2018

Penalized b-splines and their application with an in depth look at the bivariate tensor product penalized b-spline

Michael Price
Iowa State University

Penalized b-splines and their application with an in depth look at the bivariate tensor product penalized b-spline

by

Michael Joseph Price

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Cindy L. Yu, Major Professor
Arka Ghosh
David Hennessy
Ken Koehler
Huaiqing Wu

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGEMENTS

Many people have helped me while I chased the dream of graduating with a Ph'D and while some will not be mentioned here, their impact on me will not be forgotten.

For the people who helped me with emotional support, I begin by thanking my parents Joe and Maryjo who have always pushed me to do my best at anything I try. Also my brother Jonathan, as well as, my aunt Dorothy and Grandma who provided much needed encouragement and phone conversations that provided me with needed breaks from research. My fiancée Sarah Lynn Long for understanding the time requirement of this pursuit and sticking with me throughout, in addition to the unconditional encouragement by both her and her family. Finally, the Martens and Baker families who gave me a much needed break from work by letting me into their homes to enjoy many home cooked meals and to watch their children grow up.

From the academic side, my advisor Dr. Cindy Yu was invaluable for her insight and support throughout these challenging problems and without her this would not have been possible. I would also like to thank my committee members: Dr. Arka Ghosh, Dr. David Hennessy, Dr. Ken Koehler, and Dr. Haiquing Wu for their questions, comments, and suggestions. In addition, I want to thank Dr. David Hennessy and Dr. Xiaodong Du for technical and related assistance needed to make the crop insurance section possible. I also want to thank Gabriel Demuth for helping me understand complex ideas and allowing me to bounce multiple ideas off of him.

Finally I want to thank the Iowa State CSSM department for funding this endeavor of mine and providing me years of work so I could understand what it means to be a statistician.

ABSTRACT

Penalized B-splines, or P-splines, are a semiparametric method that can be used to estimate models with one or two variables and have become quite popular since they first appeared in Eilers and Marx (1996). In this dissertation, two interesting problems are investigated in the areas of crop insurance and observational studies with complex surveys using univariate and bivariate P-spline methods. Premium rates of yield insurance given by the US Department of Agriculture's Risk Management Agency are investigated to see if they are actuarially fair by comparing an estimated conditional yield density using premium data with the conditional yield density estimated using yields. A procedure is developed to estimate the conditional yield density using premium data through estimating partial derivatives of the premium rate function based on the penalized bivariate tensor product B-splines (BTPB). Xiao et al. (2012) is extended to study the asymptotic properties of partial derivatives of a penalized BTPB estimator and provide a variance estimator. The validity of the conditional yield density estimator using premium data and the variance estimation is demonstrated through simulation studies. The procedure is also applied to a crop insurance data set from Iowa to examine the actuarial fairness of the premium rates. On average, premium rates are close to our estimates and this is true for each coverage level. However, premiums for low productivity land are generally too low while those for high productivity land are generally too high. Even after subsidies, premiums for the more productive land are generally substantially higher than what they should be. A generalized method moments (GMM) estimator is considered to estimate treatment effects defined through estimation equations using an observational data set from a complex survey. It is demonstrated that the proposed estimator, which incorporates both sampling probabilities and semiparametrically estimated self-selection probabilities, gives consistent estimates of

treatment effects. The asymptotic normality of the proposed estimator is established in the finite population framework, and its variance estimation is discussed. In simulations, our proposed estimator and its variance estimator based on the asymptotic distribution are evaluated. This method is then used to estimate the effects of different choices of health insurance types on health care spending using data from the Chinese General Social Survey. The results from the simulations and the empirical study show that ignoring the sampling design weights might lead to misleading conclusions.

CHAPTER 1. INTRODUCTION

Regression analysis is the statistical technique used to predict a continuous dependent variable from a number of independent variables. While this seems straight forward, there are many different types of regression models that could be used. Theses types of regression models can be categorized into three main groupings, parametric regression, nonparametric regression, and semiparametric regression.

Linear models, generalized linear models, and nonlinear models are examples of parametric regression models, because the function that describes the relationship between the response and explanatory variables is known. However, they are often not flexible enough for describing the data at hand. With the development of computer technology and statistical software, nonparametric regression, has received more attention and recognition. Nonparametric regression differs from parametric regression in that the shape of the functional relationships between the response and the explanatory variables are not predetermined and can be adjusted to capture unusual or unexpected features of the data. The mean of a response is modeled as a smooth, but otherwise unspecified function of covariates. However, these nonparametric models tend to be sensitive to outliers and have low efficiency.

Semiparametric regression methods combine parametric and nonparametric models. They are used in situations where the fully nonparametric model may not perform well because of sensitivity to outliers or when a parametric model is known but the distribution of the errors is not known. This set up can be of substantial value when working with complex scientific problems. Semiparametric regression models reduce complex data sets to understandable summaries which when properly applied, aid in sound decision-making. They are able to retain important features of the data while discarding unimportant details.

Since semiparametric models contain a parametric component, they rely on parametric assumptions and may be misspecified and inconsistent, just like a fully parametric model.

One type of semiparametric regression model is the penalized B-spline, or P-spline. P-splines have three properties that make them popular semiparametric models.

i) The basis and the penalty are sparse allowing for efficient computation.

ii) The order of the B-spline basis and penalty are independent of each other making it possible to fine tune the mean structure without having to worry about overfitting the model.

iii) P-splines use a regression model with clearly defined coefficients making it possible to compute informative properties of the model through regression theory.

Additional development and discussion of P-splines will take place in Chapter 2.

In this dissertation, two interesting problems are investigated in the areas of crop insurance and observational studies with complex surveys using univariate and bivariate P-spline methods. Chapter 3 investigates if the premium rates of yield insurance given by the US Department of Agriculture's Risk Management Agency (RMA) are actuarially fair by comparing an estimated conditional yield density using premium data with the conditional yield density estimated using yields. A procedure to estimate the conditional yield density using premium data through estimating partial derivatives of the premium rate function based on the penalized bivariate tensor product B-splines (BTPB) is developed. Then the asymptotic distribution of general partial derivatives estimators of the penalized BTPB is derived in order to establish a variance estimator. The performance of this conditional yield density estimator and its variance estimator is verified using premium data an through simulation studies. This procedure is then applied to a crop insurance data set from Webster County, Iowa to examine the actuarial fairness of the premium rates.

Chapter 4 considers a generalized method moments (GMM) estimator to estimate treatment effects defined through estimation equations using an observational data set from a

complex survey. The proposed estimator incorporates both sampling probabilities from survey weights and semiparametrically estimated self-selection probabilities. The semiparametric method used is a univariate P-spline. The asymptotic normality of the proposed estimator is established in the finite population framework, and its variance estimation is discussed. In simulations, it is then shown that the proposed estimator and its variance estimator based on the asymptotic distribution gives consistent estimates of treatment effects. The method is then used to estimate the effects of different choices of health insurance types on health care spending using data from the Chinese General Social Survey. This analysis shows that ignoring the sampling design weights might lead to misleading conclusions.

The research in Chapter 3 has contributed to both theoretic and practical areas of study. In particular, this work has helped the development of a central limit theorem and a variance estimator for any partial derivative of penalized bivariate tensor product B-splines. It has also helped shed light into how well government subsidies are performing, in one county at least, and leads to a framework that could be used to evaluate subsidies for all agricultural counties in the United States. The research in Chapter 4 shows that the use of semiparametric methods in observational studies for complex survey statistics provides a robust way to handle misspecification of selection probabilities. Through the use of simulation studies, the danger of ignoring survey weights in observational data when using inverse probability weighted estimators is observed and show how biased results can occur under certain conditions.

This dissertation is organized as follows. Chapter 2 gives a history of P-splines. Chapter 3 discusses penalized BTPB and its application to crop insurance. The use of semiparametric methods in observational studies for complex survey statistics is covered in Chapter 4. Chapter 5 summarizes the results in this dissertation. Appendix A contains in depth proofs for Chapter 3. In depth proofs for Chapter 4 are provided in Appendix B.

CHAPTER 2. A HISTORY OF P-SPLINES

2.1 Introduction

Penalized B-splines are a semiparametric method that can be used to estimate models with one or two variables and have become quite popular since they first appeared in Eilers and Marx (1996). Because it is based on regression, the combination of a B-spline basis and a simple difference penalty lends itself well to a variety of generalizations (Eilers et al., 2015). P-splines allow for the use of a variety of additive smooth structure components, as well as allowing multiple extensions. These include penalized Bivariate Tensor Product B-splines (BTPB) which are used for modeling two-dimensional surfaces.

A look at the history of P-splines from their initial proposal up through the asymptotic theory of penalized BTPB is an informative and interesting way of understanding how these splines evolved over time and what development lies ahead. The rest of this chapter is organized as follows. In Section 2.2 we discuss the development of the two types of splines that have helped lead to the development of P-splines, Bézier curves and B-splines. The addition of the penalty term to the B-spline in order to create the P-spline is reviewed in section 2.3. Section 2.4 covers the bivariate framework. Section 2.5 discusses P-splines extensions and applications that are potential areas of research.

2.2 Precursors to P-splines

Before computers, designs were drawn by hand on paper with various drafting tools, such as rulers, compasses, and protractors. But many shapes, such as that of a ship's bow, could not be drawn with these tools. Such shapes often needed to be drawn life-size and could not be drawn free hand. Such life size drawings were done with the help of flexible strips of wood, called splines. The splines were held in place at a number of predetermined

points, called ducks; between the ducks, the elasticity of the spline material caused the strip to take the shape that minimized the energy of bending, thus creating the smoothest possible shape that fit the constraints. The shape could be tweaked by moving the ducks (Schneider, 1996).

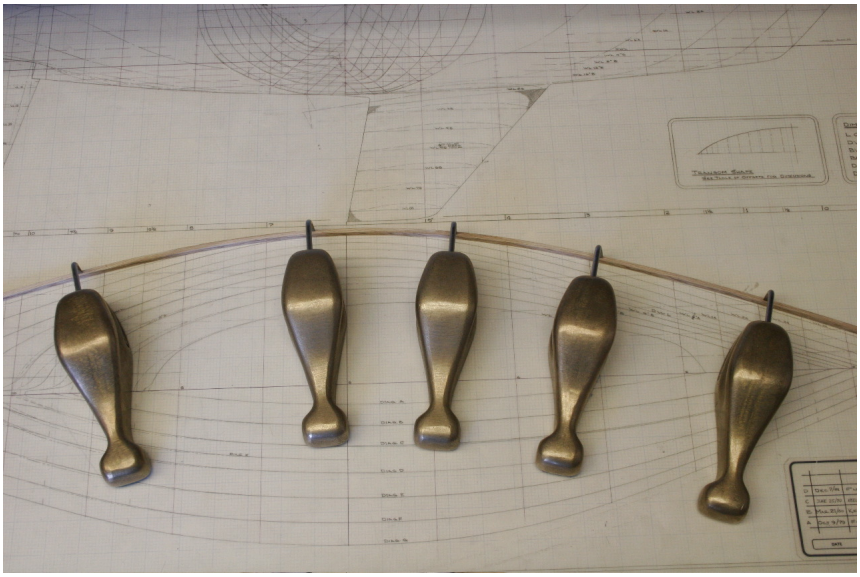

Figure 2.1 A duck and spline set.

westlawn5554X (2006)

Spline techniques in mathematics and software borrow from these ideas to form desired shapes thus giving them the name spline. A spline is a smooth polynomial function that is piecewise defined and possesses a high degree of smoothness at the places where the polynomial pieces connect. These places are known as knots instead of ducks (Schneider, 1996).

P-splines are descendants of Bézier curves and Basis splines, or B-splines. These three types of splines have three major advantages (Shene, 2014):

1. A predetermined set of control points where the spline curve generally follows the trend of these control points.

2. Geometric and numerically stable algorithms for finding points on the curve without knowing the equation of the curve.

3. Easy transitions from one dimensional curves to two dimensional surfaces since the methodology for curves applies directly to surfaces.

Bézier curves were discovered simultaneously by de Casteljau (1959) and Bézier (1968) and will be briefly discussed in section 2.2.1. Basis splines, or B-splines, were studied by Lobachevsky (1902), but a modern version, which made calculation more efficient for computers, was developed by De Boor (1976) and are discussed in section 2.2.2

2.2.1 Bézier curves

Bézier curves are special cases of B-splines. They are parametric approximations that use the Bernstein polynomials as a basis. A Bézier curve of degree d is defined as

$$r(x) = \sum_{\nu=0}^{d} \mathbf{b}_\nu b_{\nu,d}(x), 0 \leq x \leq 1.$$

where $\mathbf{b}_\nu$ are the control points and $b_{\nu,d}(x)$ is the $d+1$ Bernstein basis polynomials of degree d defined as (Patrikalakis et al., 2009)

$$b_{\nu,d}(x) = \binom{d}{\nu} x^\nu (1-x)^{d-\nu}, \quad \nu = 0, \ldots, d.$$

Lines drawn between consecutive control points of the curve form the control polygon. Unlike B-splines which pass through all control points, or knots, Bézier curves only pass through the first and last control points and are the endpoints of the fitted curve. Some nice properties Bézier curves have are as follows (Patrikalakis et al., 2009).

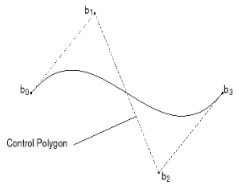

Figure 2.2 Example of a control polygon.

- Geometry invariance property: Bézier curves are invariant under translation and rotation of its control points.
- Convex hull property: For any two points in the Bézier curve, the line segment connecting these two points is contained within the domain.

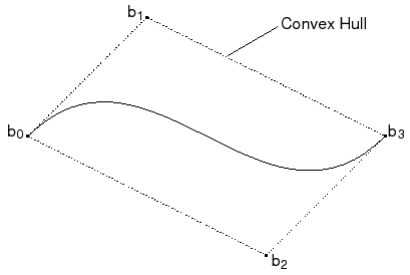

Figure 2.3 Example of a Convex Hall.

- Variation diminishing property: The number of intersection of a straight line with the Bézier curves is no greater than the number of intersection of the line with the control polynomial.

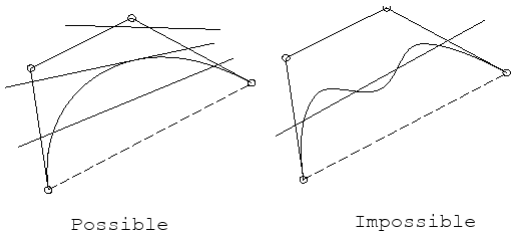

Figure 2.4 Example of the variance diminishing property of Bézier curves.

- Symmetry property: $\sum_{\nu=0}^{d} \mathbf{b}_\nu b_{\nu,d}(x) = \sum_{\nu=0}^{d} \mathbf{b}^*_{\mathbf{d}-\nu} b_{\nu,d}(1-x)$ where $\mathbf{b}_\nu = \mathbf{b}^*_{\mathbf{d}-\nu}$.

There are two major problems with Bézier curves. The first is that Bézier curves are C^1 smooth functions, so they can only be differentiated once. The second major problem with Bézier curves is the lack of local control over the curve. This means that getting the correct shape of the curve is difficult because changing a single control point in a Bézier curve will affect the entire curve (Shene, 2014).

2.2.2 B-splines

B-spline curves require more information, such as the degree of the curve and a knot vector, as well as more complex theory than Bézier curves. However, the advantages gained tend to be worth the additional information required. The advantages B-splines have over Bézier curves are (Shene, 2014):

- a B-spline curve can be a Bézier curve.
- B-spline curves satisfy all important properties that Bézier curves have.
- B-spline curves provide more control flexibility than Bézier curves can do. So the degree of a B-spline curve is not dependent on the number of control points.
- B-splines can have lower degree curves and still maintain a large number of control points.
- Control points can be moved without changing the shape of the whole curve, this means that the B-spline approximation at a point x only depends on data values near x.
- Each B-spline basis function is non-zero on only a few adjacent subintervals resulting in the B-spline functions being a local approximation method.

A B-spline approximates functions using piecewise polynomials. The Weierstrass Theorem (Weierstrass, 1885) states that there is always a polynomial arbitrarily close to any

continuous function. However the degree of this polynomial could be very large, so in order to lower the degree of the polynomial, multiple polynomials are used with each polynomial estimating only a small segment of the function to be approximated. These polynomial are connected at the knots in such a way that they achieve certain smoothness criteria, specifically so that the approximation function has C^d smoothness where d is the degree of the piecewise polynomials. This implies that the function is d-times differentiable.

Let K be the number of knots within a closed interval and d be the degree of the B-spline. Define κ_k as the location of the k-th knot, $k = (-d, \ldots, K + d + 1)$. The B-spline basis is defined recursively as:

For $s = 0$:

$$B_{k,0}(x) = \begin{cases} 1 & \text{if } \kappa_{k-1} \leq x < \kappa_k \\ 0 & otherwise \end{cases}$$

For $s = 1, 2, \cdots, d$:

$$B_{k,s}(x) = \frac{x - \kappa_{k-1}}{\kappa_{k+s-1} - \kappa_{k-1}} B_{k,s-1}(x) + \frac{\kappa_{k+s} - x}{\kappa_{k+s} - \kappa_k} B_{k+1,s-1}(x). \tag{2.1}$$

Figure 2.5 provides an illustration of a single B-spline of degree 1. A single B-spline of one degree consists of two linear pieces, centered at κ_k and spans three total knots. Figure 2.6 provides an illustration of a single B-spline of degree 3. The single B-spline of three degrees consists of four cubic pieces, centered at κ_{k+1} and spans five knots. At the three locations where the cubic pieces meet, not only are the first derivatives of the two pieces equal, but the second derivatives are as well, meaning that the base has C^3 smoothness. Figure 2.7 provides an illustration of several B-splines of degree three within a knot sequence. Some general properties of a B-spline of degree d are:

- consists of $d + 1$ polynomial pieces of degree d.
- the polynomial pieces join at d inner knots.
- derivatives up to order $d - 1$ are continuous.

- are positive on a domain spanned by $d + 1$ knots.

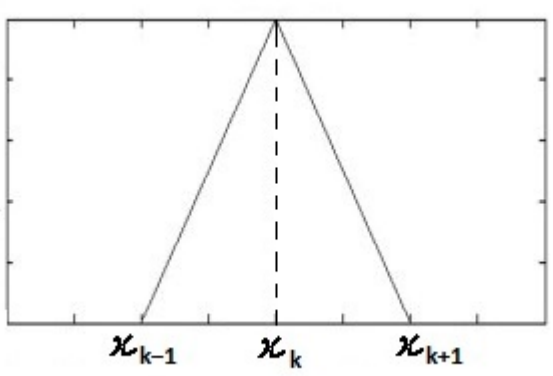

Figure 2.5 A single B-spline of degree 1.

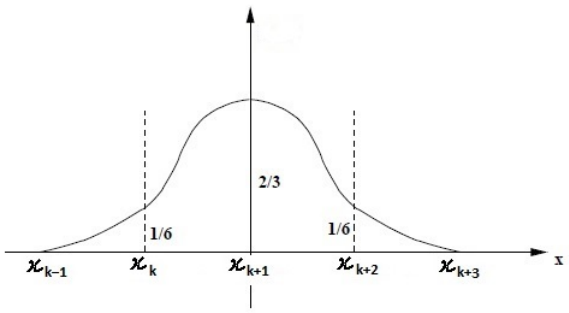

Figure 2.6 A single B-spline of degree 3.

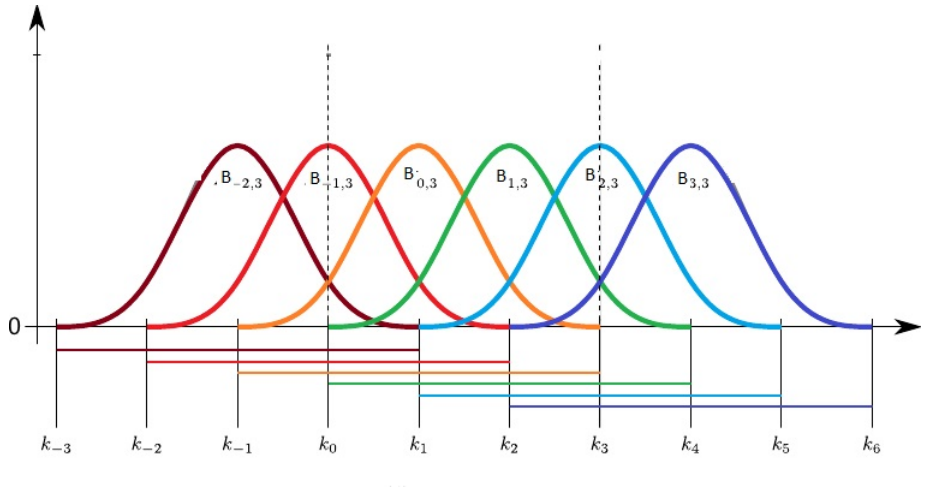

Figure 2.7 B splines of degree 3 along a knot sequence.

The final form of the curve created by a B-spline of degree d is given by

$$C(x) = \sum_{i=1}^{I} a_i B_{i,d}(x).$$

where $I = K + d + 1$ is the total number of B-spline basis being used. The control points of the B-spline curve are represented by a_i.

When interpolating data, the curve is estimated using least squares to find the optimum values of the control points. Using y_j as the observed data, the objective function to be minimized is:

$$\hat{a}_i = \underset{a_i}{\operatorname{argmin}} \sum_{j=1}^{n} \left\{ y_j - \sum_{i=1}^{I} a_i B_{i,d}(x_j) \right\}^2,$$

which results in the fitted B-spline curve to be given as:

$$\hat{c}(x) = \sum_{i=1}^{I} \hat{a}_i B_{i,d}(x).$$

Due to the structure of B-splines, the derivative of a d degree B-spline is a B-spline of degree $d-1$. Besides the change in degree, the control point values, a_i, also change. De Boor

(1978) developed a formula for the derivatives of B-splines given by:

$$\begin{aligned} C'(x) =& \Bigg(\sum_{i=1}^{I} a_i B_{i,d-1}(x) - \sum_{i=1}^{I} a_{i+1} B_{i+1,d-1}(x) \Bigg) / h \\ =& - \Bigg(\sum_{i=1}^{I} \Delta a_{i+1} B_{i,d-1}(x) \Bigg) / h, \end{aligned}$$

where $\Delta a_i = a_i - a_{i-1}$ and h is the distance between knots.

This means the second derivative is given by:

$$C''(x) = -\Bigg(\sum_{i=1}^{I} \Delta^2 a_i B_{i,d-2}(x) \Bigg) / h^2,$$

where

$$\Delta^2 a_i = \Delta \Delta a_i = a_i - 2a_{i-1} + a_{i-2}.$$

2.3 P-splines

While B-splines are a good option for estimation in goodness of fit, ease of computation, and having at least C^3 smoothness, choosing the optimal number and position of knots is a complex undertaking. Too many knots may result in overfitting the data, and modeling of noise instead of the signal. Not enough knots could lead to important features in the data being left out. Eilers and Marx (1996) proposed using a difference penalty on coefficients of adjacent B-splines and a large number of knots in place of determining the optimal number and location of knots when using a B-spline. This methodology, called a penalized B-spline, or P-spline, eliminates unneeded knots by turning their control points to near zero, and leads to a smoother fit.

If the number of interior knots, k, was to be increased to a relatively large number, the fitted curve $C(x)$ would have more variation than is reasonable given the data. In order to limit this variation, O'Sullivan (1986, 1988) created a penalty on the second derivative of the fitted curve. Using O'Sullivan (1986, 1988)'s method, the control points, a_i, are found by minimizing the objective function:

$$\hat{a}_i = \underset{a_i}{\operatorname{argmin}} \sum_{j=1}^{n} \Bigg\{ y_j - \sum_{i=1}^{I} a_i B_{i,d}(x_j) \Bigg\}^2 + \lambda \int_{x_{min}}^{x_{max}} \Bigg\{ \sum_{i=1}^{I} a_i B''_{i,d}(x) \Bigg\}^2 dx.$$

where λ controls the smoothness of fit and y_j is the observed data. When $\lambda = 0$, the P-spline fit is the same as the B-spline fit. As λ gets large, the fit is similar to that of a polynomial of degree $d - 1$.

Eilers and Marx (1996) modified this set up by having the penalty based on finite differences of the coefficients of adjacent B-splines, reducing the dimensionality of the problem from n, the number of observations, to I, the number of B-splines. In addition, O'Sullivan (1986, 1988)'s penalty is discrete but derived from the integrated squared second derivative of the fitted curve, whereas Eilers and Marx (1996)'s use a discrete penalty making it trivial to use differences of any order. The control points, a_i, of the fitted curve using Eilers and Marx (1996)'s method are found by minimizing the objective function:

$$\hat{a}_i = \underset{a_i}{\operatorname{argmin}} \sum_{j=1}^{n} \left\{ y_j - \sum_{i=1}^{I} a_i B_{i,d}(x_j) \right\}^2 + \lambda \sum_{i=m+1}^{I} \left(\Delta^m a_i \right)^2$$

where m is the order of the penalty, Δ is the difference operator, and y_j is our observed data.

2.4 Penalized Bivariate Tensor Product B-Splines

There are many problems that require the use of multiple covariates. For example modeling short-wave spectra using wavelength and temperature or modeling crop insurance data using land quality and crop coverage rate. For bivariate spline smoothing, there are two well known estimators: bivariate P-splines (Eilers and Marx, 2003; Marx and Eilers, 2005) and thin plate splines, e.g., the thin plate regression splines (Wood, 2003). More recently, a bivariate spline method that uses tensor products to combine P-splines has been developed. Using the tensor product P-spline provides the following three benefits:

- Allows for interactions between the two covariates.
- Allows fast computation, by using generalized cross validation (GCV) criterion for selecting smoothing parameters.

- Allows for the derivation of a central limit theorem.

One of the more recent developments in tensor product P-splines is the derivation of the central limit theorem development by Xiao et al. (2010). An asymptotic study of univariate penalized splines was achieved only recently (Hall and Opsomer (2005); Li and Ruppert (2008); Claeskens et al. (2009); Kauermann et al. (2009); Wang et al. (2009)). The theoretical study of penalized splines in higher dimension is more challenging, with Xiao et al. (2010) being the first to develop central limit theorems and expressions for the asymptotic mean and covariance matrix of bivariate spline estimators. Xiao et al. (2010) was able to do this by reorganizing the tensor product structure. This simplifies asymptotic analysis and helps prove that the tensor product structure is asymptotically equivalent to a kernel estimator with a product kernel. Xiao et al. (2010)'s sandwich smoother will now be shown as well as its equivalence to the tensor product structure of splines.

A general set up of data is given as follows. First assume that there is a function $\mu(x, z)$ with $(x, z) \in [0, 1]^2$ and that $y_{i,j} = \mu(x_i, z_j) + \epsilon_{i,j}, 1 \leq i \leq n_1, 1 \leq j \leq n_2$. Also assume that the points $(x_i, z_i)_{1 \leq i \leq n_1, 1 \leq j \leq n_2}$ are deterministic on a rectangular grid and that $\epsilon_{i,j}$ are random variables with mean 0 and variance σ^2. Data not following this assumption can be adjusted using binning techniques discussed in Chapter 3 that require little change to the theory. The tensor product spline of two variables is defined as:

$$\sum_{\substack{1 \leq k \leq I_1 \\ 1 \leq l \leq I_2}} a_{k,l} B^1_{k,d_1}(x) B^2_{l,d_2}(z). \tag{2.2}$$

With the data in the rectangular grid, they can be arranged into a $n_1 \times n_2$ matrix $\mathbf{Y}$. The proposed fitted values will be represented by $\hat{\mathbf{Y}}$. Let $\mathbf{A}= (a_{k,l})_{1 \leq k \leq I_1 1 \leq l \leq I_2}$ be the coefficient matrix and B^1_k and B^2_l be the B-splines of degrees d_1 and d_2 with k_1 and k_2 knots for x and z respectively. Note that this means $I_i = d_i + k_i + 1$ for $i = 1, 2$. This implies the model:

$$\hat{\mathbf{Y}} = \mathbf{B_1 A B_2^T} + \boldsymbol{\epsilon}$$

where $\mathbf{B_1} = \{B^1_{k,d_1}(x_i)\}_{1\leq k\leq I_1 1\leq i\leq n_1}$, $\mathbf{B_2} = \{B^2_{l,d_2}(z_i)\}_{1\leq l\leq I_2 1\leq i\leq n_2}$, and $\boldsymbol{\epsilon}$ is a $n_1 \times n_2$ matrix with the (i,j)th value $\epsilon_{i,j}$.

The sandwich smoother smooths along the rows and down the columns of the matrix $\mathbf{Y}$ leading to the matrix of fitted values, $\hat{\mathbf{Y}}$, defined as:

$$\hat{\mathbf{Y}} = \mathbf{S_1 Y S_2}. \tag{2.3}$$

$\mathbf{S_1}$ and $\mathbf{S_2}$ are the P-spline matrices for x and z. Even though the two P-splines are applied at the same time, this method holds one covariate fixed, while the P-spline is applied to the other covariate. This sandwich smoother is equivalent to tensor product splines with a particular penalty. Define the vec operation to be an operation that stacks the columns of a matrix into a vector. Let $\mathbf{y} = vec(\mathbf{Y})$ and $\hat{\mathbf{y}} = vec(\hat{\mathbf{Y}})$. Then apply the identity definition of the tensor product and the identity from Seber (2008) pp 240, to (2.3) to get

$$\hat{\mathbf{y}} = (\mathbf{S_2} \otimes \mathbf{S_1})\mathbf{y}.$$

This implies that the overall smoother matrix is just a tensor product of two univariate smoother matrices. S_1 and S_2 can then be defined as:

$$\mathbf{S_i} = \mathbf{B_i}(\mathbf{B_i^T B_i} + \lambda_\mathbf{i}\mathbf{D_i^T D_i})^{-\mathbf{1}}\mathbf{B_i^T}, i = 1, 2.$$

$\mathbf{B_i}$ are $n \times I$ model matrices using B-spline basis for x and z. $\mathbf{D_i}$ is a $I - m_i + 1 \times I$ difference matrix of order m_i. Where $d_{k,l} = (-1)^{l-k}\binom{m}{l-k}$ for $0 \leq l - k \leq m$ and 0 otherwise.

This allows the smoother matrix to be written as:

$$\begin{aligned}(\mathbf{S_2} \otimes \mathbf{S_1}) = (\mathbf{B_2} \otimes \mathbf{B_1})\{&\mathbf{B_2^T B_2} \otimes \mathbf{B_1^T B_1} + \lambda_1 \mathbf{B_2^T B_2} \otimes \mathbf{D_1^T D_1} \\ &+\lambda_2 \mathbf{D_2^T D_2} \otimes \mathbf{B_1^T B_1} + \lambda_1\lambda_2 \mathbf{D_2^T D_2} \otimes \mathbf{D_1^T D_1}\}^{-\mathbf{1}}(\mathbf{B_2} \otimes \mathbf{B_1})^{\mathbf{T}},\end{aligned}$$

meaning that the model uses tensor product splines with penalty:

$$\mathbf{P} = \lambda_1 \mathbf{B_2^T B_2} \otimes \mathbf{D_1^T D_1} + \lambda_2 \mathbf{D_2^T D_2} \otimes \mathbf{B_1^T B_1} + \lambda_1\lambda_2 \mathbf{D_2^T D_2} \otimes \mathbf{D_1^T D_1} \tag{2.4}$$

It can be shown that this is equivalent to the tensor product P-spline as defined in (2.2). Going back to the tensor product P-spline model, $\hat{\mathbf{Y}} = \mathbf{B_1 A B_2^T} + \boldsymbol{\epsilon}$, define $\mathbf{a} = vec(\mathbf{A})$. Then to estimate $\hat{\mathbf{a}}$, the following objective function is minimized

$$||\mathbf{Y} - \mathbf{B_1}\hat{\mathbf{A}}\mathbf{B_2^T}||_F^2 + \hat{\mathbf{a}}^\mathbf{T}\mathbf{P}\hat{\mathbf{a}}^\mathbf{T}$$

where $||\cdot||_F$ denote the Frobenius norm and $\mathbf{P}$ is defined in (2.4). $\hat{\mathbf{a}}$ satisfies the equation

$$(\mathbf{\Lambda_2} \otimes \mathbf{\Lambda_1})\hat{\mathbf{a}} = (\mathbf{B_2} \otimes \mathbf{B_1})^\mathbf{T}\mathbf{y} \tag{2.5}$$

where $\mathbf{\Lambda_i} = \mathbf{B_i^T B_i} + \lambda_\mathbf{i}\mathbf{D_i^T D_i}$ for $i = 1, 2$, or equivalently

$$\mathbf{\Lambda_1}\hat{\mathbf{A}}\mathbf{\Lambda_2} = \mathbf{B_1^T Y B_2}.$$

Which gives us the penalized estimate as

$$\hat{\mu}(x, z) = \sum_{\substack{1 \le k \le I_1 \\ 1 \le l \le I_2}} \hat{a}_{k,l} B^1_{k,d_1}(x) B^2_{l,d_2}(z).$$

Using (2.5), it follows that $\hat{\mathbf{y}} = (\mathbf{B_2} \otimes \mathbf{B_1})\hat{\mathbf{a}}$ satisfies the sandwich smoother definition given in (2.3).

$$\begin{aligned}
&\hat{\mathbf{y}} = (\mathbf{B_2} \otimes \mathbf{B_1})\hat{\mathbf{a}}, \\
&\hat{\mathbf{a}} = (\mathbf{\Lambda_2} \otimes \mathbf{\Lambda_1})^{-\mathbf{1}}(\mathbf{B_2} \otimes \mathbf{B_1})^\mathbf{T}\mathbf{y} \\
\implies &\hat{\mathbf{y}} = (\mathbf{B_2} \otimes \mathbf{B_1})(\mathbf{\Lambda_2} \otimes \mathbf{\Lambda_1})^{-\mathbf{1}}(\mathbf{B_2} \otimes \mathbf{B_1})^\mathbf{T}\mathbf{y} \\
\implies &\hat{\mathbf{Y}} = \left(\mathbf{B_1}\mathbf{\Lambda_1^{-1}}\mathbf{B_1^T}\right)\mathbf{Y}\left(\mathbf{B_2}(\mathbf{\Lambda_2^{-1}})\mathbf{B_2^T}\right) \textit{Note}: \mathbf{\Lambda_i} \text{ is symmetric} \\
\equiv &\hat{\mathbf{Y}} = \mathbf{S_1 Y S_2}
\end{aligned}$$

Since the sandwich smoother is the cross product of two matrices, the corresponding univariate smoother matrix can be adjusted into a general form that allows for sandwich smoother estimates to be calculated for any partial derivative that satisfies the smoothness constraint given below:

Let there be a function $\mu(x, z)$ that has the same set up as the Tensor Product Spline. Suppose the function of interest is $\mu^{(\nu_1,\nu_2)}(x, z)$. This function can be estimated by first getting the penalized estimate, $\hat{\mathbf{A}}$, of $\mathbf{A}$ by minimizing $||\mathbf{Y} - \mathbf{B_1 A B_2^T}||_F^2 + \mathbf{a^T P a^T}$. Then using $\hat{\mathbf{a}}$, and defining $B_d^{(\nu)}(x)$ as the ν^{th} derivative of a degree d B-spline base with respect to x, the penalized estimate for generalized partial derivative with respect to x and z is

$$\hat{\mu}^{(\nu_1,\nu_2)}(x, z) = \sum_{\substack{1\le k\le I_1 \\ 1\le l\le I_2}} \hat{a}_{k,l} B_{k,d_1}^{1,(\nu_1)}(x) B_{l,d_2}^{2,(\nu_2)}(z)$$

and the sandwich estimator can be written as $\hat{\mathbf{Y}} = \mathbf{S_1^{(\nu_1)} Y S_2^{(\nu_2)}}$, where

$$\mathbf{S_1^{(\nu_1)}} = \mathbf{B_1^{(\nu_1)}} \left(\left(\mathbf{B_1^{(\nu_1)}} \right)^{\mathbf{T}} \mathbf{B_1^{(\nu_1)}} + \lambda_1 \mathbf{D_1^T D_1} \right)^{-1} \left(\mathbf{B_1^{(\nu_1)}} \right)^{\mathbf{T}}$$

and

$$\mathbf{S_2^{(\nu_2)}} = \mathbf{B_2^{(\nu_2)}} \left(\left(\mathbf{B_2^{(\nu_2)}} \right)^{\mathbf{T}} \mathbf{B_2^{(\nu_2)}} + \lambda_2 \mathbf{D_2^T D_2} \right)^{-1} \left(\mathbf{B_2^{(\nu_2)}} \right)^{\mathbf{T}}.$$

Setting up the tensor product P-spline as a sandwich smoother and using theory from Xiao et al. (2012) allows for the development of an asymptotic theorem for partial derivatives of penalized BTPB functions, which is presented in Chapter 3.

2.5 Future Work

P-spline methodologies have expanded in many interesting directions including differential equations, and spatial statistics. When applying P-splines to differential equations, the solution can be written as a sum of B-splines (the collocation method) and use the differential equation as the penalty (Ramsay and Silverman, 2007). This means that the penalty for smoothing splines is equivalent to a differential equation in which the second derivative of the solution is zero everywhere.

In spatial data analysis, tensor product P-splines may work better than kriging, especially in estimating a trend instead of spatial interpolation (Eilers et al., 2015). This is because when attempting to estimate a covariance structure, kriging often lead to unstable

procedures since handling for non-normal data with kriging is cumbersome. P-splines can create a relatively simple covariance structure in a very stable way. Also, additional smoothing of data on large grids is problematic for kriging, but P-splines tend to handle such data with ease. Further study into exactly what conditions allow P-spline to perform better than kriging and additional studies into head to head comparisons of the two methods could be very valuable to spatial statistics.

Even after being around for over twenty years, P-splines continue to be a popular semi-parametric model and an active area of research. Not only are P-splines being applied to different areas of statistics, such as the field of survey statistics (Chapter 4), but the theoretical aspect is being expanded as well in the area of asymptotic theory and variance estimation (Chapter 3). These chapters as well as other research continue to prove Eilers et al. (2015)'s statement true that "P-splines have much more to contribute to this century" .

CHAPTER 3. AN INVESTIGATION OF ACTUARIAL FAIR CROP INSURANCE RATES USING PARTIAL DERIVATIVES OF PENALIZED BIVARIATE TENSOR PRODUCT B-SPLINES

In this chapter, premium rates of yield insurance given by the US Department of Agriculture's Risk Management Agency are investigated to see if they are actuarially fair by comparing an estimated conditional yield density using premium data with the conditional yield density estimated using yields. A procedure is developed to estimate the conditional yield density using premium data through estimating partial derivatives of the premium rate function based on the penalized bivariate tensor product B-splines (BTPB). Xiao et al. (2012) is extended to study the asymptotic properties of partial derivatives of a penalized BTPB estimator and provide a variance estimator. The validity of the conditional yield density estimator using premium data and the variance estimation is demonstrated through simulation studies. The procedure is also applied to a crop insurance data set from Iowa to examine the actuarial fairness of the premium rates. On average, premium rates are close to the estimates. This is true for each coverage level. However, premiums for low productivity land are generally too low while those for high productivity land are generally too high. Even after subsidies, premiums for the more productive land are generally substantially higher than what they should be.

3.1 Introduction

Although a wide array of technologies have changed how field crop production takes place throughout the world, it continues to be a risky business. Perils include abiotic stressors such as drought, excess heat and flooding, and also biotic stressors due to insects, fungi and bacteria. Private sector crop insurance has long been available in the United States

(Gardner, 2009) and elsewhere. However, private offerings had generally not been popular due to high administration costs. Public sector involvement in United States crop insurance markets commenced with the Agricultural Adjustment Act of 1938, which created the Federal Crop Insurance Corporation whose job was to set premiums and support insurance on the main crops in the main growing areas. However, the federal government was not well-positioned to market the contracts that were made available. Participation was very low and outcomes were unsatisfactory because only the riskiest land was enrolled making high premiums necessary in order to cover losses. This meant that growers saw little point in enrolling good land at those high premium prices.

Commencing in 1980, the government sought to address the marketing problem by entering public-private partnerships. While participation rates grew moderately during the 1980s, on average, indemnities far exceeded the sum of farmer-paid premiums and premium subsidies provided by the federal government. It was only when premium subsidy levels increased dramatically and new contract forms were offered during the 1990s that participation expanded toward levels adequate to address adverse selection. As of 2018, U.S. federal government subsidies cover between 38% and 100% of the cost of providing crop insurance and farmers can cover up to 90% of historical yield. Program costs to the federal government over the decade 2007-2016 accumulate to $71.85 billion while 312 million acres of cropland were covered by the program in 2017. Elsewhere crop insurance has also expanded exclusively through public programs or public-private partnerships (Glauber, 2015; Santeramo and Ford Ramsey, 2017). Such programs are now widely available in Canada, Japan and China while many other countries have provided support for publicly subsidized pilot programs.

Central to the program implementation in the United States is the rate-setting procedure, as implemented by the Risk Management Agency (RMA) of the USDA where historical yield data is the primary source of tract-level data (Coble et al., 2010). As required by federal legislation, whenever available, the rate-setting procedure must use an arithmetic

average of no more than ten years of historical yield data on an insured unit called Actual Production History (APH) or historical yield. The rate-setting process can vary across counties, but the only statistic that can be extracted from historical yield observations is this arithmetic mean. The algorithm used to arrive at annual rates has evolved over the years to address various concerns (Coble et al., 2010).

The main goal of this paper is to investigate if the premium rates are set to be actuarially fair. If rates are not actuarially fair, then opportunities may exist to improve the efficiency of program implementation, reducing costs to taxpayers and ensuring a more efficient use of land to produce food, energy, and environmental outputs. Actuarial fairness in premium prices is examined by comparing the conditional yield density, conditional on land quality, inferred from the crop insurance premium data with the corresponding conditional yield density estimated from the yield data. If the premium rates are actuarially fair, these two curves estimated using two different data sources are expected to be similar. This paper focuses on the estimation of the conditional yield density curve based on the premium data, which involves estimating the premium function and its second partial derivatives. Kernel density estimation is used to estimate the conditional density curve based on yield data. The relationship between the conditional yield density curve and the premium function is then discussed in detail.

A fully parametric model, such as linear or polynomial model, can be imposed on the premium price function. However, it might suffer from model misspecification. A nonparametric approach is more robust to model misspecification than fully parametric methods, but estimators based on nonparametric procedures can have poor efficiency in small samples. To leverage the advantages of both parametric and nonparametric methods and avoid the limitations of a pure procedure, a semiparametric approach is employed in this paper. Bivariate smoothing has been studied in literature, such as bivariate P-splines (Eilers and Marx, 2003; Marx and Eilers, 2005), bivariate tensor product P-spline (BTPs) (Xiao et al., 2012), and thin plate splines (Wood, 2003). Among many findings is the well-known

trade-off between computational cost and smoothness. Spline smoothing methods demand massive computation, and an unpenalized spline tends to overfit the data leading to wiggly curves despite its cheap computational cost. Xiao et al. (2012) proposes a sandwich smoother that has a tensor product structure and can be computed quickly. It is thought to be the first central limit theorem for bivariate spline estimator of any type. Motivated by the goal of the real problem mentioned above, this paper is extended to study the properties of partial derivatives of penalized bivariate tensor produce B-splines (BTPB).

This paper has some theoretical and empirical contributions. Theoretically, the asymptotic theory for any general partial derivatives of penalized BTPB estimators is developed as well as a variance estimates. The validity and applicability of this estimator is demonstrated through theoretical proofs and simulations. Empirically, this estimator is used to address the actuarial fairness question in the premium rates prescribed by the RMA. Because the relationship between the premium prices and coverage levels or land quality is unknown and highly curved, specification of an adequate fully parametric model on premium price function is difficult, and nonparametric procedures may have large variances. The penalized BTPB is used to estimate the premium function and its partial derivatives. The asymptotic theory and the variance estimation developed for the partial derivatives of penalized BTPB estimators allow for statistically sound estimation for the conditional yield density based on premium data. This estimated density curve is later compared with the conditional yield density estimated using the yield data to determine if premium prices are actuarially fair, and such statistical comparisons also shed some light on what type of adjustments the government can perform to improve the premium prices. Webster County, Iowa, a county in the center of the corn belt with generally productive land and low intra-county land variance was used for the comparison of the two conditional densities. For all coverage levels, the actual premium rates are close to what was calculated given historical data. Conditional on historical average yields, however, premiums for poorer land are generally too low. Those

for the best land are too high for the farmer to expect to recover premium costs, even after accounting for large premium subsidies.

The rest of this chapter is organized as follows. In Section 3.2, a method for inferring a range of the density function through fitting penalized BTPB to insurance premiums is developed. Section 3.3 discusses the asymptotic normality for the estimated partial derivatives of penalized BTPB as well as variance estimation. Section 3.4 demonstrates the properties of the estimated partial derivatives of penalized BTPB through simulations. Section 3.5 shows the application to insurance data from Webster County, Iowa, and discusses some implications from the results.

3.2 Set-Up

In order to use insurance premium prices to estimate crop yield density, a formula connecting premiums to crop yield density using the actuarially fair rate needs to be derived. Discussion of this relationship is in Section 3.2.1. Section 3.2.2 shows the construction of penalized BTPB and how it can be used to estimate the partial derivatives of the premium price function, thus resulting in the estimation of the conditional yield density.

3.2.1 Actuarially Fair Rate

From a consumer's point of view, an insurance contract is actuarially fair if the premiums paid are equal to the expected value of the compensation received. This expected value is defined as the probability of the insured-against event occurring multiplied by the expected compensation to be received in the event of a loss (Eeckhoudt et al., 2005). For a given plot in a given year, a common approach to modeling yield uncertainty is defined by $w = \mu(z) + \sigma(z)\epsilon$, where w is current year yield, z is called land quality, $\mu(z)$ and $\sigma(z)$ are mean and standard deviation functions depending on z, and ϵ has distribution $G(\epsilon)$ on a compact set $[\underline{\epsilon}, \bar{\epsilon}]$ with $E(\epsilon) = 0$. Average historical yield is typically used as land quality z, and is

abbreviated to APH which stands for Actual Production History, a common terminology used by the USDA.

The crop insurance payout is given by

$$Payout = p \max[xz - w, 0] = p\sigma(z) \max[\psi - \epsilon, 0], \tag{3.1}$$

where x is the chosen coverage rate, p is the insured price and is calculated by averaging futures prices for the upcoming year, and $\psi = [xz - \mu(z)]/\sigma(z)$. Note that xz is the guaranteed yield. The actuarially fair rate for a given plot is determined by the expected value of the compensation received, where the expectation is taken with respect to the distribution of yield w for the plot in a given year. So the actuarially fair premium as a function of the coverage rate (x) and the land quality (z) is calculated as

$$\begin{aligned} \mu(x,z) &= p\sigma(z)E[\max(\psi - \epsilon, 0)] = p\sigma(z)\int_{\underline{\epsilon}}^{\psi}(\psi - \epsilon)dG(\epsilon) \\ &= p\sigma(z)\psi G(\psi) - p\sigma(z)\int_{\underline{\epsilon}}^{\psi}\epsilon dG(\epsilon) = p\sigma(z)\psi G(\psi) - p\sigma(z)[\psi G(\psi) - \int_{\underline{\epsilon}}^{\psi} G(\epsilon)d\epsilon] \ , \\ &= p\sigma(z)\int_{\underline{\epsilon}}^{\psi} G(\epsilon)d\epsilon = p\int_{\mu(z)+\sigma(z)\underline{\epsilon}}^{xz} F(w|z)dw \end{aligned} \tag{3.2}$$

where $F(w|z)$ is the conditional CDF of yield given the land quality z.

Taking the second partial derivative with respect to x on both sides of equation (3.2) gives $\frac{\partial^2\mu(x,z)}{\partial x^2} = pz^2 f(xz|z)$, where $f(\cdot|z)$ is the conditional PDF of yield given the land quality z. Set $w = xz$ to obtain

$$f(w|z) = \frac{1}{pz^2}\frac{\partial^2\mu(x,z)}{\partial x^2}\bigg|_{x=\frac{w}{z}} \quad \text{for} \quad zL_x \le w \le zU_x, \tag{3.3}$$

where L_x and U_x are the lower and upper bounds of the coverage rates. From (3.3), note that estimation of the partial derivative $\frac{\partial^2\mu(x,z)}{\partial x^2}$ becomes necessary for estimating the conditional density $f(w|z)$. The restricted range, $w \in [zL_x, zU_x]$, can also be seen from the actuarially fair rate formula in (3.2), where the upper limit of the integral in the last equality indicates that the information carried by the premium prices $\mu(x,z)$ can be used to estimate $F(w|z)$ only up to xz. So for a given z, the estimated conditional density $f(w|z)$ using premium data

is calculated for this specified range, and the comparison with the estimated $f(w|z)$ using yield data is restricted in this domain also. In section 3.2.2, we discuss how to construct penalized BTPB to estimate $\frac{\partial^{v_1+v_2}\mu(x,z)}{\partial x^{v_1}\partial z^{v_2}}$ at a given (x, z) for any value v_1 and v_2. Then the estimation of $\frac{\partial^2\mu(x,z)}{\partial x^2}$ is a special case, upon setting $v_1 = 2$ and $v_2 = 0$.

3.2.2 Construction of Penalized BTPB and Estimation of $f(w|z)$

This section introduces the construction of an univariate B-spline basis, then defines the penalized BTPB. Finally derivation of how to estimate partial derivatives of premium function $\mu(x, z)$ using penalized BTPB is described.

Define a general set up as follows. Assume $\{(x_i, z_j)\}_{1\le i\le n_1, 1\le j\le n_2}$ are deterministic on a rectangular grid, and y_{ij} are the premium rates with model $y_{ij} = \mu(x_i, z_j) + \epsilon_{ij}$, where $\mu(x_i, z_j)$ is a function on the compact set $(x, z) \in [0, 1]^2$, and the ϵ_{ij}'s are independent with $E(\epsilon_{ij}) = 0$ and $Var(\epsilon_{ij}) = \sigma^2(x_i, z_j)$. The deterministic rectangular grid is needed for setting up the sandwich smoother, which allows for the development of the asymptotic distribution. In Section 3.3.2, the assumption about the deterministic grid data pattern is relaxed and the situation with random design points is considered. First define a B-spline basis for an univariate variable. A B-spline basis of order d spans the linear space of piecewise polynomials of degree $d - 1$ with continuous derivatives up to order $d - 2$. B-splines allow improvements in computational efficiency over direct use of polynomial splines (Hastie et al., 2009). Let $r = 1$ (or $r = 2$) be the index for variable x (or z) respectively. Let K_r ($r = 1$ or 2) be the number of knots within the range $(0, 1)$, $K_r^* = K_r + 1$ be the number of intervals in this range, and d_r ($r = 1$ or 2) be the degree of the B-spline, for variable x or z. In order to construct the d_r-th degree B-spline basis, define equidistantly located knots as $\kappa_k = (K_r - 1)^{-1}k, (k = -d_r + 1, \cdots, K_r + d_r + 1)$. Then the d_1-th B-spline basis for variable x is given by

$$\mathbf{B}^1(x) = (B^1_{-d_1+1,d_1}(x), B^1_{-d_1,d_1}(x), \cdots, B^1_{K_1+1,d_1}(x))^T, \tag{3.4}$$

where $B^1_{k,d_1}(x)$ are defined recursively as,

- For $s = 0$:

$$B^1_{k,0}(x) = \begin{cases} 1 & \text{if } \kappa_{k-1} \leq x < \kappa_k \\ 0 & otherwise \end{cases} \quad \text{where } k = -d_1 + 1, \cdots, K_1 + d_1 + 1;$$

- For $s = 1, 2, \cdots, d_1$:

$$B^1_{k,s}(x) = \frac{x - \kappa_{k-1}}{\kappa_{k+s-1} - \kappa_{k-1}} B^1_{k,s-1}(x) + \frac{\kappa_{k+s} - x}{\kappa_{k+s} - \kappa_k} B^1_{k+1,s-1}(x), \tag{3.5}$$

  where κ_k's are the knot locations.

Similarly, the B-spline basis of order d_2 for variable z can be specified as

$$\mathbf{B}^2(z) = (B^2_{-d_2+1,d_2}(z), B^2_{-d_2,d_2}(z), \cdots, B^2_{K_2+1,d_2}(z))^T. \tag{3.6}$$

The tensor product spline of two variables is defined as

$$\sum_{\substack{1 \leq k \leq I_1 \\ 1 \leq l \leq I_2}} a_{k,l} B^1_{k,d_1}(x) B^2_{l,d_2}(z), \tag{3.7}$$

Here $B^1_{k,d_1}(x)$ and $B^2_{l,d_2}(z)$ denote the elements in the basis functions for x and z defined in (3.4) and (3.6) respectively, $I_1 = K_1 + d_1 + 1$, $I_2 = K_2 + d_2 + 1$, and $a_{k,l}$ represents the coefficients that need to be estimated. Because the data are in a rectangular pattern, the observations can be organized into a $n_1 \times n_2$ data matrix $\mathbf{Y} = \{y_{i,j}\}_{1 \leq i \leq n_1; 1 \leq j \leq n_2}$. Define the model matrices using B-spline basis for x and z as $\mathbf{B}_1 = \{B^1_{k,d_1}(x_i)\}_{1 \leq i \leq n_1, 1 \leq k \leq I_1}$, and $\mathbf{B}_2 = \{B^2_{l,d_2}(z_j)\}_{1 \leq j \leq n_2, 1 \leq l \leq I_2}$, and the coefficient matrix $\mathbf{A} = \{a_{k,l}\}_{1 \leq k \leq I_1, 1 \leq l \leq I_2}$. Let $\mathbf{a} = Vec(\mathbf{A})$. The estimation of $\mathbf{a}$ can be obtained by minimizing the following objective function with a penalty term,

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}} ||\mathbf{Y} - \mathbf{B_1 A B}_2^T||_F^2 + \mathbf{a}^T \mathbf{P a}, \tag{3.8}$$

where $|| \cdot ||_F$ is the Frobenius Norm defined as the square root of the sum of the absolute squares of a matrix's elements and $\mathbf{P}$ is defined as

$$\mathbf{P} = \lambda_1 \mathbf{B}_2^T \mathbf{B_2} \otimes \mathbf{D}_1^T \mathbf{D_1} + \lambda_2 \mathbf{D}_2^T \mathbf{D_2} \otimes \mathbf{B}_1^T \mathbf{B_1} + \lambda_1 \lambda_2 \mathbf{D}_2^T \mathbf{D_2} \otimes \mathbf{D}_1^T \mathbf{D_1} \tag{3.9}$$

and is the penalty on the coefficients matrix which allows for the tensor-product spline model to be used while avoiding overfitting the data. Here λ_1 (or λ_2) is the smoothing parameter for variable x (or z), and the matrix $\mathbf{D}_1$ (or $\mathbf{D}_2$) is the difference matrix with order m_1 (or m_2) for variable x (or z) with the definition $\mathbf{D}_r = \{d^r_{k,l}\}_{1\le k\le I_r-m_r+1, 1\le l\le I_r}$, where $r = 1$ or 2, $d^r_{k,l} = (-1)^{l-k}\binom{m_r}{l-k}$ for $0 \le l-k \le m_r$ and 0 otherwise. The difference penalty is used to remove computational difficulty occurring when the penalty term is defined through an integral, and it controls the smoothness of the estimated bivariate function. The penalized estimated function $\hat{\mu}(x,z)$ is defined as (Xiao et al., 2012)

$$\hat{\mu}(x,z) = \sum_{\substack{1\le k\le I_1\\ 1\le l\le I_2}} \hat{a}_{k,l} B^1_{k,d_1}(x) B^2_{l,d_2}(z). \tag{3.10}$$

Define $\mu^{(v_1,v_2)}(x,z) = \frac{\partial^{(v_1+v_2)}\mu(x,z)}{\partial x^{v_1}\partial z^{v_2}}$, for any $v_1 > 0$ and $v_2 > 0$. Then the penalized BTPB estimator of $\mu^{(v_1,v_2)}(x,z)$ can be obtained as

$$\hat{\mu}^{(v_1,v_2)}(x,z) = \sum_{\substack{1\le k\le I_1\\ 1\le l\le I_2}} \hat{a}_{k,l} B^{1(v_1)}_{k,d_1}(x) B^{2(v_2)}_{l,d_2}(z), \tag{3.11}$$

where $\hat{a}_{k,l}$ is calculated in (3.8), and the derivatives of the B-splines can be obtained recursively as follows (Prochazkova, 2005). For any $k = 1, \cdots, I_1$ and $\nu = 1, \cdots, v_1$,

$$B^{1(\nu)}_{k,d_1}(x) = \frac{d_1}{\kappa_{k+d_1} - \kappa_k} B^{1(\nu-1)}_{k,d_1-1}(x) - \frac{d_1}{\kappa_{k+d_1+1} - \kappa_{k+1}} B^{1(\nu-1)}_{k+1,d_1-1}(x).$$

For any $l = 1, \cdots, I_2$ and $\nu = 1, \cdots, v_2$,

$$B^{2(\nu)}_{l,d_2}(z) = \frac{d_2}{\kappa_{l+d_2} - \kappa_l} B^{2(\nu-1)}_{l,d_2-1}(z) - \frac{d_2}{\kappa_{l+d_2+1} - \kappa_{l+1}} B^{2(\nu-1)}_{l+1,d_2-1}(z).$$

Here a B-spline function with zero derivative ($\nu = 0$) is defined to be itself. Then according to the relationship between the partial derivatives of the premium function and the conditional yield density shown in (3.3), the conditional yield density is estimated as

$$\hat{f}(w|z) = \frac{1}{pz^2}\hat{\mu}^{(2,0)}(x,z)\Big|_{x=\frac{w}{z}} \quad \text{for} \quad zL_x \le w \le zU_x. \tag{3.12}$$

3.3 Asymptotic Distributions and Variance Estimation

In order to make statistical inference for the estimator in (3.12), the asymptotic property of the derivative estimator $\hat{\mu}^{(v_1,v_2)}(x,z)$, defined in (3.11), is studied, and a variance estimator for it is developed in this section. This derivative estimator $\hat{\mu}^{(v_1,v_2)}(x,z)$ is asymptoticly equivalent to a bivariate kernel regression estimator with a product of two kernels derivatives. The idea of the equivalent kernel method was first proposed to study asymptotic analysis of smoothing splines (Rice and Rosenblatt, 1983; Silverman, 1984), and was then used to derive asymptotic properties of univariate P-splines (Li and Ruppert, 2008; Wang et al., 2009). This idea was also employed to develop the asymptotic theory for the bivariate tensor product P-splines estimators, but not for the partial derivatives (Xiao et al., 2012). Borrowing ideas from Xiao et al. (2012), the asymptotic theory for any partial derivatives of a penalized BTPB estimator under an arbitrary choice of degree and penalty is derived. Section 3.3.1 develops the asymptotic normality of $\hat{\mu}^{(v_1,v_2)}(x,z)$ under a deterministic grid pattern, and Section 3.3.2 extends the theory to deal with random design points. Section 3.3.3 discusses variance estimation. Kernel density estimation and kernel density variance estimation is covered in Section 3.3.4.

3.3.1 Asymptotic Normality of $\hat{\mu}^{(v_1,v_2)}(x,z)$

The derivation of the asymptotic distribution of $\hat{\mu}^{(v_1,v_2)}(x,z)$ proceeds in two main steps. Lemma 1 proves the equivalence between the estimator $\hat{\mu}^{(v_1,v_2)}(x,z)$ and the bivariate kernel regression estimator with the product of two kernel derivatives. Lemma 1 is analogous to Proposition 1 in Xiao et al. (2012). Theorem 1 then presents the asymptotic normality of $\hat{\mu}^{(v_1,v_2)}(x,z)$ and its asymptotic bias, using the Lemma 1 result.

Begin by defining a univariate kernel function $H_m(x)$ as $H_m(x) = (2m)^{-1}\sum_{k=1}^{m}\psi_k e^{-\psi_k|x|}$, where m is a positive integer and the ψ_k's are the m complex roots of $x^{2m}+(-1)^m = 0$ that have positive real parts. The closed form expression of the first four equivalent kernels

$H_m(x)$ are given below (Wang 2009).

$$
\begin{aligned}
H_1(x) =& \frac{1}{2}e^{-|x|} \\
H_2(x) =& \frac{1}{2\sqrt{2}}e^{-\frac{1}{\sqrt{2}}|x|}\left(\cos\frac{|x|}{\sqrt{2}}+\sin\frac{|x|}{\sqrt{2}}\right) \\
H_3(x) =& \frac{1}{6}e^{-|x|}+e^{-\frac{1}{2}|x|}\left(\frac{1}{6}\cos\frac{\sqrt{3}|x|}{2}+\frac{\sqrt{3}}{6}\sin\frac{\sqrt{3}|x|}{2}\right) \\
H_4(x) =& e^{-.9239|x|}\Big(.2310\cos\big(.3827|x|\big)+.0957\sin\big(.3827|x|\big)\Big) \\
&+e^{-.3827|x|}\Big(.0957\cos\big(.9239|x|\big)+.2310\sin\big(.9239|x|\big)\Big)
\end{aligned}
$$

Define the bivariate regression estimator with the product of kernel derivatives as

$$\mu^{*(v_1,v_2)}(x,z)=\frac{1}{n_1h_1^{v_1+1}}\frac{1}{n_2h_2^{v_2+1}}\sum_{i,j}H_{m_1}^{(v_1)}\left(\frac{x-x_i}{h_1}\right)H_{m_2}^{(v_2)}\left(\frac{z-z_i}{h_2}\right)y_{i,j},$$

where h_1 and h_2 are the bandwidths for variables x and z respectively, and $H_{m_1}^{(v_1)}(0)$ and $H_{m_2}^{(v_2)}(0)$ are taken to be the right derivatives of $H_{m_1}^{(v_1)}(x)$ and $H_{m_2}^{(v_2)}(z)$ at 0 for $v_1>0$ and $v_2>0$. The following Lemma 1 shows that $\hat{\mu}^{(v_1,v_2)}(x,z)$ is asymptotically equivalent to $\mu^{*(v_1,v_2)}(x,z)$.

Lemma 1: Assume the following conditions are satisfied.

(1) There exists a constant $\delta>0$ such that $\sup_{i,j}E\Big(|y_{i,j}|^{2+\delta}\Big)<\infty$;

(2) The regression function $\mu(x,z)$ has continuous g^{th} order derivatives where $g=\max(2m_1+v_1,2m_2+v_2)$;

(3) The variance function $\sigma^2(x,z)$ is continuous;

(4) The covariates satisfy $(x_i,z_j)=\Big(\big((i-.5)/n_1\big),\big((j-.5)/n_2\big)\Big)$;

(5) $n_1\sim cn_2$ where c is a constant;

(6) $h_1=O(n^{-\nu_1})$ and $h_2=O(n^{-\nu_2})$ for some constants $0<\nu_1,\nu_2<1$, and $(K_1^*h_1^2)^{-1}=o(1)$ and $(K_2^*h_2^2)^{-1}=o(1)$, where h_1 and h_2 are specified as $h_1=K_1^{*-1}(\lambda_1K_1^*n_1^{-1})^{1/(2m_1)}$ and $h_2=K_2^{*-1}(\lambda_2K_2^*n_2^{-1})^{1/(2m_2)}$.

Then

$$E\Big\{\hat{\mu}^{(v_1,v_2)}(x,z)-\mu^{*(v_1,v_2)}(x,z)\Big\}=O\Big(\zeta h_1^{v_1}h_2^{v_2}\Big),$$
$$Var\Big\{\hat{\mu}^{(v_1,v_2)}(x,z)-\mu^{*(v_1,v_2)}(x,z)\Big\}=o\Big((nh)^{-1}h_1^{-2v_1}h_2^{-2v_2}\Big),$$

where $n=n_1n_2$, $h=h_1h_2$, and $\zeta=O\Big[\max\big\{(K_1^*h_1)^{-2},(K_2^*h_2)^{-2}\big\}\Big]$.

The proof for Lemma 1 follows the approach of Xiao et al. (2012), but accounts for the orders derived for the derivatives and is given in Appendix A.

Theorem 1 presents the asymptotic normality of $\hat{\mu}^{(v_1,v_2)}(x,z)$, including its asymptotic bias and variance.

Theorem 1: Assume the conditions in Lemma 1 are satisfied, and $K_1^*=O(n^{\tau_1})$, $K_2^*=O(n^{\tau_2})$, $h_1=O(n^{-m_2/m_3})$ and $h_2=O(n^{-m_1/m_3})$ where $m_3=4m_1m_2+m_1+m_2$, $\tau_1>(m_1+1)m_2/m_3$ and $\tau_2>(m_2+1)m_1/m_3$. Then for any given $(x,z)\in(0,1)\times(0,1)$,

$$\sqrt{nh_1^{2v_1+1}h_2^{2v_2+1}}\Big(\hat{\mu}^{(v_1,v_2)}(x,z)-\mu^{(v_1,v_2)}(x,z)-\mu_b(x,z)\Big)\implies N\big(0,V(x,z)\big),$$

in distribution as $n_1\to\infty$ and $n_2\to\infty$, where

$$\mu_b(x,z)=(-1)^{m_1+1}h_1^{2m_1}\mu^{(2m_1+v_1,v_2)}(x,z)+(-1)^{m_2+1}h_2^{2m_2}\mu^{(v_1,2m_2+v_2)}(x,z) \tag{3.13}$$

$$V(x,z)=\sigma^2(x,z)\int\Big(H_{m_1}^{(\nu_1)}(u)\Big)^2du\int\Big(H_{m_2}^{(\nu_2)}(v)\Big)^2dv. \tag{3.14}$$

Here the asymptotic bias $\mu_b(x,z)$ is due to the difference between the true function $\mu^{(v_1,v_2)}(x,z)$ and the bivarate kernel regression estimator $\mu^{*(v_1,v_2)}(x,z)$ used for approximation. Using the result in Lemma 1, the key step in the proof of Theorem 1 is to show

$$\begin{aligned}
&\hat{\mu}^{(v_1,v_2)}(x,z)-\mu^{(v_1,v_2)}(x,z)=\hat{\mu}^{(v_1,v_2)}(x,z)-\mu^{*(v_1,v_2)}(x,z)+\mu^{*(v_1,v_2)}(x,z)-\mu^{(v_1,v_2)}(x,z)\\
&=o\bigg(\frac{1}{\sqrt{nh}}\frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\bigg)+\frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\frac{1}{n_1h_1}\frac{1}{n_2h_2}\sum_{ij}H_{m_1}^{(v_1)}\bigg(\frac{x-x_i}{h_1}\bigg)H_{m_2}^{(v_2)}\bigg(\frac{z-z_i}{h_2}\bigg)\varepsilon_{ij}\\
&+\frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\frac{1}{h_1}\frac{1}{h_2}\int\int H_{m_1}^{(v_1)}\bigg(\frac{x-u}{h_1}\bigg)H_{m_2}^{(v_2)}\big(\frac{z-v}{h_2}\big)\mu(u,v)dudv-\mu^{(v_1,v_2)}(x,z),
\end{aligned} \tag{3.15}$$

where the last two terms after the first equality in (3.15) can be shown to be $\mu_b(x,z) + o((nh)^{-1/2}h_1^{-v_1}h_2^{-v_2})$. Then the results in Theorem 1 follow. The sketch of the proof for Theorem 1 is given in Appendix A.

3.3.2 Asymptotic Distribution with Random Design Points and Variance Estimation

In this subsection, consider the situation with random design points. Assume the model $y_i = \mu(x_i, z_i) + \epsilon_i$, for $i = 1, ..., n$, where (x_i, z_i)'s are i.i.d. from a density $f(x,z)$ on the compact set $[0,1]^2$. As suggested by Xiao et al. (2012), the compact set $[0,1]^2$ is divided into $S_1 \times S_2$ equally sized rectangular grids. Let $(\tilde{x}_{i'}, \tilde{z}_{j'})$ $(i' = 1, ..., S_1; j' = 1, ..., S_2)$ be the center of the (i', j')-th grid, and $\tilde{y}_{i',j'}$ be the average of all y_i's with (x_i, z_i) falling into the (i', j')-th grid. If the (i', j')-th grid do not have data, then $\tilde{y}_{i',j'}$ is defined to be the mean of y_i's with (x_i, z_i) falling into the bins that are adjacent to the (i', j')-th grid.

Apply the same estimation method described in Section 3.2.2 on the data matrices, $\tilde{\mathbf{Y}} = \{\tilde{y}_{i',k'}\}_{1\le i'\le S_1, 1\le j'\le S_2}$, $\tilde{\mathbf{B}}_1 = \{B^1_{k,d_1}(\tilde{x}_{i'})\}_{1\le i'\le S_1, 1\le k\le I_1}$, and $\tilde{\mathbf{B}}_2 = \{B^1_{l,d_2}(\tilde{z}_{j'})\}_{1\le j'\le S_2, 1\le l\le I_2}$. The estimated coefficients of $\mathbf{a}$ is obtained by,

$$\tilde{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}} ||\tilde{\mathbf{Y}} - \tilde{\mathbf{B}}_1 \mathbf{A} \tilde{\mathbf{B}}_2^T||_F^2 + \mathbf{a}^T \tilde{\mathbf{P}} \mathbf{a}, \tag{3.16}$$

where the penalty matrix $\mathbf{P}$ is defined as

$$\tilde{\mathbf{P}} = \lambda_1 \tilde{\mathbf{B}}_2^T \tilde{\mathbf{B}}_2 \otimes \mathbf{D}_1^T \mathbf{D}_1 + \lambda_2 \mathbf{D}_2^T \mathbf{D}_2 \otimes \tilde{\mathbf{B}}_1^T \tilde{\mathbf{B}}_1 + \lambda_1 \lambda_2 \mathbf{D}_2^T \mathbf{D}_2 \otimes \mathbf{D}_1^T \mathbf{D}_1. \tag{3.17}$$

Then the penalized estimated function $\tilde{\mu}^{(v_1,v_2)}(x,z)$ is

$$\tilde{\mu}^{(v_1,v_2)}(x,z) = \sum_{\substack{1\le k\le I_1 \\ 1\le l\le I_2}} \tilde{a}_{k,l} B^{1(v_1)}_{k,d_1}(x) B^{2(v_2)}_{l,d_2}(z). \tag{3.18}$$

Applying Theorem 1 to the binned data $\tilde{\mathbf{Y}}$ with n_1 and n_2 replaced by S_1 and S_2, gives the asymptotic normality theory for $\tilde{\mu}^{(v_1,v_2)}(x,z)$.

Theorem 2: Assume the following conditions are satisfied.

Condition (1) and (2) are the same as the first two conditions in Lemma 1;

(3) The design points $\Big\{\big(x_i, z_i\big)\Big\}_{i=1}^{n}$, are independent and sampled from a distribution $F(x,z)$ with a density function $f(x,z)$ and $f(x,z)$ is positive over $[0,1]^2$ and has continuous first derivatives;

(4) Conditional on $\Big\{\big(x_i, z_i\big)\Big\}_{i=1}^{n}$, the random errors $\varepsilon_i, 1 \le i \le n$, are independent with mean 0 and conditional variance $\sigma(x_i, z_i)$;

(5) The variance function $\sigma^2(x,z)$ is twice continuously differentiable;

(6) $S \sim c_S n^{\tau}$ and $S_1 \sim c_0 S_2$ for some constants c_S, c_0, and $\tau > \dfrac{4m_1m_2}{4m_1m_2 + m_1 + m_2}$.

For any given $(x,z) \in (0,1)\times(0,1)$,

$$\sqrt{nh_1^{2v_1+1}h_2^{2v_2+1}}\Big(\tilde{\mu}^{(v_1,v_2)}(x,z) - \mu^{(v_1,v_2)}(x,z) - \mu_b(x,z)\Big) \implies N\bigg(0, \frac{V(x,z)}{f(x,z)}\bigg),$$

in distribution as $n \to \infty$ where $\mu_b(x,z)$ and $V(x,z)$ are defined in (3.13) and (3.14).

The proof uses the following decomposition

$$\begin{aligned} & \sqrt{nh}h_1^{v_1}h_2^{v_2}\Big(\tilde{\mu}^{(v_1,v_2)}(x,z) - \mu^{(v_1,v_2)}(x,z) - \mu_b(x,z)\Big) \\ = \;& \sqrt{nh}h_1^{v_1}h_2^{v_2}\bigg[\tilde{\mu}^{(v_1,v_2)}(x,z) - E\Big\{\tilde{\mu}^{(v_1,v_2)}(x,z)|\big(\underline{\mathrm{x}},\underline{\mathrm{z}}\big)\Big\}\bigg] \\ + \;& \sqrt{nh}h_1^{v_1}h_2^{v_2}\bigg[E\Big\{\tilde{\mu}^{(v_1,v_2)}(x,z)|\big(\underline{\mathrm{x}},\underline{\mathrm{z}}\big)\Big\} - \mu^{(v_1,v_2)}(x,z) - \mu_b(x,z)\bigg], \end{aligned} \tag{3.19}$$

where $\big(\underline{\mathrm{x}},\underline{\mathrm{z}}\big)$ are the design points $(x_i,z_i)_{1\le i\le n}$. Then the right-side term in (3.19) is shown to have a limiting distribution $N(0, V(x,z)/f(x,z))$, and the second term in (3.19) to have the small order $o_p(1)$.

3.3.3 Variance Estimation

The asymptotic variance in Theorem 2 is used to estimate the variance of $\tilde{\mu}^{(v_1,v_2)}(x,z)$ for any (x,z). To estimate $\sigma^2(x,z)$, a penalized BTPB estimator is applied to the data $(\tilde{\epsilon}^2_{i',j'}, \tilde{x}_{i'}, \tilde{z}_{j'})_{1\le i'\le S_1; 1\le j'\le S_2}$, where $\tilde{\epsilon}_{i',j'}$ is the mean of residuals $\hat{\epsilon}_i$'s with (x_i,z_i) falling into the (i',j')-th bin, and $\hat{\epsilon}_i = y_i - \tilde{\mu}(x_i,z_i)$ for $i = 1,...,n$. Therefore the estimator $\sigma^2(x,z)$ is

obtained by

$$\hat{\sigma}^2(x,z) = \sum_{\substack{1\le k\le I_1\\ 1\le l\le I_2}} \tilde{e}_{k,l} B^1_{k,d_1}(x) B^2_{l,d_2}(z), \tag{3.20}$$

where $\tilde{e}_{k,l}$ is estimated by replacing $\tilde{\mathbf{Y}}$ by $\tilde{\mathbf{\Omega}} = \{\tilde{\epsilon}^2_{i',j'}\}_{1\le i'\le S_1; 1\le j'\le S_2}$ in (3.16),

$$\tilde{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmin}} ||\tilde{\mathbf{\Omega}} - \tilde{\mathbf{B}}_\mathbf{1}\mathbf{E}\tilde{\mathbf{B}}_\mathbf{2}^\mathbf{T}||_\mathbf{F}^\mathbf{2} + \mathbf{e}^\mathbf{T}\tilde{\mathbf{P}}\mathbf{e}, \tag{3.21}$$

where $\tilde{\mathbf{P}}$ is defined in (3.17), $\mathbf{E} = \{e_{k,l}\}_{1\le k\le I_1; 1\le l\le I_2}$, and $\mathbf{e} = Vec(\mathbf{E})$.

The estimator of variance is defined as

$$\widehat{Var}[\tilde{\mu}^{(v_1,v_2)}(x,z)] = \frac{4}{nh_1^{2v_1+1}h_2^{2v_2+1}} \frac{\hat{\sigma}^2(x,z)}{\hat{f}(x,z)} \sum_{t=1}^{T} \left\{ H^{(v_1)}_{m_1}\left(\frac{tL}{T}\right)\right\}^2 \sum_{t=1}^{T} \left\{ H^{(v_2)}_{m_2}\left(\frac{tL}{T}\right)\right\}^2, \tag{3.22}$$

where L and T are integers large enough such that $\sum_{t=1}^{T} H^{(v_i)}_{m_i}\left(\frac{tL}{T}\right)$ $(i = 1, 2)$ is a good approximation to $\int_0^\infty H^{(v_i)}_{m_i}(u)du$, and $\hat{f}(x,z)$ is the bivariate kernel density estimator for the joint pdf $f(x,z)$. The standard bivariate Gaussian kernel is used to estimate the joint density $f(x,z)$, and bandwidth is determined by Sheather-Jones bandwidth selection (Sheather and Jones, 1991).

3.3.4 Kernel Density Estimation

To get the crop yield density using the kernel density estimator, estimate $\mu(z_i)$ and $\sigma(z_i)$ using univariate penalized B-splines according to the following algorithm:

1. $\hat{\mu}(z_i)$ is estimated using a cubic spline with ten knots and λ determined by generalized cross validation (GCV) on (w_i, z_i).

2. $\hat{\sigma}(z_i)$ is then estimating using a cubic spline with ten knots and λ determined by GCV on (γ_i, z_i) where $\gamma_i = w_i - \hat{\mu}(z)$.

3. A Gaussian smoothing kernel with bandwidth is determined by Sheather-Jones bandwidth selection used in Sheather and Jones (1991) is used to estimate $\hat{f}_\epsilon(\epsilon)$ using $\hat{\epsilon}_i$ where $\hat{\epsilon}_i = \dfrac{w_i - \hat{\mu}(z)}{\hat{\sigma}(z_i)}$.

4. Transform $\hat{f}_\epsilon(\epsilon)$ to obtain $\tilde{f}_{w|z}(w|z)$, where

$$\tilde{f}_{w|z}(w|z) = \hat{f}_\epsilon\left(\frac{w_i - \hat{\mu}(z)}{\hat{\sigma}(z_i)}\right) * \frac{1}{\hat{\sigma}(z_i)}. \tag{3.23}$$

5. To calculate $\widehat{Var}\left(\tilde{f}_{w|z}(w|z)\right)$, the delete one Jackknife method is used where

$$\widehat{Var}\left(\tilde{f}_{w|z}(w|z)\right) = \frac{n-1}{n}\sum_{b=1}^{n}\left[\tilde{f}^{[b]}_{w|z}(w|z) - \tilde{f}_{w|z}(w|z)\right]^2 \tag{3.24}$$

where $\tilde{f}^{[b]}(w|z)$ is the estimate of $\tilde{f}(w|z)$ with the b^{th} observation removed.

3.4 Simulation Studies

The main objectives of the simulation studies are to validate the conditional crop yield density estimators and evaluate their variance estimators. The conditional crop yield density estimator estimated from premium prices, $\hat{f}(w|z)$ of (3.12), using derivatives of penalized BTPB in (3.18) and variance estimator defined in (3.25) is in Section 3.4.1. The conditional crop yield density estimator estimated from historical crop yield, $\tilde{f}_{w|z}(w|z)$ of (3.23), using the kernel density estimation in Section 3.3.4 and variance estimator defined in (3.24) is in Section 3.4.2.

3.4.1 Penalized BTPB Simulation Study

The simulation set-up is specified as follows. Current year yield, w_i, is assumed to follow a model $w_i = \mu(z_i) + |z_i|^{1/5}\epsilon_i$, where $\mu(z_i)$ is the mean function and ϵ_i's are i.i.d. random errors. Consider the following two mean functions,

$$Linear : \mu(z_i) = -25 + 1.3z_i,$$

$$Quadratic : \mu(z_i) = 1.2(z_i - 50) + \frac{(z_i - 150)^2}{200},$$

and the following two random error distributions,

$$Normal : \epsilon_i \sim 10 * N(0,1);$$

$$Beta : \epsilon_i \sim 50 * [Beta(5,3) - 5/8].$$

The covariate z_i (land quality) are i.i.d. from a truncated normal distribution $N(200, 15^2)$ on interval $[100, 300]$. The beta distribution and the non-constant scaling factor $|z_i|^{1/5}$ are used because corn yields tend to have left skewed distributions (Ker and Goodwin, 2000; Du et al., 2012, 2015) and increasing yield variation as historical yield increases (Tannura et al., 2008). Based on this yield model, observed premiums are generated from $y_i = \mu(x_i, z_i) + \zeta_i$, where $\zeta_i \sim N(0, 0.1^2)$ are measurement errors, the covariate x_i (coverage rates) are i.i.d. from a uniform distribution on the discrete numbers $(0.55, 0.60, ..., 0.90, 0.95)$, mimicking real coverage rates in practice, and the true premium price is $\mu(x, z) = p \int_{z+|z|^{1/5}\underline{\epsilon}}^{xz} F_{w|z}(w|z)dw$ from (3.2). Here $p = 4$, the 2009 APH market price which is set by RMA for crop insurance purposes (Theisse, 2009), and $\underline{\epsilon}$ is -5 and $-.375$ for the Normal and Beta error distributions respectively. The conditional cdf $F_{w|z}(w|z)$ is given as

$$F_{w|z}(w|z) = \begin{cases} \Phi\left(\dfrac{w - \mu(z)}{|z|^{1/5}}, 0, 1\right) & \text{if } \epsilon_i \sim 10 * N(0,1) \\ \mathcal{I}\left(\dfrac{5}{5+3} + \dfrac{w - \mu(z)}{|z|^{1/5}}; 5, 3\right) & \text{if } \epsilon_i \sim 50 * [Beta(5,3) - 5/8], \end{cases}$$

where $\Phi(\cdot)$ is the cdf function of $N(0,1)$, and $\mathcal{I}(\cdot; 5, 3)$ is the regularized beta function. The integration in the function $\mu(x, z)$ is approximated by a Riemann sum. For each of the four combinations of mean functions and error distributions, simulate a sample of (x_i, z_i, y_i) for $i = 1, ..., 500$, and repeat the procedure to obtain $1,000$ Monte Carlo (MC) samples. For each MC sample under each scenario, based on the simulated premium data the estimated conditional density curve $\hat{f}(w|z)$ of (3.12) is computed using derivatives of penalized BTPB in (3.18). By (3.12), which was derived for the relationship between the conditional density and the partial derivative of the premium function, the point-wise standard error $\widehat{SE}[\hat{f}(w|z)]$

can be obtained from the equation:

$$\widehat{SE}[\hat{f}(w|z)] = \frac{1}{pz^2}\sqrt{\widehat{Var}\left[\tilde{\mu}^{(2,0)}\left(\frac{w}{z}, z\right)\right]}. \tag{3.25}$$

The order of spline $d = 3$ (i.e. cubic) and the degree of penalty $m = 3$ are popular choices in practice as suggested by Yoshida (2013). So set $m_i = 3$ $(i = 1, 2)$, $d_2 = 3$ and $d_1 = 5$ to allow for $\mu^{(2,0)}(x, z)$ to be estimated using cubic splines. The number of bins for x (or for z) is set to be $S_1 = 9$ (or $S_2 = 200$). These values are chosen so there is very little information loss by binning the data since x_i are i.i.d. from the uniform distribution on discrete values $(0.55, 0.60, ..., 0.95)$ and z_i are i.i.d. from $N(200, 15^2)$ truncated on $[100, 300]$. Ruppert (2002)'s recommendation for the number of knots K_i is $\min(S_i/4, 35)$. Thus $K_1 = 2$ is used for x. Because the fit changes little in test cases when K_2 varies from 35 to 10. To reduce computation time, $K_2 = 10$ was chosen. z_i values are normalized to the compact set $[0, 1]$ by subtracting of the minimum value (100) and dividing by the range (200) to ensure the asymptotic theory assumptions are met. The smoothing parameter λ_i's are determined using generalized cross validation (GCV) proposed in Golub et al. (1979).

Figure 3.1 plots the MC results of the conditional density estimator $\hat{f}(w|z)$ using the penalized BTPB defined in (3.18) for all four scenarios respectively: Linear-Normal, Linear-Beta, Quadratic-Normal, Quadratic-Beta. Five subplots in each panel give the MC results for five fixed land quality values at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of z_i. In each subplot under each panel, the solid line is the true function $f(w|z)$, the dashed line is the mean of the MC estimated $\hat{f}(w|z)$, and the lower and upper dotted lines are the 2.5^{th} and 97.5^{th} point-wise percentiles of the MC estimates. Note that in each subplot (a given z value) only part of the $\hat{f}(w|z)$ curve is estimated for the restricted range, $0.55z < w < 0.95z$, due to the reason mentioned in Section 3.2.1. The fit around the center of the data (for land quality z's that are percentiles ranging from 25^{th} to 75^{th}), appear to fit well with very small variation. At the data edges, variation tends to increase. This is expected since boundary effects are one of the limitations of spline functions because they use surrounding data for estimation and a value on the edge does not have the benefit of this surrounding data.

In order to evaluate the variance estimator performance in (3.25), the variance function $\hat{\sigma}^2(x,z)$ and the joint density $\hat{f}(x,z)$ need to be estimated. The tuning parameters for estimating $\hat{\sigma}^2(x,z)$ are chosen in the same manner as for estimating $\hat{f}(w|z)$, and the bandwidth for estimating $\hat{f}(x,z)$ is chosen by data-based selection used in Sheather and Jones (1991). The values of $L = 1,000$ and $T = 1,000$ are used in (3.22) when calculating the variance of the derivatives of the BTPB estimators.

Figure 3.2 shows coverage probability heat maps of the conditional yield density estimator using the derivatives of the BTPB for the four scenarios respectively. Using the asymptotic normality proved in Theorem 2 and the standard errors proposed in (3.25), a point-wise 95% confidence band is constructed for each MC replicate and the percentage of times that the true value is within the band is given on the heat maps. The x-axis represents the insurance coverage levels x_i, while the y-axis represents the land quality z_i. To save computation time, coverage rates were calculated only at a few locations. The values given on the heat map are the coverage rate at the midpoint of the grid cell. In each panel, lighter colors represent values close to the nominal probability of 95%, and the darker colors represent values that are either below or above the 95% nominal rate. Again the variance estimator performs well around the center of the data with performance dropping off near the data edges. However the performance on the edges tends to produce a more conservative result, with the exception being when land quality and coverage levels are high in the Quadratic-Normal scenario.

In summary, the simulation study results considered in this section confirm the effectiveness of the conditional yield density estimation based on the premium data and the derivatives of the penalized BTPB method. But it must be recognized that the estimation near the data edges may not be highly accurate due to lack of data information on boundaries.

3.4.2 Kernel Density Simulation Study

The model for the current year yield data is given by

$$w_i = \mu(z_i) + \sigma(z_i)\epsilon_i. \tag{3.26}$$

Specify the simulation set-up as follows. Current year yield, w_i, is assumed to follow a model $w_i = \mu(z_i) + |z_i|^{1/5}\epsilon_i$, where $\mu(z_i)$ is the mean function and ϵ_i's are i.i.d. random errors. Consider the following two mean functions,

$$Linear : \mu(z_i) = -25 + 1.3z,$$

$$Quadratic : \mu(z_i) = 1.2(z_i - 50) + \frac{(z_i - 150)^2}{200},$$

and the following two random error distributions,

$$Normal : \epsilon_i \sim 10 * N(0, 1);$$

$$Beta : \epsilon_i \sim 50 * [Beta(5, 3) - 5/8].$$

The methodology described in Section 3.3.4 is used to estimate the crop yield density and Jackknife variance for each MC replicate.

Figure 3.3 shows the MC mean of the kernel density estimator for each of the four set ups at fixed z_i values 120, 160, 200, 240, and 280. Dotted lines are the 2.5% and 97.5% percentile values of the MC estimates with the true value being a solid black line. The model appears to fit the true yield density well, with the exception being near the edges of the z value. Overall, however, the results seem reasonable.

Figure 3.4 shows a coverage rate heat map for the kernel density estimator. Using the Jackknife standard error and $\hat{f}(w|z)$ estimate, a 95% confidence interval was constructed for each MC replicate and the percentage of times that the true value was within the confidence interval is given on the heat map. Lighter colors are values close to the nominal coverage

rate of 95%. The x-axis run along standardized (mean centered) yield, $y_i - \mu(z_i)$ while the y-axis run along the simulated land quality (APH) values, z_i.

The nominal coverage tends to be around 95% when the linear mean structure is used. Problems do appear in the quadratic mean structure set ups where undercoverage becomes prevalent near the outer edges of the land quality values. Univariate penalized B-Spline fits with a quadratic penalty will tend to enforce linearity beyond boundary knots (Hastie and Tibshirani, 1990). This is exactly the trend seen in the quadratic mean structure of the heat maps. However, as seen in figure 3.3, specifically the 10^{th} and 90^{th} percentile graphs,the overall mean estimate does well even at values close to the data edge. This suggests that the Jackknife variance estimation of the kernel density estimator may be off.

3.5 Empirical Study

In this section the penalized BTPB is applied to crop insurance data collected in 2009 from Webster County, Iowa to investigate if the premium rates are actuarially fair. In Section 3.5.1, the penalized BTPB derivative estimators is applied to the premium data to estimate the 2009 conditional yield density. Then, in Section 3.5.2 the kernel density estimator is applied to the actual historical yield data, which runs through 2009, to estimate the corresponding conditional yield density. These two curves are then compared. If the premium rates are actuarially fair, these two curves inferred from two different data sources should agree. In Section 3.5.3, implications from the findings is discussed.

3.5.1 Application to Crop Insurance Data

To estimate the conditional density crop yield curve based on premium data, the unit-level yield insurance record data of corn for Webster County, Iowa in 2009 is used. These data are maintained by RMA. The individual insurance records contain detailed information of the insured unit, including land quality (z), coverage rates (x), actual premium prices (y), and other variables such as location, production practices and insurance choices. While

multiple coverage level choices are available, typically from 50% to 85% in 5% increments, premiums and subsidies purchased by the farmer are the only rates observed for the corresponding unit. Therefore, per acre insurance premiums and subsidies are reconstructed for all coverage levels (50%-85%) for individual insurance units in the sample following the rules established by the RMA (see Du et al. (2017) for more details). Land quality is determined by historical yield (APH) which take on values between 130 and 210 bushels per acre. There are 388 land units with 8 reconstructed per acre insurance premiums and subsides per unit giving $n = 3104$. This data was then normalized to the compact set $[0, 1]$ by subtracting off the minimum value (130) and dividing by the range (80).

To estimate $\hat{f}(w|z)$, tuning parameters are chosen for the same reasons given in Section 3.4. A B-spline base of degree $d_1 = 5$ with $S_1 = 8$ bins and $K_1 = 2$ knots is used for variable x_i. For variable z_i, a B-spline base of degree $d_2 = 3$ with $S_2 = 80$ bins and $K_2 = 20$ knots is used. The smoothing parameter λ_i's are determined by GCV, and $m = 3$ is used as the degree of penalty. The value of $p = 4$ is used and is determined by the 2009 APH market price (Theisse, 2009).

Figure 3.5 plots the estimated curve $\hat{f}(w|z)$ for five land quality values at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of z_i. In each subplot, the black solid line with black dotted bands represent the penalized BTPB estimator $\hat{f}(w|z)$ as well as its 95% confidence band based on its standard error in (3.25). Again, $T = L = 1,000$ is used for the standard error calculation. A fan effect is observed in all subplots of Figure 3.5. This is because higher yield w implies a bigger value for coverage rate x, see (3.12) where $x = w/z$. Additionally the estimated function $\hat{\sigma}^2(x, z)$ from the real data increases as the coverage rate x goes up, causing the variance estimator for the derivatives of the penalized BTPB to increase according to (3.22).

3.5.2 Application to Historical Yield Data

To estimate the conditional density yield curve based on yield data, crop insurance unit-level corn yield data for Webster County from 1990 to 2009 are obtained from RMA/USDA. The yield data contain up to 10 years of yield history for each unit insured under the federal crop insurance program over the sample period. The yield history is not consecutive for some units. There are 4318 units with $35,679$ total individual year observations.

Crop yields tend to increase over time due to technological advances and improved farming practices (Tannura et al., 2008; Xu et al., 2012). Therefore, a slight modification has to be applied to the current year yield model (3.26) in order to apply the kernel density estimator. This is done by adding cubic covariates that account for the increasing trend to the yield model used for estimation. This allows for multiple years of data to be used as the response variable w_{it} in the kernel density estimation. This means that instead of using the yield model from 3.26, the working yield model used is

$$w_{it} = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \mu(z_i) + \sigma(z_i)\epsilon_i \tag{3.27}$$

where w_{it} is the yield for unit i in year t and $\tilde{t} = 2009 - t$. Note that the years run backwards because there is a known ending date (2009) but varying starting dates. Different definitions of t will only affect the β values and not the overall fit of the model. Due to the addition of the β parameters, the estimation method from Section 3.3.4 needs to be adjusted and is given as follows.

(i) First, estimate the mean function $\mu(z_i)$ using univariate penalized B-spline. A cubic spline base $\mathbf{B}(z)$ constructed as

$$\mathbf{B}(z_i) = (B_{-d_2+1,d_2}(z), B_{-d_2,d_2}(z), \cdots, B_{K_2+1,d_2}(z))^T. \tag{3.28}$$

is augmented with the base $\mathbf{T} = [1, \tilde{t}, \tilde{t}^2, \tilde{t}^3]^T$ for time. Then the coefficients for $\mathbf{B}(z_i)$ and β_j's ($j = 0, 1, 2, 3$), can be estimated using the univariate penalized spline method by minimizing the function

$$[\hat{\mathbf{a}}, \hat{\boldsymbol{\beta}}] = \underset{\mathbf{a},\mathbf{b}}{\operatorname{argmin}} ||\mathbf{Y} - (\mathbf{T}\boldsymbol{\beta} + \mathbf{B}(z)\mathbf{a})||_F^2 + \lambda \mathbf{a}^{\mathbf{T}}\mathbf{D}\mathbf{a}, \tag{3.29}$$

where $\hat{\boldsymbol{\beta}}$ are the fitted coefficients for the cubic covariates and $\hat{\mathbf{a}}$ are the fitted coefficients for the B-spline basic. $\lambda \mathbf{a}^{\mathbf{T}}\mathbf{D}\mathbf{a}$ is a penalty that prevents overfitting.

(ii) Then obtain residuals $r_{it} = w_{it} - [\hat{\beta}_0 + \hat{\beta}_1 \tilde{t} + \hat{\beta}_2 \tilde{t}^2 + \hat{\beta}_3 \tilde{t}^3 + \hat{\mu}(z_i)]$ after estimating $\hat{\mu}(z_i)$. A penalized cubic B-spline is used to estimate the variance function $\hat{\sigma}^2(z_i)$ with r_{it}^2 as the response and z_i as the covariate.

In both step (i) and (ii), ten equally spaced knots are used, the degree of penalty $m = 3$, and the smoothing parameter λ is prescribed by GCV.

(iii) Standardize the residuals to get $\hat{\epsilon}_{it} = r_{it}/\hat{\sigma}(z_i)$, and use a kernel density estimator with Gaussian kernel to obtain the density function for $\hat{f}_\epsilon(\epsilon)$. Then the final estimator for the current yield density (t=2009) conditional on z is $\tilde{f}_{w|z}(w|z) = \hat{f}_\epsilon\left(\dfrac{w_i - \hat{\beta}_0 - \hat{\mu}(z)}{\hat{\sigma}(z)}\right)/\hat{\sigma}(z)$.

The bandwidth in step (iii) is determined by Sheather-Jones bandwidth selection (Sheather and Jones, 1991).

(iv) To calculate $\widehat{Var}\left(\tilde{f}_{w|z}(w|z)\right)$, the delete one Jackknife method is used where

$$\widehat{Var}\left(\tilde{f}_{w|z}(w|z)\right) = \frac{n-1}{n}\sum_{b=1}^{n}\left[\tilde{f}_{w|z}^{[b]}(w|z) - \tilde{f}_{w|z}(w|z)\right]^2$$

where $\tilde{f}^{[b]}(w|z)$ is the estimate of $\tilde{f}(w|z)$ with the b^{th} observation removed.

Figure 3.6 reports the estimated kernel density $\tilde{f}_{w|z}(w|z)$ for five fixed land quality values at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of z_i. This figure compares favorably with the results in Zhang (2017), which uses a density ratio estimator of crop yield distributions on data consisting of annual average county corn yields for all ninety-nine Iowa counties

from 1950 through 2010. The small confidence bands in the figure are due to the large sample in the yield data set.

For each subplot of Figure 3.5, the part of the kernel density curve $\tilde{f}_{w|z}(w|z)$ corresponding to the restricted range determined by the penalized BTPB method using premium data and its 95% confidence band is included. See the gray solid lines and gray dashed upper and lower bands in Figure 3.5. It is noticeable that the two curves inferred from two different data sources do not match. Moving across the panels Figure 3.5, the yield-inferred density curve (gray) is higher than the premium-inferred density curve (black) for low quality lands (135 APH through 145 APH). However, it tends to move lower than the yield-inferred curve tends to gets lower than the premium-inferred curve when the land quality increases (180 APH and above). According to (3.2), for a given coverage rate x and land quality z, a higher conditional distribution function leads to a higher premium price. So the statistical comparison results in Figure 3.5 imply that the premium rate setting favors owners of less productive land at the expense of those who own higher land quality. This seems to be true for lower and median coverage rates. For higher coverage rates that correspond to higher yield values (since $w = xz$), the gray curves always exceed the black curves. In next subsection, further discussion takes place about the practical implications we can draw from our findings on premium rate determination based on this statistical comparison results in Webster County.

3.5.3 Discussion

As previously mentioned, most crop insurance products offered in the United States receive generous federal subsidies. The concern here is not with any welfare losses and redistributions arising from the presence of subsidies, see, e.g., Wright (2014) or Lusk (2016). Taking subsidies as given, it is of interest to ask what implications might flow from this analysis of premium rate determination in Webster County. The findings on premium rate determination mentioned in Section 3.5.1 can be seen more clearly in Figure 3.7, which

shows the yield densities converted back to premium values for fixed values of land quality according to (3.2). In Figure 3.7, the gray lines are the premium prices calculated using (3.2) and the estimated conditional kernel density using yield data, and the dots are the observed premium prices at the coverage rate value within 2.5 bushels of the given land quality value. The interval is used in order to produce a reasonable number of data points in the graphs. Moving across the panels in Figure 3.7, it can be seen that when yield quality is low (135 APH through 145 APH), then actual premiums set by government actuaries are lower than yield-inferred premiums whereas when yield quality is high (180 APH and above) then the reverse is true.

There are several possible implications, even for a good growing region such as Webster County. Some controversy surrounds the land use response to the availability of crop insurance and to product subsidies. Weber et al. (2016) inferred that large-scale expansion of U.S. crop insurance programs during 2000-2013 had minimal effect on how much United States land was in cultivation, what was grown or how it was grown. However, most recent studies have discerned impacts. Focusing on grassland-intensive areas of the Northern Great Plains, both Feng et al. (2013) and Miao et al. (2012) established that crop insurance subsidies have increased cropland acreage. For the Cornbelt region, Claassen et al. (2017) found evidence that crop revenue insurance has slightly increased the conversion of non-cropland to cropland in addition to a somewhat larger impact on crop choices. Also in a U.S. national study, Yu et al. (2017) have estimated a 0.43% acreage response to a 10% change in crop insurance subsidy rates. Most of the additional land would be of low productivity.

For Webster County, and separate from formal subsidies, an effective 'subsidy and tax' that is channeled through premium rate-setting has been identified, where lower productivity land is 'subsidized' and higher productivity land is 'taxed'. So one implication of the findings is that, in receiving this implied subsidy through government imposed insurance rates, corn production is encouraged on lower productivity land within the county. These

lands tend to be more fragile and erosion-prone. Another implication is that, given the magnitude of the premium gap, owners of better quality land may choose not to insure their land. Without them, the pool of insured risks tilts heavily toward those who expect to receive more indemnities and program costs to the taxpayer may become large. In order to stabilize the program, premiums may have to increase with the result that intermediate quality land drop coverage in the manner of Akerlof (1970)'s argument about adverse selection and insurance market unraveling.

How subsidies affect the findings is the next discussion of interest, in particular the motivation for these subsidies. It is partly to prevent adverse selection that premium subsidies are provided in U.S. Federal crop insurance programs. Then the question becomes, are subsidies sufficient to overcome the actuarial bias?

For the years in question and for yield insurance offered to 'basic' and 'optional' units, subsidy rates are found in table 3.9 along the first row, i.e. 67% at the 50% coverage rate (Shields, 2009). Although the subsidy declines as coverage rate increases, it is generally true that the product of subsidy fraction and premium increases as coverage increases (Du et al., 2017). Figure 3.8 presents Figure 3.7 once more, with the exception that premiums have been subsidy-adjusted. These adjusted premiums, found in Table 3.9 are the ones that farmers pay, where government subsidies are transferred separately to vendors acting on the government's behalf. Under the subsidies, crop insurance becomes an even better value across all land units, where the gray curve is above all subsidy-adjusted premiums when land quality equals either 135 or 145 APH. However, the gray curve splits the net-of-subsidy premium data for intermediate APH land. For the land quality 190 APH, the only data below the curve are for the highest coverage levels while for 195 APH, no adjusted premiums are below the curve. As is the case for unsubsidized markets, for these land units the premium subsidies are insufficient to ensure that expected indemnities to be collected exceed net premiums paid. These higher productivity lands, which may have deep soils with high organic matter suitable for water and nutrient storage, are generally resilient to

weather-related crop stressors (Du et al., 2018) and so have low yield variability. The land units may be insured if the operator perceives risk to exceed this estimated risk, is very risk averse, or is required to insure by a creditor. But incentives to insure are lower than would be the case were the net premium below the gray curve.

Data from Figure 3.8 are summarized in Table 3.1 which reports the percentages of after-subsidy premiums that fall below the gray line estimate of 'actuarially fair rate'. Higher percentages imply higher motivation for taking out insurance. The overall percentage of 50.6% indicates that premium rates are close to the estimates on average. Moving down each coverage rate column, for any coverage rate as land quality increases then the percentages of after-subsidy premiums that fall below the gray line declines almost uniformly. Moving across coverage rates in the last row, the percentage generally increases on average. Thus, even after subsidies, the motivation for taking out insurance declines as land quality increases and increases as coverage rate increases. If crops from more productive land are to be insured at all then it will only be at higher coverage levels. This is consistent with what the lower subplots of Figure 3.5, where for higher coverage rates (corresponding to higher yield values) the gray curves are higher than the black curves in more productive lands (APH of 180 or above). Du et al. (2013) have shown that operators in more productive areas should prefer higher coverage levels even when crop insurance premiums are subsidized but actuarially fair. While also showing that higher productivity lands do indeed tend to have higher coverage, they did not address participation.

These findings are notable because land quality in Webster County is quite uniform when compared with other counties, even in Iowa. So there is evidence of rates-setting problems in a place that is a poor candidate for manifestation of such problems. It remains to be seen whether the patterns observed in this county apply more generally.

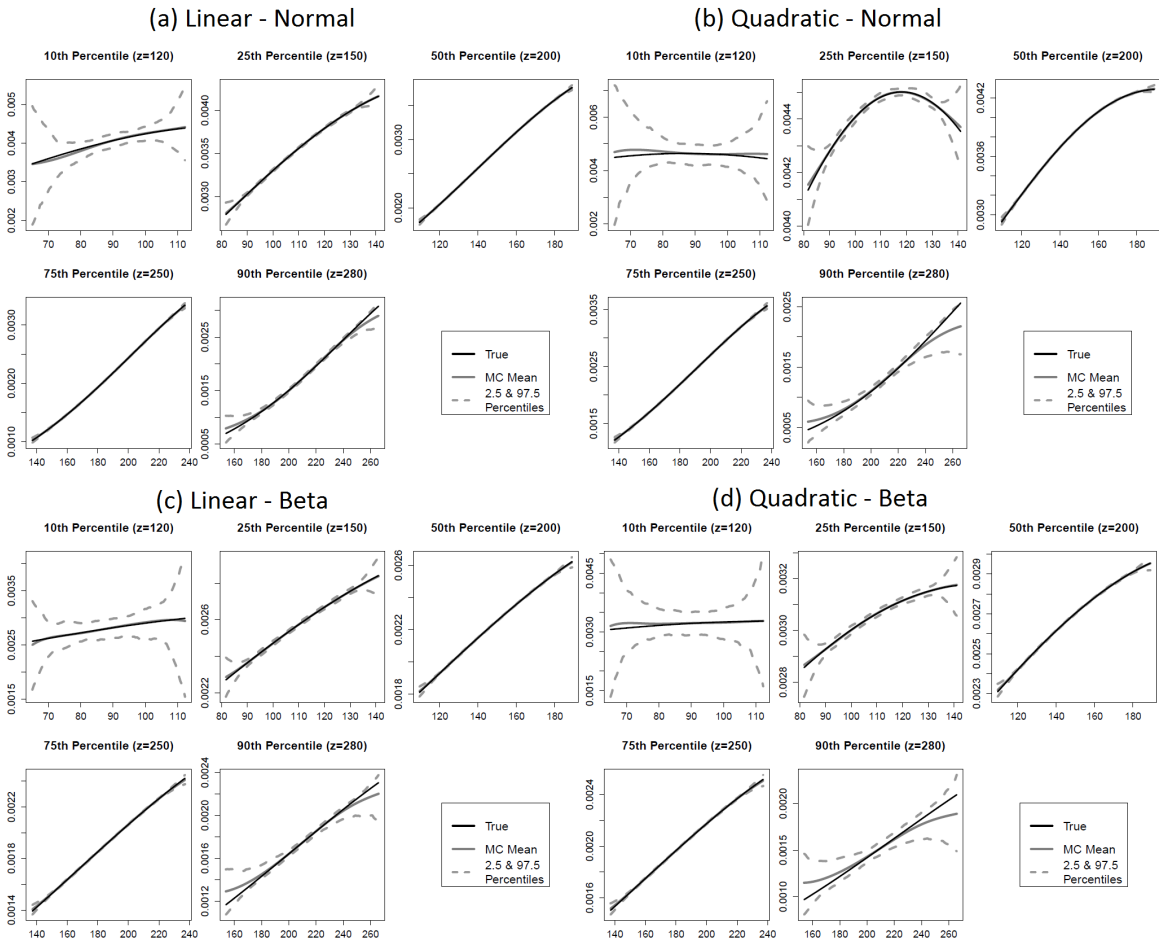

Figure 3.1 Monte Carlo means and the 2.5^{th} and 97.5^{th} percentiles of the MC estimates for $\hat{f}(w|z)$ using the penalized BTPB defined in (3.18) for all four scenarios prescribed by two mean functions and two error distributions respectively: Linear-Normal, Linear-Beta, Quadratic-Normal, Quadratic-Beta. Five subplots in each panel give the MC results for five fixed land quality values at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of z_i. The black line is the true function $f(w|z)$, the gray line is the mean of the MC estimated $\hat{f}(w|z)$, and the lower and upper dotted lines are the 2.5^{th} and 97.5^{th} point-wise percentiles of the MC estimates.

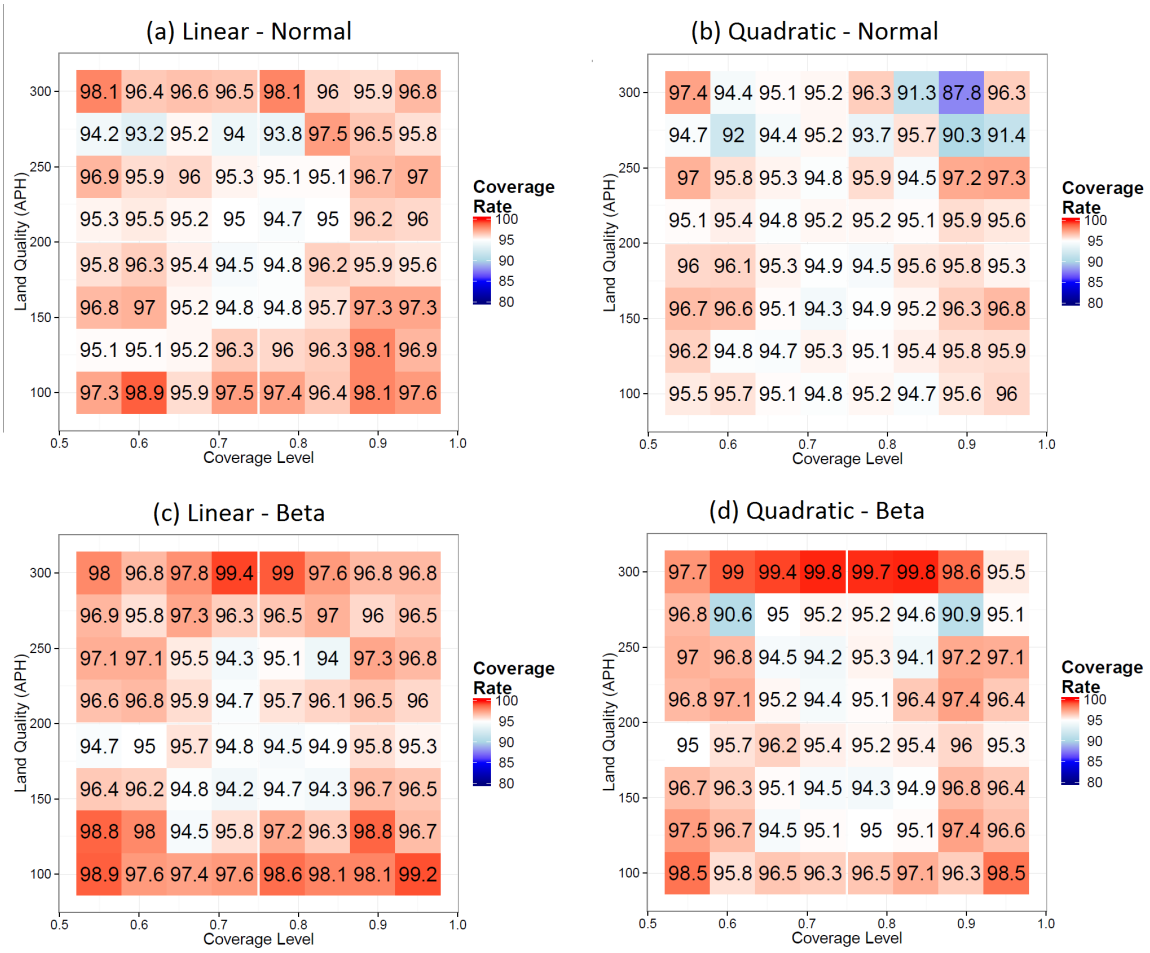

Figure 3.2 Coverage probability heat maps of the 95% confidence intervals for the conditional yield density estimator $\hat{f}(w|z)$ using the penalized BTPB in the four scenarios respectively. Using the asymptotic normality proved in Theorem 2 and the standard errors proposed in (3.25), a point-wise 95% confidence band is constructed for each MC replicate and the percentage of times that the true value is within the band is given on the heat maps. The x-axis represents the insurance coverage levels x_i, while the y-axis represents the land quality z_i. Lighter colors indicate values close to the nominal coverage rate of 95%.

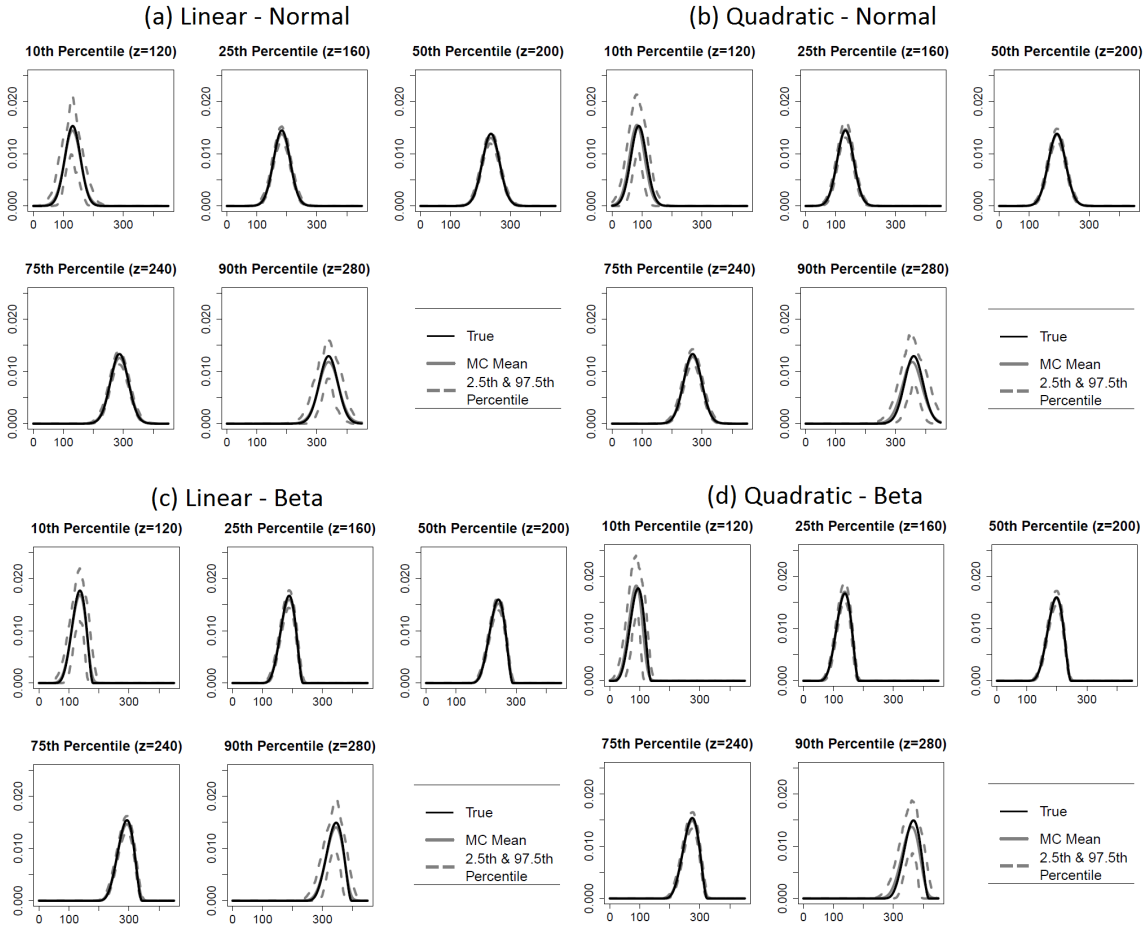

Figure 3.3 Monte Carlo means and the 2.5^{th} and 97.5^{th} percentiles of the MC estimates for $\tilde{f}(w|z)$ using kernel density estimation for all four scenarios prescribed by two mean functions and two error distributions respectively: Linear-Normal, Linear-Beta, Quadratic-Normal, Quadratic-Beta. Five subplots in each panel give the MC results for five fixed land quality values at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of z_i. The black line is the true function $f(w|z)$, the gray line is the mean of the MC estimate $\tilde{f}(w|z)$, and the lower and upper dotted lines are the 2.5^{th} and 97.5^{th} point-wise percentiles of the MC estimates.

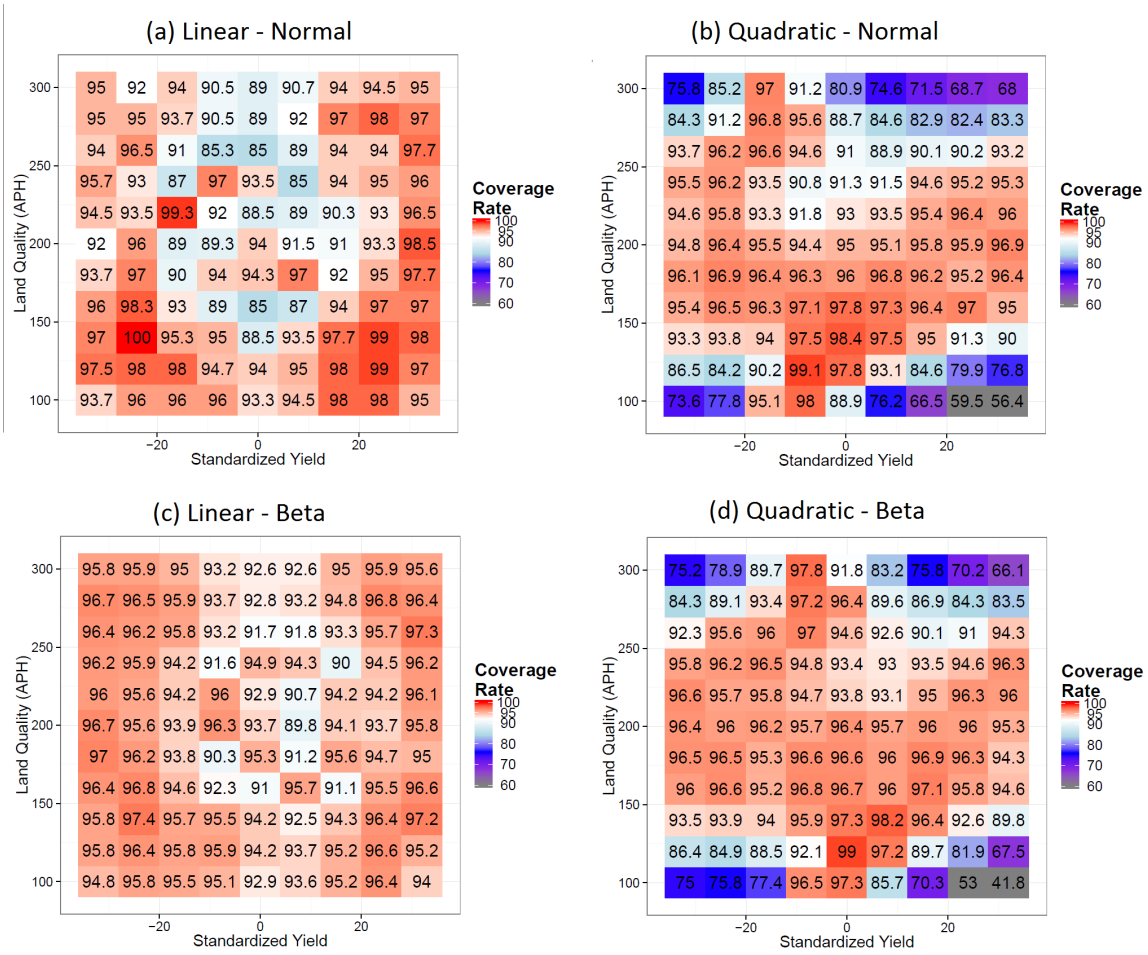

Figure 3.4 Coverage probability heat maps of the 95% confidence intervals for our conditional yield density estimator $\tilde{f}(w|z)$ using the kernel density estimation method in the four scenarios respectively. Using the delete one jackknife method and the proposed standard errors, a point-wise 95% confidence band is constructed for each MC replicate and the percentage of times that the true value is within the band is given on the heat maps. The x-axis represents the insurance coverage levels x_i, while the y-axis represents the land quality z_i. Lighter colors indicate values close to the nominal coverage rate of 95%.

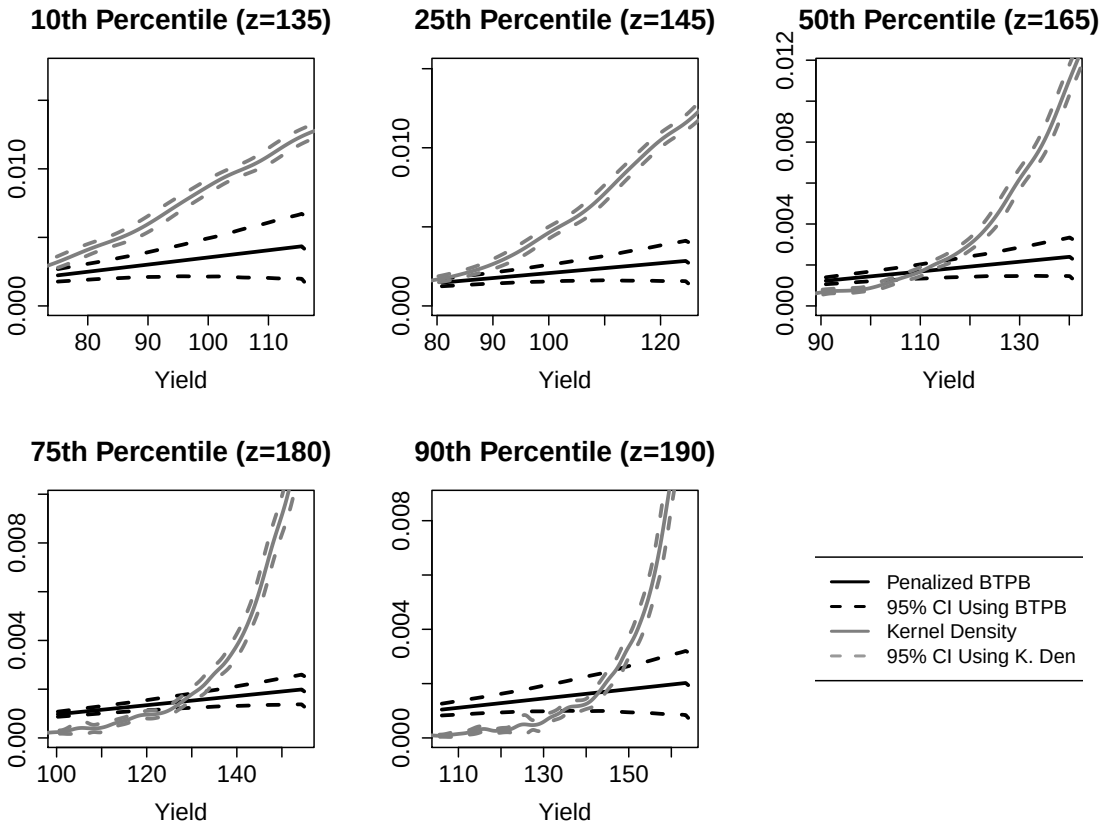

Figure 3.5 Comparison of corn yield densities in Webster County, Iowa for 2009 using different estimation methods. The black line is the penalized BTPB estimate using premium data and the black dotted lines are its 95% confidence intervals based on its asymptotic normality. The gray line represents the kernel density estimator using yield data and the gray dotted lines are 95% confidence intervals based on its Jackknife variance estimator.

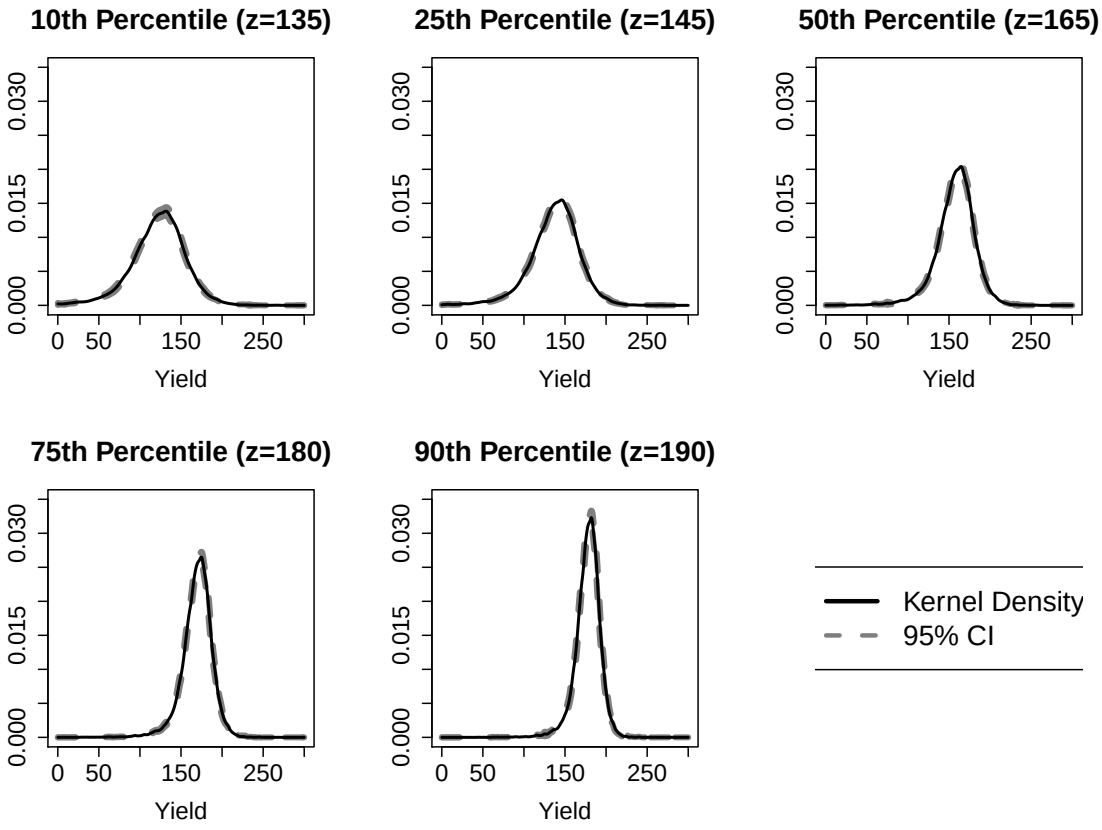

Figure 3.6 Corn yield density estimation using kernel density estimator $\tilde{f}(w|z)$. Black solid curve is the kernel density estimator using yield data, and the gray lines are its 95% confidence intervals using its Jackknife variance estimator. Five subplots are for five fixed land quality values at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of z_i.

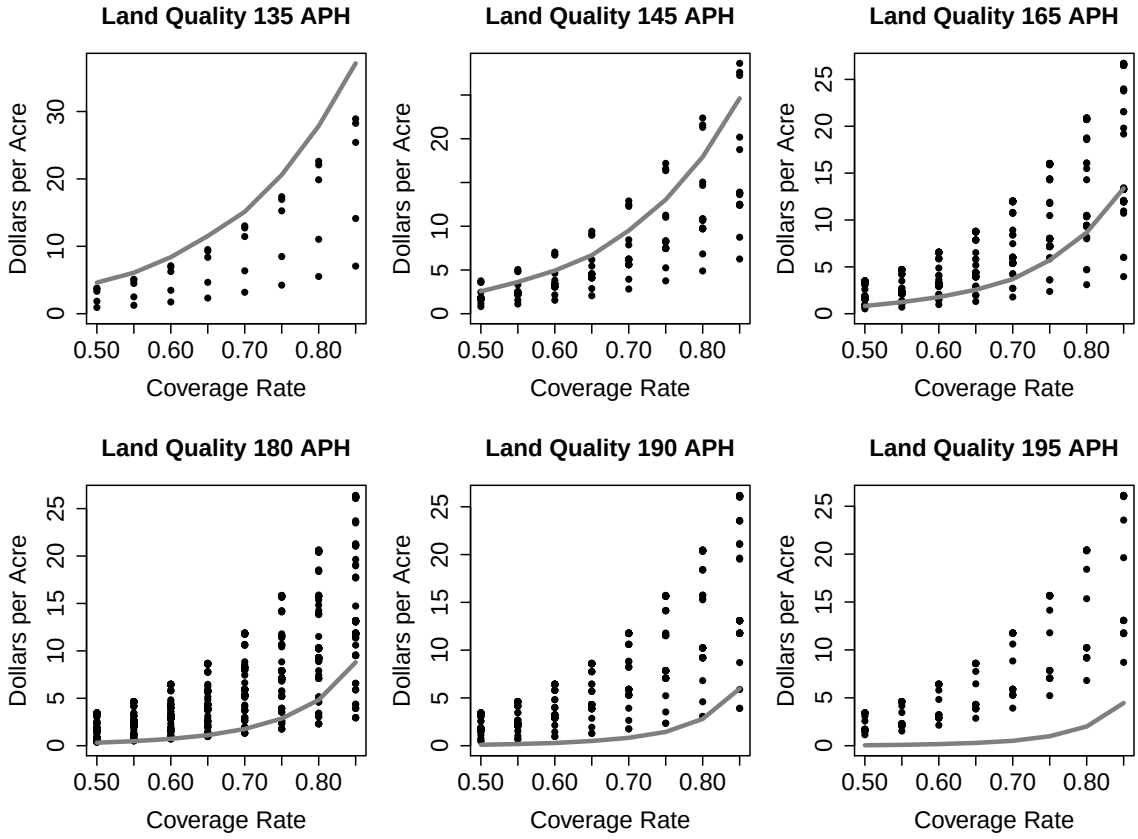

Figure 3.7 Actual Premiums and Estimated Premiums. Dots are the actual premium values for a given land quality based of coverage rate. The gray lines are the premium prices calculated using equation (3.2) and the estimated conditional kernel density $\tilde{f}(w|z)$ using yield data.

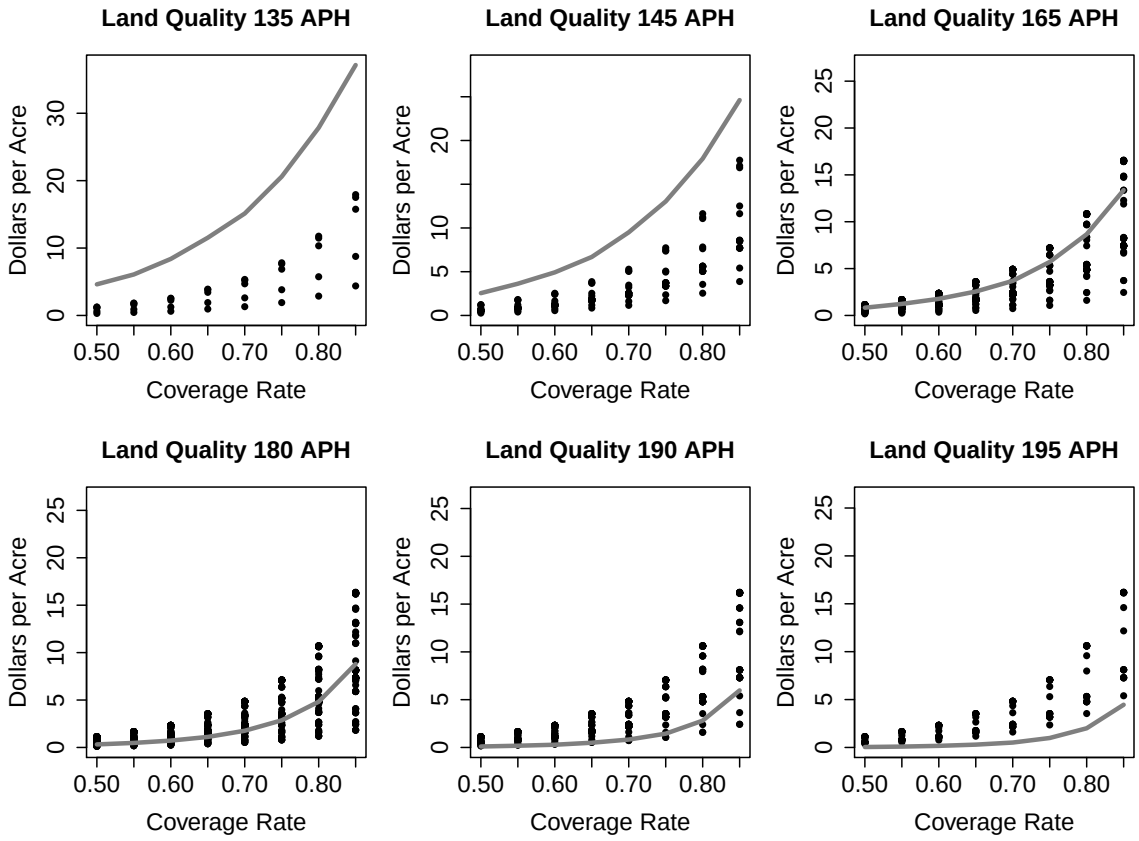

Figure 3.8 Subsidy-adjusted Premiums and Estimated Premiums. Dots are the subsidy-adjusted premium values for a given land quality based of coverage rate according to the subsidy rates in Table 3.9. The gray lines are the premium prices calculated using equation (3.2) and the estimated conditional kernel density $\tilde{f}(w|z)$ using yield data.

Figure 3.9 This table shows the subsidy rates set into law in the 2008 farm bill. These subsidy rates are in the first row. Enterprise units (row 2) and whole farm units (row 3) are also included. CAT means catastrophic and all numbers are percentages. The data in Table 3.1 and Figure 3.8 use this table to adjust the premium values. 50% guarantee gets a premium subsidy of 67%, while 55% guarantee gets a subsidy of 64% and so on Shields (2009).

Crop Insurance Premium Subsidies

(government-paid portion of premium as a percent of total premium)

Yield coverage level (%)	**CAT**	**50**	**55**	**60**	**65**	**70**	**75**	**80**	**85**
Premium subsidy (%) for most polices (including those using basic and optional units)	100	67	64	64	59	59	55	48	38
Premium subsidy (%) for enterprise units		80	80	80	80	80	77	68	53
Premium subsidy (%) for whole farm units					80	80	80	71	56

Table 3.1 Table that calculates the share of subsidized premiums that fall below the gray line by land quality and coverage rate. The second number is the total number of observations. Observations are rounded to their nearest five bushels to determine land quality. The marginal distribution for each coverage level across all land quality levels is the right most column, and the marginal distribution for each land quality across all coverage levels is the bottom row. The overall share is 50.16.

Cov Rate	.50		.55		.60		.65		.70		.75		.80		.85		Average	
L Quality	Pct	#	Pct	#	Pct	#	Pct	#	Pct	#	Pct	#	Pct	#	Pct	#	Pct	#
135	100	5	100	5	100	5	100	5	100	5	100	5	100	5	100	5	100	40
145	100	9	100	9	100	9	100	9	100	9	100	9	100	9	100	9	100	72
165	66.67	30	66.67	30	66.67	30	66.67	30	66.67	30	66.67	30	66.67	30	66.67	30	66.67	240
180	17.91	67	16.42	67	17.91	67	16.42	67	17.91	67	17.91	67	49.25	67	68.66	67	27.80	536
190	0	29	0	29	0	29	0	29	6.9	29	6.9	29	6.9	29	6.9	29	3.45	232
195	0	9	0	9	0	9	0	9	0	9	0	9	0	9	0	9	0.00	72
Average	44.07	388	44.07	388	44.85	388	44.33	388	49.74	388	53.09	388	58.76	388	62.37	388	50.16	3104

CHAPTER 4. GENERALIZED METHOD OF MOMENTS ESTIMATORS FOR MULTIPLE TREATMENT EFFECTS USING OBSERVATIONAL DATA FROM COMPLEX SURVEYS

In this chapter, a generalized method moments (GMM) estimator is considered to estimate treatment effects defined through estimation equations using an observational data set from a complex survey. It is demonstrated that the proposed estimator, which incorporates both sampling probabilities and semiparametrically estimated self-selection probabilities, gives consistent estimates of treatment effects. The asymptotic normality of the proposed estimator is established in the finite population framework, and its variance estimation is discussed. In simulations, our proposed estimator and its variance estimator based on the asymptotic distribution are evaluated. This method is then used to estimate the effects of different choices of health insurance types on health care spending using data from the Chinese General Social Survey. The results from the simulations and the empirical study show that ignoring the sampling design weights might lead to misleading conclusions.

4.1 Introduction

Observational data from a complex survey has increasingly become useful for causal inference because they can provide timely results with low cost. Survey data contains information on the treatment selections, which enables the estimation of treatment effects that cannot feasibly be evaluated with a randomized trial. In a survey, a treatment can be broadly defined as one of the survey questions, e.g. whether or not an individual has quit smoking, how often an individual does a physical exam, or what types of health insurance an individual has chosen. The existing survey data can be used to estimate effects of those treatments on health care spending, even if the health behavior or the health insurance

enrollment of an individual cannot be randomized. Also, because a well-designed survey sample is often a good representative of the target population, the treatment effect results can be generalized to the target population level if the survey weights are appropriately incorporated. Propensity score methods are well-established statistical methods that remove treatment selection bias in observational studies if the selection probability model is correctly specified (Rosenbaum and Rubin, 1983). Many observational data sets have multiple treatment options. In order to handle the complexity in multiple treatment groups, theoretical results support using the inverse of the estimated treatment selection probabilities as weights to adjust for selection bias and attain asymptotic efficiency (Hahn, 1998; Hirano et al., 2003; Cattaneo, 2010). This kind of estimator is called inverse probability weighted (IPW) estimator, and the estimated selection probabilities are called propensity scores. IPW estimators are used to address the potential confounding in observational studies as well. However, it is very common that people ignore survey weights in observational data when using the IPW estimators, yet claim that the estimated treatment effects are generalizable to the target population, causing misleading guidance in causal inference. Failure to properly account for the complex survey design may lead to biased treatment effect estimates and incorrect variance estimation.

Several authors have emphasized the importance of incorporating survey weights in their IPW estimators, e.g. DuGoff et al. (2014), Zanutto (2006), Ashmead (2014), and Ridgeway et al. (2015). The general idea is to multiply the inverse of the estimated propensity scores by the sampling design weights. Most of the papers, except for Ashmead (2014), do not provide theoretical justification for such survey adjusted estimators, and variance estimation is seldom discussed. Yu et al. (2013) proposes a semiparametric two-phase regression estimator to estimate marginal mean treatment effects in observational data sets from complex survey designs. This chapter considers a more general set up in which parameters of interest are defined through estimation equations, and uses the generalized method of moments (GMM) for parameter estimation. Similarly to Yu et al. (2013), a

connection is drawn between the two-phase sampling in survey statistics and the estimation of treatment effects from an observational database. The observational data set, denoted as A_1 (with size n), is considered as a first-phase sample from a finite population, according to a known sampling probability π_{1i} for subject i. The second-phase sampling is a partitioning of the first-phase sample (observational data set) into mutually exclusive and self-selected treatment groups, $A_{21}, ..., A_{2G}$, where G is the number of treatments. This partitioning in the second-phase can be viewed as a multinomial sampling in survey statistics, and its self-selection probabilities π_{2ig} for subject i into group g ($g = 1, ..., G$) can be estimated using the semiparametric approach in Cattaneo (2010).

The methodology in this chapter is different from DuGoff et al. (2014), Zanutto (2006), Ashmead (2014), and Ridgeway et al. (2015) in the following ways.

(i) Their papers consider two treatments, as opposed to the multi-level treatment selection presented here.

(ii) In their work, the propensity scores are estimated using a parametric linear logistic regression. Here the propensity scores, i.e. π_{2ig} are estimated through a semiparametric approach, specifically using P-splines. This setup should provide a more robust approach to the misspecification of the selection probability model.

(iii) In their work, the parameters of interest are treatment means. This chapter is interested in estimating treatment specific parameters defined through estimation equations. In addition to providing generality, defining parameters through estimation equations can facilitate variance estimation. For example, if a parameter is a function of means, such as correlation or domain mean (see more details in Section 4.2.1), the variance estimation of GMM estimators for such parameters can be easily calculated through the sandwich formula associated with the asymptotic variance for a GMM estimator. Ashmead (2014) also utilizes estimation equations in their weighting estimator.

This chapter's methodology also differs from Yu et al. (2013) in the following aspects. Yu et al. (2013) only focuses on estimating marginal treatment means, to estimate parameters defined through estimation equations (see $\hat{\boldsymbol{\theta}}_g^{(1)}$ in Section 4.2.3). This chapter proposes a second estimator to gain efficiency by incorporating the first phase and second phase means of covariates into the estimation equations (see $\hat{\boldsymbol{\theta}}_g^{(2)}$ in Section 4.2.3). This is similar to the effect of calibrating the second phase means of covariates to their first phase means seen in the optimal two-phase regression estimator discussed in Fuller (2011). Additionally, Yu et al. (2013) assumes sample missing at random (SMAR), which is commonly used in literature, while this chapter considers population missing at random (PMAR), the framework proposed in Berg et al. (2016) (see more details in Section 4.2.1. It makes sense to use PMAR assumption in the context of casual inference study using observation dataset. It is shown that when PMAR holds but SMAR fails, survey weights should be included in the estimation of π_{2ig}, i.e. the propensity scores.

Theoretical justification for the estimator is provided in a combined framework of a finite population and a superpopulation. Variance estimators are also proposed. The validity of the estimator is demonstrated through simulation studies, and it is shown that estimators that ignore design weights might be subject to biases. The feasibility of the method is explored using data from the Chinese General Social Survey to estimate the effects of different choices of health insurance types on health care spending. The rest of the chapter is organized as follows. Section 4.2 introduces the framework and the proposed estimators. Section 4.3 presents an asymptotic normality and variance estimation. Simulation studies and an empirical study are reported in Sections 4.4 and 4.5 respectively. Section 4.6 concludes.

4.2 Proposed Estimators

In this section, the estimators are introduced. Section 4.2.1 discusses the basic set-up, Section 4.2.2 introduces the semiparametric approach for estimating the self-selection probabilities, and Section 4.2.3 proposes the estimators.

4.2.1 Basic Set-up

Let U be a finite population with size N containing $(\mathbf{Y}, Z_i)$, where $i = 1, ..., N$ indexes a subject, Z_i is a covariate variable, and $\mathbf{Y} = [Y_{i1}, ..., Y_{iG}]^T$ is a vector of potential outcomes for G different treatments depending on covariate Z_i. Let δ_{1i} be the sampling indicator from the survey design, defined by $\delta_{1i} = 1$ if unit i is selected into A_1 and zero otherwise. Let π_{1i} and π_{1ij} be the first and second order inclusion probabilities of the sampling design, defined as,

$$[\pi_{1i}, \pi_{1ij}] = [Prob(\delta_{1i} = 1), Prob(\delta_{1i} = 1, \delta_{1j} = 1)].$$

Assume the sampling weights are appropriately adjusted for any non-response. If the weights are adjusted due to non-response, the method can be used with the provision that the variation from estimating $\hat{\pi}_{1i}$ is not taken into account. Let δ_{2ig} $(g = 1, ..., G)$ be the self-selection indicator of subject i selecting treatment g, defined by $\delta_{2ig} = 1$ if unit i selects treatment g and zero otherwise. The self-selection process leads to the partitioning in the second phase. Assume conditioning on a covariate X_i, the self-selection indicators $\boldsymbol{\delta}_{2i} = [\delta_{2i1}, ..., \delta_{2iG}]$ follow a multinomial distribution with probabilities,

$$\pi_{2ig} = Prob(\delta_{2ig} = 1 | X_i), \text{ for } g = 1, ..., G, \tag{4.1}$$

i.e. for any subject i,

$$\boldsymbol{\delta}_{2i} = [\delta_{2i1}, ..., \delta_{2iG}] \sim multinomial(1; \pi_{2i1}, ..., \pi_{2iG}),$$

where $\sum_{g=1}^{G} \pi_{2ig} = 1$ for any i, and $\boldsymbol{\delta}_{2i}$ is independent of $\boldsymbol{\delta}_{2j}$ for any subjects $i \neq j$. Here covariates Z_i and X_i can be totally different, or can have overlap. Separate notations are

used in order to emphasize that the outcome response variables $\mathbf{Y}$ and the self-selection indicators $\boldsymbol{\delta}_{2i}$ can depend on different sets of covariates. Discussion on how to identify Z_i and X_i practically is in Section 4.4. Both Z_i and X_i have compact supports and are observed in A_1. They are written to be univariate forms in order to reduce notation burden. It is straightforward to extend to multivariate covariates, which are considered in the simulation studies and the empirical study of this paper. Suppose that $(\mathbf{Y}, \delta_{1i}, \boldsymbol{\delta}_{2i}, X_i, Z_i); i = 1, ..., N$ are i.i.d. generated from a superpopulation ξ.

In the context of simple random sampling, a common missing at random (MAR) assumption is $\mathbf{Y} \perp \boldsymbol{\delta}_{2i} | (X_i, Z_i)$. With this MAR assumption, the selection bias can be removed by applying the propensity score method (Rosenbaum and Rubin, 1983; Hirano et al., 2003). However, in the context of a complex survey, unequal probabilities of sampling can complicate the relationship between $\mathbf{Y}$, (X_i, Z_i), $\boldsymbol{\delta}_{2i}$ and the sample inclusion indicator δ_{1i}. Even if

$$\mathbf{Y} \perp \boldsymbol{\delta}_{2i} | (X_i, Z_i), \tag{4.2}$$

holds for a specific superpopulation model,

$$\mathbf{Y} \perp \boldsymbol{\delta}_{2i} | \{(X_i, Z_i), \delta_{1i} = 1\}, \tag{4.3}$$

may not hold. Following Berg et al. (2016), assumption (4.2) is called population missing at random (PMAR), and assumption (4.3) sample missing at random (SMAR) to emphasize it depends on the realized sample (i.e. conditional on $\delta_{1i} = 1$). The SMAR has been used previously (Pfeffermann, 2011; Little, 1982). However, it is natural to consider PMAR in this context because the mechanisms underlying the selection propensity are conceptualized as inherent characteristics of the subjects in the population. For example, whether or not a person decides to stop smoking heavily depends on this person's perseverance and personality type; what types of insurance a person has chosen depends on the nature of this person's work. In these examples, the self-selection probabilities depend on subjects' inherent characteristics that have nothing to do with whether or not the subjects were selected into the survey that was typically designed for other general purposes. Berg et al.

(2016) also provides examples of situations in which PMAR may be considered reasonable. They argue that if both PMAR and SMAR hold, weights are not needed in their imputation model; however if PMAR holds but SMAR fails, it is necessary to include weights to produce consistent estimators. A situation in which PMAR holds while SMAR does not can arise if a design variable omitted from the first phase sample is related to both the sampling inclusion probabilities and the response variable. An example of such a design variable is location in a situation where design strata are functions of location, the location is correlated with the response variable, but the specific location is masked from the analyst because of concerns associated with confidentiality. Using Lemma 1 of Berg et al. (2016), two conditions of the sampling are identified and the self-selection mechanisms for which PMAR implies SMAR: (1) $\delta_{1i} \perp \mathbf{Y} | (X_i, Z_i), \boldsymbol{\delta}_{2i}$; or (2) $\boldsymbol{\delta}_{2i} \perp (\mathbf{Y}, \delta_{1i}) | (X_i, Z_i)$. The first condition states that the sampling mechanism is non-informative given covariates (X_i, Z_i) within all the second phase self-selected groups A_{2g}. The second condition states that the self-selection mechanism is independent of either $\mathbf{Y}$ or sample inclusion given (X_i, Z_i). Like Berg et al. (2016), it is suggested to include survey weights into the estimation of the self-selection probabilities π_{2ig} when SMAR fails (see Section 4.2.2). In the simulation studies, both non-informative sampling (Condition (1) above holds), and informative sampling (Condition (1) above fails) are considered.

The true parameter of interest, $\boldsymbol{\theta}_g^0$ $(g = 1, ..., G)$, is a d_θ-dimensional vector satisfying,

$$E[\mathbf{m}_g(Y_{ig}, Z_i; \boldsymbol{\theta}_g)] = 0, \tag{4.4}$$

in the superpopulation, where $\mathbf{m}_g(Y_{ig}, Z_i; \boldsymbol{\theta}_g)$, hereafter denoted as $\mathbf{m}_{ig}(\boldsymbol{\theta}_g)$ to save space. In addition to treatment marginal means, people might be interested in estimating treatment correlations or treatment domain means. For example in the empirical study, it is interesting to understand whether the correlations between annual medical expenditure and age (or household income) differ significantly across different health insurance type groups; or whether the means of annual medical expenditure for very sick people (domain means) are significantly different across health insurance type groups. The parameter defined through

(4.4) includes treatment correlations and treatment domain means as special cases. More specifically, if the parameter of interest is $\boldsymbol{\theta}_g^0 = [P_g, \mu_g, \sigma_g^2, R_g]^T$, where $P_g = Prob(Y_{ig} \leq C)$ for some C, $\mu_g = E(Y_{ig})$, $\sigma_g^2 = Var(Y_{ig})$ and $R_g = Corr(Y_{ig}, Z_i)$, then the estimation equation can be defined as,

$$\begin{aligned}
&\mathbf{m}_{ig}(\boldsymbol{\theta}_g) \\
&= [1_{Y_{ig} \leq C} - P_g, \quad Y_{ig} - \mu_g, \quad (Y_{ig} - \mu_g)^2 - \sigma_g^2, \quad (Y_{ig} - \mu_g)(Z_i - \mu_z) - R_g\sqrt{\sigma_g^2}\sqrt{\sigma_z^2}, \\
&\quad Z_i - \mu_z, \quad (Z_i - \mu_z)^2 - \sigma_z^2]^T.
\end{aligned} \tag{4.5}$$

If the parameter of interest is a treatment specific domain mean, $\boldsymbol{\theta}_g^0 = E(Y_{ig}|Z_i \leq C)$, then the estimation equation can be written as,

$$\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = [Y_{ig}1_{Z_i \leq C} - \boldsymbol{\theta}_g P_z, \quad 1_{Z_i \leq C} - P_z]^T. \tag{4.6}$$

In both examples, μ_z, σ_z^2 or P_z are nuisance parameters.

4.2.2 Semiparametric Estimation of π_{2ig}

Because of the difficulty in specifying a parametric form for π_{2ig} and the constraint, $\sum_{g=1}^{G} \pi_{2ig} = 1$, the semiparametric method in Cattaneo (2010) is used to estimate π_{2ig}. Let $\{r_k(X_i)\}_{k=1}^{\infty}$ be a sequence of known approximating functions, and assume that the generalized logit of π_{2ig} can be approximated by $R_K(X_i)^T\boldsymbol{\gamma}_{g,K}$ for $K = 1, 2, ...$, where $R_K(X_i) = [r_1(X_i), r_2(X_i), ..., r_K(X_i)]^T$ and $\boldsymbol{\gamma}_{g,K}$ is a vector of the real-valued coefficients of $R_K(X_i)$ for the g-th treatment selection. Let an estimator of the $K \times G$ matrix $\boldsymbol{\gamma}_K = [\boldsymbol{\gamma}_{1,K}, \boldsymbol{\gamma}_{2,K}, ..., \boldsymbol{\gamma}_{G,K}]$ be,

$$\widehat{\boldsymbol{\gamma}}_K = [\hat{\boldsymbol{\gamma}}_{1,K}, \hat{\boldsymbol{\gamma}}_{2,K}, ..., \hat{\boldsymbol{\gamma}}_{G,K}] = \underset{\boldsymbol{\gamma}_K | \boldsymbol{\gamma}_{1,K} = \mathbf{0}_K}{argmax} \sum_{i \in A_1} b_i w_{1i} \sum_{g=1}^{G} \delta_{2ig} log \left[\frac{e^{R_K(X_i)^T\boldsymbol{\gamma}_{g,K}}}{\sum_{g=1}^{G} e^{R_K(X_i)^T\boldsymbol{\gamma}_{g,K}}} \right], \tag{4.7}$$

where $w_{1i} = \pi_{1i}^{-1}$, and $\mathbf{0}_K$ represents a $K \times 1$ zero vector used to constrain the sum $\sum_{g=1}^{G} \hat{\pi}_{2ig} = 1$. The estimated self-selection probabilities are

$$\begin{aligned} \hat{\pi}_{2ig} &= \frac{e^{R_K(X_i)^T \widehat{\boldsymbol{\gamma}}_{g,K}}}{1+\sum_{g=2}^{G} e^{R_K(X_i)^T \widehat{\boldsymbol{\gamma}}_{g,K}}} && \text{for g=2,3,...,G} \\ &= \left(1 + \sum_{g=2}^{G} e^{R_K(X_i)^T \widehat{\boldsymbol{\gamma}}_{g,K}}\right)^{-1} && \text{for g=1.} \end{aligned} \tag{4.8}$$

This solution is that of multinomial logistic regression where the probability for each g is approximated using a linear combination of the series of the approximating functions $R_K(X_i)$. Condition II in Appendix B1 specifies assumptions about $R_K(X_i)$, π_{2ig} and K to ensure $\hat{\pi}_{2ig}$ converges to π_{2ig} fast enough. Examples of $R_K(X_i)$ include a cubic polynomial basis, $R_K(X_i) = [1, X_i, X_i^2, X_i^3]^T$, or a quadratic spline basis with q knots $R_K(X_i) = [1, X_i, X_i^2, (X_i - \kappa_1)_+^2, ..., (X_i - \kappa_q)_+^2]^T$ where $(t)_+ = t$ if $t > 0$ and 0 otherwise, and $\kappa_1, ..., \kappa_q$ are knots in the compact support of X_i.

The b_i in (4.7) is a user-specified constant that represents the properties of the sampling and the self-selecting mechanism. As discussed in Section 4.2.1, PMAR assumption does not necessarily imply SMAR assumption. If one believes SMAR assumption holds, then one can set $b_i = w_{1i}^{-1}$, which leads to unweighted estimation of $\hat{\pi}_{2ig}$. If SMAR is not satisfied, the unweighted estimator may lead to bias, and setting $b_i = 1$ is one way to attain an approximately unbiased estimator, see Berg et al. (2016) for further discussion of the choice of b_i. If it is difficult to verify SMAR assumption, so the conservative choice of $b_i = 1$ is suggested because it leads to consistent estimators under PMAR without requiring SMAR.

4.2.3 Proposed Estimators

Since the true parameter of interest $\boldsymbol{\theta}_g^0$ is defined through an estimation equation in (4.4), the GMM method with propensity scores is used for estimation. It is common that people simply ignore the sampling design weights in the first-phase and calculate a naive estimator as,

$$\hat{\boldsymbol{\theta}}_g^{nw} = \arg\min_{\boldsymbol{\theta}_g} \left[\bar{\mathbf{m}}_g^{nw}(\boldsymbol{\theta}_g)\right]^T \left[\bar{\mathbf{m}}_g^{nw}(\boldsymbol{\theta}_g)\right], \tag{4.9}$$

where

$$\bar{\mathbf{m}}_g^{nw}(\boldsymbol{\theta}_g) = \frac{1}{n}\sum_{i\in A_{2g}} \frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\hat{\pi}_{2ig}}. \tag{4.10}$$

Here the superscript 'nw' means no weight. The estimator $\hat{\boldsymbol{\theta}}_g^{nw}$ ignores the sampling weights by applying equal weights to the estimation equations in (4.10). Although it uses the propensity score $\hat{\pi}_{2ig}$ to adjust for selection biases in the second-phase, it does not account for the survey design in the first-phase, which might lead to biases and incorrect variance estimation when estimating the treatment effect parameters on the population level. This is demonstrated in the simulation studies of Section 4.4. Both Ridgeway et al. (2015) and Yu et al. (2013) analytically quantify biases caused by ignoring the survey weights in complex survey.

In order to obtain a consistent estimator for $\boldsymbol{\theta}_g^0$, the first-phase survey weights need to be included into the estimation equation. The following GMM estimator is proposed,

$$\hat{\boldsymbol{\theta}}_g^{(1)} = \arg\min_{\boldsymbol{\theta}_g} \left[\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)\right]^T \left[\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)\right], \tag{4.11}$$

where

$$\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g) = \frac{1}{N}\sum_{i\in A_{2g}} w_{1i}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\hat{\pi}_{2ig}}. \tag{4.12}$$

In order to improve efficiency, one can incorporate the information from covariate Z_i that is potentially correlated with the outcome responses into the estimation equations. For this situation a second GMM estimator is proposed as,

$$(\hat{\boldsymbol{\theta}}_g^{(2)}, \hat{\mu}_z) = \arg\min_{(\boldsymbol{\theta}_g,\mu_z)} \left[\mathbf{H}_{ng}(\boldsymbol{\theta}_g,\mu_z)\right]^T \hat{\boldsymbol{\Sigma}}_{Hg}^{-1}(\boldsymbol{\theta}_g,\mu_z) \left[\mathbf{H}_{ng}(\boldsymbol{\theta}_g,\mu_z)\right], \tag{4.13}$$

where

$$\mathbf{H}_{ng}(\boldsymbol{\theta}_g,\mu_z) = [\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g), \quad \bar{z}_{2\pi g}(\mu_z), \quad \bar{z}_{1\pi}(\mu_z)]^T, \tag{4.14}$$

$$\bar{z}_{2\pi g}(\mu_z) = \frac{1}{N}\sum_{i\in A_{2g}} w_{1i}\frac{Z_i-\mu_z}{\hat{\pi}_{2ig}} \qquad \text{and} \qquad \bar{z}_{1\pi}(\mu_z)^T = \frac{1}{N}\sum_{i\in A_1} w_{1i}(Z_i-\mu_z). \tag{4.15}$$

$\hat{\mu}_z$ is an estimator for the nuisance parameter $\mu_z^0 = E(Z_i)$ and $\hat{\mathbf{\Sigma}}_{Hg}(\boldsymbol{\theta}_g, \mu_z)$ is the variance estimator of $\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)$, which depends on the joint inclusion probabilities and is defined in (4.38). The estimator $\hat{\boldsymbol{\theta}}_g^{(2)}$ in (4.13) is connected to a two phase sampling extension of the design unbiased difference estimator proposed by Särndal et al. (2003) and Breidt et al. (2005) when $\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$.

Remark 1: It can be shown that when $\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$ and $X_i = Z_i$, the estimator $\hat{\boldsymbol{\theta}}_g^{(1)}$ in (4.11) is asymptotically equivalent to the regression estimator proposed in Yu et al. (2013).

Remark 2: The estimator $\hat{\boldsymbol{\theta}}_g^{(2)}$ in (4.13) is more efficient than the estimator $\hat{\boldsymbol{\theta}}_g^{(1)}$ in (4.11). Appendix B3 provides a sketch of proof to show that $\hat{\boldsymbol{\theta}}_g^{(2)}$ is the most efficient estimator among the class of estimators $\hat{\boldsymbol{\theta}}_g^{a}$ that use any fixed positive definite matrix $\mathbf{A}$ in the quadratic form minimization, i.e. $\hat{\boldsymbol{\theta}}_g^{a}$ is defined as

$$(\hat{\boldsymbol{\theta}}_g^a, \hat{\mu}_z^a) = \underset{(\boldsymbol{\theta}_g, \mu_z)}{\arg\min} \left[\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)\right]^T \mathbf{A^{-1}} \left[\mathbf{H_{ng}}(\boldsymbol{\theta_g}, \mu_\mathbf{z})\right]. \tag{4.16}$$

If the matrix $\mathbf{A}$ is an identity matrix, then $\hat{\boldsymbol{\theta}}_g^{a}$ obtained in (4.16) is equivalent to $\hat{\boldsymbol{\theta}}_g^{(1)}$. Therefore $\hat{\boldsymbol{\theta}}_g^{(1)}$ is expected to be less efficient than $\hat{\boldsymbol{\theta}}_g^{(2)}$, which has been confirmed by the simulation studies in Section 4.4.

Remark 3: It can be shown that when $\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$, the estimator $\hat{\boldsymbol{\theta}}_g^{(2)}$ corresponds to the optimal two phase regression estimator discussed in Fuller (2011) (Theory 2.2.4). The optimality in Fuller (2011) is in terms of achieving the minimum variance for the limiting distribution of design consistent estimators of the form, $\bar{Y}_{2p,reg} = \bar{Y}_{2\pi} + (\bar{Z}_{1\pi} - \bar{Z}_{2\pi})\hat{\boldsymbol{\beta}}$, where $[\bar{Y}_{2\pi}, \bar{Z}_{2\pi}] = (\sum_{i\in A_2} \pi_{1i}^{-1}\pi_{2i}^{-1})^{-1} \sum_{i\in A_2} (\pi_{1i}^{-1}\pi_{2i}^{-1})[Y_i, Z_i]$, $\bar{Z}_{1\pi} = (\sum_{i\in A_a} \pi_{1i}^{-1})^{-1} \sum_{i\in A_a} \pi_{1i}^{-1} Z_i$, and π_{1i} (or A_1) and π_{2i} (or A_2) are the first phase and the second phase sampling probabilities (or samples). The efficiency gain of $\bar{Y}_{2p,reg}$ over $\bar{Y}_{2\pi}$ is similar to the effect of calibrating the second phase covariate mean $\bar{Z}_{2\pi g}$ to its first phase mean $\bar{Z}_{1\pi}$.

Remark 4: It can be shown that when $\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$ and $Z_i \equiv 1$, the estimator $\hat{\boldsymbol{\theta}}_g^{(2)}$ coincides analytically with the weighting estimator discussed in Ashmead (2014) except that the propensity scores in Ashmead (2014) are estimated using a parametric logistic

regression.

Remark 5: When the population mean of Z_i is available, the estimator $\hat{\boldsymbol{\theta}}_g^{(2)}$ can be easily extended to incorporate this additional information. For example, this case can occur when there are some demographic variables available on the population level. The extended estimator can be obtained by adding one more moment $\bar{z}_N(\mu_z) = N^{-1}\sum_{i\in U}(Z_i - \mu_z)$ into the $\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)$ in (4.14). Efficiency gain should be expected since this estimator uses more information on the population level. By viewing the problem as a two-phase sampling problem, the method can be readily extended to multiple sampling phases. This extension is useful because the database A_1 can come from a larger sample within the database. This case covers the common situations where detailed treatment and outcome data is available for only a subsample of the data such as a subsample with medical chart adjudication of claims records or a subsample constructed by merging multiple sources of claims records and electronic medical records.

4.3 Asymptotic Normality and Variance Estimation

Since $\hat{\boldsymbol{\theta}}_g^{(1)}$ can be written as a special case of $\hat{\boldsymbol{\theta}}_g^{(2)}$, in Section 4.3.1 only the asymptotic normal distribution for $\hat{\boldsymbol{\theta}}_g^{(2)}$ is derived. In Section 4.3.2 a linearized variance estimator for $\hat{\boldsymbol{\theta}}_g^{(2)}$ is provided. Section 4.3.3 gives a replication variance estimator for $\hat{\boldsymbol{\theta}}_g^{(1)}$.

4.3.1 Asymptotic Normality of $\hat{\boldsymbol{\theta}}_g^{(2)}$

The asymptotic normality of $\hat{\boldsymbol{\theta}}_g^{(2)}$ is established in Theorem 1 by combining two randomizations from the finite population level and the superpopulation level. For the finite population level, consider a sequence of samples and finite populations indexed by N, where the sample size $n \to \infty$ as $N \to \infty$ (Isaki and Fuller, 1982). To define the regularity conditions, use the notation $\mathcal{F}_N$ to represent an element of the sequence of finite population with size N. To distinguish between the two randomizations, "$|\mathcal{F}_N$" is used to indicate that the reference distribution is with respect to repeated sampling conditional on the finite

population size N. For example, $E(\cdot|\mathcal{F}_N)$ and $V(\cdot|\mathcal{F}_N)$ denote the conditional mean and variance with respect to the randomization generated from repeated sampling from $\mathcal{F}_N$. Use $E_\xi(\cdot)$, $Var_\xi(\cdot)$ and $Cov_\xi(\cdot,\cdot)$ to denote mean, variance and covariance with respect to the randomization from the superpopulation ξ. The proof of Theorem 1 uses a result given in Theorem 1.3.6 of Fuller (2011) that shows how to combine two asymptotic normalities from the finite population and the superpopulation levels. Because of the importance of this theorem to the results, it is stated here as Fact 1:

Fact 1 (Theorem 1.3.6 of Fuller (2011)): Suppose θ_0 is a true parameter on a superpopulation level, θ_N is its analogous part on a finite population level, and $\hat{\theta}$ is an estimator of θ_0 calculated from a sample. If $(\hat{\theta}-\theta_N)|\mathcal{F}_N \xrightarrow{\mathcal{L}} N(0,V_{11})$ *a.s.*, and $(\theta_N-\theta_0) \xrightarrow{\mathcal{L}} N(0,V_{22})$, then, $(\hat{\theta}-\theta_0) \xrightarrow{\mathcal{L}} N(0,V_{11}+V_{22})$. Here $(\hat{\theta}-\theta_N)|\mathcal{F}_N \xrightarrow{\mathcal{L}} N(0,V_{11})$ *a.s.* means that $\hat{\theta}-\theta_N$ converges in a distribution to a random variable with the distribution of $N(0,V_{11})$ almost surely with respect to the process of repeated sampling from the sequence of finite populations as $N\to\infty$. V_{11} is the asymptotic variance of $\hat{\theta}$ on the finite population level, while V_{22} is the asymptotic variance of θ_N on the superpopulaton level.

The key step in the proof of Theorem 1 is to obtain an asymptotic equivalence of $\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)$,

$$\begin{aligned}\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g) &= \tfrac{1}{N}\textstyle\sum_{i\in A_{2g}} \frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\hat{\pi}_{2ig}} \\ &= \tfrac{1}{N}\textstyle\sum_{i\in U} \frac{\delta_{1i}\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}} - \tfrac{1}{N}\sum_{i\in U}\frac{\delta_{1i}(\delta_{2ig}-\pi_{2ig})}{\pi_{1i}\pi_{2ig}} E_\xi(\mathbf{m}_{ig}(\boldsymbol{\theta}_g)|X_i) + o_p(n^{-1/2}).\end{aligned} \tag{4.17}$$

Define

$$\mathbf{H}_{ig}(\boldsymbol{\theta}_g,\mu_z) = [\mathbf{m}_{ig}(\boldsymbol{\theta}_g), Z_i-\mu_z]^T, \tag{4.18}$$

and similarly show that an asymptotic equivalent form of $\bar{\mathbf{H}}_{2\pi g}(\boldsymbol{\theta}_g,\mu_z)$ as,

$$\begin{aligned} & \tfrac{1}{N}\textstyle\sum_{i\in A_{2g}} \frac{\mathbf{H}_{ig}(\boldsymbol{\theta}_g,\mu_z)}{\pi_{1i}\hat{\pi}_{2ig}} \\ = \; & \tfrac{1}{N}\textstyle\sum_{i\in U} \frac{\delta_{1i}\delta_{2ig}\mathbf{H}_{ig}(\boldsymbol{\theta}_g,\mu_z)}{\pi_{1i}\pi_{2ig}} - \tfrac{1}{N}\sum_{i\in U}\frac{\delta_{1i}(\delta_{2ig}-\pi_{2ig})}{\pi_{1i}\pi_{2ig}} E_\xi(\mathbf{H}_{ig}(\boldsymbol{\theta}_g,\mu_z)|X_i) + o_p(n^{-1/2}) \\ = \; & \tfrac{1}{N}\textstyle\sum_{i\in A_1} \frac{\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g,\mu_z)}{\pi_{1i}} + o_p(n^{-1/2}),\end{aligned} \tag{4.19}$$

where

$$\begin{aligned} & \boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z) \\ = \; & \mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)\tfrac{\delta_{2ig}}{\pi_{2ig}} + (1 - \tfrac{\delta_{2ig}}{\pi_{2ig}})\boldsymbol{\mu}_{Hg}(X_i; \boldsymbol{\theta}_g, \mu_z), \text{ and } \boldsymbol{\mu}_{Hg}(X_i, \boldsymbol{\theta}_g) = E_\xi(\mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)|X_i). \end{aligned} \tag{4.20}$$

Thus $\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)$ in (4.14) is written as,

$$\begin{aligned} \mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z) &= [\tfrac{1}{N}\textstyle\sum_{i\in A_{2g}} \frac{\mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)}{\pi_{1i}\hat{\pi}_{2ig}}, \tfrac{1}{N}\sum_{i\in A_1} \frac{Z_i - \mu_z}{\pi_{1i}}]^T \\ &= [\tfrac{1}{N}\textstyle\sum_{i\in A_1} \frac{\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)}{\pi_{1i}}, \tfrac{1}{N}\sum_{i\in A_1} \frac{Z_i - \mu_z}{\pi_{1i}}]^T + o_p(n^{-1/2}). \end{aligned} \tag{4.21}$$

Then the large sample theory for $\hat{\boldsymbol{\theta}}_g^{(2)}$ is derived based on the asymptotic form of $\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)$ in (4.21). Theorem 1 can now be stated as:

Theorem 1: Under the regularity conditions in Appendix B1, for any $g = 1, ...G$,

$$\sqrt{n}\left(\begin{bmatrix} \hat{\boldsymbol{\theta}}_g^{(2)} \\ \hat{\mu}_z \end{bmatrix} - \begin{bmatrix} \boldsymbol{\theta}_g^0 \\ \mu_z^0 \end{bmatrix}\right) \xrightarrow{\mathcal{L}} N(\mathbf{0}, V_g(\boldsymbol{\theta}_g^0, \mu_z^0)),$$

where

$$V_g(\boldsymbol{\theta}_g, \mu_z) = \left[\Gamma_g^T(\boldsymbol{\theta}_g)\boldsymbol{\Sigma}_{Hg}^{-1}(\boldsymbol{\theta}_g, \mu_z)\Gamma_g^T(\boldsymbol{\theta}_g)\right]^{-1}, \tag{4.22}$$

$$\Gamma_g(\boldsymbol{\theta}_g) = \left[E_\xi\left[\frac{\partial \mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)}{\partial \boldsymbol{\theta}_g}\right] \quad E_\xi\left[\frac{\partial \mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)}{\partial \mu_z}\right]; \quad \mathbf{0} \quad -1\right], \tag{4.23}$$

$$\text{and} \quad \boldsymbol{\Sigma}_{Hg}(\boldsymbol{\theta}_g, \mu_z) = \left[\boldsymbol{\Sigma}_{11}(\boldsymbol{\theta}_g, \mu_z) \quad \boldsymbol{\Sigma}_{12}(\boldsymbol{\theta}_g, \mu_z); \quad \boldsymbol{\Sigma}_{12}^T(\boldsymbol{\theta}_g, \mu_z) \quad \boldsymbol{\Sigma}_{22}(\mu_z)\right]. \tag{4.24}$$

Here the notation $[\mathbf{a_{11}}, \mathbf{a_{12}}; \mathbf{a_{21}}, \mathbf{a_{22}}]$ represents a 2×2 block matrix with blocks $\mathbf{a_{ij}}$. The term $\boldsymbol{\Sigma}_{11}(\boldsymbol{\theta}_g, \mu_z)$ in (4.24) is related to the asymptotic variance of the first element in (4.21) and is defined as,

$$\boldsymbol{\Sigma}_{11}(\boldsymbol{\theta}_g, \mu_z) = \lim_{N\to\infty} V_{\eta g,N}(\boldsymbol{\theta}_g, \mu_z) + \frac{n}{N} Var_\xi(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)), \tag{4.25}$$

$$\text{where} \quad V_{\eta g,N}(\boldsymbol{\theta}_g, \mu_z) = nN^{-2}\sum_{i\in U}\sum_{j\in U} \frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}} \boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)\boldsymbol{\eta}_{jg}^T(\boldsymbol{\theta}_g, \mu_z). \tag{4.26}$$

The term $\mathbf{\Sigma}_{22}(\mu_z)$ in (4.24) is related to the asymptotic variance of the second element in (4.21) and is defined as,

$$\mathbf{\Sigma}_{22}(\mu_z) = \lim_{N\to\infty} V_{z,N}(\mu_z) + \frac{n}{N} Var_{\xi}(Z_i), \tag{4.27}$$

$$\text{where} \qquad V_{z,N}(\mu_z) = nN^{-2} \sum_{i\in U} \sum_{j\in U} \frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}} (Z_i - \mu_z)(Z_j - \mu_z). \tag{4.28}$$

The term $\mathbf{\Sigma}_{12}(\boldsymbol{\theta}_g, \mu_z)$ in (4.24) is related to the asymptotic covariance between the two elements in (4.21) and is defined as,

$$\mathbf{\Sigma}_{12}(\boldsymbol{\theta}_g, \mu_z) = \lim_{N\to\infty} C_{\eta z,N}(\boldsymbol{\theta}_g, \mu_z) + \frac{n}{N} Cov_{\xi}(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z), Z_i), \tag{4.29}$$

$$\text{where} \qquad C_{\eta z,N}(\boldsymbol{\theta}_g, \mu_z) = nN^{-2} \sum_{i\in U} \sum_{j\in U} \frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}} \boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)(Z_j - \mu_z). \tag{4.30}$$

Equation (4.25) is connected to Fact 1 stated above, where its first term is

$$nV(N^{-1} \sum_{i\in A_1} \pi_{1i}^{-1} \boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g) | \mathcal{F}_N) \tag{4.31}$$

on the finite population corresponding to V_{11} in Fact 1, and its second term is

$$nV_{\xi}(N^{-1} \sum_{i\in U} \boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g)) \tag{4.32}$$

on the superpopulation level corresponding to V_{22} in Fact 1. The limit sign in the first term of (4.25) indicates this is the limit with respect to the process of repeated sampling from a sequence of finite population as $N \to \infty$. Similar connections can be seen in (4.27) and (4.29). The proof of Theorem 1 uses results from Pakes and Pollard (1989) (Theorems 3.2 and 3.3) which provides a general central limit theorem for estimators defined by minimization of the length of a vector valued random criterion function. The justification of Theorem 1 takes into account the finite population asymptotic framework and the semiparametric estimation of $\hat{\pi}_{2ig}$. The asymptotic equivalence of $\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)$ in (4.17) is analytically similar to the mathematical forms of the doubly robust (DR) estimators when $\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$, see

Kim and Haziza (2014), Haziza and Rao (2006), Tan (2010), and Robins et al. (2007). One difference is that the consistency of the DB estimators requires one of the response model and the outcome model to be correctly specified, while our estimators estimate both the self-selection probabilities π_{2ig} and the outcome model semiparametrically. The regularity conditions on the sample design and tuning parameters for the semiparametric estimation are provided in the Appendix B1, and an outline of the proof for Theorem 1 can be found in Appendix B2.

4.3.2 Variance Estimation Based on The Asymptotic Normality

We use the asymptotic variance $V_g(\boldsymbol{\theta}_g^0, \mu_z^0)$ in (4.22) to estimate the variance of $\hat{\boldsymbol{\theta}}_g^{(2)}$. To estimate $\boldsymbol{\Sigma}_{Hg}(\boldsymbol{\theta}_g, \mu_z)$, an estimator of $\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)$ is obtained by,

$$\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z) = \mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)\frac{\delta_{2ig}}{\hat{\pi}_{2ig}} + (1 - \frac{\delta_{2ig}}{\hat{\pi}_{2ig}})\hat{\boldsymbol{\mu}}_{Hg}(X_i; \boldsymbol{\theta}_g, \mu_z), \tag{4.33}$$

where $\boldsymbol{\mu}_{Hg}(X_i, \boldsymbol{\theta}_g)$ is also estimated semiparametrically using the same bases $R_K(X_i)$, i.e.

$$\hat{\boldsymbol{\mu}}_{Hg}(X_i; \boldsymbol{\theta}_g, \mu_z) = \hat{\boldsymbol{\beta}}_g^T(\boldsymbol{\theta}_g, \mu_z)R_K(X_i), \text{ and} \tag{4.34}$$

$$\hat{\boldsymbol{\beta}}_g(\boldsymbol{\theta}_g, \mu_z) = (\sum_{i\in A_{2g}} \pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1}R_K(X_i)R_K(X_i)^T)^{-1} \sum_{i\in A_{2g}} \pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1}R_K(X_i)\mathbf{H}_{ig}^T(\boldsymbol{\theta}_g, \mu_z). \tag{4.35}$$

An estimator of $V_g(\boldsymbol{\theta}_g^0, \mu_z^0)$ is calculated as follows,

$$\hat{V}_g(\hat{\boldsymbol{\theta}}_g^{(2)}, \hat{\mu}_z) = \left[\hat{\Gamma}_g^T(\hat{\boldsymbol{\theta}}_g^{(2)})\hat{\boldsymbol{\Sigma}}_{Hg}^{-1}(\hat{\boldsymbol{\theta}}_g^{(2)}, \hat{\mu}_z)\hat{\Gamma}_g^T(\hat{\boldsymbol{\theta}}_g^{(2)})\right]^{-1}, \tag{4.36}$$

where

$$\hat{\Gamma}_g(\boldsymbol{\theta}_g) = \frac{1}{N}\left[\sum_{i\in A_{2g}} w_{1i}\hat{\pi}_{2ig}^{-1}\frac{\partial \mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)}{\partial \boldsymbol{\theta}_g} \quad \sum_{i\in A_{2g}} w_{1i}\hat{\pi}_{2ig}^{-1}\frac{\partial \mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)}{\partial \mu_z}; \quad \mathbf{0} \quad -1\right], \tag{4.37}$$

$$\text{and} \quad \hat{\boldsymbol{\Sigma}}_{Hg}(\boldsymbol{\theta}_g, \mu_z) = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{11}(\boldsymbol{\theta}_g, \mu_z) & \hat{\boldsymbol{\Sigma}}_{12}(\boldsymbol{\theta}_g, \mu_z); & \hat{\boldsymbol{\Sigma}}_{12}^T(\boldsymbol{\theta}_g, \mu_z) & \hat{\boldsymbol{\Sigma}}_{22}(\mu_z) \end{bmatrix}. \tag{4.38}$$

The term $\hat{\boldsymbol{\Sigma}}_{11}(\boldsymbol{\theta}_g, \mu_z)$ is estimated using

$$\hat{\boldsymbol{\Sigma}}_{11}(\boldsymbol{\theta}_g, \mu_z) = \hat{V}_{\eta g,N}(\boldsymbol{\theta}_g, \mu_z) + \frac{n}{N}\hat{Var}_{\xi}(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)), \tag{4.39}$$

$$\text{where} \quad \hat{V}_{\eta g,N}(\boldsymbol{\theta}_g, \mu_z) = nN^{-2}\sum_{i\in A_1}\sum_{j\in A_1}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{1ij}\pi_{1i}\pi_{1j}}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)\hat{\boldsymbol{\eta}}_{jg}^T(\boldsymbol{\theta}_g, \mu_z), \quad \text{and} \tag{4.40}$$

$$\begin{aligned} & \hat{Var}_{\xi}(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)) \\ = \; & \tfrac{1}{N}\textstyle\sum_{i\in A_1}\pi_{1i}^{-1}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)\hat{\boldsymbol{\eta}}_{ig}^T(\boldsymbol{\theta}_g, \mu_z) \\ - \; & \tfrac{1}{N^2}\left[\textstyle\sum_{i\in A_1}\pi_{1i}^{-1}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)\right]\left[\textstyle\sum_{i\in A_1}\pi_{1i}^{-1}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)\right]^T. \end{aligned} \tag{4.41}$$

The term $\hat{\boldsymbol{\Sigma}}_{22}(\mu_z)$ is estimated using

$$\hat{\boldsymbol{\Sigma}}_{22}(\mu_z) = \hat{V}_{z,N}(\mu_z) + \frac{n}{N}\hat{Var}_{\xi}(Z_i), \tag{4.42}$$

$$\text{where} \quad \hat{V}_{z,N}(\mu_z) = nN^{-2}\sum_{i\in A_1}\sum_{j\in A_1}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{1ij}\pi_{1i}\pi_{1j}}(Z_i-\mu_z)(Z_j-\mu_z), \quad \text{and} \tag{4.43}$$

$$\hat{Var}_{\xi}(Z_i) = \frac{1}{N}\sum_{i\in A_1}\pi_{1i}^{-1}(Z_i-\mu_z)^2 - \frac{1}{N^2}\left[\sum_{i\in A_1}\pi_{1i}^{-1}(Z_i-\mu_z)\right]\left[\sum_{i\in A_1}\pi_{1i}^{-1}(Z_i-\mu_z)\right]^T. \tag{4.44}$$

The term $\hat{\boldsymbol{\Sigma}}_{12}(\boldsymbol{\theta}_g, \mu_z)$ is estimated using

$$\hat{\boldsymbol{\Sigma}}_{12}(\boldsymbol{\theta}_g, \mu_z) = \hat{C}_{\eta z,N}(\boldsymbol{\theta}_g, \mu_z) + \frac{n}{N}\hat{Cov}_{\xi}(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z), Z_i), \tag{4.45}$$

$$\text{where} \quad \hat{C}_{\eta z,N}(\boldsymbol{\theta}_g, \mu_z) = nN^{-2}\sum_{i\in A_1}\sum_{j\in A_1}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{1ij}\pi_{1i}\pi_{1j}}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)(Z_j-\mu_z), \quad \text{and} \tag{4.46}$$

$$\begin{aligned} & \hat{Cov}_{\xi}(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z), Z_i) \\ = \; & \tfrac{1}{N}\textstyle\sum_{i\in A_1}\pi_{1i}^{-1}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)(Z_i-\mu_z) - \tfrac{1}{N^2}\left[\textstyle\sum_{i\in A_1}\pi_{1i}^{-1}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)\right]\left[\textstyle\sum_{i\in A_1}\pi_{1i}^{-1}(Z_i-\mu_z)\right]. \end{aligned} \tag{4.47}$$

To construct a joint estimator for $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_G]^T$, one can simply stack $\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)$ in the quadratic form of (4.13). Define $\mathbf{H}_i(\boldsymbol{\theta}, \mu_z)$ as the stacked vector of $\mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)'s$ in (4.18) and $\boldsymbol{\eta}_i(\boldsymbol{\theta}, \mu_z)$ as the stacked vector of $\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)'s$ in equation (4.20). The asymptotic theory and the variance estimator for $\hat{\boldsymbol{\theta}}^{(2)}$ can be derived by simply replacing $\mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)$ by $\mathbf{H}_i(\boldsymbol{\theta}, \mu_z)$ and $\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)$ by $\boldsymbol{\eta}_i(\boldsymbol{\theta}, \mu_z)$. Then an inference for the treatment effects or any linear combination of treatment parameters, $\boldsymbol{\lambda}^T\boldsymbol{\theta}$ can be obtained.

4.3.3 Replication Variance Estimation

In surveys conducted on land, for example surveys about natural resources (soil, forest, water, etc.), non-responses hardly occur. However, in surveys with high non-response rates, such as surveys conducted on people, the joint inclusion probabilities are typically not available because sampling weights have to be appropriately adjusted for non-response. After such adjustments, the joint inclusion probabilities change and are hard to be derived. In practice, a set of replicate weights are often provided instead, because (1) design weights are often adjusted due to non-response issues and a set of replicate weights are provided to account for the weight adjustment; (2) sometimes a few design variables are masked from users to keep confidentiality. An example of such design variable is location which is used for defining design strata in a study, but the specific location is omitted from the analyst because of concerns associated with confidentiality. In this subsection, the replicate weights for the Jackknife variance estimator are constructed for $\hat{\boldsymbol{\theta}}_g^{(1)}$. Note that $\hat{\boldsymbol{\theta}}_g^{(2)}$ depends on the joint inclusion probabilities π_{1ij} which are typically not available when replicate weights are provided. The Jackknife variance estimator for a two-phase sampling design discussed in Fuller (2011) and Kim et al. (2006) is proposed for use here. Assume that there is a replicate variance estimator that gives a consistent estimator for the variance of the total estimator based on the first-phase sample. The replication variance estimator is written as, $\hat{V}_{JK1}(\hat{\boldsymbol{\theta}}_1) = \sum_{b=1}^{B} c_b(\hat{\boldsymbol{\theta}}_1^{[b]} - \hat{\boldsymbol{\theta}}_1)(\hat{\boldsymbol{\theta}}_1^{[b]} - \hat{\boldsymbol{\theta}}_1)^T$, where B is the number of replicates, $\hat{\boldsymbol{\theta}}_1 = \sum_{i \in A_1} w_{1i} X_i$ is the total estimator of variable X using the first-phase sample, $\hat{\boldsymbol{\theta}}_1^{[b]} =$

$\sum_{i\in A_1} w_{1i}^{[b]} X_i$ is the estimated total for the b^{th} replicate, $w_{1i}^{[b]}$ is the b^{th} replicate weights in the first-phase, and c_b is a factor associated with replicate b such that $\hat{V}_{JK1}(\hat{\boldsymbol{\theta}}_1)$ is a consistent estimator for the variance of $\hat{\boldsymbol{\theta}}_1$. Suppose the second-phase total estimator is, $\hat{\boldsymbol{\theta}}_2 = \sum_{i\in A_2} w_{1i}\pi_{2i|1i}^{-1} X_i$, where $\pi_{2i|1i}$ is the conditional probability of selecting i for the phase 2 sample given that i is in the phase 1 sample, and A_2 is the phase 2 sample. Define the b^{th} replicate of $\hat{\boldsymbol{\theta}}_2$ as, $\hat{\boldsymbol{\theta}}_2^{[b]} = \sum_{i\in A_2} w_{1i}^{[b]}\pi_{2i|1i}^{-1} X_i$. A Jackknife variance estimator for $\hat{\boldsymbol{\theta}}_2$ can be calculated as, $\hat{V}_{JK2}(\hat{\boldsymbol{\theta}}_2) = \sum_{b=1}^{B} c_b(\hat{\boldsymbol{\theta}}_2^{[b]} - \hat{\boldsymbol{\theta}}_2)(\hat{\boldsymbol{\theta}}_2^{[b]} - \hat{\boldsymbol{\theta}}_2)^T$. Kim et al. (2006) showed that $\hat{V}_{JK2}(\hat{\boldsymbol{\theta}}_2)$ is a consistent estimator for the variance of $\hat{\boldsymbol{\theta}}_2$.

Following the idea of Fuller (2009) (Section 4.3.3), let b be the index for the deleted Jackknife groups and the corresponding replicate version of $\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)$ be,

$$\bar{\mathbf{m}}_{2\pi g}^{[b]}(\boldsymbol{\theta}_g) = \frac{1}{N}\sum_{i\in A_{2g}} w_{1i}^{[b]}(\hat{\pi}_{2ig}^{[b]})^{-1}\mathbf{m}_{ig}(\boldsymbol{\theta}_g), \tag{4.48}$$

where $\hat{\pi}_{2ig}^{[b]}$ is obtained by replacing w_{1i} by $w_{1i}^{[b]}$ in (4.7). Then the replicate estimator for $\hat{\boldsymbol{\theta}}_g^{(1)}$ is,

$$\hat{\boldsymbol{\theta}}_g^{(1)[b]} = \arg\min_{\boldsymbol{\theta}_g} \left[\bar{\mathbf{m}}_{2\pi g}^{[b]}(\boldsymbol{\theta}_g)\right]^T \left[\bar{\mathbf{m}}_{2\pi g}^{[b]}(\boldsymbol{\theta}_g)\right], \tag{4.49}$$

and the replication variance estimator for $\hat{\boldsymbol{\theta}}_g^{(1)}$ is calculated as,

$$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{(1)}) = \sum_{b=1}^{B} c_b(\hat{\boldsymbol{\theta}}_g^{(1)[b]} - \hat{\boldsymbol{\theta}}_g^{(1)})(\hat{\boldsymbol{\theta}}_g^{(1)[b]} - \hat{\boldsymbol{\theta}}_g^{(1)})^T. \tag{4.50}$$

Examples of $w_{1i}^{[b]}$ and c_b for a variety of designs are given in Särndal et al. (2003). For example, if the first-phase sample is drawn from a multi-stage cluster design, the Jackknife technique is usually applied at the primary sampling unit (PSU) levels. Assuming there are B PSUs and S_b is the b^{th} PSU deleted in the b^{th} replicate sample, the b^{th} replicate weight for the first-phase is defined as,

$$w_{1i}^{[b]} = \begin{cases} 0 & \text{if} \quad i \in S_b \\ \frac{B}{B-1} w_{1i} & \text{if} \quad i \notin S_b \end{cases}, \tag{4.51}$$

and $c_b = B^{-1}(B-1)$. As mentioned in Särndal et al. (2003), for stratified sampling designs, $w_{1i}^{[b]}$ and c_b need to be defined with care. This situation is discussed in Section 4.5 of the

empirical study. If the first phase replicate weights are provided in practice, one can directly use them as $w_{1i}^{[b]}$. One thing to note is that Kim et al. (2006) assume π_{2ig} are known in their two phase replication variance estimator. Note that the consistency theorem in Kim et al. (2006) needs to be modified to account for the variation from estimating $\hat{\pi}_{2ig}$ in the Jackknife variance estimator, which has not been done in this Chapter and is open for future research.

4.4 Simulation Study

In this section, the performance of the estimators and variance estimators are evaluated under four different simulation set-ups. Three treatment levels are considered, with a population size of $N = 10,000$ and an expected sample size of $n = 1000$. I.i.d. realizations, $(\mathbf{Y}, \delta_{1i}, \boldsymbol{\delta}_{2i}, X_i, Z_i); i = 1, ..., N$, are generated according to the following superpopulation set-ups.

(1) Covariates: simulate covariates $\mathbf{Z}_i = [Z_{1i}, Z_{2i}]$ where $Z_{1i} \sim N(2, 1)$ and $Z_{2i} \sim N(10, 1)$, and $\mathbf{X}_i = [X_{1i}, X_{2i}]$ where $X_{1i} = Z_{1i}$ and $X_{2i} \sim N(0.5, 0.3^2)$.

(2) Potential response outcomes: the superpopulation model for potential outcomes is

$$Y_{ig} = \mu_g(\mathbf{Z}_i) + \sigma_g(\mathbf{Z}_i)\epsilon_{ig},$$

where

$$\mu_g(\mathbf{Z}_i) = \beta_{g0} + \beta_{g1}(Z_{1i} - 0.5) + \beta_{g2}(Z_{1i} - 0.5)^2 + \beta_{g3}Z_{2i},$$

$\epsilon_{ig} \sim N(0, 1)$, $\sigma_g(\mathbf{Z}_i) = |\mu_g(\mathbf{Z}_i)|$, and $[\beta_{g0}, \beta_{g1}, \beta_{g2}, \beta_{g3}]$ equals to $[5, 4, 2, 1]$ for $g = 1$, $[0, 1, 0, 0]$ for $g = 2$, and $[-5, -4, -2, -0.5]$ for $g = 3$.

(3) First phase sampling: consider two sampling designs, non-informative stratification sampling and informative Poisson sampling.

- Stratification (STS): population units are sorted by values of Z_{1i}, and then the population is divided into two sub-populations U_1 and U_2 with equal sizes. Simple random

sampling is used to draw 80% of the sample from U_1 and 20% from U_2. For units in stratum s ($s = 1$ or 2), $\pi_{1i} = N_s^{-1} n_s$ and $\pi_{1ij} = \{N_s(N_s-1)\}^{-1} n_s(n_s-1)$, where n_s and N_s are the sample size and the population size in stratum s. The joint inclusion probability for two units in different strata is the product of their first order inclusion probabilities.

- Informative Poisson (Informative): the first-phase sample design is Poisson sampling with selection probability,

$$\pi_{1i} = \frac{exp(-1.5 - 2.5X_{2i} + 0.07\|\mathbf{Y}\|)}{1 + exp(-1.5 - 2.5X_{2i} + 0.07\|\mathbf{Y}\|)},$$

where $\|\mathbf{Y}\| = \sqrt{Y_{i1}^2 + Y_{i2}^2 + Y_{i3}^2}$. Modeling π_{1i} as a function of $\mathbf{Y}$ is a common way (Pfeffermann and Sverchkov, 1999) to represent joint dependence of $\mathbf{Y}$ and π_{1i} on a design variable that is not contained in (X_i, Z_i). In this specification, assume $\|\mathbf{Y}\|$ is known at the design stage of the survey, but is unavailable at the analysis stage.

(4) Second phase self-selection probability models: consider two models for π_{2ig}.

- Logit Linear (LogitLinear):

$$\pi_{2ig} = \frac{\exp(\phi_{g0} + \phi_{g1}X_{1i} + \phi_{g2}X_{2i})}{\sum_{g=1}^{G} \exp(\phi_{g0} + \phi_{g1}X_{1i} + \phi_{g2}X_{2i})},$$

where $[\phi_{g0}, \phi_{g1}, \phi_{g2}]$ equals to $[-.5, 0, 0]$ for $g = 1$, $[0.3, -0.3, -0.3]$ for $g = 2$, and $[0, -0.5, 0.5]$ for $g = 3$.

- Jump (JUMP):

$$\begin{aligned}
[\pi_{2i1}, \pi_{2i2}, \pi_{2i3}] &= [0.90, 0.05, 0.05] \quad \text{if} \quad X_{1i} + X_{2i} \geq 3 \\
&= [1/3, 1/3, 1/3] \quad \text{if} \quad 2 \leq X_{1i} + X_{2i} < 3 \\
&= [0.05, 0.05, 0.90] \quad \text{if} \quad X_{1i} + X_{2i} < 2.
\end{aligned}$$

The JUMP model violates the differentiability assumption of π_{2ig} in Condition III(1) in Appendix B1. It is deliberately included in the simulation to see if a semiparametric approach can estimate non-smooth multiple treatment selection probabilities well.

For each $i \in U$, $\boldsymbol{\delta}_{2i}$ is simulated from $multinomial(1; \pi_{2i1}, \pi_{2i2}, \pi_{2i3})$. For $i \neq j$, $\pi_{1ij} = \pi_{1i}\pi_{1j}$. For STS design, which is non-informative, SMAR holds and $b_i = w_{1i}^{-1}$ in (4.7) to estimate $\hat{\pi}_{2ig}$. For Informative design, SMAR fails and $b_i = 1$ in (4.7) to estimate $\hat{\pi}_{2ig}$.

First simulate a finite population with size N from the superpopulation and then use indicators generated in (3) and (4) to obtain the first and second phase samples. Repeat the process to produce 1000 Monte Carlo (MC) samples. 5 parameters of interest are estimated for each group, $\boldsymbol{\theta}_g = [P_g, \mu_g, \sigma_g^2, R_g, D_g]$, where $P_g = Prob(Y_{ig} \leq 0)$, $\mu_g = E(Y_{ig})$, $\sigma_g^2 = Var(Y_{ig})$ and $R_g = Corr(Y_{ig}, Z_{2i})$, and $D_g = E[E(Y_{ig}|Z_{1i} \leq 0.65)]$. The corresponding estimation equations $\mathbf{m}_{ig}(\boldsymbol{\theta}_g)$ can be found in (4.5) and (4.6). For each MC sample, calculate the following four estimators:

- $\hat{\boldsymbol{\theta}}_g^{(1)}$: the estimator defined in (4.11). When $\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$, $\hat{\boldsymbol{\theta}}_g^{(1)}$ corresponds to the estimator in Yu et al. (2013) asymptotically.
- $\hat{\boldsymbol{\theta}}_g^{(2)}$: the estimator defined in (4.13).
- $\hat{\boldsymbol{\theta}}_g^{nw}$: the estimator defined in (4.9), and is included to see what happens when the survey weights are ignored in analyses.
- $\hat{\boldsymbol{\theta}}_g^{p}$: the estimator calculated the same way as $\hat{\boldsymbol{\theta}}_g^{(1)}$, except that $\hat{\pi}_{2ig}$ are estimated using a parametric multinomial regression. This estimator is introduced in order to have plausible comparisons in context of 3 treatments between our estimators and others that use parametric logistic regression to estimate propensity scores, see DuGoff et al. (2014), Zanutto (2006), Ashmead (2014), and Ridgeway et al. (2015).

A cubic spline base of X_{1i} is used for $R_K(X_{1i})$, as suggested by Breidt et al. (2005) which mentions that setting the degree of the spline equal to 3 is a popular choice in practice. Condition II(4) in Appendix B1 gives a practical guidance for the choice of K, the number of knots in the spline. Condition II(4) requires $K = O(n^\nu)$, where ν has an upper bound $\nu \leq (4\eta + 2)^{-1}$ with $\eta = 1/2$ for spline bases. The sample size used is $n = 1000$, suggesting $n^\nu = 5.6$. The choices of $K = 5, 4, 3, 2$ are tried and the corresponding $\hat{\pi}_{2ig}$ curves are

plotted. It is found that there is not noticeable change in the $\hat{\pi}_{2ig}$ curves until K decreases to 2. So $K = 3$ is used and the locations of the three knots correspond to the 25^{th}, 50^{th}, and 75^{th} quantiles of observed X_{1i}'s. A cubic spline base for $R_K(X_{2i})$ is constructed the same way. The semiparametric bases are $R_K(\mathbf{X}_i) = [R_K^T(X_{1i}), R_K^T(X_{2i})]^T$.

If the dimension of $(\mathbf{X}_i, \mathbf{Z}_i)$ is big, in practice it is suggested to run a multinomial regression using $\boldsymbol{\delta}_{2i}$ on $(\mathbf{X}_i, \mathbf{Z}_i)$ to select covariates that are most significant, and then use them for estimation of $\hat{\pi}_{2ig}$. When using $\hat{\boldsymbol{\theta}}_g^{(2)}$, one can run a multiple linear regression of Y_{ig} on $(\mathbf{X}_i, \mathbf{Z}_i)$ in A_{2g} to identify covariates that are most useful for explaining the outcome Y_{ig}, and then add their first and second phase means in the estimation equations. It is not impossible to obtain a very small $\hat{\pi}_{2ig}$ computationally, which leads to extreme weights. A solution is to truncate such $\hat{\pi}_{2ig}$'s to a small constant L (which is set to be 0.0001 in this study), then adjust the truncated $\hat{\pi}_{2ig}$ by calibrating the second phase mean of U_i to its first phase mean, i.e. $\tilde{\pi}_{2ig} = F_g \hat{\pi}_{2ig}^t$ where $F_g = (\sum_{i \in A_1} w_{1i} U_i)^{-1} \sum_{i \in A_{2g}} w_{1i} (\hat{\pi}_{2ig}^t)^{-1} U_i$, and $\hat{\pi}_{2ig}^t$ is the truncated propensity score which equals to L if $\hat{\pi}_{2ig} < L$, otherwise remains unchanged. Here the variable U_i can be an important covariate chosen by users, or a weighted mean of $(\mathbf{X}_i, \mathbf{Z}_i)$ where weights indicate importance of the covariates. The average of the covariate $\mathbf{X}_i$ is used as U_i in both of the simulation studies and the empirical study.

Figures 4.1-4.4 show side-by-side boxplots of MC estimates of the four estimators for all treatment effects. Each figure represents one of four simulation set-ups: (STS-LogitLinear), (STS-JUMP), (Informative-LogitLinear), and (Informative-JUMP). In each subplot, the first two boxplots are for $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$, and the third and fourth boxplots are for $\hat{\boldsymbol{\theta}}_g^{p}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ respectively. When comparing our estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ with $\hat{\boldsymbol{\theta}}_g^{nw}$, $\hat{\boldsymbol{\theta}}_g^{nw}$ is highly biased in most of parameters and scenarios, due to ignoring the survey weights. The variances of $\hat{\boldsymbol{\theta}}_g^{nw}$ in general are smaller than those of $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$, which is expected especially when the survey weights are very different from each other. The coefficient of variation (CV) of the weights for the STS design is 0.75, and the CV of weights for the Informative design is 4.77. When comparing estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ with $\hat{\boldsymbol{\theta}}_g^{p}$, biases of $\hat{\boldsymbol{\theta}}_g^{p}$ are comparable to those

of $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ for the LogitLinear model because in this scenario $\hat{\boldsymbol{\theta}}_g^{p}$ correctly assumes a parametric model for π_{2ig}. However, in the situation of JUMP models, $\hat{\boldsymbol{\theta}}_g^{p}$ has larger biases than $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ because π_{2ig} is misspecified parametrically. When comparing $\hat{\boldsymbol{\theta}}_g^{(1)}$ with $\hat{\boldsymbol{\theta}}_g^{(2)}$, both of their biases are comparable in all scenarios. However, the plots show that $\hat{\boldsymbol{\theta}}_g^{(2)}$ consistently has smaller variances than $\hat{\boldsymbol{\theta}}_g^{(1)}$. The variance reduction of $\hat{\boldsymbol{\theta}}_g^{(2)}$ over $\hat{\boldsymbol{\theta}}_g^{(1)}$ indicates that efficiency gain occurs after adding the first and second phase means of covariates to the estimation equations, which confirms Remark 2. Additionally, it is promising to see that both $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ have relatively small biases even if the JUMP model fails to satisfy the differentiability assumption in the theory, indicating the semiparametric approach of estimating $\hat{\pi}_{2ig}$ works well for the non-smooth function considered. The MC results can also be found in four tables for readers who prefer to see numbers rather than figures. Tables 4.9-4.12 report biases, MC standard deviations, and the root mean square of errors of the estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ for different treatments in the four simulation scenarios: STS-LogitLinear, STS-JUMP, Informative-LogitLinear and Informative-JUMP.

Tables 4.5 and 4.6 contain the coverage probabilities of the 95% confidence intervals for $\hat{\boldsymbol{\theta}}_g^{(2)}$ based on its asymptotic normality and its linearized variance estimator in Section 4.3.2, and the coverage probabilities of the 95% confidence intervals for $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ based on the Jackknife approach discussed in Section 4.3.3. The replication variance estimator for $\hat{\boldsymbol{\theta}}_g^{nw}$ is calculated by replacing w_{1i} by N/n in (4.51). This gives inappropriate variance estimation for $\hat{\boldsymbol{\theta}}_g^{nw}$ under an unequal probability sampling, but mimics what people do when they ignore survey weights. To create the Jackknife replicates, $B = 1000$ and deleting one unit at a time are used. The coverage probabilities for $\hat{\boldsymbol{\theta}}_g^{(2)}$ using the linearized variance estimator seem to work well, except for the marginal mean μ_g under (STS-LogitLinear) and the marginal proportion P_g under (STS-JUMP). The rest of coverage probabilities are reasonably close to the nominal size 95%. The Jackknife variance estimator of $\hat{\boldsymbol{\theta}}_g^{(1)}$ gives very good coverage probabilities. However the coverage probabilities for $\hat{\boldsymbol{\theta}}_g^{nw}$ using the Jackknife variance estimation are far away from the nominal size, especially under

the Informative-JUMP model where the coverage probabilities are severely underestimated. Those under-coverages are due to the biases in $\hat{\boldsymbol{\theta}}_g^{nw}$, or inappropriate variance estimation, or both.

The simulation studies demonstrate the validity of the estimators and variance estimators.

4.4.1 Comparison Between This Jackknife Variance Estimator and the Generalized Jackknife Variance Estimator

Berger and Skinner (2005) propose the generalized Jackknife variance estimator that incorporates the joint inclusion probabilities. The generalized Jackknife that uses π_{1ij} (delete 1) given by Berger and Skinner (2005) is defined as,

$$\hat{V}_{JK}^{(G)}(\hat{\theta}_g) = \sum_{j\in A_1}\sum_{i\in A_2} \frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1ij}} u_i u_j \tag{4.52}$$

where

$$u_j = \left(1 - \tilde{w}_{1j}\right)\left(\hat{\theta}_g - \hat{\theta}_{g(j)}\right) \qquad (j \in A_1),$$

$$\tilde{w}_{1i} = \frac{\frac{1}{\pi_{1i}}}{\sum_{k\in A_1}\frac{1}{\pi_{1k}}} (\text{Hajek Weights}),$$

and $\hat{\theta}_{g(j)}$ is obtained by using $\hat{\theta}^{(1)}$ estimator and defining

$$w_{1i}^{[j]} = \begin{cases} 0 & \text{if} \quad i = j \\ \pi_{1i}^{-1} & \text{if} \quad i \neq j \end{cases}. \tag{4.53}$$

The results from the Jackknife variance estimator and the results from the generalized Jackknife method in Berger and Skinner (2005) are compared using the four simulation set-ups. Under all 4 simulation scenarios considered in this chapter, it was found that both of the methods give very close coverage probabilities. See Tables 4.1 through 4.4 below. Only results for the Jackknife variance estimator are discussed in the rest of this chapter.

Table 4.1 **STS-LogitLinear**: The coverage probabilities of the 95% constructed intervals for the 5 estimated parameters using our Jackknife variance estimator (the first column), and the generalized Jackknife variance estimator (the second column).

	Jackknife	Generalized	Jackknife	Generalized	Jackknife	Generalized
P_g	95.1	95.1	94.0	94.1	94.1	94.2
μ_g	95.4	95.4	95.0	95.0	94.1	94.1
σ_g^2	94.7	94.7	94.8	94.8	94.3	94.3
R_g	94.7	94.7	95.1	95.1	94.3	94.4
D_g	94.8	94.8	95.1	95.1	95.9	95.8

Table 4.2 **STS-JUMP**: The coverage probabilities of the 95% constructed intervals for the 5 estimated parameters using our Jackknife variance estimator (the first column), and the generalized Jackknife variance estimator (the second column).

	Jackknife	Generalized	Jackknife	Generalized	Jackknife	Generalized
P_g	92.8	92.7	92.3	92.2	95.4	95.4
μ_g	95.3	95.3	93.3	93.3	95.3	95.3
σ_g^2	93.0	92.9	96.6	96.7	96.1	96.1
R_g	94.6	94.6	95.2	95.3	95.8	95.9
D_g	95.0	95.0	93.5	93.4	96.6	96.7

Table 4.3 **Informative-LogitLinear**: The coverage probabilities of the 95% constructed intervals for the 5 estimated parameters using our Jackknife variance estimator (the first column), and the generalize Jackknife variance estimator (the second column).

	Jackknife	Generalized	Jackknife	Generalized	Jackknife	Generalized
P_g	95.2	95.1	96.4	96.2	94.7	94.7
μ_g	96.4	96.2	95.4	95.4	95.5	95.4
σ_g^2	94.1	94.1	95.1	95.1	96.2	96.1
R_g	96.0	95.9	94.8	94.8	95.4	95.4
D_g	96.2	96.1	95.7	95.6	94.9	94.9

Table 4.4 **Informative-JUMP**: The coverage probabilities of the 95% constructed intervals for the 5 estimated parameters using our Jackknife variance estimator (the first column), and the generalized Jackknife variance estimator (the second column).

	Jackknife	Generalized	Jackknife	Generalized	Jackknife	Generalized
P_g	96.3	95.6	94.7	94.8	96.9	95.9
μ_g	97.2	96.1	97.1	96.0	93.9	94.4
σ_g^2	97.0	96.0	92.3	93.6	96.5	95.7
R_g	93.2	94.1	95.0	95.0	95.4	95.2
D_g	95.0	95.0	94.3	94.6	96.0	95.5

4.5 Empirical Study

The feasibility of the estimators discussed in Section 4.2.3 is investigated by estimating the mean annual medical expenditures under different choices of health insurance types in China. The data is from the Chinese General Social Survey (CGSS) conducted by the National Survey Research Center at the Renming University of China in 2010. The population consisted of all Chinese adults (18+) in mainland China. A sample of 12,000 adults was drawn for the base questionnaire and a subsample of 4,000 adults was drawn for the health care questionnaire. Data were collected by in-person interviews. The sample for the CGSS survey was selected using a multi-stage cluster sampling design. In the first stage, the primary sampling units (PSUs) were districts which were divided into 2 strata. Stratum 1 contained 67 districts in 5 major cities (Shanghai, Beijing, Guangzhou, Shenzhen and Tianjin), and stratum 2 contained 2795 districts in the rest of the area of China. In both strata, a probability proportional to size (PPS) design with the resident population size as the size variable was used to select the PSUs (40 PSUs were selected in stratum 1, and 100 PSUs were selected in stratum 2). In the second stage, the secondary sampling units (SSUs) were communities. A PPS design with resident population size as the size variable was used to select 2 SSUs within each selected PSU in stratum 1 and 4 SSUs within each selected PSU in stratum 2. In the third stage, the ultimate sampling units (USUs) were households. In each selected SSU, 25 households were drawn by a systematic sampling method. Then a respondent was selected randomly within each household. Totally 12,000 households responded to the base questionnaire. Then every third household respondent in each SSU was selected to answer the health care questionnaire. The subsample of 4,000 was used in our investigation.

The response variable in the study is the annual medical expenditure. The treatment variable is the health insurance type (public health insurance, private health insurance, and no health insurance). Public health insurance is sponsored by Chinese government and is the main health insurance type in China. Six relevant covariates are chosen to

study from the health care questionnaire: age, household register (urban, rural, other), annual household income, physical condition (healthy, just so-so/or a little sick, sick, very sick), chronic disease (yes, no), and treatment to illness (self-treatment, go to hospital, no treatment). Due to some non-response units, the final data had a sample size of 3,866. The data weights were adjusted to deal with the non-response issue.

The following parameters of interest are estimated, $\boldsymbol{\theta}_g^0 = E(Y_{ig})$ and $\boldsymbol{\theta}_g^0 = E(Y_{ig}|I_{di} = 1)$ where I_{di} is the indicator for the domain of interest that contains respondents who have sick or very sick physical condition. When estimating $\hat{\pi}_{2ig}$, use $b_i = 1$ in (4.7) to obtain conservative estimates since it is difficult to verify SMAR assumption. For comparison, the results using $b_i = w_{1i}^{-1}$ in (4.7) are also reported.

Estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ are calculated and the Jackknife variance estimator discussed in Section 4.3.3 is used to calculate their standard errors. $\hat{\boldsymbol{\theta}}_g^{(2)}$ is not included into the empirical study because π_{1ij} are not available. Since the design is a stratified multi-stage cluster design, the districts (PSUs) in different strata are used as the deleted Jackknife groups S_b. The Jackknife variance estimator is,

$$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{(1)}) = \sum_{h=1}^{2} \frac{B_h - 1}{B_h} \sum_{b=1}^{B_h} (\hat{\boldsymbol{\theta}}_g^{(1)[b]} - \hat{\boldsymbol{\theta}}_g^{(1)})(\hat{\boldsymbol{\theta}}_g^{(1)[b]} - \hat{\boldsymbol{\theta}}_g^{(1)})^T, \tag{4.54}$$

where $\hat{\boldsymbol{\theta}}_g^{(1)[b]}$ is the minimizer of (4.49) and the replicate weight in the first-phase is defined as,

$$w_{1i}^{[b]} = \begin{cases} 0 & \text{if} \quad i \in S_b \\ \pi_{1i}^{-1} & \text{if} \quad i \notin S_b \text{ and } h(i) \neq h(b) \\ \frac{B_h}{B_h - 1}\pi_{1i}^{-1} & \text{if} \quad i \notin S_b \text{ and } h(i) = h(b). \end{cases} \tag{4.55}$$

Here $h(i)$ is the stratum to which unit i belongs, $h(b)$ is the stratum where the b^{th} deleted group S_b belongs, and $[B_1, B_2] = [40, 100]$. The replicate estimator $\hat{\boldsymbol{\theta}}_g^{nw[b]}$ for the estimator $\hat{\boldsymbol{\theta}}_g^{nw}$ without survey weights and the variance estimator $\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{nw})$ can be obtained in the same way by simply replacing π_{1i} by nN^{-1} in (4.55). A spline base of degree 2 with 8 equally spaced knots in the data range is constructed for the two continuous variables (age

and annual household income). Dummy variables are created for the remaining categorical variables and added to the model.

Tables 4.7 and 4.8 contain the estimated treatment mean effects and estimated treatment domain mean effects for physical condition, along with standard errors (in parentheses) and 95% confidence intervals (in brackets), for $b_i = 1$ and $b_i = w_{1i}^{-1}$ cases respectively. The treatment effect estimates in Table 4.7(a) indicate that, when the data weights are neglected, the estimated mean medical expenditure of the public health insurance group is not significantly different from that of the no health insurance group. However, when the data weights are incorporated, the public health group is found to spend significantly more on the medical expenses than the no health insurance group. This makes sense because people who have no health insurance might be reluctant to spend money to see doctors. This trend is also seen in the domain treatment effects estimates in Table 4.7(b). In addition, when the data weights are neglected for the treatment mean effect estimates, the estimated mean medical expenditure of the private health insurance group is significantly different from that of the no insurance group. Incorporating the data weights finds these estimated means not significantly different. Table 4.8 gives the same story as Table 4.7 when comparing the public health insurance group versus the private health insurance group, and comparing the public health insurance group versus the no health insurance group. However, when comparing the private health group with the no insurance group, there is a significant difference in the treatment mean effect for both estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ as reported in Table 4.8. Note that the standard errors of the unweighted estimator are not consistently smaller than those of the weighted estimator because the variation of weights in the real data is small (the CV = 0.45).

This study demonstrates that the proposed estimators are feasible in real data application and suggests that ignoring the weights of an observational data might lead to a misleading conclusion.

4.6 Conclusions

GMM estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ were considered to estimate treatment effects defined through an estimation equation in an observational data set that is a sample drawn by a complex survey design. The estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ include both the first-phase sampling probabilities and the estimated second-phase selection probabilities to remove the biases due to ignoring unequal sampling design in the first-phase and the selection biases in the second-phase. The self-selection probabilities are estimated using a semiparametric approach in Cattaneo (2010) to deal with the situation with multiple treatments. Our simulation studies demonstrate that neglecting the first-phase design and handling only treatment selection could lead to erroneous treatment effect estimation. The proposed estimator is designed to handle multiple treatments and do not require strong model assumption of the selection probability as in a fully parametric solution. The estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ can be readily extended to multiple sampling phases as well when the data set is a subsample of a larger survey sample.

Table 4.5 **Stratification**: The coverage probabilities of the 95% constructed intervals for the 5 estimated parameters using the linearized variance estimator $\hat{V}_L(\hat{\boldsymbol{\theta}}_g^{(2)})$ for $\hat{\boldsymbol{\theta}}_g^{(2)}$, and the Jackknife variance estimators $\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{(1)})$ and $\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{nw})$ for $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ respectively.

(a) STS-LogitLinear

	Trt1			Trt 2			Trt 3		
	$\hat{V}_L(\hat{\boldsymbol{\theta}}_g^{(2)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{(1)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{nw})$	$\hat{V}_L(\hat{\boldsymbol{\theta}}_g^{(2)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{(1)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{nw})$	$\hat{V}_L(\hat{\boldsymbol{\theta}}_g^{(2)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{(1)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{nw})$
P_g	93.6	95.2	59.8	94.3	94.0	57.9	92.4	94.1	56.2
μ_g	95.4	95.5	58.7	95.2	95.1	57.9	88.1	94.1	62.3
σ_g^2	92.5	94.7	61.7	94.4	94.9	60.4	92.2	94.3	58.2
R_g	94.2	94.7	57.6	92.1	95.1	60.3	95.9	94.3	59.1
D_g	92.4	94.8	56.7	95.1	95.1	58.8	92.6	95.9	62.2

(b) STS-JUMP

	Trt1			Trt 2			Trt 3		
	$\hat{V}_L(\hat{\boldsymbol{\theta}}_g^{(2)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{(1)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{nw})$	$\hat{V}_L(\hat{\boldsymbol{\theta}}_g^{(2)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{(1)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{nw})$	$\hat{V}_L(\hat{\boldsymbol{\theta}}_g^{(2)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{(1)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{nw})$
P_g	89.2	92.8	71.9	95.2	92.3	80.4	92.5	95.4	50.0
μ_g	92.2	95.3	73.0	95.6	93.3	76.8	94.5	95.3	79.3
σ_g^2	94.2	93.0	56.5	93.3	96.6	83.9	95.3	96.1	86.0
R_g	95.3	94.6	60.7	93.8	95.2	81.0	92.3	95.8	61.1
D_g	92.9	95.0	50.0	96.6	93.2	61.7	94.1	96.6	28.2

Table 4.6 **Informative**: The coverage probabilities of the 95% constructed intervals for the 5 estimated parameters using the linearized variance estimator $\hat{V}_L(\hat{\boldsymbol{\theta}}_g^{(2)})$ for $\hat{\boldsymbol{\theta}}_g^{(2)}$, and the Jackknife variance estimators $\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{(1)})$ and $\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{nw})$ for $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ respectively.

(a) Informative-LogitLinear

	Trt1			Trt 2			Trt 3		
	$\hat{V}_L(\hat{\boldsymbol{\theta}}_g^{(2)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{(1)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{nw})$	$\hat{V}_L(\hat{\boldsymbol{\theta}}_g^{(2)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{(1)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{nw})$	$\hat{V}_L(\hat{\boldsymbol{\theta}}_g^{(2)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{(1)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{nw})$
P_g	94.3	95.2	58.9	94.1	96.4	49.3	95.4	94.7	49.2
μ_g	95.2	96.4	50.6	92.5	95.4	46.2	95.1	95.5	49.3
σ_g^2	92.0	94.1	46.2	95.4	95.1	50.6	93.7	96.2	44.4
R_g	94.9	96.0	45.1	90.6	94.8	42.8	93.8	95.4	38.3
D_g	93.4	96.2	59.3	93.8	95.7	48.1	93.1	94.9	45.7

(b) Informative-JUMP

	Trt1			Trt 2			Trt 3		
	$\hat{V}_L(\hat{\boldsymbol{\theta}}_g^{(2)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{(1)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{nw})$	$\hat{V}_L(\hat{\boldsymbol{\theta}}_g^{(2)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{(1)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{nw})$	$\hat{V}_L(\hat{\boldsymbol{\theta}}_g^{(2)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{(1)})$	$\hat{V}_{JK}(\hat{\boldsymbol{\theta}}_g^{nw})$
P_g	94.4	96.3	6.6	97.7	94.7	72.4	92.9	96.9	0.0
μ_g	92.5	97.2	0.0	91.1	97.1	2.0	92.2	93.9	0.0
σ_g^2	94.9	97.0	0.0	95.5	92.3	2.0	94.6	96.5	20.6
R_g	92.7	93.2	41.6	95.2	95.0	71.4	94.3	95.4	48.4
D_g	92.2	95.0	0.0	91.0	94.3	81.9	95.0	96.0	0.0

Table 4.7 **Empirical Study With Weights in Estimation of $\hat{\pi}_{2ig}$**: The treatment effect estimates using estimators $\hat{\boldsymbol{\theta}}_g^{nw}$ and $\hat{\boldsymbol{\theta}}_g^{(1)}$ defined in Section 4.2.3. The parameter of interests are $\boldsymbol{\theta}_g^0 = E(Y_{ig})$ and $\boldsymbol{\theta}_g^0 = E(Y_{ig}|I_{di} = 1)$ where I_{di} is the indicator for the domain of interest that contains respondents who have sick or very sick physical condition. The standard errors are in parentheses and calculated using the Jackknife variance estimator, and the 95% confidence intervals are in brackets.

(a) Treatment Mean Effect Estimates for $\boldsymbol{\theta}_g^0 = E(Y_{ig})$

Estimators	Public - Private	Public - No Insurance	Private - No Insurance
$\hat{\boldsymbol{\theta}}_g^{(1)}$	1349.57 (215.90) [926.40 1772.74]	309.408 (28.23) [254.07 364.74]	-1040.165 (698.47) [-2409.17 328.84]
$\hat{\boldsymbol{\theta}}_g^{nw}$	1210.57 (353.50) [517.71 1903.44]	-21.45 (29.17) [-78.61779 35.71]	-1232.03 (56.56) [-1342.88 -1121.18]

(b) Treatment Domain Mean Effect Estimates for $\boldsymbol{\theta}_g^0 = E(Y_{ig}|I_{di} = 1)$

Estimators	Public - Private	Public - No Insurance	Private - No Insurance
$\hat{\boldsymbol{\theta}}_g^{(1)}$	3214.18 (32.22) [3151.03 3277.34]	811.56 (38.69) [735.73 887.39]	-2402.62 (46.48) [-2493.73 -2311.52]
$\hat{\boldsymbol{\theta}}_g^{nw}$	3320.93 (9.97) [3301.39 3340.47]	4.49 (2.69) [-0.77 9.76]	-3316.43 (240.85) [-3788.50 -2844.37]

Table 4.8 **Empirical Study Without Weights in Estimation of $\hat{\pi}_{2ig}$**: The treatment effect estimates using estimators $\hat{\boldsymbol{\theta}}_g^{nw}$ and $\hat{\boldsymbol{\theta}}_g^{(1)}$ defined in Section4.2.3. The parameter of interests are $\boldsymbol{\theta}_g^0 = E(Y_{ig})$ and $\boldsymbol{\theta}_g^0 = E(Y_{ig}|I_{di} = 1)$ where I_{di} is the indicator for the domain of interest that contains respondents who have sick or very sick physical condition. The standard errors are in parentheses and calculated using the Jackknife variance estimator, and the 95% confidence intervals are in brackets.

(a) Treatment Mean Effect Estimates for $\boldsymbol{\theta}_g^0 = E(Y_{ig})$

Estimators	Public - Private	Public - No Insurance	Private - No Insurance
$\hat{\boldsymbol{\theta}}_g^{(1)}$	1301.04 (150.81)	298.02 (42.79)	-1003.02 (169.31)
	[1005.45 1596.63]	[214.15 381.89]	[-1334.87 -671.17]
$\hat{\boldsymbol{\theta}}_g^{nw}$	1205.295 (259.68)	-13.23 (55.84)	-1218.52 (260.12)
	[696.32 1714.27]	[-122.68 96.22]	[-1728.36 -708.68]

(b) Treatment Domain Mean Effect Estimates for $\boldsymbol{\theta}_g^0 = E(Y_{ig}|I_{di} = 1)$

Estimators	Public - Private	Public - No Insurance	Private - No Insurance
$\hat{\boldsymbol{\theta}}_g^{(1)}$	2519.35 (239.67)	829.45 (87.41)	-1689.90 (257.46)
	[2049.60 2989.10]	[658.13 1000.77]	[-2194.52 -1185.28]
$\hat{\boldsymbol{\theta}}_g^{nw}$	3207.10 (17.14)	4.092 (4.30)	-2343.00 (180.83)
	[3173.51 3240.69]	[-4.34 12.52]	[-2697.43 -1988.57]

Table 4.9 STS-LogitLinear: Biases, Monte Carlo Standard Deviations, and Root Mean Square of Errors for estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ for different treatments.

		True Values	Bias				MC Std. Error				RMSE			
			$\hat{\boldsymbol{\theta}}_g^{(1)}$	$\hat{\boldsymbol{\theta}}_g^{(2)}$	$\hat{\boldsymbol{\theta}}_g^{p}$	$\hat{\boldsymbol{\theta}}_g^{nw}$	$\hat{\boldsymbol{\theta}}_g^{(1)}$	$\hat{\boldsymbol{\theta}}_g^{(2)}$	$\hat{\boldsymbol{\theta}}_g^{p}$	$\hat{\boldsymbol{\theta}}_g^{nw}$	$\hat{\boldsymbol{\theta}}_g^{(1)}$	$\hat{\boldsymbol{\theta}}_g^{(2)}$	$\hat{\boldsymbol{\theta}}_g^{p}$	$\hat{\boldsymbol{\theta}}_g^{nw}$
Trt 1	P_g	0.1584	0.0021	0.0019	-0.0013	0.0208	0.0247	0.0222	0.0208	0.0162	0.0021	0.0019	0.0013	0.0208
	μ_g	16.0002	-0.0432	-0.0259	-0.1013	1.8214	2.0933	1.2560	1.8165	1.3627	0.0432	0.0259	0.1013	1.8214
	σ_g^2	287.8674	5.3876	3.2325	-5.8530	107.4503	122.7497	73.6498	104.5842	80.4343	5.3876	3.2325	5.8530	107.4503
	R_g	0.0287	0.0063	0.0038	-0.0057	0.0692	0.0739	0.0443	0.0669	0.0525	0.0063	0.0038	0.0057	0.0692
	D_g	12.8079	0.0084	0.0076	0.1019	2.0162	1.7492	1.5742	2.0998	1.4850	0.0084	0.0076	0.1019	2.0162
Trt 2	P_g	0.2042	0.0014	0.0012	-0.0007	0.0233	0.0273	0.0245	0.0234	0.0173	0.0014	0.0012	0.0007	0.0233
	μ_g	1.4991	0.0027	0.0016	-0.0037	0.1256	0.1434	0.0860	0.1176	0.0940	0.0027	0.0016	0.0037	0.1256
	σ_g^2	4.2514	0.0723	0.0434	-0.0138	0.5336	0.5942	0.3565	0.5419	0.4133	0.0723	0.0434	0.0138	0.5336
	R_g	-0.0012	0.0026	0.0016	-0.0042	0.0592	0.0718	0.0431	0.0605	0.0453	0.0026	0.0016	0.0042	0.0592
	D_g	0.7003	0.0022	0.0020	-0.0045	0.0885	0.0964	0.0868	0.0882	0.0666	0.0022	0.0020	0.0045	0.0885
Trt 3	P_g	0.8415	-0.0010	-0.0009	-0.0016	0.0303	0.0291	0.0262	0.0238	0.0232	0.0010	0.0009	0.0016	0.0303
	μ_g	-16.0011	0.0743	0.0446	0.0458	2.3737	2.3841	1.4305	1.8297	1.8205	0.0743	0.0446	0.0458	2.3737
	σ_g^2	288.1261	10.3364	6.2018	-4.9257	134.6220	142.6968	85.6181	109.9074	99.0553	10.3364	6.2018	4.9257	134.6220
	R_g	-0.0291	-0.0016	-0.0009	0.0000	0.0883	0.0884	0.0530	0.0767	0.0650	0.0016	0.0009	0.0000	0.0883
	D_g	-12.8001	-0.0782	-0.0704	-0.0333	1.1742	1.3449	1.2104	1.3474	0.9442	0.0782	0.0704	0.0333	1.1742

Table 4.10 STS-JUMP: Biases, Monte Carlo Standard Deviations, and Root Mean Square of Errors for estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ for different treatments.

		True Value	Bias				MC Std. Error				RMSE			
			$\hat{\boldsymbol{\theta}}_g^{(1)}$	$\hat{\boldsymbol{\theta}}_g^{(2)}$	$\hat{\boldsymbol{\theta}}_g^{p}$	$\hat{\boldsymbol{\theta}}_g^{nw}$	$\hat{\boldsymbol{\theta}}_g^{(1)}$	$\hat{\boldsymbol{\theta}}_g^{(2)}$	$\hat{\boldsymbol{\theta}}_g^{p}$	$\hat{\boldsymbol{\theta}}_g^{nw}$	$\hat{\boldsymbol{\theta}}_g^{(1)}$	$\hat{\boldsymbol{\theta}}_g^{(2)}$	$\hat{\boldsymbol{\theta}}_g^{p}$	$\hat{\boldsymbol{\theta}}_g^{nw}$
	P_g	0.1583	0.0039	0.0035	-0.0316	0.0212	0.0358	0.0322	0.0657	0.0263	0.0039	0.0035	0.0316	0.0212
	μ_g	27.4995	-0.0402	-0.0241	-5.0323	1.4714	2.0711	1.2426	4.0882	1.6818	0.0402	0.0241	5.0323	1.4714
Trt 1	σ_g^2	972.9215	3.6327	2.1796	-337.3423	91.2239	112.2363	67.3418	199.9725	72.2744	3.6327	2.1796	337.3423	91.2239
	R_g	0.0312	-0.0001	-0.0000	0.0710	0.0835	0.0816	0.0490	0.1168	0.0631	0.0001	0.0000	0.0710	0.0835
	D_g	19.5189	-0.1568	-0.1411	-2.7680	3.2622	2.8777	2.5899	4.4791	2.0695	0.1568	0.1411	2.7680	3.2622
	P_g	0.2042	0.0067	0.0060	0.0050	0.0217	0.0594	0.0534	0.0500	0.0420	0.0067	0.0060	0.0050	0.0217
	μ_g	1.4991	0.0292	0.0175	-0.0404	0.1401	0.3245	0.1947	0.2480	0.2105	0.0292	0.0175	0.0404	0.1401
Trt 2	σ_g^2	4.2514	0.3046	0.1828	-0.4651	0.2893	1.2892	0.7735	1.0852	0.8574	0.3046	0.1828	0.4651	0.2893
	R_g	-0.0012	0.0174	0.0104	-0.0103	0.0621	0.1657	0.0994	0.1312	0.1139	0.0174	0.0104	0.0103	0.0621
	D_g	0.7003	-0.0014	-0.0013	0.1188	0.1204	0.1572	0.1415	0.1920	0.1028	0.0014	0.0013	0.1188	0.1204
	P_g	0.8415	-0.0064	-0.0057	0.0305	0.0421	0.0447	0.0403	0.0683	0.0304	0.0064	0.0057	0.0305	0.0421
	μ_g	-22.5016	-0.0360	-0.0216	-6.4636	-1.6614	3.9907	2.3944	9.8176	2.7043	0.0360	0.0216	6.4636	1.6614
Trt 3	σ_g^2	722.1559	10.4905	6.2943	164.8906	2.4137	166.0244	99.6147	452.1543	126.7102	10.4905	6.2943	164.8906	2.4137
	R_g	-0.0178	0.0046	0.0028	0.0789	0.0917	0.1724	0.1034	0.2320	0.0725	0.0046	0.0028	0.0789	0.0917
	D_g	-14.5078	0.0191	0.0172	0.3821	1.8477	1.1725	1.0552	1.0119	0.8569	0.0191	0.0172	0.3821	1.8477

Table 4.11 Informative-LogitLinear: Biases, Monte Carlo Standard Deviations, and Root Mean Square of Errors for estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ for different treatments.

		True Values	Bias				MC Std. Error				RMSE			
			$\hat{\boldsymbol{\theta}}_g^{(1)}$	$\hat{\boldsymbol{\theta}}_g^{(2)}$	$\hat{\boldsymbol{\theta}}_g^{p}$	$\hat{\boldsymbol{\theta}}_g^{nw}$	$\hat{\boldsymbol{\theta}}_g^{(1)}$	$\hat{\boldsymbol{\theta}}_g^{(2)}$	$\hat{\boldsymbol{\theta}}_g^{p}$	$\hat{\boldsymbol{\theta}}_g^{nw}$	$\hat{\boldsymbol{\theta}}_g^{(1)}$	$\hat{\boldsymbol{\theta}}_g^{(2)}$	$\hat{\boldsymbol{\theta}}_g^{p}$	$\hat{\boldsymbol{\theta}}_g^{nw}$
	P_g	0.1584	0.0356	0.0320	-0.0038	0.0176	0.1018	0.0916	0.0672	0.0126	0.0356	0.0320	0.0038	0.0176
	μ_g	16.0002	1.7118	1.0271	-0.7409	1.8568	17.7836	10.6701	4.7429	1.4249	1.7118	1.0271	0.7409	1.8568
Trt 3	σ_g^2	287.8674	32.6937	19.6162	0.2747	157.2096	931.2460	558.7476	188.9476	117.0559	32.6937	19.6162	0.2747	157.2096
	R_g	0.0287	0.0440	0.0264	-0.0104	0.0599	0.2635	0.1581	0.1180	0.0454	0.0440	0.0264	0.0104	0.0599
	D_g	12.8079	0.6684	0.6015	-0.2855	2.4909	6.4879	5.8391	2.5865	1.8438	0.6684	0.6015	0.2855	2.4909
	P_g	0.2042	0.0042	0.0038	0.0043	0.0210	0.0832	0.0749	0.0675	0.0158	0.0042	0.0038	0.0043	0.0210
	μ_g	1.4991	0.0545	0.0327	-0.0279	0.1393	0.5712	0.3427	0.3522	0.1030	0.0545	0.0327	0.0279	0.1393
Trt 3	σ_g^2	4.2514	0.2298	0.1379	-0.0802	0.9065	2.1921	1.3153	0.9470	0.6563	0.2298	0.1379	0.0802	0.9065
	R_g	-0.0012	0.0047	0.0028	-0.0057	0.0711	0.1773	0.1064	0.1238	0.0515	0.0047	0.0028	0.0057	0.0711
	D_g	0.7003	-0.0091	-0.0082	-0.0066	0.0920	0.1770	0.1593	0.1347	0.0691	0.0091	0.0082	0.0066	0.0920
	P_g	0.8415	-0.0102	-0.0091	-0.0124	0.0135	0.0713	0.0642	0.0642	0.0105	0.0102	0.0091	0.0124	0.0135
	μ_g	-16.0011	-0.8802	-0.5281	-0.6803	1.9857	8.4257	5.0554	5.6402	1.4650	0.8802	0.5281	0.6803	1.9857
Trt 3	σ_g^2	288.1261	16.3605	9.8163	-9.0152	187.3405	316.0102	189.6061	198.6736	140.2728	16.3605	9.8163	9.0152	187.3405
	R_g	-0.0291	0.0039	0.0024	-0.0028	0.0995	0.2294	0.1376	0.1413	0.0683	0.0039	0.0024	0.0028	0.0995
	D_g	-12.8001	-0.4545	-0.4090	-0.1292	1.2369	3.1189	2.8070	2.1273	0.8995	0.4545	0.4090	0.1292	1.2369

Table 4.12 Informative-JUMP: Biases, Monte Carlo Standard Deviations, and Root Mean Square of Errors for estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ for different treatments.

		True Values	Bias				MC Std. Error				RMSE			
			$\hat{\boldsymbol{\theta}}_g^{(1)}$	$\hat{\boldsymbol{\theta}}_g^{(2)}$	$\hat{\boldsymbol{\theta}}_g^{p}$	$\hat{\boldsymbol{\theta}}_g^{nw}$	$\hat{\boldsymbol{\theta}}_g^{(1)}$	$\hat{\boldsymbol{\theta}}_g^{(2)}$	$\hat{\boldsymbol{\theta}}_g^{p}$	$\hat{\boldsymbol{\theta}}_g^{nw}$	$\hat{\boldsymbol{\theta}}_g^{(1)}$	$\hat{\boldsymbol{\theta}}_g^{(2)}$	$\hat{\boldsymbol{\theta}}_g^{p}$	$\hat{\boldsymbol{\theta}}_g^{nw}$
	P_g	0.1583	0.0216	0.0194	-0.0377	-0.0866	0.0938	0.0844	0.0947	0.0276	0.0216	0.0194	0.0377	0.0866
	μ_g	27.4995	0.1657	0.0994	-7.7266	25.7351	9.4555	5.6733	6.3092	2.0310	0.1657	0.0994	7.7266	25.7351
Trt 1	σ_g^2	972.9215	235.7866	141.4720	-530.3687	901.6896	621.9053	373.1432	253.3727	148.6468	235.7866	141.4720	530.3687	901.6896
	R_g	0.0312	-0.0111	-0.0067	0.0353	0.1395	0.2412	0.1447	0.1865	0.0929	0.0111	0.0067	0.0353	0.1395
	D_g	19.5189	2.2467	2.0221	-3.7765	25.0799	8.7762	7.8986	4.6672	2.7718	2.2467	2.0221	3.7765	25.0799
	P_g	0.2042	0.0139	0.0125	0.0253	0.0406	0.1135	0.1021	0.1041	0.0447	0.0139	0.0125	0.0253	0.0406
	μ_g	1.4991	0.5548	0.3329	0.0823	0.7385	2.1396	1.2838	0.5119	0.2124	0.5548	0.3329	0.0823	0.7385
Trt 2	σ_g^2	4.2514	0.7456	0.4473	0.4392	4.4726	4.1332	2.4799	1.7006	1.3357	0.7456	0.4473	0.4392	4.4726
	R_g	-0.0012	0.0205	0.0123	0.0169	0.1023	0.3361	0.2017	0.1747	0.1227	0.0205	0.0123	0.0169	0.1023
	D_g	0.7003	-0.0368	-0.0331	-0.0476	0.0504	0.3322	0.2989	0.3527	0.1212	0.0368	0.0331	0.0476	0.0504
	P_g	0.8415	-0.0016	-0.0014	0.0870	0.1234	0.0592	0.0533	0.0529	0.0084	0.0016	0.0014	0.0870	0.1234
	μ_g	-22.5016	0.9339	0.5604	-27.9144	-18.7933	6.7477	4.0486	26.0017	2.5644	0.9339	0.5604	27.9144	18.7933
Trt 3	σ_g^2	722.1559	17.4960	10.4976	1068.3398	614.4755	253.0702	151.8421	1156.2205	265.7647	17.4960	10.4976	1068.3398	614.4755
	R_g	-0.0178	0.0063	0.0038	-0.0400	0.1043	0.1875	0.1125	0.2908	0.0729	0.0063	0.0038	0.0400	0.1043
	D_g	-14.5078	-0.2283	-0.2054	0.1097	-5.3154	2.3139	2.0825	1.4489	0.7884	0.2283	0.2054	0.1097	5.3154

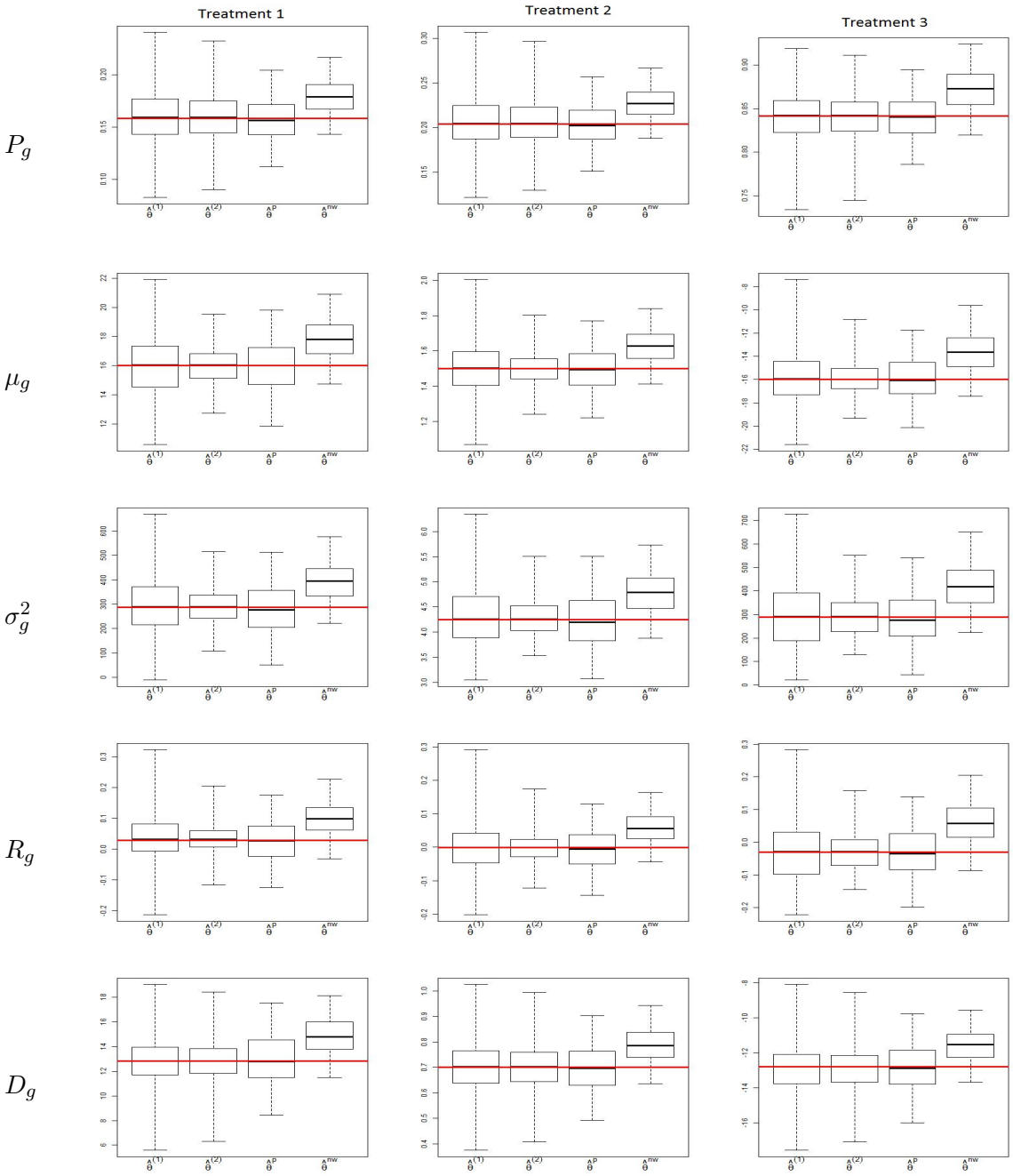

Figure 4.1 **STS-LogitLinear**: Boxplots of MC estimates of the four estimators for all treatments. Each row represents a parameter, and each column represents a treatment. In each subplot, the four boxplots are for $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ respectively in order. The horizontal line is located at the value of the true treatment effect.

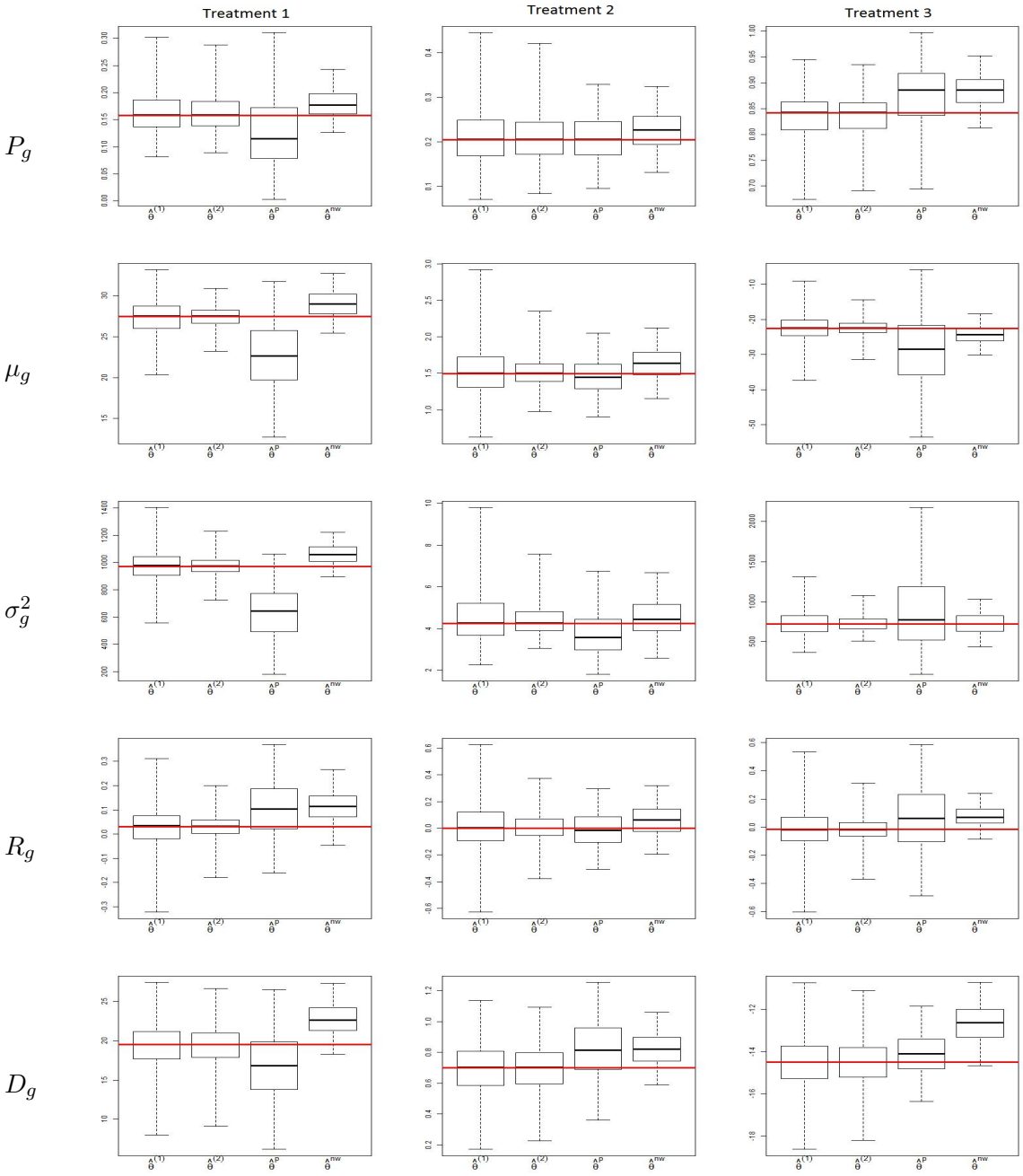

Figure 4.2 **STS-JUMP**: Boxplots of MC estimates of the four estimators for all treatments. Each row represents a parameter, and each column represents a treatment. In each subplot, the four boxplots are for $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ respectively in order. The horizontal line is located at the value of the true treatment effect.

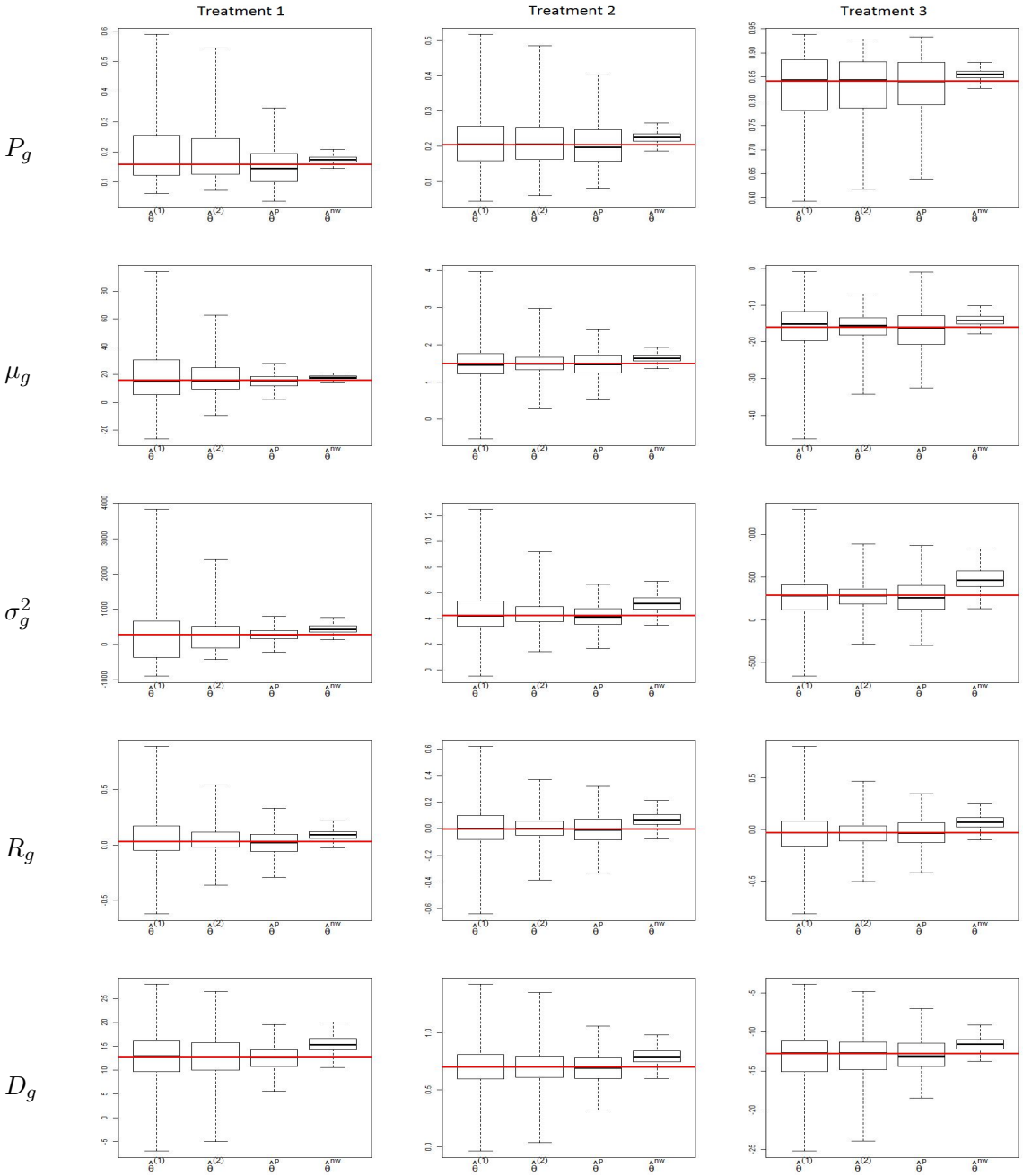

Figure 4.3 **Informative-LogitLinear**: Boxplots of MC estimates of the four estimators for all treatments. Each row represents a parameter, and each column represents a treatment. In each subplot, the four boxplots are for $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ respectively in order. The horizontal line is located at the value of the true treatment effect.

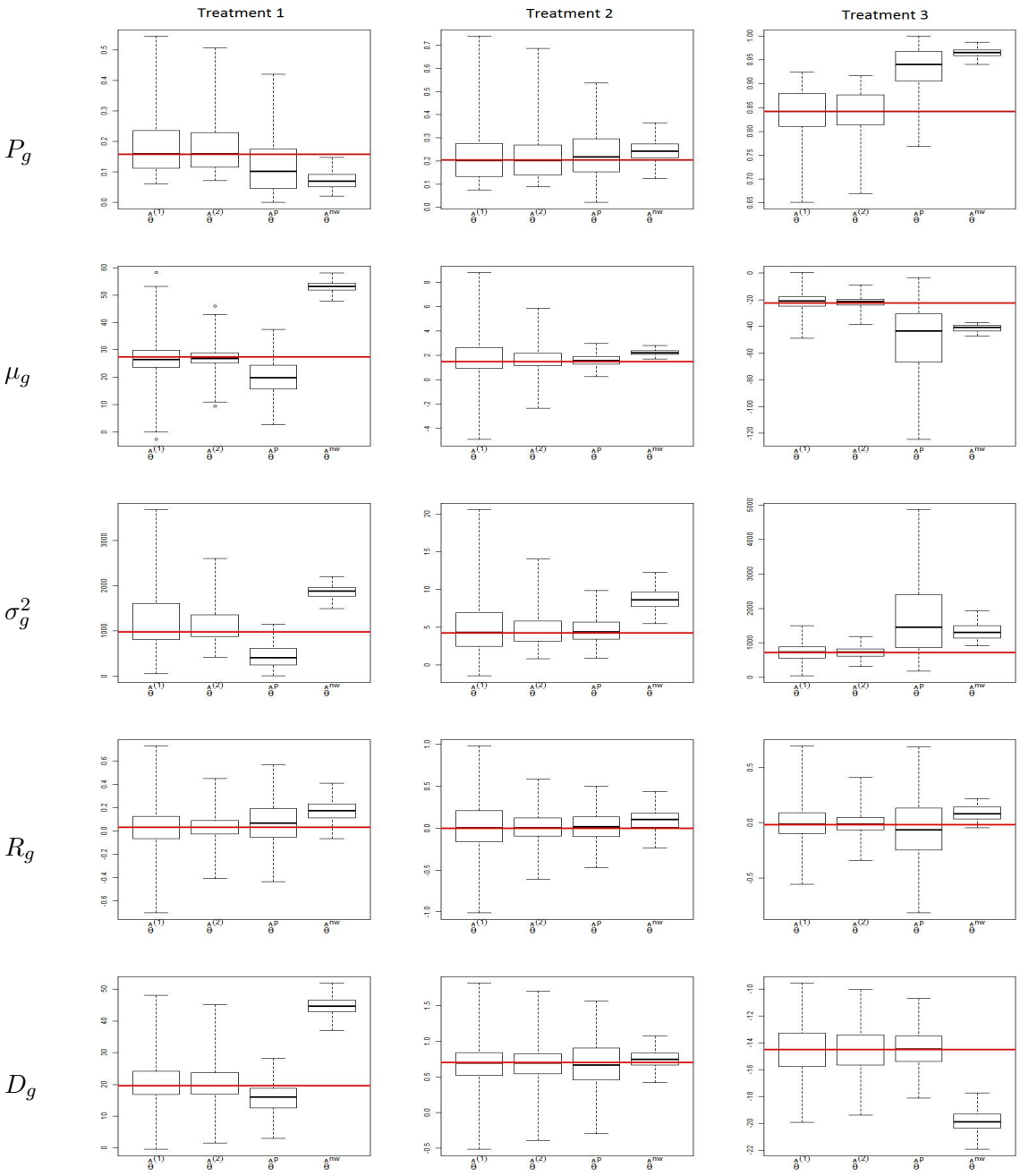

Figure 4.4 **Informative-JUMP**: Boxplots of MC estimates of the four estimators for all treatments. Each row represents a parameter, and each column represents a treatment. In each subplot, the four boxplots are for $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ respectively in order. The horizontal line is located at the value of the true treatment effect.

CHAPTER 5. SUMMARY AND DISCUSSION

In this work theoretical and practical aspects of penalized b-spline smoothing, a semi-parametric smoothing technique which gained much popularity over the last two decades, were considered. In Chapter 2, P-splines were presented as a smoothing technique that combines different splines basis, penalties and knots to provide a very flexible technique for fitting data. How these splines were developed and some practical applications of P-splines and extensions to bivariate smoothing and variance estimation were also discussed.

In Chapter 3, theoretical and empirical contributions were made using the penalized Bivariate Tensor Product B-spline (BTPB). Theoretically, the asymptotic theory for any general partial derivatives of penalized BTPB estimators was developed as well as a variance estimator. The validity and applicability of this estimator was verified through theoretical proofs and simulations. Empirically, this estimator was used to address the actuarial fairness question in the premium rates prescribed by the Risk Management Agency. This was done by applying the estimator to data from Webster County, Iowa, a county in the center of the corn belt with generally productive land and low intra-county land variance. This estimated density curve was then compared with the conditional yield density estimated using the historical yield data estimated through the use of kernel density to determine if premium prices are actuarially fair.

For all coverage levels, the actual premium rates are close to what was calculated given historical data. Conditional on historical average yields, however, premiums for poorer land were too low in general, and those for the best land are too high for the farmer to expect to recover premium costs, even after accounting for large premium subsidies. However, as land quality in Webster County is quite uniform when compared with other counties, even in Iowa, the county is not good candidate for identifying rate discrepancies induced

by heterogeneity in cropped land productivity. It remains to be seen whether the patterns observed in this one county apply more generally.

P-splines were applied to the area of observations studies for complex surveys in Chapter 4. The main focus of Chapter 4 was on the inclusion of survey weights in observation data when using inverse probability weighted (IPW) estimator and not the application of P-splines. However, P-splines do play a big role in estimating the self-selection probabilities. It is very common that people ignore survey weights in observational data when using the IPW estimators, yet claim that the estimated treatment effects are generalizable to the target population, causing misleading guidance in causal inference. Failure to properly account for the complex survey design may lead to biased treatment effect estimates and incorrect variance estimation. The general idea is to multiply the inverse of the estimated propensity scores by the sampling design weights. Chapter 4 considered a set up in which parameters of interest are defined through estimation equations, and uses the generalized method of moments (GMM) for parameter estimation. The observational data set is considered as a first-phase sample from a finite population and has a known sampling probability. The second-phase sampling is a partitioning of the first-phase sample (observational data set) into mutually exclusive and self-selected treatment groups. This partitioning in the second-phase can be viewed as a multinomial sampling in survey statistics, and its self-selection probabilities are estimated using P-splines.

The feasibility of the method was investigated by estimating the mean annual medical expenditures under different choices of health insurance types in China. The data is from the Chinese General Social Survey (CGSS). The response variable in the study was the annual medical expenditure, with treatment variables being health insurance type (public health insurance, private health insurance, and no health insurance). It was found that when the data weights are neglected, the estimated mean medical expenditure of the public health insurance group is not significantly different from that of the no health insurance group. However, when the data weights are incorporated, the public health group is found

to spend significantly more on the medical expenses than the no health insurance group. In addition, when the data weights are neglected for the treatment mean effect estimates, the estimated mean medical expenditure of the private health insurance group is significantly different from that of the no insurance group, while incorporating the data weights finds these estimated means not significantly different. This study demonstrates that the estimators presented are feasible in real data application and suggests that ignoring the weights of an observational data might lead to a misleading conclusion.

While P-splines are by no means near the end of their development cycle, this work has made significant progress in pushing them towards that end. In doing so, P-splines practicality has greatly increased, particularly for bivariate uses and variance estimation, as seen in Chapter 3. In addition, it is plain to see that p-splines have much more to contribute to this century (Eilers et al., 2015).

Bibliography

Akerlof, G. A. (1970). The market for lemons: Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500.

Ashmead, R. D. (2014). *Propensity Score Methods for Estimating Causal Effects from Complex Survey Data*. PhD thesis, The Ohio State University.

Berg, E., Kim, J.-K., and Skinner, C. (2016). Imputation under informative sampling. *Journal of Survey Statistics and Methodology*, 4(4):436–462.

Berger, Y. G. and Skinner, C. J. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):79–89.

Bézier, P. (1968). *How Renault uses numerical control for car body design and tooling*. Society of Automotive Engineers.

Breidt, F., Claeskens, G., and Opsomer, J. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika*, 92(4):831–846.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155(2):138–154.

Claassen, R., Langpap, C., and Wu, J. (2017). Impacts of federal crop insurance on land use and environmental quality. *American Journal of Agricultural Economics*, 99(3):592–613.

Claeskens, G., Krivobokova, T., and Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3):529–544.

Coble, K. H., Knight, T. O., Goodwin, B. K., Miller, M. F., Rejesus, R. M., and Duffield, G. (2010). A comprehensive review of the rma aph and combo rating methodology: Final report.

De Boor, C. (1976). Splines as linear combinations of b-splines. a survey. Technical report, Wisconsin Univ Madison Mathematics Research Center.

De Boor, C. (1978). *A Practical Guide to Splines*. Number v. 27 in Applied Mathematical Sciences. Springer-Verlag.

de Casteljau, P. (1959). Courbes et surfaces a poles. Technical report, S. A. Andre Citroen, Paris.

Du, X., Feng, H., and Hennessy, D. A. (2017). Rationality of choices in subsidized crop insurance markets. *American Journal of Agricultural Economics*, 99(3):732–756.

Du, X., Feng, H., Hennessy, D. A., and Arrora, G. (2018). Land resilience and tail dependence among crop yield distributions. *American Journal of Agricultural Economics*, aax082:https://doi.org/10.1093/ajae/aax082.

Du, X., Hennessy, D. A., and Feng, H. (2013). A natural resource theory of us crop insurance contract choice. *American Journal of Agricultural Economics*, 96(1):232–252.

Du, X., Hennessy, D. A., and Yu, C. L. (2012). Testing days conjecture that more nitrogen decreases crop yield skewness. *American Journal of Agricultural Economics*, 94(1):225–237.

Du, X., Yu, C. L., Hennessy, D. A., and Miao, R. (2015). Geography of crop yield skewness. *Agricultural Economics*, 46(4):463–473.

DuGoff, E. H., Schuler, M., and Stuart, E. A. (2014). Generalizing observational study results: applying propensity score methods to complex surveys. *Health services research*, 49(1):284–303.

Eeckhoudt, L., Gollier, C., and Schlesinger, H. (2005). *Economic and Financial Decisions Under Risk*. Princeton paperbacks. Princeton University Press.

Eilers, P. H. and Marx, B. D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems*, 66(2):159 – 174.

Eilers, P. H., Marx, B. D., and Durbán, M. (2015). Twenty years of p-splines. *SORT-Statistics and Operations Research Transactions*, 39(2):149–186.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statist. Sci.*, 11(2):89–121.

Feng, H., Hennessy, D. A., and Miao, R. (2013). The effects of government payments on cropland acreage, conservation reserve program enrollment, and grassland conversion in the dakotas. *American Journal of Agricultural Economics*, 95(2):412–418.

Fuller, W. A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4):933–944.

Fuller, W. A. (2011). *Sampling statistics*, volume 560. John Wiley & Sons.

Gardner, B. L. (2009). *American Agriculture in the Twentieth Century: How It Flourished and What It Cost*. Harvard University Press Cambridge.

Glauber, J. (2015). Agricultural insurance and the world trade organization. Technical report, International Food Policy Research Institute, Washington DC. Discussion Paper 01473.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331.

Hall, P. and Opsomer, J. D. (2005). Theory for penalised spline regression. *Biometrika*, 92(1):105–118.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag New York, 2 edition.

Haziza, D. and Rao, J. N. (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology*, 32(1):53.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96.

Kauermann, G., Krivobokova, T., and Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):487–503.

Ker, A. P. and Goodwin, B. K. (2000). Nonparametric estimation of crop insurance rates revisited. *American Journal of Agricultural Economics*, 82(2):463–478.

Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statist. Sinica*, 24:375–94.

Kim, J. K., Navarro, A., and Fuller, W. A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American statistical association*, 101(473):312–320.

Li, Y. and Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, 95(2):415–436.

Little, R. J. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77(378):237–250.

Lobachevsky, N. (1902). *Pangeometrie*. Leipzig, W. Engelmann,.

Lorentz, G. (1986). Approximation of functions, chelsea publ. *Co., New York*.

Lusk, J. L. (2016). Distributional effects of crop insurance subsidies. *Applied Economic Perspectives and Policy*, 39(1):1–15.

Marx, B. D. and Eilers, P. H. (2005). Multidimensional penalized signal regression. *Technometrics*, 47(1):13–22.

Messer, K. and Goldstein, L. (1993). A new class of kernels for nonparametric curve estimation. *Ann. Statist.*, 21(1):179–195.

Miao, R., Hennessy, D. A., and Feng, H. (2012). The effects of crop insurance subsidies and sodsaver on land use change. *Journal of Agricultural and Resource Economics*, 41(2):247–65.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1(4):502–518.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing*, 9(2):363–379.

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society*, pages 1027–1057.

Patrikalakis, N. M., Maekawa, T., and Cho, W. (2009). Definition of bézier curve and its properties. http://web.mit.edu/hyperbook/Patrikalakis-Maekawa-Cho/node12.html.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? why is it a problem? how can we approach it. *Survey Methodology*, 37(2):115–136.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 166–186.

Prochazkova, J. (2005). Derivative of b-spline function. In *Proceedings of the 25th Conference on Geometry and Computer Graphics, Prague, Czech Republic*.

Ramsay, J. O. and Silverman, B. W. (2007). *Applied functional data analysis: methods and case studies*. Springer.

Rice, J. and Rosenblatt, M. (1983). Smoothing splines: Regression, derivatives and deconvolution. *Ann. Statist.*, 11(1):141–156.

Ridgeway, G., Kovalchik, S. A., Griffin, B. A., and Kabeto, M. U. (2015). Propensity score analysis with survey weighted data. *Journal of Causal Inference*, 3(2):237–249.

Robins, J., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. *Statistical Science*, 22(4):544–559.

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

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4):735–757.

Santeramo, F. G. and Ford Ramsey, A. (2017). Crop insurance in the eu: Lessons and caution from the us. *EuroChoices*, 16(3):34–39.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Schneider, P. J. (1996). Nurb curves: A guide for the uninitiated. *develop*, (25):48–74.

Seber, G. A. (2008). *A matrix handbook for statisticians*, volume 15. John Wiley & Sons.

Sheather, S. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53:683–690.

Shene, C.-K. (2014). Introduction to computing with geometry notes. Website.

Shields, D. A. (2009). Federal crop insurance: Background and issues. Library of Congress, Congressional Research Service.

Silverman, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics*, 12(3):898–916.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682.

Tannura, M. A., Irwin, S. H., and Good, D. L. (2008). Weather, technology, and corn and soybean yields in the us corn belt. *Marketing and outlook research report*, Department of Agricultural and Consumer Economics:University of Illinois at Urbana–Champaign.

Theisse, K. (2009). Crop insurance decisions. http://www.cornandsoybeandigest.com/crop-insurance-decisions-0.

Wang, X., Shen, J., and Ruppert, D. (2009). Local Asymptotics of P-Spline Smoothing. *ArXiv e-prints*.

Weber, J. G., Key, N., and ODonoghue, E. (2016). Does federal crop insurance make environmental externalities from agriculture worse? *Journal of the Association of Environmental and Resource Economists*, 3(3):707–742.

Weierstrass, K. (1885). On the analytic representability of so-called arbitrary functions of a real variable : First indication. In *Session reports of the Royal PREUSSIAN Academy of Sciences at Berlin*, pages 633–639.

westlawn5554X (2006). https://www.boatdesign.net/threads/spline-weights-too-heavy-costly-to-delivery.9244/page-2.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

Wright, B. D. (2014). Multiple peril crop insurance. *Choices*, 29(3).

Xiao, L., Li, Y., Apanasovich, T. V., and Ruppert, D. (2012). Local Asymptotics of P-splines. *ArXiv e-prints*.

Xiao, L., Li, Y., and Ruppert, D. (2010). Fast Bivariate Penalized Splines: the Sandwich Smoother. *ArXiv e-prints*.

Xu, Z., Hennessy, D., Sardana, K., and Moschini, G. (2012). The realized yield effect of gm crops: Us maize and soybean. *Essays on GMO effects on crop yields, the effects of pricing errors on implied volatilities and smoothing for seasonal time series with a long cycle*, page 3.

Yoshida, T. (2013). Asymptotics for penalized spline estimators in quantile regression. *Communications in Statistics - Theory and Methods*, page doi:10.1080/03610926.2013.765477.

Yu, C., Legg, J., Liu, B., et al. (2013). Estimating multiple treatment effects using two-phase semiparametric regression estimators. *Electronic Journal of Statistics*, 7:2737–2761.

Yu, J., Smith, A., and Sumner, D. A. (2017). Effects of crop insurance premium subsidies on crop acreage. *American Journal of Agricultural Economics*.

Zanutto, E. L. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of data Science*, 4(1):67–91.

Zhang, Y. Y. (2017). A density-ratio model of crop yield distributions. *American Journal of Agricultural Economics*, 99(5):1327–1343.

APPENDIX A. CHAPTER 3 APPENDIX

In this appendix, the proofs for Lemma 1, Theorem 1 and Theorem 2 for chapter 3 can be found in sections A1, A2, and A3 respectively.

A1: Proof of Lemma 1

By the proof of proposition 1 in Xiao et al. (2012),

$$\hat{\mu}(x,z) - \mu^*(x,z) = \sum_{i,j} b_{i,j}(x,z) y_{i,j}, \tag{A1.1}$$

where

$$b_{i,j}(x,z) = \frac{1}{n_1 h_1} H_{m_1}\left(\frac{x-x_i}{h_1}\right) d_{i,2}(z) + \frac{1}{n_2 h_2} H_{m_2}\left(\frac{z-z_i}{h_2}\right) d_{i,1}(x) + \tilde{b}_{i,j}(x,z),$$

$$d_{i,1}(x) = \sum_{k,r} B_k^1(x) B_r^1(x_i) S_{k,r,x} - \frac{1}{n_1 h_1} H_{m_1}\left(\frac{x-x_i}{h_1}\right),$$

$$d_{i,2}(z) = \sum_{l,s} B_l^2(z) B_s^2(z_i) S_{l,s,z} - \frac{1}{n_2 h_2} H_{m_2}\left(\frac{z-z_i}{h_2}\right).$$

The terms $S_{k,r,x}$, $S_{l,s,z}$, and $\tilde{b}_{i,j}(x,z)$ are defined as follows. Let $q_1 = \max(d_1, m_1)$ and $q_2 = \max(d_2, m_2)$. There exist vectors $S_{k,x}$ and a constant $c_1 > 0$ so that for $q_1 < j < I_1 - q_1$, $S_{k,x}^\top \Lambda_{1,j} = \delta_{k,j}$, and for $1 \le j \le q_1$, or $I_1 - q_1 \le j \le I_1, S_{k,x}^\top \Lambda_{1,j} = O\Big[\exp\big\{ - c_1 h_1^{-1} \min(x, 1-x)\big\}\Big]$. Here $\Lambda_{1,j}$ is the j^{th} column of $\Lambda_1 = B_1^T B_1 + \lambda_1 D_{m_1}^T D_{m_1}$ and $\delta_{k,j} = 1$ if $j = k$ and 0 otherwise. Similarly, there exist vector $S_{l,z}$ and a constant $c_2 > 0$ so that for $q_2 < j < I_2 - q_2$, $S_{l,z}^\top \Lambda_{2,j} = \delta_{l,j}$, and for $1 \le j \le q_2$, or $I_2 - q_2 \le j \le I_2, S_{l,z}^\top \Lambda_{2,j} = O\Big[\exp\big\{ - c_2 h_2^{-1} \min(z, 1-z)\big\}\Big]$.

Define $\tilde{a}_{k,l} = \Big(S_{l,z} \otimes S_{k,x}\Big)^{\top} \Big(\Lambda_2 \otimes \Lambda_1\Big)\hat{a}$, $c = \min\big\{c_1 \min(x, 1-x), c_2 \min(z, 1-z)\big\}$, and $\tilde{b}_{i,j,k,l}$ such that

$$\tilde{a}_{k,l} - \hat{a}^{*}_{k,l} = \sum_{i,j} \tilde{b}_{i,j,k,l}(x,z) y_{i,j}.$$

So

$$\hat{\mu}(x,z) = \sum_{i,j} y_{i,j} \Big[\big\{ \sum_{k,r} B^1_k(x) B^1_r(x_i) S_{k,r,x} \big\} \big\{ \sum_{l,s} B^2_l(z) B^2_s(z_i) S_{l,s,z} \big\} + \tilde{b}_{i,j}(x,z) \Big],$$

where

$$\tilde{b}_{i,j}(x,z) = \sum_{k,l} \tilde{b}_{i,j,k,l} B^1_k(x) B^2_l(z),$$

and

$$\tilde{b}_{i,j}(x,z) = O\Big[\exp\big\{ - c \min(h_1^{-1}, h_2^{-1}) \big\} \Big].$$

Since $h_1 = O\big(n^{-v_1}\big)$ and $h_2 = O\big(n^{-v_2}\big)$, $\tilde{b}_{i,j}(x,z) = n^{-1} o(\zeta)$.

Based on equation (A1.1),

$$\hat{\mu}^{(v_1,v_2)}(x,z) - \mu^{*(v_1,v_2)}(x,z) = \sum_{i,j} b^{(v_1,v_2)}_{i,j}(x,z) y_{i,j},$$

and

$$b_{i,j}^{(v_1,v_2)}(x,z) = \underbrace{\frac{1}{n_1 h_1}\frac{1}{h_1^{v_1}} H_{m_1}^{(v_1)}\left(\frac{x-x_i}{h_1}\right) d_{i,2}^{(v_2)}(z)}_{\widehat{=}A_{ij}}$$
$$+ \underbrace{\frac{1}{n_2 h_2}\frac{1}{h_2^{v_2}} H_{m_2}^{(v_2)}\left(\frac{z-z_i}{h_2}\right) d_{i,1}^{(v_1)}(x)}_{\widehat{=}B_{ij}}$$
$$+ \underbrace{d_{i,1}^{(v_1)}(x) d_{i,2}^{(v_2)}(z)}_{\widehat{=}C_{ij}} + \underbrace{\tilde{b}_{i,j}^{(v_1,v_2)}(x,z)}_{\widehat{=}D_{ij}}. \tag{A1.2}$$

Hence,

$$E\Big\{\hat{\mu}^{(v_1,v_2)}(x,z) - \mu^{*(v_1,v_2)}(x,z)\Big\} = \sum_{i,j} b_{i,j}^{(v_1,v_2)}(x,z)\mu(x_i,z_i), \tag{A1.3}$$

and

$$Var\Big\{\hat{\mu}^{(v_1,v_2)}(x,z) - \mu^{*(v_1,v_2)}(x,z)\Big\} = \sum_{i,j} \left\{b_{i,j}^{(v_1,v_2)}(x,z)\right\}^2 \sigma^2(x_i,z_i). \tag{A1.4}$$

For the term with D_{ij}:

$$\sum_{i,j} |D_{ij}\mu(x_i,z_j)| = \sum_{i,j} |\tilde{b}_{i,j}^{(v_1,v_2)}(x,z)\mu(x_i,z_j)| = o(\zeta), \tag{A1.5}$$

where the last equality holds because $\tilde{b}_{i,j}^{(v_1,v_2)}(x,z) = n^{-1}o(\zeta)$.

For the term with A_{ij}: Begin by showing how to obtain the expression for $d_{i,2}^{(v_2)}(z)$. Define $\tilde{\lambda}_i = \lambda_i K_i^* n_i^{-1}$. By proposition 1 of Xiao et al. (2012), there exist $\phi_1, \phi_2 < \infty$ such that,

$$n_1 h_1 \sum_{k,r} B_k^1(x) B_r^1(x_i) S_{k,r,x} = H_{m_1}\left(\frac{x-x_i}{h_1}\right) + \delta_{d_1 > m_1} O_{11} + \exp\left(-\phi_2 \frac{|x-x_i|}{h_1}\right) O_{12} \tag{A1.6}$$

where

$$O_{11} = O\Big(\tilde{\lambda}_1^{-2+\frac{1}{2m_1}}\Big) + \delta_{\left\{|x-x_i|<\frac{\phi_1}{K_1^*}\right\}} O\Big(\tilde{\lambda}_1^{-\frac{d_1}{d_1-m_1}+\frac{1}{2m_1}}\Big),$$
$$O_{12} = O\Big(\tilde{\lambda}_1^{-\frac{1}{m_1}}\Big) + \delta_{\{m_1=1\}}\delta_{\left\{|x-x_i|\le(d_1+1)\tilde{\lambda}_1^{-\frac{1}{2m_1}}\right\}} O\Big(\tilde{\lambda}_1^{-\frac{1}{2m_1}}\Big);$$

Similarly, there exist $\phi_3, \phi_4 < \infty$ such that,

$$n_2h_2\sum_{l,s} B_l^2(z)B_s^2(z_i)S_{l,s,z} = H_{m_2}\Big(\frac{z-z_i}{h_2}\Big) + \delta_{d_2>m_2}O_{21} + \exp\Big(-\phi_4\frac{|z-z_i|}{h_2}\Big)O_{22} \quad \text{(A1.7)}$$

where

$$O_{21} = O\Big(\tilde{\lambda}_2^{-2+\frac{1}{2m_2}}\Big) + \delta_{\left\{|z-z_i|<\frac{\phi_2}{K_2^*}\right\}} O\Big(\tilde{\lambda}_1^{-\frac{d_2}{d_2-m_2}+\frac{1}{2m_2}}\Big),$$
$$O_{22} = O\Big(\tilde{\lambda}_2^{-\frac{1}{m_2}}\Big) + \delta_{\{m_2=1\}}\delta_{\left\{|z-z_i|\le(d_2+1)\tilde{\lambda}_2^{-\frac{1}{2m_2}}\right\}} O\Big(\tilde{\lambda}_2^{-\frac{1}{2m_2}}\Big);$$

By equation (A1.6),

$$n_2h_2\sum_{l,s} B_l^{2(v_2)}(z)B_s^2(z_i)S_{l,s,z} = \frac{1}{h_2^{v_2}}H_{m_2}^{(v_2)}\Big(\frac{z-z_i}{h_2}\Big) + \delta_{d_2>m_2}O_{21} + \frac{(-\phi_4)^{v_2}}{h_2^{v_2}}\exp\Big(-\phi_4\frac{|z-z_i|}{h_2}\Big)O_{22}$$

Hence,

$$d_{i,2}^{(v_2)}(z) = \frac{1}{n_2h_2}\frac{1}{h_2^{v_2}}H_{m_2}^{(v_2)}\Big(\frac{z-z_i}{h_2}\Big) + \frac{1}{n_2h_2}\delta_{d_2>m_2}O_{21} + \frac{1}{n_2h_2}\frac{(-\phi_4)^{v_2}}{h_2^{v_2}}\exp\Big(-\phi_4\frac{|z-z_i|}{h_2}\Big)O_{22}$$
$$- \frac{1}{n_2h_2}\frac{1}{h_2^{v_2}}H_{m_2}^{(v_2)}\Big(\frac{z-z_i}{h_2}\Big).$$

Then

$$A_{ij} = \frac{1}{n_1h_1}\frac{1}{h_1^{v_1}}H_{m_1}^{(v_1)}\Big(\frac{x-x_i}{h_2}\Big)\Big[\frac{1}{n_2h_2}\delta_{d_2>m_2}O_{21} + \frac{1}{n_2h_2}\frac{(-\phi_4)^{v_2}}{h_2^{v_2}}\exp\Big(-\phi_4\frac{|z-z_i|}{h_2}\Big)O_{22}\Big]$$

Using similar arguments in proposition 1 of Xiao et al. (2012), obtain

$$\sum_{i,j}\Big|\frac{1}{h_1^{v_1}}\frac{1}{n_1h_1}\frac{1}{n_2h_2}H_{m_1}^{(v_1)}\Big(\frac{x-x_i}{h_2}\Big)\delta_{d_2>m_2}O_{21}\mu(x_i,z_i)\Big| = O(\zeta h_1^{-v_1}) = O(\zeta h_1^{-v_1}h_2^{-v_2})$$

and

$$\sum_{i,j}\Big|\frac{1}{h_1^{v_1}}\frac{(-\phi_4)^{v_2}}{h_2^{v_2}}\frac{1}{n_1h_1}\frac{1}{n_2h_2}H_{m_1}^{(v_1)}\Big(\frac{x-x_i}{h_2}\Big)\exp\Big(-\phi_4\frac{|z-z_i|}{h_2}\Big)O_{22}\mu(x_i,z_i)\Big| = O(\zeta h_1^{-v_1}h_2^{-v_2})$$

Therefore

$$\sum_{i,j}\Big|A_{ij}\mu(x_i,z_i)\Big| = O(\zeta h_1^{-v_1}h_2^{-v_2}) \tag{A1.8}$$

For the term with B_{ij}: First derive

$$d_{i,1}^{(v_1)}(x) = \frac{1}{n_1h_1}\frac{1}{h_1^{v_1}}H_{m_1}^{(v_1)}\Big(\frac{x-x_i}{h_1}\Big) + \frac{1}{n_1h_1}\delta_{d_1>m_1}O_{11} + \frac{1}{n_1h_1}\frac{(-\phi_2)^{v_1}}{h_1^{v_1}}\exp\Big(-\phi_2\frac{|x-x_i|}{h_1}\Big)O_{12}$$
$$-\frac{1}{n_1h_1}\frac{1}{h_1^{v_1}}H_{m_1}^{(v_1)}\Big(\frac{x-x_i}{h_1}\Big).$$

Then

$$B_{ij} = \frac{1}{n_2h_2}\frac{1}{h_2^{v_2}}H_{m_2}^{(v_2)}\Big(\frac{z-z_i}{h_2}\Big)\Big[\frac{1}{n_1h_1}\delta_{d_1>m_1}O_{11} + \frac{1}{n_1h_1}\frac{(-\phi_2)^{v_1}}{h_1^{v_1}}\exp\Big(-\phi_2\frac{|x-x_i|}{h_1}\Big)O_{12}\Big].$$

It can be shown that

$$\sum_{i,j}\Big|\frac{1}{h_2^{v_2}}\frac{1}{n_1h_1}\frac{1}{n_2h_2}H_{m_2}^{(v_2)}\Big(\frac{z-z_i}{h_1}\Big)\delta_{d_1>m_1}O_{21}\mu(x_i,z_i)\Big| = O(\zeta h_2^{-v_2}) = O(\zeta h_1^{-v_1}h_2^{-v_2}),$$

and

$$\sum_{i,j}\Big|\frac{1}{h_2^{v_2}}\frac{(-\phi_2)^{v_1}}{h_1^{v_1}}\frac{1}{n_1h_1}\frac{1}{n_2h_2}H_{m_2}^{(v_2)}\Big(\frac{z-z_i}{h_2}\Big)\exp\Big(-\phi_2\frac{|x-x_i|}{h_1}\Big)O_{12}\mu(x_i,z_i)\Big| = O(\zeta h_1^{-v_1}h_2^{-v_2})$$

Therefore

$$\sum_{i,j}\Big|B_{ij}\mu(x_i,z_i)\Big| = O(\zeta h_1^{-v_1}h_2^{-v_2}) \tag{A1.9}$$

For the term with C_{ij}:

$$\begin{aligned} d_{i,1}^{(v_1)}(x)d_{i,2}^{(v_2)}(z) =& \Big[\frac{1}{n_1h_1}\delta_{d_1>m_1}O_{11} + \frac{1}{n_1h_1}\frac{(-\phi_2)^{v_1}}{h_1^{v_1}}\exp\Big(-\phi_2\frac{|x-x_i|}{h_1}\Big)O_{12}\Big] \\ &\times\Big[\frac{1}{n_2h_2}\delta_{d_2>m_2}O_{21} + \frac{1}{n_2h_2}\frac{(-\phi_4)^{v_2}}{h_2^{v_2}}\exp\Big(-\phi_4\frac{|z-z_i|}{h_2}\Big)O_{22}\Big]. \end{aligned}$$

It is shown that

$$\sum_{i,j}\Big|\frac{1}{n_1h_1}\frac{1}{n_2h_2}\delta_{d_1>m_1}O_{11}\delta_{d_2>m_2}O_{21}\mu(x_i,z_i)\Big| = O(\zeta^2) = O(\zeta) = O(\zeta h_1^{-v_1}h_2^{-v_2}),$$

$$\sum_{i,j}\Big|\frac{(-\phi_4)^{v_2}}{h_2^{v_2}}\frac{1}{n_1h_1}\frac{1}{n_2h_2}\delta_{d_1>m_1}O_{11}\exp\Big(-\phi_4\frac{|z-z_i|}{h_2}\Big)O_{22}\mu(x_i,z_i)\Big| = O(\zeta^2h_2^{-v_2}) = O(\zeta h_1^{-v_1}h_2^{-v_2}),$$

$$\sum_{i,j}\Big|\frac{(-\phi_2)^{v_1}}{h_1^{v_1}}\frac{1}{n_1h_1}\frac{1}{n_2h_2}\exp\Big(-\phi_2\frac{|x-x_i|}{h_1}\Big)O_{12}\delta_{d_2>m_2}O_{21}\mu(x_i,z_i)\Big| = O(\zeta^2h_1^{-v_1}) = O(\zeta h_1^{-v_1}h_2^{-v_2}),$$

and

$$\begin{aligned} &\sum_{i,j}\Big|\frac{(-\phi_2)^{v_1}}{h_2^{v_1}}\frac{(-\phi_4)^{v_2}}{h_2^{v_2}}\frac{1}{n_1h_1}\frac{1}{n_2h_2}\exp\Big(-\phi_2\frac{|x-x_i|}{h_1}\Big)O_{12}\exp\Big(-\phi_4\frac{|z-z_i|}{h_2}\Big)O_{22}\mu(x_i,z_i)\Big| \\ &= O(\zeta^2h_1^{-v_1}h_2^{-v_2}) \\ &= O(\zeta h_1^{-v_1}h_2^{-v_2}). \end{aligned}$$

Thus

$$\sum_{i,j}\Big|C_{ij}\mu(x_i,z_i)\Big| = O(\zeta h_1^{-v_1}h_2^{-v_2}) \tag{A1.10}$$

Combining equations (A1.2), (A1.3), (A1.5), (A1.8), (A1.9), (A1.10), it has been showed that

$$E\Big\{\hat{\mu}^{(v_1,v_2)}(x,z)-\mu^{*(v_1,v_2)}(x,z)\Big\}=\sum_{i,j}b_{i,j}^{(v_1,v_2)}(x,z)\mu(x_i,z_i)=O(\zeta h_1^{-v_1}h_2^{-v_2}).$$

Now to show $Var\Big\{\hat{\mu}^{(v_1,v_2)}(x,z)-\mu^{*(v_1,v_2)}(x,z)\Big\}=o\Big(\big(nh\big)^{-1}h_1^{-2v_1}h_2^{-2v_2}\Big)$:

$$\begin{aligned}Var\Big\{\hat{\mu}^{(v_1,v_2)}(x,z)-\mu^{*(v_1,v_2)}(x,z)\Big\}&=\sum_{i,j}\big\{b_{i,j}^{(v_1,v_2)}(x,z)\big\}^2\sigma^2(x_i,z_i)\\&=\sum_{i,j}(A_{ij}+B_{ij}+C_{ij}+D_{ij})^2\sigma^2(x_i,z_i),\end{aligned}$$

where $A_{ij},B_{ij},C_{ij},$ and D_{ij} are defined in equation (A1.2).

Here only $\sum\limits_{i,j}A_{ij}^2\sigma^2(x_i,z_i)=o\Big(\big(nh\big)^{-1}h_1^{-2v_1}h_2^{-2v_2}\Big)$ is shown, the rest of the terms have the same or smaller orders.

$$\begin{aligned}A_{ij}^2&=\frac{1}{(n_1h_1)^2}\frac{1}{h_1^{2v_1}}\bigg\{H_{m_1}^{(v_1)}\bigg(\frac{x-x_i}{h_1}\bigg)\bigg\}^2\bigg[\frac{1}{n_2h_2}\delta_{d_2>m_2}O_{21}+\frac{1}{n_2h_2}\frac{(-\phi_4)^{v_2}}{h_2^{v_2}}\exp\bigg(-\phi_4\frac{|z-z_i|}{h_2}\bigg)O_{22}\bigg]^2\\&=\underbrace{\frac{1}{h_1^{2v_1}}\frac{1}{(n_1h_1)^2}\frac{1}{(n_2h_2)^2}\bigg\{H_{m_1}^{(v_1)}\bigg(\frac{x-x_i}{h_1}\bigg)\bigg\}^2\delta_{d_2>m_2}O_{21}^2}_{\widehat{=}E_{1,ij}}\\&+\underbrace{\frac{1}{h_1^{2v_1}}\frac{1}{h_2^{2v_2}}\frac{1}{(n_1h_1)^2}\frac{(\phi_4)^{2v_2}}{(n_2h_2)^2}\bigg\{H_{m_1}^{(v_1)}\bigg(\frac{x-x_i}{h_1}\bigg)\bigg\}^2\exp\bigg(-2\phi_4\frac{|z-z_i|}{h_2}\bigg)O_{22}^2}_{\widehat{=}E_{2,ij}}\\&+\underbrace{\frac{1}{h_1^{2v_1}}\frac{1}{(n_1h_1)^2}\frac{2(-\phi_4)^{v_2}}{(n_2h_2)^2}\bigg\{H_{m_1}^{(v_1)}\bigg(\frac{x-x_i}{h_1}\bigg)\bigg\}^2\delta_{d_2>m_2}O_{21}\exp\bigg(-\phi_4\frac{|z-z_i|}{h_2}\bigg)O_{22}}_{\widehat{=}E_{3,ij}}.\end{aligned}$$

It can be shown that

$$\sum_{i,j}\Big|E_{1,ij}\sigma^2(x_i,z_i)\Big| = \frac{1}{h_1^{2v_1}}\frac{1}{nh}o\Big(\zeta\Big) = o\Big((nh)^{-1}h_1^{-2v_1}h_2^{-2v_2}\Big),$$

$$\sum_{i,j}\Big|E_{2,ij}\sigma^2(x_i,z_i)\Big| = \frac{1}{h_1^{2v_1}}\frac{1}{h_2^{2v_2}}\frac{1}{nh}o\Big(\zeta\Big) = o\Big((nh)^{-1}h_1^{-2v_1}h_2^{-2v_2}\Big),$$

$$\sum_{i,j}\Big|E_{3,ij}\sigma^2(x_i,z_i)\Big| = \frac{1}{h_1^{2v_1}}\frac{1}{nh}o\Big(\zeta^2\Big) = o\Big((nh)^{-1}h_1^{-2v_1}h_2^{-2v_2}\Big).$$

Therefore $\sum_{i,j}\Big|A_{ij}^2\sigma^2(x_i,z_i)\Big| = o\Big((nh)^{-1}h_1^{-2v_1}h_2^{-2v_2}\Big)$. The similar arguments can be used to show the orders of the other terms, so they are skipped and it is concluded that

$$Var\Big\{\hat{\mu}^{(v_1,v_2)}(x,z) - \mu^{*(v_1,v_2)}(x,z)\Big\} = o\Big((nh)^{-1}h_1^{-2v_1}h_2^{-2v_2}\Big).$$

A2: Proof of Theorem 1

By the results in lemma 1,

$$\hat{\mu}^{(v_1,v_2)}(x,z) - \mu^{*(v_1,v_2)}(x,z) = o\bigg(\frac{1}{\sqrt{nh}}\frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\bigg).$$

Write

$$\begin{aligned}
&\hat{\mu}^{(v_1,v_2)}(x,z) - \mu^{(v_1,v_2)}(x,z)\\
=&\hat{\mu}^{(v_1,v_2)}(x,z) + \mu^{*(v_1,v_2)}(x,z) - \mu^{*(v_1,v_2)}(x,z) - \mu^{(v_1,v_2)}(x,z)\\
=&o\bigg(\frac{1}{\sqrt{nh}}\frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\bigg) + \frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\frac{1}{n_1h_1}\frac{1}{n_2h_2}\sum_{ij}H_{m_1}^{(v_1)}\bigg(\frac{x-x_i}{h_1}\bigg)H_{m_2}^{(v_2)}\bigg(\frac{z-z_i}{h_2}\bigg)\varepsilon_{ij}\\
+&\frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\frac{1}{n_1h_1}\frac{1}{n_2h_2}\sum_{ij}H_{m_1}^{(v_1)}\bigg(\frac{x-x_i}{h_1}\bigg)H_{m_2}^{(v_2)}\bigg(\frac{z-z_i}{h_2}\bigg)\mu(x_i,z_i) - \mu^{(v_1,v_2)}(x,z).
\end{aligned}$$

By Lemma 4 of Xiao et al. (2012), define

$$\mu_0^{(v_1,v_2)}(x,z) = \frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\frac{1}{n_1 h_1}\frac{1}{n_2 h_2}\sum_{ij} H_{m_1}^{(v_1)}\left(\frac{x-x_i}{h_1}\right)H_{m_2}^{(v_2)}\left(\frac{z-z_i}{h_2}\right)\mu(x_i,z_i)$$

$$- \frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\frac{1}{h_1}\frac{1}{h_2}\int\int H_{m_1}^{(v_1)}\left(\frac{x-u}{h_1}\right)H_{m_2}^{(v_2)}(v)\mu(u,v)dudv$$

$$= \frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}O\left\{\max\left(\frac{1}{n_1^2 h_1^2},\frac{1}{n_2^2 h_2^2}\right)\right\} = \frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}o(h_1^{2m_1}) = o\left(\frac{1}{\sqrt{nh}}\frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\right).$$

Since proposition 3.4 of Messer and Goldstein (1993) is crucial for the following derivation, it is stated here.

Proposition 3.4 of Messer and Goldstein (1993):

$$\int_{-\infty}^{\infty} t^g H_m^{(v)}(t)dt = \begin{cases} 0 & 0 \le g < 2m+v, g \ne v \\ (-1)^v v! & g = v \\ (-1)^{m+v+1}(2m+v)! & g = 2m+v \end{cases}$$

Using Taylor expansion and the above properties of $H_m(t)$, it can be shown

$$
\begin{aligned}
&\frac{1}{h_1}\frac{1}{h_2}\int\int H_{m_1}^{(v_1)}\Big(\frac{x-u}{h_1}\Big)H_{m_2}^{(v_2)}\Big(\frac{z-v}{v}\Big)\mu(u,v)dudv\\
=&\int\int H_{m_1}^{(v_1)}\big(\tilde{u}\big)H_{m_2}^{(v_2)}\big(\tilde{v}\big)\mu(x-h_1\tilde{u},z-h_2\tilde{v})d\tilde{u}d\tilde{v}\\
=&\int\int H_{m_1}^{(v_1)}\big(\tilde{u}\big)H_{m_2}^{(v_2)}\big(\tilde{v}\big)\Big[\sum_{g_1=0}^{\infty}\sum_{g_2=0}^{\infty}\frac{1}{g_1!g_2!}\mu^{(g_1,g_2)}(x,z)(-h_1\tilde{u})^{g_1}(-h_2\tilde{v})^{g_2}\Big]d\tilde{u}d\tilde{v}\\
=&\frac{\mu^{(v_1,v_2)}(x,z)}{v_1!v_2!}\int\int H_{m_1}^{(v_1)}\big(\tilde{u}\big)H_{m_2}^{(v_2)}\big(\tilde{v}\big)(-h_1\tilde{u})^{v_1}(-h_2\tilde{v})^{v_2}d\tilde{u}d\tilde{v}\\
+&\frac{\mu^{(v_1,2m_2+v_2)}(x,z)}{v_1!(2m_2+v_2)!}\int\int H_{m_1}^{(v_1)}\big(\tilde{u}\big)H_{m_2}^{(v_2)}\big(\tilde{v}\big)(-h_1\tilde{u})^{v_1}(-h_2\tilde{v})^{2m_2+v_2}d\tilde{u}d\tilde{v}\\
+&\frac{\mu^{(2m_1+v_1,v_2)}(x,z)}{(2m_1+v_1)!v_2!}\int\int H_{m_1}^{(v_1)}\big(\tilde{u}\big)H_{m_2}^{(v_2)}\big(\tilde{v}\big)(-h_1\tilde{u})^{2m_1+v_1}(-h_2\tilde{v})^{v_2}d\tilde{u}d\tilde{v}\\
+&\frac{\mu^{(2m_1+v_1,2m_2+v_2)}(x,z)}{(2m_1+v_1)!(2m_2+v_2)!}\int\int H_{m_1}^{(v_1)}\big(\tilde{u}\big)H_{m_2}^{(v_2)}\big(\tilde{v}\big)(-h_1\tilde{u})^{2m_1+v_1}(-h_2\tilde{v})^{2m_2+v_2}d\tilde{u}d\tilde{v}\\
+&o\Big(h_1^{2m_1+v_1}h_2^{2m_2+v_2}\Big)\\
=&\mu^{(v_1,v_2)}(x,z)h_1^{v_1}h_2^{v_2}+\mu^{(v_1,2m_2+v_2)}(x,z)(-1)^{m_2+1}h_1^{v_1}h_2^{2m_2+v_2}\\
+&\mu^{(2m_1+v_1,v_2)}(x,z)(-1)^{m_1+1}h_1^{2m_1+v_1}h_2^{v_2}\\
+&\mu^{(2m_1+v_1,v_2)}(x,z)(-1)^{m_1+1}(-1)^{m_2+1}h_1^{2m_1+v_1}h_2^{2m_2+v_2}+o\Big(h_1^{2m_1+v_1}h_2^{2m_2+v_2}\Big).
\end{aligned}
$$

So

$$
\begin{aligned}
&\frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\frac{1}{h_1}\frac{1}{h_2}\int\int H_{m_1}^{(v_1)}\Big(\frac{x-u}{h_1}\Big)H_{m_2}^{(v_2)}\Big(\frac{z-v}{h_2}\Big)\mu(u,v)dudv\\
=&\mu^{(v_1,v_2)}(x,z)+(-1)^{m_2+1}h_2^{2m_2}\mu^{(v_1,m_2+v_2)}(x,z)+(-1)^{m_1+1}h_1^{2m_1}\mu^{(m_1+v_1,v_2)}(x,z)\\
+&o\Big(h_1^{2m_1}\Big)+o\Big(h_2^{2m_2}\Big).
\end{aligned}
$$

(A2.1)

It is easy to verify that $o\Big(h_1^{2m_1}\Big) = o\Big(\frac{1}{\sqrt{nh}}\Big) = o\Big(\frac{1}{\sqrt{nh}}\frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\Big)$, and $o\Big(h_2^{2m_2}\Big)=$ $o\Big(\frac{1}{\sqrt{nh}}\Big)$
$= o\Big(\frac{1}{\sqrt{nh}}\frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\Big)$.

Thus

$$\hat{\mu}^{(v_1,v_2)}(x,z) - \mu^{(v_1,v_2)}(x,z) = \frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\frac{1}{nh}\sum_{ij} H_{m_1}^{(v_1)}\Big(\frac{x-x_i}{h_1}\Big)H_{m_2}^{(v_2)}\Big(\frac{z-z_i}{h_2}\Big)\varepsilon_{ij}$$
$$+ \mu_0^{(v_1,v_2)}(x,z) + \mu_b(x,z) + o\Big(\frac{1}{\sqrt{nh}}\frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\Big).$$

where $\mu_b(x,z) = (-1)^{m_2+1}h_2^{2m_2}\mu^{(v_1,m_2+v_2)}(x,z) + (-1)^{m_1+1}h_1^{2m_1}\mu^{(m_1+v_1,v_2)}(x,z)$.

Therefore,

$$\sqrt{nh}h_1^{v_1}h_2^{v_2}\Big(\hat{\mu}^{(v_1,v_2)}(x,z) - \mu^{(v_1,v_2)}(x,z) - \mu_b(x,z)\Big) \stackrel{d}{=} \frac{1}{\sqrt{nh}}\sum_{ij} H_{m_1}^{(v_1)}\Big(\frac{x-x_i}{h_1}\Big)H_{m_2}^{(v_2)}\Big(\frac{z-z_i}{h_2}\Big)\varepsilon_{ij}.$$

By an argument similar to one in the proof of lemma 1,

$$Var\Big(\frac{1}{\sqrt{nh}}\sum_{ij} H_{m_1}^{(v_1)}\Big(\frac{x-x_i}{h_1}\Big)H_{m_2}^{(v_2)}\Big(\frac{z-z_i}{h_2}\Big)\varepsilon_{ij}\Big)$$
$$= \frac{1}{nh}\sum_{ij}\Big\{H_{m_1}^{(v_1)}\Big(\frac{x-x_i}{h_1}\Big)\Big\}^2\Big\{H_{m_2}^{(v_2)}\Big(\frac{z-z_i}{h_2}\Big)\Big\}^2\sigma^2(x_i,z_i)$$
$$= \sigma^2(x,z)\int\int\Big\{H_{m_1}^{(v_1)}\Big(\frac{x-u}{h_1}\Big)\Big\}^2\Big\{H_{m_2}^{(v_2)}\Big(\frac{z-v}{h_2}\Big)\Big\}^2 du\ dv + o(1)$$
$$= V(x,z) + o(1)$$

A3: Proof of Theorem 2

Denote the design points $\{x_i, z_i\}_{i=1}^n$ by $(\underline{x},\underline{z})$. Apply Theorem 1 to the binned data $\tilde{Y}$ with n_1, n_2 replaced by S_1, S_2,

$$E\Big\{\tilde{\mu}^{(v_1,v_2)}(x,z)|\big(\underline{x},\underline{z}\big)\Big\} = \frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\frac{1}{S_1h_1}\frac{1}{S_2h_2}\sum_{i',j'} G_{i',j'}E\Big\{\tilde{y}_{i',j'}|\big(\underline{x},\underline{z}\big)\Big\}$$

$$Var\Big\{\tilde{\mu}^{(v_1,v_2)}(x,z)|\big(\underline{x},\underline{z}\big)\Big\} = \frac{1}{h_1^{2v_1}}\frac{1}{h_2^{2v_2}}\frac{1}{(S_1h_1)^2}\frac{1}{(S_2h_2)^2}\sum_{i',j'} G_{i',j'}^2 Var\Big\{\tilde{y}_{i',j'}|\big(\underline{x},\underline{z}\big)\Big\}$$

where $G_{i',j'} = H_{m_1}^{(v_1)}\left(\frac{x-\tilde{x}_{i'}}{h_1}\right)H_{m_2}^{(v_2)}\left(\frac{z-\tilde{z}_{j'}}{h_2}\right) + Shh_1^{v_1}h_2^{v_2}b_{i',j'}^{(v_1,v_2)}(x,z)$, $S = S_1 \times S_2$, and $h = h_1 \times h_2$.

By Theorem 2 of Xiao et al. (2012),

$$\sup_{k,l}\left|\frac{n}{S}Var\{\tilde{y}_{i',j'}|\big(\underline{x},\underline{z}\big)\} - \frac{\sigma^2(\tilde{x}_{i'},\tilde{z}_{j'})}{f(\tilde{x}_{i'},\tilde{z}_{j'})}\right| = o_p(1).$$

So

$$Var\Big\{\tilde{\mu}^{(v_1,v_2)}(x,z)|\big(\underline{x},\underline{z}\big)\Big\} = \frac{1}{h_1^{2v_1}}\frac{1}{h_2^{2v_2}}\frac{1}{(Sh)^2}\frac{S}{h}\sum_{i',j'}\frac{n}{S}Var\Big\{\tilde{y}_{i',j'}|\big(\underline{x},\underline{z}\big)\Big\}G_{i',j'}^2,$$

and

$$\left|Var\Big\{\tilde{\mu}^{(v_1,v_2)}(x,z)|\big(\underline{x},\underline{z}\big)\Big\} - \frac{1}{Sh}\frac{S}{nh}\frac{1}{h_1^{2v_1}}\frac{1}{h_2^{2v_2}}\sum_{i',j'}\frac{\sigma^2(\tilde{x}_{i'},\tilde{z}_{j'})}{f(\tilde{x}_{i'},\tilde{z}_{j'})}G_{k,l}^2\right| = \frac{1}{Sh}\frac{S}{nh}\frac{1}{h_1^{2v_1}}\frac{1}{h_2^{2v_2}}\sum_{k,l}o_p(1)G_{i',j'}^2$$

$$= o_p\left(\frac{S}{nh}\frac{1}{h_1^{2v_1}}\frac{1}{h_2^{2v_2}}\right).$$

And because

$$\frac{1}{S_1 h_1}\frac{1}{S_2 h_2}\sum_{i',j'}\frac{\sigma^2(\tilde{x}_{i'},\tilde{z}_{j'})}{f(\tilde{x}_{i'},\tilde{z}_{j'})} = \frac{\sigma^2(x,z)}{f(x,z)}\int\int H_{m_1}^{(v_1)}(u)H_{m_2}^{(v_2)}(v)dudv + o_p(1)$$

$$= \frac{V(x,z)}{f(x,z)} + o_p(1),$$

then

$$Var\Big\{\tilde{\mu}^{(v_1,v_2)}(x,z)|\big(\underline{\mathrm{x}},\underline{\mathrm{z}}\big)\Big\} = \frac{S}{nh}\frac{1}{h_1^{2v_1}}\frac{1}{h_2^{2v_2}}\frac{V(x,z)}{f(x,z)} + o_p\Big(\frac{S}{nh}\frac{1}{h_1^{2v_1}}\frac{1}{h_2^{2v_2}}\Big). \tag{A3.1}$$

Similarly, by Theorem 2 of Xiao et al. (2012),

$$\sup_{i',j'}\Big|E\Big\{\tilde{y}_{i',j'}|\big(\underline{\mathrm{x}},\underline{\mathrm{z}}\big)\Big\} - \mu(\tilde{x}_{i'},\tilde{z}_{j'})\Big| = O_p\Big(S^{-\frac{1}{2}}\Big),$$

then

$$\begin{aligned}\Big|E\Big\{\tilde{\mu}^{(v_1,v_2)}(x,z)|\big(\underline{\mathrm{x}},\underline{\mathrm{z}}\big)\Big\} - \frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\frac{1}{Sh}\sum_{i',j'}\mu\Big(\tilde{x}_{i'},\tilde{z}_{j'}\Big)\Big| &= \frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\frac{1}{Sh}\sum_{i',j'}O_p\Big(S^{-\frac{1}{2}}\Big)|G_{i',j'}| \\ &= O_p\Big(S^{-\frac{1}{2}}\frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\Big).\end{aligned} \tag{A3.2}$$

It is easy to show

$$\frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\frac{1}{Sh}\sum_{i',j'}\mu\Big(\tilde{x}_{i'},\tilde{z}_{j'}\Big)G_{i',j'} = \mu^{(v_1,v_2)}(x,z) + \mu_b(x,z) + o_p\Big(\frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\frac{1}{\sqrt{nh}}\Big). \tag{A3.3}$$

Combining (A3.1) and (A3.2),

$$E\Big\{\tilde{\mu}^{(v_1,v_2)}(x,z)|\big(\underline{\mathrm{x}},\underline{\mathrm{z}}\big)\Big\} - \mu^{(v_1,v_2)}(x,z) = O_p\Big(S^{-\frac{1}{2}}\frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\Big) + \mu_b(x,z) + o_p\Big(\frac{1}{h_1^{v_1}}\frac{1}{h_2^{v_2}}\frac{1}{\sqrt{nh}}\Big).$$

Because $h_1 \sim n^{-\frac{m_2}{m_3}}$, $h_2 \sim n^{-\frac{m_1}{m_3}}$, and $S = n^{\tau}$ where $\tau > \dfrac{4m_1m_2}{m_3}$, it can be shown that $nh = o\Big(n^{\frac{4m_1m_2}{m_3}}\Big)$, thus $O_p\Big(\dfrac{1}{\sqrt{S}}\dfrac{1}{h_1^{v_1}}\dfrac{1}{h_2^{v_2}}\Big) = o_p\Big(\dfrac{1}{h_1^{v_1}}\dfrac{1}{h_2^{v_2}}\dfrac{1}{\sqrt{nh}}\Big)$.

Then

$$
\begin{aligned}
&\sqrt{nh}h_1^{v_1}h_2^{v_2}\Big(\tilde{\mu}^{(v_1,v_2)}(x,z)-\mu^{(v_1,v_2)}(x,z)-\mu_b(x,z)\Big)\\
=&\sqrt{nh}h_1^{v_1}h_2^{v_2}\left[\tilde{\mu}^{(v_1,v_2)}(x,z)-E\Big\{\tilde{\mu}^{(v_1,v_2)}(x,z)|\big(\underline{\mathrm{x}},\underline{\mathrm{z}}\big)\Big\}\right]\\
+&\sqrt{nh}h_1^{v_1}h_2^{v_2}\left[E\Big\{\tilde{\mu}^{(v_1,v_2)}(x,z)|\big(\underline{\mathrm{x}},\underline{\mathrm{z}}\big)\Big\}-\mu^{(v_1,v_2)}(x,z)-\mu_b(x,z)\right]\\
\overset{d}{=}&\sqrt{nh}h_1^{v_1}h_2^{v_2}\left[\tilde{\mu}^{(v_1,v_2)}(x,z)-E\Big\{\tilde{\mu}^{(v_1,v_2)}(x,z)|\big(\underline{\mathrm{x}},\underline{\mathrm{z}}\big)\Big\}\right]\\
\sim&N\left(0,\frac{V(x,z)}{f(x,z)}\right).
\end{aligned}
$$

The asymptotic distribution $N\left(0,\dfrac{V(x,z)}{f(x,z)}\right)$ is obtained by equation (A3.1).

APPENDIX B. CHAPTER 4 APPENDIX

The regularity conditions of Theorem 1 in Chapter 4 are presented in B1. The proof of Theorem 1 is found in B2. B3 provides a sketch of proof to show that $\hat{\boldsymbol{\theta}}_g^{(2)}$ is the most efficient estimator among the class of estimators $\hat{\boldsymbol{\theta}}_g^a$ that use any fixed positive definite matrix $\mathbf{A}$ in the quadratic form minimization

B1: Regularity Conditions of Theorem 1

The notation of $|\cdot|$ represents the norm of a matrix, defined as $|A| = \sqrt{trace(A'A)}$ and the notation of $\|\cdot\|$ denotes the sup-norm in all arguments for functions.

Regular conditions on the sample designs in both phases are given here. The following notations, I_i, π_i and π_{ij}, denote the sampling indicator, the first and second inclusion probabilities either for the first-phase design or for the second-phase design. For example, $I_i = \delta_{1i}$ or $I_i = \delta_{2ig}$ for any g, and $\pi_i = \pi_{1i}$ or $\pi_i = \pi_{2ig}$ for any g, depending on whether the design if the first-phase design or the second-phase design.

Condition I:

(1) Any variable v_i such that $E[|v_i|^{2+\delta}] < \infty$, where $\delta > 0$, satisfies $\sqrt{n}(\bar{v}_{HT} - \bar{v}_N)|\mathcal{F}_N \xrightarrow{\mathcal{L}} N(0, V_\infty)$ *a.s.* , where $(\bar{v}_{HT}, \bar{v}_N) = N^{-1}\sum_{i=1}^N(\pi^{-1}v_iI_i, v_i)$, $V_\infty = lim_{N\to\infty}V_N$, and $V_N = nV(\bar{v}_{HT}|\mathcal{F}_N)$ is the conditional variance of the Horvitz-Thompson estimator, $\bar{v}_{HT}$, given $\mathcal{F}_N$.

(2) $nN^{-1} \to f_\infty \in [0, 1]$.

(3) There exist constant C_1, C_2 and C_3 such that $0 < C_1 \leq nN^{-1}\pi_i^{-1} < \infty$, and $|n(\pi_{ij} - \pi_i\pi_j)\pi_i^{-1}\pi_j^{-1}| \leq C_3 < \infty$ *a.s.*

Condition I(1) and I(2) are regular conditions assumed for a survey design in a finite population framework. Condition I(3) is used in Fuller (2009). The part of condition I(3) related to the joint selection probabilities is used in the proofs to bound sums of

covariance induced by the sample design. Condition I(3) holds for simple random sampling, where $(\pi_{ij} - \pi_i\pi_j)\pi_i^{-1}\pi_j^{-1} = n^{-1}(n-1)(N-1)^{-1}N - 1$, and for Poisson sampling, where $(\pi_{ij} - \pi_i\pi_j)\pi_i^{-1}\pi_j^{-1} = 0$, and can hold for cluster sampling and stratified sampling. Fuller (2009) explains that a designer has the control to ensure condition I(3). Note that for the second-phase design in this situation, $(\pi_{2ij,g} - \pi_{2ig}\pi_{2jg})\pi_{2ig}^{-1}\pi_{2jg}^{-1} = 0$ for any g because the second-phase design is a multinomial extension of Poisson sampling.

Next regular conditions on the tuning parameters of the semiparametric basis are given. For simplicity, consider the special case of power series and spline series.

Condition II:

(1) The smallest eigenvalue of $E[R_K(X_i)R_K(X_i)']$ is bounded away from zero uniformly in K.

(2) There exists a sequence of constant $\zeta(K)$ such that $\|R_K(X_i)\| \leq \zeta(K)$ for $K \to \infty$ and $\zeta(K)K^{1/2}n^{-1/2} \to 0$.

(3) For all g, $\pi_{2ig}(X_i)$ and $\boldsymbol{\mu}_{mg}(X_i, \boldsymbol{\theta}_g) = E[\mathbf{m}_{ig}(\boldsymbol{\theta}_g)|X_i]$ are s-time differentiable with $sd_x^{-1} \geq 5\eta/2 + 1/2$, where d_x is the dimension of X_i, and $\eta = 1$ or $\eta = 1/2$ depending on whether power series or spline series are used as basis function.

(4) $K = O(n^\nu)$ with $4sd_x^{-1} - 6\eta \geq \nu^{-1} \geq 4\eta + 2$, where $\eta = 1$ or $\eta = 1/2$ depending on whether power series or spline series are used as basis function.

Condition II(1) and II(2) are standard assumptions and are automatically satisfied in the case of power series or spline series. Condition II(3) and II(4) describe the minimum smoothness required as a function of the dimension of X and the choice of basis, and the relationship between the sample size and the number of bases. Under II(3) and II(4), by Lorentz (1986), there exists a K-vector $\boldsymbol{\gamma}_{g,K}^*$ for any g such that

$$\left\| log(\frac{\pi_{2ig}(X)}{1 - \sum_{g=2}^{G}\pi_{2ig}(X)}) - R_K^T(X)\boldsymbol{\gamma}_{g,K}^* \right\| = O(K^{-\frac{s}{\nu}}), \tag{B1.1}$$

where $R_K^T(X)\boldsymbol{\gamma}_{g,K}^*$ is the best L_∞ approximation for the logarithm of the odds ratio of treatment g to the base treatment. The property (B1.1) is used to derive the convergence

rate of $\hat{\pi}_{2ig}$ to π_{2ig} as follows,

$$\|\hat{\pi}_{2ig} - \pi_{2ig}\| = O_p(\xi(K)K^{1/2}n^{-1/2} + \xi(K)K^{1/2}K^{-s/d_x}) = o_p(1). \tag{B1.2}$$

For details, see Theorem B-1 of Cattaneo (2010).

The regular conditions on the estimation equation function $\mathbf{m}_{ig}(Y_{ig}, Z_i; \boldsymbol{\theta}_g)$ are as follows.

Condition III:

(1) $\mathbf{m}_{ig}(Y_{ig}, Z_i; \boldsymbol{\theta}_g)$ is differentiable with respect to $\boldsymbol{\theta}_g$.

(2) Both $\mathbf{m}_{ig}(Y_{ig}, Z_i; \boldsymbol{\theta}_g)$ and its first derivative with respect to $\boldsymbol{\theta}_g$ have bounded $2 + \delta$ moments. More specifically, $E[|h(Y_i, Z_i; \boldsymbol{\theta})|^{2+\delta}] < M$, where $h(Y_i, Z_i; \boldsymbol{\theta})$ denote an element of $\mathbf{m}_{ig}(Y_{ig}, Z_i; \boldsymbol{\theta}_g)$ or an element of its first derivative with respect to $\boldsymbol{\theta}_g$.

(3) $V_\xi(\mathbf{M}_{x,i}(\boldsymbol{\theta}))$ is a positive definite matrix uniformly in $\boldsymbol{\theta}$, where

$$\mathbf{M}_{x,i} = [X_i, \mathbf{m}_1(Y_{i1}, X_i; \boldsymbol{\theta}_1), ..., \mathbf{m}_G(Y_{iG}, X_i; \boldsymbol{\theta}_1)]^T.$$

(4) $\Gamma_g(\boldsymbol{\theta}_g^0)$ is full rank.

(5) Assume that $\bar{h}_{HT}(\boldsymbol{\theta}) - \bar{h}_N(\boldsymbol{\theta})$ converges to 0 uniformly in $\boldsymbol{\theta}$, where $\bar{h}_{HT}(\boldsymbol{\theta}) = N^{-1}\sum_{i=1}^N I_i \pi_i^{-1} h_i(Y_i, Z_i; \boldsymbol{\theta})$, $\bar{h}_N(\boldsymbol{\theta}) = N^{-1}\sum_{i=1}^N h_i(Y_i, Z_i; \boldsymbol{\theta})$, and $h_i(Y_i, Z_i; \boldsymbol{\theta})$ has the same interpretation as in condition III(2) above. This condition means that for all $\epsilon > 0$, there exists a $\delta > 0$ such that $Pro(|\bar{h}_{HT}(\boldsymbol{\theta}) - \bar{h}_N(\boldsymbol{\theta})| > \epsilon) < \delta$, for all N greater than some value M, and for all $\boldsymbol{\theta}$.

B2: Proof of Theorem 1

The proof of Theorem 1 proceeds in two steps. The first step is to show that the asymptotic equivalence of $\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)$,

$$\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g) = \frac{1}{N}\sum_{i\in U} \frac{\delta_{1i}\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}} - \frac{1}{N}\frac{\delta_{1i}(\delta_{2ig} - \pi_{2ig})}{\pi_{1i}\pi_{2ig}}\boldsymbol{\mu}_{mg}(X_i; \boldsymbol{\theta}_g) + o_p(n^{-1/2}), \tag{B2.1}$$

where $\boldsymbol{\mu}_{mg}(X_i;\boldsymbol{\theta}_g) = E_\xi((\mathbf{m}_{ig}(\boldsymbol{\theta}_g)|X_i)$. In order to show (B2.1), first decompose $\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)$ into

$$\begin{aligned}
\tfrac{1}{N}\textstyle\sum_{i\in A_{2g}} \frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\hat{\pi}_{2ig}} &= \tfrac{1}{N}\textstyle\sum_{i\in A_1}\left\{\frac{\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\hat{\pi}_{2ig}} - \frac{\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}} + \frac{\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}^2}(\hat{\pi}_{2ig}-\pi_{2ig})\right\} \\
&+ \tfrac{1}{N}\textstyle\sum_{i\in A_1}\left\{-\frac{\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}^2}(\hat{\pi}_{2ig}-\pi_{2ig}) + \frac{\boldsymbol{\mu}_{mg}(X_i;\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}}(\hat{\pi}_{2ig}-\pi_{2ig})\right\} \\
&+ \tfrac{1}{N}\textstyle\sum_{i\in A_1}\left\{-\frac{\boldsymbol{\mu}_{mg}(X_i;\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}}(\hat{\pi}_{2ig}-\pi_{2ig}) + \frac{\boldsymbol{\mu}_{mg}(X_i;\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}}(\delta_{2ig}-\pi_{2ig})\right\} \\
&+ \tfrac{1}{N}\textstyle\sum_{i\in A_1}\left\{\frac{\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}} - \frac{\boldsymbol{\mu}_{mg}(X_i;\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}}(\delta_{2ig}-\pi_{2ig})\right\}.
\end{aligned} \tag{B2.2}$$

By the result in (B1.2), the first three terms in (B2.2) can be shown to have order $o_p(n^{-1/2})$ asymptotically, which leads to equation (B2.1). Similar arguments can be used to show $\bar{\mathbf{H}}_{2\pi g}(\boldsymbol{\theta}_g) = \frac{1}{N}\sum_{i\in A_1}\pi_{1i}^{-1}\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g) + o_p(n^{-1/2})$. The justification of those orders follows Cattaneo (2010).

The second step is to show the following two conditions of Pakes and Pollard (1989) hold: (1) $sup_{\boldsymbol{\theta}_g\in\Theta}|\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g) - E(\mathbf{m}_{ig}(\boldsymbol{\theta}_g))| = o_p(1)$, and (2) for every sequence of real numbers $\delta_n \to 0$, $sup_{|\boldsymbol{\theta}_g-\boldsymbol{\theta}_g^0|\le\delta_n}|\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g) - E(\mathbf{m}_{ig}(\boldsymbol{\theta}_g)) - \bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g^0)| = o_p(n^{-1/2})$. By equation (B2.1), it is shown that

$$\begin{aligned}
E(\bar{\mathbf{m}}_{2\pi g} - E(\mathbf{m}_g(\boldsymbol{\theta}_g))^2 &= E(\tfrac{1}{N}\textstyle\sum_{i\in U}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)\delta_{1i}\delta_{2ig}}{\pi_{1i}\pi_{2ig}} \\
&- \tfrac{1}{N}\textstyle\sum_{i\in U}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)(\delta_{2ig}-\pi_{2ig})}{\pi_{2ig}} - E(\mathbf{m}_{ig}(\boldsymbol{\theta}_g)))^2 + o(n^{-1/2}) \\
&\le 2T_{1N} + 2T_{2N} + o(n^{-1/2}),
\end{aligned} \tag{B2.3}$$

where

$$T_{1N} = E(\frac{1}{N}\sum_{i\in U}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)\delta_{1i}\delta_{2ig}}{\pi_{1i}\pi_{2ig}} - E(\mathbf{m}_{ig}(\boldsymbol{\theta}_g)))^2$$

and

$$T_{2N} = E(\frac{1}{N}\sum_{i\in U}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)(\delta_{2ig}-\pi_{2ig})}{\pi_{2ig}})^2.$$

It is easy to show $T_{1N} = O(N^{-1})$ and $T_{2N} = O(N^{-1})$.

$$\begin{aligned} T_{1N} &= Var(\tfrac{1}{N}\textstyle\sum_{i\in U}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)\delta_{1i}\delta_{2ig}}{\pi_{1i}\pi_{2ig}}) \\ &= \tfrac{1}{N}Var(X) + E(\tfrac{1}{N^2}\textstyle\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}}\frac{\mathbf{m}_{jg}^T(\boldsymbol{\theta}_g)}{\pi_{1j}}) \\ &+ E(\tfrac{1}{N^2}\textstyle\sum_{i\in U}(\frac{1}{\pi_{2ig}}-1)\frac{\delta_{1i}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)\mathbf{m}_{ig}^T(\boldsymbol{\theta}_g)}{\pi_{1i}^2}) \\ &= O(\tfrac{1}{N}) \end{aligned} \tag{B2.4}$$

$$\begin{aligned} T_{2N} &= E[\tfrac{1}{N^2}\textstyle\sum_{i\in U}\sum_{j\in U}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)(\delta_{2ig}-\pi_{2ig})}{\pi_{2ig}}\frac{\mathbf{m}_{jg}^T(\boldsymbol{\theta}_g)(\delta_{2jg}-\pi_{2jg})}{\pi_{2jg}}] \\ &= O(\tfrac{1}{N}) + E\{\tfrac{1}{N^2}\textstyle\sum_{i\neq j}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)\mathbf{m}_{jg}^T(\boldsymbol{\theta}_g)E[\frac{(\delta_{2ig}-\pi_{2ig})(\delta_{2jg}-\pi_{2jg})}{\pi_{2ig}\pi_{2jg}}|A_1]\} = O(\tfrac{1}{N}) \end{aligned} \tag{B2.5}$$

Then $E(\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g) - E(\mathbf{m}_g(\boldsymbol{\theta}_g)))^2 = O(\frac{1}{N}) \Longrightarrow \bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g) - E(\mathbf{m}_g(\boldsymbol{\theta}_g)) = o_p(1)$.

Condition (1) of Pakes and Pollard (1989) holds. Similarly, it can be shown that

$$sup_{(\boldsymbol{\theta}_g,\mu_z)}|\bar{\mathbf{H}}_{2\pi g}(\boldsymbol{\theta}_g,\mu_z) - E(\mathbf{H}_{ig}(\boldsymbol{\theta}_g,\mu_z))| = o_p(1).$$

By equation (B2.1), it can also be shown that

$\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g) - E(\mathbf{m}_g(\boldsymbol{\theta}_g)) - \bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g^0) = T_{3N} - T_{4N} + o_p(n^{-1/2})$,

where

$T_{3N} = \frac{1}{N}\sum_{i\in U}\frac{(\mathbf{m}_{ig}(\boldsymbol{\theta}_g)-\mathbf{m}_{ig}(\boldsymbol{\theta}_g^0))\delta_{1i}\delta_{2ig}}{\pi_{1i}\pi_{2ig}} - E(\mathbf{m}_{ig}(\boldsymbol{\theta}_g) - \mathbf{m}_{ig}(\boldsymbol{\theta}_g^0))$ and

$T_{4N} = \frac{1}{N}\sum_{i\in U}\frac{E[(\mathbf{m}_{ig}(\boldsymbol{\theta}_g)-\mathbf{m}_{ig}(\boldsymbol{\theta}_g^0))|X](\delta_{2ig}-\pi_{2ig})}{\pi_{2ig}}$. When $|\boldsymbol{\theta}_g - \boldsymbol{\theta}_g^0| \leq \delta_n$,

$$\begin{aligned} E(T_{3N}^2) &= \tfrac{1}{N}Var(\mathbf{m}_{ig}(\boldsymbol{\theta}_g) - \mathbf{m}_{ig}(\boldsymbol{\theta}_g^0)) \\ &+ E[\tfrac{1}{N^2}\textstyle\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)-\mathbf{m}_{ig}(\boldsymbol{\theta}_g^0)}{\pi_{1i}}\frac{\mathbf{m}_{jg}(\boldsymbol{\theta}_g)-\mathbf{m}_{jg}(\boldsymbol{\theta}_g^0)}{\pi_{1j}}] \\ &+ E[\tfrac{2}{N^2}\textstyle\sum_{i\in U}(\frac{1}{\pi_{2ig}}-1)\frac{(\mathbf{m}_{ig}(\boldsymbol{\theta}_g)-\mathbf{m}_{ig}(\boldsymbol{\theta}_g^0))^2}{\pi_{1i}}] \leq \tfrac{1}{N}O(\delta_n^2) = o(\tfrac{1}{N}) \end{aligned} \tag{B2.6}$$

$$\begin{aligned} E(T_{4N}^2) &\leq E[\tfrac{1}{N^2}\textstyle\sum_{i\in U}E(\mathbf{m}_{ig}(\boldsymbol{\theta}_g) - \mathbf{m}_{ig}(\boldsymbol{\theta}_g^0)|X)^2] \\ &\leq E\tfrac{1}{N}E[(\mathbf{m}_{ig}(\boldsymbol{\theta}_g) - \mathbf{m}_{ig}(\boldsymbol{\theta}_g^0))^2|X] \leq \tfrac{1}{N}O(|\boldsymbol{\theta}_g - \boldsymbol{\theta}_g^0|^2) = o(\tfrac{1}{N}). \end{aligned} \tag{B2.7}$$

Then $T_{3N} = o_p(n^{-1/2})$ and $T_{4N} = o_p(n^{-1/2})$ when $|\boldsymbol{\theta} - \boldsymbol{\theta}^0| \leq \delta_n$, thus Condition (2) of Pakes and Pollard (1989) is verified. Similarly, for every sequence of real numbers $\delta_n \to 0$,

$$sup_{\left\| \begin{bmatrix} \boldsymbol{\theta}_g \\ \mu_z \end{bmatrix} - \begin{bmatrix} \boldsymbol{\theta}_g^0 \\ \mu_z^0 \end{bmatrix} \right\| \leq \delta_n} |\bar{\mathbf{H}}_{2\pi g}(\boldsymbol{\theta}_g, \mu_z) - E(\mathbf{H}_{ig}(\boldsymbol{\theta}_g)) - \bar{\mathbf{H}}_{2\pi g}(\boldsymbol{\theta}_g^0, \mu_z^0)| = o_p(n^{-1/2}). \quad \text{(B2.8)}$$

For a vector $c = [c_1, c_2]^T$, it is known $|c| \leq \sqrt{2}(|c_1| + |c_2|)$. Therefore,

$$sup_{(\boldsymbol{\theta}_g, \mu_z)} |\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z) - E(\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z))| = o_p(1), \quad \text{(B2.9)}$$

and

$$sup_{\left\| \begin{bmatrix} \boldsymbol{\theta}_g \\ \mu_z \end{bmatrix} - \begin{bmatrix} \boldsymbol{\theta}_g^0 \\ \mu_z^0 \end{bmatrix} \right\| \leq \delta_n} |\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z) - E(\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)) - \mathbf{H}_{ng}(\boldsymbol{\theta}_g^0, \mu_z^0)| = o_p(n^{-1/2}). \quad \text{(B2.10)}$$

Condition (1) and (2) of Pakes and Pollard (1989) in terms of $\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)$ can be verified. The details of the proof can be obtained upon request.

B3: Proof that $\hat{\boldsymbol{\theta}}_g^{(2)}$ is the most efficient estimator

Similar proof in Theorem 1 can be used to show that the asymptotic variance of $(\hat{\boldsymbol{\theta}}_g^a, \hat{\mu}^a)$ is

$$\tilde{V}_g(\boldsymbol{\theta}_g^0, \mu_z^0) = (\Gamma_g^T(\boldsymbol{\theta}_g^0)\mathbf{A}^{-1}\Gamma_g(\boldsymbol{\theta}_g^0))^{-1}\Gamma_g^T(\boldsymbol{\theta}_g^0)\mathbf{A}^{-1}\Sigma_{Hg}^{-1}(\boldsymbol{\theta}_g^0, \mu_z)\mathbf{A}^{-1}\Gamma_g(\boldsymbol{\theta}_g^0)(\Gamma_g^T(\boldsymbol{\theta}_g^0)\mathbf{A}^{-1}\Gamma_g(\boldsymbol{\theta}_g^0))^{-1}. \quad \text{(B3.1)}$$

Comparing this asymptotic variance of $(\hat{\boldsymbol{\theta}}_g^a, \hat{\mu}^a)$ in (B3.1) with $V_g(\boldsymbol{\theta}_g^0, \mu_z^0)$ in (4.22), gives

$$\tilde{V}_g(\boldsymbol{\theta}_g^0, \mu_z^0) - V_g(\boldsymbol{\theta}_g^0, \mu_z^0) = C \cdot C^T \geq \mathbf{0}, \text{ where} \quad \text{(B3.2)}$$

$$\begin{aligned} C \;=\; & (\Gamma_g^T(\boldsymbol{\theta}_g^0)\mathbf{A}^{-1}\Gamma(\boldsymbol{\theta}_g^0))^{-1}\Gamma_g^T(\boldsymbol{\theta}_g^0)\mathbf{A}^{-1}\Sigma_{Hg}^{1/2}(\boldsymbol{\theta}_g^0, \mu_z) \\ & \times(\mathbf{I} - \Sigma_{Hg}^{-1/2}(\boldsymbol{\theta}_g^0, \mu_z)\Gamma_g(\boldsymbol{\theta}_g^0)(\Gamma_g^T(\boldsymbol{\theta}_g^0)\Sigma_{Hg}^{-1}(\boldsymbol{\theta}_g^0, \mu_z)\Gamma_g(\boldsymbol{\theta}_g^0))^{-1}\Gamma_g(\boldsymbol{\theta}_g^0)\Sigma_{Hg}^{-1/2}(\boldsymbol{\theta}_g^0, \mu_z)). \end{aligned}$$

Thus the asymptotic variance for $\hat{\boldsymbol{\theta}}^{(2)}$ should be smaller than the asymptotic variance of $\hat{\boldsymbol{\theta}}^{(1)}$.