. Subgroups are created based on the three categorical variables, with two levels for gender and

ethnicity and three levels for region, resulting to $G = 12$ subgroups. The frequencies of the groups are shown in Table 4.19.

Subgroup	**Characteristic**	***Dulaglutide***	***Liraglutide***	**Total**
1	North America , Female , Hispanic	22	21	43
2	South America , Female , Hispanic	15	16	31
3	Europe , Female , Hispanic	2	1	3
4	North America , Male , Hispanic	20	19	39
5	South America , Male , Hispanic	5	7	12
6	Europe , Male , Hispanic	4	2	6
7	North America , Female , Not Hispanic	29	24	53
8	South America , Female , Not Hispanic	0	0	0
9	Europe , Female , Not Hispanic	82	77	159
10	North America , Male , Not Hispanic	17	23	40
11	South America , Male , Not Hispanic	0	0	0
12	Europe , Male , Not Hispanic	79	86	165

Table 4.19: Frequency of Categories of *Lilly* Data

Since subgroups 3, 6, 8, and 11 are sparse, 9 observations from these groups are discarded for further analysis. This results to a total of $n^* = 542$ subjects with $n_A^* = 269$ units for treatment A and $n_B^* = 273$ for treatment B. For the 8 remaining subgroups, the test for equality of mean vectors with unknown Σ is implemented to verify the balance of the treatment and control groups with respect to the continuous variables. The results of the multivariate testing based on Yao's multivariate test statistic suggest that there is balance between groups in the subgroups with respect to the continuous variables. This implies that there is no systematic difference in the covariates between the groups in all the subgroups. Therefore, the data may be

likened to a completely randomized setup; thus, direct estimation can be made in each subgroup to generate treatment effect. Table 4.20 reflects the estimates of the treatment effect in each subgroup as calculated by the difference in mean response values of the two treatment groups.

Subgroup	n_A	n_B	**Mean Difference**	**Standard Error**
1	22	21	-0.1915	0.3399
2	15	16	0.0117	0.2931
4	20	19	-0.1718	0.3746
5	5	7	-0.5200	0.6042
7	29	24	0.1818	0.2116
9	82	77	-0.1482	0.1165
10	17	23	0.0201	0.2327
12	79	86	-0.1673	0.1194

Table 4.20: Subgroup Estimates of *Lilly* Data

The test of homogeneity is then performed on the treatment effects of all subgroups prior to applying meta-analysis for the estimation of the ATE. Using the quantities in Table 4.20, a test statistic of

$$\chi^2 = 3.2626$$

is calculated. The rejection region states that the null hypothesis is rejected if $\chi^2 > \chi^2_{0.05,7}$. Therefore, given the critical region of $\chi^2_{0.05,7} = 14.0671$, the null hypothesis is not rejected. Because the assumption of homogeneity is satisfied, the weighted combination calculation is used. The ATE estimator is given by

$$\tilde{\Delta} = -0.1043$$

with the standard error of $\hat{\sigma}(\tilde{\Delta}) = 0.0682$. This estimate is shown to be comparable with the classical test on difference between two population means, where the ATE is calculated as

$$\tilde{\Delta} = -0.1054$$

with the standard error $\hat{\sigma}(\tilde{\Delta}) = 0.0737$. Although, the estimates are comparable, the swapping method provides an advantage of eliminating probable bias induced by covariates.

The propensity scores methods were also applied to the *Lilly* data for comparison. The results are shown in Table 4.21. The results suggest that the proposed method generates comparable results to the z-test and other propensity score methods except stratification, where larger difference and standard errors are calculated. Also, all confidence intervals include 0 which imply that non-inferiority of *dulaglutide* over *liraglutide*. However, it can also be observed that the "swapping" method has the smallest standard error which shows the advantage of using the proposed method.

Mean Difference	**Propensity Scores**			**Swapping Method**
	Stratification	Matching	IPTW	
-0.1054 (0.0737) (-0.2501, 0.0393)	-0.1599 (0.3601) (-0.8657, 0.5459)	-0.1037 (0.0739) (-0.2485, 0.0411)	-0.1093 (0.0699) (-0.2463, 0.0277)	-0.1043 (0.0682) (-0.2380, 0.0294)

Table 4.21: ATE, SE and CI based on Swapping and Propensity Score Methods of *Lilly* Data

We created a modified data set to mimic imbalance between groups for investigating the performance of the swapping method. To do this, 36% and 35% of the subjects under Treatment A in subgroups 9 and 12, respectively, were dropped. Only subjects with ages lower than the median age 56 were considered for drop-outs. Meanwhile, 50% of the subjects in subgroup 9 and 35% of the subjects in subgroup 12 under Treatment B with ages higher than 56 were dropped. Under this mechanism, imbalance is induced; thus, the swapping mechanism may be applied to the modified subgroups. The Yao's Test performed on subgroups 9 and 12 establishes the imbalance with p-values 0.0492 and 0.0478, respectively. Treatment effect estimates on subgroups 9 and 12 are based on the swapping method while the treatment effects for remaining subgroup differences are based on the direct difference of mean response values. The result of the swapping procedure is shown in Table 4.22.

Subgroup	n_A	n_B	**Difference**	**Error**
1	22	21	-0.1915	0.3399
2	15	16	0.0117	0.2931
4	20	19	-0.1718	0.3746
5	5	7	-0.5200	0.6042
7	29	24	0.1818	0.2116
9*	71	64	-0.1422	0.1413
10	17	23	0.0201	0.2327
12*	79	86	-0.2776	0.1456

* : Modified groups with drop-out rates to allow the implementation of the swapping method

Table 4.22: Frequency of Categories under Modified Data Set of *Lilly* Data

The homogeneity testing procedure suggests that a random effects model is

not necessary with

$$\chi^2 = 4.2816.$$

An average treatment effect estimate of

$$\tilde{\Delta} = -0.1181$$

with standard error $\hat{\sigma}(\tilde{\Delta}) = 0.0772$, is calculated for the modified data set. The result suggests that the proposed swapping procedure is able to produce comparable ATE estimates with acceptable standard errors to the full data set even with the presence of drop-out.

The estimated ATE is compared with those derived from propensity scores methods as well as the classical test of two population mean difference. The results are shown in Table 4.23. The results reflect that the proposed method is the most comparable to the estimate based on the full data set, with the relatively small standard error. The results support that the swapping method can correct the bias due to the potential non-randomization induced even by an open-label study design.

In general, the data analyses performed show that the use of the swapping method yields comparable results to standard propensity score methods and the natural estimator for the difference in two population means. However, it can be observed that smaller standard errors are realized when the swapping method is implemented. This may be attributed to the elimination of bias induced by systematic differences in covariates through the proposed method.

Mean Difference	**Propensity Scores**			**Swapping Method**
	Stratification	Matching	IPTW	
-0.1489 (0.0803) (-0.3066, 0.0088)	-0.0838 (0.3691) (-0.8072, 0.6396)	-0.0512 (0.0775) (-0.2031, 0.1007)	-0.0827 (0.0736) (-0.2270, 0.0616)	-0.1181 (0.0772) (-0.2694, 0.0332)

Table 4.23: ATE, SE and CI based on Swapping and Propensity Score Methods of Modified *Lilly* Data

It can also be observed that the "swapping" method produces similar ATE estimates with or without drop-outs, as evident by the small absolute difference between the "swapping" estimates reflected in Tables 4.21 and 4.23. In both data sets, the proposed method produces small standard errors and the establish the non-inferiority of one drug over the other.

4.6 Summary of Results

The simulations and data analyses on two real-life data sets suggest that the proposed "swapping" procedure is able to provide reasonable estimates under the following conditions:

(a) There is a common parameter β^* such that $\beta_T^* = \beta_C^* = \beta^*$. Intuitively, this suggests that the covariates affect the outcome in the same way; thus, the effects of the treatments are reflected only in the model coefficients.

(b) Ignorability can be reached through sequential balancing in the categorical and continuous covariates.

(c) Residual bias for potential outcome distribution can be reasonably modeled parametrically by the continuous variables; hence, along with assumption (b), the distribution of the treatment and the control are exchangeable.

(d) Sequential balancing in the categorical and continuous variables imply simultaneous balancing across all categorical and continuous variables.

(e) The treatment and the control have homogeneous effect sizes, β_T^* and β_C^*, so

that the two estimates can be combined according to the formula presented in Section 3. This is probable because the possible values of the covariates belong to the same space with some overlap. Hence, the two estimates may be combined accordingly.

Also, it is emphasized that in the simulations, although the proposed method is able to capture the true mean and provide smaller standard errors and absolute value difference from the true mean in comparison to propensity score methods, this does not imply that the latter is not able to handle the data at hand properly. The results are as such due to the low prevalence rate of the treatment units in the simulation setup and thus, one-to-one matching and stratification discard several observations in the data. As a consequence, almost distinct data sets are used to compare the proposed procedure and the propensity scores methods.

Chapter 5

Subject Profile Analysis

5.1 Overview

In reality, covariates could potentially affect the response to a degree where it is only effective on a set of subjects with certain covariate features. Hence, it would also be of interest to look into the behavior of the ATE given certain constraints on the covariates. In this section, exploratory analysis is performed to assess the treatment effect based on the subject covariate profile. This is performed via simultaneous inference of regression models where the ATE is perceived in the context of confidence bands. Such analysis will aid in determining the aspects of a subject's covariate profile for which the treatment illustrates positive or negative effects while ensuring perfect balance with respect to the categorical covariates. Balanced subgroups are formulated using the available information on categorical variables, similar to the swapping mechanism described in the previous chapters. Confidence bands, which reflect the magnitude of difference between treatment and control responses, are then calculated to describe the differences in outcomes between treatment and control groups with respect to the remaining continuous variables.

5.2 Simultaneous Inference of Regression Models

The average treatment effect can be assessed using regression models of the treatment and control groups which describe the relationship of the treatment outcome Y on a same set of predictors $x_1, x_2, \ldots, x_p$. In general, suppose the two linear regression models are

$$\mathbf{Y}_i = \mathbf{X_i}\boldsymbol{\beta_i} + \mathbf{e_i}, \quad i = T, C \tag{5.1}$$

where $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,n_i})'$ is a vector of treatment outcomes, $\mathbf{X}_i$ is an $n_i \times (p+1)$ design matrix with full column rank and the first column is a vector of $\mathbf{1}'$s while the l-th column is $(x_{i,1,l}, x_{i,2,l}, \ldots, x_{i,n_i,l})'$, $\boldsymbol{\beta}_i = (\beta_{i,0}, \beta_{i,1}, \ldots, \beta_{i,p})'$ is a vector of unknown regression coefficients and $\mathbf{e}_i = (e_{i,1}, e_{i,2}, \ldots, e_{i,p})$ is a vector of random errors where $e_{i,j} \sim N(0, \sigma^2), i = 1, 2, j = 1, 2, \ldots, n_i$. Under this setup, $\mathbf{Y}_T$ and $\mathbf{Y}_C$ are the observed outcomes of treatment and control groups, respectively, that depends on the same p covariates. While it is important to determine if the two models are different, the magnitude of dissimilarity of these regression lines is of more interest for the estimation of the average treatment effect.

To assess the magnitude of difference between the two models, a simultaneous confidence band

$$\boldsymbol{x'\beta_T - x'\beta_C} = (1, x_1, x_2, \ldots, x_p)\boldsymbol{\beta}_T - (1, x_1, x_2, \ldots, x_p)\boldsymbol{\beta}_C \tag{5.2}$$

to bound the difference between the two models over the whole covariate space is generated [17]. To estimate this band, suppose $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{T}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{C}}$ and $\boldsymbol{\beta} = \boldsymbol{\beta_T} - \boldsymbol{\beta_C}$. It can be verified that $\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \boldsymbol{\sigma^2\Delta})$ where $\boldsymbol{\Delta} = (\boldsymbol{X_T'X_T})^{-1} + (\boldsymbol{X_C'X_C})^{-1}$, $\dfrac{\hat{\boldsymbol{\sigma}}}{\boldsymbol{\sigma}} \sim \sqrt{\dfrac{\chi_v^2}{v}}$ with $v = n_T + n_C - 2(p+1)$ and $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\sigma}}$ are independent random variables.

Hoel [11] and Scheffe [31, 32] generalized an exact $1-\alpha$ simultaneous confidence band for $\boldsymbol{x'\beta_T - x'\beta_C}$ over the whole covariate space as

$$\boldsymbol{x'\beta_T - x'\beta_C} \in \boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C} \pm \sqrt{(p+1)f^{\alpha}_{p+1,v}}\boldsymbol{\hat{\sigma}}\sqrt{\boldsymbol{x'\Delta x}} \quad \forall \quad \mathbf{x}_{(0)} \in \mathbb{R}^p,$$

where $f^{\alpha}_{p+1,v}$ is the upper α point of an $F-$distribution with degrees of freedom $p+1, v$ as defined above and $\hat{\sigma}^2 = \dfrac{n_T - p - 1}{n_T + n_C - 2(p+1)}\hat{\sigma}^2_T + \dfrac{n_C - p - 1}{n_T + n_C - 2(p+1)}\hat{\sigma}^2_C$ with $\hat{\sigma}^2_T$ and $\hat{\sigma}^2_C$ being the respective mean square residuals of the treatment and control regression models. If the two models are the same, then the difference in treatment outcome is the zero hyperplane $\mathbf{x'0}$ which is included in the confidence band with probability $1-\alpha$. This specification results to a size α test of the hypotheses $H_0 : \boldsymbol{\beta}_T = \boldsymbol{\beta}_C$ against $H_a : \boldsymbol{\beta}_T \neq \boldsymbol{\beta}_C$ using the confidence where the null is rejected if and only if $\mathbf{x'0}$ is outside the band for at least one $\mathbf{x}_{(0)} \in \mathbb{R}^p$.

However, in many real-life applications, the values of the covariates do not span the entire $\mathbb{R}^p$ space. In such cases, estimating confidence bands as above is inefficient and inappropriate particularly if the models hold only over a restricted region. In fact, Stewart [30] discussed possible drawbacks of visualizing the bands over the entire $\mathbb{R}^p$. This led to the development of a more useful confidence band for the difference between two models over a restricted region of the covariates. A two-sided constant-width simultaneous confidence band for $\boldsymbol{x'\beta_T - x'\beta_C}$ over the covariate region $\chi_r = \{(x_1, x_2, \ldots, x_p) : a_i \leq x_i \leq b_i, i = 1, 2, \ldots, p\}$ has the form

$$\boldsymbol{x'\beta_T - x'\beta_C} \in \boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C} \pm c\boldsymbol{\hat{\sigma}}\sqrt{\boldsymbol{x'\Delta x}} \quad \forall \quad \mathbf{x}_{(0)} \in \chi_r, \tag{5.3}$$

where c is a critical constant chosen so that the simultaneous confidence level of the band is $1-\alpha$ [17]. The confidence level of this simultaneous confidence band is

given by $P(S < c)$, where

$$\begin{aligned} S &= \sup_{\mathbf{x}_{(0)} \in \chi_r} \frac{|\boldsymbol{x'}(\hat{\boldsymbol{\beta}}_{\boldsymbol{T}} - \boldsymbol{\beta}_{\boldsymbol{T}} - \hat{\boldsymbol{\beta}}_{\mathbf{C}} + \boldsymbol{\beta}_{\boldsymbol{C}})|}{(\hat{\sigma})\sqrt{x'\Delta x}} \\ &= \sup_{\mathbf{x}_{(0)} \in \chi_r} \frac{|(\boldsymbol{Px})'(\boldsymbol{P^{-1}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{T}} - \boldsymbol{\beta}_{\boldsymbol{T}} - \boldsymbol{\beta}_{\boldsymbol{C}} + \hat{\boldsymbol{\beta}}_{\boldsymbol{C}})/\hat{\boldsymbol{\sigma}})|}{\sqrt{(\boldsymbol{Px})'(\boldsymbol{Px})}} \\ &= \sup_{\mathbf{x}_{(0)} \in \chi_r} \frac{||(\boldsymbol{Px})'\boldsymbol{T}|}{||\boldsymbol{Px}||} \\ &= \sup_{\mathbf{v} \in \mathbf{C}(\mathbf{P}, \chi_\mathbf{r})} \frac{|\boldsymbol{v'T}|}{||\boldsymbol{v}||} \end{aligned} \tag{5.4}$$

with $\mathbf{P}$ being the square matrix of $\boldsymbol{\Delta}, \boldsymbol{T} \sim T_{p+1,v}$ and $\mathbf{C}(\mathbf{P}, \chi_\mathbf{r}) = \{\lambda \mathbf{P}\mathbf{x} : \lambda \geq 0, \mathbf{x}_{(\mathbf{0})} \in \chi_\mathbf{r}\} = \{\lambda(\mathbf{p}_0 + x_1\mathbf{p}_1 + \ldots + x_p\mathbf{p}_p), x_i \in [a_i, b_i]\}$ for $i = 1, \ldots, p$. $\mathbf{C}(\mathbf{P}, \chi_\mathbf{r})$ can be viewed as the cone spanned by these vectors. It is apparent that if $p \geq 1$, then the derivation of the distribution of S and essentially the critical constant c becomes non-trivial. Simulation-based methods have been presented to calculate c and has shown to be as close to the exact values as possible under sufficiently large number of replications [18].

It is noted that the distribution of the pivotal quantity S is independent of the unknown parameters σ and β but is dependent on the bounds $[a_i, b_i]$ as well as the design matrix $\mathbf{X}$. Lui [16] described the complicated relationship of these components in a general setting.

In this study, confidence bands based on a restricted covariate space is considered. These bands are used to determine the average treatment effect based on the available subject profile information. Simulation-based methods to estimate c were also applied. With these bands, the behavior of the ATE is presented based on the remaining continuous covariates.

5.3 Proposed Exploratory Analysis

Given the G^* non-sparse subgroups that were formed given the different levels of the categorical variables, regression models were fitted and simultaneous regression inference were performed. Confidence bands given a restricted region are constructed on subgroups with sufficiently large number of subjects. These bands will determine the magnitude of treatment effect difference between the treatment and control groups. If the zero hyperplane is contained in the band, then there is no significant difference in treatment effect between the two groups for a certain group of subjects with similar categorical covariate features. On one hand, if the zero hyperplane is not included in the band, it is indicative of a significant treatment effect between the two groups. The magnitude and direction of the difference will likewise be determined in these bands. Through this procedure, one is able to detect for which group of subjects will the treatment effect difference be significant (either in the positive or negative direction) or insignificant.

For continuous covariates with a narrow range for $[a_i, b_i]$, one may investigate the subject profile by looking into the behavior of each independent variable considered in the confidence band while holding the other variables fixed. This will aid in understanding the behavior of treatment difference with respect to a single independent, continuous variable under fixed values of the remaining predictors.

This exploratory mechanism is advantageous for several reasons. First, partial balance is achieved due to the creation of independent subgroups based on the categorical variables. This eliminates any bias induced by such variables since units

in each generated subgroups are made as similar as possible with respect to their categorical characteristics. Second, more extensive analysis of the subjects' covariate profiles is performed. This provides a comprehensive idea on the behavior of the covariates for which positive and negative ATE are realized. Consequently, medical practitioners are presented with a means of identifying patients for which the intended treatment will be effective or not. However, a limitation of this procedure is that it is only applicable for small number of covariates although this is typical of observational studies.

5.4 Illustration

To illustrate the proposed exploratory procedure, we revisit the data sets considered in Sections 4.5.1 (NSW Data) and 4.5.2 (*Lilly* Clinical Trial Data). We then look into the behavior of the average treatment effect on the different subgroups formed from the categorical variables of the given data set.

5.4.1 National Supported Work Data

Given the 4 categorical variables of this data set, 7 non-sparse groups were formed. Confidence bands, illustrating the magnitude in the earning difference between subjects who participated in the NSW job training program and those who did not, were generated using the continuous predictors age and number of years in school. The Lalonde dataset is best modeled with these main effects and a quadratic term on age [15]. This same relationship is used in modeling the real earnings of the

treatment and control groups for each of the 7 subgroups formed. The results of the regression modeling procedure are shown in Table 5.1. Information on the restricted covariate region of interest of the independent variables age (x_1) and number of years in school (x_2) is also presented in the same table.

Subgroup	**Model Estimates**	**Covariate Region**
1	$\hat{\mathbf{y}}_T = 91111 - 4477X_1 - 1365X_2 + 69.43X_1^2$ $\hat{\mathbf{y}}_C = -395310 + 40276X_1 - 10393X_2 - 757.48X_1^2$ $\hat{\beta} = (486420, -44753, 9028.2, 826.9)$	$x_1 : [20, 41]$ $x_2 : [12, 14]$
2	$\hat{\mathbf{y}}_T = -59637 + 3622X_1 + 942.2X_2 - 54.03X_1^2$ $\hat{\mathbf{y}}_C = -13265 + 1010X_1 + 278.2X_2 - 17.84X_1^2$ $\hat{\beta} = (-46372, 2612, 664.0, -36.2)$	$x_1 : [18, 46]$ $x_2 : [12, 16]$
6	$\hat{\mathbf{y}}_T = -339270 + 11101X_1 + 15105X_2 - 183.5X_1^2$ $\hat{\mathbf{y}}_C = -102010 + 6671X_1 + 139.4X_2 - 103.82X_1^2$ $\hat{\beta} = (-237260, 4429.7, 14966, -79.7)$	$x_1 : [23, 42]$ $x_2 : [12, 14]$
9	$\hat{\mathbf{y}}_T = -25544 + 1287X_1 + 1749.5X_2 - 22.38X_1^2$ $\hat{\mathbf{y}}_C = -4341 + 2097X_1 - 1683.9X_2 - 39.29X_1^2$ $\hat{\beta} = (-21203, -810.1, 3434.4, 16.92)$	$x_1 : [17, 38]$ $x_2 : [7, 11]$
10	$\hat{\mathbf{y}}_T = -2308 + 322.4X_1 + 269.8X_2 - 5.06X_1^2$ $\hat{\mathbf{y}}_C = 3499 - 167.9X_1 + 244.7X_2 + 3.67X_1^2$ $\hat{\beta} = (-5807, 490.3, 25.01, -8.74)$	$x_1 : [17, 55]$ $x_2 : [3, 11]$
11	$\hat{\mathbf{y}}_T = 141450 - 10180X_1 - 3471.8X_2 + 238.7X_1^2$ $\hat{\mathbf{y}}_C = -22538 + 1524X_1 + 1148.7X_2 - 27.01X_1^2$ $\hat{\beta} = (163990, -11704, -4620.5, 265.7)$	$x_1 : [17, 50]$ $x_2 : [4, 11]$
14	$\hat{\mathbf{y}}_T = -4526 + 952X_1 - 210.7X_2 - 15.7X_1^2$ $\hat{\mathbf{y}}_C = 61464 - 3860X_1 - 113.7X_2 + 63.6X_1^2$ $\hat{\beta} = (-65991, 4812, -97.01, -79.4)$	$x_1 : [19, 46]$ $x_2 : [4, 11]$

Table 5.1: Regression Estimates of NSW Data

Upon generation of the regression model estimates for both groups, the critical constant c was simulated using the proposed procedure of Lui, et al [18]. It is noted that the values of $\hat{\beta}, \sigma^2$ and $\mathbf{\Delta}$ are based on the design-matrix. These components

were calculated using MATLAB. The same program was used in generating the graphs of the confidence bands for each subgroup.

Based on the figures of the simultaneous confidence bands, shown in Figure 5.1 to Figure 5.7, it can be observed that the zero hyperplane is included in the band which implies that there is no significant treatment difference between the two groups for all subgroups except subgroup 1 where the treatment difference between the two groups is shown to be negative in some portion of the surface of x_1 and x_2.

Table 5.2 shows the p-value of the simultaneous inference in regression, which support the graphs of the confidence band as evident by the p-values greater than $\alpha = 0.05$. Based on the p-value, it can be deduced that for subgroup 1, the zero hyperplane $\mathbf{x}'\mathbf{0}$ is outside the band for at least one $\mathbf{x}_{(\mathbf{0})}$ in χ_r. Graphically, it can be observed that this conclusion occurs on lower values of age and number of years in school. This shows that those who have attended the NSW job training still tended to have lower earnings than those who did not avail of the training for the younger respondents who have not attended much years in school. For the remaining subgroups, the p-value suggests that $\mathbf{x}_{(\mathbf{0})}$ is contained in all possible $\mathbf{x_0}$ in the restricted region.

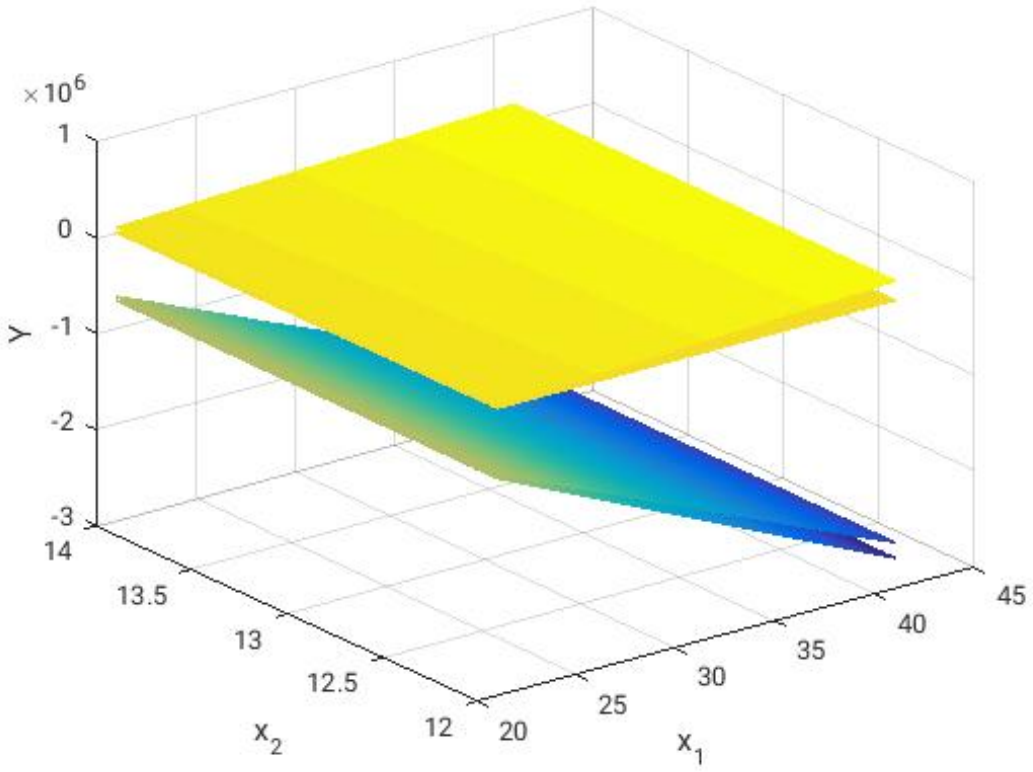

Figure 5.1: Simultaneous Confidence Band of Y, ATE, for Subgroup 1 on the restricted covariate region $x_1 : [20, 41]$ and $x_2 : [12, 14]$

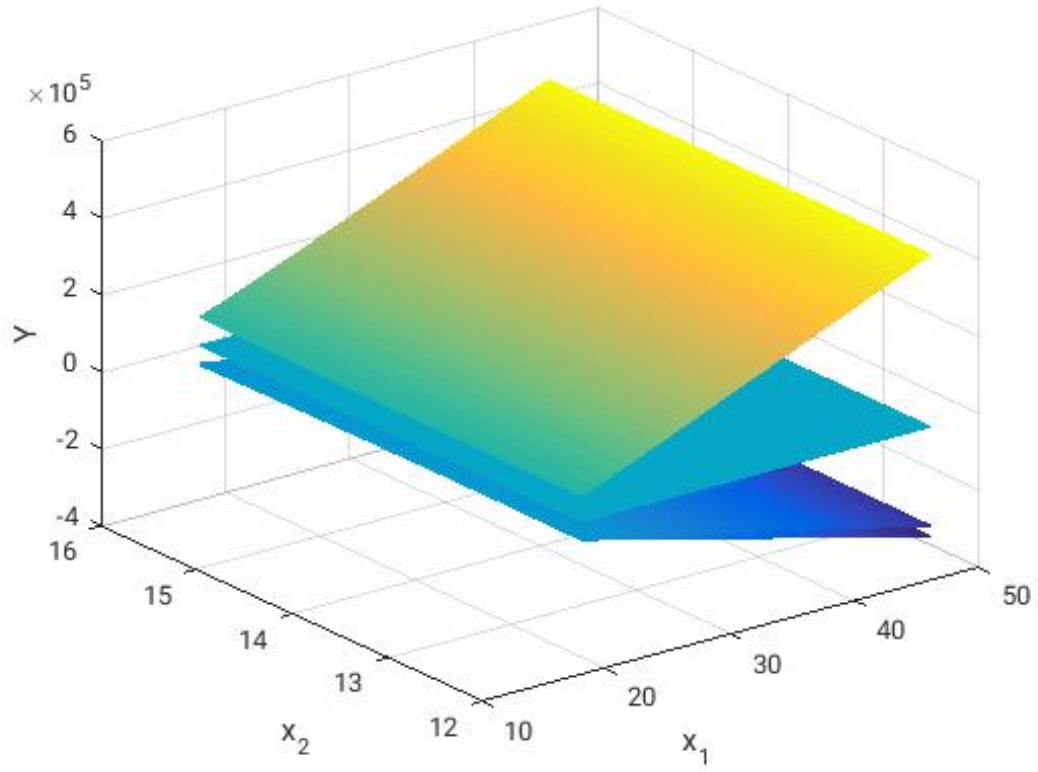

Figure 5.2: Simultaneous Confidence Band of Y, ATE, for Subgroup 2 on the restricted covariate region $x_1 : [18, 46]$ and $x_2 : [12, 16]$

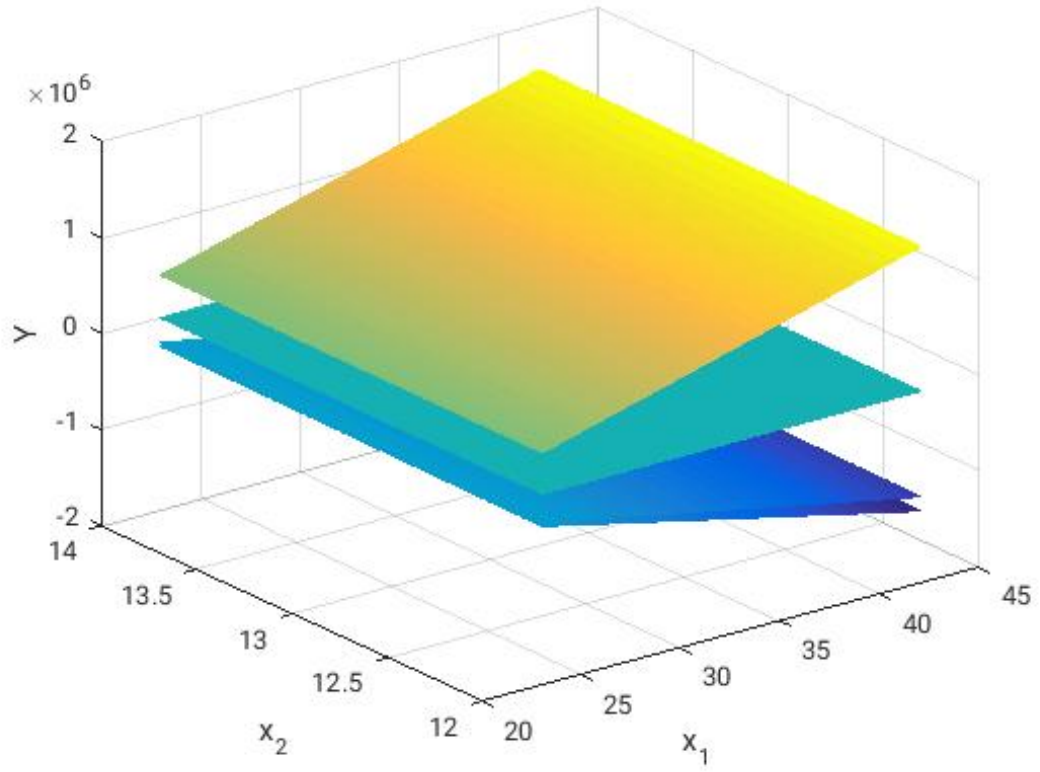

Figure 5.3: Simultaneous Confidence Band of Y, ATE, for Subgroup 6 on the restricted covariate region $x_1 : [23, 42]$ and $x_2 : [12, 14]$

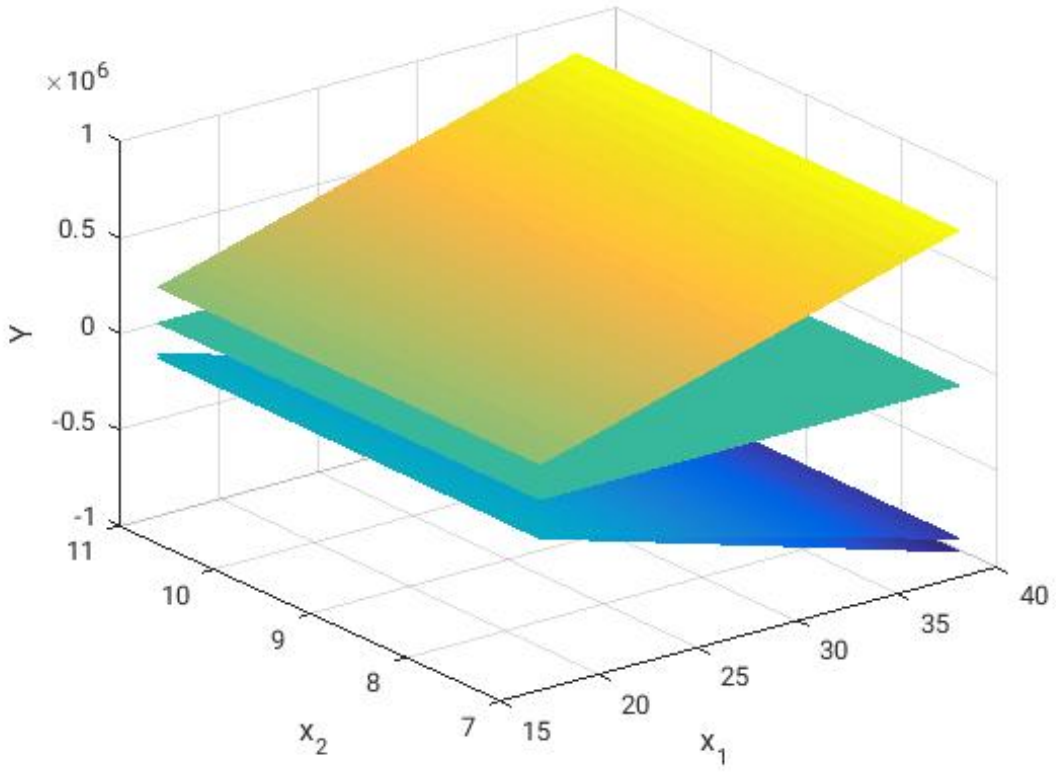

Figure 5.4: Simultaneous Confidence Band of Y, ATE, for Subgroup 9 on the restricted covariate region $x_1 : [17, 38]$ and $x_2 : [7, 11]$

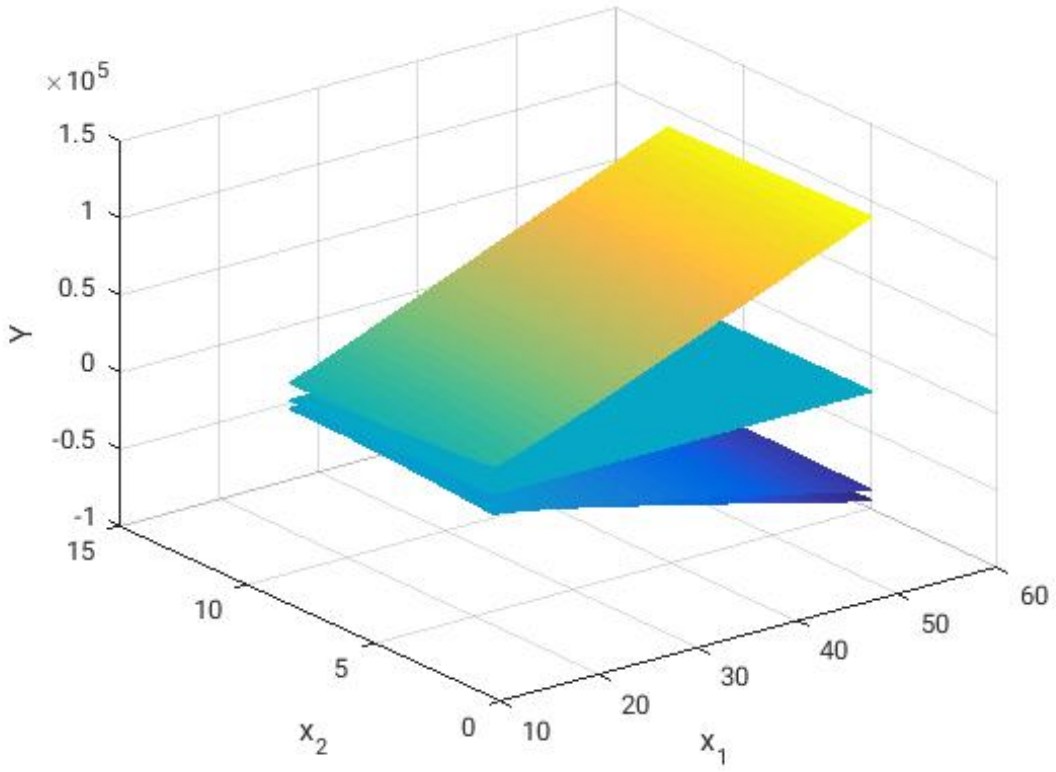

Figure 5.5: Simultaneous Confidence Band of Y, ATE, for Subgroup 10 on the restricted covariate region $x_1 : [17, 55]$ and $x_2 : [3, 11]$

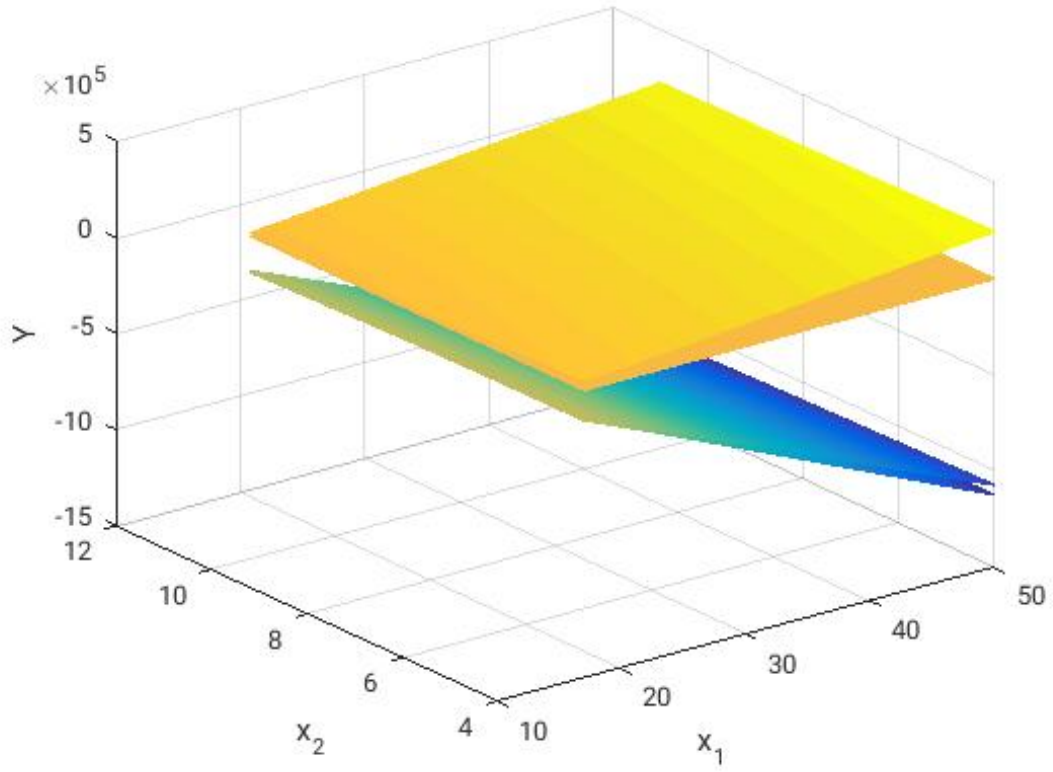

Figure 5.6: Simultaneous Confidence Band of Y, ATE, for Subgroup 11 on the restricted covariate region $x_1 : [17, 50]$ and $x_2 : [4, 11]$

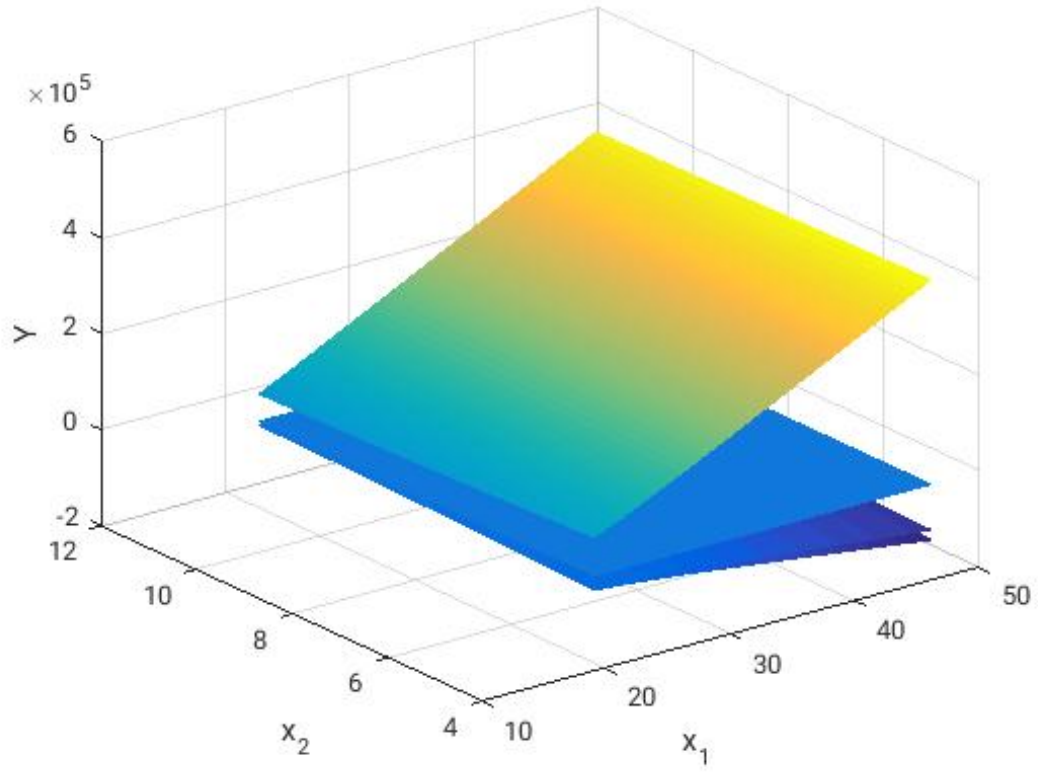

Figure 5.7: Simultaneous Confidence Band of Y, ATE, for Subgroup 14 on the restricted covariate region $x_1 : [19, 46]$ and $x_2 : [4, 11]$

Subgroup	p-value	Conclusion
1	0.0308	Reject Ho
2	0.1568	Do not reject Ho
6	0.1057	Do not reject Ho
9	0.6005	Do not reject Ho
10	0.8425	Do not reject Ho
11	0.0658	Do not reject Ho
14	0.1163	Do not reject Ho

Table 5.2: Simultaneous Inference p-values on Subgroups of NSW Data

This behavior is compared to the classical z-test for testing means of two independent populations. The mean differences of the subgroups are reflected in Table 4.17. The negative mean difference calculated for subgroup 1 is likewise observed on the confidence band. However, the confidence band suggests that the difference in regression model fit between the two groups are significantly different. This is not reflected on the z-test carried out for subgroup 1 (p-value = 0.3795). For the remaining subgroups, consistent results are observed as depicted by the non-rejection of the null hypothesis for the t-test and the inclusion of the zero hyperplane in the confidence bands.

Given that the bounds for the number of years of education for this data set is narrow, one may be able to explore the ATE by looking into the behavior of the confidence bands as a function of one variable in each of the subgroup. By finding the solutions of the confidence bands under a fixed covariate x_j, one is able to determine the value of $x_i, i \neq j$ for which the direction of the treatment effect changes. The

choice of the independent variable to be held fixed is based on the covariate region, where x_j with the narrowest region is chosen. In the NSW data set, an ATE function in terms x_1 is generated by holding the number of education (x_2) fixed because the x_2-region is narrower than the x_1- region. For a fixed covariate, the ideal scenario is for the confidence bands to illustrate a positive treatment effect over the entire covariate regions of interest. This implies that for the training will increase the salary of all subjects across all age and number of years of education. Although, this result may not necessarily hold true across all subgroups. In some instances, varying directions may occur and thus, the fixed value for which the function with a positive treatment effect is not clearly identified. While this occurrence may be true about this exploratory analysis, it will, on one hand, provide a good idea of the subject profile for which the treatment is effective or not.

Tables 5.3 - 5.9 reflect the forms of confidence bands for fixed values of x_2 while Figures 5.8 - 5.14 show the corresponding quadratic function of x_1 given a constant x_2. Specifically, the minimum, mean and maximum values of x_2 were considered as fixed points. The tables reflect the forms of each confidence band, including the simulated critical constant c and regression estimates.

For subgroup 1, it can be observed that given a maximum possible value of x_2, a positive treatment effect is guaranteed across all possible values of x_1, as shown in Table 5.3 and Figure 5.8. However, when the minimum and mean values are considered to be the fixed points, the ATE is positive or negative on some interval of x_1. Real solutions are also calculated, as reflected in the table, which implies that for some value of x_1 and under a fixed value of x_2, no treatment effect is observed.

This suggests that a respondent with a subject profile defined by subgroup 1, that is non-black, non-hispanic, single and has an educational degree with 14 years spent in school, will most likely have positive earnings given the job training regardless of their age. This relationship is more clearly illustrated in Figure 5.8.

x_2	**Confidence Band:** $\boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C \pm c\hat{\sigma}\sqrt{x'\Delta x}}$	**Solution on** χ_R
12	$826x_1^2 - 44753x_1 + 594760$ $\pm 4.37(2.78)\sqrt{\frac{33x_1^4-3000x_1^3+152400x_1^2-2760000x_1+18850000}{5000}}$	$x_{1+} = 23.39, 30.93$ $x_{1-} = 23.46, 30.68$
12.23	$826x_1^2 - 44753x_1 + 596836.44$ $\pm 4.37(2.78)\sqrt{\frac{33x_1^4-3000x_1^3+152446x_1^2-2762300x_1+18878129}{5000}}$	$x_{1+} = 23.82, 30.31$ $x_{1-} = 23.63, 30.59$
14	$826x_1^2 - 44753x_1 + 612816$ $\pm 4.37(2.78)\sqrt{\frac{33x_1^4-3000x_1^3+152800x_1^2-2780000x_1+19130000}{5000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$

Table 5.3: Confidence Bands for Subgroup 1 under fixed x_2

For subgroup 2, it can be observed that across all the possible range of x_2, the ATE can be non-negative or negative. This implies that there is no guarantee on which ranges of x_1 and x_2 is the treatment effective in increasing earnings, for black, non-hispanic, single and with degree. These results are reflected in the solutions of the confidence bands fixed at x_2, as shown in Table 5.4. However, figure 5.9 demonstrates that the behavior of the ATE on the restricted region of x_1 and fixed values of x_2 considered tends more towards the positive direction. Hence, it can be deduced that most of the black, non-hispanic, single respondents regardless of degree status will gain positive effects from the job training.

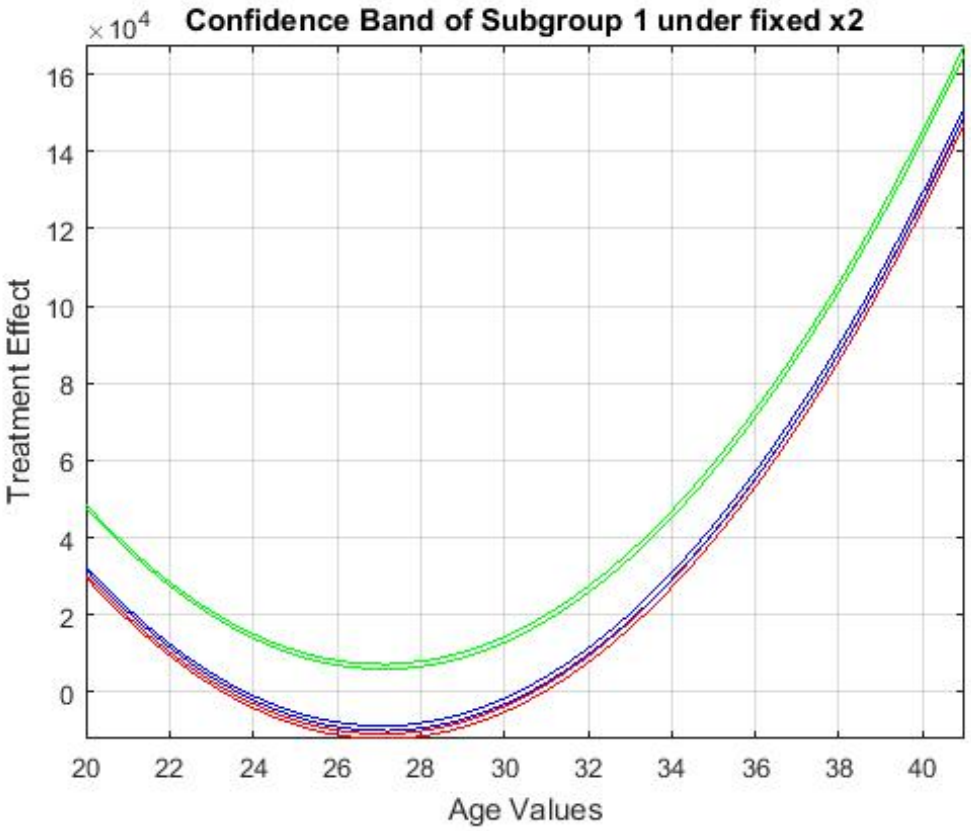

Figure 5.8: Simultaneous Confidence Band of Y, ATE, for Subgroup 1 on fixed minimum, mean and maximum values of x_2. The red, blue and green confidence bands indicate the behavior of ATE as a function of x_1 given a fixed value of x_2 at the minimum, mean and maximum, respectively.

x_2	**Confidence Band:** $\boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C \pm c\hat{\sigma}\sqrt{x'\Delta x}}$	**Solution on** χ_R
12	$-36x_1^2 + 2612x_1 - 38404$ $\pm(3.12)(2.58)\sqrt{\frac{39x_1^4-4000x_1^3+164600x_1^2-2980000x_1+19820000}{500000}}$	$x_{1+} = 20.48, 52.07$ $x_{1-} = 20.50, 52.04$
12.37	$-36x_1^2 + 2612x_1 - 38158.32$ $\pm(3.12)(2.58)\sqrt{\frac{78x_1^4-8000x_1^3+329792x_1^2-5989600x_1+40053549}{1000000}}$	$x_{1+} = 20.27, 52.34$ $x_{1-} = 20.28, 52.22$
16	$-36x_1^2 + 2612x_1 - 35748$ $\pm(3.12)(2.58)\sqrt{\frac{39x_1^4-4000x_1^3+167800x_1^2-3140000x_1+23580000}{500000}}$	$x_{1+} = 18.29, 54.28$ $x_{1-} = 18.31, 54.23$

Table 5.4: Confidence Bands for Subgroup 2 under fixed x_2

Meanwhile, for subgroup 6, there is a guaranteed behavior realized when x_2 is fixed at the minimum, according to the results shown in Table 5.5. Figure 5.10 displays this relationship as well. The graph suggests that for black, non-hispanic, married respondents with a degree, the ATE is negative over all possible domain of x_1 when x_2 is at the minimum. On one hand, the behavior of the ATE is always positive when x_2 is held fixed at the maximum for all possible values of x_1. This change in behavior may be an indicator of the effect of the number of years in school on the direction of the treatment effect for this group of respondents. It shows that positive earnings are achieved after the training for respondents who have stayed for 14 years in school while an opposite behavior is observed for those who have stayed for 12 years.

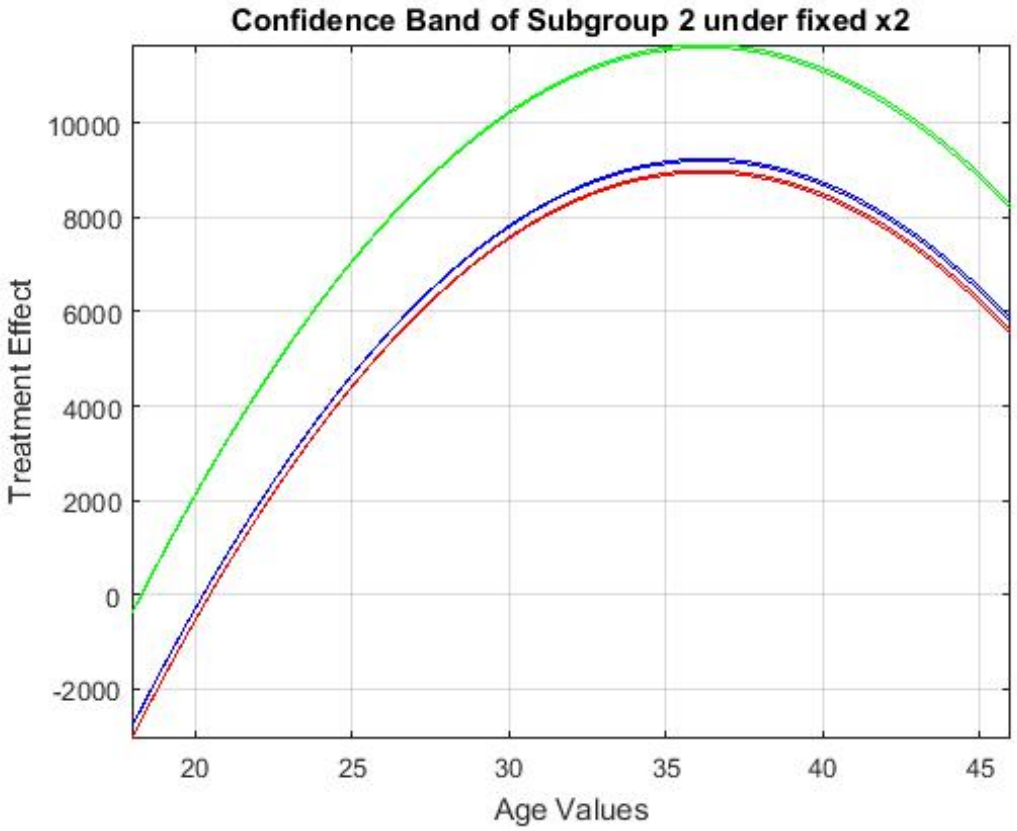

Figure 5.9: Simultaneous Confidence Band of Y, ATE, for Subgroup 2 on fixed minimum, mean and maximum values of x_2. The red, blue and green confidence bands indicate the behavior of ATE as a function of x_1 given a fixed value of x_2 at the minimum, mean and maximum, respectively.

x_2	**Confidence Band:** $\boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C \pm c\hat{\sigma}\sqrt{x'\Delta x}}$	**Solution on** χ_R
12	$-79x_1^2 + 4430x_1 - 57665$ $\pm(3.96)(3.29)\sqrt{\frac{3x_1^4-375x_1^3+18595x_1^2-3980000x_1+3132500}{3125}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
12.47	$-79x_1^2 + 4430x_1 - 50630.98$ $\pm(3.96)(3.29)\sqrt{\frac{960x_1^4-120000x_1^3+5959424x_1^2-127961600x_1+1012176329}{1000000}}$	$x_{1+} = 15.70, 40.42$ $x_{1-} = 16.04, 39.99$
14	$-79x_1^2 + 4430x_1 - 27733$ $\pm(3.96)(3.29)\sqrt{\frac{3x_1^4-375x_1^3+18715x_1^2-406000x_1+3270250}{3125}}$	$x_{1+} = 7.10, 48.97$ $x_{1-} = 7.27, 48.82$

Table 5.5: Confidence Bands for Subgroup 6 under fixed x_2

On the other hand, the confidence band of subgroup 9, which is composed of subjects who are non-black, non-hispanic, single and non-degree holder, a positive treatment effect is guaranteed at a fixed, maximum value of x_2. This is apparent in the solutions of its corresponding band reflected in Table 5.6. This behavior is also true for all values of x_1 given the fixed mean value of x_2. At the minimum, however, it can be established that the training has a negative effect when the number of years in school is fixed at 7 years. This means that subjects who did not spend much time in school and did not attend the training tend to have higher earnings; thus, the training is not beneficial for this group of subjects. This behavior is illustrated also in Figure 5.11.

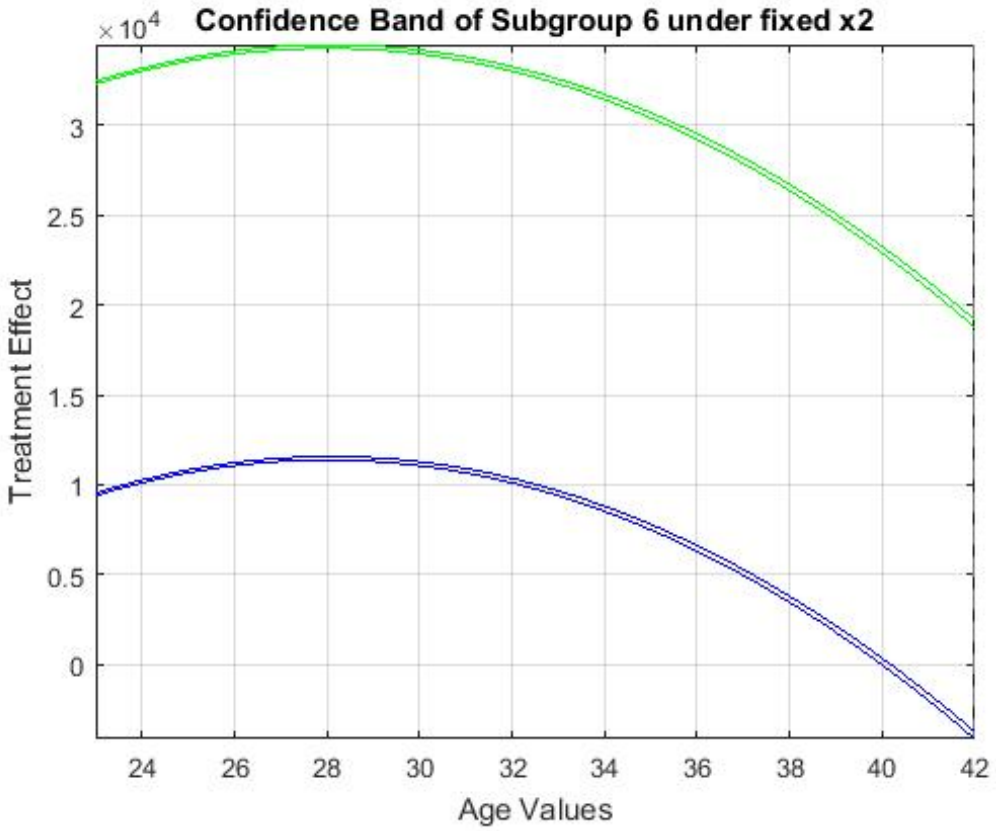

Figure 5.10: Simultaneous Confidence Band of Y, ATE, for Subgroup 6 on fixed minimum, mean and maximum values of x_2. The red, blue and green confidence bands indicate the behavior of ATE as a function of x_1 given a fixed value of x_2 at the minimum, mean and maximum, respectively.

x_2	**Confidence Band:** $\boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C \pm c\hat{\sigma}\sqrt{x'\Delta x}}$	**Solution on** χ_R
7	$17x_1^2 - 811x_1 + 2843$ $\pm(3.67)(1.67)\sqrt{\frac{9x_1^4-800x_1^3+28034x_1^2-418200x_1+2301800}{10000}}$	$x_{1+} = 3.98, 44.12$ $x_{1-} = 4.29, 43.85$
9.67	$17x_1^2 - 811x_1 + 12014.45$ $\pm(3.67)(1.67)\sqrt{\frac{450x_1^4-40000x_1^3+1409977x_1^2-21257100x_1+118143412}{500000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
11	$17x_1^2 - 811x_1 + 16583$ $\pm(3.67)(1.67)\sqrt{\frac{9x_1^4-800x_1^3+28282x_1^2-428600x_1+2401800}{10000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$

Table 5.6: Confidence Bands for Subgroup 9 under fixed x_2

For subgroup 10, it is shown that the solutions of the confidence bands for any of the fixed values of x_2 considered is around the lower extremes of x_1, as reflected in Table 5.7. This suggests that the change of behavior in the effectiveness of the treatment is reflected more among younger subjects. Figure 5.12 shows that the ATE is dominantly positive, at least for values of x_1 for which is it defined; hence, the treatment is effective. Based on the graph, we note that observations could be made only on some values of x_1 as the confidence bands under any of the fixed values of x_2 are not defined in the entire domain of x_1.

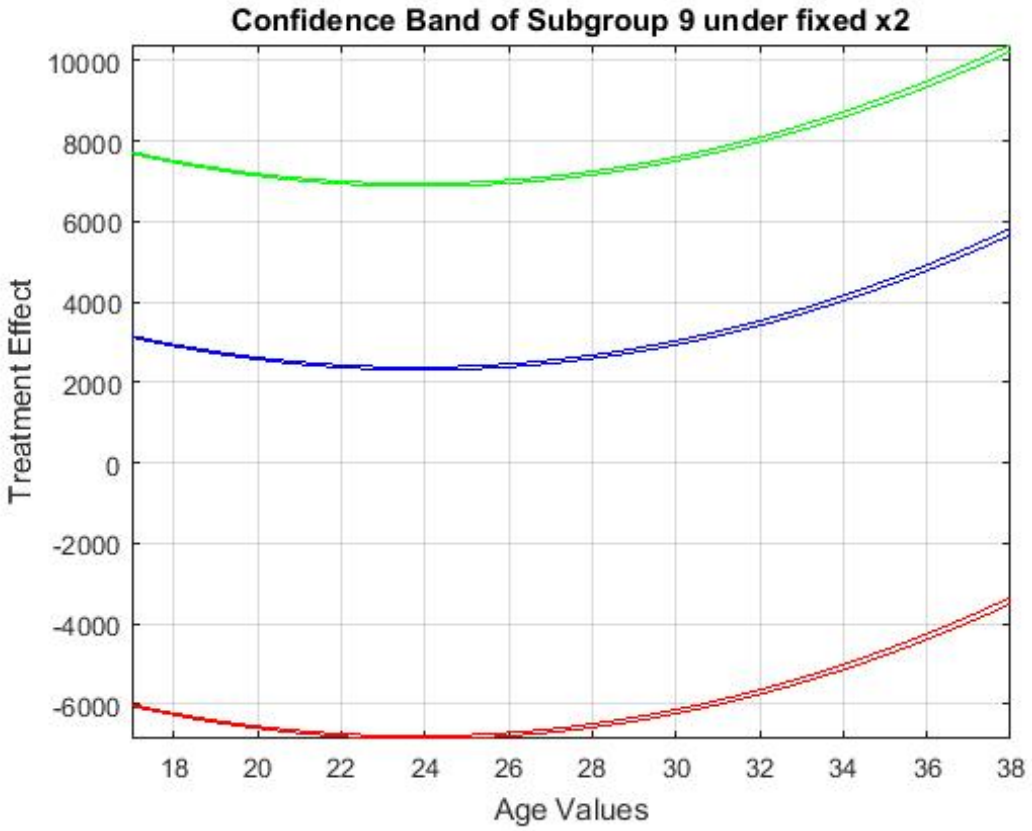

Figure 5.11: Simultaneous Confidence Band of Y, ATE, for Subgroup 9 on fixed minimum, mean and maximum values of x_2. The red, blue and green confidence bands indicate the behavior of ATE as a function of x_1 given a fixed value of x_2 at the minimum, mean and maximum, respectively.

x_2	**Confidence Band:** $\boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C \pm c\hat{\sigma}\sqrt{x'\Delta x}}$	**Solution on** χ_R
3	$-9x_1^2 + 490x_1 - 5733$ $\pm(3.06)(1.17)\sqrt{\frac{3x_1^4-400x_1^3+16980x_1^2-309900x_1+2430200}{1000000}}$	$x_{1+} = 17.01$ $x_{1-} = 17.03$
9.61	$-9x_1^2 + 490x_1 - 5567.75$ $\pm(3.06)(1.17)\sqrt{\frac{150x_1^4-20000x_1^3+868920x_1^2-16557400x_1+121042193}{50000000}}$	$x_{1+} = 16.15$ $x_{1-} = 16.16$
11	$-9x_1^2 + 490x_1 - 5533$ $\pm(3.06)(1.17)\sqrt{\frac{3x_1^4-400x_1^3+17460x_1^2-335500x_1+2503800}{1000000}}$	$x_{1+} = 15.987$ $x_{1-} = 15.993$

Table 5.7: Confidence Bands for Subgroup 10 under fixed x_2

Looking at subgroup 11, real solutions are calculated within the range of x_1 for the mean and maximum. This indicates that there is a change in direction of ATE for some values of x_1; hence, the effect of the training on subjects that are non-black, hispanic, single and no degree cannot be easily determined when they have spent 9.61 and 11 years in school. It is interesting to note, on one hand, that the training yields positive effect for subjects who share the same categorical profile but has spent only 3 years in school. These observations are shown in Table 5.8 and Figure 5.13.

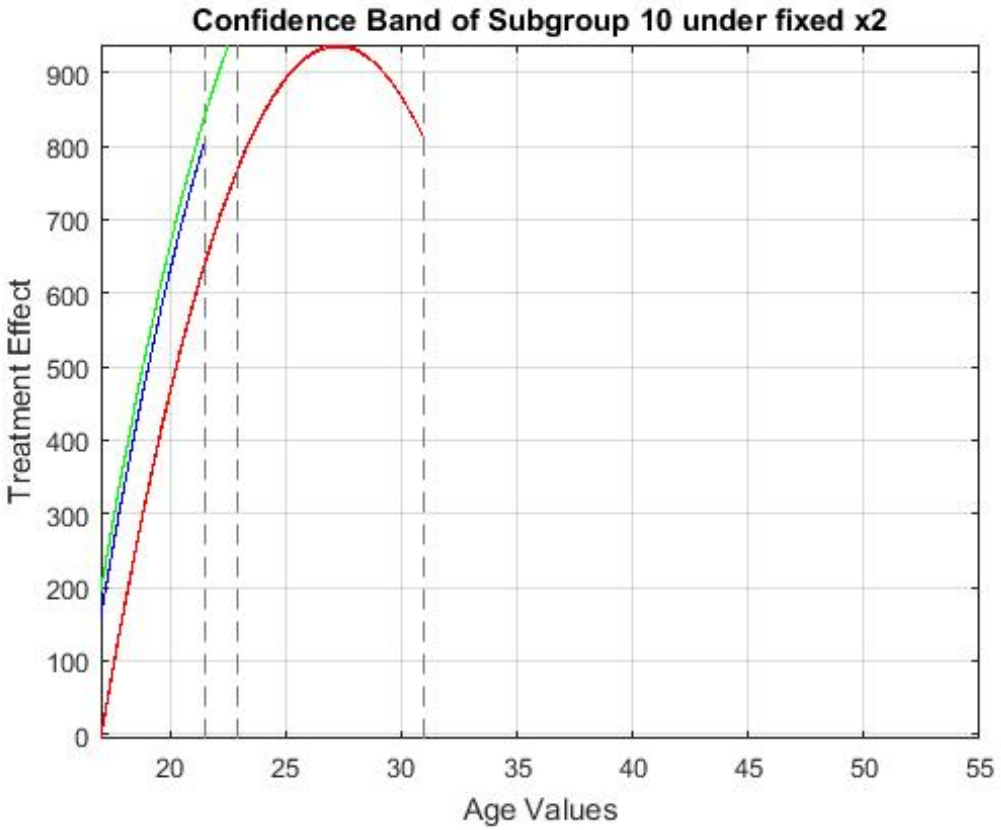

Figure 5.12: Simultaneous Confidence Band of Y, ATE, for Subgroup 10 on fixed minimum, mean and maximum values of x_2. The red, blue and green confidence bands indicate the behavior of ATE as a function of x_1 given a fixed value of x_2 at the minimum, mean and maximum, respectively.

x_2	**Confidence Band:** $\boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C \pm c\hat{\sigma}\sqrt{x'\Delta x}}$	**Solution on** χ_R
4	$266x_1^2 - 11704x_1 + 145505$ $\pm(3.34)(3.20)\sqrt{\frac{21x_1^4-2000x_1^3+74900x_1^2-1180000x_1+7960000}{50000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
8.90	$266x_1^2 - 11704x_1 + 122862.1$ $\pm(3.34)(3.20)\sqrt{\frac{21x_1^4-2000x_1^3+74165x_1^2-1555700x_1+7422225}{50000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
11	$266x_1^2 - 11704x_1 + 113158$ $\pm(3.34)(3.20)\sqrt{\frac{21x_1^4-2000x_1^3+73850x_1^2-1145000x_1+7522500}{50000}}$	$x_{1+} = 15.56, 29.05$ $x_{1-} = 15.52, 29.11$

Table 5.8: Confidence Bands for Subgroup 11 under fixed x_2

For the last subgroup, subgroup 14, the solutions of the confidence bands on a fixed x_2 at the minimum, mean and maximum suggest that there is no guaranteed ATE behavior detected as shown in Table 5.9 because the solutions fall in the range of x_1. Similar to other subgroups, this is indicative of a change in treatment effect within the domain; thus, no guaranteed observation on the behavior of ATE may be established. Although, Figure 5.14 suggests that majority of the ATE behavior realized is towards the negative direction. It shows that respondents who are black, hispanic, married and has no degree tend to have a negative treatment effect, at least for those with ages that are at the extremes and regardless of the number of years in school since this relationship holds true for all three fixed points of x_2 considered.

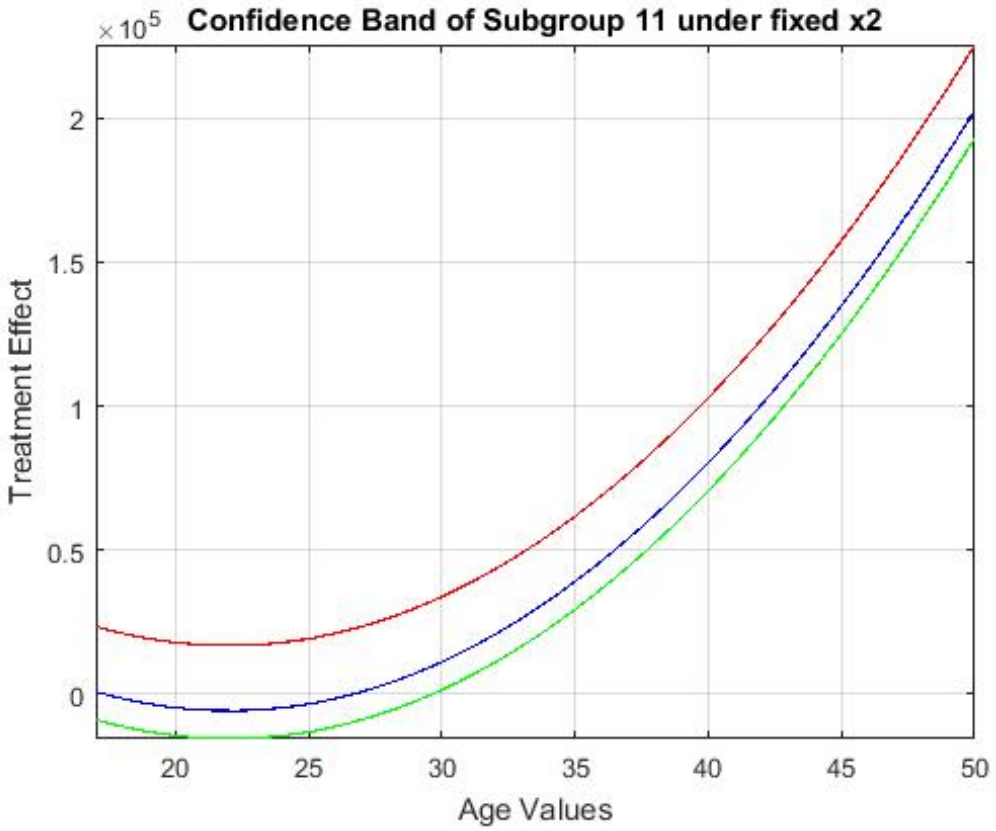

Figure 5.13: Simultaneous Confidence Band of Y, ATE, for Subgroup 11 on fixed minimum, mean and maximum values of x_2. The red, blue and green confidence bands indicate the behavior of ATE as a function of x_1 given a fixed value of x_2 at the minimum, mean and maximum, respectively.

x_2	**Confidence Band:** $\boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C \pm c\hat{\sigma}\sqrt{x'\Delta x}}$	**Solution on** χ_R
4	$-80x_1^2 + 4812x_1 - 66374$ $\pm(3.19)(2.78)\sqrt{\frac{x_1^4-140x_1^3+6859x_1^2-119000x_1+1025000}{25000}}$	$x_{1+} = 21.42, 38.74$ $x_{1-} = 21.44, 38.71$
9.90	$-80x_1^2 + 4812x_1 - 66940.4$ $\pm(3.19)(2.78)\sqrt{\frac{2x_1^4-280x_1^3+13995x_1^2-255700x_1+2180685}{50000}}$	$x_{1+} = 21.81, 38.36$ $x_{1-} = 21.88, 38.26$
11	$-80x_1^2 + 4812x_1 - 67046$ $\pm(3.19)(2.78)\sqrt{\frac{x_1^4-140x_1^3+7025x_1^2-129500x_1+1116000}{25000}}$	$x_{1+} = 21.92, 38.23$ $x_{1-} = 21.94, 38.21$

Table 5.9: Confidence Bands for Subgroup 14 under fixed x_2

Although there is no unifying behavior of the average treatment effect across the various subgroups on a specified covariate region, these analyses remain to be helpful in identifying the behavior of the treatment effect given certain information on the covariates. In this specific example, the expected treatment effect of a subject is identifiable based on information of their gender, marital status, ethnicity, age and number of years in school.

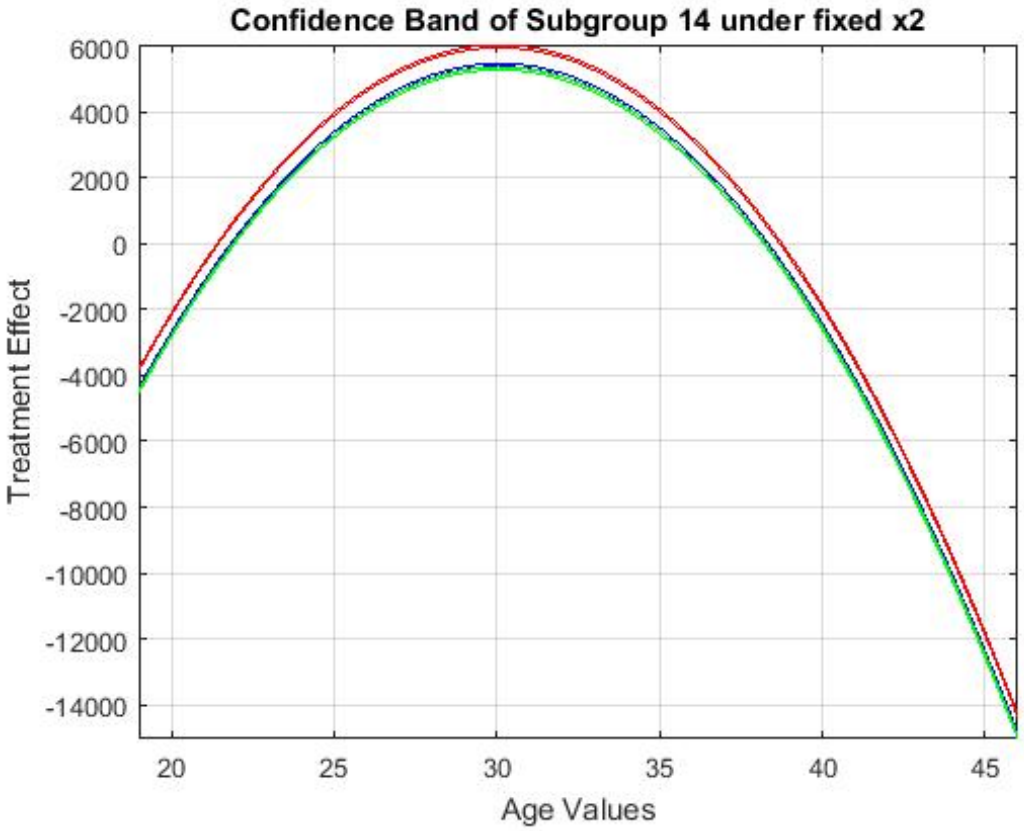

Figure 5.14: Simultaneous Confidence Band of Y, ATE, for Subgroup 14 on fixed minimum, mean and maximum values of x_2. The red, blue and green confidence bands indicate the behavior of ATE as a function of x_1 given a fixed value of x_2 at the minimum, mean and maximum, respectively.

5.4.2 *Lilly* Clinical Trial Data

For the *Lilly* clinical trial data, 8 non-sparse subgroups are formed using the categorical variables. The magnitudes of difference in hemoglobin A1C levels between *dulaglutide* and *liraglutide* are illustrated using confidence bands. As an initial step, regression estimates are generated. A linear regression model was used to estimate the mean hemoglobin A1C level for each of the treatment groups. The results are shown in Table 5.10. Also, the covariate regions of the independent variables, age (x_1) and body mass index (BMI)(x_2), are reflected in the same table.

Subgroup	Model Estimates	Covariate Region
1	$\hat{\mathbf{y}}_T = 5.3096 - 0.0005X_1 + 0.0601X_2$ $\hat{\mathbf{y}}_C = 6.8927 + 0.0045X_1 + 0.0087X_2$ $\hat{\beta} = (-1.5830, -0.0050, 0.0514)$	$x_1 : [28, 72]$ $x_2 : [19.8, 44.2]$
2	$\hat{\mathbf{y}}_T = 3.3269 + 0.0116X_1 + 0.0078X_2$ $\hat{\mathbf{y}}_C = 1.9487 - 0.0102X_1 + 0.1578X_2$ $\hat{\beta} = (1.3782, 0.0218, -0.0800)$	$x_1 : [23, 67]$ $x_2 : [26.3, 41.7]$
4	$\hat{\mathbf{y}}_T = 5.7410 + 0.0361X_1 - 0.0207X_2$ $\hat{\mathbf{y}}_C = 3.5169 + 0.0149X_1 + 0.0995X_2$ $\hat{\beta} = (2.2240, 0.0212, -0.1202)$	$x_1 : [35, 74]$ $x_2 : [20.8, 44.3]$
5	$\hat{\mathbf{y}}_T = 9.1803 - 0.0807X_1 + 0.0487X_2$ $\hat{\mathbf{y}}_C = 3.2242 - 0.0030X_1 + 0.1292X_2$ $\hat{\beta} = (5.9561, -0.0776, -0.0806)$	$x_1 : [41, 67]$ $x_2 : [24.5, 34.5]$
7	$\hat{\mathbf{y}}_T = 6.0661 - 0.0132X_1 + 0.0349X_2$ $\hat{\mathbf{y}}_C = 6.6703 - 0.0049X_1 - 0.0023X_2$ $\hat{\beta} = (-0.6042, -0.0083, 0.0372)$	$x_1 : [33, 75]$ $x_2 : [24.6, 47.2]$
9	$\hat{\mathbf{y}}_T = 5.0199 + 0.0101X_1 + 0.0270X_2$ $\hat{\mathbf{y}}_C = 6.8766 + 0.0127X_1 - 0.0300X_2$ $\hat{\beta} = (-1.8567, -0.0026, 0.0569)$	$x_1 : [35, 77]$ $x_2 : [21.8, 44.8]$
10	$\hat{\mathbf{y}}_T = 7.8321 - 0.0050X_1 - 0.0342X_2$ $\hat{\mathbf{y}}_C = 4.5625 + 0.0238X_1 + 0.0142X_2$ $\hat{\beta} = (3.2696, -0.0288, -0.0484)$	$x_1 : [38, 74]$ $x_2 : [24.1, 43.6]$
12	$\hat{\mathbf{y}}_T = 6.8136 - 0.0106X_1 + 0.0094X_2$ $\hat{\mathbf{y}}_C = 9.0976 - 0.0198X_1 - 0.0363X_2$ $\hat{\beta} = (-2.2840, 0.0092, 0.0456)$	$x_1 : [35, 73]$ $x_2 : [21.4, 45.7]$

Table 5.10: Regression Estimates of *Lilly Data*

Based on the linear regression estimates in Table 5.10 , confidence bands are created for the specified covariate regions in each subgroup. To do so, the critical constant c is simulated and the data-based components σ and $\mathbf{\Delta}$ are calculated. Figures 5.15 - 5.22 show the generated confidence bands. Looking at the surfaces of each subgroup, we observe that the zero hyperplane is inside the band for all $\mathbf{x}_{(\mathbf{0})}$ in

the covariate regions for all 8 subgroups. This suggests that the average treatment effect is insignificant for all groups.

The conclusions drawn from the confidence bands can be further established using the p-value of the simultaneous inference in regression models, shown in Table 5.11. For all subgroups considered, the p-values are significantly greater than $\alpha = 0.05$ which lead to the non-rejection of the null hypothesis that $\beta_A - \beta_B = 0$. This establishes further the insignificant ATE between the two drugs; thus, *dulaglutide* is non-inferior to *liraglutide* in stabilizing hemoglobin A1C levels among type II diabetes patients.

Subgroup	**p-value**	**Conclusion**
1	0.8282	Do not reject Ho
2	0.6638	Do not reject Ho
4	0.2584	Do not reject Ho
5	0.7311	Do not reject Ho
7	0.6018	Do not reject Ho
9	0.0940	Do not reject Ho
10	0.7061	Do not reject Ho
12	0.1028	Do not reject Ho

Table 5.11: Simultaneous Inference p-values on Subgroups of *Lilly* Data

Although the regions of the covariates considered are relatively wider than the NSW data in Section 5.4.1, the behavior of the average treatment effect was still explored by holding one independent variable fixed and formulating the difference in treatment effect as a function of the other variable. Since BMI (x_2) has a narrower

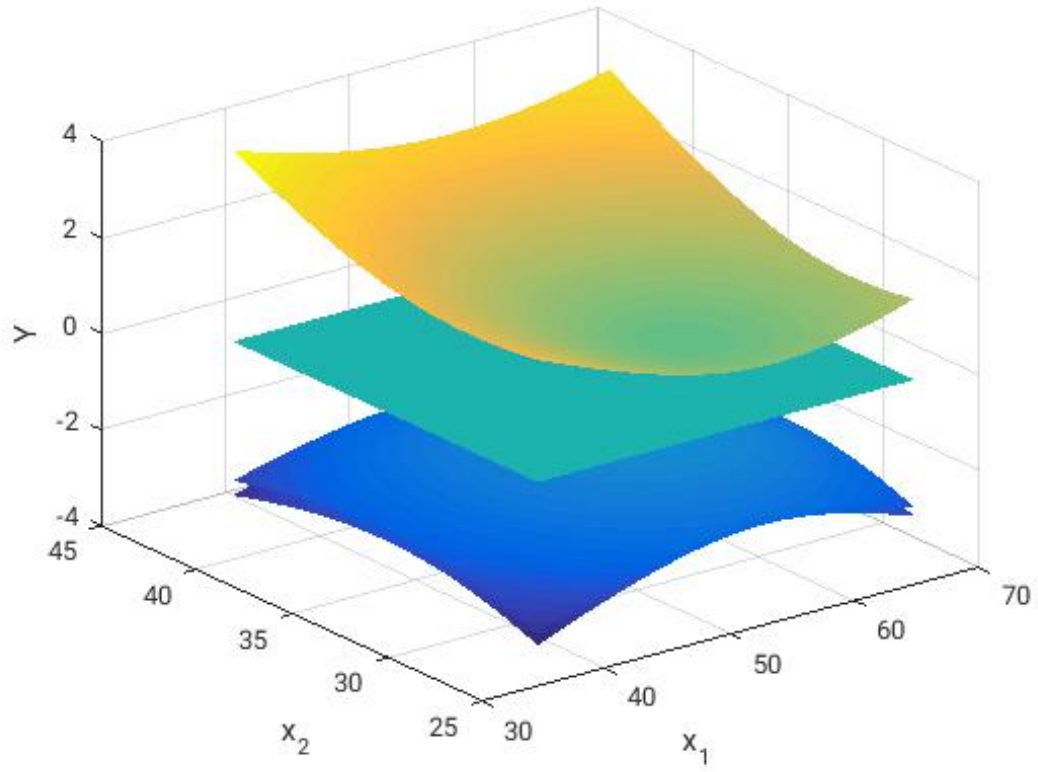

Figure 5.15: Simultaneous Confidence Band of Y, ATE, for Subgroup 1 on the restricted covariate region $x_1 : [28, 72]$ and $x_2 : [19.8, 44.2]$

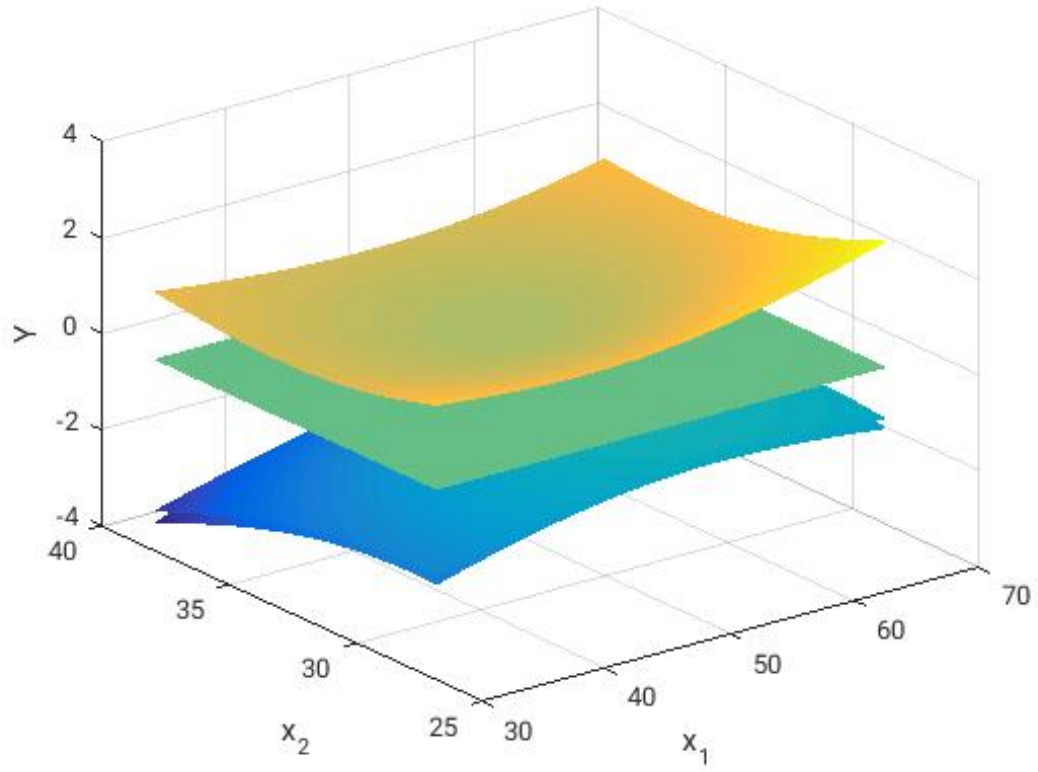

Figure 5.16: Simultaneous Confidence Band of Y, ATE, for Subgroup 2 on the restricted covariate region $x_1 : [23, 67]$ and $x_2 : [26.3.8, 41.7]$

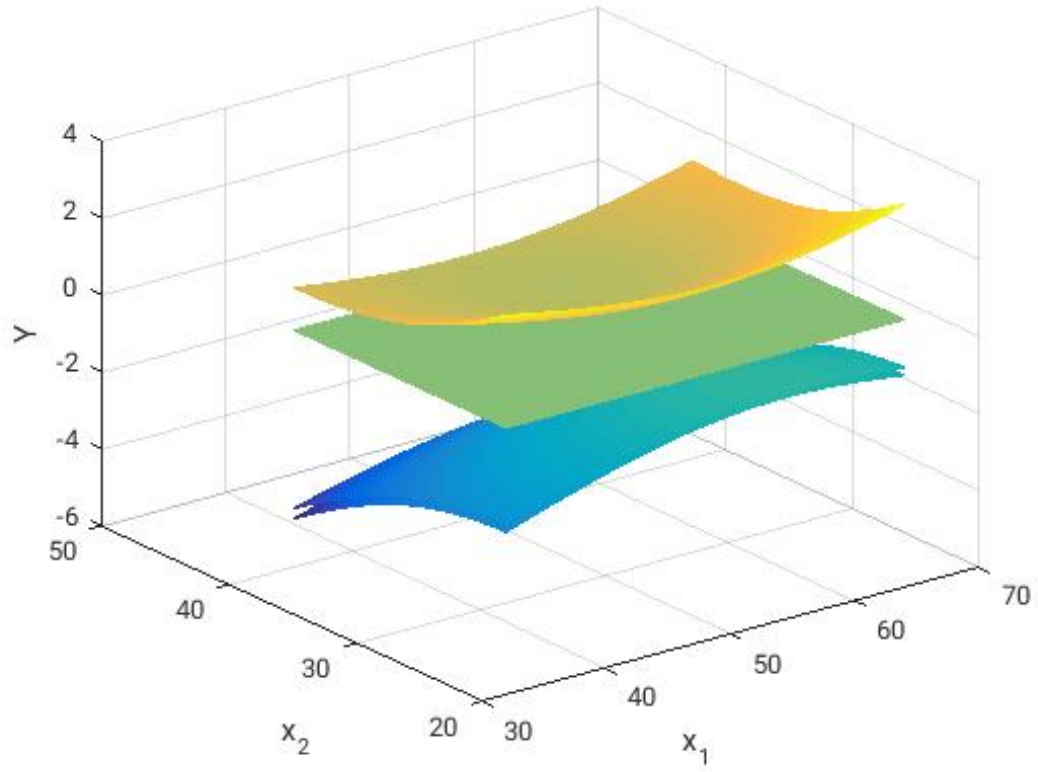

Figure 5.17: Simultaneous Confidence Band of Y, ATE, for Subgroup 4 on the restricted covariate region $x_1 : [35, 74]$ and $x_2 : [20.8, 44.3]$

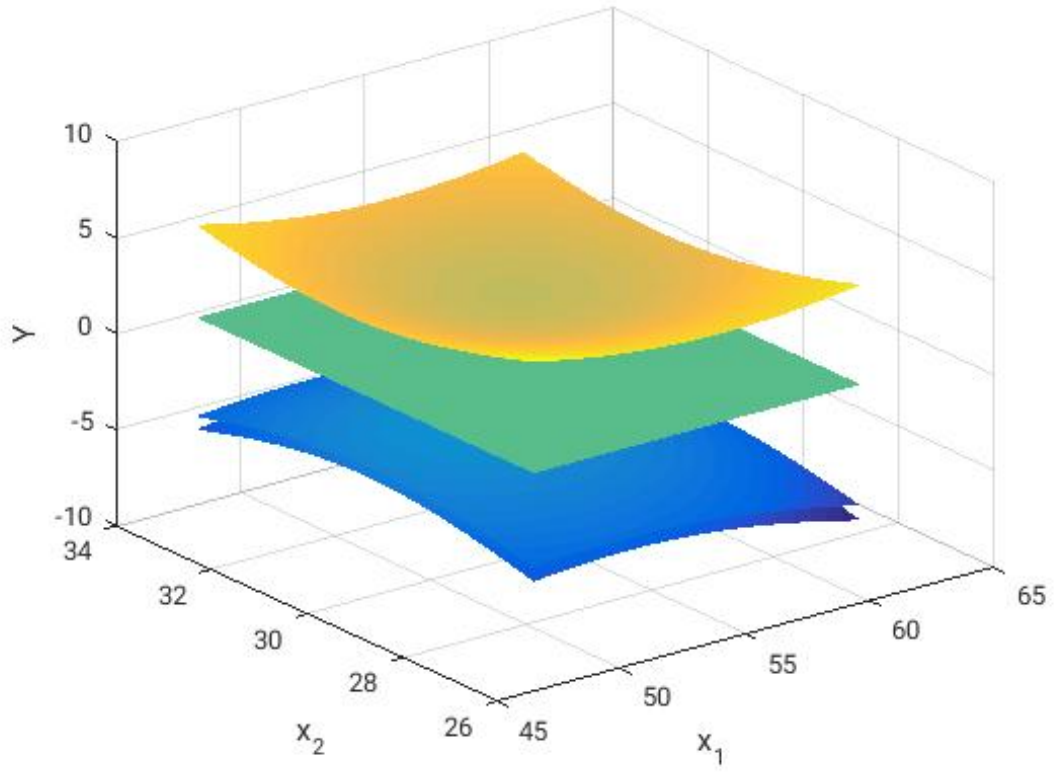

Figure 5.18: Simultaneous Confidence Band of Y, ATE, for Subgroup 5 on the restricted covariate region $x_1 : [41, 67]$ and $x_2 : [24.5, 34.5]$

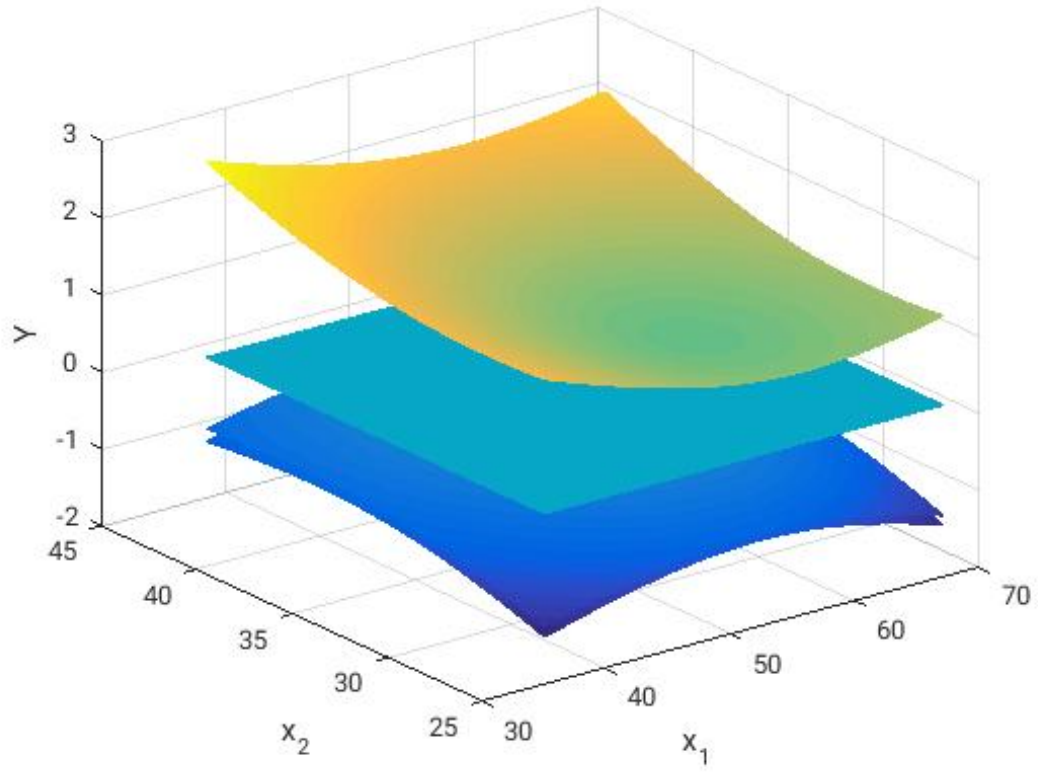

Figure 5.19: Simultaneous Confidence Band of Y, ATE, for Subgroup 7 on the restricted covariate region $x_1 : [33, 75]$ and $x_2 : [24.6, 47.2]$

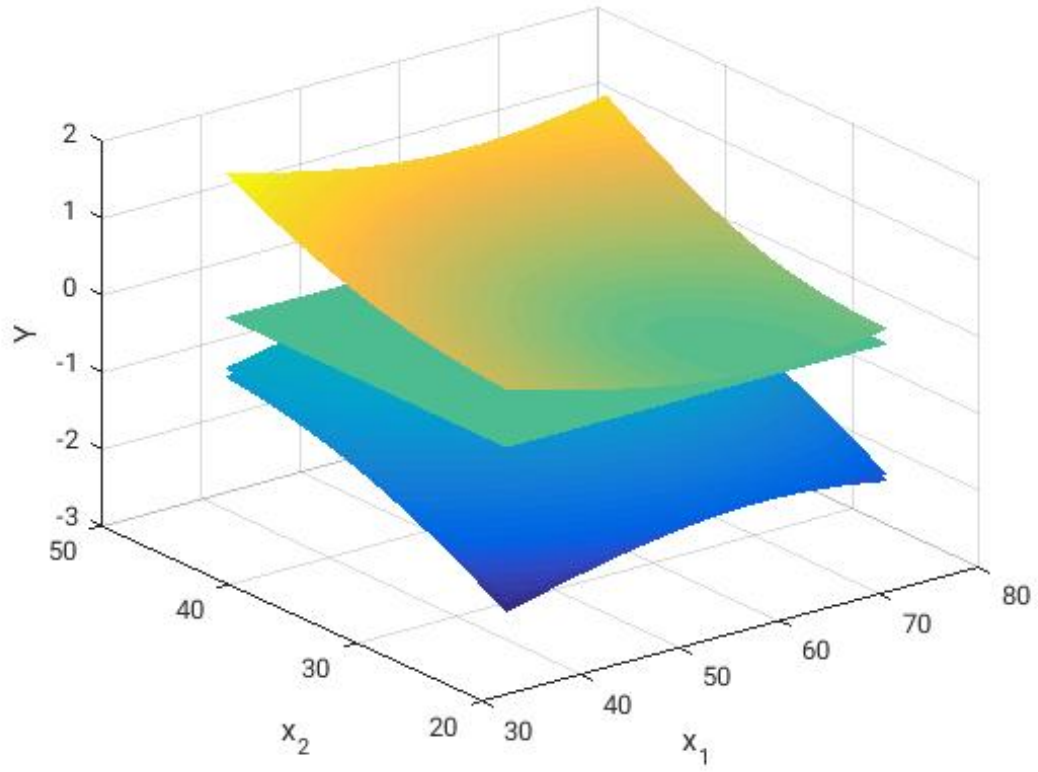

Figure 5.20: Simultaneous Confidence Band of Y, ATE, for Subgroup 9 on the restricted covariate region $x_1 : [35, 77]$ and $x_2 : [21.8, 44.8]$

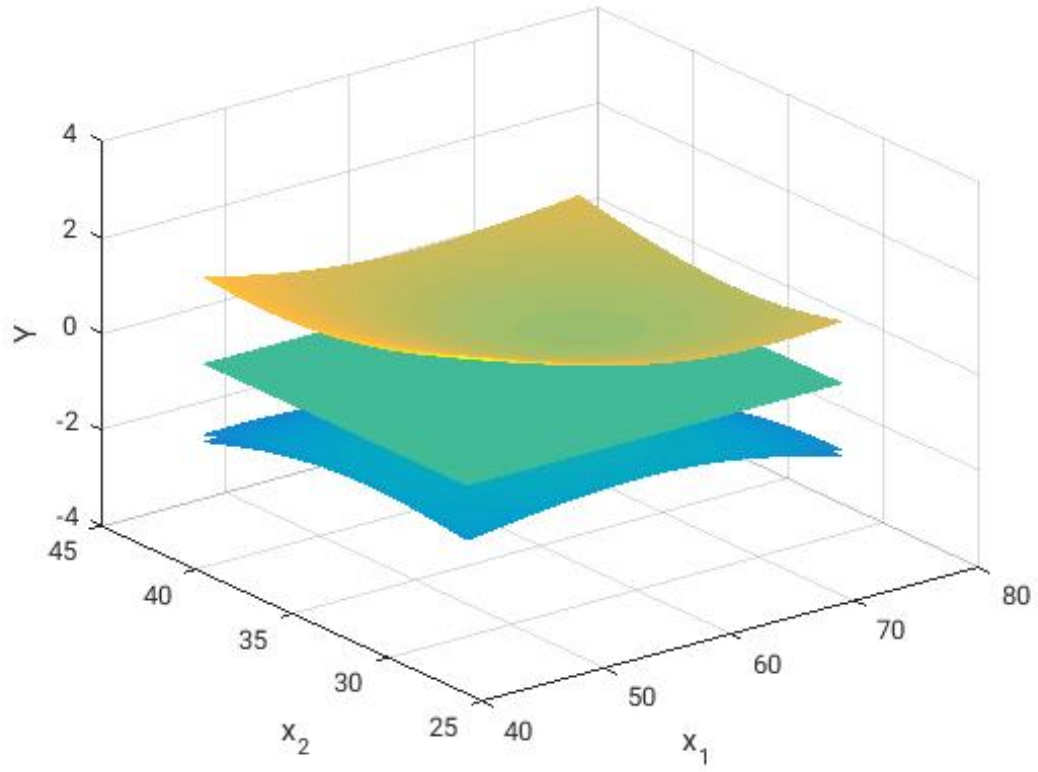

Figure 5.21: Simultaneous Confidence Band of Y, ATE, for Subgroup 10 on the restricted covariate region $x_1 : [38, 74]$ and $x_2 : [24.1, 43.6]$

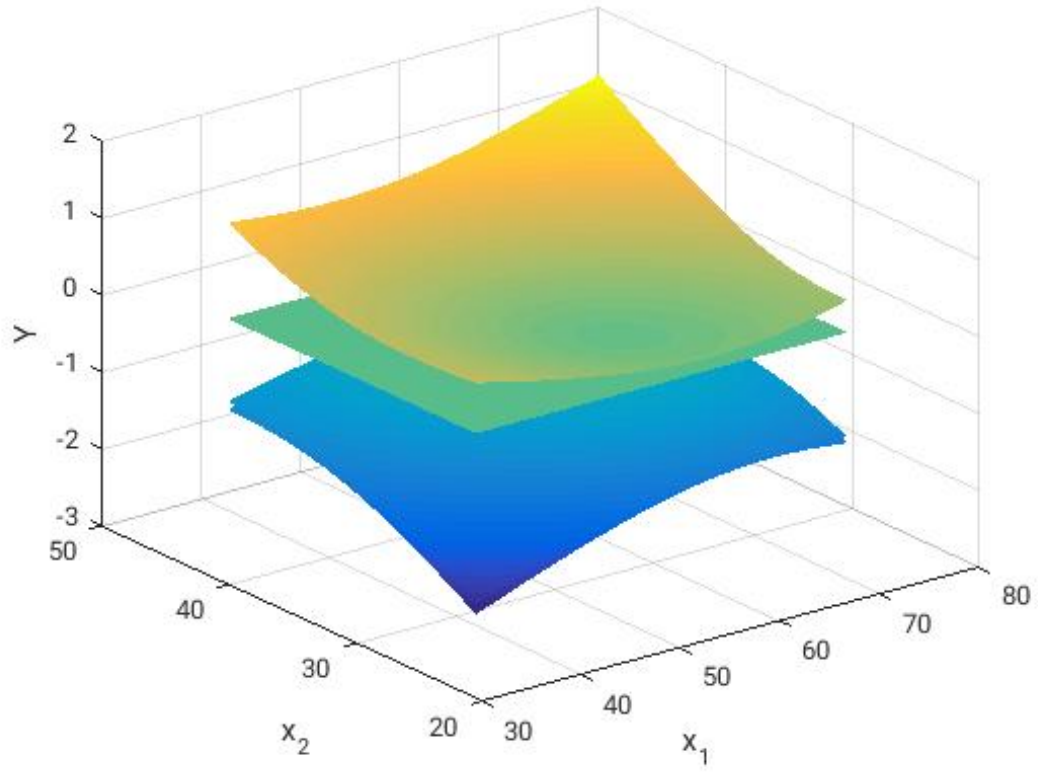

Figure 5.22: Simultaneous Confidence Band of Y, ATE, for Subgroup 12 on the restricted covariate region $x_1 : 35, 73]$ and $x_2 : [21.4, 45.7]$

range than age (x_1), a treatment effect function in terms of age (x_1) by holding (x_2) fixed, was derived. The fixed values of x_2 were also set at the minimum, mean and maximum BMI for each subgroup. These results are presented for each subgroup in Tables 5.12 - 5.19. In these tables, one may also be able to identify the simulated critical constant c and corresponding σ of the data set in each subgroup.

Looking at the solutions of these bands at a fixed x_1, it can be deduced that for any value of x_1 on the given fixed values of x_2, the ATE is non-zero for subgroups 1, 4, 5, 7, 9 and 12. For the remaining subgroups 2 and 10, there are real solutions calculated that for specific values of x_1. This indicates that there is a change in ATE behavior across some interval of x_1 given the fixed values of x_2 considered. Hence, no absolute behavior of the treatment effect can be determined for these subgroups. Figures 5.23 - 5.30 reflect the behavior of the response as a function of x_1. The results show that the relationship of x_1 and the treatment effect is depicted by hyperbolic bands that include the x_1-axis in most subgroups.

For subgroup 1, no solution for the confidence band as a function of x_1 is calculated, as shown in Table 5.12. Given the hyperbolic relationship between Y and x_1, this suggests a consistent behavior in the ATE for majority of the groups, as shown in Figure 5.23. This implies that hyperbolic bands generated for each fixed value of x_2 envelop the x_1-axis; thus, one treatment does not perform better over the other. It is noticeable, however, that the corresponding hyperbolic band for the fixed mean value of x_2 provides the narrowest band among the 3 fixed values of x_2 considered. These hyperbolic bands also tend to be narrower towards the mean of the observed x_1 but wider near the two ends. This indicates that a wider range of

disparity in hemoglobin levels are observed for younger or older patients between the two groups. However, with the inclusion of $y = 0$ in the bands, the superiority of a drug still cannot be established.

x_2	**Confidence Band:** $\boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C \pm c\hat{\sigma}\sqrt{x'\Delta x}}$	**Solution on** χ_R
19.8	$\frac{-250x_1-28269}{50000} \pm (2.2614)(1.1440)\sqrt{\frac{200x_1^2-21785x_1+701094}{125000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
31.4	$\frac{-250x_1+1543}{50000} \pm (2.2614)(1.1440)\sqrt{\frac{200x_1^2-21205x_1+583876}{125000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
44.2	$\frac{-250x_1+34439}{50000} \pm (2.2614)(1.1440)\sqrt{\frac{200x_1^2-20565x_1+661444}{125000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$

Table 5.12: Confidence Bands for Subgroup 1 under fixed x_2

Meanwhile, subgroup 2 displays a different behavior. From the results shown in Table 5.13, real solutions for the upper band are calculated. This suggests that for some interval of x_1, consistent behavior of ATE is observed. Based on its corresponding figure, Figure 5.24, it can be observed that given a fixed x_2 at the minimum, the ATE is negative for $x_1 \in [23, 56.26]$. This means that for female subjects from South America with Hispanic descent and with BMI = 26.8, *liraglutide* produces increased hemoglobin A1C levels. For the mean and maximum values of x_2, although calculations show that the hyperbolic band has real solutions, these values are outside the covariate region of interest of x_1. As shown in Figure 5.24, the ATE is likewise negative. Therefore, under the same categorical subject profile mentioned and BMI equal to 31.7 and 41.7, *liraglutide* provides higher hemoglobin levels. Generally for this subgroup, we can observe the inferiority of *dulaglutide* as a treatment.

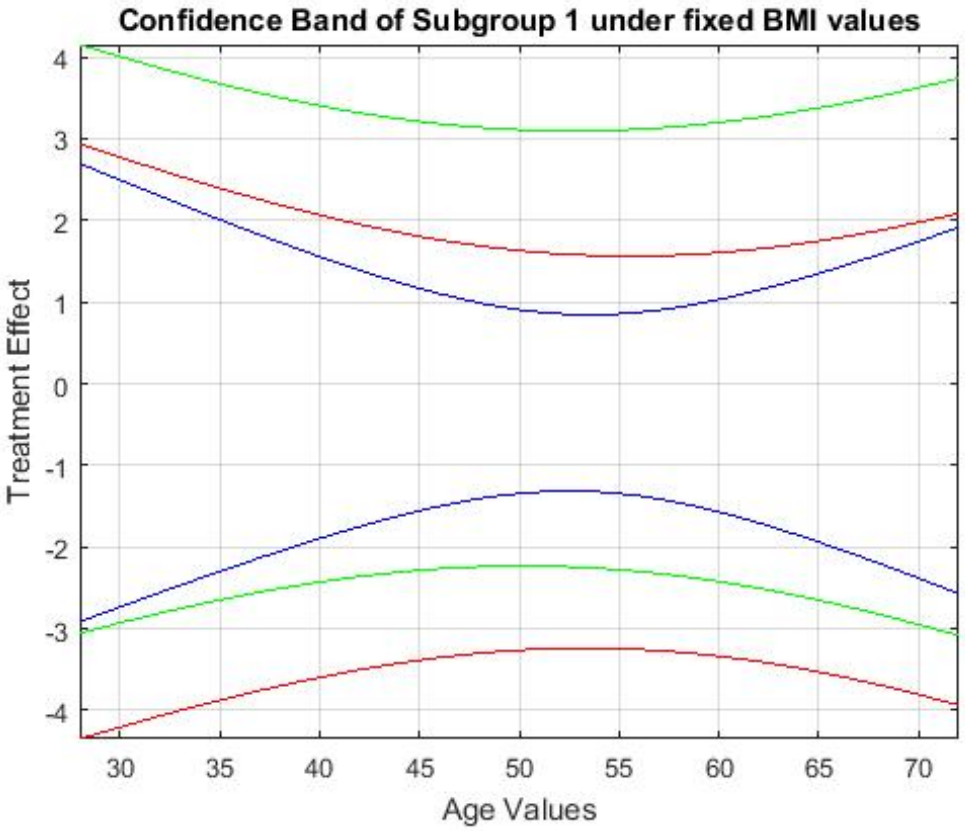

Figure 5.23: Simultaneous Confidence Band of Y, ATE, for Subgroup 1 on fixed minimum, mean and maximum values of x_2. The red, blue and green confidence bands indicate the behavior of ATE as a function of x_1 given a fixed value of x_2 at the minimum, mean and maximum, respectively.

It is also interesting to note that among the three fixed values of x_2, the mean generates the narrowest hyperbolic band.

x_2	**Confidence Band:** $\boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C \pm c\hat{\sigma}\sqrt{x'\Delta x}}$	**Solution on** χ_R
26.8	$\frac{109x_1-13209}{5000} \pm (2.6702)(0.7357)\sqrt{\frac{300x_1^2-28080x_1+759951}{250000}}$	$x_{1+} = 20.32, 56.26$ $x_{1-} = \phi$
31.7	$\frac{109x_1-16884}{5000} \pm (2.6702)(0.7357)\sqrt{\frac{1200x_1^2-119180x_1+3128739}{1000000}}$	$x_{1+} = 2.21, 73.03$ $x_{1-} = \phi$
41.7	$\frac{109x_1-24384}{5000} \pm (2.6702)(0.7357)\sqrt{\frac{1200x_1^2-133180x_1+4964139}{1000000}}$	$x_{1+} = -13.02, 85.53$ $x_{1-} = \phi$

Table 5.13: Confidence Bands for Subgroup 2 under fixed x_2

Table 5.14 and Figure 5.25 illustrate the behavior of the covariates for subgroup 4 which consists of male, Hispanic respondents from North America. Based on the solutions of the hyperbolic bands, it can be concluded that the bands provided for the ATE estimate contains 0. This suggests that one drug is non-inferior over the other. The width of the bands for the minimum and maximum values of x_2 are comparable, while the confidence band on a fixed maximum value of x_2 is considerably wide. In all three fixed values though, the band is narrower towards the mean of the observed x_1-values. The wide band may be an indicator that for subjects with this subject profile and has high BMI, *liraglutide* gives higher hemoglobin levels as shown in the majority of band being negative. This behavior is not quite apparent in the fixed minimum and mean values of x_2, because the band also contains comparable range in the positive side.

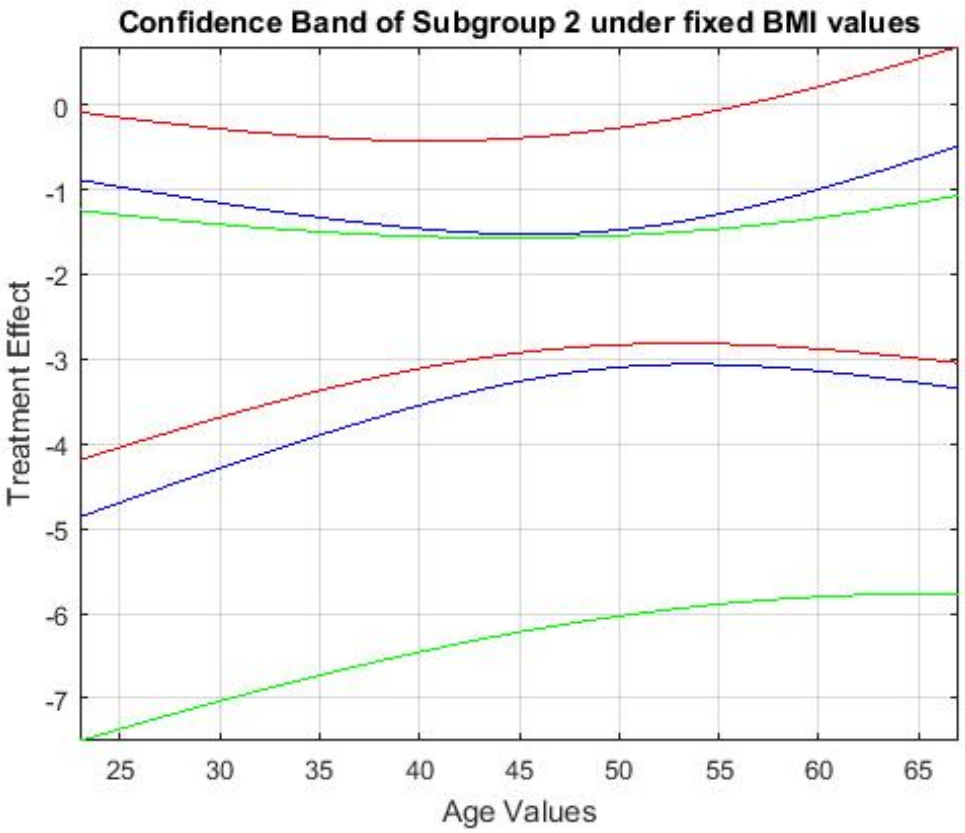

Figure 5.24: Simultaneous Confidence Band of Y, ATE, for Subgroup 2 on fixed minimum, mean and maximum values of x_2. The red, blue and green confidence bands indicate the behavior of ATE as a function of x_1 given a fixed value of x_2 at the minimum, mean and maximum, respectively.

x_2	**Confidence Band:** $\boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C \pm c\hat{\sigma}\sqrt{x'\Delta x}}$	**Solution on** χ_R
20.8	$\frac{-33x_1-3511}{10000} \pm (2.6551)(1.1259)\sqrt{\frac{6x_1^2-646x_1+19615}{5000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
30.5	$\frac{-66x_1-20971}{20000} \pm (2.6551)(1.1259)\sqrt{\frac{24x_1^2-2532x_1+72727}{20000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
44.3	$\frac{-330x_1-252929}{100000} \pm (2.6551)(1.1259)\sqrt{\frac{600x_1^2-60540x_1+2046151}{500000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$

Table 5.14: Confidence Bands for Subgroup 4 under fixed x_2

For Subgroup 5, no real solution for the hyperbolic bands is calculated, as shown in Table 5.15. This means that *dulaglutide* and *liraglutide* are not significantly different in increasing hemoglobin levels. Figure 5.26 shows this behavior as well. For this subgroup, the band is widest for a fixed minimum value of x_2 while the fixed mean is narrowest. All three hyperbolic bands are narrowest at the middle values of x_1. However, in all cases, superiority of one drug is not established.

x_2	**Confidence Band:** $\boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C \pm c\hat{\sigma}\sqrt{x'\Delta x}}$	**Solution on** χ_R
24.5	$\frac{-1554x_1+79677}{20000} \pm (3.2799)(1.2179)\sqrt{\frac{416x_1^2-44792x_1+1300293}{40000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
30.2	$\frac{-777x_1+35250}{10000} \pm (3.2799)(1.2179)\sqrt{\frac{2600x_1^2-281660x_1+7671387}{250000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
34.5	$\frac{-1554x_1+63577}{20000} \pm (3.2799)(1.2179)\sqrt{\frac{416x_1^2-45272x_1+1271693}{40000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$

Table 5.15: Confidence Bands for Subgroup 5 under fixed x_2

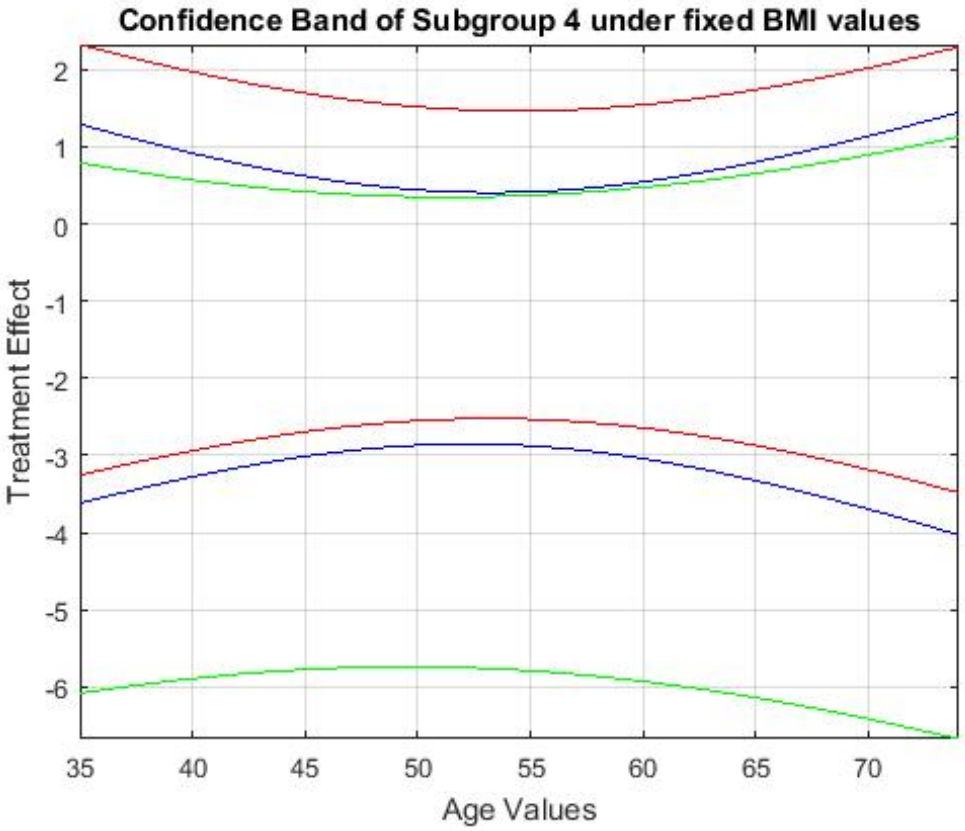

Figure 5.25: Simultaneous Confidence Band of Y, ATE, for Subgroup 4 on fixed minimum, mean and maximum values of x_2. The red, blue and green confidence bands indicate the behavior of ATE as a function of x_1 given a fixed value of x_2 at the minimum, mean and maximum, respectively.

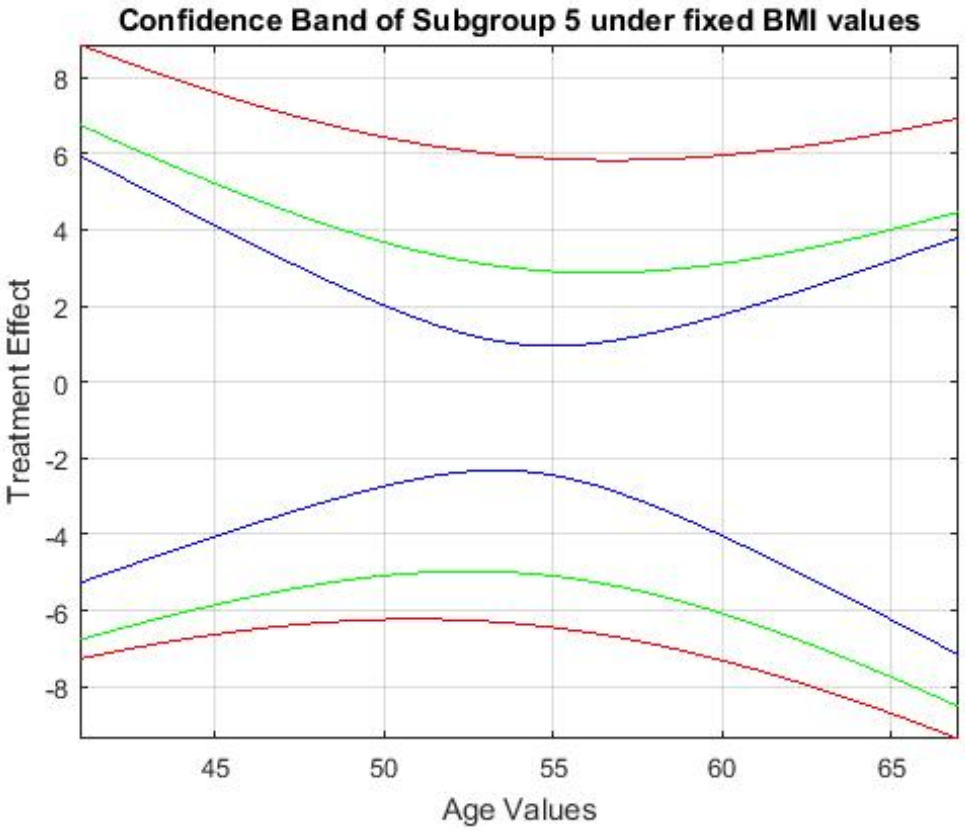

Figure 5.26: Simultaneous Confidence Band of Y, ATE, for Subgroup 5 on fixed minimum, mean and maximum values of x_2. The red, blue and green confidence bands indicate the behavior of ATE as a function of x_1 given a fixed value of x_2 at the minimum, mean and maximum, respectively.

Table 5.16 and Figure 5.27 show the result for Subgroup 7, comprising of female, non-hispanic subjects from North America. With the non-real solutions of the hyperbolic bands and their corresponding behavior reflected in Figure **??**, it is observed that the two drugs considered have a non-inferior effect for this type of subject profile. Like the previous subgroups, the bands are narrowest at the extreme ends of the observed values of x_1 and tend to have narrower effect towards the center, given any of the fixed effect of x_2.

x_2	**Confidence Band:** $\boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C \pm c\hat{\sigma}\sqrt{x'\Delta x}}$	**Solution on** χ_R
24.6	$\frac{-415x_1+15546}{50000} \pm (2.6176)(0.7585)\sqrt{\frac{250x_1^2-28181x_1+867449}{250000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
33.8	$\frac{-415x_1+32658}{50000} \pm (2.6176)(0.7585)\sqrt{\frac{250x_1^2-28043x_1+806821}{250000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
47.2	$\frac{-415x_1+57582}{50000} \pm (2.6176)(0.7585)\sqrt{\frac{250x_1^2-27842x_1+915361}{250000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$

Table 5.16: Confidence Bands for Subgroup 7 under fixed x_2

Also, *dulaglutide* and *liraglutide* are shown to be non-inferior for female, non-hispanic Europeans as shown in Table 5.17 and Figure 5.28. However, a closer look at the hyperbolic bands show that for a fixed value of x_2 at the minimum, majority of the band lies on the negative treatment effect. Meanwhile, when x_2 is fixed at the maximum, this band tends to move upward, causing majority of the difference in the treatment groups to be positive. This suggests that under this subject profile, patients with lower BMI will have increased hemoglobin levels when taking *liraglutide* while those who have higher BMI will have higher hemoglobin

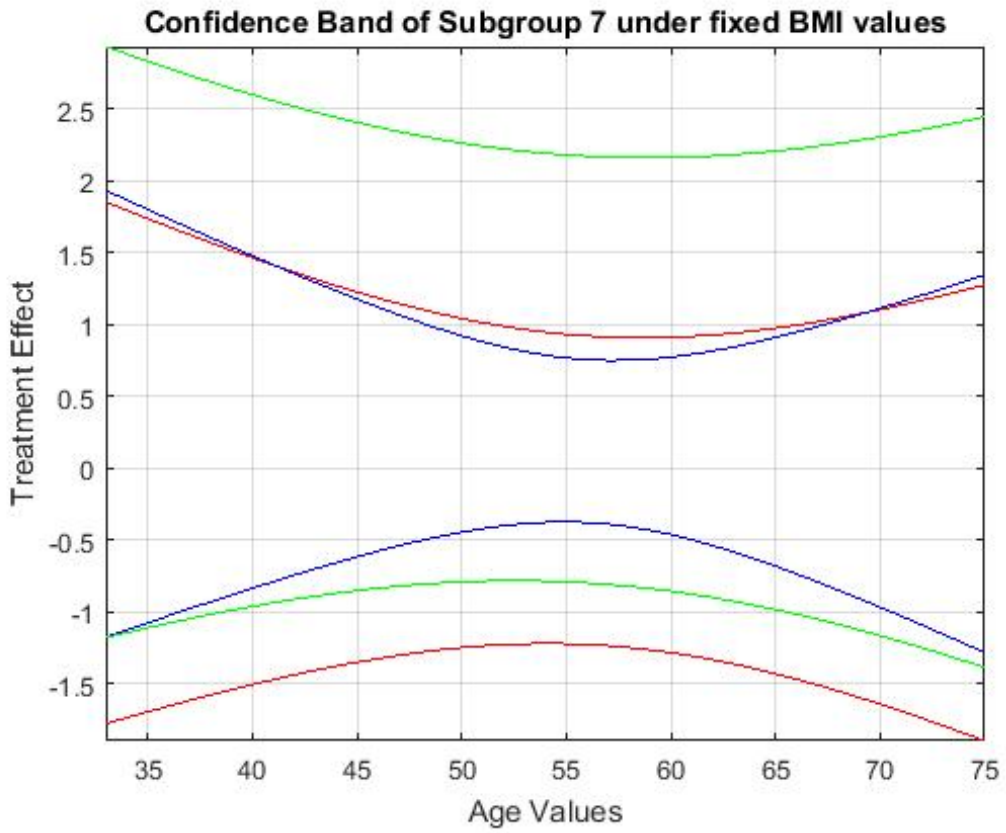

Figure 5.27: Simultaneous Confidence Band of Y, ATE, for Subgroup 7 on fixed minimum, mean and maximum values of x_2. The red, blue and green confidence bands indicate the behavior of ATE as a function of x_1 given a fixed value of x_2 at the minimum, mean and maximum, respectively.

levels when taking *dulaglutide*. At the mean value of x_2, the hyperbolic band does not tend to favor the positive nor the negative direction.

x_2	**Confidence Band:** $\boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C \pm c\hat{\sigma}\sqrt{x'\Delta x}}$	**Solution on** χ_R
21.8	$\frac{-26x_1-6141}{10000} \pm (2.5760)(0.7112)\sqrt{\frac{125x_1^2-14919x_1+512988}{250000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
32.8	$\frac{-26x_1+129}{10000} \pm (2.5760)(0.7112)\sqrt{\frac{125x_1^2-14424x_1+451883}{250000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
44.8	$\frac{-26x_1+6969}{10000} \pm (2.5760)(0.7112)\sqrt{\frac{125x_1^2-13884x_1+474923}{250000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$

Table 5.17: Confidence Bands for Subgroup 9 under fixed x_2

Table 5.18 and Figure 5.29 show the results derived for subgroup 10. This subgroup is comprised of male, non-hispanic respondents from North America. The superiority of one drug is still not established for this group. However, it is interesting to note that x_1 is undefined for a fixed minimum x_2. Therefore, for subjects with this profile and has BMI = 24.1, we are not able to provide any observation of the behavior of Y in terms of x_1. On the available hyperbolic bands, the mean continues to produce narrower bands than the maximum.

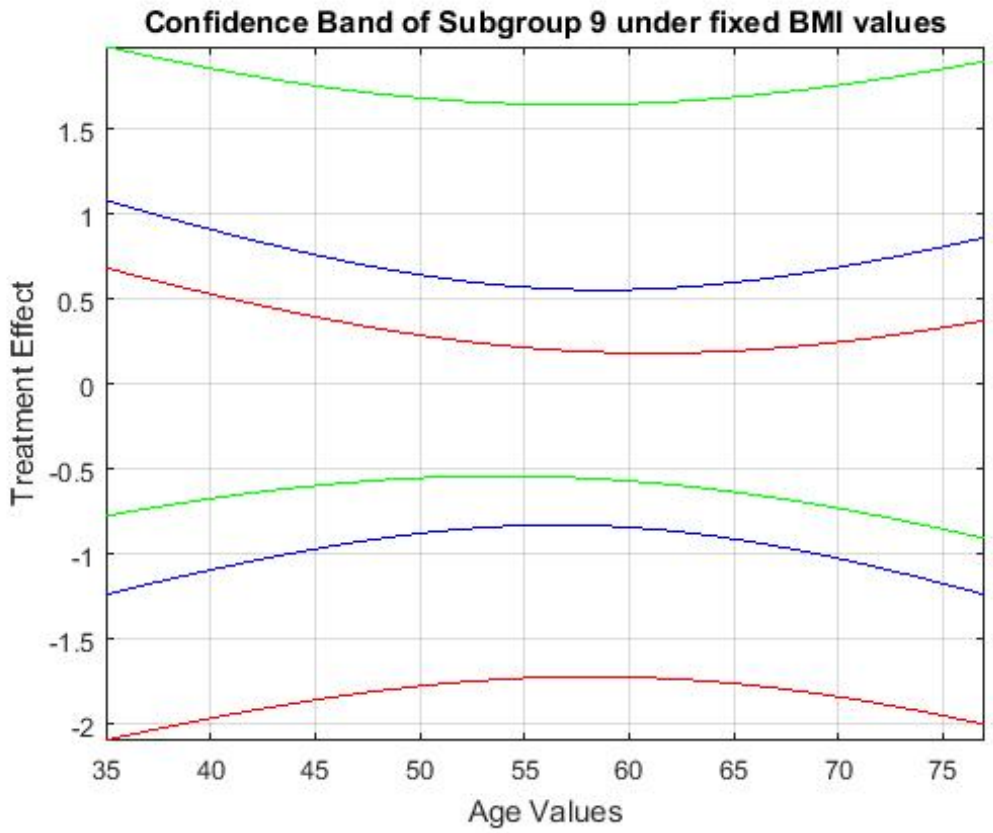

Figure 5.28: Simultaneous Confidence Band of Y, ATE, for Subgroup 9 on fixed minimum, mean and maximum values of x_2. The red, blue and green confidence bands indicate the behavior of ATE as a function of x_1 given a fixed value of x_2 at the minimum, mean and maximum, respectively.

x_2	**Confidence Band:** $\boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C \pm c\hat{\sigma}\sqrt{x'\Delta x}}$	**Solution on** χ_R
24.1	$\frac{-720x_1+52579}{25000} \pm (2.6416)(0.7334)\sqrt{\frac{700x_1^2-77865x_1+1830141}{500000}}$	$x_{1+} = 77.58$ $x_{1-} = 27.13$
32.7	$\frac{-720x_1+42173}{25000} \pm (2.6416)(0.7334)\sqrt{\frac{1400x_1^2-161940x_1+4815169}{1000000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
43.6	$\frac{-90x_1+3623}{3125} \pm (2.6416)(0.7334)\sqrt{\frac{350x_1^2-35580x_1+1115039}{250000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$

Table 5.18: Confidence Bands for Subgroup 10 under fixed x_2

For the last subgroup, the same observation is made about the non-inferiority of the two drugs being studied. Results are shown in Table 5.19 and Figure 5.30. While no treatment effect is realized on all three bands, we note that majority of the band generated under fixed minimum and mean values of x_2 tend towards the negative direction. Thus, it can be observed that for male, non-hispanic Europeans with low to moderate BMI, the drug *liraglutide* may be a more effective treatment in increasing hemoglobin levels among diabetic patients.

x_2	**Confidence Band:** $\boldsymbol{x'\hat{\beta}_T - x'\hat{\beta}_C \pm c\hat{\sigma}\sqrt{x'\Delta x}}$	**Solution on** χ_R
21.4	$\frac{460x_1-65301}{50000} \pm (2.5840)(0.7386)\sqrt{\frac{25x_1^2-2918x_1+102006}{50000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
32.6	$\frac{460x_1-39709}{50000} \pm (2.5840)(0.7386)\sqrt{\frac{25x_1^2-2582x_1+72438}{50000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$
45.7	$\frac{920x_1-19551}{100000} \pm (2.5840)(0.7386)\sqrt{\frac{100x_1^2-8756x_1+246915}{200000}}$	$x_{1+} = \phi$ $x_{1-} = \phi$

Table 5.19: Confidence Bands for Subgroup 12 under fixed x_2

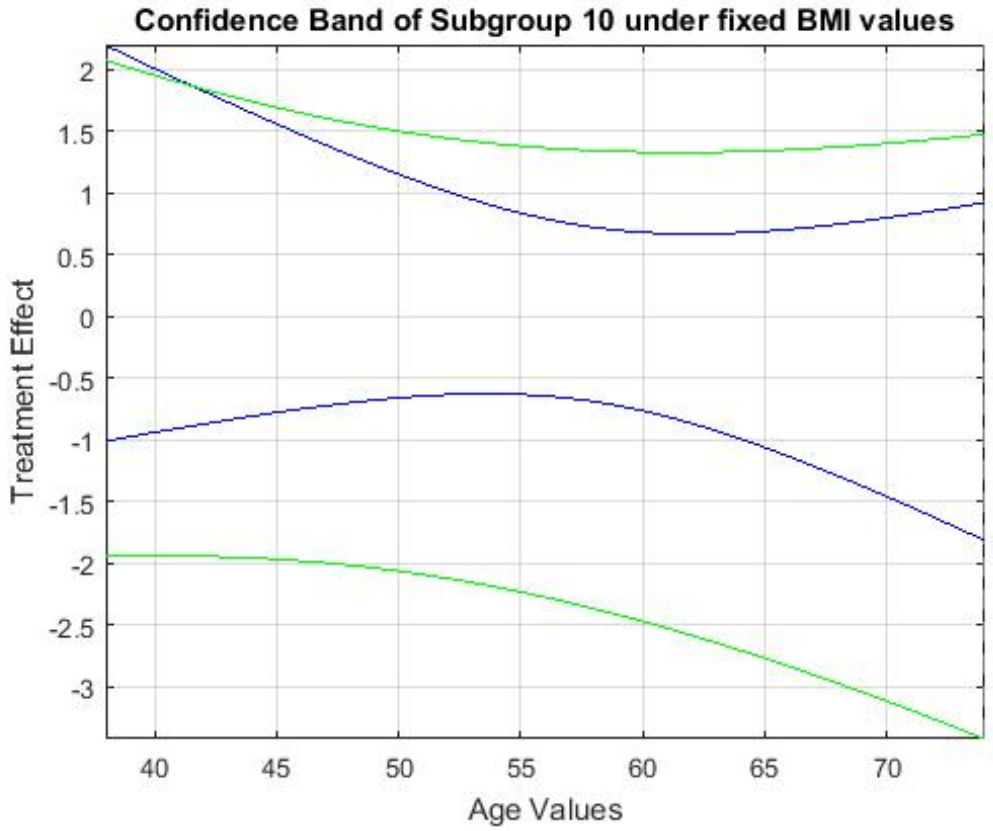

Figure 5.29: Simultaneous Confidence Band of Y, ATE, for Subgroup 10 on fixed minimum, mean and maximum values of x_2. The red, blue and green confidence bands indicate the behavior of ATE as a function of x_1 given a fixed value of x_2 at the minimum, mean and maximum, respectively.

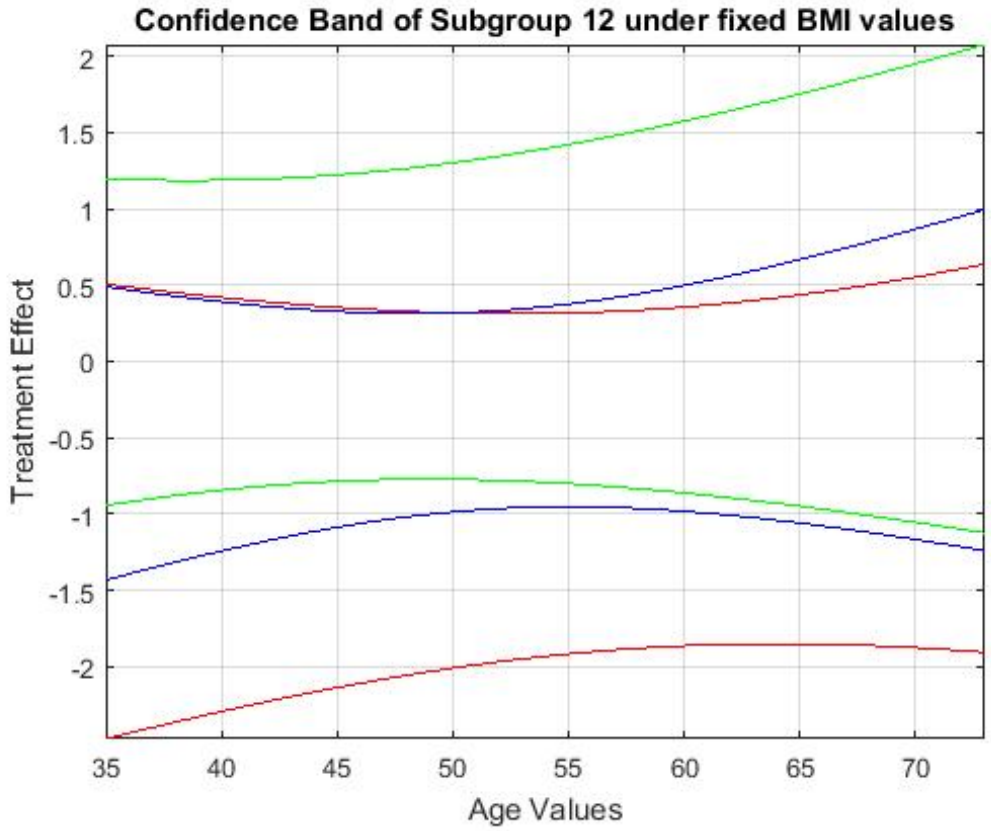

Figure 5.30: Simultaneous Confidence Band of Y, ATE, for Subgroup 12 on fixed minimum, mean and maximum values of x_2. The red, blue and green confidence bands indicate the behavior of ATE as a function of x_1 given a fixed value of x_2 at the minimum, mean and maximum, respectively.

In general, as illustrated in Figure 5.23 to Figure 5.30, there is an insignificant treatment effect given the range of x_1 of the different subgroups under a fixed x_2. This supports the surfaces generated as confidence bands over the entire space of interest of x_1 and x_2.

As a summary, the results of the exploratory analysis of the *Lilly* data set suggest that there is no significant treatment effect between the two drugs. This means that the drug *dulaglutide* is non-inferior to *liraglutide* in stabilizing the hemoglobin A1C levels of type II diabetes patients. However, with this analysis, we are able to identify the treatment effect behavior given some information on a subject's region of origin, civil status, gender, ethnicity, BMI and age.

Chapter 6

Conclusion

In this dissertation, an alternative method for estimation of average treatment effect (ATE) is exploited. The proposed method first requires balancing among the categorical variables resulting in several homogeneous subgroups and then performing a detailed statistical analysis within each subgroup for drawing valid inference about ATE. A new method designated as the "swapping" method is introduced for this purpose and used for subsequent inference. A homogeneous test for the ATEs across the homogeneous subgroups is performed and appropriate meta-analysis methods (fixed or random) is used for the eventual estimation of the overall ATE.

The novel method of simultaneous inference in regression is also used to explain analysis of ATE based on subject covariate profile. Two real data sets are utilized as illustrations of our proposed methods.

Appendix A

Proof of the distributional properties of the proposed estimator

The distributional properties of $\underset{\sim}{\delta}$ for a fixed class i of the treatment group is given by

$$\underset{\sim}{\delta} = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_{n_i} \end{pmatrix} \sim N_{n_i}\left(\alpha \underset{\sim}{1}, \Psi\right) \tag{A.1}$$

with

$$\begin{aligned}
\alpha = E(\delta_j) &= \left[\beta_{0,T}^{'} + \beta_1(v_{j1} - \bar{v}_1) + \ldots + \beta_p(v_{jp} - \bar{v}_p)\right] - \\
&\quad \left[\beta_{0,C}^{'} + \beta_1(v_{j1} - \bar{w}_1) + \ldots + \beta_p(v_{jp} - \bar{w}_p)\right] \\
&= (\beta_{0,T}^{'} - \beta_{0,C}^{'}) + \beta_1(\bar{w}_1 - \bar{v}_1) + \ldots + \beta_p(\bar{w}_p - \bar{v}_p) \\
&= (\beta_{0,T} + \beta_1\bar{v}_1 + \ldots + \beta_p\bar{v}_p - \beta_{0,C} - \beta_1\bar{w}_1 - \ldots - \beta_p\bar{w}_p) + \\
&\quad \beta_1(\bar{w}_1 - \bar{v}_1) + \ldots + \beta_p(\bar{w}_p - \bar{v}_p) \\
&= \beta_{0,T} - \beta_{0,C} + \beta_1(\bar{v}_1 - \bar{w}_1) + \ldots + \beta_p(\bar{v}_p - \bar{w}_p) + \\
&\quad \beta_1(\bar{w}_1 - \bar{v}_1) + \ldots + \beta_p(\bar{w}_p - \bar{v}_p) \\
&= \beta_{0,T} - \beta_{0,C}.
\end{aligned} \tag{A.2}$$

For the elements of the Ψ matrix, consider unit j of the ith subgroup where

the observed and estimated values are

$$\begin{aligned} U_j &= \beta_{0,T}^{'} + \beta_1(v_{j1} - \bar{v}_1) + \ldots + \beta_p(v_{jp} - \bar{v}_p) \\ \tilde{U}_j &= \hat{\beta}_{0,C}^{'} + \hat{\beta}_1(v_{j1} - \bar{w}_1) + \ldots + \hat{\beta}_p(v_{jp} - \bar{w}_p) \\ &= \bar{Z} + \frac{1}{\sigma_T^2}(v_{j1} - \bar{w}_1)\underset{\sim}{d}_{T1}\underset{\sim}{U} + \frac{1}{\sigma_C^2}(v_{j1} - \bar{w}_1)\underset{\sim}{d}_{C1}\underset{\sim}{Z} + \\ &\quad \ldots + \frac{1}{\sigma_T^2}(v_{jp} - \bar{w}_p)\underset{\sim}{d}_{Tp}\underset{\sim}{U} + \frac{1}{\sigma_C^2}(v_{jp} - \bar{w}_p)\underset{\sim}{d}_{Cp}\underset{\sim}{Z}, \end{aligned} \tag{A.3}$$

respectively. Note that, the quantity $Cov(U_j, \tilde{U}_j)$ can be calculated, while picking only those terms in the sum that involves U_j, as

$$\begin{aligned} Cov(U_j, \tilde{U}_j) &= Cov\left(U_j, \frac{(v_{j1} - \bar{w}_1)}{\sigma_T^2} d_{Tj1}U_j + \ldots + \frac{(v_{jp} - \bar{w}_p)}{\sigma_T^2} d_{Tjp}U_j\right) \\ &= \frac{(v_{j1} - \bar{w}_1)}{\sigma_T^2} d_{Tj1}Var(U_j) + \ldots + \frac{(v_{jp} - \bar{w}_p)}{\sigma_T^2} d_{Tj1}Var(U_p) \\ &= (v_{j1} - \bar{w}_1)d_{Tj1} + \ldots + (v_{jp} - \bar{w}_p)d_{Tjp} \\ &= \sum_{l=1}^{p}(v_{jl} - \bar{w}_l)d_{Tjl} \\ &= \underset{\sim}{v}_j^{'}\underset{\sim}{d}_{Tj}^{*} \end{aligned} \tag{A.4}$$

where $\underset{\sim}{v}_j = \begin{pmatrix} (v_{j1} - \bar{w}_1) \\ \vdots \\ (v_{jp} - \bar{w}_p) \end{pmatrix}$ and $\underset{\sim}{d}_{Tj}^{*} = \begin{pmatrix} d_{Tj1} \\ \vdots \\ d_{Tjp} \end{pmatrix}$. Therefore, the variance of the effect size is given by

$$\begin{aligned} Var(\delta_j) &= Var(U_j - \tilde{U}_j) = Var(U_j) + Var(\tilde{U}_j) - 2Cov(U_j, \tilde{U}_j) \\ &= \sigma_T^2 + \underset{\sim}{v}_j^{*'}\Lambda_c\underset{\sim}{v}_j^{*} - 2\underset{\sim}{v}_j^{'}\underset{\sim}{d}_{Tj}^{*} \end{aligned} \tag{A.5}$$

where Λ_C is the $(p+1) \times (p+1)$ dispersion matrix of $\tilde{U}_j$ and v_j^{*}, v_j and $\underset{\sim}{d}_{Tj}^{*}$ are as in the main text.

Following the same rationale as above for the calculation of the off-diagonal elements of Ψ and considering only the terms of $\tilde{U}_j$ that involves $U_{j'}$ for $j \neq j'$, the $Cov(U_{j'}, \tilde{U}_j)$ can be calculated as

$$\begin{aligned} Cov(U_{j'}, \tilde{U}_j) &= Cov\left(U_{j'}, \frac{(v_{j1} - \bar{w}_1)}{\sigma_T^2} d_{Tj'1} U_{j'} + \ldots + \frac{(v_{jp} - \bar{w}_p)}{\sigma_T^2} d_{Tj'p} U_{j'}\right) \\ &= \frac{(v_{j1} - \bar{w}_1)}{\sigma_T^2} d_{Tj'1} Var(U_{j'}) + \ldots + \frac{(v_{jp} - \bar{w}_p)}{\sigma_T^2} d_{Tj'p} Var(U_{j'}) \\ &= (v_{j1} - \bar{w}_1) d_{Tj'1} + \ldots + (v_{jp} - \bar{w}_p) d_{Tj'p} \\ &= \sum_{l=1}^{p} (v_{jl} - \bar{w}_l) d_{Tj'l} \\ &= \underset{\sim}{v}_j^{'} \underset{\sim}{d}^*_{Tj'} \end{aligned} \tag{A.6}$$

where $\underset{\sim}{v}_j$ is as previously defined and $\underset{\sim}{d}^*_{Tj'} = \begin{pmatrix} d_{Tj'1} \\ \vdots \\ d_{Tj'p} \end{pmatrix}$. It can also be shown that the $Cov(U_j, \tilde{U}_{j'})$ can be calculated as

$$\begin{aligned} Cov(U_j, \tilde{U}_{j'}) &= Cov\left(U_j, \frac{(v_{j'1} - \bar{w}_1)}{\sigma_T^2} d_{Tj1} U_j + \ldots + \frac{(v_{j'p} - \bar{w}_p)}{\sigma_T^2} d_{Tjp} U_j\right) \\ &= \frac{(v_{j'1} - \bar{w}_1)}{\sigma_T^2} d_{Tj1} Var(U_j) + \ldots + \frac{(v_{j'p} - \bar{w}_p)}{\sigma_T^2} d_{Tjp} Var(U_j) \\ &= (v_{j'1} - \bar{w}_1) d_{Tj1} + \ldots + (v_{j'p} - \bar{w}_p) d_{Tjp} \\ &= \sum_{l=1}^{p} (v_{j'l} - \bar{w}_l) d_{Tjl} \\ &= \underset{\sim}{v}_{j'}^{'} \underset{\sim}{d}^*_{Tj}. \end{aligned} \tag{A.7}$$

For the $Cov(\tilde{U}_j, \tilde{U}_{j'})$, recall that $\tilde{U}_j$ and $\tilde{U}_{j'}$ can be written as

$$\tilde{U}_j = \bar{Z} + \begin{pmatrix} (v_{j1} - \bar{w}_1) & (v_{j2} - \bar{w}_2) & \ldots & (v_{jp} - \bar{w}_p) \end{pmatrix} \underset{\sim}{\hat{\beta}}^* \tag{A.8}$$

$$\tilde{U}_{j'} = \bar{Z} + \begin{pmatrix} (v_{j'1} - \bar{w}_1) & (v_{j'2} - \bar{w}_2) & \ldots & (v_{j'p} - \bar{w}_p) \end{pmatrix} \underset{\sim}{\hat{\beta}}^* \tag{A.9}$$

so that the covariance of $\tilde{U}_j$ and $\tilde{U}_{j'}$ can be computed as

$$\begin{aligned} Cov(\tilde{U}_j, \tilde{U}_{j'}) &= Cov(\bar{Z} + \begin{pmatrix} (v_{j1} - \bar{w}_1) & \dots & (v_{jp} - \bar{w}_p) \end{pmatrix} \underset{\sim}{\hat{\beta}}^*, \\ &\quad \bar{Z} + \begin{pmatrix} (v_{j'1} - \bar{w}_1) & \dots & (v_{j'p} - \bar{w}_p) \end{pmatrix} \underset{\sim}{\hat{\beta}}^*) \\ &= Cov(\bar{Z}, \bar{Z}) + Cov(\bar{Z}, \begin{pmatrix} (v_{j1} - \bar{w}_1) & \dots & (v_{jp} - \bar{w}_p) \end{pmatrix} \underset{\sim}{\hat{\beta}}^*) + \\ &\quad Cov(\bar{Z}, \begin{pmatrix} (v_{j'1} - \bar{w}_1) & \dots & (v_{j'p} - \bar{w}_p) \end{pmatrix} \underset{\sim}{\hat{\beta}}^*) + \\ &\quad Cov(\begin{pmatrix} (v_{j1} - \bar{w}_1) & \dots & (v_{jp} - \bar{w}_p) \end{pmatrix} \underset{\sim}{\hat{\beta}}^*, \begin{pmatrix} (v_{j'1} - \bar{w}_1) & \dots & (v_{j'p} - \bar{w}_p) \end{pmatrix} \underset{\sim}{\hat{\beta}}^*). \end{aligned} \tag{A.10}$$

Looking into the components of the covariance computation, it is easily established that

$$Cov(\bar{Z}, \bar{Z}) = \frac{\sigma_C^2}{m_i} \tag{A.11}$$

and

$$\begin{aligned} &Cov(\begin{pmatrix} (v_{j1} - \bar{w}_1) & \dots & (v_{jp} - \bar{w}_p) \end{pmatrix} \underset{\sim}{\hat{\beta}}^*, \begin{pmatrix} (v_{j'1} - \bar{w}_1) & \dots & (v_{j'p} - \bar{w}_p) \end{pmatrix} \underset{\sim}{\hat{\beta}}^*) \\ &\quad = \begin{pmatrix} (v_{j1} - \bar{w}_1) & \dots & (v_{jp} - \bar{w}_p) \end{pmatrix} Var(\underset{\sim}{\hat{\beta}}^*) \begin{pmatrix} (v_{j'1} - \bar{w}_1) \\ \vdots \\ (v_{j'p} - \bar{w}_p) \end{pmatrix} \\ &\quad = \begin{pmatrix} (v_{j1} - \bar{w}_1) & \dots & (v_{jp} - \bar{w}_p) \end{pmatrix} \left(\sum_{i=T,C} \Sigma_i^{-1} \right)^{-1} \begin{pmatrix} (v_{j'1} - \bar{w}_1) \\ \vdots \\ (v_{j'p} - \bar{w}_p) \end{pmatrix}. \end{aligned} \tag{A.12}$$

The two quantities can be combined and simplified as

$$Cov\left(\bar{Z},\bar{Z}\right)+Cov(\left(\begin{array}{cccc}(v_{j1}-\bar{w}_1) & \ldots & (v_{jp}-\bar{w}_p)\end{array}\right)\underset{\sim}{\hat{\beta}}^*,\left(\begin{array}{cccc}(v_{j'1}-\bar{w}_1) & \ldots & (v_{j'p}-\bar{w}_p)\end{array}\right)\underset{\sim}{\hat{\beta}}^*\Bigg)$$

$$=\frac{\sigma_C^2}{m_i}+\left(\begin{array}{cccc}(v_{j1}-\bar{w}_1) & \ldots & (v_{jp}-\bar{w}_p)\end{array}\right)\left(\sum_{i=T,C}\Sigma_i^{-1}\right)^{-1}\left(\begin{array}{c}(v_{j'1}-\bar{w}_1)\\ \vdots \\ (v_{j'p}-\bar{w}_p)\end{array}\right)$$

$$=\underset{\sim}{v}_j^{*'}\Lambda_C^*\underset{\sim}{v}_{j'}^*, \tag{A.13}$$

where $\Lambda_C^*=\begin{pmatrix}\frac{\sigma_C^2}{m_i} & 0 & 0 & \ldots & 0\\ 0 & & & & \\ \vdots & & \left(\sum_{i=T,C}\Sigma_i^{-1}\right)^{-1} & & \\ 0 & & & & \end{pmatrix}$. Furthermore, it is computed that

$$Cov(\bar{Z},\left(\begin{array}{cccc}(v_{j1}-\bar{w}_1) & \ldots & (v_{jp}-\bar{w}_p)\end{array}\right)\underset{\sim}{\hat{\beta}}^*)$$

$$=Cov\left(\frac{Z_1+\ldots+Z_{m_i}}{m_i},\left(\begin{array}{cccc}(v_{j1}-\bar{w}_1) & \ldots & (v_{jp}-\bar{w}_p)\end{array}\right)\left(\begin{array}{c}\frac{1}{\sigma_T^2}d_{T1}U+\frac{1}{\sigma_C^2}d_{C1}Z\\ \vdots \\ \frac{1}{\sigma_T^2}d_{Tp}U+\frac{1}{\sigma_C^2}d_{Cp}Z\end{array}\right)\right)$$

$$=Cov\left(\frac{Z_1+\ldots+Z_{m_i}}{m_i},\left(\begin{array}{cccc}(v_{j1}-\bar{w}_1) & \ldots & (v_{jp}-\bar{w}_p)\end{array}\right)\left(\begin{array}{c}\frac{1}{\sigma_C^2}d_{C11}Z_1+\ldots+\frac{1}{\sigma_C^2}d_{Cm_i1}Z_{m_i}\\ \vdots \\ \frac{1}{\sigma_C^2}d_{C1p}Z_1+\ldots+\frac{1}{\sigma_C^2}d_{Cm_ip}Z_{m_i}\end{array}\right)\right)$$

$$= \frac{1}{m_i} Cov(Z_1 + \ldots + Z_{m_i}, (v_{j1} - \bar{w}_1) \left[\frac{1}{\sigma_C^2} d_{C11} Z_1 + \ldots + \frac{1}{\sigma_C^2} d_{Cm_i1} Z_{m_i} \right] +$$

$$+ \ldots + (v_{jp} - \bar{w}_p) \left[\frac{1}{\sigma_C^2} d_{C1p} Z_1 + \frac{1}{\sigma_C^2} d_{C2p} Z_2 + \ldots + \frac{1}{\sigma_C^2} d_{Cm_ip} Z_{m_i} \right])$$

$$= \frac{1}{m_i} [Cov(Z_1, (v_{j1} - \bar{w}_1) \frac{1}{\sigma_C^2} d_{C11} Z_1) + \ldots + Cov(Z_1, (v_{jp} - \bar{w}_p) \frac{1}{\sigma_C^2} d_{C1p} Z_1) +$$

$$\ldots + Cov(Z_{m_i}, (v_{j1} - \bar{w}_1) \frac{1}{\sigma_C^2} d_{Cm_i1} Z_{m_i}) + \ldots + Cov(Z_{m_i}, (v_{jp} - \bar{w}_p) \frac{1}{\sigma_C^2} d_{Cm_ip} Z_{m_i})]$$

$$= \frac{1}{m_i} [(v_{j1} - \bar{w}_1) d_{C11} + \ldots + (v_{jp} - \bar{w}_p) d_{C1p} + \ldots + (v_{jp} - \bar{w}_p) d_{Cm_ip}]$$

$$= \frac{1}{m_i} \sum_{q=1}^{m_i} \underset{\sim}{v}_j' \underset{\sim}{d}^*_{Cq} \tag{A.14}$$

Under the same formulations as above, the

$$Cov(\bar{Z}, \begin{pmatrix} (v_{j'1} - \bar{w}_1) & (v_{j'2} - \bar{w}_2) & \ldots & (v_{j'p} - \bar{w}_p) \end{pmatrix} \underset{\sim}{\hat{\beta}}^*) = \frac{1}{m_i} \sum_{q=1}^{m_i} \underset{\sim}{v}_{j'}' \underset{\sim}{d}^*_{Cq}. \tag{A.15}$$

These quantities result to the calculation of $Cov(\tilde{U}_j, \tilde{U}_{j'})$ as

$$Cov(\tilde{U}_j, \tilde{U}_{j'}) = \underset{\sim}{v}_j^{*'} \Lambda_c \underset{\sim}{v}^*_{j'} + \frac{1}{m_i} \sum_{q=1}^{m_i} \underset{\sim}{v}_j' \underset{\sim}{d}^*_{Cq} + \frac{1}{m_i} \sum_{q=1}^{m_i} \underset{\sim}{v}_{j'}' \underset{\sim}{d}^*_{Cq}. \tag{A.16}$$

Hence, the covariance of the effect sizes δ_j and $\delta_{j'}$ can be derived as

$$Cov(\delta_j, \delta_{j'}) = Cov(U_j - \tilde{U}_j, U_{j'} - \tilde{U}_{j'})$$

$$= Cov(U_j, U_{j'}) - Cov(U_j, \tilde{U}_{j'}) - Cov(U_{j'}, \tilde{U}_j) + Cov(\tilde{U}_j, \tilde{U}_{j'})$$

$$= -\underset{\sim}{v}_{j'}' \underset{\sim}{d}^*_{Tj} - \underset{\sim}{v}_j' \underset{\sim}{d}^*_{Tj'} + \underset{\sim}{v}_j^{*'} \Lambda_c^* \underset{\sim}{v}^*_{j'} + \frac{1}{m_i} \sum_{q=1}^{m_i} \underset{\sim}{v}_j' \underset{\sim}{d}^*_{Cq} + \frac{1}{m_i} \sum_{q=1}^{m_i} \underset{\sim}{v}_{j'}' \underset{\sim}{d}^*_{Cq} \tag{A.17}$$

As a summary, it is established that

$$\underset{\sim}{\delta} = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_{n_i} \end{pmatrix} \sim N_{n_i} \left(\alpha \underset{\sim}{1}, \Psi \right)$$

with

$$\alpha = \beta_{0,T} - \beta_{0,C}$$

$$Var(\delta_j) = \sigma_T^2 + \underset{\sim}{v}_j^{*'} \Lambda_c \underset{\sim}{v}_j^* - 2\underset{\sim}{v}_j^{'} \underset{\sim}{d}_{Tj}^*$$

$$Cov(\delta_j, \delta_{j'}) = -\underset{\sim}{v}_{j'}^{'} \underset{\sim}{d}_{Tj}^* - \underset{\sim}{v}_j^{'} \underset{\sim}{d}_{Tj'}^* + \underset{\sim}{v}_j^{*'} \Lambda_c^* \underset{\sim}{v}_{j'}^* + \frac{1}{m_i} \sum_{q=1}^{m_i} \underset{\sim}{v}_j^{'} d_{Cq}^* + \frac{1}{m_i} \sum_{q=1}^{m_i} \underset{\sim}{v}_{j'}^{'} d_{Cq}^*.$$

Bibliography

[1] P.C. Austin,"An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate Behavioral Research* **46**(3) (2011): 399-424.

[2] P.C. Austin, "The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments," *Statistics in Medicine* **33**(7)(2014): 1242-1258.

[3] P.C. Austin, "The performance of different propensity score methods for estimating relative risks," *Journal of Clinical Epidemiology* **61**(6) (2008): 537-545.

[4] P.C. Austin and T. Schuster, "The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: A simulation study," *Statistical Methods in Medical Research* **25**(5) (2016): 2214-2237.

[5] S.O. Becker and A. Ichino, "Estimation of average treatment effects based on propensity scores," *The Stata Journal* **2**(4) (2002): 1-19.

[6] R. DerSimonian and N.M. Laird,"Evaluating the effect of coaching on SAT scores: a meta-analysis," *Harvard Educational Review* **53**(1) (1983): 1-15.

[7] A. Gelman and X.L. Meng, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (John Wiley & Sons, Inc., Hoboken, New Jersey, 2004), Chap. 4.

[8] E.M. Hade and B. Lu, "Bias associated with using the estimated propensity score as a regression covariate," *Statistics in Medicine* **33**(1) (2014): 74-87.

[9] J. Hartung, G. Knapp and B.K. Sinha, *Statistical Meta-Analysis with Applications* (John Wiley & Sons, Inc., Hoboken, New Jersey, 2008), Chap. 4.

[10] J. Hartung, G. Knapp and B.K. Sinha, *Statistical Meta-Analysis with Applications* (John Wiley & Sons, Inc., Hoboken, New Jersey, 2008), Chap. 7.

[11] P.G. Hoel, *"Confidence regions for linear regression,"* Proceedings of the second Berkeley Symposium University of California Press, 1951.

[12] G.W. Imbens, "Nonparametric estimation of average treatment effects under exogeneity: a review," *Review of Economics and Statistics* **86**(1)(2004):4-29.

[13] S. Johansen, "The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression," *Biometrika* **67**(1) (1980): 85-92.

[14] K. Krishnamoorthy and J. Yu, "Modified Nel and Van der Merwe test for the multivariate Behrens-Fisher problem," *Statistica and Probability Letters* **66**(2) (2004): 161-169.

[15] R. Lalonde, "Causal effects in non-experimental studies: reevaluating the evaluation of training programs," *Journal of the American Statistical Association* **94**(448) (1999): 1053-1062.

[16] W. Liu, *Simultaneous inference in regression* (CRC Press Taylor & Francis Group, Boca Raton, Florida, 2011), Chap. 3.

[17] W. Liu, *Simultaneous inference in regression* (CRC Press Taylor & Francis Group, Boca Raton, Florida, 2011), Chap. 5.

[18] W. Liu, M. Jamshidian, Y. Zhang and J. Donnelly, "Simulation-based simultaneous confidence bands in multiple linear regression with predictor variables constrained in intervals," *Journal of Computational and Graphical Statistics* **14**(2) (2005): 459-484.

[19] J. Mandel and R.C. Paule, "Interlaboratory evaluation of a material with unequal number of replicates," *Analytical Chemistry* **42**(11) (1970): 1194-1197.

[20] Q. Mcnemar, "Note on the sampling error of the difference between correlated proportions or percentages." *Psychometrika* **12**(2) (1947): 153-157.

[21] D.G. Nel and C.A. Van der Merwe, "A solution to the multivariate Behrens-Fisher problem," *Communications in Statistics:Theory and Methods* **15**(12) (1986): 3719-3735.

[22] S.M. Perkins, W. Tu, M.G. Underhill, X.H. Zhou and M.D. Murray, "The use of propensity scores in pharmacoepidemiologic research," *Pharmacoepidemiology and Drug Safety* **9**(2) (2000): 93-101.

[23] P.S.R.S. Rao, J. Kaplan and W.G. Cochran, "Estimators for the one-way random effects model with unequal error variances," *Journal of the American Statistical Association* **76**(373) (1981): 89-97.

[24] P.R. Rosenbaum and D.B. Rubin, "Central role of the propensity score in observational studies for causal effects," *Biometrika* **70**(1) (1983): 41-55.

[25] P.R. Rosenbaum and D.B. Rubin, "Reducing bias in observational studies using subclassification on the propensity score," *Journal of American Statistical Association* **79**(387) (1984): 516-524.

[26] P.R. Rosenbaum and D.B. Rubin, "Constructing a control group using multivariate matched sampling methods that incorporate the propensity scores," *The American Statistician* **39**(1) (1985): 33-38.

[27] P.R. Rosenbaum, "Model-based direct adjustment," *Journal of the American Statistical Association* **82**(398)(1987): 387-394.

[28] D.B. Rubin and N. Thomas, "Matching using estimated propensity scores: relating theory to practice," *Biometrics* **52**(1) (1996): 249-264.

[29] J.L. Schafer and J. Kang, "Average causal effects from non-randomized studies: a practical guide and simulated example," *Psychological Methods* **13**(4)(2008): 279-313.

[30] P.W. Stewart, "The graphical advantages of finite interval confidence band procedures," *Communications in Statistics - Theory and Methods* **20**(12) (1991): 3975-3993.

[31] H. Scheffe, "A method for judging all constraints in analysis of variance," *Biometrika* **40** (1953): 87-104.

[32] H. Scheffe, *The Analysis of Variance* (Wiley, 1959)

[33] K. Sidik and J.N. Jonkman, "A simple confidence interval for meta-analysis," *Statistics in Medicine* **21** (2002): 3153-3159.

[34] J. Park and B.K. Sinha, "Some aspects of multivariate Behrens-Fisher problem," *Calcutta Statistical Association Bulletin, 61 (Special 6th Triennial Proceedings Volume)* **61** (2009): 241-244.

[35] R.A. Stone, D.S. Obrosky, D.E. Singer, W.N. Kapoor, M.J. Fine and The Pneumonia Patient Outcomes Research Team (PORT) Investigators, "Propensity score adjustment for pre-treatment differences between hospitalized and ambulatory patients with community-acquired pneumonia," *Medical Care* **33**(4 Suppl) (1995): AS56-66 .

[36] E.J. Williamson, A. Forbes, I.R. White, "Variance reduction in randomised trials by inverse probability weighting using the propensity score," *Statistics in Medicine* **33**(5) (2014): 721-737.

[37] Y. Yao, "An approximate degrees of freedom solution to the multivariate Behrens-Fisher problem," *Biometrika* **52**(1-2) (1965): 139-148.